COMPUTER SCIENCE 4373
Assignment #6

Points: 100 **Weight**: 2%
Due: Friday, Oct. 16, 2020 at 11:50 pm on BlackBoard
Note: Late assignment will not be accepted without instructor's pre–approval.

Instruction: This assignment should be completed individually. Please make sure your answer is legible (and preferably formatted using MS Words or LaTeX/LyX). If a question requires you to follow an algorithm, show a clear trace of the algorithm. If the algorithm is iterative, show the details in the first two iterations. For each of the remaining iterations, show the status of the algorithm at the end of the iteration. Please also submit to the BlackBoard a single zip file your_name_hwk06.zip which should contain your solution in a PDF, a Word document, or a Jupyter Notebook (with narrative formatted in markdown cells), the program source code, the input data, and the output of your program.

Notice: *Given the nature of on-line course, we will require you to practice using Words, Markdown, or HTML to write and format your homework solutions (**no scanned smeared image please**). It will prepare you for taking the on-line exams, where only a Words style editor (with HTML support) is available.*
Suggestion: *If the trace of an algorithm involves a lot of calculations, you may want to show the details in the first place, and then write some scripts to perform the calculation for subsequent cases. The Jupyter notebook will be very handy for such cases.*

The following questions are based on this dataset (table) from an employee database (also in data file hwk06-01,csv).

department	status	age	salary	count
sales	senior	31..35	46k..50k	30
sales	junior	26..30	26k..30k	40
sales	junior	31..35	31k..35k	40
systems	junior	21..25	46k..50k	20
systems	senior	31..35	66k..70k	5
systems	junior	26..30	46k..50k	3
systems	senior	41..45	66k..70k	3
marketing	senior	36..40	46k..50k	10
marketing	junior	31..35	41k..45k	4
secretary	senior	46..50	36k..40k	4
secretary	junior	26..30	26k..30k	6

The data is a summary of the original data table. For example, the first row indicates that 30 employees in the sales department has an age between 31 and 35 inclusive and a salary between 46K and 50K inclusive. The attribute status is the class label.

1. [25] Write Python code to perform the following tasks.

   (a) Read data from hwk06-01.csv into a DataFrame df1.

   (b) Create a new DataFrame df2 by replicating each row of df1 with the number of copies as indicated in the count column. For example, the first row in df1 should appear 30 times in df2, second row in df1 40 times in df2, etc.

   (c) Perform the one-of-kinds encoding on the categorical columns. For example, if a categorical attribute has three values "A", "B", and "C", you will encode "A" as 0, "B" as 1, and "C" as 2. One way to do this in Python is to save the values in a list and use the list index as the coding.

   (d) Make sure that df2 does not have the column count.

2. [25] Write Python code to use sklearn.tree.DecisionTreeClassifier to learn a decision tree using the df2 as the training data, and then use the decision tree to predict the status of a user provided unseen data, such as

$$t =< department : systems, status :?, age : 28, salary : 50K >$$

   Your program needs to convert the actual age and salary values into the codes for the corresponding ranges before using the decision tree to predict the status. Also, use the graphviz package to display the learned decision tree.

3. [25] Write Python code to create another DataFrame df3 from df1 as follows.

   (a) Replicate rows as described above

   (b) Convert values in the age and salary columns to random values drawn from the specific range for each row. For example, suppose the age of a row is "31..35", replace it by a random integer between 31 and 35 inclusively.

   (c) Perform the same encoding on column department

   (d) Make sure there is no column count

4. [25] Write Python code to use sklearn.naive_bayes to learn a Guassian Naive Bayes classifier using df3 as the training data, and use the learned predictive model to predict the status of a user provided unseen data, for example,

$$t =< department : systems, status :?, age : 28, salary : 50K >$$

   Again, you need to encode the department.