Journal of Information Technology Education: Innovations in Practice **Volume 9, 2010**

A Tools-Based Approach to Teaching Data Mining Methods

Musa J. Jafar
West Texas A&M University
Canyon, TX, USA

mjafar@mail.wtamu.edu

Executive Summary

Data mining is an emerging field of study in Information Systems programs. Although the course content has been streamlined, the underlying technology is still in a state of flux. The purpose of this paper is to describe how we utilized Microsoft Excel's data mining add-ins as a front-end to Microsoft's Cloud Computing and SQL Server 2008 Business Intelligence platforms as back-ends to teach a senior level data mining methods class. The content presented and the hands on experience gained have broader applications in other areas, such as accounting, finance, general business, and marketing. Business students benefit from learning data mining methods and the usage of data mining tools and algorithms to analyze data for the purpose of decision support in their areas of specialization.

Our intention is to highlight these newly introduced capabilities to faculty currently teaching a business intelligence course. Faculty interested in expanding their teaching portfolio to the data mining and the business intelligence areas may also benefit from this article.

This set of integrated tools allowed us to focus on teaching the analytical aspects of data mining and the usage of algorithms through practical hands-on demonstrations, homework assignments, and projects. As a result, students gained a conceptual understanding of data mining and the application of data mining algorithms for the purpose of decision support. Without such a set of integrated tools, it would have been prohibitive for faculty to provide comprehensive coverage of the topic with practical hands-on experience.

The availability of this set of tools transformed the role of a student from a programmer of data mining algorithms to a business intelligence analyst. Students now understand the algorithms and use tools to perform (1) elementary data analysis, (2) configure and use data mining computing engines to build, test, compare and evaluate various mining models, and (3) use the mining models to analyze data and predict outcomes for the purpose of decision support. If it was not for the underlying technologies that we used, it would have been impossible to cover such material in a one-semester course and provide students with much needed hands-on experience in data mining.

Finally, what we presented is how to utilize the cloud as a computing platform that transformed the role of a student from "doing low-level IT" in a data mining course to a business intelligence analyst using tools to analyze data for the purpose of decision support.

Editor: John Beachboard

Keywords: Data mining, Decision Support, Business Intelligence, Excel Data mining Add-ins, Cloud Computing.

Introduction

Data mining is the process of discovering useful and previously unknown information and relationships in large data sets (Campos, Stengard, & Milenova, 2005; Tan, Steinbach, & Kumar, 2006). Accordingly, data mining is the purposeful use of information technology to implement algorithms from machine learning, statistics, and artificial intelligence to analyze large data sets for the purpose of decision support.

The field of data mining grew out of limitations in standard data analysis techniques (Tan et al., 2006). Advancements in machine learning, pattern recognition, and artificial intelligence algorithms coupled with computing trends (CPU power, massive storage devices, high-speed connectivity, and software academic initiatives from companies like Microsoft, Oracle, and IBM) enabled universities to bring data mining courses into their curricula (Jafar, Anderson, & Abdullat, 2008b). Accordingly, Computer Science and Information Systems programs have been aggressively introducing data mining courses into their curricula (Goharian, Grossman, & Raju, 2004; Jafar, Anderson, & Abdullat 2008a; Lenox & Cuff, 2002; Saquer, 2007).

Computer Science programs focus on the deep understanding of the mathematical aspects of data mining algorithms and their efficient implementation. They require advanced programming and data structures as prerequisites for their courses (Goharian et al., 2004; Musicant, 2006; Rahal, 2008).

Information Systems programs on the other hand, focus on the data analysis and business intelligence aspects of data mining. Students learn the theory of data mining algorithms and their applications. Then they use tools that implement the algorithms to build mining models to analyze data for the purpose of decision support. Accordingly, a first course in programming, a database management course, and a statistical data analysis course suffice as prerequisites. For Information Systems programs, a data centric, algorithm understanding and process-automation approach to data mining similar to Jafar et al. (2008a) and Campos et al. (2005) is more appropriate. A data mining course in an Information Systems program has an (1) analytical component, (2) a tools-based, hands-on component ,and (3) a rich collection of data sets.

(1) The analytical component covers the theory and practice of the lifecycle of a data mining analysis project, elementary data analysis, market basket analysis, classification and prediction (decision trees, neural networks, naïve Bayes, logistic regression, etc.), cluster analysis and category detection, testing and validation of mining models, and finally the application of mining models for decision support and prediction. Textbooks from Han and Kamber (2006) and Tan et al. (2006) provide a comprehensive coverage of the terminology, theory, and algorithms of data mining.

(2) The hands-on component requires the use of tools to build projects based on the algorithms learned in the analytical component. We chose Microsoft Excel with its data mining add-in(s) as the front-end and Microsoft's Cloud Computing and SQL Server 2008 data mining computing engines as the back-end. Microsoft Excel is ubiquitous. It is a natural front-end for elementary data analysis and presentation of data. Its data mining add-in(s) are available as a free download. The add-in(s) are automatically configured to send data to Microsoft's Cloud Computing engine server. The server performs the necessary analysis and receives analysis results back into Excel to present them in tabulated and chart formats. Using wizards, the add-in(s) are easily configured to connect to a SQL Server 2008 running analysis services to send data and receive analysis results back into Excel for presentation. The add-in(s) provide a rich wizard-based, uniform graphical user interface to manage the data, the data mining models, the configurations, and the pre and

post view of data and mining models. They provide a workbench for an end-to-end life-cycle support of a data mining project.

(3) A rich collection of data sets to demonstrate the capabilities of the algorithms and to provide practical hands-on experience, homework assignments and projects is vital. We chose to use the Iris data set, the Mushrooms data set, and the Bikebuyers data set. The Iris and the Mushrooms data sets are public domain data sets available from the UCI repository (University of California Irvine, 2009). In this paper, we use the Iris and the Mushrooms data sets for elementary data analysis, classification, and clustering analysis. The Iris data set attributes are quantitative. It is composed of 150 records of: **Iris(sepalLength, sepalWidth, petalLength, petalWidth, iris-Type)** for a total of 4 attributes in each record. The length and width attributes are in cm and the classification (iris-type) is *setosa*, *versicolor*, or *virginica*. The Mushrooms data set attributes are qualitative. It is composed 8,124 records of: **Mushroom(capShape, capSurface,…, odor, ring-Type, habitat, gillSize, ….., classification)** for a total of 21 attributes in each record. The classification of a mushroom is either *poisonous* or *edible*. The Bikebuyers data set is available from Microsoft Corporation in support of their Business Intelligence set of tools. We will use this data set for market basket analysis. The data set is composed of 121,300 records of: **Bike-Buyer(salesOrderNumber, quantity, productName, model, subcategory, category**) for a total of 6 attributes in each record.

The Data Mining Process

Given a data set, the data mining process starts with elementary data analysis. It allows an analyst to understand the characteristics of the attributes of the data set (dependency, ranges, max, min, count, average, standard deviation, quartiles, outliers, etc.). The data set is then divided into a training data set and a testing and validation data set (holdout). The training data set is used to build the mining structure and associated mining models. Using accuracy charts, the holdout set is used to test the accuracy and the efficiency of the mining models. If a model is valid and its accuracy is acceptable, it is then used for prediction (Figure 1). The data mining process is very elabo-

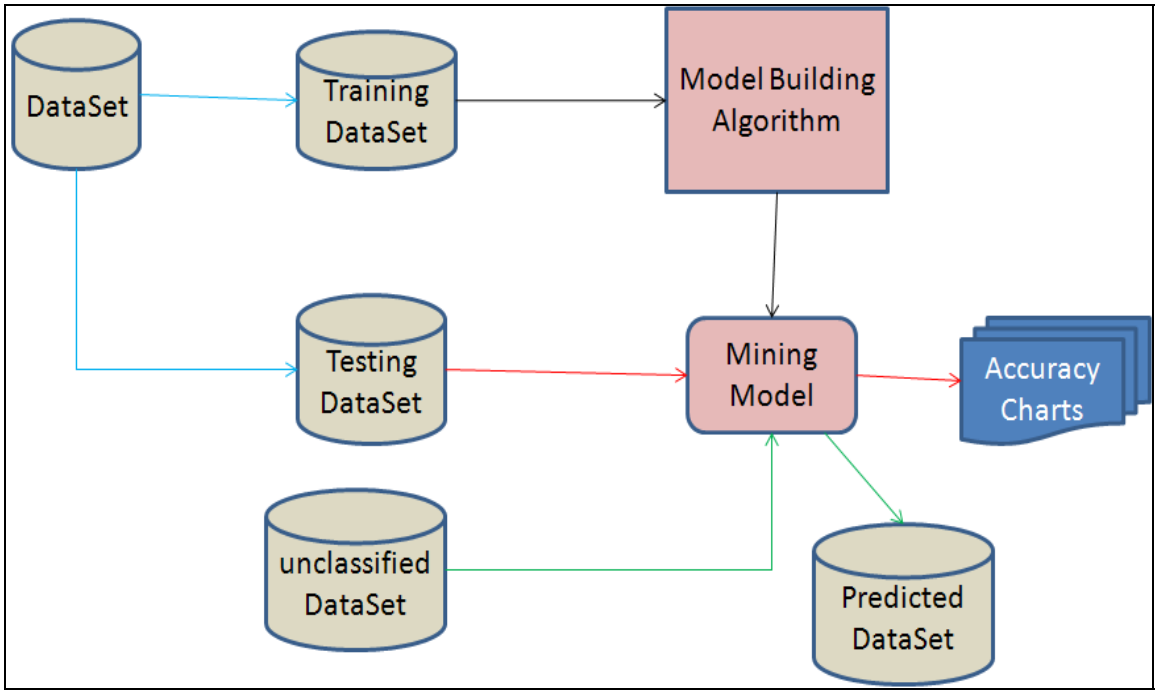


Figure 1: High Level Flow of Data Mining Activities

rate. It requires extensive front-end processing and presentation of results. In the past, our students had to write their own macro(s) to randomly split a data set into training and testing sets. With Excel's data mining add-ins, this task is performed automatically through configuration. It allowed both faculty and students to focus on the configuration and analysis tasks instead of writing random number generator macros to split the data.

