ISSN: 2141 – 3290
www.wojast.com

STUDENTS' DATASET MINING FOR ACADEMIC PERFORMANCE RISK PATTERNS IDENTIFICATION

INYANG[1], U. G. AND UMOREN[2] M. U.
[1]*Department of Computer Science,*
Faculty of Science, University of Uyo, Uyo, Nigeria.
e-mail: udoiinyang@yahoo.com
[2]*Department of Mathematics & Statistics,*
Faculty of Science, University of Uyo, Uyo, Nigeria.

ABSTRACT: Experience shows that, the number of students who fail to complete their studies on schedule is on the increase. Hence, the need to increase the number of students who successfully complete their studies on schedule by identifying at-risk students, at an early stage and then develop a plan for minimizing poor performance in the courses that contribute significantly to failure to complete studies on schedule. This will enable such students to adjust before they fail to graduate on schedule. In this paper, we seek to improve the quality of students' performances by discovering the courses that have strong correlation with students' duration of studies and identify risk patterns in the form of rules between these courses and the status of students after spending the stipulated minimum period of a programme. We used feature ranking algorithm with Tschuprow's T as the test statistic. The threshold of 0.0001 and 0.25 for p-value and test statistic respectively, were used to prune insignificant courses. The parameters for the risk pattern mining are minsup =30%, minconf =75% and lift $\geq$100%, a supervised rule generator based on apriori algorithm was the rule generation tool. Seven rules were generated and evaluated. The results show that the rules are reliable and strong enough to be used for the early identification and creation of support services for at-risk students.

INTRODUCTION

The evolution of information technology has made the collection, processing, transfer and storage of huge amount of data easier and cheaper to meet the increasing demand for information. As huge amount of data is being collected and stored in various formats (records, files, documents, images, sound, videos, scientific data) traditional statistical techniques and database management tools are no longer adequate for analyzing them, hence the need for proper and efficient knowledge extraction tool such as data mining [Kumar and Chadha, 2011]. Data mining, aims at discovering useful information from large collection of data. The main attribute of data mining is that it includes Knowledge Discovery in Databases (KDD) which is a nontrivial process of identifying valid, novel, potentially useful and understandable patterns in data repositories, thereby contributing to the prediction of outcome trends by profiling performance attributes that support effective decision making [Ogor, 2007].

Pattern mining is a data mining method that involves finding existing patterns in data. In this context, pattern means association. The original motivation for searching association rules came from the desire to analyze market transaction data, that is, to examine customers' behavior in terms of the purchased products [Liu, 2007]. The discovery of association rules is an important task in the process of data mining. The general objective is to find frequent co-occurrences of items within a set of transactions.

Students' result repository is a large data bank of students' raw scores and grades in different courses enrolled for during their years of attendance in an institution [Chandra and Nandhini, 2010]. Educational data can be personal or academic; it can be useful in the following ways; provision of an understanding of students' behaviour, assistance of instructors, improvement of teaching, evaluation and improvement of e-learning systems [Romero and Ventura, 2007].

Experience shows that the number of students who fail to complete their studies on schedule is on the increase. Considering the high cost of education; students who spend extra year(s) in school risk dropping out of school because of financial constraints. Also, spending more than the stipulated minimum duration of a programme, on account of poor academic performance, increases the affected students' chances of having a class of degree below second class lower. This extra year(s) may significantly contribute to the affected students' age, upon graduation, exceeding the minimum age of preference by employers. Hence, there is need to increase the number of students who successfully complete their studies on schedule by identifying, at-risk students, at an early stage. This will enable such students to adjust before they fail to graduate on schedule or drop out. In this paper, we seek to improve the quality of students' performances by discovering the courses that have strong correlation with students' duration of studies. Secondly, identify patterns in the form of rules between these poor performance courses and the status of students after the stipulated minimum period of a programme.

