Semester Project for DATA2

Kıvanç Gördü

Estimated time: 5 hours; Time spent: 15 hours

1. Aim and dataset

The aim of this project is to examine the relationship between features and entities in general, including the impact of *diagnosis* variable to the rest. Diagnosis column consist of M and B, respectively, the initials of Malignant and Benign. By studying diagnosis data with the other features, we will be clarifying in which cases the measurement results may remark people's health condition. The research will be conducted in frame of Data Expletory Analysis and Descriptive Statistics.

The models used in this article includes boxplot, scatterplot, correlation chart, k-means and logistic regression. Applied libraries are tidyverse, cluster, Hmisc, plotpy, ggfortify, factoextra, NbClust, ggpubr, dplyr, PerformanceAnalytics, ggplot2.

Diagnosis data stands for categorical data with binary values. The rest of the columns are quantitative with continues values. It includes "X" column where there is only *NaN* registers and it there is no regarding explanation have been found. Therefore, his column will be excluded in further steps with default *slice* method of R. I will also change the index arrangement of all the dataset to provide a better visualization, note that index of patients is not included in the table above for the purposes of better visualization However, it is going to be included when we will be interpreting the entities. The table has 569 rows and 32 columns. The dataset is constructed table in format of CSV file.

Table 1

Breast Cancer Dataset

diagnosis	M	M	M	⋮	M	M	M	B
radius_mean	17.99	20.57	19.69		21.56	20.13	16.6	20.6
texture_mean	10.38	17.77	21.25		22.39	28.25	28.08	29.33
perimeter_mean	122.8	132.9	130		142	131.2	108.3	140.1
area_mean	1001	1326	1203		1479	1261	858.1	1265
smoothness_mean	0.118	0.085	0.11		0.111	0.098	0.085	0.118
compactness_mean	0.278	0.079	0.16		0.116	0.103	0.102	0.277
...								
compactness_worst	0.666	0.187	0.424		0.192	0.309	0.868	0.064
concavity_worst	0.712	0.242	0.45		0.322	0.34	0.939	0
concave.points_worst	0.265	0.186	0.243	⋮	0.163	0.142	0.265	0
symmetry_worst	0.46	0.275	0.361		0.257	0.222	0.409	0.287
fractal_dimension_worst	0.119	0.089	0.088		0.066	0.078	0.124	0.07
X	NA	NA	NA		NA	NA	NA	NA

2. Descriptive statistics of the dataset

For the initial step, we will be looking to descriptive statistics summary. This method helps us to describe and understand the features of a specific dataset by giving short summaries about the measures of the data.

Table 2

Statistical Summary of Columns

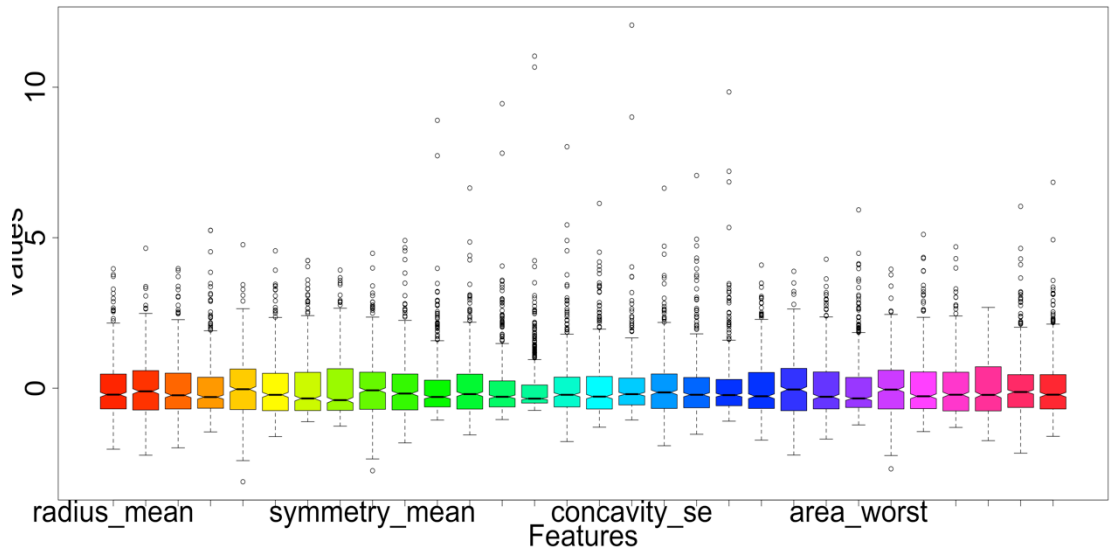
	radius_mean	texture_mean	perimeter_mean	area_mean	⋮	concavity_worst	concave.points_worst	symmetry_worst	fractal_dimension_worst
Min	6.981	9.71	43.79	143.5		0	0	0.16	0.06
1st Quarter	11.7	16.17	75.17	420.3		0.11	0.06	0.25	0.07
Median	13.37	18.84	86.24	551.1		0.23	0.1	0.28	0.08
Mean	14.127	19.29	91.97	654.89		0.27	0.11	0.29	0.08
3rd Quarter	15.78	21.8	104.1	782.7		0.38	0.16	0.32	0.09
Max	28.11	39.28	188.5	2501		1.25	0.29	0.66	0.21