For the rest of the paper, we will take a hands-on approach to data mining with examples. Each section will be composed of a theory component of data mining followed by a practice component. We will review the basic topics taught in a standard data mining course. These topics are (1) elementary data analysis and outlier detection, (2) market basket analysis or association rules analysis, (3) classification and prediction, and (4) clustering analysis. The last section of the paper is a summary and conclusions section. The topics, subtopics and, terminology used can be found in standard data mining text books such as Han and Kamber (2006) and Tan et al. (2006).

Elementary Data Analysis

Elementary data analysis is the first basic step in data mining. It allows an analyst to understand the intricacies of a data set, the characteristics of each attribute, and the dependencies between attributes.

The Theory

Students learn the concepts of elementary data analysis through book chapters, lecture notes, lectures, and homework assignments. Students learn to: (1) Classify the data type of each attribute (quantitative, qualitative, continuous, discrete, or binary) and its scale of measure (nominal, ordinal, interval, or ratio); (2) Produce summary statistics for each quantitative attribute (mean, median, mode, min, max, quartiles, etc.); (3) Visualize data through histograms, scatter plots, and box plots; (4) Produce hierarchical data analysis through pivot tables and pivot charts; (5) Produce and analyze the various correlation matrices and key influencers of attributes; (6) In preparation for data mining, attributes and their values may also need to be relabeled, grouped, or normalized.

The Practice

The hands-on practice of elementary data analysis is performed in Excel. Most of the elementary data analysis tasks can be performed in Excel's charting, sorting, tables, and pivot tables functions (King, 2009; Tang, 2008). The data analysis add-in tools allow students to generate descriptive statistics, correlation matrices, histograms, percentiles, and scatter plots. Using wizards, a student can easily produce summary statistics similar to those in Figures 2, 3, and 4.

In Excel, the table, pivot table, and charts tools allow students to perform various hierarchical analyses and relabeling of data. Elementary data analysis of the Iris data set is shown in Figure 3. It is easy to see that all petal lengths, petal widths, and sepal lengths of the *setosa*(s) are at the low end of the scale. The petal lengths and petal widths of the *virginica*(s) are at the high end of the scale. Filters can also be added on top of a row or column contents to produce hierarchical representation of the data to provide more resolution during data analysis.

	SepalLength	SepalWidth	PetalLength	PetalWidth
Mean	5.84	3.05	3.76	1.20
Standard Error	0.07	0.04	0.14	0.06
Median	5.80	3.00	4.35	1.30
Mode	5.00	3.00	1.50	0.20
Standard Deviation	0.83	0.43	1.76	0.76
Sample Variance	0.69	0.19	3.11	0.58
Kurtosis	−0.55	0.29	−1.40	−1.34
Skewness	0.31	0.33	−0.27	−0.10
Range	3.60	2.40	5.90	2.40
Minimum	4.30	2.00	1.00	0.10
Maximum	7.90	4.40	6.90	2.50
Sum	876.50	458.10	563.80	179.80
Count	150.00	150.00	150.00	150.00
Confidence Level(95	0.13	0.07	0.28	0.12

	SepalLength	SepalWidth	PetalLength	PetalWidth
SepalLength	0.68			
SepalWidth	−0.04	0.19		
PetalLength	1.27	−0.32	3.09	
PetalWidth	0.51	−0.12	1.29	0.58

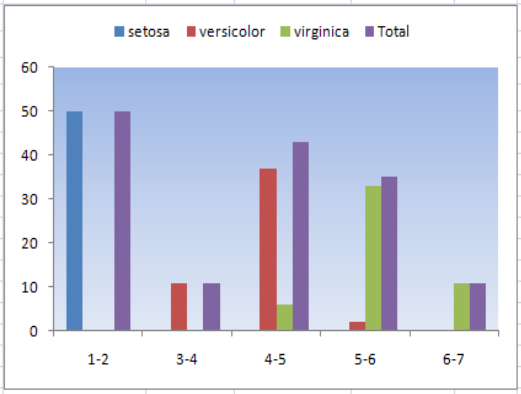
Figure 2: Descriptive Statistics and Correlation Matrix Results from Excel

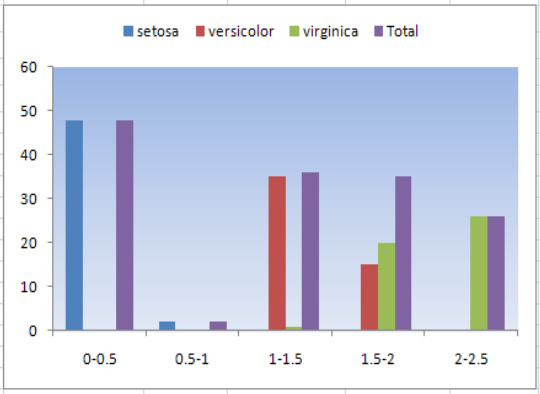
Using Excel's data mining add-ins, we can also analyze the overall key influencers of an iris type. We can measure the relative impact of each attribute value on the classification of an iris (Figure 4). We can also perform pair wise comparisons between the different classifications. Based on an analytical model, the key influencers tool automatically breaks the range of a continuous attribute into intervals while determining the influencers of the iris type. The algorithm discovered that a *petalWidth < 0.4125* strongly favored the *setosa* classification. A *petalWidth* in the range of *[0.4125, 1.33]* strongly favored a *versicolor* classification and a *petalLength >=5.48* strongly favored a *virginica* classification. A student can use this visual analysis and presentation of the key influencers to build the initial classification rules for an expert system. Excel tools allow students to perform pair wise discrimination for key influencers of the different types of classifications. The length of the "Relative Impact" bar indicates the relative importance of each attribute range to the corresponding classification (Figure 4).

Excel's data exploration tool allows users to interactively produce histograms and configure bucket counts. The data clean-up tool permits users to interactively create line charts and specify ranges for outliers of numeric data. With the data sampling tool users interactively divide the data into different random samples. Users are able to interactively re-label data into ranges such as low, medium, high with the re-labeling tool.

Petal Length	setosa	versicolor	virginica	Total
1–2	50			50
3–4		11		11
4–5		37	6	43
5–6		2	33	35
6–7			11	11
Total	50	50	50	150

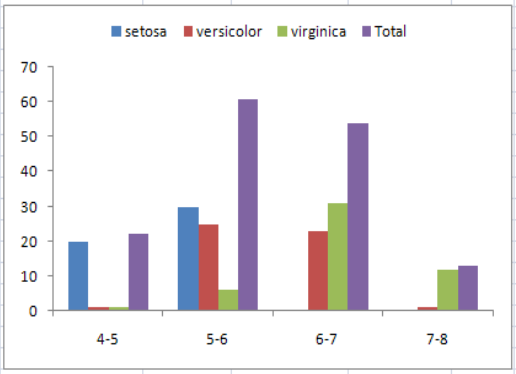
Petal Width	setosa	versicolor	virginica	Total
0–0.5	48			48
0.5–1	2			2
1–1.5		35	1	36
1.5–2		15	20	35
2–2.5			26	26
2.5–3			3	3
Grand Total	50	50	50	150





Sepal Length	setosa	versicolor	virginica	Total
4–5	20	1	1	22
5–6	30	25	6	61
6–7		23	31	54
7–8		1	12	13
Grand Total	50	50	50	150

Sepal Width	setosa	versicolor	virginica	Total
0–0.5	1	9	1	11
0.5–1	1	25	20	46
1–1.5	27	16	26	69
1.5–2	17		3	20
2–2.5	4			4
2.5–3	50	50	50	150



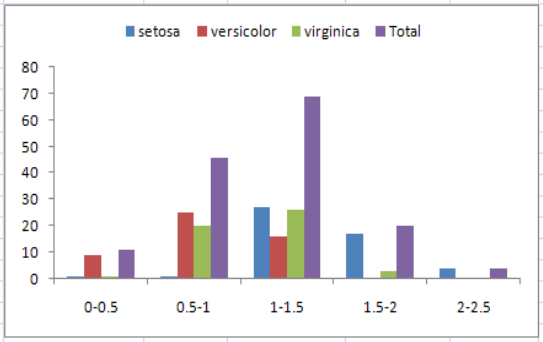


Figure 3: Pivot Tables and Analysis Charts

Key Influencers Report for 'Iris'