DATASET DESCRIPTION AND PREPROCESSING

The dataset used for training and analysis, consists of six sets of Bachelor of Science, Computer Science graduands. The input variables are performances of students in first year courses, both first and second semesters. The status of a student, after the stipulated five years (10 semesters), which is the minimum duration of the programme, is the target variable. Each Student's performance in a course, measured by their score in the course, is an aggregation of continuous assessment score and the examination scores. Continuous assessment includes assignments, class work, tests, seminars and so on. It constitutes 30% of the total score while examination is 70%. The various grades, G(x) associated with scores are described in equation 1.

$$G(x) = \begin{cases} \text{“}A\text{”} & ; \quad x > 69 \\ \text{“}B\text{''} & ; \quad 60 \leq x < 70 \\ \text{“}C\text{''} & ; \quad 50 \leq x < 60 \\ \text{“}D\text{''} & ; \quad 45 \leq x < 50 \\ \text{“}E\text{'} & ; \quad 40 \leq x < 45 \\ \text{“}F\text{'} & ; \quad x < 40 \end{cases} \qquad (1)$$

One of the requirements for graduation is a minimum of 'E' grade in all the compulsory courses prescribed for students in the programme. Upon satisfying the requirements, students graduate with the class of degree corresponding to their graduating Cumulative Grade Point Average. Any student, who fails to complete his studies on account of poor academic performance, is said to *'spill'*; such students are referred to as '*Spillover"* students. This category of students spends more than the minimum duration for graduation.

In this work, our interest is on students scores < 50% and the status of the student on expiration of the minimum period of studentship. We intend to discover hidden and important relationship and patterns between student performances in year one courses and their status. To achieve this, we use linguistic terms 'Satisfactory", 'Average' and 'Poor' to describe grades of scores while '*Graduated*' and '*Spillover*' describe students' status after the expiration of the minimum programme duration. The grade, G(x) associated with score, x, is presented in Equation 2.

$$G(x) = \begin{cases} \text{“}Satisfactory\text{”} & ; \quad x > 59 \\ \text{“}Average\text{''} & ; \quad 50 \leq x < 59 \\ \text{“}Poor\text{''} & ; \quad x < 50 \end{cases} \qquad (2)$$

Based on equation 2, we transform the scores to obtain discrete attributes suitable for processing by apriori algorithm. Students that failed to complete their studies on account of voluntary withdrawal and those who have missing result(s) in any of the courses were excluded from the dataset. We collected 496 complete records for Bachelor of Science, Computer Science students admitted from 1999 to 2005 in University of Uyo, Nigeria. The dataset was randomly split into two; 423 records as the training dataset and 73 as the test dataset.

DATA PRE-PROCESSING

Data pre-processing is aimed at filtering out redundant or irrelevant attributes from the original data [Tsai and Chen, 2010] and also rank attributes based on their contribution to the target values as well as increasing the speed and accuracy of models in the prediction task [Paris et al., 2010; Inyang et al., 2009. We used feature ranking algorithm provided in Tanagra 1.4.41 for feature selection. The test statistic is Tschuprow's *T*, a measure of association between each of the input variables and the target variable based on Chi square [Liebetrau, 1983]. The threshold of 0.0001 and 0.25 for p-value and test statistic respectively, were used to prune insignificant courses. A rank of all the courses based on the test statistic is shown in Table 1.

Table 1: Rank of courses based on weights and p-values

S/N	Attribute	Statistic	p-value
1	MTH 122	0.311190	0.000000
2	PHY 111	0.241510	0.000000
3	MTH 121	0.237389	0.000000
4	PHY 121	0.224498	0.000000
5	CSC 111	0.201018	0.000006
6	CSC 121	0.199666	0.000007
7	GST 121	0.197956	0.000008
8	GST 122	0.182187	0.000049
9	CHM 111	0.169214	0.000191
10	PHY 112	0.140604	0.002704
11	MTH 111	0.133720	0.004756
12	GST 114	0.130606	0.006084
13	PHY 122	0.127220	0.007899
14	GST 111	0.122606	0.011151
15	CHM 121	0.101855	0.044911
16	BIO 111	0.058158	0.363610
17	GST 112	0.033099	0.720591

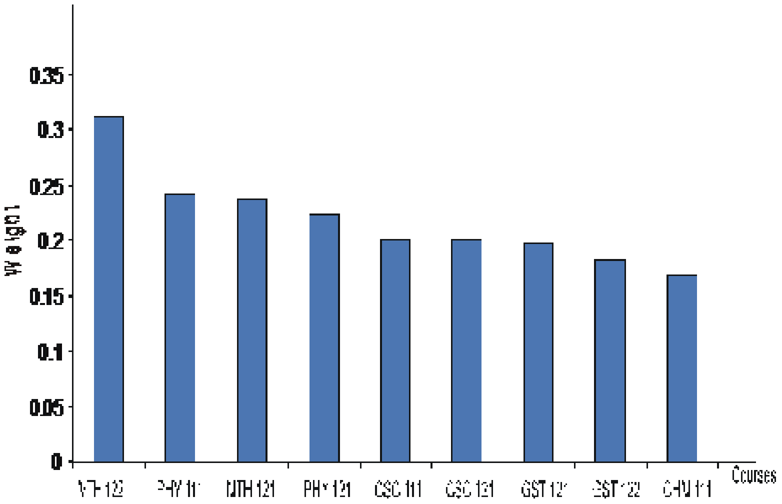


Figure 1: Selected attributes with their relevance value to the status of students

