CS 4661: Introduction to Data Science
Dr. M. Pourhomayoun
Homework4
Due Date: Wed, Nov 15

Up to 2 students can team up to work on this homework. One student should submit the homework on behalf of the team. Make sure to include the name/CIN of everyone on every submitted file. Write and submit your python codes in "Jupyter Notebook" to perform the following tasks. Make sure to provide proper descriptions as MarkDown for each section of your code describing what this part of the code is supposed to do!

Question1: Cancer Diagnosis Using Machine Learning

Write and submit your python codes in "Jupyter Notebook" to perform the following tasks:

In this homework, we work with a real dataset from UCI Dataset.

a- Read the dataset file "Cancer.csv" (you should download it from CSNS), and store it in a Pandas DataFrame. Check out the dataset. As you see, the dataset includes 9 numerical features. The last column is the binary label ("1" means it is a malignant cancer, "0" means it is a benign tumor). You will use all 9 features in this homework.

b- Use sklearn functions (see class tutorials for details) to split the dataset into testing and training sets with the following parameters: **test_size=0.3, random_state=2**.

c- Use "Decision Tree Classifier" to predict Cancer based on the training/testing datasets that you built in part (b). Then, calculate and report the accuracy of your classifier. Use this command to define your tree:
my_DecisionTree = DecisionTreeClassifier(random_state=2).

d- Use scikit-learn "Random Forest" classifier to predict Cancer based on the training/testing datasets that you built in part (b). Then, calculate and report the accuracy of your classifier. Use this command to import and define your classifier:
from sklearn.ensemble import RandomForestClassifier
my_RandomForest =
RandomForestClassifier(n_estimators = 19, bootstrap = True, random_state=2)

Similar to previous syntax, use **my_RandomForest.fit** for training your random forest classifier and **my_RandomForest.predict** for prediction.

Question2: predict the probability of Heart Disease

Write and submit your python codes in "Jupyter Notebook" to perform the following tasks. Make sure to provide proper descriptions as MarkDown for each section of your code.

a- In this question, we work with a simplified version of Heart dataset (remember that this dataset is a little different from what you have used in HW2). Read the dataset file "Hearts_short.csv" (you should download it from CSNS), and assign it to a Pandas DataFrame.

b- Generate the feature matrix and label vector (AHD). Then, normalize (scale) the features.

c- Split the dataset into testing and training sets with the following parameters: test_size=0.2, random_state=3.

d- Use Logistic Regression Classifier to **predict** Heart Disease occurrence based on the training/testing datasets that you built in part(c). Then, compute and report the **Accuracy**.

Now, Use Logistic Regression Classifier to **predict the probability** of Heart Disease based on the training/testing datasets that you built in part (c) (you have to use "my_logreg.**predict_proba**" method rather than "my_logreg.**predict**"). Then, Plot the **Roc Curve** for this classifier, and also Compute the **AUC** (Area Under Curve for ROC).