Key Influencers and their impact over the values of 'Iris'

Filter by 'Column' or 'Favors' to see how various columns influence 'Iris'

Column	Value	Favors	Relative Impact
PetalWidth	< 0.4125333333	Iris-setosa	
PetalLength	< 1.8284842842	Iris-setosa	
SepalLength	< 5.1178229176	Iris-setosa	
SepalWidth	3.263544748 - 3.7617159688	Iris-setosa	
SepalWidth	>= 3.7617159688	Iris-setosa	
PetalWidth	0.4125333333 - 1.3334644818	Iris-versicolor	
PetalLength	4.107475968 - 4.8017055472	Iris-versicolor	
PetalLength	1.8284842842 - 4.107475968	Iris-versicolor	
PetalWidth	1.3334644818 - 1.5863925018	Iris-versicolor	
SepalWidth	< 2.6057250916	Iris-versicolor	
SepalLength	5.1178229176 - 5.90955178	Iris-versicolor	
SepalWidth	2.6057250916 - 3.050809412	Iris-versicolor	
PetalLength	>= 5.482571752	Iris-virginica	
PetalWidth	>= 2.0939782008	Iris-virginica	
PetalWidth	1.5863925018 - 2.0939782008	Iris-virginica	
PetalLength	4.8017055472 - 5.482571752	Iris-virginica	
SepalLength	>= 7.042132968	Iris-virginica	
SepalLength	6.4071383016 - 7.042132968	Iris-virginica	
SepalLength	5.90955178 - 6.4071383016	Iris-virginica	
SepalWidth	2.6057250916 - 3.050809412	Iris-virginica	

Discrimination between factors leading to 'Iris-setosa' and 'Iris-versicolor'

Filter by 'Column' to see how different values favor 'Iris-setosa' or 'Iris-versicolor'

Column	Value	Favors Iris-setosa	Favors Iris-versicolor
PetalWidth	< 0.4125333333		
PetalLength	< 1.8284842842		
SepalLength	< 5.1178229176		
PetalLength	4.107475968 - 4.8017055472		
PetalWidth	0.4125333333 - 1.3334644818		
PetalWidth	1.3334644818 - 1.5863925018		
SepalWidth	3.263544748 - 3.7617159688		
SepalLength	5.90955178 - 6.4071383016		
PetalLength	1.8284842842 - 4.107475968		
SepalWidth	< 2.6057250916		
SepalWidth	2.6057250916 - 3.050809412		
SepalWidth	>= 3.7617159688		
SepalLength	6.4071383016 - 7.042132968		
PetalWidth	1.5863925018 - 2.0939782008		
PetalLength	4.8017055472 - 5.482571752		

Figure 4: Key Influencers Analysis Results

Market Basket Analysis

Market basket analysis allows a retailer to understand the purchasing behavior of customers. It predicts products that customers may purchase together. It allows retailers to bundle products, offer promotions on products or suggest products that have not yet been added to the basket. Market basket analysis can be used to analyze the browsing behavior of students inside a course management system by modeling each visit as a market basket and the click stream of a student as a set of items inside a basket.

The Theory

Students learn the theoretical foundations and concepts of association analysis through book chapters, lecture notes, lectures, and homework assignments. Students learn conditional probabil-

ity concepts and Bayesian statistics, the concepts of item sets, item set support and its calculation, frequent item sets, closed frequent item sets, association rules, rule support and its calculations, rule confidence and its calculations, rule strength and its calculations, rule importance and its calculations, correlation analysis and the lift of association rules and their calculations, apriori and general algorithms for generating frequent item sets from a market basket set, apriori and general algorithms for generating association rules from a frequent item set. Given a small market basket and a set of thresholds, students should be able to use algorithms to manually (apriori and confidence-based pruning) generate the pruned item set, detect closed item sets, generate rules and calculate their support, and the confidence and importance of rules as shown in the activity diagram in Figure 5. The cited textbooks present algorithms in complex English-like structures with mathematical notations. It is more helpful to students when faculty visually present an algorithm by flowcharting it as a UML-based activity diagram and then use an example to demonstrate the algorithm in action. Figure 5 is an activity diagram of the apriori algorithm for discovering fre-

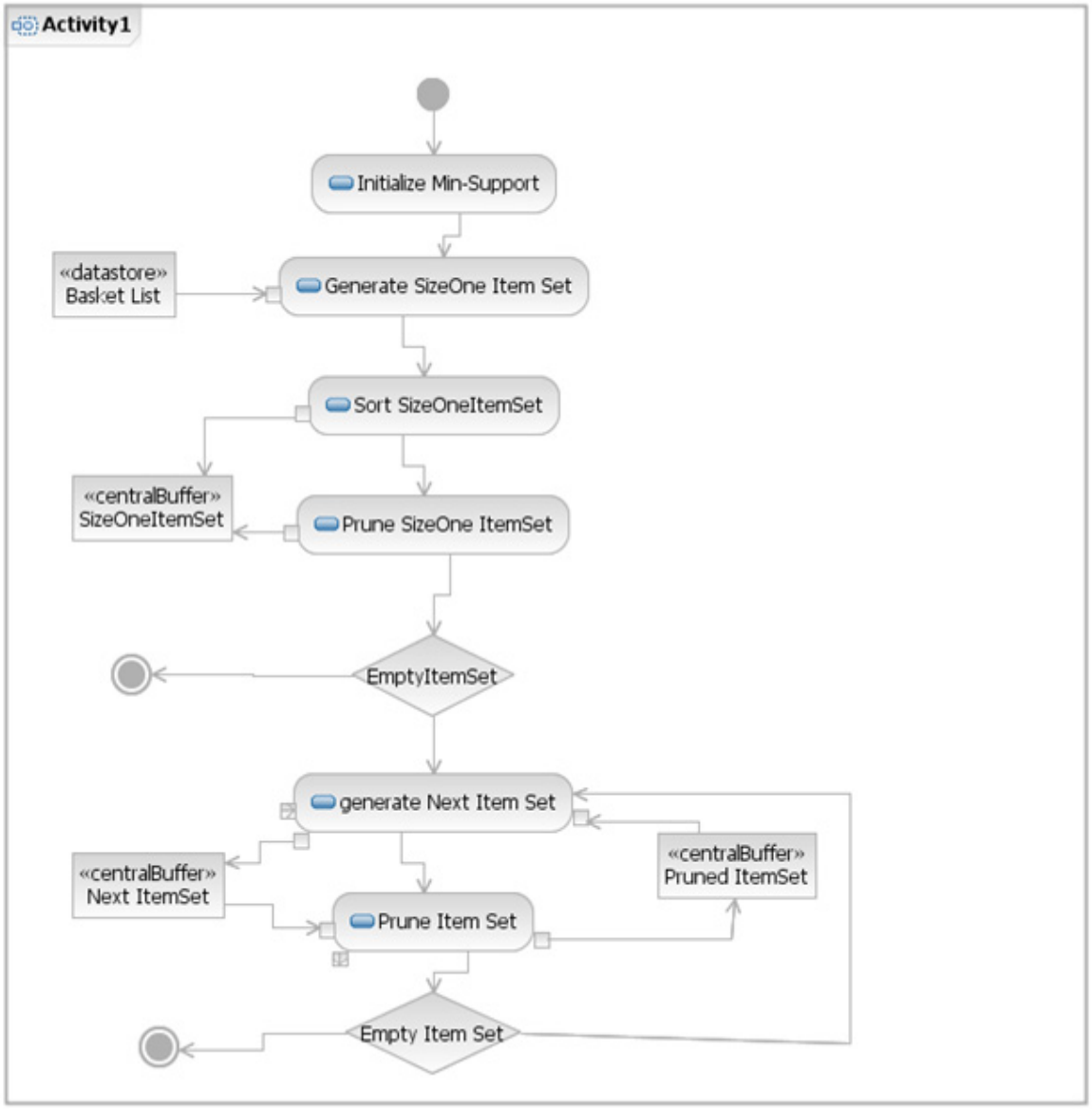


Figure 5: An Activity Diagram of the Apriori Item Set Generation Algorithm

quent item sets (item sets with two or more items). Figure 6 is an example implementation of the algorithm as it applies to an item set of purchases. In Figure 6 we started with 6 transactions and then used the apriori algorithm to generate all the frequent item sets with a minimum support of 2. We stopped when no item sets with the minimum threshold support could be generated. Item sets with support less than 2 (the red item sets) were successively eliminated.

