Original Article

Selecting a suitable method of data mining for successful forecasting

Received (in revised form): 31st October 2011

Yas Alsultanny

was an Assistant Professor at Baghdad University, and the head of the Remote Sensing Department at the Scientific Research Council in Baghdad from 1997 to 2000. From 2000 he was an Associate Professor at the Applied Science University in Jordan, and in 2007 he became an Associate Professor at the Arabian Gulf University in Bahrain. He has spoken at many conferences.

ABSTRACT The aim of using data mining in the education field is to enhance educational performance, by using the six school indicators, which are defined in this article to identify useful guidelines that can improve school performance. Knowledge discovery requires a clear methodology that can be successfully applied in the education sector. This can be obtained from the use of the CRoss-Industry Standard Process for Data Mining (CRISP-DM). The CRISP-DM was used in this article to implement data mining for knowledge discovery from the database of 104 schools. Three methods of data mining, Naïve Bayes, Nearest Neighbor and the C4.5 decision tree, are implemented on the school data. The results showed that the C4.5 decision tree is significantly more accurate compared with the other methods.

Journal of Targeting, Measurement and Analysis for Marketing (2011) **19**, 207–225. doi:10.1057/jt.2011.21

Keywords: data mining; knowledge discovery; CRISP; school performance; forecasting; decision tree

INTRODUCTION

Database systems have had great success during the past two decades. More and more data are being collected and saved in databases; a database with a petabyte (1024 terabytes) of data is no longer uncommon. Finding useful information in these databases has become an important focus of many enterprises, and more and more attention has turned to data mining as a key component of such information discovery. Data mining algorithms and visualization tools are being used to find important patterns in data and to create useful forecasts. This technology is being virtually

applied in all business sections, including banking, telecommunication, manufacturing, marketing and e-commerce.¹

Data mining is not specific to one type of media or data, and should be applicable to any kind of information repository. However, algorithms and approaches may differ when applied to different types of data. Indeed, the challenges presented by different types of data vary significantly. Data mining is being put into use and studied for databases, including relational databases, object-relational databases and object-oriented databases, data warehouses, transactional databases, unstructured and semi-structured repositories such as the World Wide Web, advanced databases such as spatial databases, multimedia databases, time-series databases and textual databases, and even flat files. A data

---

Correspondence: Yas Alsultanny
College of Graduate Studies, Arabian Gulf University, Manama,
PO Box 26671, Kingdom of Bahrain
E-mail: alsultanny@hotmail.com

warehouse is a repository of data collected from multiple data sources (often heterogeneous), and is intended to be used as a whole under the same unified schema. Data warehouses are usually modeled by a multi-dimensional data structure.

KNOWLEDGE DISCOVERY

Knowledge discovery has been defined as the 'non-trivial extraction of implicit, previously unknown and potentially useful information from data'. It is a process of which data mining forms just one part. Figure 1 shows a slightly idealized version of the complete knowledge discovery process. Data come in, possibly from many sources. They are integrated and placed in some common data store. Part of them is then taken and pre-processed into a standard format. These 'prepared data' are then passed to a data mining algorithm that produces an output in the form of rules or some other kind of 'patterns'. These are then interpreted to give new and potentially useful knowledge. This brief description makes it clear that although the data mining algorithms are central to knowledge discovery, they are not the whole story. The pre-processing of the data and the interpretation (as opposed to the blind use) of the results are both of great importance.²

The main purpose of data mining is to extract patterns from the data at hand, increase their intrinsic value and transfer the data to knowledge.¹ Figure 2 shows that in the

Figure 1: Data mining is the core of the knowledge discovery process.
Source: Bramer (2007)².

Figure 2: Data, information and knowledge.
Source: Turban *et al* (2010)³.

information technology context, knowledge is very distinct from data and information.³

DATA MINING TASKS

Data mining tasks can be grouped into the following main categories:

- Classification
- Clustering
- Association rules
- Regression
- Forecasting

RELATED STUDIES

Urbancic *et al*,⁴ discussed some aspects of the international offer of machine learning: data mining and decision support education, based on investigations of the materials available on the world-wide web. The aim of this investigation was to provide starting points for an analysis that could be used to enhance the educational offer in this area in general, and knowledge transfer from academia to industry and the business sector in particular, in addition to general findings concerning the materials collected such as contents, target audience and teaching materials.

Behrouz *et al*⁵ presented an approach to classifying students in order to predict their final grade based on features extracted from logged data in an education web-based system. Their design implemented a series of pattern classifiers to compare their performance on an online course dataset. A combination of multiple classifiers leads to a significant improvement in classification performance.

Schumann⁶ proved that the CRoss-Industry Standard Process for Data Mining (CRISP-DM), a non-proprietary data mining process that was

developed for and is currently used in the business world, can be transferred to educational settings and provide a start-to-end structure that is capable of producing operationally actionable information to address the student achievement questions of educational leaders.

Häubscher *et al*⁷ collected data from interactions with an educational system, which are only useful if the patterns can be meaningfully interpreted in the context of the student–system interaction.

Romero *et al*⁸ compared different data mining methods and techniques for classifying students based on their model usage data and the final marks obtained in their respective courses. They developed a specific mining tool for making the configuration and execution of data mining techniques easier for instructors. They used real data from seven model courses with Cordoba University students.

Pittman⁹ proposed to expand the understanding of how educational institutions might benefit from including data mining processes and multivariate analysis to inform student retention strategies. Her study applied data mining techniques to student demographic and behavioral data in an institution of higher education, providing a detailed description of the data mining process.

DATA MINING IMPLEMENTATION AND ANALYSIS

There are many data mining techniques that can be used to evaluate a school's performance. The following classification technique is used:

- Naïve Bayes Classification;
- *k*-Nearest Neighbor Classification;
- Decision tree Classification.

The CRISP-DM is used in this article,¹⁰ which consists mainly of the following tasks:¹¹

- Business Understanding;
- Data Understanding;
- Data Preparation;
- Modeling;
- Evaluation;
- Deployment.

Figure 3 shows the phases and tasks for CRISP-DM.

It is important for educational leaders and decision makers to classify the schools according to their performance, which is based on several contributing indicators in order to help school administrations to correct their problems.

Figure 3: Phases and tasks for CRISP-DM.

Figure 4: Designed data mining block diagram.

