CSC 404/604, Fall 2024, Assignment 10

handed out Nov. 18, due Nov. 25 at 11:59pm on Canvas

In this assignment, we will learn how to use two different classifiers: k-Nearest Neighbors and Decision Trees on two data sets: Iris and Breast Cancer. You will also test whether you agree with some of the research findings concerning the dataset you already used in Assignments 5, 8 and 9 which related to chances of survival after heart failures.

Here are the datasets you will use in this assignment:

Iris: https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_iris.html

Mushrooms: https://www.kaggle.com/datasets/uciml/mushroom-classification

Heart Failure: https://archive.ics.uci.edu/dataset/519/heart+failure+clinical+records

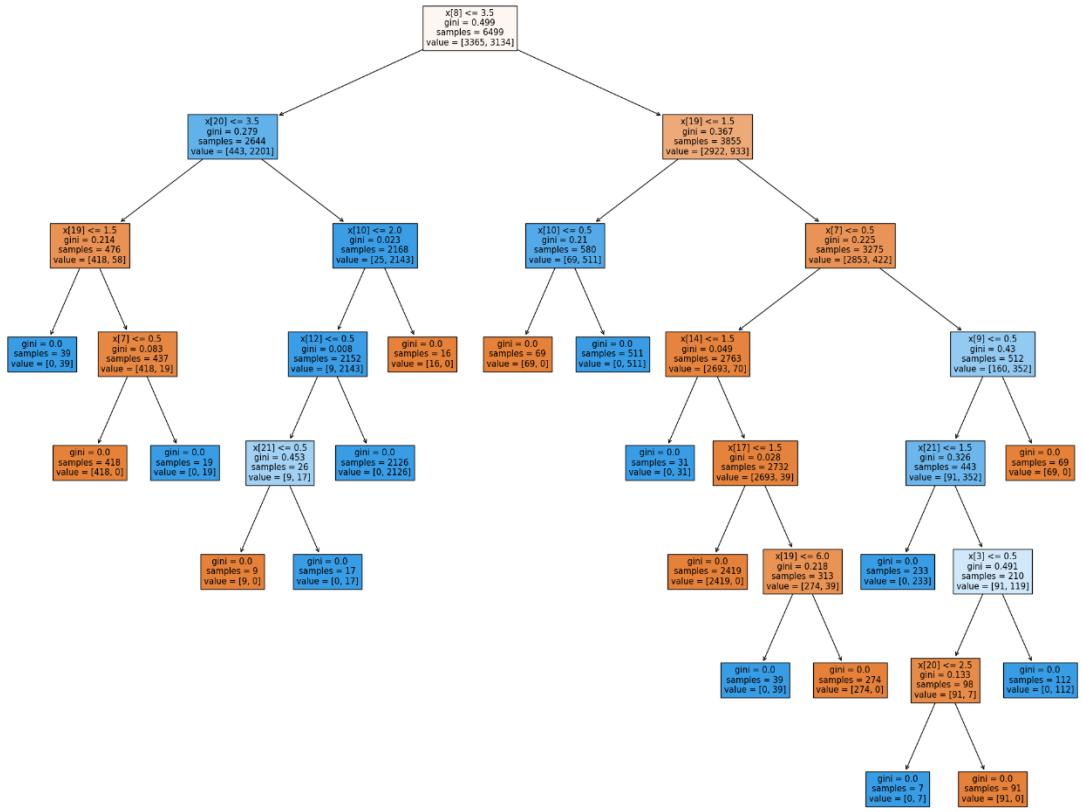
Here is the description of the assignment:

The first part of this assignment asks you to train the k-Nearest-Neighbor's algorithm and Decision Tree algorithm on the Iris data set. A description of this data set can be found at: https://www.kaggle.com/datasets/uciml/iris. Because it takes only four features, it is easy for us to experiment with it so as to understand what machine learning does for us. In particular, after training the two classifiers on the data set, I will ask you to find instances where the two classifiers agree and instances where they disagree on the class of examples. Here is an example where my trained classifiers disagreed:

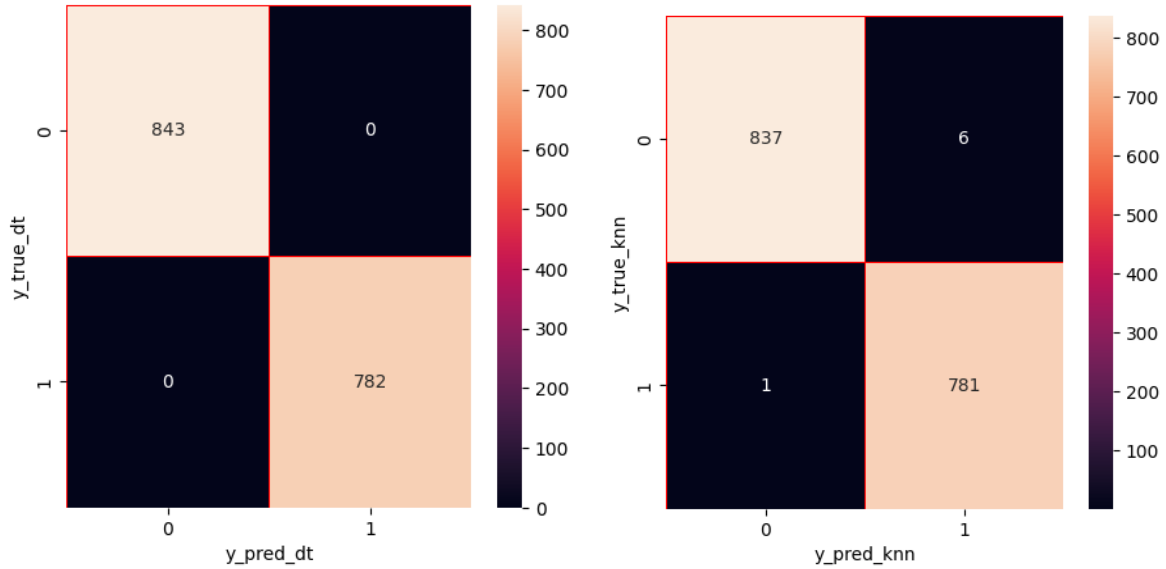
```
[[5.  2.9 3.  0.6]]
KNN Prediction:
['versicolor']

 DT Prediction:
['setosa']
```

Explain what may be going on with your classifiers and the data you are testing them on.

Moving on to the Mushrooms data set, you are asked to train a decision tree classifier on it, visualize the tree you have learned, and show the confusion matrix you obtain when training on a training set and testing on a separate testing set. You are also asked to show the confusion matrix you obtain from the K-NN Classifier (Note: I used k=11, a bad parameter, to force the classifier to classify the data badly (and really differently from the Decision Tree!). The difficulty with the mushroom data set is that you need to apply some data preparation steps the data in order to run the algorithms. See https://www.milindsoorya.com/blog/mushroom-dataset-analysis-and-classification-python. Here are some results I obtained:



And the decision tree's confusion matrix on the left; knn (k=11)'s on the right



After obtaining results of this kind, take a moment to analyze them and explain how such results can be helpful.

In the last part of the assignment, you will work with a more realistic data set: the heart failure data set you have already encountered in assignments 5, 8, and 9. Here again, you will be asked to train a Decision Tree and k-nearest neighbors and test them on a testing set separate from the training set. In the paper by Chicco et al., 202, associated with that data set, the authors claim that two features: serum creatinine

and ejection fraction were not only sufficient to predict heart failure survival, but, in fact, helped obtain better predictive results. We will test whether we agree with this assessment by testing our two classifiers with all 12 features + the class DEATH_EVENT and with only serum creatinine and ejection fraction + the class DEATH_EVENT. You will run both classifiers on both the full and reduced data set and show the results using the classification report feature.

I have done so using the decision tree classifier (with entropy) and the results I obtained confirm the paper's hypothesis. (See below). See if you can reproduce these results (not necessarily exactly), and add new ones for the k-NN classifiers.

Full set of features:

	precision	recall	f1-score	support
0	0.83	0.84	0.83	57
1	0.72	0.70	0.71	33
accuracy			0.79	90
macro avg	0.77	0.77	0.77	90
weighted avg	0.79	0.79	0.79	90

Reduce set of features (only two features):

	precision	recall	f1-score	support
0	0.93	0.94	0.93	68
1	0.81	0.77	0.79	22
accuracy			0.90	90
macro avg	0.87	0.86	0.86	90
weighted avg	0.90	0.90	0.90	90

For your reference, the paper I cited above can be found at:

https://www.semanticscholar.org/paper/Machine-learning-can-predict-survival-of-patients-Chicco-Jurman/e64579d8593140396b518682bb3a47ba246684eb