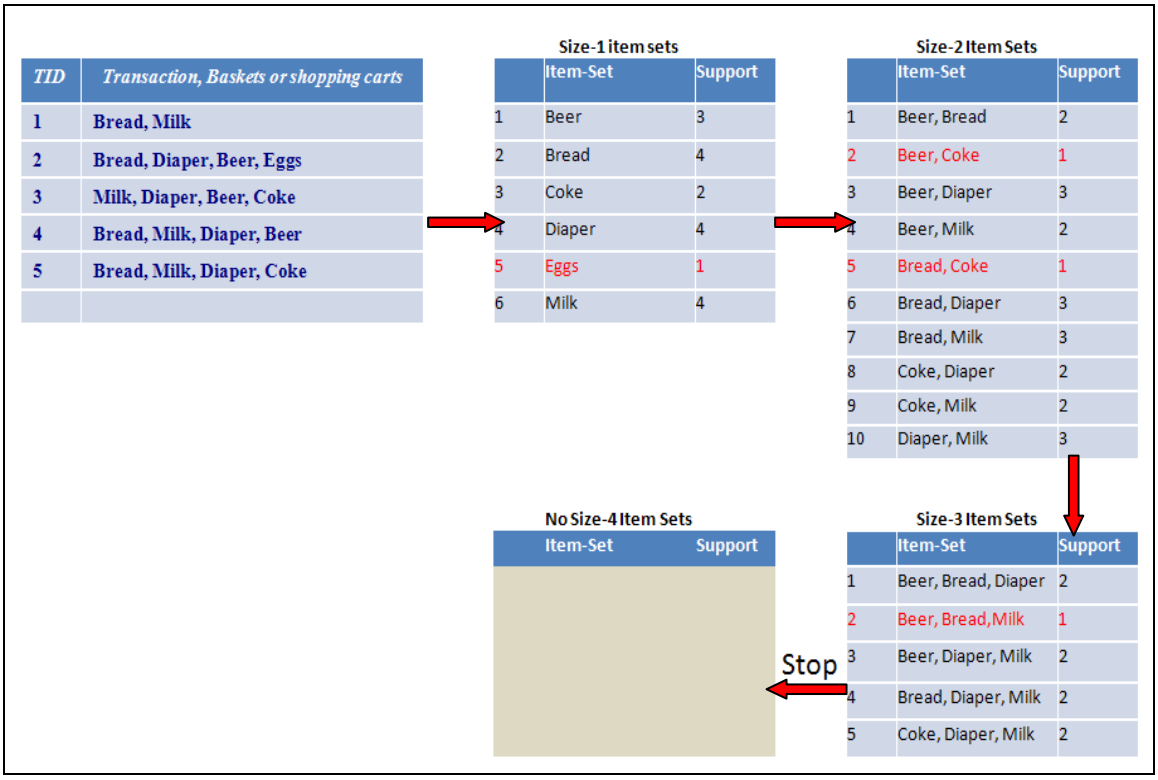


Figure 6: An Example Implementation of the Apriori Algorithm

The Practice

The Bikebuyers data set has 31,450 sales orders with 121,300 recorded items for 266 different products spanning 35 unique categories and 107 different models. Each record is a sales order that describes the details of the items sold (**salesOrderNumber**, **quantity**, **productName**, **modelName**, **subcategoryName**, **categoryName**). Figure 7 is a sample of the data set. According to the sample, *SO43659* has 12 different products as follows: one "Mountain-100 Black, 42", three "Mountain-100 Black, 44"… and four "Sport-100 Helmet Blue".

After learning the theory, students use this large data set to perform market basket analysis. They use wizards to build and configure mining structures based on the "Model Name" of the items sold. The wizards allow students to configure the parameters, thresholds, and probabilities for item sets and association rules for the market basket analysis. Students can run multiple association analysis scenarios and analyze the item sets and their rules. Figure 8 is the output of a market basket analysis. It is composed of three tab-groups: (1) the association rules that were predicted, (2) the frequent item sets that were computed (detailed in Figure 11), and (3) the dependency network between the items. Figure 8 shows the output after executing a calibrated association analysis algorithm with minimum probability of 0.4 and importance threshold of 0.23. The algorithm concluded that those who bought an "All-Purpose Bike Stand" and a "HL Road Tire" also bought a "Tire Tube" with a probability of *0.941* and an importance of *1.080*. In other words, **All-**

Purpose Bike Stand & HL Road Tire → Road Tire Tube (0.941, 1.08).

For a rule such as A→B, the importance is measured by calculating the $\log \frac{prob(B/A)}{Prob(B/\neg A)}$.

SO	Qty	ProductName	ModelName	SubCategoryName	CategoryName
SO43659	1	Mountain-100 Black, 42	Mountain-100	Mountain Bikes	Bikes
SO43659	3	Mountain-100 Black, 44	Mountain-100	Mountain Bikes	Bikes
SO43659	1	Mountain-100 Black, 48	Mountain-100	Mountain Bikes	Bikes
SO43659	1	Mountain-100 Silver, 38	Mountain-100	Mountain Bikes	Bikes
SO43659	1	Mountain-100 Silver, 42	Mountain-100	Mountain Bikes	Bikes
SO43659	2	Mountain-100 Silver, 44	Mountain-100	Mountain Bikes	Bikes
SO43659	1	Mountain-100 Silver, 48	Mountain-100	Mountain Bikes	Bikes
SO43659	3	Long-Sleeve Logo Jersey, M	Long-Sleeve Logo Jersey	Jerseys	Clothing
SO43659	1	Long-Sleeve Logo Jersey, XL	Long-Sleeve Logo Jersey	Jerseys	Clothing
SO43659	6	Mountain Bike Socks, M	Mountain Bike Socks	Socks	Clothing
SO43659	2	AWC Logo Cap	Cycling Cap	Caps	Clothing
SO43659	4	Sport-100 Helmet, Blue	Sport-100	Helmets	Accessories
SO43660	1	Road-650 Red, 44	Road-650	Road Bikes	Bikes
SO43660	1	Road-450 Red, 52	Road-450	Road Bikes	Bikes
SO43661	1	HL Mountain Frame - Black, 48	HL Mountain Frame	Mountain Frames	Components
SO43661	1	HL Mountain Frame - Black, 42	HL Mountain Frame	Mountain Frames	Components
SO43661	2	HL Mountain Frame - Black, 38	HL Mountain Frame	Mountain Frames	Components
SO43661	4	AWC Logo Cap	Cycling Cap	Caps	Clothing
SO43661	4	Long-Sleeve Logo Jersey, L	Long-Sleeve Logo Jersey	Jerseys	Clothing
SO43661	2	HL Mountain Frame - Silver, 46	HL Mountain Frame	Mountain Frames	Components
SO43661	3	Mountain-100 Black, 38	Mountain-100	Mountain Bikes	Bikes
SO43661	2	Mountain-100 Black, 48	Mountain-100	Mountain Bikes	Bikes
SO43661	2	Sport-100 Helmet, Blue	Sport-100	Helmets	Accessories
SO43661	2	HL Mountain Frame - Silver, 48	HL Mountain Frame	Mountain Frames	Components
SO43661	4	Mountain-100 Black, 42	Mountain-100	Mountain Bikes	Bikes
SO43661	2	Mountain-100 Silver, 44	Mountain-100	Mountain Bikes	Bikes
SO43661	2	Long-Sleeve Logo Jersey, XL	Long-Sleeve Logo Jersey	Jerseys	Clothing
SO43661	2	Mountain-100 Black, 44	Mountain-100	Mountain Bikes	Bikes
SO43661	5	Sport-100 Helmet, Black	Sport-100	Helmets	Accessories
SO43662	3	Road-650 Red, 52	Road-650	Road Bikes	Bikes
SO43662	5	Road-650 Black, 52	Road-650	Road Bikes	Bikes
SO43662	2	LL Road Frame - Red, 62	LL Road Frame	Road Frames	Components
SO43662	4	Road-450 Red, 58	Road-450	Road Bikes	Bikes
SO43662	3	LL Road Frame - Red, 44	LL Road Frame	Road Frames	Components
SO43662	5	Road-650 Red, 44	Road-650	Road Bikes	Bikes
SO43662	3	Road-650 Black, 58	Road-650	Road Bikes	Bikes
SO43662	2	Road-650 Black, 44	Road-650	Road Bikes	Bikes
SO43662	1	Road-150 Red, 56	Road-150	Road Bikes	Bikes
SO43662	1	Road-450 Red, 44	Road-450	Road Bikes	Bikes
SO43662	3	Road-650 Red, 48	Road-650	Road Bikes	Bikes
SO43662	1	ML Road Frame - Red, 48	ML Road Frame	Road Frames	Components
SO43662	6	Road-450 Red, 52	Road-450	Road Bikes	Bikes

Figure 7: A Sample of the Data Set

Figure 9 is an Excel export of the rules tab in Figure 8. Both the importance and probability of each rule are presented as a bar and a numeric value for validation, further analysis, sorting, filtering, etc. The user interface also allows an analyst to select a rule and drill through it to the record cases associated with that rule. Figure 10 is a drill through the top rule of Figure 8. Figure 11 is an Excel export of the item sets tab of Figure 8. Using the dependency network tab, students can also explore the item sets generated and the strength of dependencies between items. Students then store their results in work sheets for further data analysis.