The following indicators have been selected and used in this article:

$$\text{Students Per Class (SPC)} = \frac{\sum \text{Students count}}{\sum \text{Class count}} \quad (1)$$

$$\text{Students Per Teacher (SPT)} = \frac{\sum \text{Students count}}{\sum \text{Teachers count}} \quad (2)$$

$$\begin{aligned} \text{Success Rate of the School (SRS)} \\ = \frac{\sum \text{Success rate of successful students}}{\sum \text{Successful students}} \end{aligned} \quad (3)$$

$$\begin{aligned} \text{Percentage of Successful Students in} \\ \text{School (PSSS)} = \frac{\sum \text{Successful students}}{\sum \text{All students}} \end{aligned} \quad (4)$$

$$\begin{aligned} \text{Experience Rate of School (ERS)} \\ = \frac{\sum \text{The total experience of school staff}}{\sum \text{All school staff}} \end{aligned} \quad (5)$$

$$\begin{aligned} \text{STability Rate of the School (STRS)} \\ = \frac{\sum \text{Experienced teachers}}{\sum \text{All teachers}} \end{aligned} \quad (6)$$

Figure 4 shows the data mining process diagram that was designed and implemented in this article.

Data that are collected from the schools are constantly increased with time. These data are raw and must be converted into a form to extract useful knowledge from them, and this knowledge is essential for decision-makers working in the educational field. Data were collected from a data warehouse for the academic year 2007–2008. The data for 104 intermediate schools were used in this article. The data warehouse contains data about the staff, schools, exams and students. These tables of the warehouse were constructed for processing. Figure 5 shows the samples to the tables used in building the data warehouse.

The data were distributed among 170 tables and stored in the Microsoft Access database. This database is classified according to the following category:

- *Staff data* (contain detailed information on employee, level of education, specialization, the university, personal information and training courses).
- *School data* (contain detailed information about schools and their geographical location, the school library, laboratories, school supplies and much more information).
- *Exam data* (contain detailed information on all of the students' exams, statistics, analysis of results, graphs for students, and information to assist in the issuance of monthly reports on the student).

Figure 5: Sample of the schema database warehouse.

- *Student data* (contain detailed information about students, their parents, contact information, health, social data, official papers and many other things).

The six indicators from Equations (1) to (6) were used to measure the *School Performance*; these indicators enabled the production of the dataset that was used with data mining. Table 1 shows the sample of the result to these dataset. These

samples of the 104 schools were selected from the warehouse that was built in this study.

The suggested procedure to classify the schools' performance is as follows:

1. Sorting the value of SPC and SPT attributes in ascending order and the rest of the attributes in descending order.
2. The dataset was divided into *three groups*: the first group with 26 readings, the second group

with 52 readings and the third group with 26 readings with the assumption that the values are distributed by normal distribution.

3. The period band for each class was identified through the initial and final values of each attribute. For example, in Table 2 the excellent band (first group) for attribute SPC is between (11.14) and (20.30), the value (11.14) will be the initial value, while (20.30) will be the last value. Table 2 shows the period bands for all the attributes.
4. Generate classification rules for dataset by using the previous step. Table 3 shows these rules.
5. Classify the dataset according to the classifications of Table 3. Table 4 shows the sample of the training dataset after classification.
6. Replace the following:
  - *Good* values replaced by the number 1.
  - *Very good* values replaced by the number 2.
  - *Excellent* values replaced by the number 3.
7. The total score was calculated for the six indicators so that the total score cannot be less than 6 (by assuming all the attributes had values of 1) and the total score cannot be more than 18 (by assuming all the attributes had values of 3).

8. Total score was divided into three bands as follows:

- Between 6 and 10 points classified as *good*.
- Between 11 and 14 points classified as *very good*.
- Between 15 and 18 points classified as *excellent*.

Table 5 shows the sample of the dataset after applying steps 6–8.

The dataset has been divided into two groups, the first group used to train the model and the second used to test this model. The first group (training dataset) consists of 80 instances (records), approximately 77 percent of the dataset, and

Table 1: The six indicators of the schools

School no	SPC	SPT	SRS	PSSS	ERS	STRS
1	20.30	9.23	83.80	91.13	4542.00	100.00
2	17.64	8.08	81.41	86.08	4226.00	95.83
3	33.50	14.36	81.88	99.75	2655.00	67.86
4	34.08	14.77	83.40	80.36	2179.00	80.00
5	35.23	14.77	79.87	91.27	2602.00	74.19
6	19.91	9.95	81.79	85.84	2135.00	86.36
7	28.78	11.77	82.95	74.52	3906.00	100.00
8	39.25	20.48	88.98	89.17	3977.00	82.61
9	31.70	15.10	85.49	93.38	3419.00	90.48
10	28.50	14.25	82.84	88.72	2678.00	82.14
...	...	...	...	...	...	...
...	...	...	...	...	...	...
...	...	...	...	...	...	...

Table 3: Classification rule for dataset

Indicators (tables attributes)	Class 1	Class 2	Class 3
	Good (G)	Very good (VG)	Excellent (EX)
SPC	>30.5	>20.3 AND =<30.5	<=20.3
SPT	>13.63	>8.71 AND =<13.63	<=8.71
SRS	<81.09	>=81.09 AND <85.87	>=85.87
PSSS	<80.36	>=80.36 AND <95.08	>=95.08
ERS	<2386	>=2386 AND <3928	>=3928
STRS	<60.00	>=60.00 AND <90.91	>=90.91

Table 4: Sample of training dataset after classification

No	SPC	SPT	SRS	PSSS	ERS	STRS
1	EX	VG	VG	VG	EX	EX
2	EX	EX	G	VG	EX	EX
3	G	G	VG	EX	VG	VG
4	G	G	VG	G	G	VG
5	G	G	G	VG	VG	VG
6	EX	VG	G	VG	G	VG
7	VG	VG	VG	G	VG	EX
8	G	G	EX	VG	EX	VG
9	G	G	VG	VG	VG	VG
10	VG	G	VG	VG	VG	VG
...	...	...	...	...	...	...
...	...	...	...	...	...	...
...	...	...	...	...	...	...