In Figure 2, the counts of spillover students are higher than those that graduated. Comparatively, the proportions of students that fail to complete their courses on schedule are more in the poor grades of courses, followed by average and satisfactory grade. This means that, the selected courses are strongly associated with spillover students.

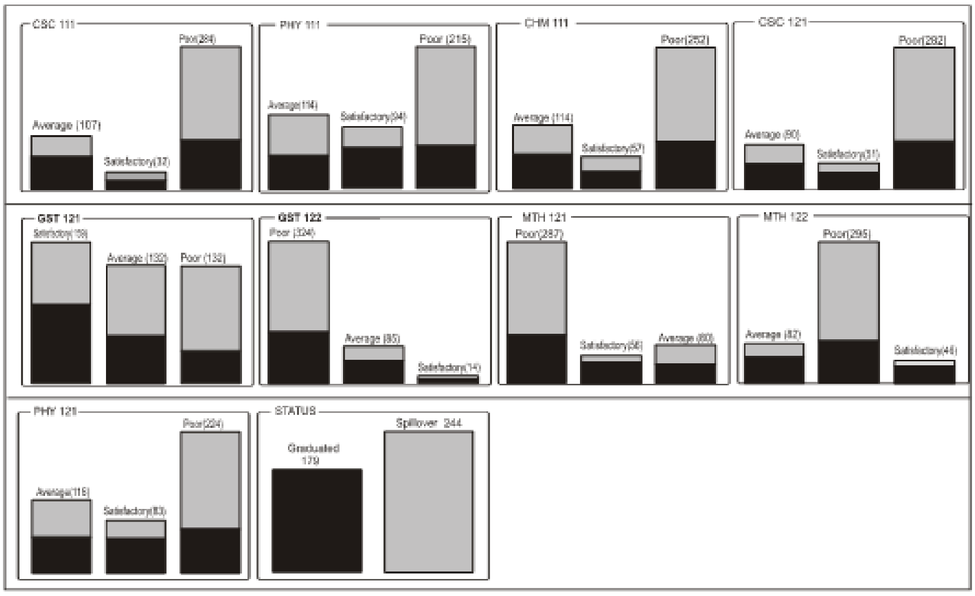


Figure 2: Significant courses showing their importance value to the status of students

The relationship of each 1-itemset with the student's status is an implication of the form X→T (where X represents courses and T is the status of the students) and measured in terms of support, confidence and lift. The support (s), confidence (c) and lift (I) are derived from equations 3, 4 and 5 as in [Tan et al., 2006]. The result is presented in Table 2.

$$\text{Support, s(X→T)} = \frac{\sigma(X \cup T)}{N} \tag{3}$$

$$\text{Confidence, c(X→T)} = \frac{\sigma(X \cup T)}{\sigma(X)} \tag{4}$$

$$\text{Lift, } I\text{(X→T)} = \frac{c(X \rightarrow T)}{s(T)} \tag{5}$$

Table 2: Relationship between input attribute and target attribute

S/N	X→T	Support	Confidence	Lift
1	GST 122→T	0.487	0.636	1.102
2	MTH 122→T	0.485	0.695	1.205
3	MTH 121→T	0.447	0.659	1.142
4	CSC 111→T	0.442	0.658	1.141
5	CSC 121→T	0.440	0.660	1.143
6	CHM 111→T	0.392	0.659	1.142
7	PHY 121→T	0.366	0.692	1.200
8	PHY 111→T	0.355	0.698	1.209
9	GST 121→T	0.225	0.720	1.248

In Table 2, T represents the spillover status of students. The confidence and lift of each rule is above 60% and 100% respectively. This means that the rules are effective; therefore these courses are suitable for rule mining. Hence, we wish to find the patterns in terms of rules using GST 122, MTH 122, MTH 121, CSC 111, CSC 121, CHM 111, PHY 121, PHY 111 and GST 121 as the frequent itemset.