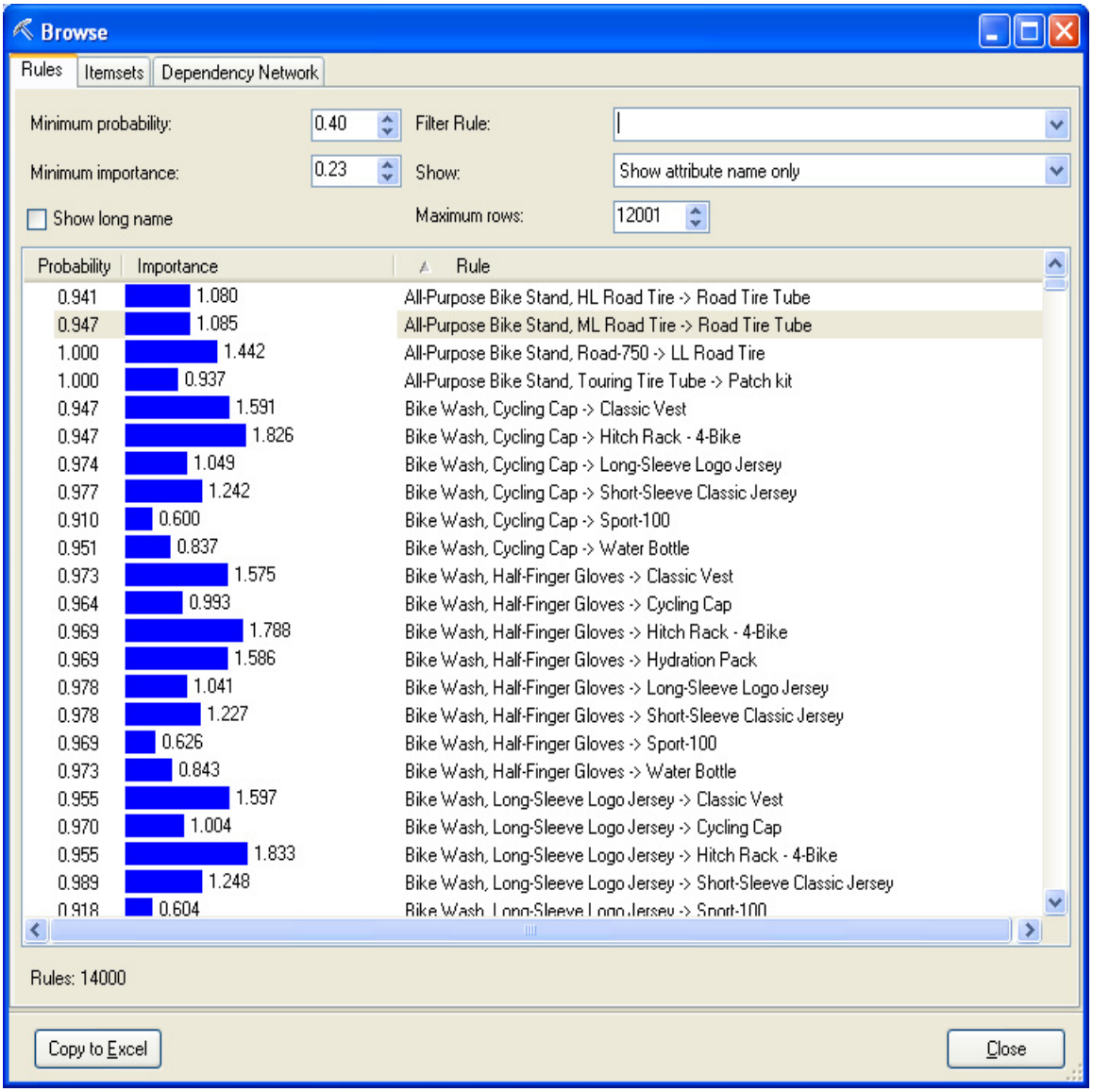


Figure 8: The Main Screen of the Market Basket Analysis

Associate By ModelName

Rules

Probability	Importance	Rule
0.941	1.080	All-Purpose Bike Stand, HL Road Tire -> Road Tire Tube
0.947	1.085	All-Purpose Bike Stand, ML Road Tire -> Road Tire Tube
1.000	1.442	All-Purpose Bike Stand, Road-750 -> LL Road Tire
1.000	0.937	All-Purpose Bike Stand, Touring Tire Tube -> Patch kit
0.947	1.591	Bike Wash, Cycling Cap -> Classic Vest
0.947	1.826	Bike Wash, Cycling Cap -> Hitch Rack - 4-Bike
0.974	1.049	Bike Wash, Cycling Cap -> Long-Sleeve Logo Jersey
0.977	1.242	Bike Wash, Cycling Cap -> Short-Sleeve Classic Jersey
0.910	0.600	Bike Wash, Cycling Cap -> Sport-100
0.951	0.837	Bike Wash, Cycling Cap -> Water Bottle
0.973	1.575	Bike Wash, Half-Finger Gloves -> Classic Vest
0.964	0.993	Bike Wash, Half-Finger Gloves -> Cycling Cap
0.969	1.788	Bike Wash, Half-Finger Gloves -> Hitch Rack - 4-Bike
0.969	1.586	Bike Wash, Half-Finger Gloves -> Hydration Pack
0.978	1.041	Bike Wash, Half-Finger Gloves -> Long-Sleeve Logo Jersey
0.978	1.227	Bike Wash, Half-Finger Gloves -> Short-Sleeve Classic Jersey
0.969	0.626	Bike Wash, Half-Finger Gloves -> Sport-100
0.973	0.843	Bike Wash, Half-Finger Gloves -> Water Bottle
0.955	1.597	Bike Wash, Long-Sleeve Logo Jersey -> Classic Vest
0.970	1.004	Bike Wash, Long-Sleeve Logo Jersey -> Cycling Cap
0.955	1.833	Bike Wash, Long-Sleeve Logo Jersey -> Hitch Rack - 4-Bike
0.989	1.248	Bike Wash, Long-Sleeve Logo Jersey -> Short-Sleeve Classic Jersey
0.918	0.604	Bike Wash, Long-Sleeve Logo Jersey -> Sport-100
0.963	0.843	Bike Wash, Long-Sleeve Logo Jersey -> Water Bottle
0.922	0.796	Bike Wash, Mountain Bottle Cage -> Water Bottle
0.901	1.364	Bike Wash, Road-550-W -> Road-350-W
0.912	1.215	Bike Wash, Road-550-W -> Road-750
0.868	0.774	Bike Wash, Road-550-W -> Water Bottle
0.948	1.594	Bike Wash, Short-Sleeve Classic Jersey -> Classic Vest
0.967	1.003	Bike Wash, Short-Sleeve Classic Jersey -> Cycling Cap
0.948	1.830	Bike Wash, Short-Sleeve Classic Jersey -> Hitch Rack - 4-Bike
0.981	1.053	Bike Wash, Short-Sleeve Classic Jersey -> Long-Sleeve Logo Jersey

Figure 9: An Excel Export of the Rules Tab

Drill through for model 'Associate By ModelName'

Cases Classified to:
All-Purpose Bike Stand, HL Road Tire -> Road Tire Tube

	SalesOrderNumber	ModelName Table.ModelName
1	SO51179	Road-250
2	SO51179	HL Road Tire
3	SO51179	Road Tire Tube
4	SO51179	All-Purpose Bike Stand
5	SO53997	HL Road Tire
6	SO53997	Road Tire Tube
7	SO53997	Patch kit
8	SO53997	All-Purpose Bike Stand
9	SO54601	Road-250
10	SO54601	HL Road Tire
11	SO54601	Road Tire Tube
12	SO54601	Patch kit
13	SO54601	All-Purpose Bike Stand
14	SO54780	Road Tire Tube
15	SO54780	HL Road Tire
16	SO54780	All-Purpose Bike Stand
17	SO54780	Short-Sleeve Classic Jersey
18	SO55560	HL Road Tire
19	SO55560	Road Tire Tube
20	SO55560	All-Purpose Bike Stand
21	SO55560	Classic Vest
22	SO55984	Road Tire Tube
23	SO55984	HL Road Tire
24	SO55984	All-Purpose Bike Stand
25	SO56027	Road-250
26	SO56027	Road Tire Tube
27	SO56027	HL Road Tire
28	SO56027	All-Purpose Bike Stand
29	SO56162	HL Road Tire
30	SO56162	Road Tire Tube
31	SO56162	All-Purpose Bike Stand
32	SO56347	HL Road Tire
33	SO56347	Road Tire Tube
34	SO56347	All-Purpose Bike Stand
35	SO57738	Road-250
36	SO57738	HL Road Tire

Figure 10: A Drill through the Top Association Rule Showing 36 out of 100 Cases

Associate By ModelName

Itemsets

Support	Size	Itemset
5194	1	Sport-100
3251	1	Water Bottle
3086	1	Mountain-200
2385	1	Patch kit
2340	1	Cycling Cap
2163	1	Mountain Tire Tube
2146	1	Long-Sleeve Logo Jersey
1737	1	Road-250
1654	1	Road Tire Tube
1480	1	Fender Set - Mountain
1474	1	Short-Sleeve Classic Jersey
1443	1	Half-Finger Gloves
1396	1	Mountain Bottle Cage
1369	1	Road-550-W
1291	1	Road-750
1264	1	Road-150
1223	1	Road-650
1185	1	Road Bottle Cage
1172	2	Mountain Bottle Cage, Water Bottle
1153	1	Touring-1000
1057	2	Road Bottle Cage, Water Bottle
1025	1	Touring Tire Tube
1004	2	Water Bottle, Sport-100
999	1	Women's Mountain Shorts
994	2	Long-Sleeve Logo Jersey, Sport-100
976	2	Long-Sleeve Logo Jersey, Cycling Cap
969	1	HL Mountain Tire
955	2	Cycling Cap, Sport-100
936	1	Bike Wash
927	1	Road-350-W
919	2	Mountain Tire Tube, Sport-100
815	1	ML Mountain Tire
803	2	Half-Finger Gloves, Sport-100
778	1	Classic Vest
761	1	Hydration Pack
744	1	LL Road Tire
705	2	Cycling Cap, Water Bottle
676	2	Mountain-200, Sport-100
673	1	Touring-3000
667	3	Long-Sleeve Logo Jersey, Cycling Cap, Sport-100

Figure 11: An Excel Export of the Item Sets Tab

Classification and Prediction

Classification and prediction is the most elaborate part of a data mining course. “Classification is the task of learning a target function that maps each attribute set X to one of the predefined class labels Y” (Tan et al., 2006). In an introductory course in data mining, students learn decision trees, naïve Bayes, neural networks, and logistic regression models.

