Machine Learning HW 5

(1) 編寫一個 Python3 程序以讀取 HW5 數據文件（”hw5_cancer.csv”)。整體共計 569 數據集 + header. 除了header, HW5 數據文件每行是1個數據集(dataset). 每個數據集(每行)包含 30個 features and 1個 classification (0 是 惡性, 1 是 良性). 每個數據都用逗號分隔.

Write a Python code to read HW5 data file (”hw5_cancer.csv”)。There are 569 datasets, 1 dataset per line, plus a header line in the CSV file. For each dataset, there are 30 features and 1 classification (0 as malignant and 1 as benign). Data is separated by a comma.

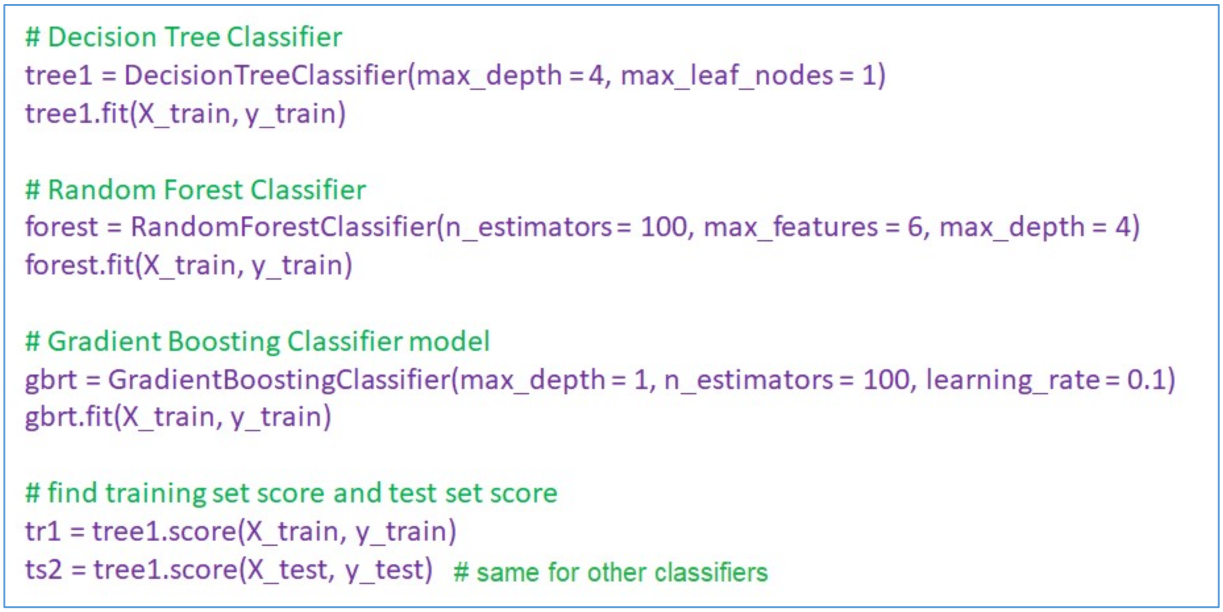
(2) Use 426 (75%) datasets as training data and the remaining as the test data.

(3) Use (a) **Decision tree**, (b) **Random Forests**, (3) **Gradient Boosted Regression Trees**, to train your model with the training and test data.

(4) You are to find the model which makes the test score above 0.958 while the training score less than 0.990. Any model will do.

(5) You can import the corresponding classifiers as shown below. Sample Python codes using 3 classifiers are given below.

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import GradientBoostingClassifier
```



```
# Decision Tree Classifier
tree1 = DecisionTreeClassifier(max_depth = 4, max_leaf_nodes = 1)
tree1.fit(X_train, y_train)

# Random Forest Classifier
forest = RandomForestClassifier(n_estimators = 100, max_features = 6, max_depth = 4)
forest.fit(X_train, y_train)

# Gradient Boosting Classifier model
gbrt = GradientBoostingClassifier(max_depth = 1, n_estimators = 100, learning_rate = 0.1)
gbrt.fit(X_train, y_train)

# find training set score and test set score
tr1 = tree1.score(X_train, y_train)
ts2 = tree1.score(X_test, y_test)  # same for other classifiers
```

6). 估計所需時間：2-4小時

(7). 截止時間: 在2021年12月3日上課之前提交 zipped 的 python 程序 (“ yourID_name _HW5.py”) 和 print-screen 圖 (“ yourID_name_HW5_cancer.jpg”) of the test score output and the parameters which meet the requirement as “yourID_name_HW5.zip”. Submit anyway even if you cannot meet the target.

(8). 截止時間: Before the class time on 12/3. Just submit your Python code and best result – even if you cannot find the model.