Orysya Stus
4.13.2017

DSE 220 Homework 1

Question 1
Prior to preprocessing, there are 178 rows in the dataframe.
After dropping missing labels, there are 168 rows in the dataframe.
After dropping rows with more than 7 missing features (>7 missing features), there are 154 rows in the dataframe and 14 columns in the dataframe.

Question 2
After removing features >50% of missing values, there are 154 rows in the dataframe and 13 columns in the dataframe. The column/feature ['Ash'] was dropped.

Stds before filling nulls for the features with missing values
Magnesium 14.884800
Flavanoids 0.994727
dtype: float64

Stds after fillings nulls for the features with missing values
Magnesium 14.440377
Flavanoids 0.873573
dtype: float64

Question 3
Outliers are seen for alcohol if the values of alcohol < 10. There are 5 instances in total. These instances should be dropped.
After dropping alcohol related outliers, there are 149 rows of data.
After dropping proline related outliers, there are 148 rows of data.

Basis for outlier removal:
Furthermore, the outliers in this feature distribution are not as obvious. In order to determine, if there are any outliers we will keep only the ones that are within +3 to -3 standard deviations in the column.

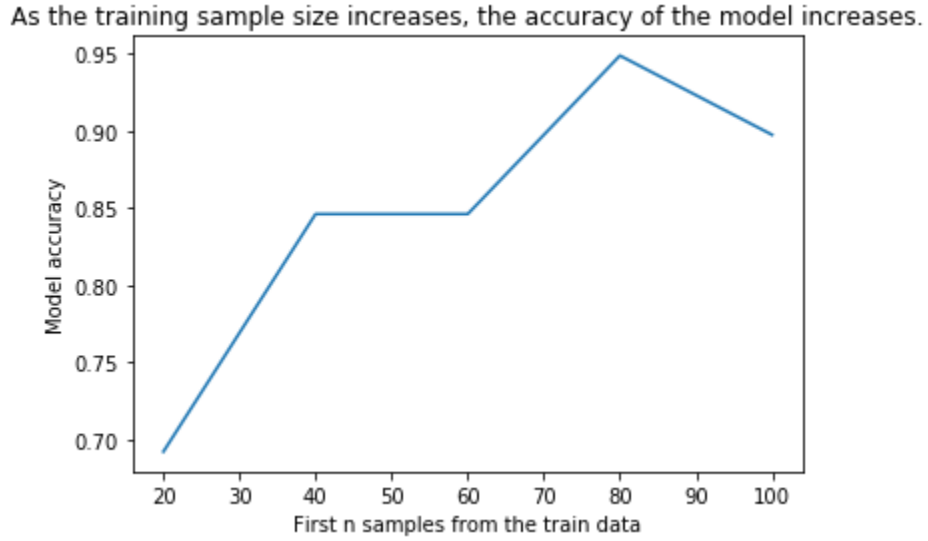
Question 4
For criterion = gini the validation accuracy = 0.974358974359
For criterion = entropy the validation accuracy = 0.948717948718
The best criterion is gini with a validation accuracy of 0.974358974359

The test accuracy using criterion = gini is 0.74358974359

Question 5
For min_samples_split = 2 the validation accuracy = 0.974358974359
For min_samples_split = 5 the validation accuracy = 0.974358974359
For min_samples_split = 10 the validation accuracy = 0.923076923077
For min_samples_split = 20 the validation accuracy = 0.948717948718
The best min_samples_split is 2 with a validation accuracy of 0.974358974359

The test accuracy using min_samples_split = 2 is 0.74358974359

Question 6
Using the first 20 samples, the validation accuracy is 0.641025641026
Using the first 40 samples, the validation accuracy is 0.820512820513
Using the first 60 samples, the validation accuracy is 0.871794871795
Using the first 80 samples, the validation accuracy is 0.948717948718
Using the first 100 samples, the validation accuracy is 0.948717948718



Question 7
For Euclidean distance and k = 3, the test accuracy is 0.871794871795

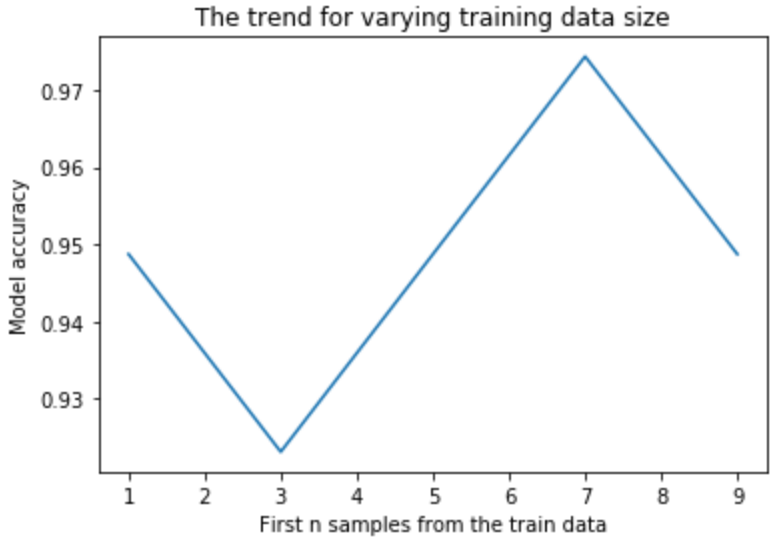
Question 8
For metric = euclidean the validation accuracy = 0.923076923077
For metric = manhattan the validation accuracy = 0.948717948718
For metric = chebyshev the validation accuracy = 0.923076923077
The best metric is manhattan with a validation accuracy of 0.948717948718

The test accuracy using metric = manhattan is 0.974358974359

Question 9
For k = 1 the validation accuracy = 0.948717948718
For k = 3 the validation accuracy = 0.923076923077
For k = 5 the validation accuracy = 0.948717948718
For k = 7 the validation accuracy = 0.974358974359
For k = 9 the validation accuracy = 0.948717948718
The best k is 7 with a validation accuracy of 0.974358974359

The test accuracy using k = 7 is 0.923076923077

Orysya Stus
4.13.2017



Question 10

Using the first 20 samples, the validation accuracy is 0.948717948718
Using the first 40 samples, the validation accuracy is 1.0
Using the first 60 samples, the validation accuracy is 1.0
Using the first 80 samples, the validation accuracy is 1.0
Using the first 100 samples, the validation accuracy is 0.923076923077