PATTERN MINING AND EVALUATION

The parameters for the rule mining are minsup =20%, minconf =75% and lift =100%, a supervised rule generator based on apriori algorithm provided in Tanagra 1.4.41 was the rule generation tool. The attributes listed in Table 2 are the input variables, while students' status served as the target variable. A rank of the rules in order of their lift is presented in Table 3

Table 3: Rule set for predicting students' status

Rule No.	**Rules**	**Support (s)**	**Confidence (c)**	**Lift (I)**
1	MTH 122, CSC 111,PHY 121→T	0.262	0.816	1.415
2	MTH 122,PHY 121,PHY 111→T	0.206	0.813	1.410
3	MTH 121, CSC 11, PHY 121→T	0.229	0.789	1.367
4	MTH 122,GST 121 →T	0.201	0.787	1.364
5	GST 122, PHY 121,PHY 111 →T	0.206	0.784	1.359
6	MTH 122, CSC 121,PHY 121 →T	0.262	0.782	1.355
7	PHY 121,PHY 111 →T	0.227	0.780	1.353
8	CSC 111, CSC 121, PHY 121 →T	0.239	0.777	1.347
9	GST 122, MTH 122, PHY 121 →T	0.277	0.775	1.343
10	GST 122, CSC 111, PHY 121 →T	0.258	0.773	1.340
11	MTH 122, MTH 121, PHY 121 →T	0.248	0.772	1.338
12	CSC 111, PHY 121 →T	0.296	0.772	1.338
13	MTH 122, MTH 121, CSC 111 →T	0.336	0.768	1.331
14	MTH 122, PHY 121 →T	0.312	0.767	1.330
15	MTH 122, CHM 111, PHY 121 →T	0.217	0.767	1.329
16	CSC 111, CHM 111, PHY 121 →T	0.208	0.765	1.327
17	GST 122,MTH 122,PHY 111 →T	0.300	0.760	1.318
18	MTH 122,CSC 111,CHM 111 →T	0.291	0.759	1.316
19	GST 122, MTH 122,CSC 111 →T	0.364	0.759	1.315
20	GST 122, MTH 122, MTH 121 →T	0.371	0.758	1.315
21	GST 122,MTH 122,CSC 121 →T	0.364	0.755	1.309
22	GST 122, MTH 122, CHM 111 →T	0.312	0.754	1.308
23	MTH 122, MTH 121,PHY 111 →T	0.274	0.753	1.306
24	MTH 122, CSC 111, PHY 111 →T	0.279	0.752	1.303
25	GST 122, MTH 121, PHY 111 →T	0.279	0.752	1.303
26	MTH 121, CSC 121, PHY 121 →T	0.227	0.750	1.300

Each rule in Table 3 is a pattern; a total of 26 rules were extracted and ranked based on their lift. The consequence, T, is the **Spillover** status of students. In this context, the confidence of a rule specifies the probability of a co-occurrence of poor performances in courses listed in the antecedent part of the rule with the Spillover status of students. The lift specifies the effectiveness of the rules. Examples of the interpretation of the rules are as follows:

Rule 5: If a student's performance in {GST 122, PHY 121 and PHY 111} is **Poor** then the student is 78.4% likely to be a ***Spill*over** student.

Rule 10: If a student's score in GST 122, CSC 111 and PHY 121 is less than 50 (grades F, E and D), then the probability that such student will spend more than the minimum stipulated years before graduation is 77.3%.

In terms of risk associated with each rule, we used linguistic set {*Very high, High, Low*} to categorize the patterns based on support(s) and confidence (c) as follows:

$$\text{Risk} = \begin{cases} \text{'Very High'} & ;\ s \geq 3.0\ ,\ c \geq 0.75 \\ \text{'High'} & ;\ 2.2 < s \leq 2.5\ ,\ 0.72 < c < 0.75 \\ \text{'Low'} & ;\ 2.0 < s \leq 2.2\ ;\ c \leq 0.72 \end{cases} \qquad (6)$$

We are interested in the patterns with *very high* risk; we therefore present these set of rules in Table 4:

