A Data-Driven Exploration of Factors Affecting Student Performance in a Third-Level Institution

Michael G. Madden[1], William Lyons[2] and Ita Kavanagh[2]

[1] College of Engineering & Informatics,
National University of Ireland, Galway, Ireland
michael.madden@nuigalway.ie.
[2] Department of Information Technology,
Limerick Institute of Technology, Moylish Park, Limerick, Ireland
bill.lyons@lit.ie, ita.kavanagh@lit.ie.

Abstract. This paper describes an application of data mining techniques to the analysis of student academic records, collected at Limerick Institute of Technology, with the goal of acquiring clearer, evidence-based understanding of how a variety factors affect students' examination performance. To this end, a comprehensive dataset has been prepared. It has been analysed using a variety of machine learning/data mining techniques, in order to examine it from multiple perspectives. Of the techniques used, Bayesian networks have been found to be best in terms of yielding results that are comprehensible and meaningful. The results of this work provide a useful snapshot of factors affecting performance for the student group analysed, as well as demonstrating a process by which other institutions may analyse their own student groups.

1 Introduction

Academic institutions are increasingly required to monitor their performance and the performance of their students. This gives rise to a need to collate, analyse and interpret data, in order to have evidence to inform academic policies that are aimed at, for example, improving student retention rates, allocating teaching and support resources, or creating intervention strategies to mitigate factors that may affect student performance adversely. There are a number of reasons for this:

1. Third level education should aim to maximise the potential of each student. Therefore, a careful examination of student outcomes against some benchmark or expected outcome may provide evidence as to whether student potential is being realised. Such insights may also help the college prioritise scarce resources, to focus them on specific problem areas.
2. Institutions have an obligation to deliver value for money to the bodies that fund them (in Ireland, funding comes primarily from the state, with some also from the students).
3. Institutions are often judged by the quality of the awards that they provide; e.g. the more honours level graduates a course provides, the better the course is perceived to be. This provides additional incentive for institutions to take proactive steps to support students.

The purpose of this work is to examine how data mining techniques may be applied to the task of analysing and interpreting data that has been drawn from an academic institution's record systems. The specific goals are:

- To demonstrate how data mining techniques may be applied to this domain;
- To identify which data mining algorithms are most suitable to the task;
- To examine the data mining results, to see what can be learned from them.

As described in Section 3, rule induction, tree induction, and Bayesian network induction techniques are all applied to the data, in order to discover which are the most useful in order to achieve insightful analysis of the data.

At the outset, a caveat should be noted. We do not believe that students' results are a deterministic function of attributes recorded in a college's information system; there are many factors, either unmeasured or unmeasurable, that may affect results, such as motivation, aptitude, interest and work ethic. Nonetheless, our working hypothesis is that it may be possible to identify some factors that are both recorded and relevant. This hypothesis will be tested by examining the predictive accuracy of models induced from the data.

2 Data Description

Data for this work was drawn from Limerick Institute of Technology's electronic record system. Records were prepared for all students registered in all four schools in LIT in 2004 and 2005. To maintain anonymity, student names, IDs and dates of birth were not extracted, though the date of birth was used to compute Mature Status (age > 25). Some effort was required to transform multiple tables into a single flat file, and to check the consistency of records extracted. Table 1 lists the attributes extracted for each student. This table indicates the available data rather that an ideal list of attributes. For example, anecdotal evidence suggests that there is a strong correlation between student attendance and overall performance. However, attendance data is not stored in the electronic record system, so this cannot be validated.

Student personality is also not recorded. Other studies, such as by Chamorro-Premuzic & Furnham [3] consider the influence of personality using much smaller datasets and with the student's permission. Furnham *et al.* [4] indicate that student personality type may hold the key to why students behave the way they do. Attendance levels provide a clue to one of the five basic personality traits of conscientiousness and they note that "*conscientiousness appears to be a consistent predictor of occupation performance throughout a variety of settings*" and define traits of conscientiousness as "achievement striving", "dutifulness", "order" and "responsibility". Interestingly, they conclude that conscientiousness may be a better predictor of academic performance than standard intelligence tests.

Another factor that is not considered in this study is the influence of the College's Learning Support Unit (LSU), a drop-in facility that is available to provide extra tuition support for students having difficulty with a particular subject or with the course they are studying. Students' usage of the LSU is recorded but is not linked centrally to individual student records.

