BBM469: Data Intensive Applications Lab. **(Due: 26/04/2020)**

Submission Assignment #2

Student Name: Okan ALAN, *Student No:* 21526638

1 Introduction

In this report, I discuss my approaches to given tasks and data structures/algorithms that I used. This report containing the used model and analysis of the result.

In this assignment, our goal was to learn clustering and classification. During this time, they expected us to learn algorithms in those topics.

2 Dataset

The dataset is too understandable. Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. A few of the images can be found at [Web Link]

2.1 Attribute Information

- ID number
- Diagnosis (M = malignant, B = benign)

This data set contains 3 different variations of the below attributes. These are mean, standard error, and worst/largest mean values. I thought I should use select one of three to avoid overfitting.

- radius : mean of distances from center to points on the perimeter
- texture : standard deviation of gray-scale values
- perimeter : size of the core tumor
- area
- smoothness : local variation in radius lengths
- compactness : perimeter*perimeter / area - 1.0
- concavity : severity of concave portions of the contour
- concave points : number of concave portions of the contour
- symmetry
- fractal dimension : ”coastline approximation” - 1

There are 569 records and it consists of 357 benign, 212 malignant records.

2.2 Preprocessing Dataset

Firstly, when I read the dataset, I realized there 2 columns are no effect of results. One of them is the first column ID number because we don't interest in the ill person number. The second column is the last column created by the reading library because of the extra comma at the end of the line.

Secondly, I enumerated the string values with integer because we had been learned in the lesson. This conversation helps to computer to understand data more eligible. In this data, we have "diagnosis" columns which consist of M(malignant) and B(bening) letters. I changed M and B respectively 0 and 1.

Then, if you look outline of our dataset, you can see, it is divided into 3 parts such as mean, standard error, and worst. I chose the first one. I dropped the standard error and the worst/largest mean columns because these are correlated.

This section's last process is to generate a correlation map. These 2 figures are shows us how related selected 2 variables. According to them, I divided mean features into high and low correlated features. High correlated features; radius-mean,perimeter-mean,area-mean,compactness-mean,concavity-mean,concave points-mean. Others are low features. In lessons, I learned highly correlated variables caused to overfitting. Therefore I selected low correlated features with one of the high correlated features. I prefered to use "concave parameter_mean" because when you look the correlation map 2 intensive areas circled with red line. These are "radius-mean, perimeter-mean, area-mean" and "compactness-mean, concavity-mean, concave points-mean". I examined them, I realized that "concave parameter_mean" is the most related among themselves. I explained which feature is how effect the result.

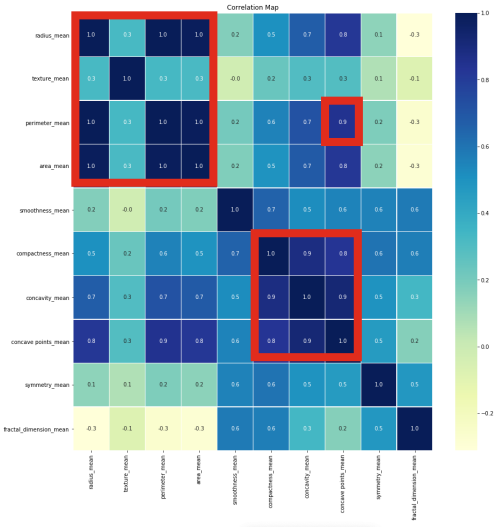

Figure 1: Correlation Map

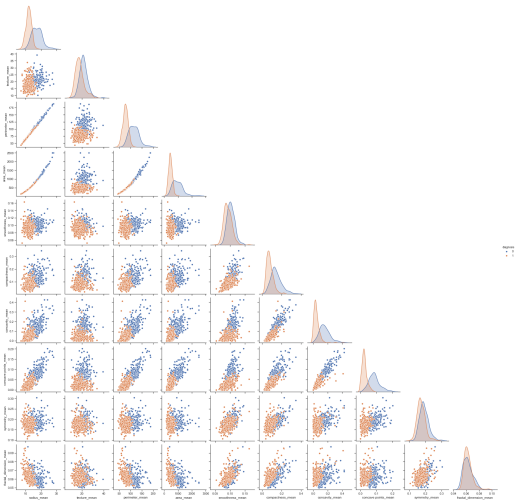

Figure 2: Correlation via Pair Plotting

3 Clustering

Clustering data means that collecting the similar data records into one cluster. Clustering is a unsupervised learning model. Our dataset is already labeled data therefore I am not totaly sure my clustering results. I am not sure how okay it is to apply clustering to this data. My clustering method was KMeans.

