Assignment 2

Garcia, Jorge A.
Department of Physics
New Mexico State University
(Dated: October 4, 2019)

Course: C S 508

Instructor: Dr. Tuan Le

I. HOLDOUT

The script "prob1_holdout.py" loads the data from the "glass.csv" file, and uses the *train_test_split* function from the *scikit-learn* library to do an 80-20 split of the data into training and testing sets. The new sets are then saved in their own CSV files in the "data" folder.

II. DECISION TREE CLASSIFIER

The code for this section can be found in the "prob2_trees.py" file. It uses the *DecisionTreeClassifier* class from the *scikit-learn* library. The trees are trained 50 times to obtain a statistical description (average and standard deviation) of its performance and take into account the random initialization which can converge differently.

A. Tree Impurity: Entropy

The resulting training and testing accuracy for the entropy impurity trees can be seen in Figure 1. As the maximum depth increases, the training data is eventually perfectly fitted. The accuracy for the testing data is not as promising, eventually remaining constant at ∼68%; a clear result of overfitting. The best model is obtained by having a max tree depth of 7, which is the point that maximizes the training set accuracy.

B. Tree Impurity: Gini Index

The resulting training and testing accuracy for the gini index impurity trees can be seen in Figure 2. As with the entropy impurity trees, the training accuracy increases with the maximum depth eventually being able to perfectly fit the training data. The testing accuracy increases up until a maximum tree depth of 4, after which it decreases and remains constant at 68% as well.

III. K-NEAREST NEIGHBOR CLASSIFIER

The code for this section can be found in the "prob3_knn.py" file. It uses the *KNeighborsClassifier* function from the *scikit-learn* library.

Plotted in Figures 3 - 5 is the testing and training accuracy of the K-Nearest Neighbors algorithm using Euclidean, Manhattan and Cosine metrics accordingly. The same overall trend is followed throughout: the training set is predicted perfectly when considering a single neighbor (which must be the case; it is the data point itself) and decreases as the number of neighbors increase. The highest testing accuracy, for the three metrics, can be seen when considering a single neighbor. As the number of neighbors increase, an overall decrease in accuracy is seen.