Classification and prediction is a five-step process: (1) select a classification algorithm, (2) calibrate the parameters of the algorithm, (3) feed a training data set to the algorithm to learn a classification model, (4) load a test data set to the learned model to measure its accuracy, (5) use the learned model to predict previously unknown classifications. Model fitting is an iterative process (steps 2, 3, and 4). Algorithm parameters are calibrated and fine-tuned during the process of finding a satisfactory mining model. Through confusion matrices and accuracy charts, students compare the performance of various mining models to select an appropriate model.

The Theory

Students learn the theoretical foundation and concepts of classification analysis through book chapters, lecture notes, lectures, and homework assignments. In our course, we cover the following areas of classification and prediction:

- Decision tree algorithms: Students learn information gain concepts, best entry selection for a tree-split, entropy, Gini index, and classification error measures.
- Naïve Bayes algorithms: Students learn conditional, prior and posterior probability, independence, and correlation between attributes.
- Neural networks algorithms: Students learn the “simple” concepts of back propagation, nodes, and layers (the mathematical theory of neural networks is beyond the scope of the course).
- Logistic regression analysis: Students learn the difference between a standard linear regression and a logistic regression.

The Practice

For the practice, we used the Mushrooms data set from the UCI repository (University of California Irvine, 2009). Figure 12 is a partial sample of the data set. The class column is the classification column with values edible or poisonous.

First students perform elementary data analysis on the data set. They explore (1) the characteristics of the attributes and their sets of legal values, and (2) elementary classifications, histograms, and groupings using pivot tables and pivot charts. Figure 13 is a pivot chart of the profiles of some of the attributes. It shows the distribution of the classification of a mushroom broken down by attribute type and value. The second row indicates that the data set has 3343 records of:
bruises = none mushrooms with
prob(bruises = none/edible) = 0.35 &
prob(bruises = bruises/poisonous) = 0.84

ID	class	capShape	capSurface	capColor	bruises	odor	gillAttachment	gillSpacing	gillSize	gillColor
1	poisonous	flat	scaly	brown	none	fishy	free	close	narrow	buff
2	poisonous	convex	smooth	buff	bruises	foul	free	close	broad	chocolate
3	edible	convex	scaly	brown	bruises	almond	free	close	broad	white
4	edible	convex	scaly	brown	bruises	none	free	close	broad	white
5	edible	knobbed	smooth	brown	none	none	attached	close	broad	yellow
6	poisonous	knobbed	scaly	red	none	spicey	free	close	narrow	buff
7	edible	bell	smooth	white	none	none	free	crowded	broad	gray
8	poisonous	flat	smooth	white	bruises	pungent	free	close	narrow	black
9	poisonous	flat	fibrous	yellow	none	foul	free	close	broad	chocolate
10	edible	flat	fibrous	brown	bruises	none	free	close	broad	brown
11	edible	flat	fibrous	gray	none	none	free	crowded	broad	brown
12	edible	flat	fibrous	white	none	none	free	crowded	broad	brown
13	edible	flat	smooth	brown	none	none	free	close	broad	white
14	edible	flat	fibrous	brown	bruises	none	free	close	broad	white
15	edible	bell	smooth	brown	none	none	attached	close	broad	orange
16	poisonous	knobbed	scaly	red	none	fishy	free	close	narrow	buff
17	edible	flat	fibrous	brown	none	none	free	crowded	broad	chocolate

Figure 12: Sample Records of the Mushroom Data Set

Attributes	States	Population (A	edible	poisonou
Size		5687	2929	2758
bruises	none	3343	35 %	84 %
bruises	bruises	2344	65 %	16 %
capColor	brown	1594	30 %	26 %
capColor	gray	1261	24 %	20 %
capColor	red	1058	16 %	22 %
capColor	yellow	770	9 %	18 %
capColor	white	729	18 %	8 %
capColor	buff	115	1 %	3 %
capColor	pink	106	1 %	2 %
capColor	cinnamon	28	1 %	0 %
capColor	...	...	...	...
capShape	convex	2566	47 %	43 %
capShape	flat	2205	37 %	40 %
capShape	knobbed	572	6 %	15 %
capShape	bell	320	10 %	1 %
capShape	sunken	21	1 %	0 %
capShape	conical	3	0 %	0 %
capSurface	scaly	2291	35 %	46 %
capSurface	smooth	1781	28 %	35 %
capSurface	fibrous	1614	37 %	19 %
gillAttachment	free	5534	95 %	100 %
gillAttachment	attached	153	5 %	1 %
gillColor	buff	1203	0 %	44 %
gillColor	pink	1047	20 %	16 %
gillColor	white	839	23 %	6 %
gillColor	brown	713	22 %	3 %
gillColor	gray	542	6 %	14 %
gillColor	chocolate	518	5 %	14 %
gillColor	purple	343	11 %	1 %
gillColor	black	284	8 %	2 %
gillColor	...	...	...	...

Figure 13: Sample Attribute Profiles of the Data Set

For the purpose of classification analysis, we built a mining structure and four mining models (a decision tree model, a naïve Bayes model, a neural network model, and a logistic regression model). Then we compared the performance of these models using the holdout set, accuracy charts, and a classification matrix.

The mining structure

Students used wizards to create and configure a mining structure. This involved (1) the inclusion and exclusion of attributes, (2) the configuration of the characteristics of each attribute (key, data type, content type) and (3) the split percentage of data into training and testing (holdout) sets.

Decision tree mining model.

To build a decision tree model, students configured parameters support such as information gain, scoring methods, and type of tree split method. Data and configuration are then sent to the data mining computing engine. A decision tree-mining model with drill through, legends, and display capabilities for each node is generated and sent back to Excel. Figure 14 is an example of a decision tree that has been generated. Students can view the details of the mining legend of each node and drill through it to the underlying data set that supported that node. For example, tracing the "stalkSurfaceBelowRing not = 'scaly'" node, the classification rule is:

If **odor** = 'none' **&** sporePrintColor = 'white' **&**
ringNumber != 2 **&** stalkSurfaceBelowRing != 'scaly'

Then Prob(mushroom is edible) = 85.6 **&**
Prob(mushroom is poisonous) = 14.4

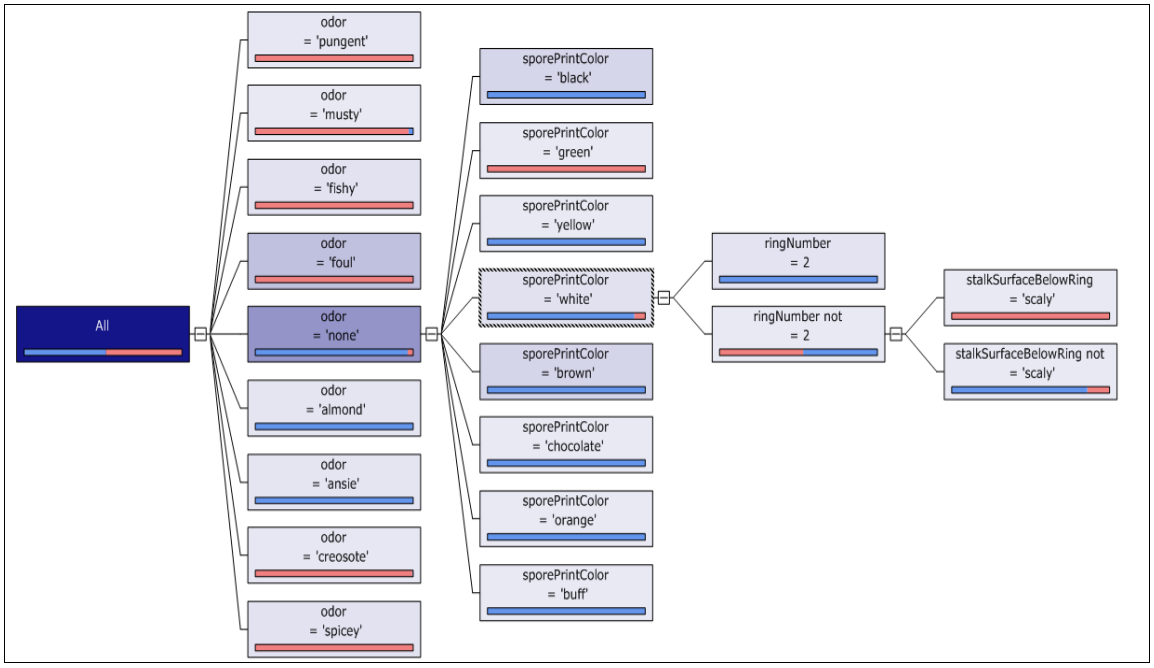


Figure 14: A Decision Tree Classification of the Mushroom Data Set

Naïve Bayes, logistic regression and neural network models

Similarly, students used wizards to configure parameters and to build a naïve Bayes, a logistic regression, and a neural network classification model. The models display discrimination tables that show each attribute value, the classification it favors, and a bar chart as a measure of support. Figure 15 is the naïve Bayes output for the same mining structure as in the decision tree. It dis-

