Assignment #2

Due date: **December 2**

Total Score: 100 points

Objective: Data Analysis using Decision Tree Induction

Background

Assume that a local distributor (company) has collected a data set from their customers' spending in two business categories, *Restaurant* and *Retail*. The data set consists of the amount of their customers' spending in US dollar through the company in 6 different product types such as *Fresh*, *Milk*, *Grocery*, *Frozen*, *Detergent*, and *Delicassen*.

A schema of the data set is defined as follow:
Dataset (BusinessCategory, Fresh, Milk, Grocery, Frozen, Detergent, Delicassen)

- Business categories are represented by an integer value either 1 for Restaurant or 2 for Retail business. Data types for all product types are dollar amount.
- This data set in CSV format, "**Dataset.csv**" is posted on the course page.

Two managers, operation manager and marketing manager in this company are interested in knowing what product types and how much each business, Restaurant or Retail tends to spend. This information is useful to operation manager to maintain proper level of inventory. For example, if they see more restaurant owners buying certain types of products, they need to order more products on the types in advance before the product types become out of stock. To marketing manager, this information is useful to decide which business area (Restaurant or Retail) they need to focus on their next marketing strategies or how they need to negotiate with the producers (manufacturers) of certain product types for better deal to increase the profit margin.

Required activities

Write **a brief report** that summarizes your data analysis activities and results including following elements:

(1) Your name(s) and contact email address(es); the percentage contribution to this assignment if this assignment was completed by a team. If a team cannot reach a consensus on the individual contribution, include the individual's claimed percent contribution with a brief description on specific tasks performed by each member.

(2) A brief description about the software tool you used or program you implemented.

(3) The results of your data analysis including the following elements:

(a) The name of the class/decision column (out of 7 columns in this data set) for your analysis

(b) A brief explanation of the process of your analysis listing a series of steps taken, specifying input and output if there is any, to produce the final decision tree from this data set

(c) A snapshot of the final decision learned. If the tree is too big, show only the important part of the tree based on your judgment (to meet managers' expectations).

(d) Give two rules you can obtain from the decision tree that can be used to explain the nature of the data. If a rule is too big, you can show only the important part again based on your judgment.

(e) A brief explanation of knowledge learned from the data analysis in layman terms so that two managers can understand.

Grading policy

The grade will depend on the quality of report that clearly demonstrates your research, the quality of analysis results, the level of understanding on the related subjects, effort, and writing.

----------**How to Submit this Assignment**----------

Upload your report in **Word** format (**NOT PDF**) to Titanium. Unless the source code is your own implementation, **DO NOT** include the source code. Instead, clearly specify the source of the programs you used.
