**Classification using different classification tools**

(NumPy and data visualization packages are allowed.)

(SKLearn tools and models are allowed.)

Requirement:

Select a public dataset with multiple classes. Use at least three different classifiers to classify them. Compare and discuss the classification performance of different models that you selected. Discuss what could affect the performance for different classifiers (e.g. values of hyperparameters that you used) and how they affect the performance.

1. **Select a dataset**: select one from iris.csv or MNSt.csv or

any public dataset (include your csv file in your submission if you use another dataset)

* decisionTreeClassifier\_iris.py => iris.csv dataset
* knnLogReg\_iris.py => iris.csv dataset
* linear\_reg\_diabetes.py => diabetes dataset

1. **Split the dataset**:

Please randomly partition the data into training and testing datasets. For example, 70% for training, and 30% for testing.

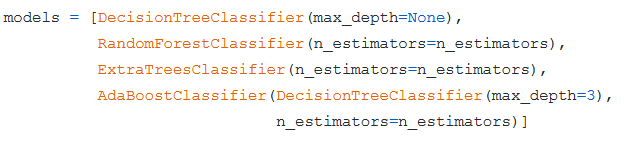
**If you want to better select the values of some hyperparameters, you will need to do validation** (not required, but suggested). You will need to partition the data into these three splits. For example, 70% for training, 10% for validation and 20% for testing. Report your final performance only using the testing dataset**. (2pts)**



Hyperparameters:



1. Select at least **three different types** of classifiers based on different techniques (linear SVM and nonlinear SVM belong to **one type**) and fit the training set to your models. You might need to convert the string label values to numbers using “LabelEncoder in sklearn”. **(2pts)**

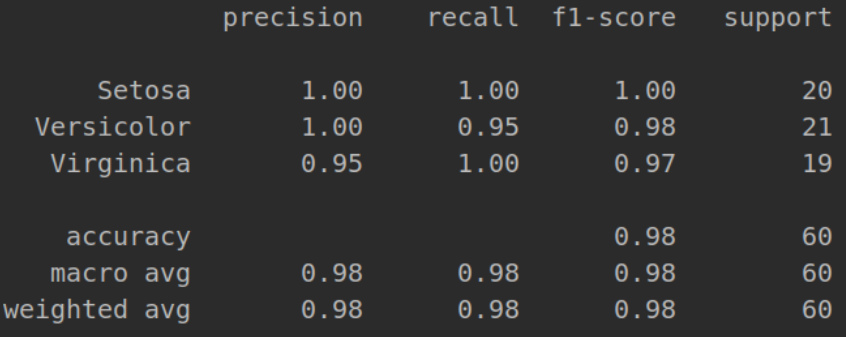




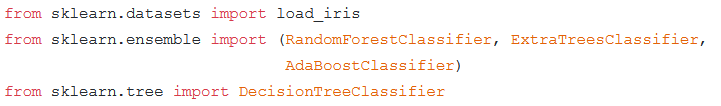


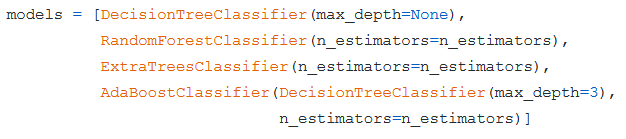


1. **Train your Classifiers** and perform validation if needed using different hyperparameters values
2. **Evaluate**: Once your classifier is trained, you should evaluate the classification performance (accuracy, precision, recall, F-measure) on the test set. These scores need to be included in your report. **(5pts)**



1. **Repeat the training and testing for different classifiers you selected.** Write down all classifiers that you selected and write down model resource for all of them in your report. If some classifier was defined by your own function, please include that function definition (code) in your report. **(5pts)**















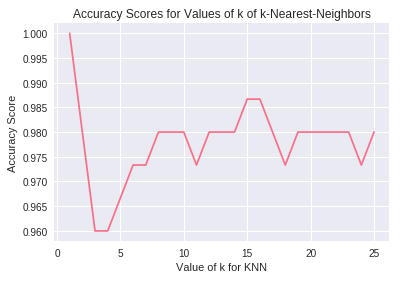
1. **Which is the best classifier and why?** Answer or analysis of this question needs to be included in your report. **(3pts)**

My analysis of this questions leads me to the point of view KNN (K-Nearest Neighbors) is a perfect classifier. However, the problem is that as the complexity of the data increases, KNN just starts being less effective. Therefore, the question of what classifier is the "best" classifier usually comes down to your data, and most of the time, the way you preprocess you data is more important than the classifier you use. In summation, the classifier is only going to be as good as the data you put in it.

1. **What affect the classification performance?** Answer or analysis of this question needs to be included in your report. **(3pts)**

*You could try different sizes of the training/testing data, different hyperparameters’ values, Or even different datasets using different classifiers.*

After trying a multitude of different k-values, the lower the value the more complex the model. I also plotted the different k-values to illustrate this…



1. **Bonus up to 5 pts.** (Anything that was not listed above.)

<https://github.com/GitNickProgramming/cs3120_MidtermProject>

I did a bunch of documentation, as well as different model types. This might be easier to grade if you use the link to my GitHub page.

**Submission**:

File 1: Report

File 2: Code (+dataset if you used another dataset that not in blackboard)

Write a report to describe /answer required questions of your design.

Upload your code with comments as a separate .py or zip file.

**e.g.**

**File1: MidtermHW\_FirstnameLastname.doc/.pdf (this is the report)**

**+**

**File2: MidtermHW\_ FirstnameLastname.py (this is the code. only “.py” files accepted.**

**OR**

**MidtermHW\_ FirstnameLastname.zip if you have multiple “.py” files or another dataset.**