A Tools-Based Approach To Teaching Data Mining Methods

Musa Jafar
mjafar@mail.wtamu.edu
CIS Department, West Texas A&M University
Canyon, TX 79018

Russell Anderson
russellkanderson@gmail.com

Abstract

In this paper, we describe how we used Microsoft Excel's data mining add-ins and cloud computing components to teach our senior data mining class. The tools were part of a larger set of tools that we used as part of SQL Server Business Intelligence Development Studio. We also demonstrate the ease of use of these tools to teach a course in data-mining methods with focus on elementary data analysis, data mining algorithms and the usage of the algorithms to analyze data in support of decision-making and business intelligence. The tools allow faculty to focus on the analytical aspects of the algorithms, data mining analysis and practical hands-on homework assignments and projects. The tools allow students to gain conceptual understanding of data mining, hands-on practical experience in data mining algorithms using and analysis of data using data mining tools for the purpose of decision support without having to write large amounts of code to implement the algorithms. We also demonstrate that without such tools, it would have been impossible for a faculty to provide a comprehensive coverage of the topic in a first course in data mining methods. The availability of such tools transform the role of a student from a programmer of data mining algorithms to a business intelligence analyst who understands the algorithms and uses a set of tools that implement these algorithms to analyze data for the purpose of decision support.

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

1. INTRODUCTION

Computer Science and Information Systems programs have been aggressively introducing data mining methods courses into their curriculum, (Lenox & Cuff, 2002; Goharian, Grossman, & Raju, 2004; Saquer, 2007; Jafar, Anderson, & Abdullat, 2008), outlined course content in data mining that are consistent with (Lenox & Cuff, 2002). Computer Science programs have been focusing on the "deep understanding of data mining, instead of simply using tools" (Goharian, Grossman, & Raju, 2004; Musicant, 2006; Rahal, 2008). They focus on the algorithmic aspect of data mining and the efficient implementation of the algorithms. They require advanced programming and data structures knowledge as prerequisites (Musicant, 2006; Rahal, 2008). However, for Bachelor of Business Administration (BBA) in Information Systems, we see the data analysis and the business intelligence aspects of data mining as the focus. Students learn the theoretical concepts and use data sets and tools that implement data mining algorithms to analyze data. We require a first programming course, a database management course, and a statistical data analysis course as prerequisites. Accordingly, the deep understanding of the algorithms, their implementation and the efficiency of implementation is more appropriate for a computer science program. For BBA students, a data centric, algorithm understanding and process-automation approach to data mining similar to Campos, Stengard, & Milenova (2005) is more appropriate.

For the tools part, we chose Microsoft Excel with its data mining add-in(s) as the front-end and cloud computing and SQL Server 2008 as the back-end. Microsoft Excel is pervasive, it is available on almost every college desktop, with its data presentation capabilities, charting, functions and macro support; it is a natural front-end environment for data analysis. It provided us with a self-contained computing environment with back-end support (add-ins, cloud computing and SQL Server Connectivity). Others (Tang, 2008) have chosen XMLMiner with spreadsheet support for pre-and post data mining analysis. Oracle Corporation and IBM provide similar capabilities also. However, Microsoft Excel with its data mining add-ins and support is more pervasive on academic and individual desktops (No additional installation is required).

In the early 2000, data mining tools and technologies were hard to learn, acquire and teach. In the past three years however, these technologies became available for universities through the combination of commercial and open source academic initiatives at a minimal cost (Jafar, Anderson, & Abdullat, 2008) outlines various commercial initiatives and their coverage. In summary, data mining theory has streamlined, its computing technology matured and the tools are available for free or at a minimal cost for academic programs. Information Systems programs could use an approach similar to the database design course approach for teaching data mining courses. In a database design course, we teach students the theory of database design which may include relational algebra, relational calculus, and includes entity-relationship modeling, normalization, transaction management, relational model and indexing. We use tools such as Visio enterprise, IBM Rational, ER-Win or MySQL Workbench for modeling, and a relational engine such as Oracle, SQL Server, IBM-DB2 or MySQL Database Management System for the hands-on component.

2. BACKGROUND

Data mining for the purpose of decision support is **not** the process of defining, designing and developing efficient algorithms and their implementations. It is the process of (1) consolidating large data sets into a minable data set, (2) using the mineable data set to train model building algorithms to generate analysis and prediction mining models, (3) validating the capabilities of the mining models, and then (4) using the mining models for the purpose of decision support. In summary, 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).

Usually a data set is divided into a training data set and a testing data (holdout) set. The training data set is used to build the mining structure and associated mining models. The testing data set is used to test the accuracy of the mining models. If a model is valid and its accuracy is acceptable, it is then used for prediction. Figure 1 of the appendix is a visualization of the data mining process. It is worth mentioning that in the past, our students had to write their own macro(s) to randomly split the data into training and test data sets. However with the Excel data mining add-ins, the data mining wizards allow students to perform this task automatically and through configuration. That allows us both faculty and students to focus on the data analysis task instead of writing random number generator(s) macros to perform the splitting task (it saved us a homework assignment).

In this paper we will take a hands-on approach to data mining with examples. Each section will be composed of the theory component of data mining followed by the practice component. For the practice components, we will 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), we will use these two data sets for elementary data analysis, classification and clustering analysis. 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 (association analysis). The Iris data set attributes are quantitative; the mushroom data set attributes are qualitative.

The Iris data set is composed of 150 records of:

Iris(sepal-length, sepal-width, petal-length, petal-width, iris-type)

for a total of 4 attributes for each Iris type. The length and width attributes are in centimeters. The classification (Iris-type) are Setosa, Versicolor, or Virginica.

The Mushrooms data set is composed of 8,124 records of:

Mushroom(capShape, capSurface,…, odor, ringType, habitat, gillSize, ….., classification)

for a total of 21 attributes in each record. All the attributes are qualitative and the classification of a mushroom is either Poisonous or Edible.

The Bikebuyers data set is composed of 121,300 records of:

BikeBuyer(SalesOrder,Quantity, ProductName, Model, Subcategory,Category).

For the rest of this paper we cover the basic topics covered in a standard data-mining course. The standard topics are elementary data analysis and outlier detection, association rules (market basket analysis), classification algorithms and cluster analysis. The topics, subtopics and the terminology used are found in standard data mining textbooks. The textbooks of (Han & Kamber, 2006; Tan, Steinbach, & Kumar, 2006) are standard for a data mining methods course. In the next section, we explore elementary data analysis, followed by association analysis then classification and then cluster analysis. The last section is a summary and conclusions section. All the figures and tables are in the appendix.

3. ELEMENTARY DATA ANALYSIS

The Theory

The data mining process has been defined as an Extract, Transform and Load (ETL) process. Elementary data analysis is a step that precedes (ETL). It is the first basic step in data mining. It allows data miners to understand the intricacies of the data set. The data miner needs to have domain knowledge of the dataset, knowledge of the characteristics of each attribute and where possible dependencies between attributes. Data miners need to be able to perform elementary data analysis on the data set under consideration. They should be able 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).
(3)Visualize the data through histograms, scatter plots, quartile plots and box plots,
(4)Produce hierarchical data analyses through pivot tables and pivot charts.
(5)Produce and analyze the various correlation matrices and key influencers of the attributes.

Finally, in preparation for data mining, the data may need to be relabeled, grouped or normalized.

The Practice

The hands-on practice of elementary data analysis is performed in Excel. Excel is a natural fit for elementary data analysis, with its charting, sorting, table and pivot table capability most of the elementary data analysis tasks can be easily performed from within Excel (Tang, 2008). Excel and Excel tools are heavily used in analyzing data for the purpose of decision support. The data analysis add-in tools allow students to generate descriptive statistics, produce correlation matrices, histograms, percentiles and scatter plots. Using wizards, within minutes, a student can produce summary statistics similar to those in Figure 2, Figure 3 and Figure 4 of the appendix.

The table, pivot table and charting tools allow students to perform various hierarchical analyses and relabeling of the data. Figure 3 is a sample of pivot tables and pivot table charts that can be produced easily from within Excel. From the pivot charts accompanying the pivot tables, it is easy to see that all petal lengths and sepal lengths of the Setosa(s) are small; the Virginica(s) dominate the high end of the petal length and sepal lengths. Filters can also be added on top of the row and column contents to produce hierarchical representation of the data for the purpose of elementary data analysis.

Using the data mining add-ins, we can analyze the overall key influencers of the Iris classification with the relative impact of each attribute value set (Figure 4). We can also perform pair-wise comparisons between the different classifications. The key Influencers tool automatically break the range of a continuous attribute into intervals while determining the key influencers of the Iris type. Based on an analytical model, the algorithm decided that PetalWidth < 0.4125 strongly favors the Setosa classification, a petalwidth in the range of [0.4125, 1.33] strongly favors a Versicolor classification and a petallength >=5.48 strongly favors a Virginical classification. A student can use this visual analysis and presentation of the key influencers to build expert system rules for a classification decision support system. The key influencers tool allows students to perform pair wise discrimination for key influencers of the different types of the classifications. The length of the bar charts to the right indicates the relative importance of each attribute range.

The data exploration tools allow users to interactively produce histograms and configure bucket counts. The data clean-up tools allow users to interactively produce line charts and specify ranges for outliers of numeric data. The data-sampling tools allow users to interactively divide the dataset into multiple random samples. The re-labeling tool allows users to interactively re-label data into ranges such as low, medium and high.

4. ASSOCIATION RULES (MARKET BASKET ANALYSIS)

Market basket analysis allows a retailer to understand the purchasing behavior of customers and predict 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 also be used to analyze 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

Through book chapters, lecture notes and lectures, students learn theoretical foundations and concepts of association analysis. Students learn conditional probability concepts and Baysian statistics. They learn 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 lift of association rules and their calculations, *a priori* and general algorithms for generating frequent item sets from a market basket set, *a priori* and general algorithms for generating association rules from frequent item sets. Faculty may also design exam questions and homework problem solving assignments to emphasize these concepts. For example, given a small market basket and a set of thresholds, students should be able to manually use association analysis algorithms (*a priori* and confidence-based pruning) to generate the pruned item set, detect closed item sets, generate rules and calculate the support, confidence and importance of rules as shown in the activity diagram in Figure 5. The two books (Han & Kamber, 2006; Tan, Steinbach, & Kumar, 2006) that we have used for the past three years have a tendency to write algorithms in complex English-like structures with a lot of mathematical notations. It is helpful when faculty visualize an algorithm by flowcharting it as an activity diagram then use an example to demonstrate the algorithm in action. Figure 5 is an activity diagram of the *a priori* algorithm for discovering frequent itemsets (size two or more) and Figure 6 is an example implementation of the algorithm as it applies to an item-set of purchases. In Figure 6 we start with 6 transactions, then we use the *a priori* algorithm to generate all the frequent item-sets with minimum threshold support of 2. We stop when no item-sets with the minimum threshold support can be generated.

The Practice

The Bikebuyers data set has 31,450 sales orders with 121,300 recorded items for 266 different products that spans across 35 unique categories and 107 different models. Each record has a sales Order that describes the details of the items sold (Product Name, Quantity, Model Name, Subcategory Name and Category Name). Figure 7 is a sample of the data set, According to this sample, Sales Order 43659 has 12 different products as follows: One Mountain-100 Black, 42, 3 Mountain-100 Black, 44, …, and 4 Sport-100 Helmet Blue.