plays the attributes, their values, and the level of contribution of each value to the favored classification. Figure 16 is the logistic regression model output, and Figure 17 is the neural network model output of the same mining structure. It is also possible to build multiple decision tree mining models with different calibrations and attributes and then compare the models for accuracy and validity.

Attributes	Values	Favors edible	Favors poisonous
odor	none		
odor	foul		
stalkSurfaceAboveRing	silky		
gillColor	buff		
gillSize	broad		
gillSize	narrow		
ringType	pendant		
sporePrintColor	chocolate		
ringType	large		
bruises	none		
bruises	bruises		
stalkSurfaceAboveRing	smooth		
population	several		
sporePrintColor	brown		
sporePrintColor	black		
gillSpacing	close		
gillSpacing	crowded		
sporePrintColor	white		
habitat	paths		
odor	spicey		
odor	fishy		

Figure 15: Naïve Bayes: Attribute Discrimination of Class

Attribute	Value	Favors poisonous	Favors edible
sporePrintColor	green		
ringType	flaring		
odor	creosote		
odor	pungent		
stalkColorAboveRing	yellow		
stalkColorBelowRing	yellow		
sporePrintColor	purple		
capShape	sunken		
capShape	conical		
odor	musty		
capColor	cinnamon		
sporePrintColor	buff		
odor	foul		
sporePrintColor	chocolate		
capColor	purple		
stalkRoot	rooted		
odor	none		

Figure 16: Logistic Regression: Attribute Discrimination of class

Attribute	Value	Favors edible	Favors poisonous
sporePrintColor	green		
odor	creosote		
odor	pungent		
gillColor	green		
ringType	flaring		
capColor	green		
odor	foul		
sporePrintColor	purple		
stalkColorAboveRing	yellow		
stalkColorBelowRing	orange		
capShape	sunken		
odor	none		
stalkColorBelowRing	yellow		
capColor	cinnamon		
ringNumber	2		
population	abundant		
population	numerous		
capColor	purple		
ringType	none		

Figure 17: Neural Network: Attribute discrimination of Class

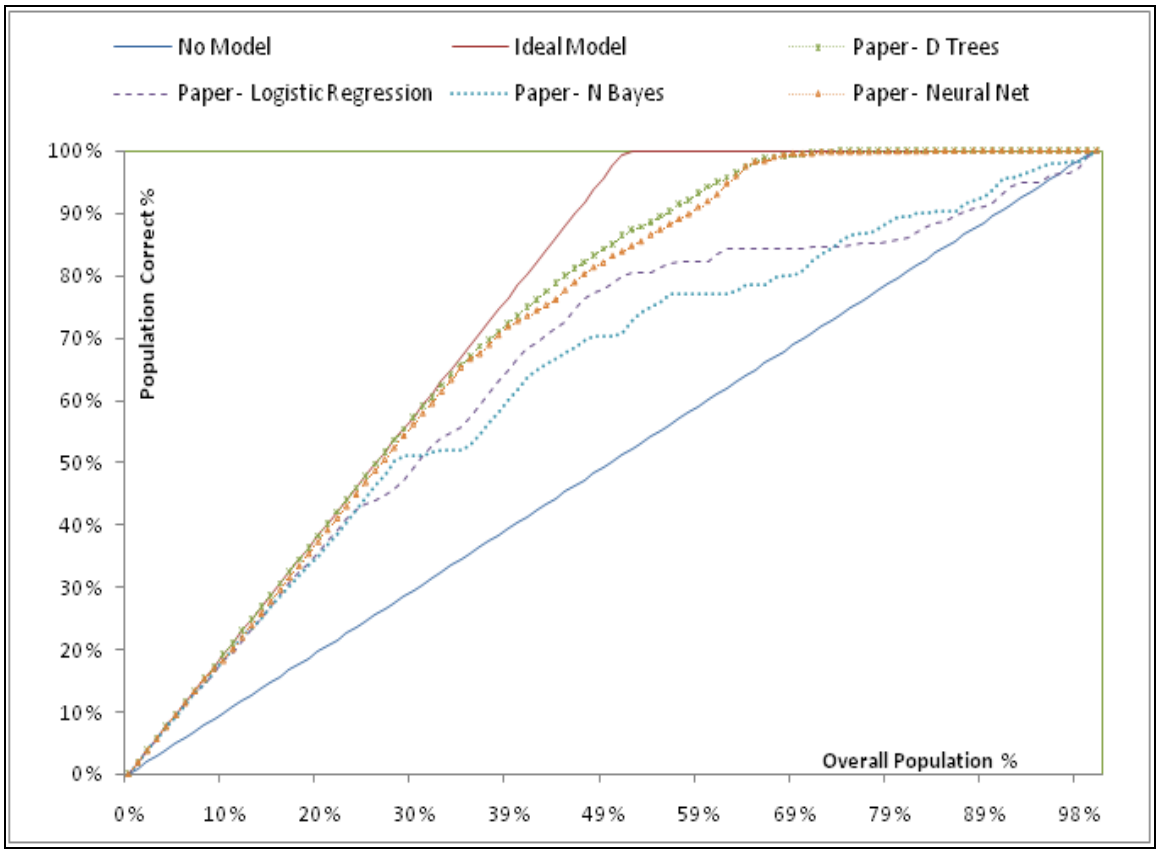


Figure 18: Accuracy (Lift) Chart of 4 the Classification Models for the Poisonous class

Model validation

We demonstrated the simplicity of building a classification mining structure and four mining models for that mining structure. We would like to measure the accuracy of each mining model and compare their performance. Students use wizards to build accuracy charts for each of the four mining models. Figure 18 is the output of such a process. The straight-line from the origin *(0, 0)* to *(100%, 100%)* is the random predictor model. The broken-line from *(0, 0)* to *(49%, 100%)* to *(100%, 100%)* is the ideal model predictor. The ideal predictor will correctly predict every classification. From the graph, the ideal line implies that 49% of the mushrooms in the testing data set are poisonous. The other curves are the decision tree (close to the ideal line), the neural network, the logistic regression, and the naïve Bayes model predictors. Analyzing the chart, the decision tree outperformed the rest of the models. The naïve Bayes model performed worse than the rest of the three models.

Cluster Analysis

Finally, students learn how to perform cluster analysis of data, which is a form of unsupervised learning and grouping of data records.

The Theory

Students learn the theoretical foundation and concepts of cluster analysis through book chapters, lecture notes, lectures, and homework assignments. The topics covered include: (1) distance and weighted distance between objects, (2) similarity measures and weighted similarity measures between data types (nominal, ordinal, interval, and ratio), (3) various k-norms (k= 1, 2 and ∞), (4) simple matching coefficient, cosine, Jacard, and correlation, (5) various center-based clustering algorithms such as the k-means, bisection k-means, and (6) density-based clustering algorithms such as DBSCAN.

The Practice

For the practice, we used the Iris and the Mushrooms data sets. The Iris data set provides an all-numeric ratio scale measure. The Mushrooms data set provides an all-qualitative categorical scale measure. Using wizards, students configured the attributes of interest the maximum number of clusters, the split methods, clustering algorithm to use, and minimum cluster size. Figure 19 is the output of a clustering run. The characteristics of each predicted category is displayed. The relative importance of the range of values of an attribute to the classification of the category is presented as a bar chart. The longer the bar, the stronger the relative importance of the corresponding attribute value for that category.

Students also learn how to perform classification through hierarchical clustering of data. From Figure 20 students could see that category 2 and 3 are well clustered around the *setosa* and the *versicolor* irises. However, category 1 has a mix of *versicolor* (14 records) and *virginica* (50 records) irises. Students then use category 1 records to perform more clustering on the records of this category to extract clear separation criterion between the clusters.

Similarly, we performed (auto detect) hierarchical clustering analysis against the Mushrooms data set. Nine categories or clusters were detected. Mapping the clusters against the poisonous and edible classifications, categories 1, 2, 3, and 5 produced a perfect fit. Figure 21 shows the classification matrix of the first iteration of the cluster analysis.