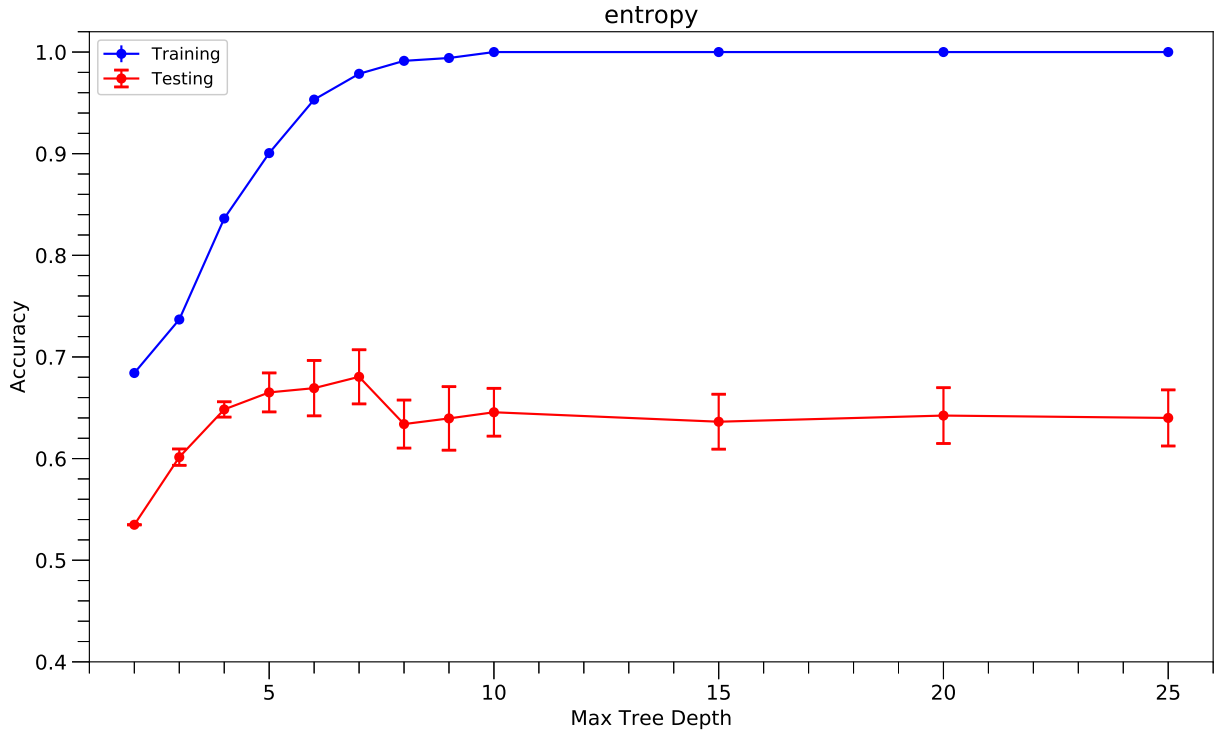

FIG. 1. Training and testing accuracy of decision trees using entropy impurity for different maximum tree depths.

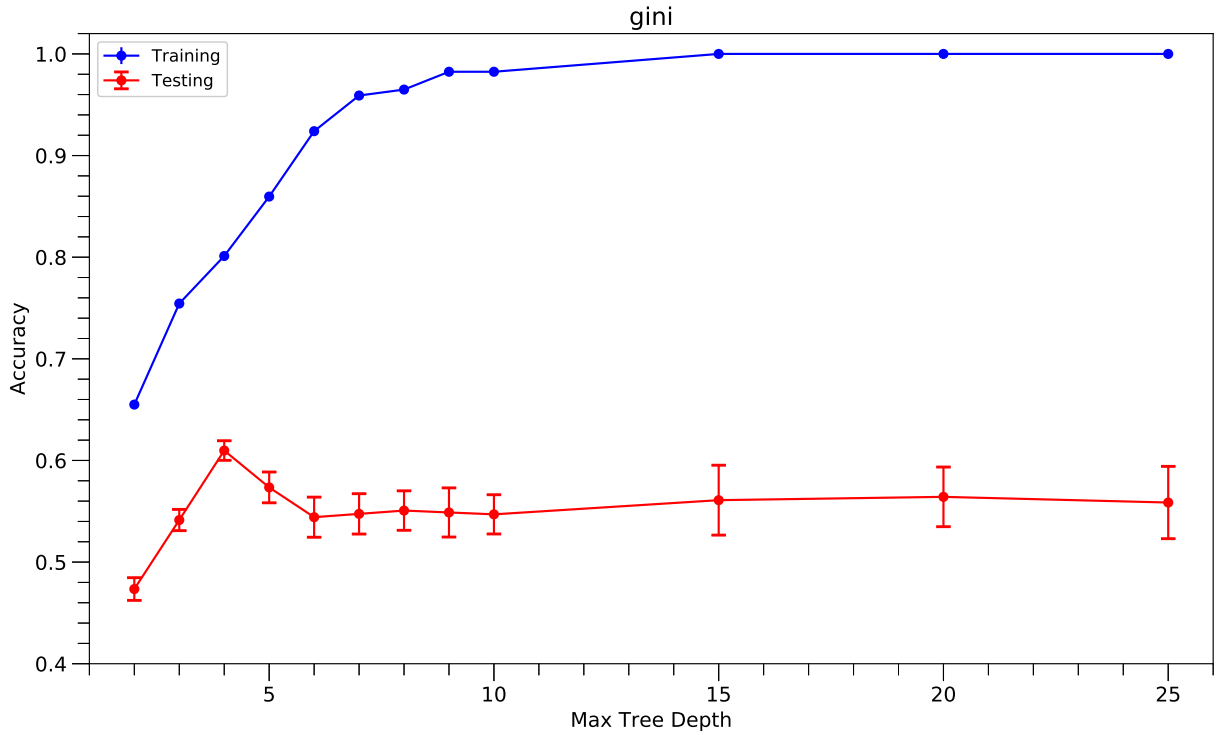

FIG. 2. Training and testing accuracy of decision trees using entropy impurity for different maximum tree depths.