The below images contain original data cluster results. The left side is normal clusters, the right side is Kmeans applied clusters. Normal cluster is according to the diagnosis column but Kmean applied clusters were created by the KMeans algorithm. I can say that there is no huge difference between KMeans clustering and the original cluster. There are only a few records' that are exchanged cluster. Below figures are belong to original dataset but there is same result for original and normalized dataset. You can see the result in detail in the codes.

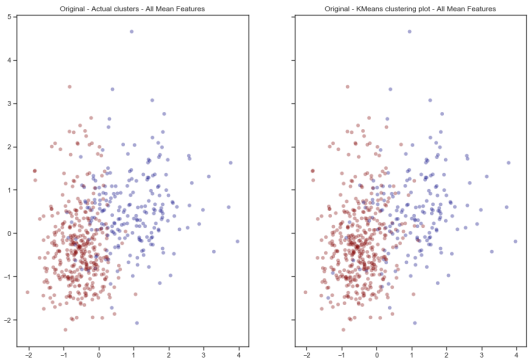

Figure 3: Clustering with all mean features

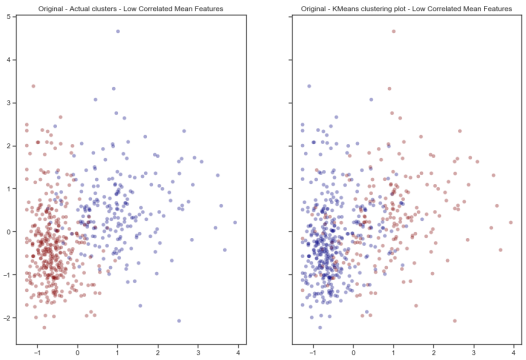

Figure 4: Clustering with selected features

4 Splitting Data

I splited the dataset into 20% test and 80% train.

5 Normalization

Actually, you wanted us to clustering the normalized at step 6. But I read a few articles about normalization, I normalized original dataset after splitting it into train and test. The results are almost the same original dataset. There are 6 options; test/train cluster with all features/high correlated/low correlated. Related figures are in code. Why we should normalization after splitting? Because, if we normalize the dataset before splitting dataset, there would be a data leakage. We might lose some important data. Additionally, I normalized the train and test separately.

6 Classification

My classification method was Random Forest. Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction

6.1 Results

If I hadn't deleted the standard error and the worst/largest mean columns, my accuracy would be 1-3% less than the results below. I am putting my results to below.

	Original	Normalized
All Mean Features	0.9217391304347826	0.9130434782608695
Selected Features	0.9652173913043478	0.9478260869565217

Table 1: Accuracy Table

The best score belongs to the original data with selected features. When I normalized the original dataset, my result was decreased. Probably If I used the SVm instead of Random Forest, the normalized test set's would take higher score.

7 Effect of Selected Features

I used shap library to understand the effect of features in my model. Each position on the x-axis is an instance of the data. Red SHAP values increase the prediction, blue values decrease it. Below figure belongs to random forest model that is trained with original data including all features.

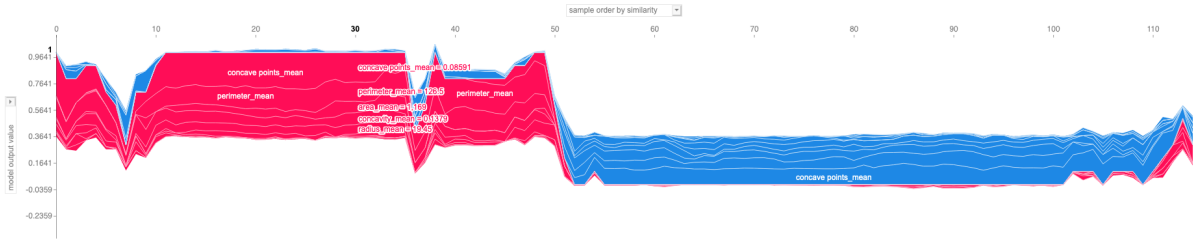

Figure 5: Effect of features on the model

According the above figure the "concave points-mean,perimeter-mean" are highly effect the result in our model.