**Data Mining Basics KEY**

Formal Assessment

**Objective:** To determine your level of understanding of the nature of data mining and your ability todescribe data mining tools and techniques.

**Please answer the following questions (be specific and detailed)**.

1. What is the definition of data mining?

Data Mining is the process of analyzing data from different perspectives and summarizing it into useful information.

1. Of the two manual methods of data mining that were described in the lesson, pick one and identify and explain it.
* **Bayes’ Theorem**= Thomas Bayes, 1700s. Probability measures a degree of belief andBayes’ Theorem links the degree of belief in a proposition before and after accounting for evidence. Example: For example, suppose somebody proposes that a biased coin is twice as likely to land heads than tails. Degree of belief in this might initially be 50%. The coin is then flipped a number of times to collect evidence. Belief may rise to 70% if the evidence supports the proposition.
* **Regression Analysis**, **1800s**= A statistical process for estimating the relationships among

variables…between one dependent variable and one or more independent variables. It helps one understand how the typical value of the dependent variable (or 'Criterion Variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed

1. What are three reasons someone would use data mining?
	* To determine market trends
	* To save money
	* To make money
	* To determine future consumer spending
	* To analyze consumer spending habits
	* To help see determine patterns
	* To save time
2. Of the six different levels of analysis used in data mining, pick one and explain it in your own words.
	1. Artificial neural networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure.
	2. Genetic algorithms: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.
	3. Decision trees: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.
	4. Nearest neighbor method: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k 1). Sometimes called the k-nearest neighbor technique.
	5. Rule induction: The extraction of useful if-then rules from data based on statistical significance.
	6. Data visualization: The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

What are two very important questions someone should ask and answer before purchasing any data mining software?

* + - How big is/will be your database?
		- How complex are/will be your queries?

**Each question will be worth 20 points.**