After learning the theory, students use this large data set to perform market basket analysis and focus their time and effort on the analysis task and the decision support aspects of it. For the task at hand students use wizards to build an association analysis mining structure based on the Model Names of the items sold. With the data mining add-ins to Excel, and using wizards 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 association rules. Figure 8 is the output of the 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 (Figure 11) and (3) the dependency network between the items. The figure shows the output after executing the calibrated association analysis algorithm. The algorithm concluded that those who bought all-purpose bike stand and an HL Road Tire also bought a tire Tube with a probability of 0.941 and an importance of 1.080. i.e. the association rule is: **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 $\log \frac{prob(B/A)}{Prob(B/\sim A)}$.

Figure 9 is an Excel export of Figure 8 (the item sets tab) for further analysis. The user interface allows students to select a rule and drill through it to the record cases associated with that rule, Figure 10 is a drill through of the top rule of Figure 8. Figure 11 is an Excel export of the item-sets tab of Figure 8. It can be seen that the bar charts from the data-mining add-ins are exported as conditional formatting data bars in Excel. Students can also explore the item sets generated and the strength of dependencies between items. Students then store their results into work sheets for further data analysis.

5. CLASSIFICATION ANALYSIS AND PREDICTION

This is the most elaborate part of a course in data mining. Generally speaking, "classification is the task of assigning objects to one of several predefined categories". Formally, "classification is the task of learning a target function that maps each attribute set X to one of the predefined class labels Y" (Tan, Steinbach, & Kumar, 2006). In an introductory course in data mining, students usually learn decision trees, Naïve Bayes, Neural Networks and Logistic regressions models.

The Classification process is a four step process: (1) Select the classification algorithm and specify its parameters. (2) Feed a training data set to the algorithm to learn a classification model. (3) Feed a test data set to the learned model to measure its accuracy. (4) Use the learned model to predict previously unknown classes. In most cases, the mining model fitting is an iterative process; algorithm parameters are calibrated and fine-tuned during the process of finding a satisfactory mining model. Through confusion matrices and lift charts, students usually compare the performance of various mining algorithms and models to select an appropriate algorithm and an associated model.

The Theory

Through book chapters, lecture notes and lectures, students learn the theoretical foundation and concepts of classification analysis. Usually, classification analysis is divided into four areas:

- Decision tree algorithms where students learn information gain concepts, best entry selection for a tree-split, entropy, Gini index and classification error measures.
- Naïve Bayes algorithms where students learn conditional, prior and posterior probability, independence and correlation between attributes.

- Neural Networks algorithms where students learn the "simple" concepts of back propagation, nodes and layers (the details of how neural networks work and the theory behind it is beyond the scope of such a course).
- Logistic regression where students learn the difference between standard linear regression and logistic regression where the classification is qualitative usually (Binomial distribution) and the algorithm to predict the probability of one of these values (not to predict a value on the continuum of the corresponding numerical outcomes).

The Practice

For the practice, we use the Mushrooms data set. It is a public domain data set from the UCI database. The data set is used to predict whether a mushroom is poisonous or edible. Figure 12 is partial sample of the data set.

First students perform elementary data analysis on the data set to (1) find out the characteristics of the attributes and their data ranges, (2) elementary classifications, histograms and groupings using pivot tables and pivot charts. Figure 13 is a pivot chart that details the distribution of the classifications of a mushroom broken down by attribute. With pivot tables and charts, students can build numerous hierarchical histograms to understand the characteristics of the data.

For the purpose of classification analysis we build a mining structure and four mining models (a decision tree model, a Bayes model, a neural network model and a logistic regression model). Then we compare the performance of these models using lift charts and a classification matrix.

The Mining Structure: Students use the wizards of the data mining tools to create and configure a mining structure. This involves the inclusion and exclusion of attributes, the configuration of the characteristics of each attribute (key, data type and content type, split percentage of data into training and testing).

Associated Decision Tree Mining Model: Students configure the parameters support, information gain scoring methods, the type of tree split, etc. then a decision tree-mining model with drill through, legends and display capabilities for each node is generated, Figure 14 is an example of a decision tree, students can drill through each node to the underlying data set that supports that node. For example, the model classified the deep bottom branch of the tree as follows:

If odor = 'none' **&**
  sporePrintColor = 'white' **&**
  ringNumber not = 2 **&**
  stalkSurfaceBelowRing not = 'scaly'
Then Prob(mushroom is edible) = 85.6 **&**
  Prob(mushroom is poisonous) = 14.4

Associated Naïve Bayes, Logistic Regression and Neural Network Models: Similarly, students configure parameters and use wizards to build a naïve Bayes, logistic regression, and a neural network classification models. 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 displays the attributes, their values and the level of contribution of that value to the favored classification. Similarly Figure 16 is the logistic regression model output and Figure 17 is the neural network model output.

Associated Model Validation: In this section, we demonstrate the simplicity of performing model validation using the four classification algorithms. Students use wizards to build the accuracy charts (Figure 18) for each of the four models. The straight-line from the origin (0, 0) to (100%, 100%) is the random predictor. The broken line from (0, 0) to (49%, 100%) to (100%, 100%) is the ideal model predictor, it 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 outperforms the rest of the models; the Naïve Bayes model performs worse than the other three models. Students then use their analytical skills to compare the models relative to the ideal model and random model.

6. CLUSTER ANALYSIS AND CATEGORY DETECTION

The theory

Finally, students learn how to perform cluster analysis of data which is a form of unsupervised classification. Through chapters, lectures, class presentations and "paper and pencil" homework assignments, students learn the concept of distance and weighted distance between object. Students learn the concept of similarity measures and weighted similarity measures between data types (nominal, ordinal, interval and ratio) and data entities. Measures such as the various k^{th} norms (k= 1, 2 and ∞), simple matching coefficient, cosine, Jacard and correlation are learned. Students learn various center-based clustering algorithms such as the K-means, Bisection K-means. Students also learn density-based clustering algorithms such as DBSCAN.

The Practice

For the practice we use the Iris and the Mushrooms data set. The Iris set provides an all numeric ratio scales measure. The Mushrooms data set provides an all qualitative categorical scales measure. Using wizards, students configure the attributes of interest the maximum number of clusters, the split methods, clustering algorithm to use, minimum cluster size, etc. Figure 19 is the output of a clustering run, the characteristics of each of category are displayed.

Students also learn how to perform classification through hierarchical clustering of data. For example, from Figure 20, students can see that category two and three are well clustered around the Setosa and the Versicolor classifications. However, category one has a mix of Versicolor (14 records) and Viriginica (50 records). Students then learn to filter Category one out and perform more clustering on the records of this category to extract clear separation criteria between the clusters. Since clustering is a non-supervised learning process, category detection is not applicable. Since we know that the iris dataset has three distinct categories, we use pivot tables to demonstrate the accuracy of the algorithm. Category one can be clustered again to refine it.

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

The following histogram shows the characteristics of each cluster, the longer the bar the stronger the influence of the corresponding attribute value. Keep in mind that categories 1 and 3 produce poisonous classifications and categories 2 and 5 produce edible classifications. The records of these categories (1, 2, 3 and 5) are filtered out and another classification is performed. Two iterations later a perfect match is produced. Accordingly, students are able to learn to perform hierarchical iterative clustering on the data.

7. SUMMARY AND CONCLUSION

Data mining and data analysis for the purpose of decision support is a fast growing area of computing. In the early 2000(s), a data mining methods course was taught as a pure research topics in computer science. However, with the maturity of the discipline, the convergence of algorithms and the availability of computing platforms, students can learn data mining methods as a problem solving discipline that strengthens their analytical skills. The theory has matured, standard textbooks are published and the accompanying technology implements the same basic algorithms. Given the academic initiatives of companies like Oracle, IBM and Microsoft, Information Systems programs are capable of providing a computing platform in support of data mining methods courses. We strongly recommend extending the Information Systems curriculum to include a data mining track for up to three 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 that a track in data mining methods, potentially enhances the career opportunities of Information Systems students. Iit is a sustainable growth area that is natural to a BBA in Information Systems program. BBA in 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.

Finally, what we presented is not a course about Excel add-ins. It is as much of a course about Excel as a database course is about Oracle, SQL Server, DB2 or MySQL database management systems, or a business statistics course is about SAS, SPSS or R. 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 the much needed hands-on experience in data mining. It is neither the intension of this paper nor that of the course is to teach hands-on excel. We teach the theory of data mining and the underlying algorithms.

8. REFERENCES

Campos, M. M., Stengard, P. J., & Milenova, B. L. (2005). Data-Centric Automated Data Mining. International C*onference on Machine Learning and Application.* IEEE Computer Society.

Goharian, N., Grossman, D., & Raju, N. (2004). Extending the Undergraduate Computer Science Curriculum to Include Data Mining. *International Conference on Information Technology: Coding and Computing*, *2.*

Han, J., & Kamber, M. (2006). *Data Mining* Concepts *and Techniques.* Elsevier Inc.

Jafar, M. J., Anderson, R. R., & Abdullat, A. A. (2008). Data Mining Methods Course for Computer Information Systems Students. *Information Systems Education Journal , 6(48)*.

Jafar, M. J., Anderson, R. R., & Abdullat, A. A. (2008). Software Academic Initiatives: A Framework for supporting a Contemporary Information Systems Academic Curriculum. *Information Systems Education Journal , 6(55)*.

Lenox, T. L., & Cuff, C. (2002). Data Mining Methods Course for Computer Information Systems Students. *Information Systems Education Conference.*

Letsche. (2007). Service Learning Outcomes in an Undergraduate Data Mining Course. *Midwest Instruction and Computing* Symbosium*.*

Musicant, D. R. (2006). A data Mining course for computer science: primary sources and implementations. *37th SIGCSE technical symposium on Computer science education.* ACM.

Rahal, I. (2008). Undergraduate research experiences in data mining. *39th SIGCSE technical symposium on Computer science.* ACM.

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

Tan, Steinbach, & Kumar. (2006). *Introduction to Data Mining.* Peasrson Education Inc.

Tang, H. (2008). A Simple Approach of Data Mining in Excel. *4th International Conference on* Wireless *Communication, Networking and Mobile Computing*, (pp. 1-4). Dalian.