Table 2: Period band classification

	SPC		SPT		SRS		PSSS		ERS		STRS	
	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
EX	11.14	20.30	4.33	8.71	85.87	94.66	95.08	100.00	3928	5844	90.91	100.00
VG	20.30	30.50	8.71	13.63	81.09	85.87	80.36	95.08	2386	3928	60.00	90.91
G	30.50	58.25	13.63	33.29	76.43	81.09	47.44	80.36	1295	2386	12.00	60.00

Table 5: Sample of dataset classification used in training data mining

No.	SPC	SPT	SRS	PSSS	ERS	STRS	Total score	School ranks
1	3	2	2	2	3	3	15.00	EX
2	3	3	1	2	3	3	15.00	EX
3	1	1	2	3	2	2	11.00	VG
4	1	1	2	1	1	2	8.00	G
5	1	1	1	2	2	2	9.00	G
6	3	2	1	2	1	2	11.00	VG
7	2	2	2	1	2	3	12.00	VG
8	1	1	3	2	3	2	12.00	VG
9	1	1	2	2	2	2	10.00	G
10	2	1	2	2	2	2	11.00	VG
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...

the second group (testing dataset) consists of 24 instances (records), approximately 23 percent of the dataset.

NAÏVE BAYES CLASSIFICATION

Naïve Bayes is one of the most efficient and effective inductive learning algorithms for machine learning and data mining. Its competitive performance in classification is surprising.¹² The probability of an *event*, for example that the school receives excellent performance, is a number from 0 to 1 inclusive, with 0 indicating ‘impossible’ and 1 indicating ‘certain’. A probability of 0.7 means that if a long series of trials was conducted, for example if the *School Performance* recorded yearly for N years, the expectation that the school will get excellent performance calculated by $0.7 \times N$ years. The longer the series of trials, the more reliable this estimate is likely to be.²

Given a set of k mutually exclusive and exhaustive classifications $c_1, c_2, \dots, c_k$, which have prior probabilities $P(c_1), P(c_2), \dots, P(c_k)$, respectively, and n attributes $a_1, a_2, \dots, a_n$, which for a given instance have values $v_1, v_2, \dots, v_n$, respectively, the posterior probability of class c_i occurring for the specified instance can be shown to be proportional to

$$P(c_{\downarrow} i) \times P(a_{\downarrow} 1 = v_{\downarrow} 1 \text{ and } a_{\downarrow} 2 = v_{\downarrow} 2 \dots \text{ and } a_{\downarrow} n = v_{\downarrow} n | c_{\downarrow} i) \quad (7)$$

Making the assumption that the attributes are independent, the value of this expression can be calculated by using the product

$$P(c_{\downarrow} i) \times P(a_{\downarrow} 1 = v_{\downarrow} 1 | c_{\downarrow} i) \times P(a_{\downarrow} 2 = v_{\downarrow} 2 | c_{\downarrow} i) \times \dots \times P(a_{\downarrow} n = v_{\downarrow} n | c_{\downarrow} i) \quad (8)$$

The product will be calculated for each value of i from 1 to k and the classification that has the largest value will be chosen. Three mutually exclusive and exhaustive events can be defined as follows:

- E_1 – School has a *good* performance.
- E_2 – School has a *very good* performance.
- E_3 – School has *excellent* performance.

From the training dataset in Table 5, the probability of an event is usually indicated by a capital letter P , to get the following:

$$\begin{aligned} P(E_1) &= 18/80 = 0.225 \\ P(E_2) &= 54/80 = 0.675 \\ P(E_3) &= 8/80 = 0.100 \end{aligned}$$

The values 18, 54 and 8 represent the number of schools that was obtained at the ranks of *Good*, *Very Good* and *Excellent*, respectively. The number of schools used as a training set is 80. Each of these probabilities is between 0 and 1 inclusive, as it has to be qualified as a probability. They also satisfy a second important condition: the sum of the ranks of the three probabilities has to be 1, because precisely one of the events must always occur. In this case,

$$P(E_1) + P(E_2) + P(E_3) = 1 \quad (9)$$

Table 6 shows the *School Performance* obtained by calculating how many times each attribute occurs with each value (*Good*, *Very good* and *Excellent*) to determine the *School Performance*. For example, the SPC is *Very good* for 39 schools.

- Six of them have *School Performance* = *Good*.
- Thirty-two of them have *School Performance* = *Very good*.
- One of them has *School Performance* = *Excellent*.

Table 6: Conditional and prior probabilities for dataset

Indicators (attributes)	School performance		
	G	VG	EX
SPC			
G	12/18=0.667	5/54=0.093	0/8=0.000
VG	6/18=0.333	32/54=0.593	1/8=0.125
EX	0/18=0.000	17/54=0.315	7/8=0.875
SPT			
G	13/18=0.722	7/54=0.130	0/8=0.000
VG	3/18=0.167	31/54=0.574	3/8=0.375
EX	2/18=0.111	16/54=0.296	5/8=0.625
SRS			
G	6/18=0.333	9/54=0.167	1/8=0.125
VG	12/18=0.667	28/54=0.519	3/8=0.375
EX	0/18=0.000	17/54=0.315	4/8=0.500
PSSS			
G	5/18=0.278	8/54=0.148	0/8=0.000
VG	13/18=0.722	24/54=0.444	5/8=0.625
EX	0/18=0.000	22/54=0.407	3/8=0.375
ERS			
G	4/18=0.222	20/54=0.370	2/8=0.250
VG	14/18=0.778	32/54=0.593	0/8=0.000
EX	0/18=0.000	2/54=0.037	6/8=0.750
STRS			
G	5/18=0.278	19/54=0.352	1/8=0.125
VG	12/18=0.667	27/54=0.500	3/8=0.375
EX	1/18=0.056	8/54=0.148	4/8=0.500
Prior probability	18/80=0.225	54/80=0.675	8/80=0.100

Table 7: School performance for a new school

No.	SPC	SPT	SRS	PSSS	ERS	STRS	School performance
81	VG	VG	VG	VG	EX	EX	Unknown

The cells in the table simply count these occurrences for all possible values of each attribute and probabilities that the event will occur. The prior probability value in the final row (prior probability row) in Table 6 shows the total number of occurrences of *Excellent*, *Very good* and *Good*, respectively.

As an example, suppose a new school with the values that are shown in Table 7. How to recognize the *School Performance* for this school?

$$\text{School Performance} = \text{Good} \\ 0.333 \times 0.167 \times 0.667 \times 0.722 \times 0.000 \times 0.056 = 0$$

$$\text{School Performance} = \text{Very good} \\ 0.593 \times 0.574 \times 0.519 \times 0.444 \times 0.037 \times 0.148 \\ = 0.00043$$

$$\text{School Performance} = \text{Excellent} \\ 0.125 \times 0.375 \times 0.375 \times 0.625 \times 0.750 \times 0.500 \\ = 0.004065$$

The largest value is for class = *Excellent*

The target value is calculated by the probability, the highest probability value will be selected, this value represent the performance that needs determination.