Table 1. Description of attributes in the student dataset.

Attribute	**Description**
LivesNearby	Derived from county of residence, contains a value of 'Y' or 'N'; where Y corresponds to the nearest counties, Limerick and Clare.
Grant Status	'Y' or 'N' indicates whether a student is in receipt of a grant or not. Intended to give some indication of socio-economic background.
NativeEnglish	Derived from nationality of student. There are very few students with Native English = N: 6 in 2005 and 9 in 2004.
Mature	'Y' or 'N'; whether the student is considered to be in the academic category of 'Mature'(age > 25).
Gender	'F' for Female, 'M' for Male.
Level	Indicates the award level of the course in which the student is enrolled. 6 = National Certificate, 7 = Ordinary Level Degree, 8 = Honours Degree.
Student Type	An indication of the type of student enrolled. One of: New First Time; Continuing; Add On; Repeat (Full Time); Repeat (Exam Only); Transfer; Access.
Registration	Related to Student Type: one of Registered, Repeat, Withdrawn.
School	Indicates which of the 4 L.I.T. schools the student is enrolled in.
Department	The department in which the student is enrolled. Each department belongs to one school.
Time Status	'F' for full-time, 'P' for part-time. There are very few part-time students in the dataset.
Major Code	The major subject area of the course in which the student is enrolled.
Course	The specific course on which the student is enrolled.
Course Year	Year of the course on which the student is enrolled: values 1-4.
Block Description	Combination of Course and Course Year.
CAO Points[1]	CAO Points for the student (discretised into bands of 100 points where algorithms require discrete data). Only available in the system for Year 1 and Year 2 students; 'Unknown' otherwise.
Semester	Whether the subject is semesterised.
Application Type	Closely linked to Student Type. Values are: CAO Applicant; Internal Direct; External Direct; External Agency; NTCB Applicant.
CAO Preference	The ranking of the student's course in the student's CAO application. Available only when CAO points are available; 'Unknown' otherwise.
Result	Overall academic result of the student in the academic year: G01 (1st Class Honours); G21 (2nd Class Honours Grade 1); G22 (2nd Class Honours Grade 2); GPS (Pass); GFL (Fail)

It is likely that, if other attributes were available, they would impact on the relationships and interactions discovered. Nonetheless, the data available to this study forms a useful basis for the work.

The distribution of Result grades in the data is listed in Table 2. As it shows, there are 3459 records for 2004 and 3358 records for 2005. The most commonly awarded grade is 2nd Class Honours Grade II (G22).

[1] CAO is the Central Applications Office, an agency that manages admissions to all third-level institutions in Ireland. Students' points are computed from their second-level Leaving Certificate examination. Students list the courses they would like to apply for, in order of preference, and are offered a place on the highest-preference course for which they have sufficient points, which in turn depends on availability of places and demand for them.

Table 2. Distribution of grades for 2004 and 2005 data.

2004 Data	*#Results*	*% of Total*	*2005 Data*	*#Results*	*% of Total*
G01	363	10.49	G01	380	11.32
G21	1020	29.49	G21	1034	30.79
G22	1147	33.16	G22	1103	32.85
GPS	545	15.76	GPS	505	15.04
GFL	384	11.10	GFL	336	10.01
Total	3459	100.00	*Total*	3358	100.00

The distribution of results can be used to establish the baseline performance for classifiers for this application. A trivial classifier that always predicts the majority class (G22) will have a classification accuracy of 32.9% on the 2005 data. A 'real' classifier would have to outperform this if it is not just acting at random.

To test whether classifiers can capture general trends in results, even if not predicting them exactly, we define an alternative score which we call *G±1:* we give a score of 1 if the correct grade is predicted or 0.5 if the grade is one higher or lower, and express the answer as a percentage of all cases. Thus, on the 2005 data, the trivial G22 classifier would score 1103 + 1034*0.5 + 505*0.5 = 1872.5 out of 3358, or 55.8%.

3 Experiments

3.1 Experimental Procedure

The dataset described in Section 2 was analysed using a variety of different algorithms, in order to assess their utility in this application:

- Decision tree induction using C4.5
- Classification rule learning using PART and Prism (not discussed further for reasons of space, and because conclusions were similar to those of C4.5)
- Association rule learning using APRIORI
- Bayesian network learning using hill-climbing search.