University of California Irvine. (2009). *UCI Machine* Learning *Repository*. Retrieved 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*.*

9. APPENDIX

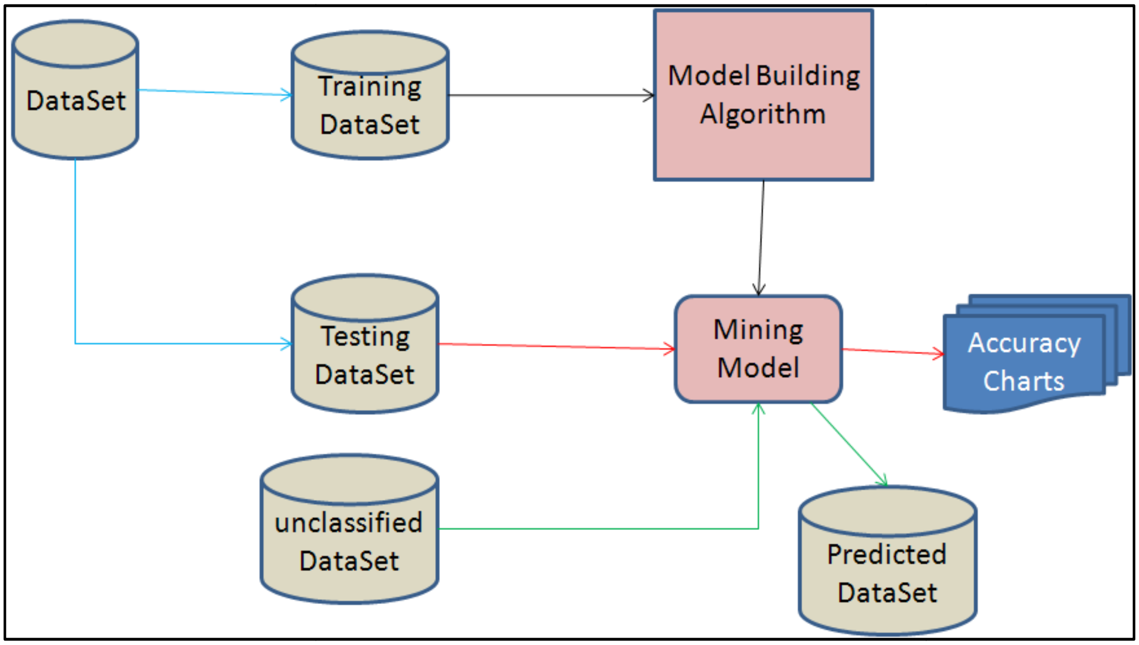
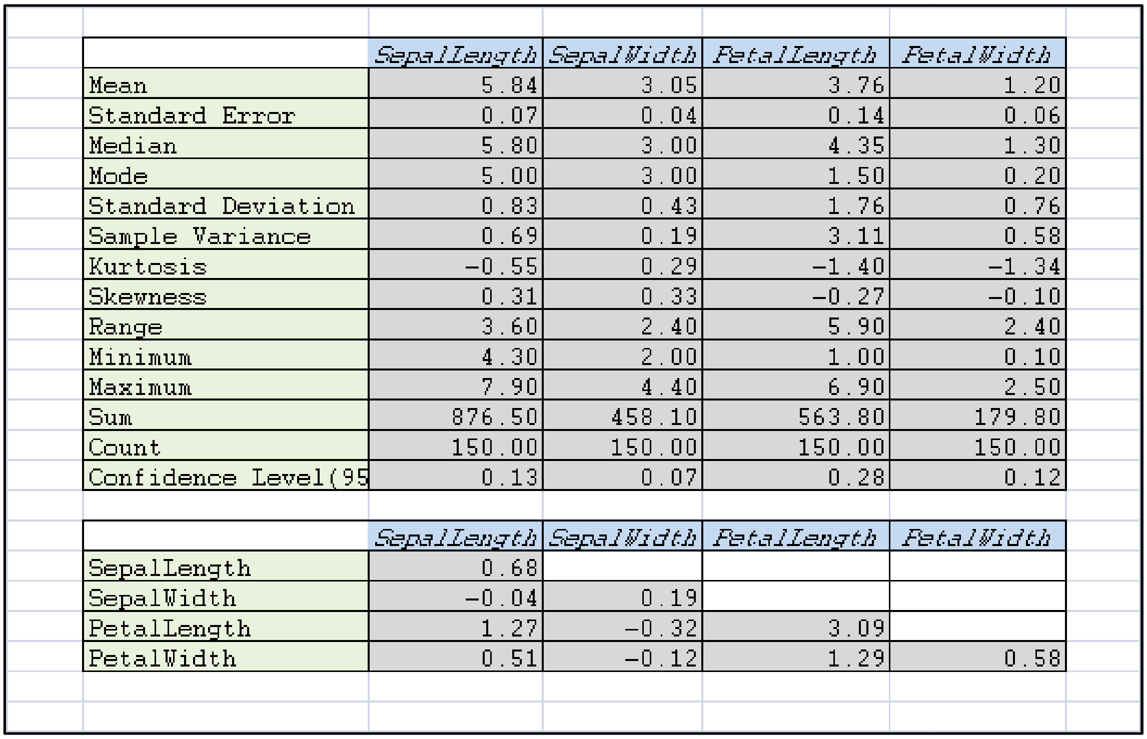


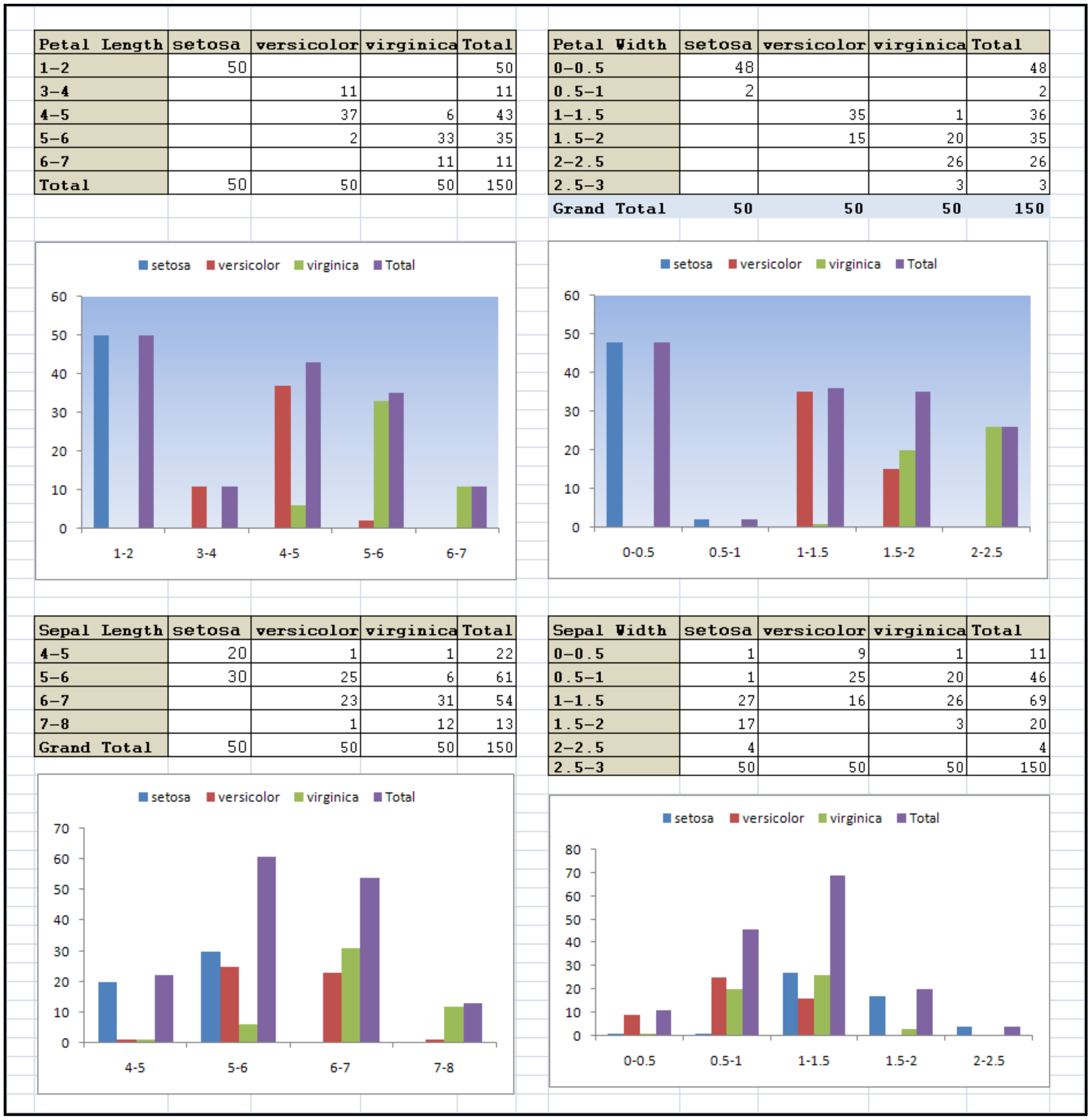
Figure 1 High level flow of data mining activities



	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



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

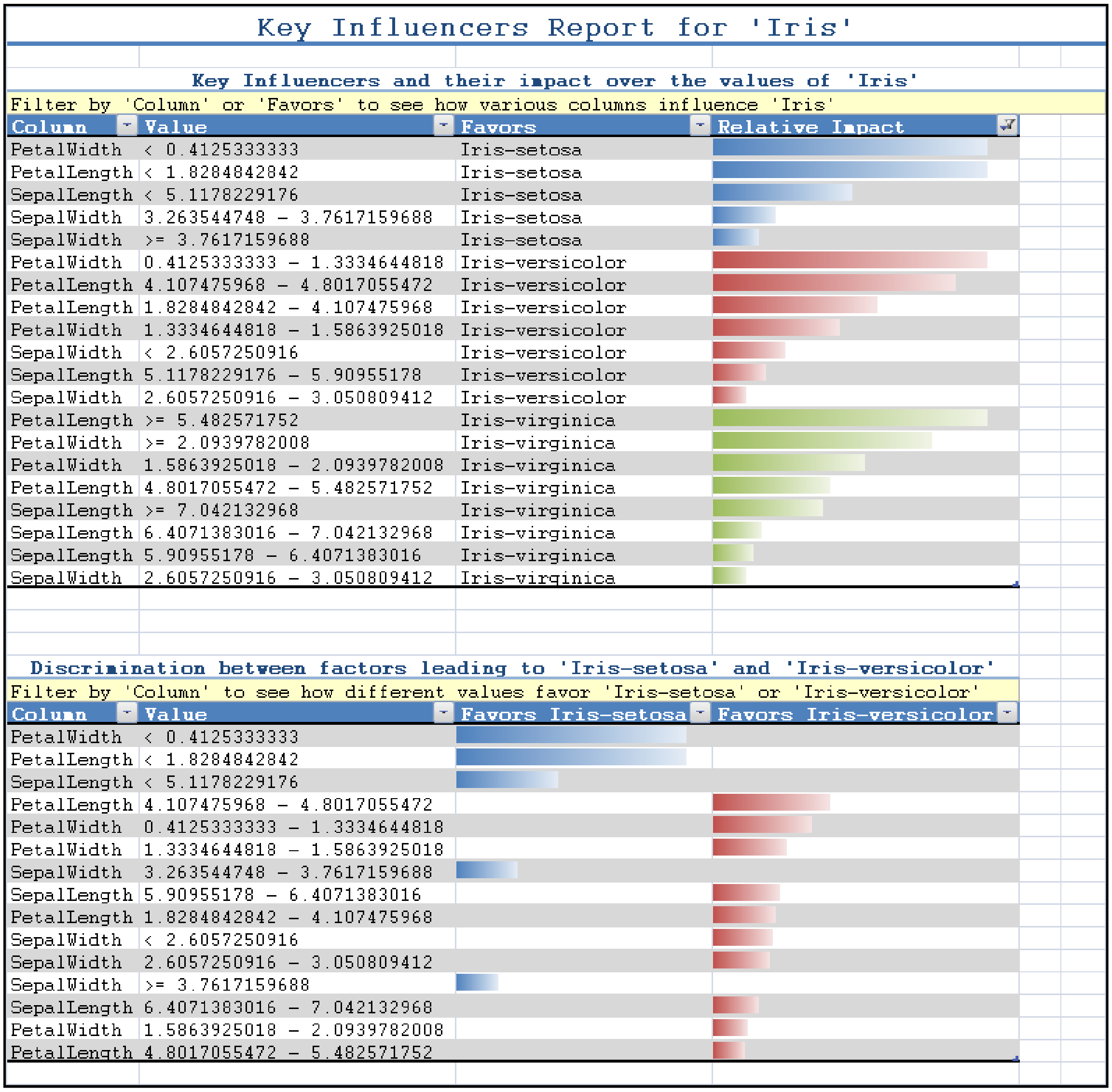
Petal Width	setosa	versicolor	virginica	Total
0-0.5	48			48
0.5-1	2			2
1-1.5		35	1	36
1.5-2		15	20	35
2-2.5			26	26
2.5-3			3	3
Grand Total	50	50	50	150