Table 4 : Very high risk rules for students' status prediction

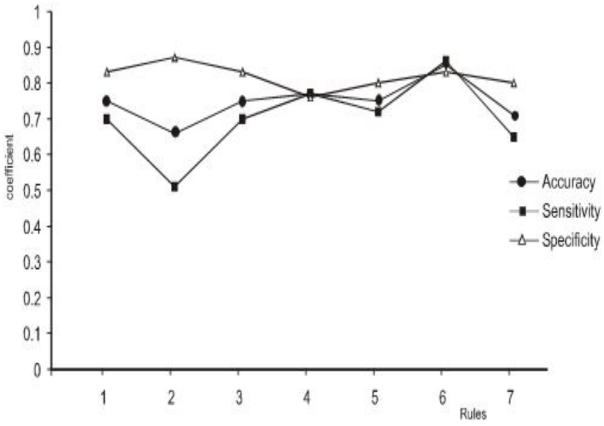
Rule No	Rules	Support	Confidence
1	MTH 122, MTH 121, CSC 111 →T	0.336	0.768
2	MTH 122, PHY 121 →T	0.312	0.767
3	GST 122, MTH 122, PHY 111 →T	0.300	0.760
4	GST 122, MTH 122, CSC 111 →T	0.364	0.759
5	GST 122, MTH 122, MTH 121 →T	0.371	0.758
6	GST 122, MTH 122, CSC 121 →T	0.364	0.755
7	GST 122, MTH 122, CHM 111 →T	0.312	0.754

The pattern of poor performances in the rules described in Table 4 is with very high risk, indicating that students that perform poorly in these courses have high chances of spending extra year(s) before graduation. The support and confidence of these rules are high making the rules to have high interestingness values. The ruleset in Table 4 were evaluated with the test dataset, the confusion matrix and rule evaluation measures are as presented in Table 5. True Positive (TP) is all spillover students correctly classified as spillover students. False Positive (FP) represents all graduates incorrectly classified as spillover students while True Negative (TN) indicates graduated students that were classified as graduates. Sensitivity, the True Positive Rate (TPR) determines each rule's test performance on classifying spillover instances correctly among all spillover samples in the test dataset. False Positive Rate (FPR), on the other hand, defines how many incorrect predictions of graduated cases occurred among all graduated students in the test dataset.

Table 5: Confusion matrix with rule evaluation measures

Rule No	Rule	TP	FN	FP	TN	Accu-racy	Sensi-tivity	Speci-ficity
1	MTH 122, MTH 121, CSC 111 →T	30	13	5	25	0.75	0.70	0.83
2	MTH 122, PHY 121 →T	22	21	4	26	0.66	0.51	0.87
3	GST 122, MTH 122, PHY 111 →T	30	13	5	25	0.75	0.70	0.83
4	GST 122, MTH 122, CSC 111→T	33	10	7	23	0.77	0.77	0.76
5	GST 122, MTH 122, MTH 121 →T	31	12	6	24	0.75	0.72	0.8
6	GST 122, MTH 122, CSC 121 →T	37	6	5	25	0.85	0.86	0.83
7	GST 122, MTH 122, CHM 111 →T	28	15	6	24	0.71	0.65	0.8

The accuracy measures the fraction of predicted spillover cases that are actually spillover students. From Table 5, a high accuracy is accompanied by high sensitivity and specificity of rules. The relationship between accuracy, specificity and sensitivity is as shown in Figure 3.



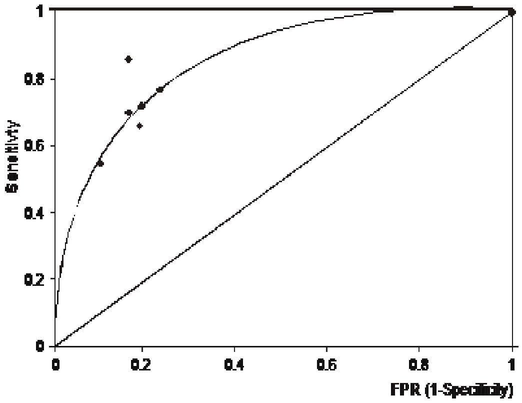


Fig. 3: Relationship between rules evaluation measures

Fig. 4: ROC Curve of *very high* risks rules evaluation

The specificity of the rules is relatively higher than accuracy and sensitivity, which implies very low FPR. The results of accuracy, sensitivity and specificity prove that the rules are efficient and reliable for prediction of students' status. The Receivers Operating Characteristics (ROC) curve of the rules evaluated with the test dataset is shown in Figure 4. In Figure 4, each point is a coordinate (FPR,TPR) and represents an instance of a rule. The ROC curve in Figure 4 is closer to the perfect classification coordinate (0,1), which indicates a better performance of the rules [Fawcett, 2006]. It also shows that any increase in sensitivity will be accompanied by a decrease in specificity.