Table 8 calculated by using the Naïve Bayes classification algorithm, where Table 5 used to determine the rank in each column of Table 8, the posterior probabilities to each possible rank used to test the dataset.

The predictive accuracy of a model is estimated by the following equation:

$$\text{Predictive accuracy } p = C / N \quad (10)$$

where C: correctly classified instances; N: total number of instances in test set.

Correctly classified=18	Incorrectly classified=6
Therefore, the Predictive accuracy	$p = (18/24) \times 100 = 75$ percent
	$1 - p = (6/24) \times 100 = 25$ percent

Nearest Neighbor classification is mainly used when all attribute values are continuous, although it can be modified to deal with categorical attributes. The idea is to estimate the classification of an unseen instance by using the classification of the instance or instances that are *closest* to it, Table 9 shows the dataset of nearest neighbor classification.

Basic k-Nearest Neighbor Classification Algorithm

- Find the *k* training instances that are closest to the unseen instance.
- Take the most commonly occurring classification for these *k* instances.

Table 8: Classification by Naïve Bayes algorithm

No	SPC	SPT	SRS	PSSS	ERS	STRS	Actual school performance	Posterior probabilities	Predicted school performance	Error
1	VG	VG	VG	VG	EX	EX	VG	EX VG G 0.004065 0.00043 0	EX	✓
2	G	G	EX	VG	EX	VG	VG	EX VG G 0 0.0000311 0	VG	
3	EX	EX	VG	G	EX	EX	EX	EX VG G 0 0.0000393 0	VG	✓
4	G	G	G	VG	VG	VG	G	EX VG G 0 0.0002634 0.0601026	G	
5	G	G	G	G	VG	EX	G	EX VG G 0 0.0000260 0.0019264	G	
6	VG	VG	G	G	VG	G	G	EX VG G 0 0.0017514 0.0011114	VG	✓
7	G	G	VG	G	VG	EX	G	EX VG G 0 0.0000809 0.0038527	G	
8	VG	VG	G	G	VG	VG	G	EX VG G 0 0.0024888 0.0026673	G	
9	VG	VG	G	G	EX	EX	VG	EX VG G 0 0.0000461 0	VG	
10	G	VG	VG	VG	EX	EX	VG	EX VG G 0 0.0000672 0	VG	
11	VG	VG	G	VG	EX	VG	VG	EX VG G 0.0010300 0.0004667 0	EX	✓
12	EX	EX	EX	VG	EX	VG	EX	EX VG G 0.0480652 0.0002417 0	EX	
13	VG	VG	EX	EX	EX	EX	EX	EX VG G 0.0032959 0.0002394 0	EX	
14	VG	VG	VG	VG	EX	EX	VG	EX VG G 0.0041199 0.0004302 0	EX	✓
15	G	VG	G	G	EX	EX	VG	EX VG G 0 0.0000072 0	VG	
16	VG	VG	G	G	EX	EX	VG	EX VG G 0 0.0000461 0	VG	
17	G	G	VG	VG	EX	EX	VG	EX VG G 0 0.0000152 0	VG	
18	VG	VG	G	G	EX	EX	VG	EX VG G 0 0.0000461 0	VG	

Table 8 continued

No	SPC	SPT	SRS	PSSS	ERS	STRS	Actual school performance	Posterior probabilities	Predicted school performance	Error
19	G	VG	VG	G	EX	EX	VG	EX VG G 0.0000224 0	VG	
20	VG	VG	EX	VG	EX	VG	VG	EX VG G 0.0041199 0.0008815 0	EX	✓
21	VG	G	G	G	VG	VG	G	EX VG G 0 0.0005620 0.0115582	G	
22	VG	VG	VG	G	EX	EX	VG	EX VG G 0 0.0001434 0	VG	
23	G	VG	EX	VG	EX	VG	VG	EX VG G 0 0.0001377 0	VG	
24	VG	EX	VG	G	EX	VG	VG	EX VG G 0 0.0002498 0	VG	

On the basis of the k -Nearest Neighbor classification algorithm and Table 9, the following question can be asked:

What should classification be if there is the following unseen instance?

SPC	SPT	Class
25	10	????

The answer for this question is found by calculating the distance between the point (25,10) and other points and then finding the nearest five points (see Figure 6). There are many possible ways of measuring the distance between the two instances with n attribute values, or equivalently between two points in n -dimensional space. The three conditions on any measure of distance implemented by using the following (Bramer, 2007);

1. The distance of any point A from itself is zero, that is $\text{dist}(A,A) = 0$;
2. The distance from A to B is the same as the distance from B to A , that is, $\text{dist}(A,B) = \text{dist}(B,A)$ (the symmetry condition).
3. The intuitive idea is that 'the shortest distance between any two points is a straight line'.

The condition says that for any two points A, B and Z $\text{dist}(A,B) \leq \text{dist}(A,Z) + \text{dist}(Z, \text{points}, B)$.

There are many possible distance measures, but the most popular is almost certainly the Euclidean distance.

$$\text{Euclidean Distance} = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2} \quad (11)$$

By applying Equation (11) to the point (25,10) and the value from Table 9, the five nearest points to the point (25,10) are classified as *very good*, as shown in Table 10.

MIN-MAX NORMALIZATION

Min-Max normalization works by seeing how much greater the field value is than the minimum value $\min(X)$ and scaling this difference by the range. That is

$$X^* = \frac{X - \min(X)}{\text{range}(X)} = \frac{X - \min(X)}{\max(X) - \min(X)} \quad (12)$$

By applying Equation (12) to the dataset, the normalization can be obtained. Table 11 shows the sample of dataset normalization.