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

■ setosa ■ versicolor ■ virginica ■ Total
0
10
20
30
40
50
60
1-2
3-4
4-5
5-6
6-7
■ setosa ■ versicolor ■ virginica ■ Total
0
10
20
30
40
50
60
0-0.5
0.5-1
1-1.5
1.5-2
2-2.5
■ setosa ■ versicolor ■ virginica ■ Total
0
10
20
30
40
50
60
70
4-5
5-6
6-7
7-8
■ setosa ■ versicolor ■ virginica ■ Total
0
10
20
30
40
50
60
70
80
0-0.5
0.5-1
1-1.5
1.5-2
2-2.5

Figure 3 A Screenshot of Pivot Tables and Charts Analysis To



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

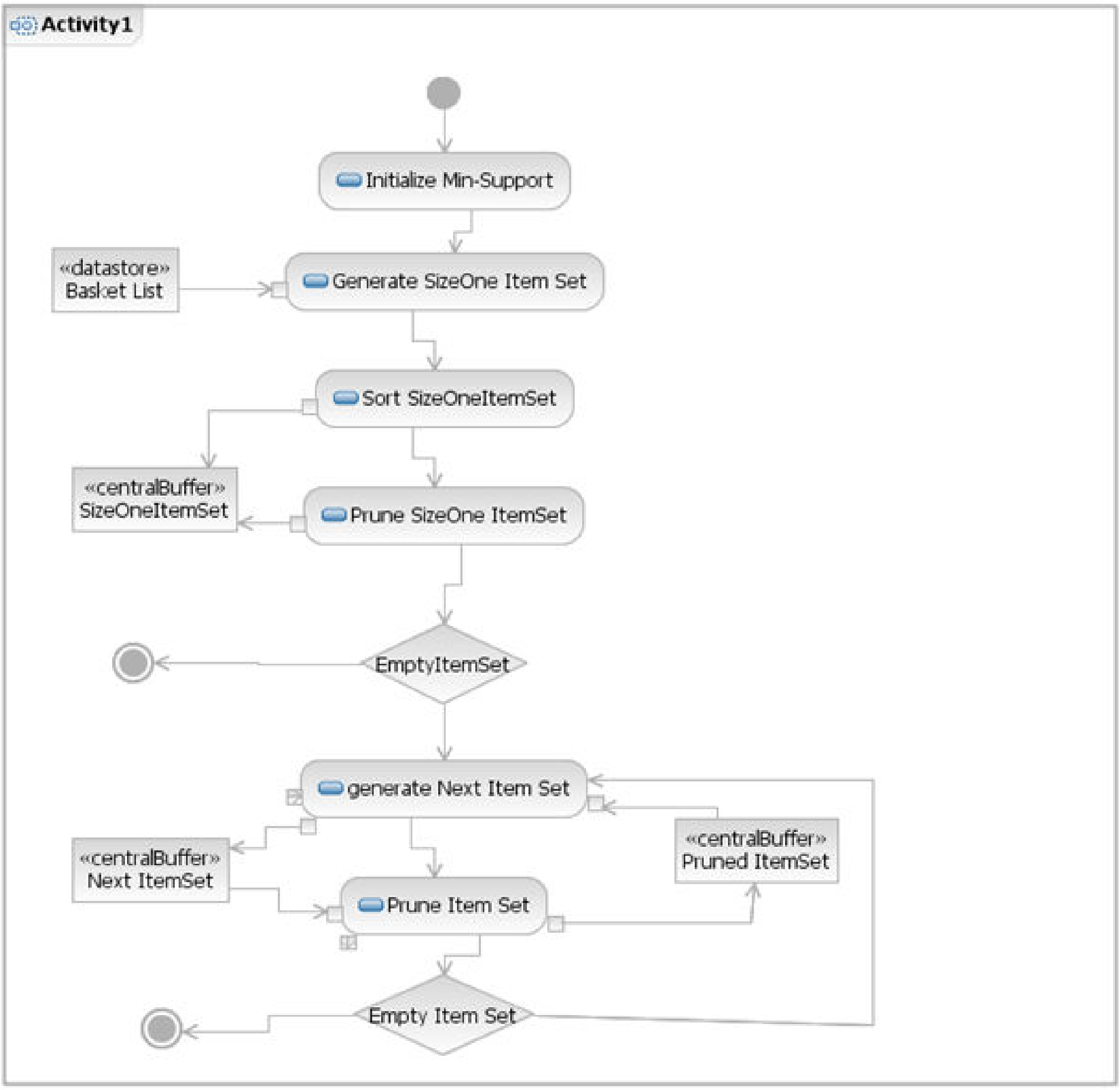
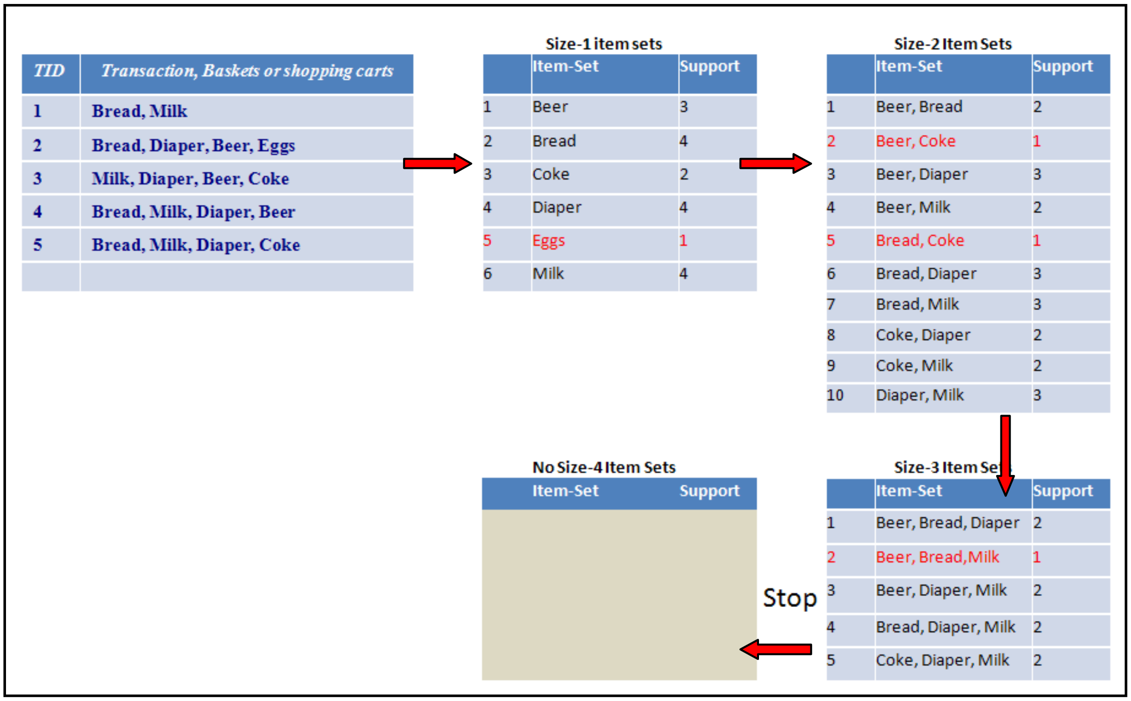


Figure 5 Activity diagram of the Apriori Item set generation Algorithm



TID	Transaction, Baskets or shopping carts
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Size-1 item sets

	Item-Set	Support
1	Beer	3
2	Bread	4
3	Coke	2
4	Diaper	4
5	Eggs	1
6	Milk	4

Size-2 Item Sets

	Item-Set	Support
1	Beer, Bread	2
2	Beer, Coke	1
3	Beer, Diaper	3
4	Beer, Milk	2
5	Bread, Coke	1
6	Bread, Diaper	3
7	Bread, Milk	3
8	Coke, Diaper	2
9	Coke, Milk	2
10	Diaper, Milk	3

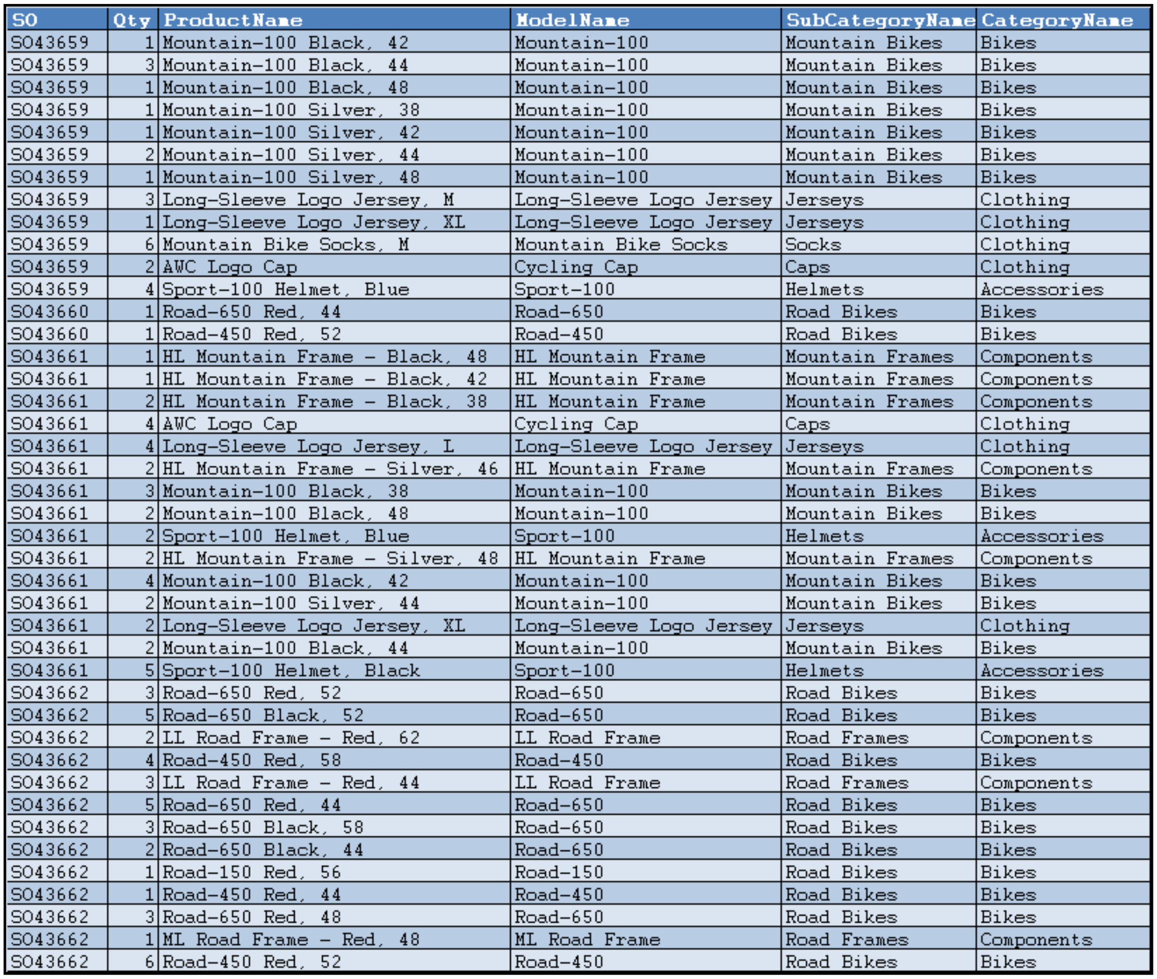
Size-3 Item Sets

	Item-Set	Support
1	Beer, Bread, Diaper	2
2	Beer, Bread,Milk	1
3	Beer, Diaper, Milk	2
4	Bread, Diaper, Milk	2
5	Coke, Diaper, Milk	2

