Introduction to Data Science

Homework 2

40 points

1. Use ssh to connect to bl2.cs.fsu.edu and login with your *bl2* user name you used for Homework 1. You can do it either from the internal network (any machine in cs.fsu.edu domain) or outside (e.g. from shell.cs.fsu.edu)
2. Consider file traindata.csv outside your home folder (../traindata.csv). It contains synthetic labeled data in a form of a table, each row having a patient who did or did not have a heart attack in the past. It is a synthetic dataset composed for this assignment, it does not have any real medical data.
    a. Each row has more than 10 attributes and the last attribute (*Result*) – a label (*yes/no*), indicating whether a patient did or did not have a heart attack in the past.
3. Separate a part of the file traindata.csv (for example 1/10th) and save it as a separate file - testdata.csv. Remove this 1/10th from traindata.csv and do not use it for step #4.
4. Browse the file (traindata.csv) to get familiar with its contents and **manually** design a formula that can calculate the label value, based on the attributes values.

   IMPORTANT!!: **DO NOT USE ANY EXISTING MODELS AND DO NOT DEVELOP ANY MODELS BY YOURSELF**. If you do (let's say take and train an already existing Decision Tree or any other model, then take the automatically generated rules and use them to help you solve this assignment), then you'll lose all credit for this assignment.

    a. Start from something simple, for example:
        i. if (Age>=54 and Gender=1) then *Result*=no

           else if (Age<40 and Gender=0) *Result*=yes;

       Note that this is not the final rule that you need to turn in, but just an example of how the design process can be started.

   b. As you start designing and trying out your rule, you will see it works for some rows (i.e. the *Result* generated by your rule is the same as the *Result* value already present in the row). Try changing or enriching your rules with more attributes, so the final rule performs better (i.e. makes less mistakes compared to the *Result* value present in the rows).

   c. Make iterations to optimize your rule to make it perform as accurate as possible.

   d. You can use any suitable programming language to execute your rule on the test set.

5. When you finish changing your rule, calculate its *accuracy* on testset.csv (see step #3) by using the following formula:

   a. $100\% * \frac{\#rows\ in\ the\ test\ set - \#incorrect\ labels\ generated\ by\ your\ rule\ on\ the\ test\ set}{\#rows\ in\ the\ test\ set}$

What to turn-in

Turn in a PDF file having the following items:

1. (25 points) The rule you designed to label the data rows
2. (5 points) Explanation of your rule and the intuition you used to design it
3. (10 points) The accuracy of your rule calculated on testdata.csv separated from the main file – traindata.csv in step #3.
4. Do not forget to write your name at the beginning of the file.

Good luck!