The 2004 and 2005 datasets were analysed separately, for two reasons. Firstly, CAO information is available only for students who started in 2004 or later, i.e. First Year students only in 2004, but First and Second Year students in 2005. Therefore, the CAO information is distributed differently for the two years. Secondly, many of the same students have records relating to their 2004 and 2005 examinations, so it would not be correct to treat them as independent records. All results presented in this paper use the 2005 dataset.

To measure the validity of the models constructed using various classification algorithms, 10-fold cross validation was used to compute confusion matrices, accuracy and G±1 scores. Then, all data for a year was used to construct a final classification model, for discursive purposes. The popular Weka machine learning software package, as described by Witten and Frank [8], was used for all experiments.

3.3 Decision Trees

The C4.5 decision tree induction algorithm [7] is used for this analysis. Results were found to be relatively sensitive to specifying the minimum number of instances in each leaf of the tree. Table 3 shows an extract from the tree constructed when this is set to 20; with a larger setting, the trees created are more compact but lose inductive power, whereas with a smaller setting, the trees become so large as to be difficult to interpret. (In the table, the values *(X/Y)* are the numbers of records at a leaf that support/contradict the decision.)

Table 3. Excerpt from decision tree induced from 2005 data (132 leaves in total)

BlockDescription = Electronic_Eng_Level_7_Y3: GPS (21/13)	
BlockDescription = Civil_Engineering_Level_7__Y3: G22 (60/42)	
… … … …	
BlockDescription = Cons_Engr/Mgmt_Level_8_Y2	
… … … …	
BlockDescription = Video_and_Sound_Tech_L6__Y2: G21 (14/8)	
BlockDescription = Construction_Level_6_Y2	
| **CAOPoints <= 275: GPS (25/16)**	**(1)**
| **CAOPoints > 275: G22 (39/22)**	
BlockDescription = Civil_Engineering_Level_6__Y1	
| **CAOPoints <= 345: GPS (34/18)**	**(2)**
| **CAOPoints > 345: G22 (37/15)**	
BlockDescription = Software_Development_L8__Y1: G22 (18/10)	
BlockDescription = Business_Stud_Acc/Fin_L6_Y2	
| CAOPoints <= 320: GPS (30/11)	
| CAOPoints > 320	
| | **Gender = M: GPS (21/13)**	**(3)**
| | **Gender = F: G21 (23/13)**	
BlockDescription = Design_(Communications)_L7: G21 (30/15)	

This tree is quite shallow, and is dominated by clauses of the form: *if BlockDescription = A, Result = B*. Since BlockDescription is a combination of course and year, these clauses simply correspond to statements about average grades achieved by different student groups.

A small number of clauses, highlighted in bold and numbered in the table, are more interesting. Clauses (1) and (2) indicate that students with higher CAO points at entry tend to do better in Construction Studies Year 2 and Civil Engineering Year 2. Clause (3) indicates that among Business Studies students with relatively high CAO points, female students tend to perform substantially better than males: more males get Pass grades while more females get 2:1 Honours.

It is interesting to look at the confusion matrix for this classifier, shown in Table 4. On the diagonal, highlighted in bold, are the values that are correctly predicted, so for example, there are 1034 records with Result = G21, and 461 of these are predicted correctly. Overall, the predictive accuracy in this case is 36.8%, which is somewhat better than would be achieved by always predicting the majority class (G22, 32.9%).

On the other hand, the confusion matrix shows that the majority of predictions are off by only one grade; for example, of the misclassified G21 results, 90 are classified as G01 and another 387 are classified as G22. The G±1 Score for this classifier is 57.9%, which is an improvement on the trivial G22 classifier's score of 55.8%.

Table 4. Confusion matrix for C4.5 decision tree for 2005 Results dataset

	Predicted				
Actual	**G01**	**G21**	**G22**	**GPS**	**GFL**
G01	**87**	144	111	22	16
G21	90	**461**	387	62	34
G22	51	371	**524**	102	55
GPS	24	101	222	**124**	34
GFL	19	83	125	69	**40**

Similar trends are found when using the other classification algorithms. We interpret this as indicating that the factors being considered in this study are able to predict a general trend in student performance, but that unrecorded factors (perhaps such as motivation, attendance, aptitude and work level) result in a student performing better or worse than the trend would indicate.