No Size-4 Item Sets

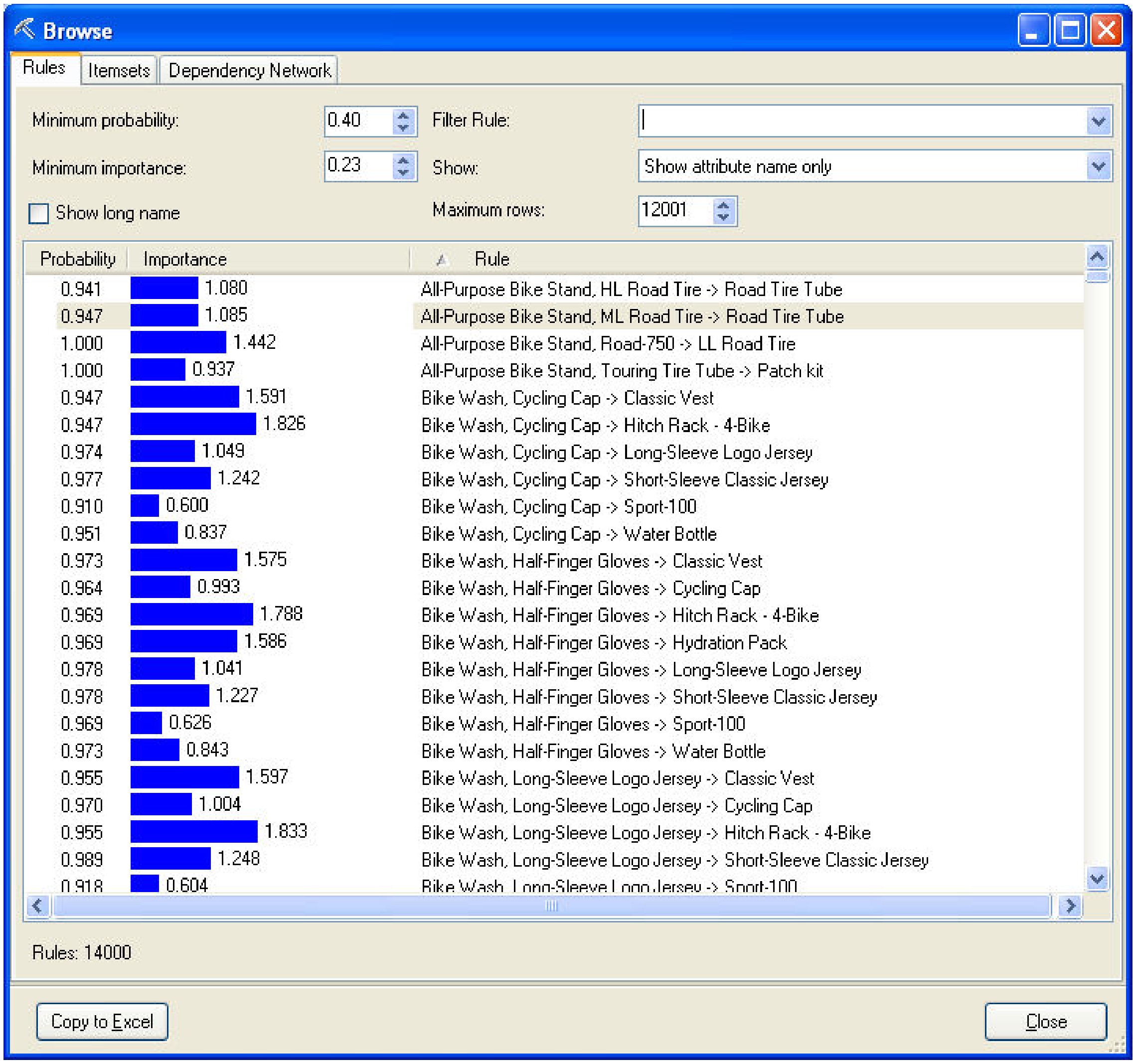
Item-Set	Support

Figure 6 Apriori Algorithm implementation example



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



Minimum probability: 0.40

Filter Rule:

