Homework 4

Instructions. This homework is due by Friday, September 29th at 5pm. You can edit this file and copy/paste into it by opening it in Libreoffice on Linux, PDFfiller Google App, Acrobat, Microsoft Word and others. If you use the internet for information, please assess the verity of the source and document it as a source. Please turn in a pdf version of the written solutions, as well as a zip file of all code you use to solve the problems. Please name each piece of code by the question it is intended to answer or otherwise clearly indicate to me how to run the code to get each answer.

1. There are two files on Blackboard and `hpc-class`, `yield.train.csv` and `yield.test.csv`, that contain information on crop yields and various potential predictors (features), as listed below:

    1. County name
    2. Year (1995 – 2004)
    3. Drought Index (DI)
        - Normal conditions (0)
        - Wet conditions (> 0)
        - Drought conditions (< 0)
    4. Yield: the response variable
        - Very Low
        - Low
        - High
        - Very High
    5. Soil: the quality of the soil, higher is better
    6. Rainfall (RF): rainfal recorded for the period May–Sep (tenths of millimeters)
    7. TMin, TMax: average mininum or maximum temperature recorded for the period May–Sep (tenths of Celsius)

    (a) Train a decision tree using `sklearn.tree.DecisionTreeClassifier` using the training data and assessing performance on the test data. Assess the performance using Matthew's correlation coefficient, accessible via `sklearn.metrics.matthews_corrcoef()`.

    i. The Gini impurity index for a group of observations in a multi-class classification problems is

$$G = \sum_{i=1}^{n_c} p_i(1 - p_i),$$

    where n_c is the number of classes, and p_i is the proportion of the observations in the ith class. The importance of split s is

$$I_s = G_{s,\text{parent}} - G_{s,\text{child 1}} - G_{s,\text{child 2}},$$

where $G_{s,\text{parent}}$ is the Gini index of the parent node, and the others are the Gini indices of the daughter nodes. The **Gini importance** of a feature is obtained by averaging the importance of all splits involving that feature, possibly weighting by the number of affected observations. What feature has the largest Gini importance in the tree?
(Hint: see the sklearn.tree.DecisionTreeClassifier manual).

ii. What maximum tree depth achieves the best performance? Beware of randomness in the answer!

iii. What causes the randomness in the estimated trees from multiple runs on the same data?

(b) Train a random forest using `sklearn.ensemble.RandomForestClassifier` using the training data and assessing performance on the test data.

i. Why does increasing the `n_estimators` argument reduce the variability in the estimated random forests from multiple runs?

ii. How does restricting the maximum depth of the tree affect performance? Thoughts why?

iii. How does the fraction of features considered at each split affect performance? Thoughs why?

iv. Compute and display the Confusion matrix from the best random forest result you have obtained. Does the classifier have particular problem distinguishing certain outcomes and not others? Why might that be?

2. Apply the decision tree or random forest to the permissive, nonpermissive virus sequence data to determine (1) if you can distinguish permissive and nonpermissive cells based only on the virus fragments they contain and (2) what motifs (nucleotide patterns) best distinguish the two types of cells. This is an open-ended question. You must identify useful features, choose on of your two machine learning techniques, and choose its parameter settings. Justify and critique your choices. Analyze and interpret your results.

Hint. I do not know what you learned in lab, but I found the pandas Python library particularly helpful for massaging the data. In particular, I learned a lot from this blog post.