Table 9: Dataset for Nearest Neighbor

No.	SPC	SPT	Class	No.	SPC	SPT	Class
1	29.2000	12.6957	VG	53	21.8000	10.3810	VG
2	20.3000	9.2273	EX	54	22.2222	9.0909	EX
3	17.6364	8.0833	EX	55	31.0000	13.6400	VG
4	33.5000	14.3571	VG	56	32.5000	13.5417	VG
5	34.0769	14.7667	VG	57	26.5000	8.8333	EX
6	35.2308	14.7742	VG	58	26.7778	10.9546	VG
7	19.9091	9.9545	EX	59	26.9444	13.4722	VG
8	28.7778	11.7727	VG	60	26.2222	11.2381	VG
9	39.2500	20.4783	G	61	34.5000	14.6364	VG
10	35.1111	16.2051	VG	62	23.6667	10.1429	VG
11	31.7000	15.0952	VG	63	31.2000	12.4800	VG
12	28.5000	14.2500	VG	64	29.9000	10.6786	VG
13	31.5556	12.9091	VG	65	28.7143	14.8889	VG
14	33.4444	15.8421	VG	66	27.0588	11.5000	VG
15	22.0000	11.0000	VG	67	21.1111	10.0000	VG
16	29.5714	12.9375	VG	68	28.6667	12.2857	VG
17	21.4444	10.7222	VG	69	20.4444	9.2000	EX
18	15.0000	7.1429	EX	70	24.2500	8.0833	EX
19	30.2778	12.1111	VG	71	30.8333	12.3333	VG
20	34.7778	17.3889	VG	72	28.8333	8.6500	EX
21	27.5714	12.0625	VG	73	48.5385	21.7586	G
22	30.3077	11.9394	VG	74	28.7778	14.0000	VG
23	27.4000	11.7429	VG	75	18.0000	4.5000	EX
24	27.1818	13.0000	VG	76	17.0000	10.2000	EX
25	33.0625	16.0303	VG	77	23.2500	11.1600	VG
26	31.1364	15.5682	VG	78	21.1429	14.8000	VG
27	19.8333	7.4375	EX	79	22.0000	7.3333	EX
28	27.6667	11.3182	VG	80	17.7500	7.1000	EX
29	26.9333	16.1600	VG	81	14.3333	6.1429	EX
30	36.8000	17.2500	VG	82	22.5455	7.5152	EX
31	34.2000	14.2500	VG	83	20.5000	11.1818	VG
32	27.0000	10.8000	VG	84	22.0000	11.0000	VG
33	32.3200	15.8431	VG	85	38.4167	15.8966	VG
34	24.4444	10.4762	VG	86	29.6667	9.8889	EX
35	58.2500	33.2857	G	87	14.7778	6.6500	EX
36	22.3333	7.4444	EX	88	19.4444	13.4615	EX
37	25.6667	6.4167	EX	89	18.6667	6.2222	EX
38	23.7500	11.8750	VG	90	19.5000	8.2105	EX
39	34.3333	18.7273	VG	91	22.4444	9.1818	EX
40	33.3182	15.5957	VG	92	17.3333	6.5000	EX
41	30.5000	13.5556	VG	93	32.3333	16.1667	VG
42	26.7143	10.1081	VG	94	19.6667	6.2105	EX
43	17.0000	7.8462	EX	95	13.0000	6.5000	EX
44	24.0000	8.8889	EX	96	19.8889	7.1600	EX
45	30.8333	13.2143	VG	97	13.2000	5.5000	EX
46	28.8000	9.0000	EX	98	20.6667	6.8889	EX
47	13.2667	6.6333	EX	99	30.2778	12.1111	VG
48	25.6667	11.0000	VG	100	18.1667	13.6250	EX
49	20.6000	9.3636	EX	101	15.2500	8.7143	EX
50	11.1429	4.3333	EX	102	27.3333	13.3878	VG
51	18.5000	8.4091	EX	103	20.3750	9.5882	EX
52	17.4000	9.1579	EX	104	19.8889	8.9500	EX

The formula for Euclidean distance between points $(a_1, a_2, \dots, a_n)$ and $(b_1, b_2, \dots, b_n)$ in n -dimensional space can be generalized by the following equation:

$$\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \dots + (a_n - b_n)^2} \quad (13)$$

Figure 6: Two-dimensional representation of dataset in Table 9.

Table 10: The five nearest points to the point (25, 10)

SPC	SPT	Class	Distance
24.444	10.476	VG	0.732
25.667	11.000	VG	1.202
23.667	10.143	VG	1.341
24.000	8.889	EX	1.495
26.714	10.108	VG	1.718

By using equation (13) and the training dataset of the five-Nearest Neighbor, the result is as shown in Table 12.

To calculate the predictive accuracy for the k -Nearest Neighbor classification model, equation (10) was used.

Correctly classified instances = 18

Incorrectly classified instances = 6

Therefore, the Predictive accuracy

$$p = (18/24) \times 100 = 75 \text{ percent}$$

$$1 - p = (6/24) \times 100 = 25 \text{ percent}$$

DECISION TREE

Decision trees are powerful and popular for both classification and prediction. The attractiveness of tree-based methods is largely a result of the fact that decision trees represent rules. Rules can readily be expressed in English so that humans can understand them; they can also be expressed in a database access language such as SQL to retrieve records in a particular category. Decision trees are also useful for exploring data to gain insight into the relationships of a large number of candidate input variables to a target variable. Because decision trees combine both data exploration and modeling, they are a powerful first step in the modeling process even when building the final model using some other technique.¹³

Table 11: Sample of normalization dataset

No	School code	SPC	SPT	SRS	PSSS	ERS	STRS
1	53395	0.19439	0.169034	0.404087	0.831311	0.713783	1
2	53401	0.137845	0.129523	0.272819	0.735228	0.644317	0.952652
3	53402	0.474602	0.346217	0.298756	0.995268	0.298967	0.63474
4	53403	0.486849	0.360362	0.382171	0.626383	0.194328	0.772727
5	53405	0.511343	0.360622	0.188805	0.833848	0.287316	0.706745
6	53406	0.186091	0.194154	0.293728	0.730705	0.184656	0.845041
7	53412	0.374358	0.256953	0.357815	0.515209	0.573972	1
8	53416	0.596664	0.557637	0.688683	0.794003	0.58958	0.802372
9	53420	0.436391	0.371711	0.497229	0.873971	0.466916	0.891775
10	53421	0.368461	0.342516	0.351287	0.785439	0.304023	0.797078
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...