Category Nam	Row Count
Category 1	64
Category 2	50
Category 3	36

Category Characteristics

Category	Column	Value	Relative I
Category 1	PetalWidth	High:1.58 – 2.09	
Category 1	PetalLength	High:4.80 – 5.48	
Category 1	SepalLength	High:6.40 – 7.04	
Category 1	PetalLength	Very High:>= 5.48	
Category 1	PetalWidth	Very High:>= 2.09	
Category 1	SepalLength	Very High:>= 7.04	
Category 1	SepalLength	Medium:5.90 – 6.40	
Category 2	PetalLength	Very Low:< 1.82	
Category 2	PetalWidth	Very Low:< 0.41	
Category 2	SepalLength	Very Low:< 5.11	
Category 2	SepalWidth	High:3.26 – 3.76	
Category 2	SepalWidth	Very High:>= 3.76	
Category 3	PetalWidth	Low:0.41 – 1.33	
Category 3	PetalLength	Low:1.82 – 4.10	
Category 3	SepalWidth	Very Low:< 2.60	
Category 3	PetalLength	Medium:4.10 – 4.80	
Category 3	SepalLength	Low:5.11 – 5.90	
Category 3	SepalWidth	Low:2.60 – 3.05	
Category 3	PetalWidth	Medium:1.33 – 1.58	

Figure 19: Category Characteristics of Iris Data

Count of id				
Row Labels	Iris-setosa	Iris-versicolor	Iris-virginica	Total
Category 1		14	50	64
Category 2	50			50
Category 3		36		36
Total	**50**	**50**	**50**	**150**

Figure 20: Accuracy Clustering Matrix

Categories	edible	poisonous	Grand Total
Category 1		1728	1728
Category 2	1728		1728
Category 3		1296	1296
Category 4	704	256	960
Category 5	864		864
Category 6	96	480	576
Category 7	320	72	392
Category 8	304	48	352
Category 9	192	36	228
Grand Total	**4208**	**3916**	**8124**

Figure 21: Mapping Detected Categories to Classifications

Figure 22 shows the characteristics of each cluster. Categories 1 and 3 produced poisonous classifications while categories 2 and 5 produced edible classifications. Excel add-ins allow us to re-label category names. For example, we could re-label category 1 as poisonous-1. The records of categories 1, 2, 3, and 5 were filtered out and another classification was performed on the records of the remaining categories. Two iterations later, a perfect match was produced.

Category	Column	Value	Relative Importance
Category 1	gillColor	buff	
Category 1	sporePrintColor	white	
Category 1	stalkRoot	cup	
Category 1	gillSize	narrow	
Category 1	ringType	evanescent	
Category 1	population	several	
Category 1	stalkShape	tapering	
Category 1	bruises	none	
Category 1	odor	spicey	
Category 1	odor	fishy	
Category 2	habitat	woods	
Category 2	bruises	bruises	
Category 2	odor	none	
Category 2	stalkRoot	bulbuous	
Category 2	ringType	pendant	
Category 2	stalkShape	tapering	
Category 2	stalkColorAboveRing	gray	
Category 2	stalkColorBelowRing	gray	
Category 2	stalkSurfaceAboveRing	smooth	
Category 2	gillSize	broad	
Category 2	gillColor	purple	
Category 2	population	solitary	
Category 2	sporePrintColor	black	
Category 3	ringType	large	
Category 3	sporePrintColor	chocolate	
Category 3	odor	foul	
Category 3	stalkSurfaceAboveRing	silky	
Category 3	stalkShape	enlarging	
Category 3	stalkRoot	bulbuous	
Category 3	stalkColorAboveRing	buff	
Category 3	stalkColorBelowRing	buff	
Category 3	stalkColorAboveRing	brown	
Category 3	bruises	none	
Category 5	stalkRoot	equal	
Category 5	gillSpacing	crowded	
Category 5	population	abundant	
Category 5	habitat	grasses	
Category 5	odor	none	
Category 5	ringType	evanescent	
Category 5	stalkSurfaceAboveRing	fibrous	
Category 5	stalkSurfaceBelowRing	fibrous	
Category 5	stalkColorBelowRing	white	
Category 5	stalkColorAboveRing	white	
Category 5	bruises	none	

Figure 22: Characteristics of the three Clusters that Produced a Perfect Match

Summary and Conclusions

Data mining and data analysis for the purpose of decision support is a fast growing area of computing. In the early 2000s, a data mining methods course was taught as a pure research topic in Computer Science. With the maturity of the discipline, the convergence of the data mining algorithms and the availability of computing platforms, students can now learn data mining methods as a problem solving discipline. The theory has matured. Standard textbooks are published. Accompanying technologies that implement the same basic algorithms are available at nominal costs through academic initiatives from companies like Oracle, IBM, and Microsoft. Accordingly, Information Systems programs are capable of providing a computing platform in support of their data mining methods courses. We strongly recommend extending the Information Systems curriculum to include a data mining track of two or more courses.

Walstrom, Schmbach, & Crampton (2008) provided an in-depth survey of 300 students enrolled in an introductory business course justifying their reasons for not choosing Information Systems as an area of specialization. We do see a track in data mining methods enhances the career opportunities of Information Systems students. It is a sustainable growth area that is natural to an Information Systems program. Information Systems students should be able to represent, consolidate, and analyze data using data mining tools to provide organizations with business intelligence for the purpose of decision support.

The content presented and the hands on experience gained have broader applications. In the past, the authors have taught the Introduction to Management Information Systems course. Accounting, finance, general business, and marketing students benefited from the market basket analysis and the decision tree homework assignments. It was easy for the students to analyze large sets of data and produce business intelligence reports.

Finally, if it was not for the underlying technologies that we used, it would have been impossible to cover such material in a one-semester course and provide students with much needed hands-on experience in data mining. What we presented is how to utilize the cloud as a computing platform that transformed the role of a student from "doing low-level IT" in a data mining course to a business intelligence analyst using tools to analyze data for the purpose of decision support (Markoff, 2009).

References

Campos, M. M., Stengard, P. J., & Milenova, B. L. (2005). Data-centric automated data mining. *Proceedings of the Fourth International Conference on Machine Learning and Applications*. IEEE Computer Society, 97-104.

Goharian, N., Grossman, D., & Raju, N. (2004). Extending the undergraduate computer science curriculum to include data mining. *Proceeding of the International Conference on Information Technology: Coding and Computing*, 2, P.251.

Han, J., & Kamber, M. (2006). *Data mining concepts and techniques.* Boston: Elsevier.

Jafar, M. J., Anderson, R. R., & Abdullat, A. (2008a). Data mining methods course for computer information systems students. *Information Systems Education Journal*, 6(48). Retrieved from http://isedj.org/6/48/ISEDJ.6%2848%29.Jafar.pdf

Jafar, M. J., Anderson, R. R., & Abdullat, A. (2008b). Software academic initiatives: A framework for supporting a contemporary information systems academic curriculum. *Information Systems Education Journal*, 6(55). Retrieved from http://isedj.org/6/55/index.html

King, M. A. (2009). A realistic data warehouse project: An integration of Microsoft Access and Microsoft Excel advanced features and skills. *Journal of Information Technology Education, 8*, IIP-091-104. Retrieved from http://www.jite.org/documents/Vol8/JITEv8IIP091-104King696.pdf

Lenox, T. L., & Cuff, C. (2002). Development of a data mining course for undergraduate students. *Proceedings of the Information Systems Education Conference,* 19.

Musicant, D. R. (2006). A data mining course for computer science: Primary sources and implementations. *37th SIGCSE technical symposium on Computer Science Education.* ACM, 538-542.

Markoff, J. (2009, Dec. 14). A deluge of data shapes a new era of computing. *The New York Times.* Retrieved Dec. 14, 2009, from http://www.nytimes.com/2009/12/15/science/15books.html?_r=1&scp=3&sq=cloud%20computing%20microsoft&st=cse

Rahal, I. (2008). Undergraduate research experiences in data mining. *39th SIGCSE Technical Symposium on Computer Science Education,* ACM, 461-465.

Saquer, J. (2007). A data mining course for computer science and non-computer science students. *Journal of Computing Sciences in Colleges*, *22*(4) 109-114.

Tan, P., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining.* Boston: Pearson Education.

Tang, H. (2008). A simple approach of data mining in Excel. *4th International Conference on Wireless Communication, Networking and Mobile Computing*, 1-4.

University of California Irvine. (2009). *UCI machine learning repository*. Retrieved May 10, 2009, from http://archive.ics.uci.edu/ml/

Walstrom, K. A., Schmbach, T. P., & Crampton, W. J. (2008). Why are students not majoring in information systems? *Journal of Information Systems Education, 19*(1), 43-52.

Biography

Musa J. Jafar is Gensler Professor of Computer Information Systems at West Texas A&M University. He holds a Ph.D. in Systems Engineering, an M.S. in Management Information Systems, and an M.S. in Systems Engineering from the University of Arizona at Tucson. He also holds an M.S. in Mathematics from the American University of Beirut. He worked as a senior software engineer and architect for IBM, U.S. West Communications and Bell Communications Research. Additional information can be found at: http://faculty.cis.wtamu.edu/jafar/