Minimum importance: 0.23

Show: Show attribute name only

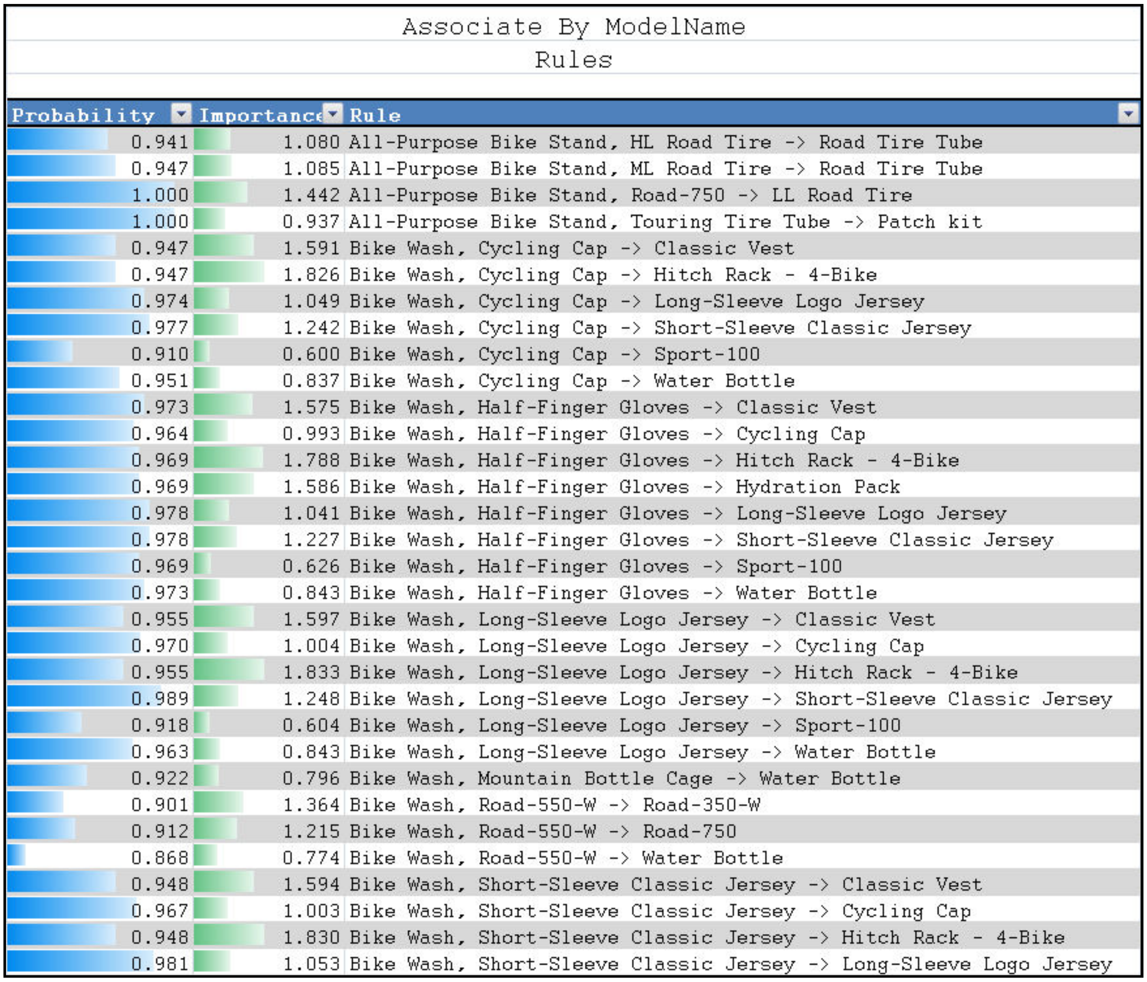
☐ Show long name

Maximum rows: 12001

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

Rules: 14000

Figure 8 A Screenshot of the Analysis

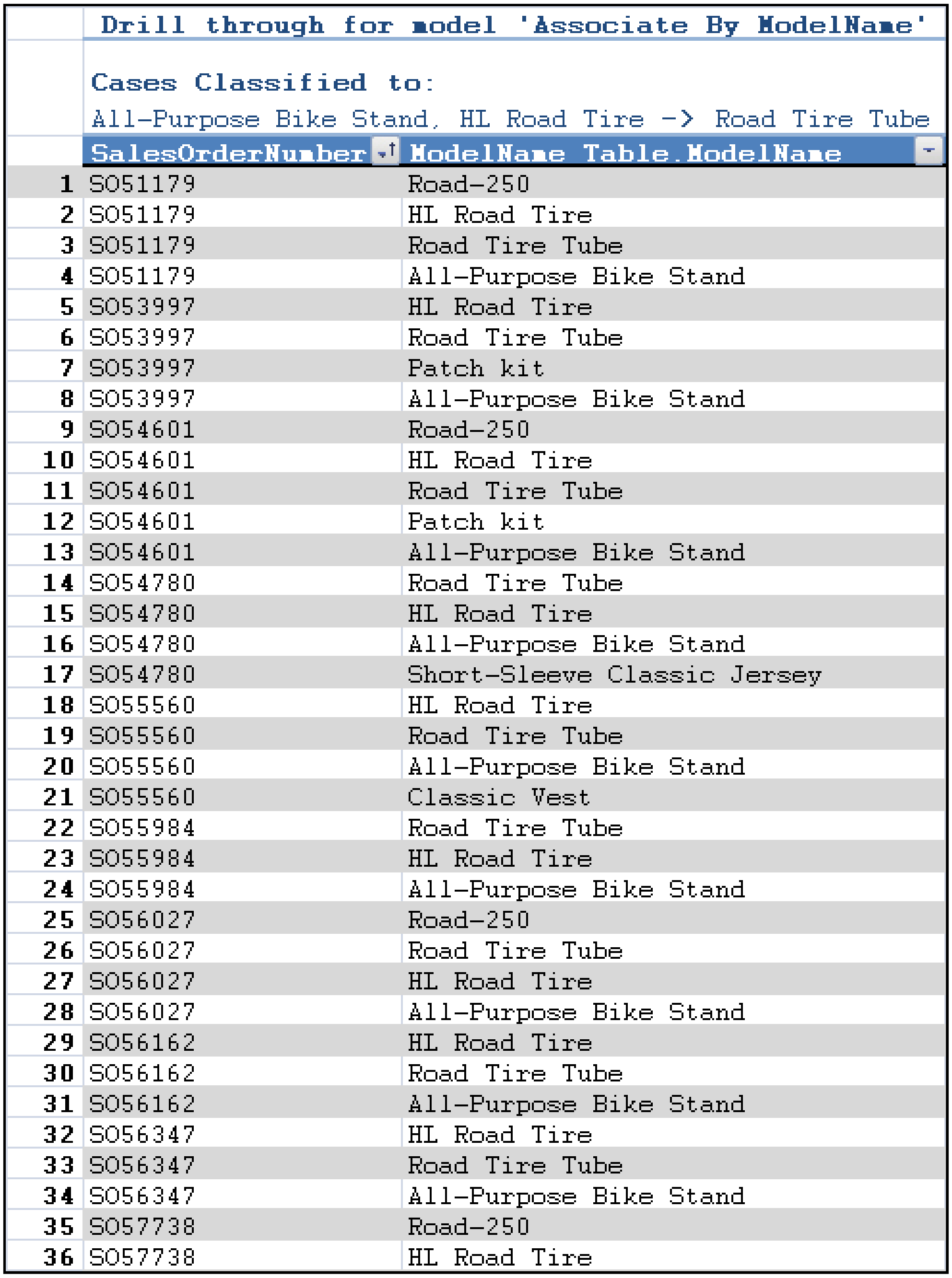


Associate By ModelName

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 Association Rules



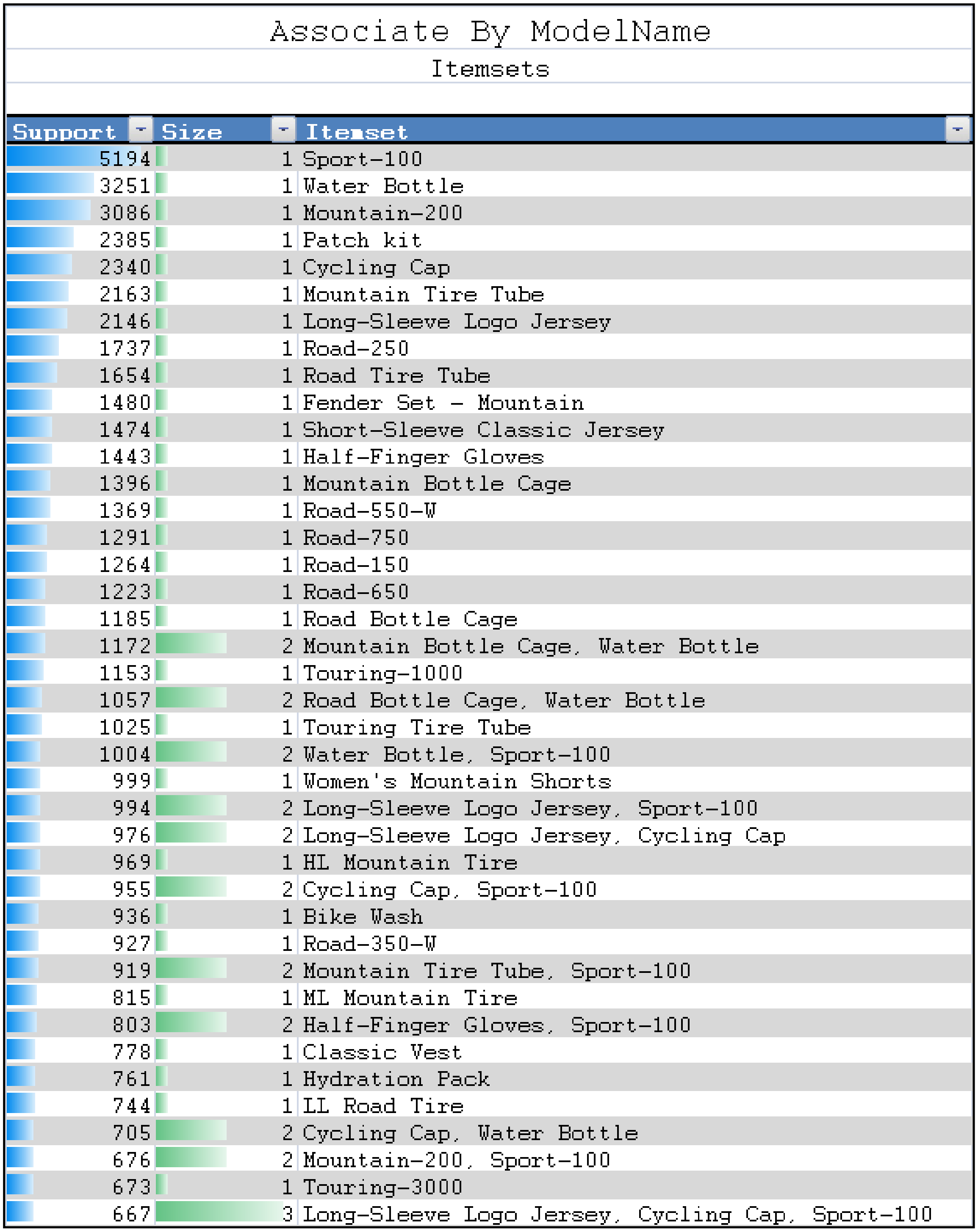
Drill through for model 'Associate By ModelName'

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

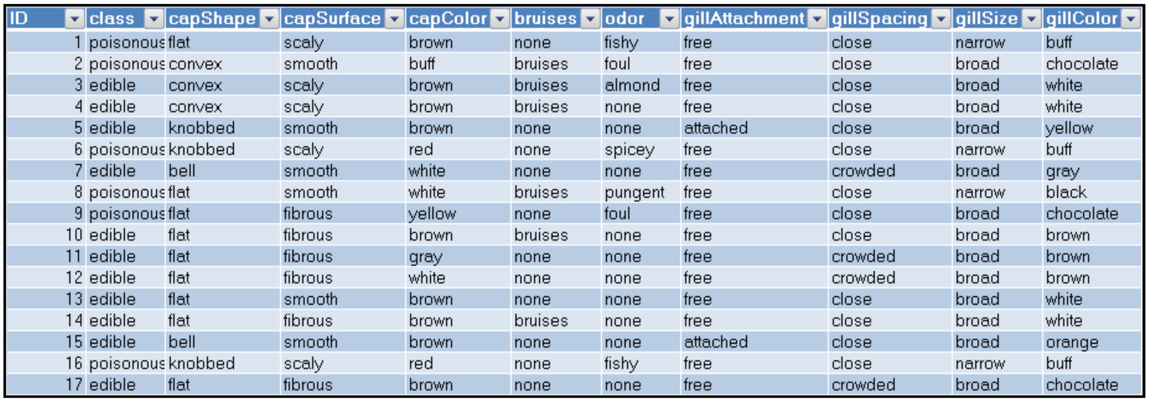


Associate By ModelName

Itemsets

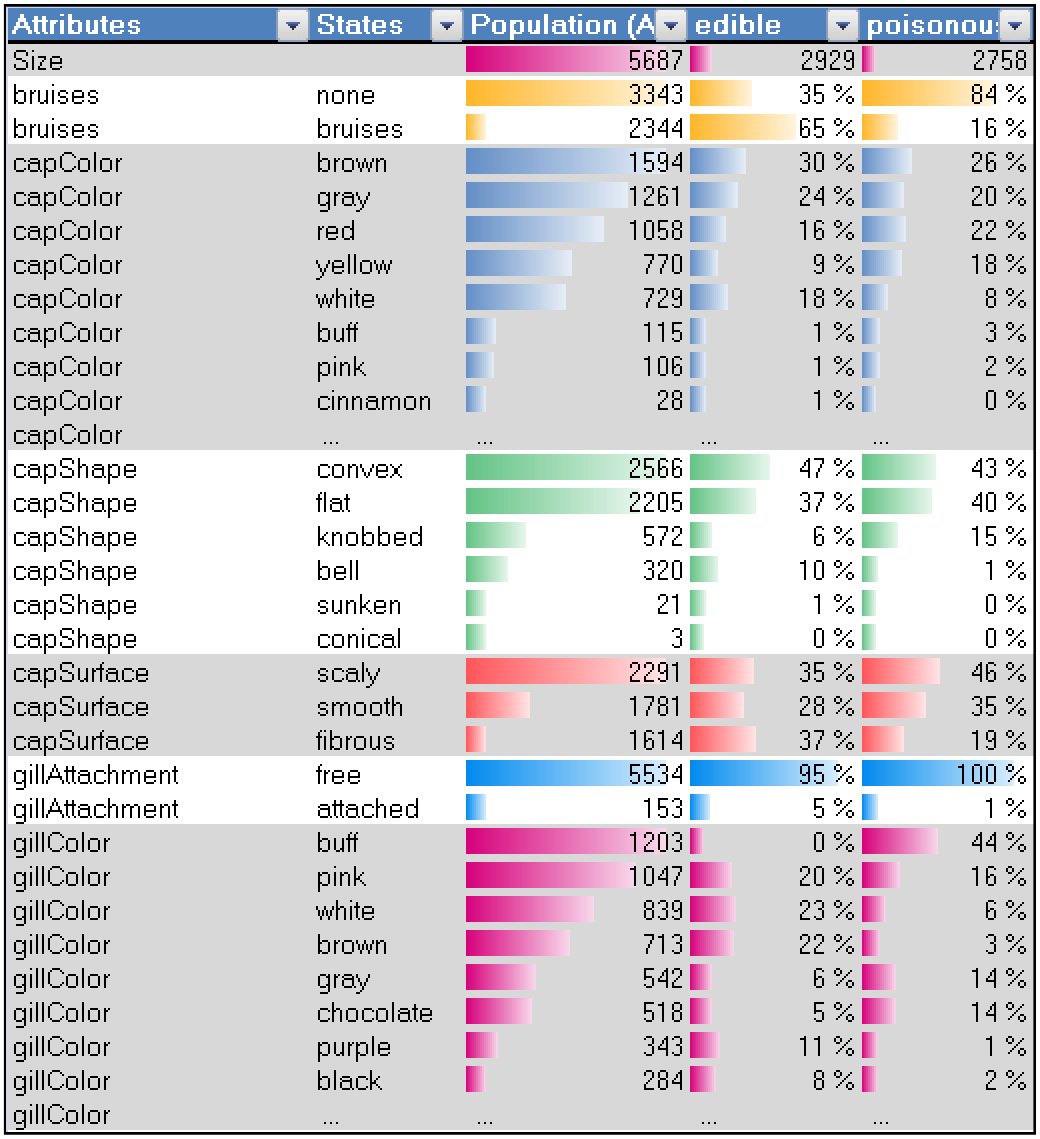
Support	Size	Itemset
5194	1	Sport-100
3251	1	Water Bottle
3086	1	Mountain-200
2385	1	Patch kit
2340	1	Cycling Cap
2163	1	Mountain Tire Tube
2146	1	Long-Sleeve Logo Jersey
1737	1	Road-250
1654	1	Road Tire Tube
1480	1	Fender Set - Mountain
1474	1	Short-Sleeve Classic Jersey
1443	1	Half-Finger Gloves
1396	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 A Sample generated List of Item Sets with their support and size