Table 12: The five-Nearest Neighbor result

No.	SPC	SPT	SRS	PSSS	ERS	STRS	School class	Five Nearest Neighbor		Predicted school	Error
								Distance	Ranking point		
1	0.3833	0.2888	0.3290	0.6938	0.7639	1.0000	VG	0.2644 0.2777 0.2920 0.3185 0.3272	VG EX VG VG EX	VG	
2	0.5088	0.4100	0.5460	0.6568	0.6637	0.8543	VG	0.2772 0.2913 0.3104 0.3114 0.3405	VG G G VG VG	VG	
3	0.0819	0.0970	0.4543	0.4546	1.0000	0.9459	EX	0.4923 0.4971 0.5204 0.5424 0.5554	EX EX EX EX VG	EX	
4	0.4496	0.3975	0.2740	0.7928	0.4770	0.7549	G	0.1814 0.2176 0.2258 0.2290 0.2340	G VG G G G	G	
5	1.0000	1.0000	0.0076	0.3795	0.5511	1.0000	G	0.3862 0.3997 0.4069 0.4182 0.4472	G VG G G G	G	
6	0.2676	0.2605	0.2054	0.5594	0.4783	0.2898	G	0.3862 0.3997 0.4069 0.4182 0.4472	G VG G G G	G	
7	0.4923	0.4972	0.3681	0.6167	0.4753	0.9483	G	0.1909 0.2769 0.2951 0.2983 0.2987	G G VG G VG	G	

Table 12 *continued*

No.	SPC	SPT	SRS	PSSS	ERS	STRS	School class	Five Nearest Neighbor		Predicted school	Error
								Distance	Ranking point		
8	0.4109	0.3185	0.2309	0.3996	0.5179	0.7790	G	0.1494 0.2585 0.2942 0.3538 0.3688	VG VG VG VG G	VG	✓
9	0.2729	0.1573	0.1006	0.3500	0.6516	0.9158	VG	0.2744 0.3560 0.3684 0.3707 0.4309	VG VG VG VG VG	VG	
10	0.4180	0.3067	0.4394	0.8766	0.9046	0.9188	VG	0.3395 0.4288 0.4474 0.4539 0.4623	EX EX G VG EX	EX	✓
11	0.3748	0.1612	0.2812	0.7490	0.6795	0.6804	VG	0.2706 0.2766 0.2975 0.3141 0.3246	VG VG VG EX VG	VG	
12	0.0451	0.0794	0.5519	0.8470	0.8608	0.8864	EX	0.2949 0.3023 0.3452 0.3909 0.4548	EX EX EX EX VG	EX	
13	0.3083	0.2303	0.5360	0.9176	0.7727	0.9459	EX	0.2190 0.3009 0.3414 0.3548 0.3691	EX EX VG EX G	EX	
14	0.2008	0.1737	0.4729	0.6952	0.7767	1.0000	VG	0.1651 0.2474 0.2594 0.2816 0.3486	EX EX EX EX VG	EX	✓
15	0.4534	0.3181	0.2020	0.2624	0.7195	0.9527	VG	0.2662 0.3487 0.4163 0.4326 0.5474	VG VG VG VG G	VG	
16	0.3260	0.1554	0.0826	0.6171	0.8380	0.9369	VG	0.3521 0.3985 0.4153 0.4230 0.4315	EX VG VG EX EX	EX	
17	0.4958	0.3559	0.5133	0.7361	0.6944	0.8967	VG	0.2736 0.2925 0.3239 0.3306 0.3437	G VG VG G VG	VG	
18	0.2659	0.2007	0.2448	0.6182	0.6595	0.9594	VG	0.1903 0.1954 0.2174 0.2747 0.2786	EX VG VG EX VG	VG	

Table 12 *continued*

No.	SPC	SPT	SRS	PSSS	ERS	STRS	School class	Five Nearest Neighbor		Predicted school	Error
								Distance	Ranking point		
19	0.4258	0.2814	0.4083	0.6402	0.6557	0.9091	VG	0.1871 0.1907 0.2205 0.2703 0.3189	VG VG VG VG VG	VG	
20	0.3982	0.2192	0.6447	0.7200	0.8411	0.7565	VG	0.3976 0.4165 0.4320 0.4357 0.4410	VG EX VG EX EX	EX	✓
21	0.3730	0.3646	0.2696	0.6403	0.5494	0.6633	G	0.2310 0.2373 0.2553 0.2582 0.2597	G VG G VG VG	VG	✓
22	0.3379	0.2475	0.3524	0.5906	0.6291	0.9432	VG	0.1158 0.1285 0.2445 0.2491 0.2805	VG VG VG VG VG	VG	
23	0.4180	0.2763	0.6694	0.8766	0.7109	0.8485	VG	0.3170 0.3320 0.3367 0.3602 0.3676	G EX VG VG VG	VG	
24	0.3755	0.1491	0.5177	0.5931	0.6142	0.8295	VG	0.3976 0.5547 0.3363 0.2090 0.5355	G EX VG VG VG	VG	

Decision trees are an intuitive and widely used type of influence diagram. The basic goal of a decision tree is to find an optimal set of yes-or-no questions that ultimately leads to a correct classification, or probability. The tree must have meaningful criteria for choosing questions, and derive answers from the training dataset.¹⁴

Decision tree C4.5 algorithm

The decision tree C4.5 algorithm is Quinlan's extension of his own ID3 algorithm for generating decision trees.¹⁵ The C4.5 algorithm recursively visits each decision node, selecting the optimal split, until no further splits are possible.

The C4.5 algorithm uses the concept of information gain or entropy reduction to select the optimal split. Suppose the variable X whose k possible values have probabilities $p_1, p_2, \dots, p_k$. What is the smallest number of bits, on average

per symbol, needed to transmit a stream of symbols representing the values of X observed? The answer is called the *entropy* of X and is defined as

$$H(X) = -\sum_j p_j \log_2(p_j) \quad (14)$$

C4.5 uses this concept of entropy as follows. Suppose that we have a candidate split S , which partitions the training dataset T into several subsets, $T_1, T_2, \dots, T_k$.

The mean information requirement can then be calculated as the weighted sum of the entropies for the individual subsets, as follows:

$$H_s(T) = \sum_{i=1}^k p_i H_s(T_i) \quad (15)$$

Where P_i represents the proportion of records in subset i . Define the information gain to be

$$\text{gain}(S) = H(T) - H_S(T) \quad (16)$$

That is, the increase in information produced by partitioning the training data T according to this candidate split S . At each decision node, C4.5 chooses the optimal split to be the split that has the greatest information gain, $\text{gain}(S)$.