Further experiments were conducted using different settings and with the BlockDescription attribute removed from the dataset. Overall, however, decision tree induction and classification rule induction did not yield many insights in this application.

3.4 Association Rules

The APRIORI algorithm [1] was used to discover association rules in the data, with settings such that only rules that cover at least 10% of the dataset and that correctly classify 90% of the instances are selected. Table 5 show the ten best rules found.

Table 5. Best 10 APRIORI rules from 2005 dataset

No.	**Rule**
1.	NativeEnglish=Y [3352] ==> TimeStatus=F [3348] conf:(1)
2.	Registration=RG [3226] ==> TimeStatus=F [3222] conf:(1)
3.	NativeEnglish=Y Registration=RG [3220] ==> TimeStatus=F [3216] (conf: 1)
4.	TimeStatus=F [3354] ==> NativeEnglish=Y [3348] conf:(1)
5.	Registration=RG [3226] ==> NativeEnglish=Y [3220] conf:(1)
6.	TimeStatus=F Registration=RG [3222] ==> NativeEnglish=Y [3216] (conf: 1)
7.	Registration=RG [3226] ==> NativeEnglish=Y TimeStatus=F [3216] (conf: 1)
8.	TimeStatus=F [3354] ==> Registration=RG [3222] conf:(0.96)
9.	NativeEnglish=Y [3352] ==> Registration=RG [3220] conf:(0.96)
10.	NativeEnglish=Y TimeStatus=F [3348] ==> Registration=RG [3216] (conf: 0.96)

On reviewing these rules, some interesting (although obvious) relationships are found. Many, however, are related to the NativeEnglish attribute (computed from the student's nationality) and given that the large majority of records have NativeEnglish = Y, the usefulness of the rules is questionable. The Timestatus column, indicating whether a student is full- or part-time, also emerges in a number of the rules, but again, there are only 20 part-time students in the dataset. Assuming almost all native

English speakers in the college are almost all Irish, the rules in Table 5 may be interpreted as follows: (1) most Irish students are full-time; (2) most registered students are full-time; (3) Most registered Irish students are full-time; and so on.

After experimenting with various settings and options, selected attributes were removed in order to determine whether this would lead to more useful rules. The removed attributes were those that had very few occurrences in the dataset or whether directly predictable from other attributes. Table 6 lists some of the rules generated when the attributes NativeEnglish, Timestatus, Semester, CAO Preference, and School were removed. Along with each rule is an interpretation in italics.

Table 6. Sample APRIORI rules from 2005 dataset with reduced attributes

No.	Rule
1.	LivesNearby=Y StudentType=New_First_Time [631] ==> Mature=N CourseYear=Y1 [590] (conf: 0.94) *Most first time students who live nearby are not mature and attend year 1 of a course.*
2.	MajorCode=Construction_Engineering/Mgmt [187] ==> Gender=M Registration=RG [175] (conf: 0.94) *Most registered students for Construction Engineering are Male*.
3.	Gender=F StudentType=New_First_Time [466] ==> Mature=N [440] (conf: 0.94) *Most new first time female students are not mature applicants.*

These are accurate, thereby helping to validate the methodology, but they do not offer much new. In addition, a limitation of association rules in this application is that they do not necessarily relate to the question of interest in this study: factors that influence student results.

3.5 Bayesian Networks

A standard hill-climbing algorithm [5] is used to construct the Bayesian network; Bouckaert [2] describes the implementation. This is a search-and-score approach; each candidate network is scored using the minimum description length (MDL) score, which evaluates the network's likelihood relative to the dataset, with a penalty term to favour less complex networks over more complex ones. For details, please refer to Heckerman [5]. The search procedure is to start with the empty network and successively apply local operations that greedily improve the MDL score maximally, until a local minimum is found. The local operations applied are arc insertion, arc deletion and arc reversal. Note that this search procedure does not require a node ordering to be specified. We specify that each node should have no more than 3 parents (this helps to restrict the network to identify only the strongest interactions between nodes), but do not otherwise constrain the network structure.

Figure 1 shows a Bayesian network structure learned from the 2005 data. It is useful to bear some points in mind when interpreting a network such as this:

1. An arc A → B can be interpreted as "A is correlated with B"; conversely; the absence of an arc between A and C can be interpreted as "A is not directly correlated with C".