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

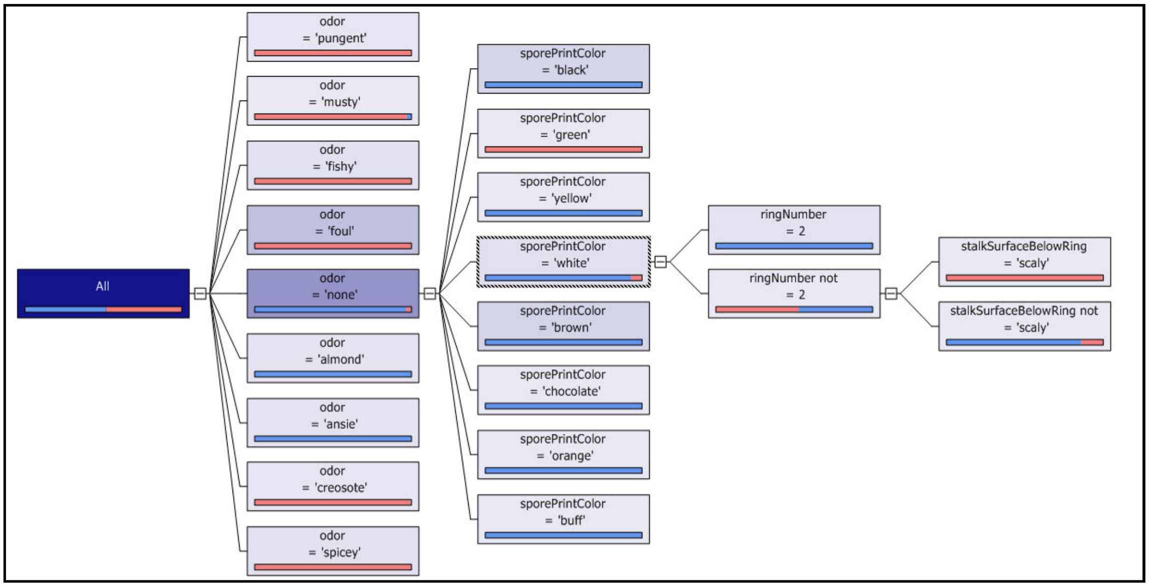
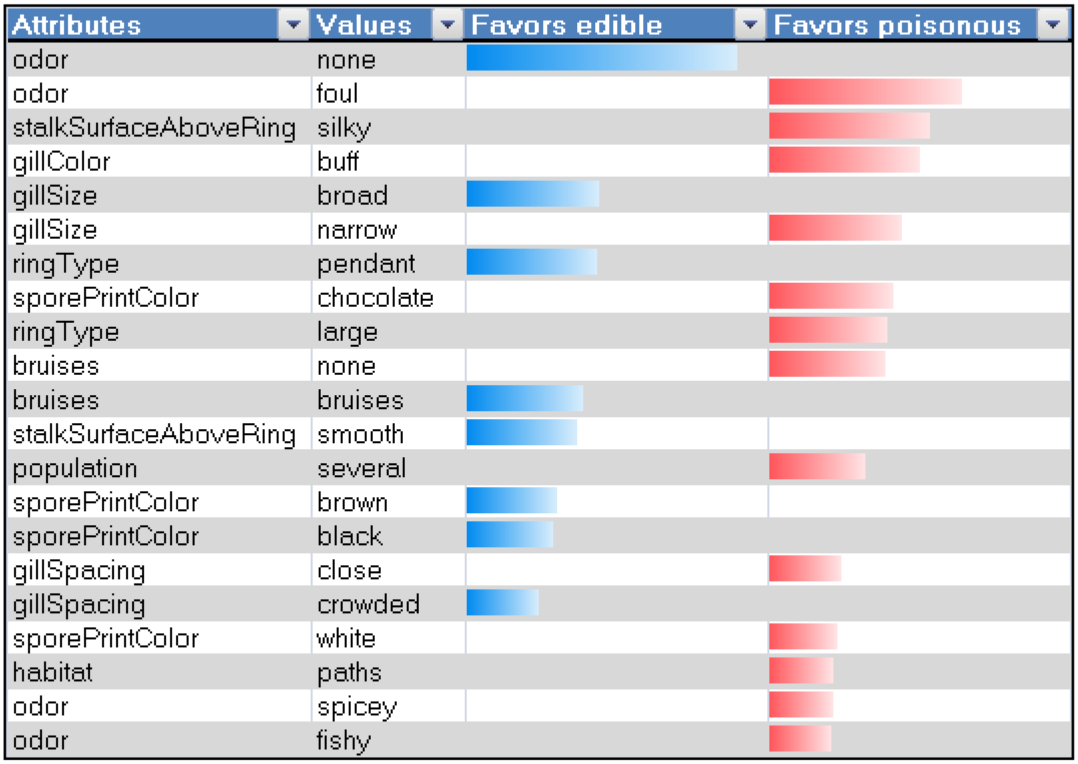
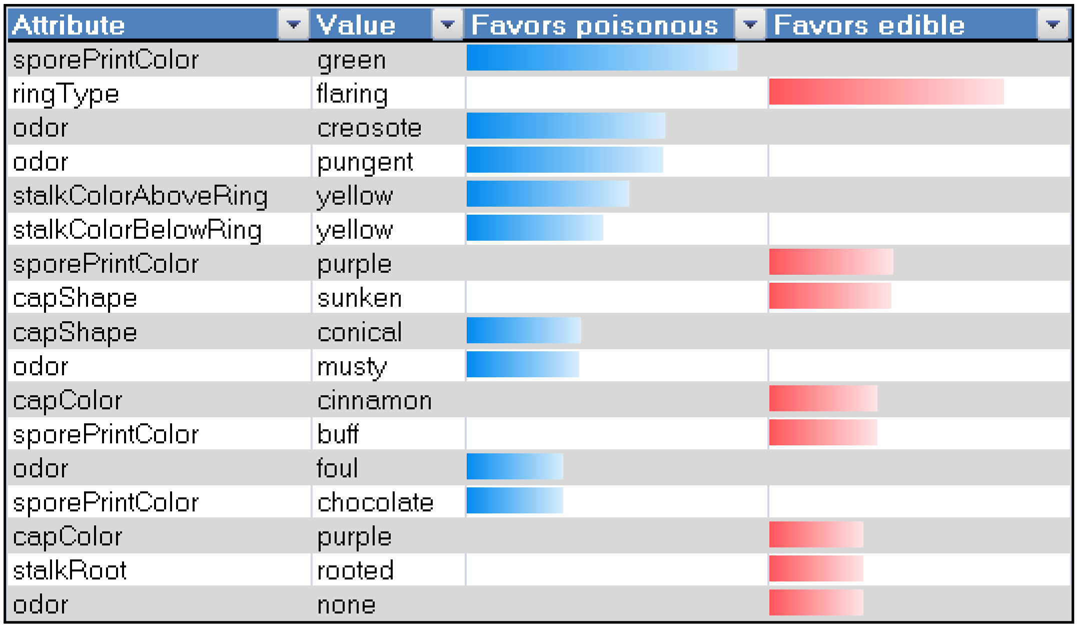

All
odor = 'pungent'
odor = 'musty'
odor = 'fishy'
odor = 'foul'
odor = 'none'
odor = 'almond'
odor = 'ansie'
odor = 'creosote'
odor = 'spicey'
sporePrintColor = 'black'
sporePrintColor = 'green'
sporePrintColor = 'yellow'
sporePrintColor = 'white'
sporePrintColor = 'brown'
sporePrintColor = 'chocolate'
sporePrintColor = 'orange'
sporePrintColor = 'buff'
ringNumber = 2
ringNumber not = 2
stalkSurfaceBelowRing = 'scaly'
stalkSurfaceBelowRing not = 'scaly'

Figure 14 A Decision Tree Classification of the Mushroom Data Set



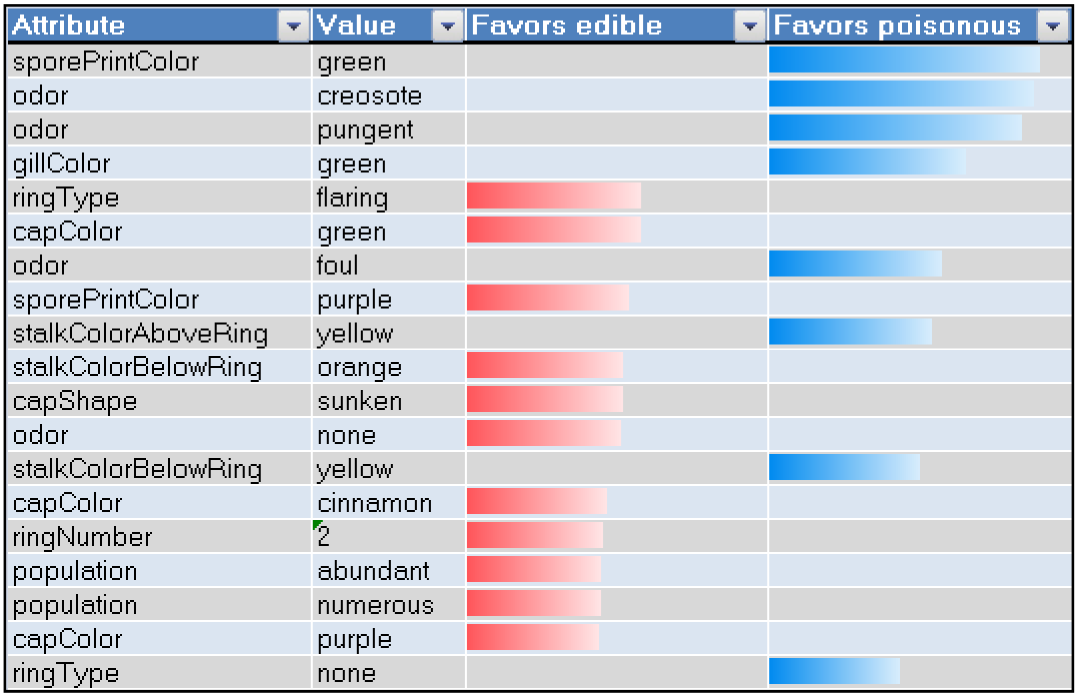
Attributes	Values	Favors edible	Favors poisonous
odor	none	🟦	
odor	foul		🟥
stalkSurfaceAboveRing	silky		🟥
gillColor	buff		🟥
gillSize	broad	🟦	
gillSize	narrow		🟥
ringType	pendant	🟦	
sporePrintColor	chocolate		🟥
ringType	large		🟥
bruises	none		🟥
bruises	bruises	🟦	
stalkSurfaceAboveRing	smooth	🟦	
population	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