Given a set S of cases, C4.5 first grows an initial tree using the divide-and-conquer algorithm as follows:

- If all the cases in S belong to the same class or S is small, the tree is a leaf labeled with the most frequent class in S .
- Otherwise, choose a test based on a single attribute with two or more outcomes. Make this test the root of the tree with one branch for each outcome of the test, partition S into corresponding subsets $S_1, S_2, \dots$ according to

the outcome for each case, and apply the same procedure recursively to each subset.

Equations (14) and (15) are used for the training dataset. Table 13 shows the optimal split that has the greatest information gain is (0.3694); this is achieved with the split by attribute SPC.

Table 14 shows the information gain with the second split node, which was the attribute SRS with maximum Information Gain value (0.4151). Figure 7 shows the final decision tree.

Table 15 shows the result of testing the test dataset of the last 24 instances on the C4.5 decision tree model obtained in Figure 7.

To calculate the predictive accuracy for the C4.5 decision tree Classification model, by using equation (9). Equation (10) was used to calculate the predictive accuracy for the C4.5 decision tree Classification model.

Correctly classified instances = 16

Incorrectly classified instances = 8

Table 13: Information gain for the first split node in C4.5 decision tree

Attribute	School rank			Sum	$P_i H_S(T_i)$	$H_S(T)$	Information Gain= $H(T) - H_S(T)$
	G	VG	EX				
SPC							
G	12	5	0	17	0.8740	0.8297	0.3694
VG	6	32	1	39	0.7852		
EX	0	17	7	24	0.8709		
SPT							
G	13	7	0	20	0.9341	0.9347	0.2645
VG	3	31	3	37	0.8016		
EX	2	16	5	23	1.1492		
SRS							
G	6	9	1	16	1.2476	1.0708	0.1284
VG	12	28	3	43	1.1849		
EX	0	17	4	21	0.7025		
PSSS							
G	5	8	0	13	0.9612	1.0307	0.1685
VG	13	24	5	42	1.3505		
EX	0	22	3	25	0.5294		
ERS							
G	4	20	2	26	0.9913	0.9130	0.2861
VG	14	32	0	46	0.8865		
EX	0	2	6	8	0.8113		
STRS							
G	5	19	1	25	0.9510	1.1275	0.0716
VG	12	27	3	42	1.1981		
EX	1	8	4	13	1.2389		
School rank	18	54	8	80	$H(T)=1.1991$		

Table 14: Information gain for the second split node in C4.5 decision tree

Attribute	School class			Sum	$P_i H_S(T_i)$	$H_S(T)$	Information gain
	G	VG	EX				
SPC=EX							
SPT							
G	—	—	—	0	0.0000	0.8708	0.0001
VG	—	5	2	7	0.8631	—	—
EX	—	12	5	17	0.8740	—	—
SRS							
G	—	4	1	5	0.7219	0.8620	0.0089
VG	—	7	3	10	0.8813	—	—
EX	—	6	3	9	0.9183	—	—
PSSS							
G	—	3	—	3	0.0000	0.7564	0.1145
VG	—	6	5	11	0.9940	—	—
EX	—	8	2	10	0.7219	—	—
ERS							
G	—	10	2	12	0.6500	0.4875	0.3833
VG	—	6	—	6	0.0000	—	—
EX	—	1	5	6	0.6500	—	—
STRS							
G	—	9	1	10	0.4690	0.6324	0.2385
VG	—	7	2	9	0.7642	—	—
EX	—	1	4	5	0.7219	—	—
School rank (SPC=EX)		17	7	24	$H(T)=0.8709$	—	—
SPC=VG							
SPT							
G	1	3	—	4	0.8113	0.7381	0.0471
VG	3	25	1	29	0.6907	—	—
EX	2	4	—	6	0.9183	—	—
SRS							
G	4	5	—	9	0.9911	0.5892	0.1960
VG	2	19	—	21	0.4537	—	—
EX	—	8	1	9	0.5033	—	—
PSSS							
G	2	5	—	7	0.8631	0.6471	0.1381
VG	4	15	—	19	0.7425	—	—
EX	—	12	1	13	0.3912	—	—
ERS							
G	3	10	—	13	0.7793	0.5991	0.1860
VG	3	22	—	25	0.5294	—	—
EX	—	—	1	1	0.0000	—	—
STRS							
G	4	9	—	13	0.8905	0.6802	0.1049
VG	2	17	1	20	0.7476	—	—
EX	—	6	—	6	0.0000	—	—
School rank (SPC=VG)	6	32	1	39	$H(T)=0.7852$	—	—
SPC=G							
SPT							
G	12	4	—	16	0.8113	0.7636	0.1104
VG	—	1	—	1	0.0000	—	—
EX	—	—	—	0	0.0000	—	—
SRS							
G	2	—	—	2	0.0000	0.4588	0.4151
VG	10	2	—	12	0.6500	—	—
EX	—	3	—	3	0.0000	—	—
PSSS							
G	3	—	—	3	0.0000	0.5727	0.3013
VG	9	3	—	12	0.8113	—	—
EX	—	2	—	2	0.0000	—	—
ERS							
G	1	—	—	1	0.0000	0.7382	0.1358
VG	11	4	—	15	0.8366	—	—
EX	—	1	—	1	0.0000	—	—

Table 14 continued

Attribute	School class			Sum	$P_i H_S(T_i)$	$H_S(T)$	Information gain
	G	VG	EX				
STRS							
G	1	1	—	2	1.0000	0.8313	0.0427
VG	10	3	—	13	0.7793	—	—
EX	1	1	—	2	1.0000	—	—
School rank (SPC=G)	12	5	—	17	$H(T)=0.8740$	—	—

Figure 7: The decision tree model.

Therefore, the *Predictive accuracy*

$$p = (16/24) \times 100 = 66.667 \text{ percent}$$

$$1 - p = (8/24) \times 100 = 33.333 \text{ percent}$$

The C4.5 decision tree model was implemented by using data of 80 schools (out of 104 schools) and applying the following indicators: SPC, SPT, SRS, PSSS, ERS and STRS. The model was then tested by using the 24 schools (out of 104 schools). The results showed that there are 16 schools whose predicted *School Performance* is identical to the actual school performance, while there are eight schools whose predicted *School Performance* is different from the actual school performance. Predictive accuracy was 66.667 percent for the C4.5 decision tree model, which is an acceptable percentage in the implementation of data mining projects.