2. An arc between two nodes may indicate a positive or negative correlation, or indeed a correlation that holds only for some values of the two variables; examination of the table of probabilities for each arc is required to fully understand the nature of the correlation.
3. As described by Pearl [6], the Markov blanket of a node *x* is the union of *x*'s direct parents, *x*'s direct children and all direct parents of *x*'s direct children. The Markov blanket of *x* is one of its Markov boundaries, meaning that *x* is unaffected by nodes outside the Markov blanket. Thus, when examining the factors thataffect a node, all nodes within its Markov blanket are relevant and all outside it are irrelevant.

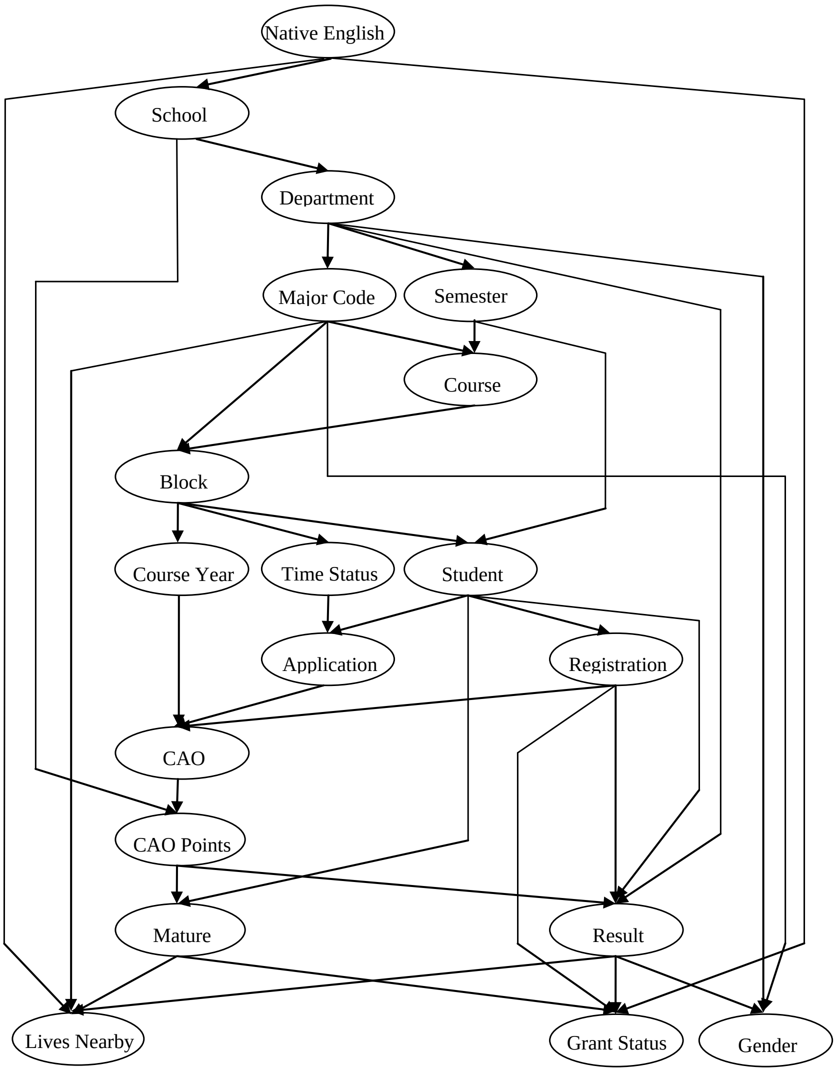

Fig. 1. Bayesian Network Induced from the 2005 Dataset

Examining the network, it can be seen that there are several features that are “obvious”, which, while they do not contribute to understanding the domain, demonstrate that the network can discover meaningful relationships in the data. These include:

- School is related to Department, which is related to Major Code, which is related to Course
- Block Description (i.e. the specific year of the specific course) is linked to Course Year, Course and Major Code
- Semester (i.e. whether there is semesterisation) is linked to Course and Department
- Native English (which is computed from the student’s home country) is linked to Lives Nearby (i.e. whether the student comes from the nearby counties of Limerick or Clare)
- CAO Points is linked to Mature status, since CAO points are not recorded for mature students
- Registration status and Native English are both linked to Grant status, since repeat students are not eligible to receive grants, nor are students from outside the European Union