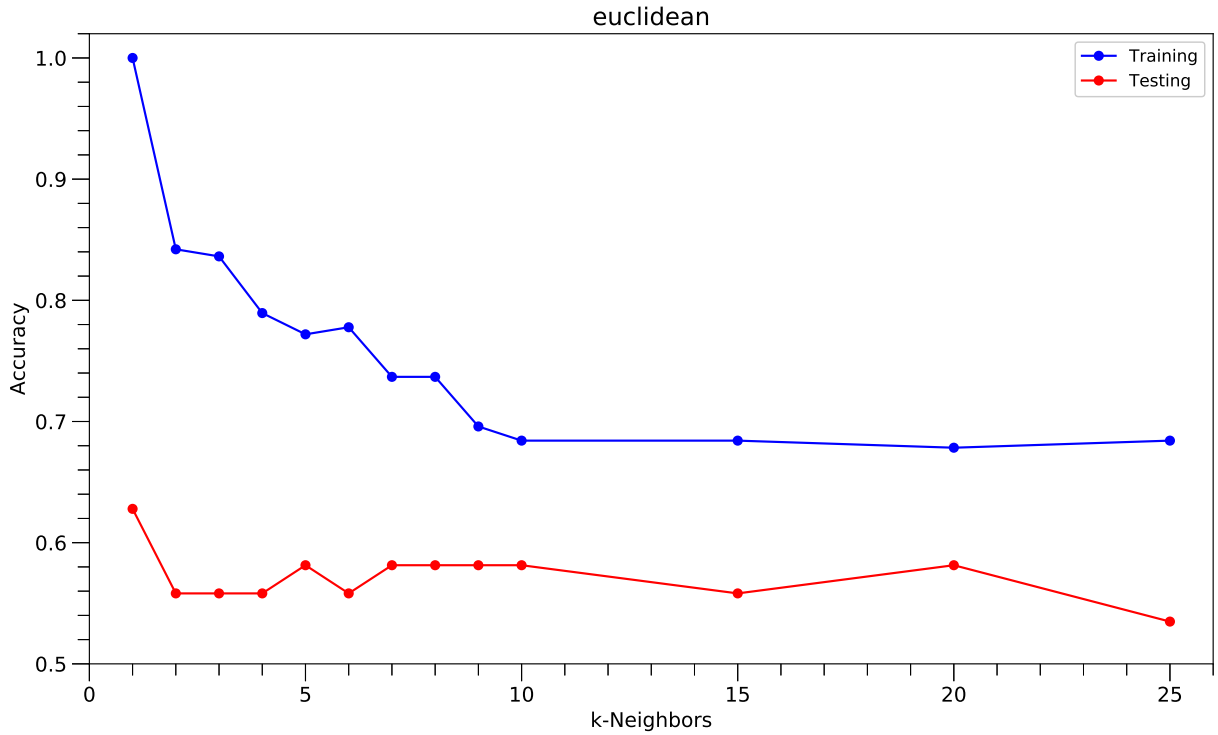

FIG. 3. Training and testing accuracy of K-Nearest Neighbors using the Euclidean metric and increasing neighbors.