ACCURACY EVALUATION

After testing the three methods of data mining (Nearest Neighbor, Naïve Bayes and the C4.5 decision tree), the three methods were valued by measuring accuracy, *Weka version 3.6* was one of the data mining tools used in accuracy comparison. *Weka* was developed at the University of Waikato in New Zealand. The

dataset used with the *Weka* is the same set of data that were used in the three methods, but converted by equation (13) to values between 0 and 1. Figure 8 shows the *Weka* system test option.

Table 16 shows the results of testing each method of data mining by using different test modes. The results showed that the C4.5 decision tree had a high accuracy rate (95.1923 percent) compared with the other methods, when using the test mode (evaluate on training data).

CONCLUSION

Data mining, also called Knowledge Discovery in Databases, is the field of discovering useful information from large amounts of data. In recent years, there has been increasing interest in the use of data mining to investigate scientific questions within educational research, an area of inquiry termed educational data mining. Educational data mining is the area of scientific inquiry centered on the development of methods for making discoveries within the unique kinds of data that come from educational settings, and using those methods to better understand students and the settings that they learn in.

The results were obtained by applying the normal method in the analysis by using Microsoft 2007; the results showed that the test method (evaluate on training data) achieved the accuracy (87.5 percent, 85.576 percent and 95.192 percent) of the method's Nearest Neighbor, Naïve Bayes and the C4.5 decision tree, respectively. Therefore, the C4.5 decision tree algorithm has a high accuracy rate compared with the other models, which is recommended to be used with education data processing and testing.

Table 15: The C4.5 decision tree result

No	SPC	SPT	SRS	POSSS	ERS	STRS	School class	Predicted school	Error
1	VG	VG	VG	VG	EX	EX	VG	VG	—
2	G	G	EX	VG	EX	VG	VG	VG	—
3	EX	EX	VG	G	EX	EX	EX	EX	—
4	G	G	G	VG	VG	VG	G	G	—
5	G	G	G	G	VG	EX	G	G	—
6	VG	VG	G	G	VG	G	G	VG	✓
7	G	G	VG	G	VG	EX	G	G	—
8	VG	VG	G	G	VG	VG	G	VG	✓
9	VG	VG	G	G	EX	EX	VG	VG	—
10	G	VG	VG	VG	EX	EX	VG	G	✓
11	VG	VG	G	VG	EX	VG	VG	VG	—
12	EX	EX	EX	VG	EX	VG	EX	EX	—
13	VG	VG	EX	EX	EX	EX	EX	VG	✓
14	VG	VG	VG	VG	EX	EX	VG	VG	—
15	G	VG	G	G	EX	EX	VG	G	✓
16	VG	VG	G	G	EX	EX	VG	VG	—
17	G	G	VG	VG	EX	EX	VG	G	✓
18	VG	VG	G	G	EX	EX	VG	VG	—
19	G	VG	VG	G	EX	EX	VG	G	✓
20	VG	VG	EX	VG	EX	VG	VG	VG	—
21	VG	G	G	G	VG	VG	G	VG	✓
22	VG	VG	VG	G	EX	EX	VG	VG	—
23	G	VG	EX	VG	EX	VG	VG	VG	—
24	VG	EX	VG	G	EX	VG	VG	VG	—

Figure 8: Weka system test option.

Table 16: Accuracy of results by using WEKA

Test mode	Method		
	Five-Nearest Neighbor	Naïve Bayes	C4.5 decision tree
Ten-fold cross-validation	84.6154	78.8462	74.0385
Evaluate on training data	87.5000	85.5769	95.1923
Split 70.0 percent train, remainder test	67.7419	70.9677	58.0645
Split 77.0 percent train, remainder test	75.0000	79.1667	50.0000
Split 80.0 percent train, remainder test	76.1905	80.9524	71.4286
Supplied test set (manual calculation result)	75.0000	75.0000	66.6667

REFERENCES

- 1 Tang, Z. and MacLennan, J. (2005) *Data Mining with SQL Server 2005*. Indianapolis, IN: Wiley.
- 2 Bramer, M.A. (2007) *Principles of Data Mining*. London: Springer.
- 3 Turban, E.M. and Wetherbe, J. (2010) *Information Technology for Management: Transforming Business in the Digital Economy*. New York: John Wiley.
- 4 Urbancic, T., Skrjanc, M. and Flach, P. (2002) Web-based analysis of data mining and decision support education. *AI Communications* 15(4): 199–204.
- 5 Behrouz, M.B. and Punch, W.F. (2003) Using genetic algorithms for data mining optimization in an educational web-based system. GECCO 2003 Genetic and Evolutionary Computation Conference. Chicago, IL: Springer-Verlag, pp. 2252–2263.
- 6 Schumann, J.A. (2005) Data mining methodologies in educational organizations. PhD thesis, University of Connecticut, CT, USA.
- 7 HÄubscher, R., Puntambekar, S. and Nye, A.H. (2007) Domain specific interactive data mining. International Conference on User Modeling UM2007. Corfu, Greece: User Modelling Inc, pp. 81–90.
- 8 Romero, C., Ventura, S., Espejo, P.G. and Hervás, C. (2008) Data mining algorithms to classify students. Educational Data Mining Conference EDM 2008, Montreal, Canada, pp. 182–185.
- 9 Pittman, K. (2008) Comparison of data mining techniques used to predict student retention. PhD thesis, Nova Southeastern University, FL, USA.
- 10 The Cross Industry Standard Process for Data Mining Blog, <http://crispdm.wordpress.com/about-crisp-dm-2/>, accessed November 2011.
- 11 Chapman, P. et al (2000) CRISP-DM 1.0: Step-by-Step data mining guide, www.nancygrady.info/CRISP-DM.pdf.
- 12 Zhang, H. (2004) The optimality of Naive Bayes. The 17th International FLAIRS Conference, Miami Beach, FL: Florida Artificial Intelligence Research Symposium.
- 13 Berry, M.J. and Linoff, G.S. (2004) *Data Mining Techniques for Marketing Sales and Customer Relationship Management*, 3rd edn. Indianapolis, IN: Wiley Publishing.
- 14 Murray, E. (2005) Using decision trees to understand student data. The 22nd International Conference on Machine Learning, Bonn, Germany.
- 15 Quinlan, J.R. (1994) C4.5: Programs for machine learning. *Machine Learning* 16(3): 235–240.