There are several other features that are not quite so obvious but are logical:

- Gender is linked to Major Code and Department; this is because the ratio of male to female students varies quite widely between courses. For example, there tend to be more female students in the School of Art and Design than in the School of the Built Environment.
- CAO Points is linked to CAO Preference, for two reasons: firstly, if a student did not enter via the CAO, both Points and Preference will be coded as ‘unknown’; secondly, students with higher points are more likely to get their first preference
- CAO Points is linked to School, because for admission into some schools (such as Art and Design), ‘bonus’ CAO points are awarded for some factors (such as the student’s art portfolio).

The nodes in the Markov blanket of the Result node are:

CAO Points   Department   Major Code   Grant Status
Lives Nearby   Native English   Registration   Mature   Gender

These indicate, for example, that results vary by Department and Major Code (as was noted in the decision tree’s focus on Block Description), and that CAO Points are correlated with results. Registration status (which indicates whether a student is repeating) is also correlated, which is not surprising but did not emerge in earlier analyses. Gender is also shown to be a factor, though as mentioned above, gender balance varies between schools and average grades awarded also varies between schools. Further analysis would be required to explore the influence on results of the other factors identified here: Grant Status, Native English and Lives Nearby.

The classification accuracy of this Bayesian network is 36.9% and its G±1 Score is 58.5%. These results are slightly better than those of the decision tree analysis presented earlier, and somewhat better than the performance of the trivial G22 classifier.

Overall, while the Bayesian network analysis discovered some obvious relationships in the data, it also discovered some less obvious correlations. Because of the compact nature of Bayesian networks, with one node for each variable in the domain, they do not become more difficult to interpret when constructed with larger datasets. The one noteworthy limitation of the approach is that, at least when using the induction algorithm that has been used in this work, variables must be categorical. In this dataset, the only numerical variable was CAO Points, which had to be discretised into bands of 100 points.

4. Conclusions

The purpose of this work was to examine how data mining techniques may be applied to the task of analysing and interpreting data that has been drawn from an academic institution's record systems. Of the algorithms considered, most performed with similar accuracy, predicting about 37% of results correctly, which is higher than guessing the majority class (32.8%), and their G±1 Score was around 58%, whereas the trivial majority classifier's G±1 Score is 55.8%. We interpret this as indicating that the factors being considered in this study are able to predict a trend in student performance to a limited extent, but that factors not reflected in the data often result in a student performing better or worse than the trend would indicate.

Bayesian networks were found to be the most useful analysis tool for this application. The networks provide a clear and comprehensible graphical overview of relationships in the data, finding the obvious relationships, confirming relationships found by other algorithms and also finding other relationships that were not highlighted by the other algorithms. In addition, their predictive accuracy was as good as any other techniques tried, or better. For these reasons, the authors recommend the use of Bayesian networks for future data exploration applications such as the one addressed in this paper.

References

1. Agrawal, R., Mannila, H., Srikant, R., Toivonen, H. and Verkamo, A.I. "Fast Discovery of Association Rules." Advances in Knowledge Discovery and Data Mining. Menlo Park, CA: AIII Press/MIT Press. (1996)
2. Bouckaert, R. "Bayesian network classifiers in Weka." Available at www.cs.waikato.ac.nz/~remco/weka_bn/. (2005)
3. Chamorro-Premuzic, T., Furnham, A. "Personality predicts academic performance: Evidence from two longitudinal studies on university students." Journal of Research in Personality, 37. (2003)
4. Furnham, A., Chamorro-Premuzic, T., McDougall, F. "Personality, cognitive ability, and beliefs about intelligence as predictors of academic performance". Learning and Individual Differences, 14 49-66. (2003)
5. Heckerman, D. "A Tutorial on Learning with Bayesian Networks." Technical Report MSR-TR-95-06, Microsoft Corporation, Redmond. (1996)
6. Pearl, J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Francisco, USA. (1988)
7. Quinlan, J.R. C4.5: Programs for Machine Learning. Morgan Kaufman, S.F., USA. (1993)
8. Witten, I.H. and Frank, E. Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco, USA. (2005)