Subsequently, we apply the boxplot graph of our columns below. Boxplot is a rough view of all the columns and their correlation. Upon our methods settings, we will scale them before using them. The reason behind is that the range of the features highly differ, and we want to put them into one frame with scaling.

We also used color variable equivalent to rainbow and actived the notch feature. Besides, the graph set to be horizontal to ease the comparison of the features. Starting line and the ending line demonstrate the minimum and maximum points. In figure 1, we note that the shortest range belongs to *area_se* and the longest to *compactness_mean* columns. The tiny circles are the outliers that are remarkably out of the normal scale. We may realize that each column has plenty of them. One of the most is in the case of *area*_se. Besides, in the middle columns, we also realize that outliers very much differ from each other. Quartile means one fourth of the data. The whispers construct the 1st and 4th quartiles also the first lower and upper piece do so for the 2nd and 3rd quartiles of each variable. The cutting line indicates *Interquartile Range (IQR)* that is so-called *median*. Median is the border of %50 of entities. Notches are the curves of boxplots. Notches in the boxes show approximate 95% confidence intervals for the medians. That means, if the notches overlap each other, we can claim that it is less likely that their medians differ. *Radius_mean* and *perimeter_mean* share the similar median, max-min points and density, meaning that they might correlate. If we take this argument and search the definition of radius and diameter then, we find that radius is the half of diameter, so the argument is accurate. Afterward, we look at *symmetry_worst* and *fractal_dimension_worst*, their medians overlap, their density of box plot (including the shape and curve) match. We may deduce that they are almost proportionally identical except a slight difference of max-min points so in their range.

