Assignment #2 (total score: 70)

Due date: **December 1**
To be completed by a team

Background

A car manufacturing company wants to know what kind of cars will be desirable to their customers. So, the company has collected data that describe the major car characteristics including car price, maintenance cost, comfortability (based on the number of doors and passenger capacity), trunk size, and estimated safety of a car.

The schema of the data set is defined as follows:
Cars (*CarPrice, MaintCost, NumDoors, NumPersons, TrunkSize, Safety,* ***Rating***)

In this data set, there are **6 attributes** including CarPrice, MaintCost, NumDoors, NumPersons, TrunkSize, and Safety. **Rating** defines the **decision classes**, each of which represents the **target concept** (bad, acceptable, good, and excellent).

Possible values for each attribute are as follows:

- *CarPrice*: vhigh, high, med, low where vhigh means very high
- *MaintCost*: vhigh, high, med, low
- *NumDoors*: 2, 3, 4, more where more represents 5 or more doors
- *NumPersons*: 2, 4, more where more represents 5 or more persons
- *TrunkSize*: small, med, big
- *Safety*: low, med, high
- *Rating*: bad, acc, good, excl where acc represents "acceptable" and excl represents "excellent"

The data set in CSV format, "**Cars.csv**" is posted on the course page.

The company wants to know **which car characteristics** lead customers to rate whether a car is **bad**, **acceptable**, **good**, and **excellent**. This information is important for the company to decide the type of car they will produce in the future.

Required activities

Analyze the data using a **decision tree induction approach** to learn the customer preferences in terms of ratings and **write a brief report** that summarizes your data analysis activities and results including:

(1) Your name(s) and contact email address(es). The percentage contribution made by each member when 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 including the names of the tool and developer (or URL). If you wrote the program yourself, you may clearly say so.

(3) The results of your data analysis:

(a) A brief explanation of your analysis process listing the steps taken to produce the final decision tree from this data set

(b) 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 (that may meet the company's expectations).

(c) Two of the most important rules from the decision tree based on your judgement (that may meet the company's expectations). If a rule is too big, you can show only the important part based on your judgment.

(d) A brief recommendation for the company in layman terms so that a management team of the company can understand

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.

Grading policy

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