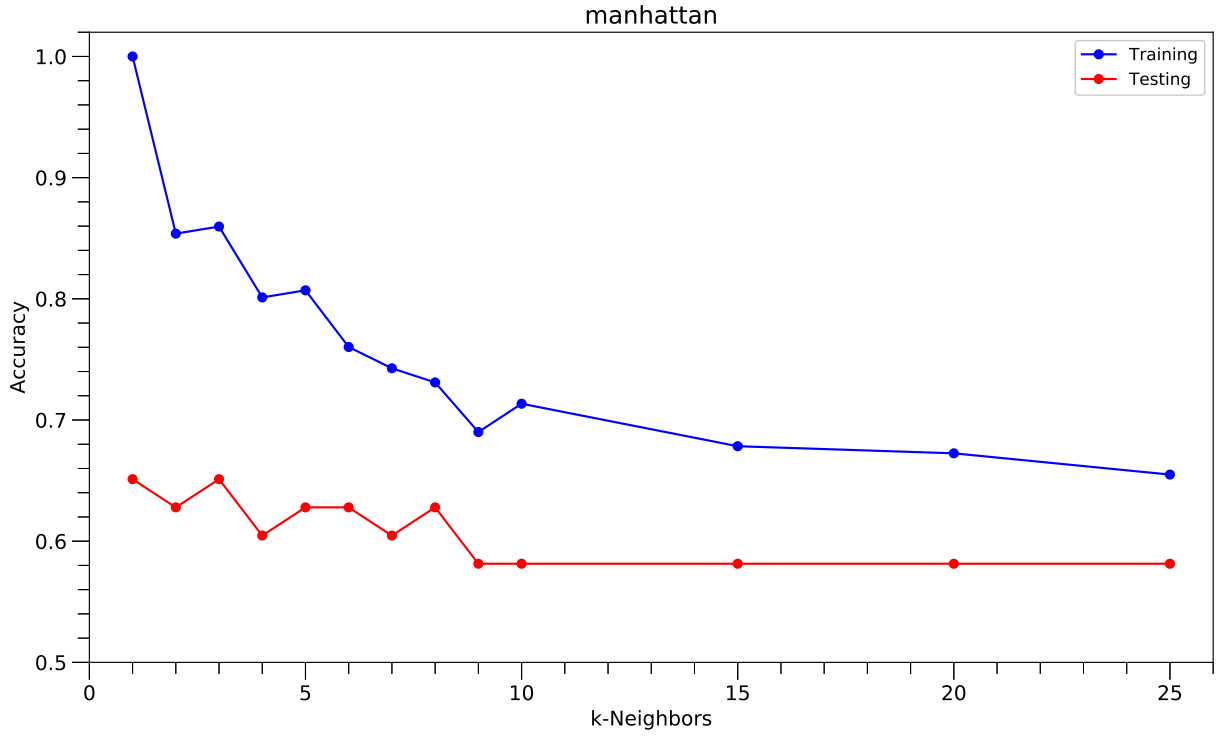

FIG. 4. Training and testing accuracy of K-Nearest Neighbors using the Manhattan metric and increasing neighbors.

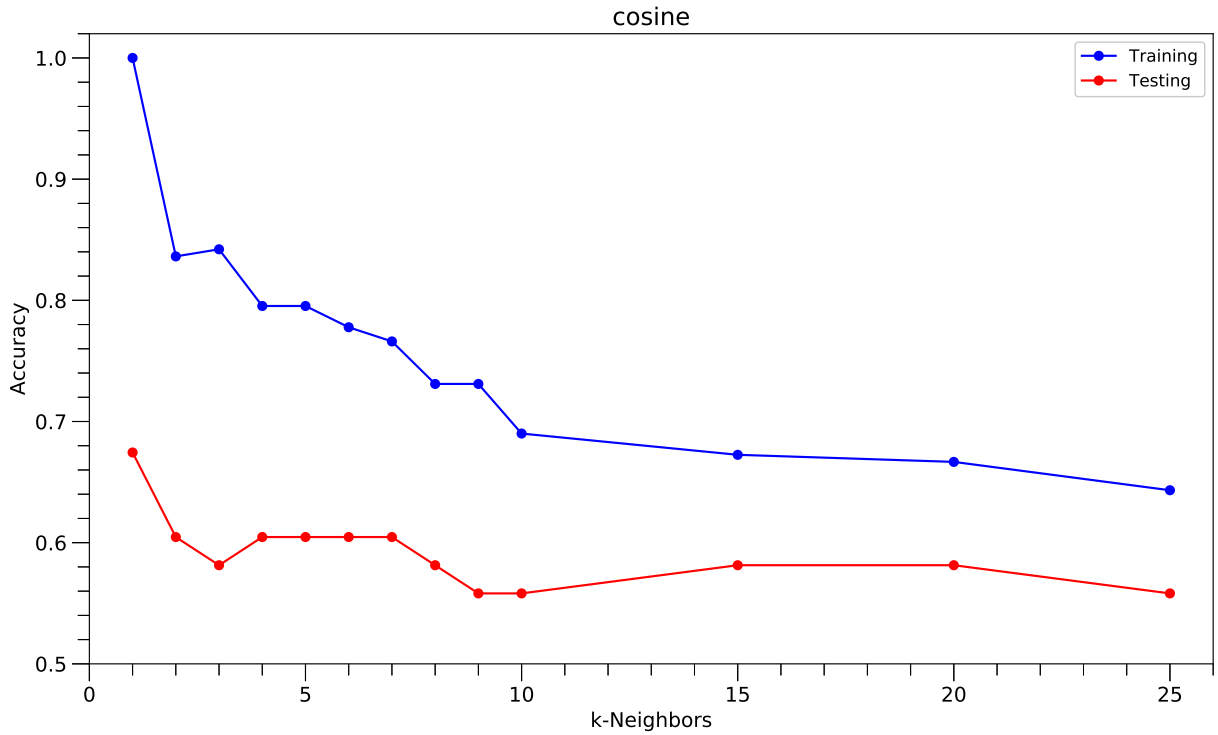

FIG. 5. Training and testing accuracy of K-Nearest Neighbors using the Cosine metric and increasing neighbors.