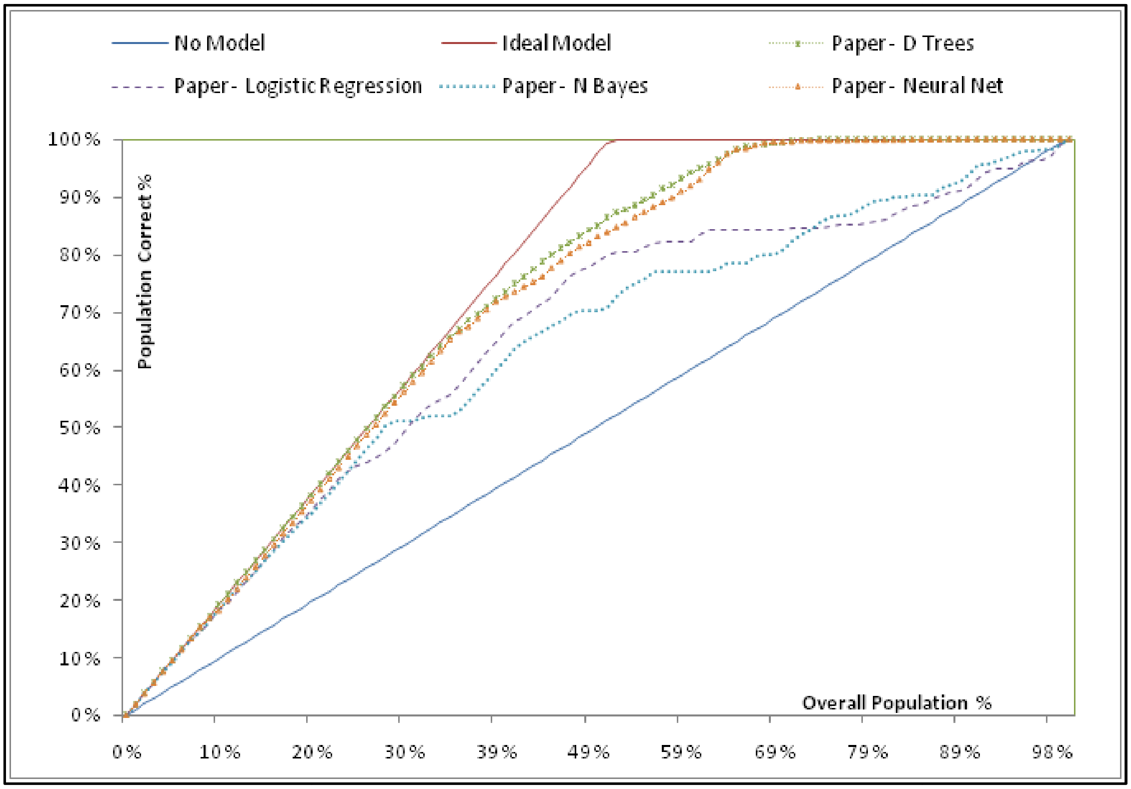
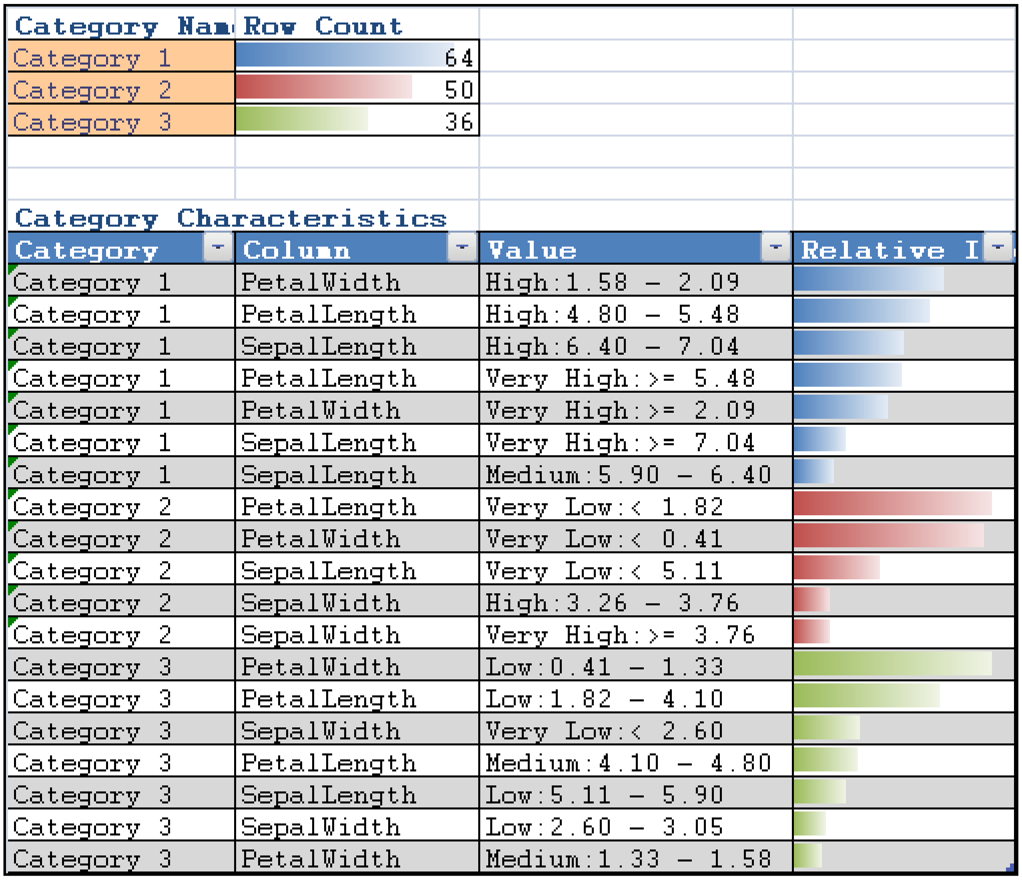


Figure 18 Accuracy Chart of 4 the Classification Models for the Poisonous class

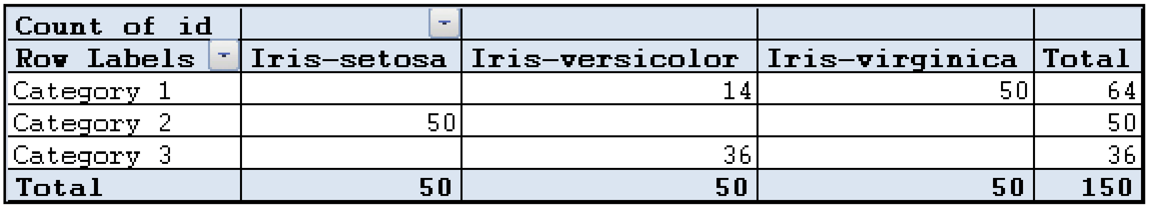


Category Nam	Row Count
Category 1	64
Category 2	50
Category 3	36

Category Characteristics

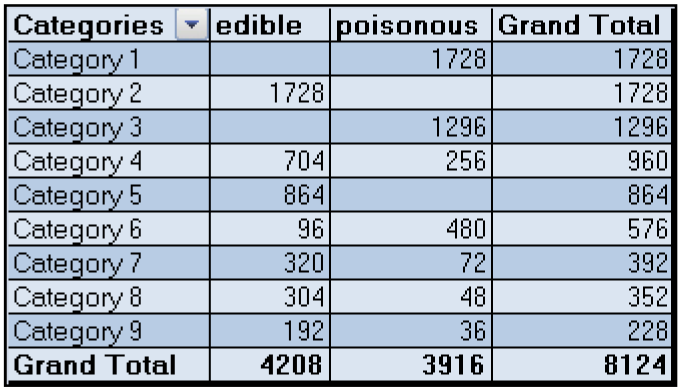
Category	Column	Value	Relative I
Category 1	PetalWidth	High:1.58 – 2.09	
Category 1	PetalLength	High:4.80 – 5.48	
Category 1	SepalLength	High:6.40 – 7.04	
Category 1	PetalLength	Very High:>= 5.48	
Category 1	PetalWidth	Very High:>= 2.09	
Category 1	SepalLength	Very High:>= 7.04	
Category 1	SepalLength	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



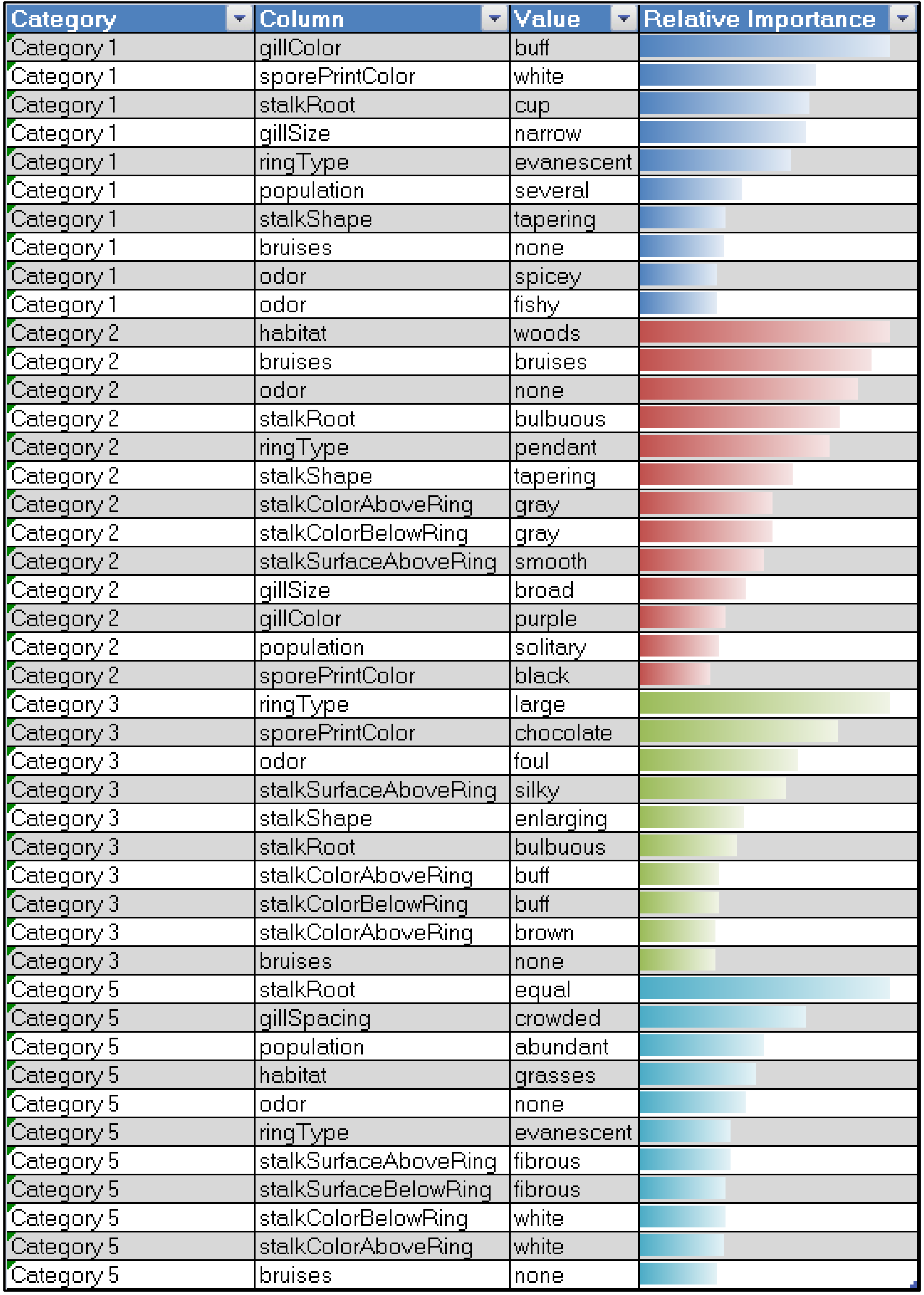
Count of id				
Row Labels	Iris-setosa	Iris-versicolor	Iris-virginica	Total
Category 1		14	50	64
Category 2	50			50
Category 3		36		36
Total	**50**	**50**	**50**	**150**

Figure 20Accuracy 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



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 3 clusters that produced perfect match