Team Project Midterm Report

10.25.2019

Cyrus Baker, Lucas Herrmann, Ivan Nieto
CS 488 - Data Mining

Data

https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29

The data we are using is provided by the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. It contains various data on breast cancer tumors including their cell size and shape, thickness and other attributes. The final attribute denotes whether or not the tumor was benign or malignant which is what we will be trying to predict.

Data Mining Task

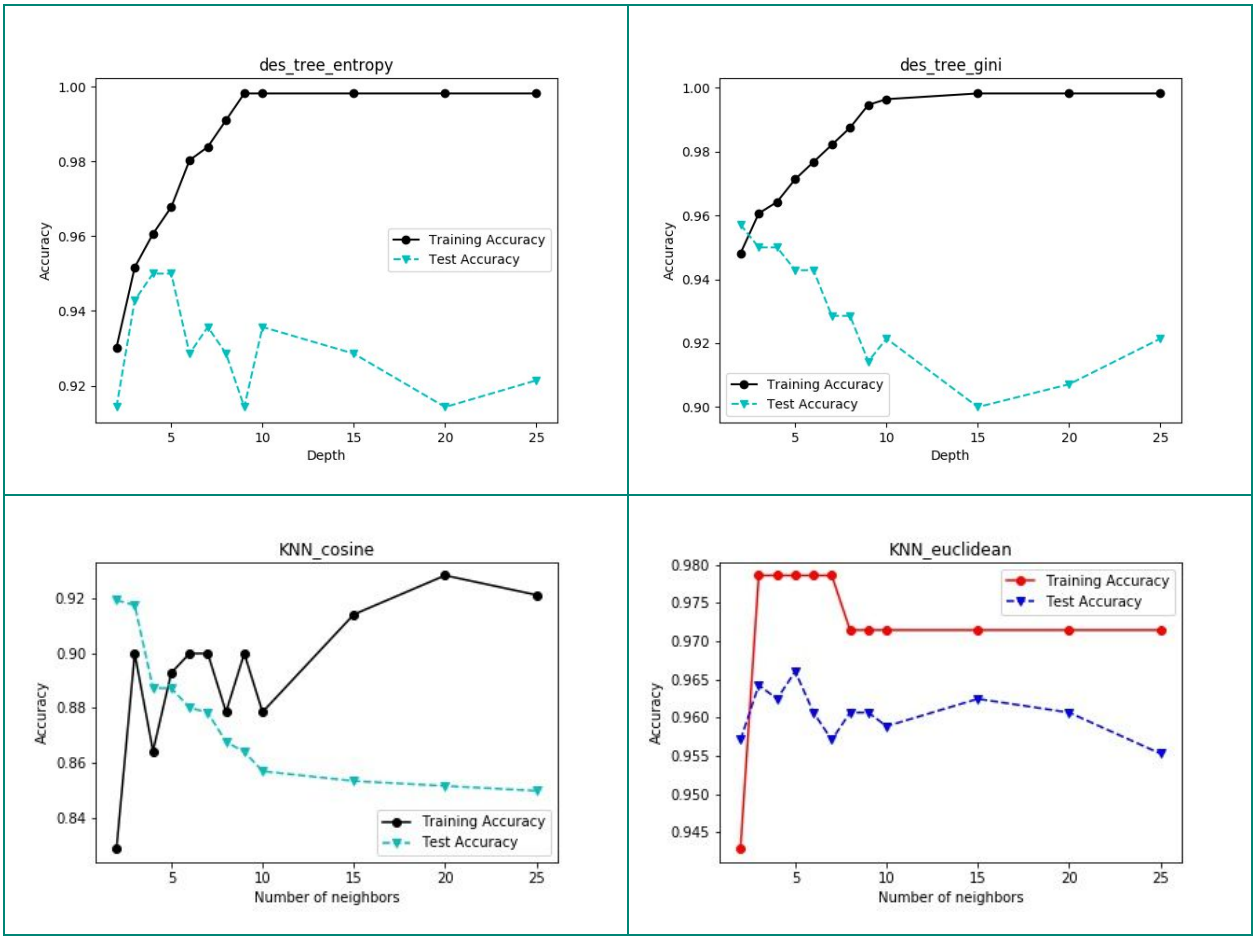
We are performing a classification tasks on our breast cancer data set. We use the first few attributes (clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitoses) to predict the classification of the tumor. To achieve this classification, we have chosen to train and test the data using a decision tree classifier and a KNN classifier.

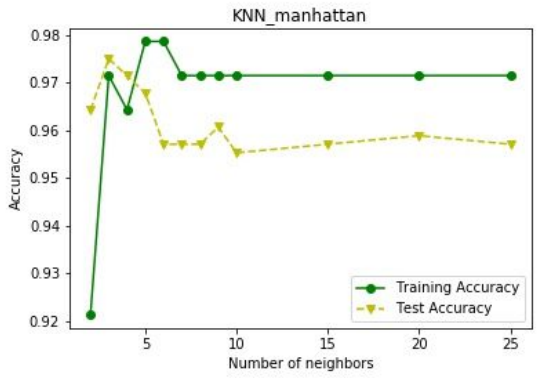
The data from the study has the following instances

```
Data columns (total 11 columns):
    Sample Code #                   699 non-null int64
    Clump Thickness                 699 non-null int64
    Uniformity of Cell Size         699 non-null int64
    Uniformity of Cell Shape        699 non-null int64
    Marginal Adhesion               699 non-null int64
    Single Epithelial Cell Size     699 non-null int64
    Bare Nuclei                     699 non-null object
    Bland Chromatin                 699 non-null int64
    Normal Nucleoli                 699 non-null int64
    Mitoses                         699 non-null int64
    Class - 2=Benign 4=Malignant    699 non-null int64
```

Progress

To date, we first cleaned the data by removing the bare nuclei attribute from the data set as well as the ID attribute. We then split the data into test and training data, and constructed both a KNN classifier and a decision tree classifier. To achieve this, we used the pandas and numpy python libraries to process the data, and the sklearn library to construct our classifiers much like we did in our second homework.





Difficulties/Challenges

One of our main challenges was with the data itself. At first we tried to run the data in the same way as we did in the homework, but the classifiers kept having errors with the data. In the end we found that one of the attributes in the data had the type 'object' rather than type 'int64' like the rest of the data. For the time being we have decided to run the classifiers without that attribute until the time that we can figure out how to change that object's type.

Schedule

From here on out, we must examine the classification methods and use our test data to determine the more accurate of the two methods using the data that we have so far collected. Once selected, we will begin preparation of our final report and presentation to report our findings.