RULE ANALYSIS

The results show that the 7 rules are interesting since their support and confidence are high. Out of the 7 rules, MTH 122 appears in all the rules while GST 122 has a count of 71.4%; this signifies the importance of these courses in the programme under study. It implies that students who perform poorly in these courses have a probability, determined by other courses in the pattern, of having an extra year. For example {GST 122, MTH 122, CSC 111} with (s = 0.364, c = 0.759) indicates a probability of 0.76 of spending extra year(s) before graduation. There is need to pay more attention to the teaching of these courses. Also students who fail these courses should not be allowed to register higher level courses in which they are prerequisites. For example CSC 111 and CSC 121 are prerequisites to most core departmental courses. Students who perform poorly in these courses should be monitored by counselors, staff advisers and lecturers. More attention and support services should be given to this category of students to reduce their risk of spending extra year before graduation. The performance evaluation of these rules shows high accuracy. The sensitivity, accuracy and specificity reveal that the rules have very low FPR, therefore are very reliable.

CONCLUSION

Predicting students' performance is useful in identifying students who are likely to perform poorly in their studies. Association rule mining which has been used to perform important analysis in the educational environment, for decisions to enhance educational standards, was employed in this work. A dataset of 423 complete records of B. Sc. Computer Science graduates was the training dataset, 17 first year courses were used as input variables and status of students after the minimum duration of studies was the target variable. We identified and ranked the input variables based on their percentage contribution to the duration of students in school. We used feature ranking algorithm with Tschuprow's T as the test statistic. The threshold of 0.0001 and 0.25 for p-value and test statistic respectively, were used to prune insignificant courses. The interestingness measures for risk pattern mining were minimum support of 30%, minimum confidence of 75% and minimum lift threshold of 100%, a supervised rule generator based on apriori algorithm was the rule generation tool. 7 risk patterns were extracted and evaluated for interestingness using accuracy, sensitivity and specificity as measures. The results show that the rules are reliable and strong enough to be used for the early identification, creation of support services for at-risk students and development a plan for minimizing poor performances in the identified courses. This rule set can be used in the knowledge base of expert system in the domain of educational data mining

REFERENCES

Chandra, E. and Nandhini, K (2010): Knowledge Mining from Student Data. European Journal of Scientific Research. 47 (1) pp156-163

Fawcett, T. (2006): An introduction to ROC analysis. Pattern Recognition Letters. (27) pp 861–874.

Inyang, U. G.; Njungbwen, E. and Inyang, M. U. (2009): Design of An Analytic Hierarchy Process Based Decision Support System for Residential Property Renting. ICASTOR Journal of Mathematical Sciences. 3(2) 183-195

Kumar, V. and Chadha, A. (2011): An Empirical Study of the Applications of Data Mining Techniques in Higher Education. International Journal of Advanced Computer Science and Applications. 2(3) pp 80-84

Liebetrau, A. (1983): Measures of Association (Quantitative Applications in the Social Sciences). Sage Publications

Liu, B. (2007): Web Data Mining: Exploring Hyperlinks, Contents and Usage Data. Springer

Ogor, E. N (2007): Student Academic Performance Monitoring and Evaluation Using Data Mining Techniques. Fourth Congress of Electronics, Robotics and Automotive Mechanics. IEEE Computer Society. pp 354-359

Paris, I. H. M.; Affendey, S. L. and Mustapha, N (2010): Improving Academic Performance Prediction using Voting Technique in Data Mining. World Academy of Science, Engineering and Technology (62) pp820-823

Romero, C., and Ventura, S. (2007): Educational Data mining: A Survey from 1995 -2005, Expert systems with applications (33) pp 135-146

Tan, Pang-Ning, Steinbach, M and Kumar, V. (2006): Introduction to Data Mining. Addison-Wesley

Tsai, Chih-Fong and Chen, M. (2010): Variable Selection by Association Rules for Customer Churn Prediction of Multimedia on Demand. Expert Systems with Applications. 37(3) pp 2006-2015.