Figure 1

Boxplot of Each Column

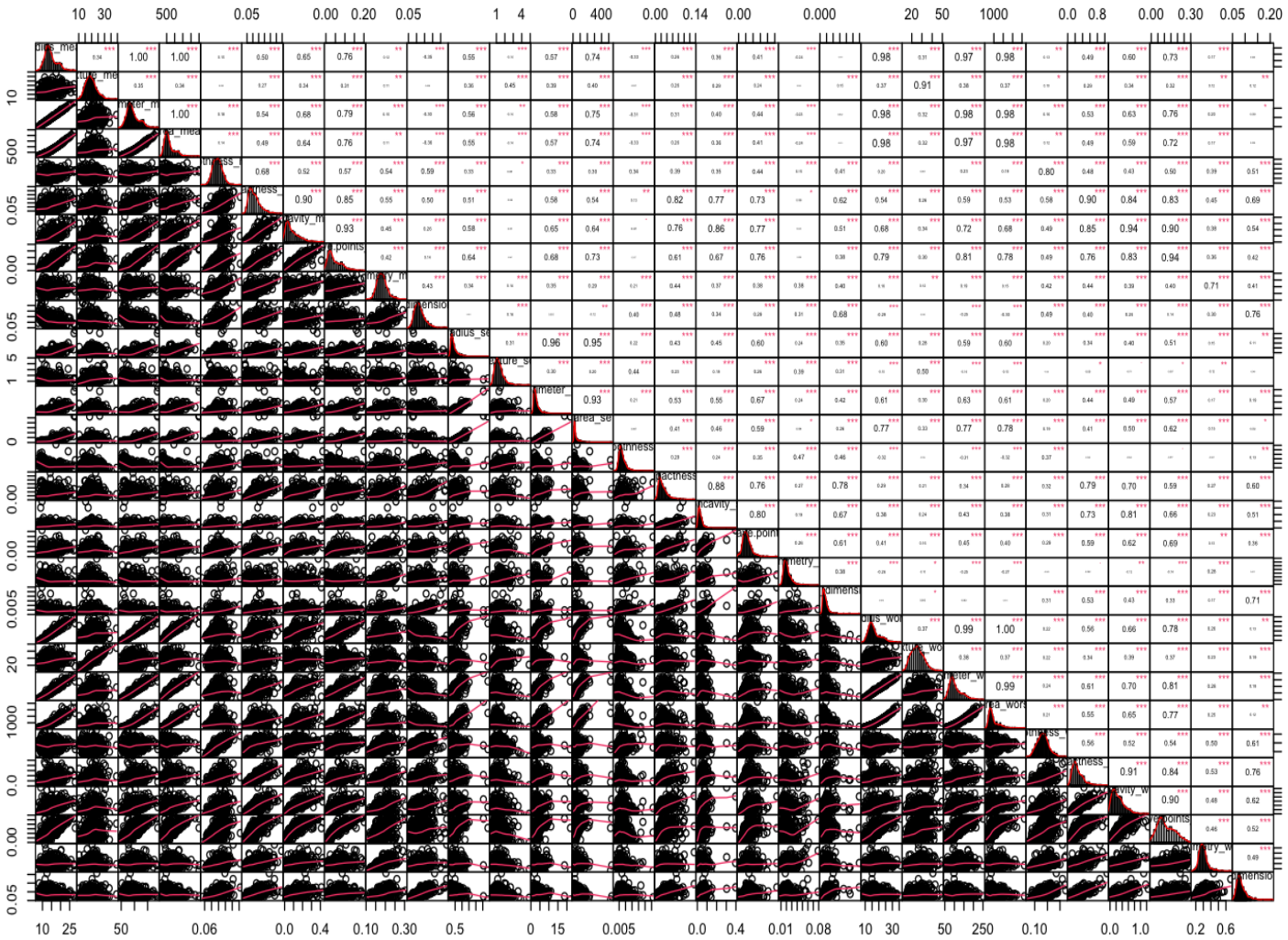


3. Mutual relationships of features

First and foremost, it is important to remember that correlation does not mean there must be a causation. There might be, but we would never know. It merely means that there is a factor that effect both variables more or less equally. There are two typical correlations methods, the first is *Pearson's correlation* and the latter is called *Spearman's* correlation. Person's correlation is used to measure linear correlation ranges from -1 to +1. 0 means no correlation. Negative and positive correlations are expressed with -1 and +1. The higher correlation means the closer number to +1 or -1. Positive correlation vectors are extended upside, negative correlation vectors face bottom. On the other hand, Spearman's correlation measure both the linearity and monotony between two variables. If the Pearson's coefficient is a perfect -1 or +1, Spearman's correlation coefficient will be the same perfect value unless there are repeating data values. The same extension logic applies asymmetrically to monotony vectors but not exactly as it does in Person's, because vectors are drawn by two factors. Thus, Person's correlation is more reliable than Spearman's correlation in terms of linearity. The other difference is that Spearman's correlation functions with rank-ordered register whereas Person's correlation works with raw values. There is not a complete true method for dataset. We looked at both correlations and decided to proceed with Spearman's correlation because we want to take monotony into consideration, so it is more sensitive.

Figure 2

Correlation of Each Column



Although, the graph is very small and detailed, we see each column as a separate box. For each cross of the columns, we have a correlation graph (Spearman's in our case), correlation value and the p-value. P-value is index of significance that helps us to determine whether the correlation can be. By the convention, if the p-value is %5 or lower, there is only 5% chance that results from the sample occurred due to chance and it is significant possible so there *may* be a correlation. We say that the If p-value is higher than %5, correlation is not statistically possible. In the graph, we encounter the stars on the top of numbers. When the p-value is significant, we have more than at least one start. We can almost certainly assume that three-star correlations are very much close to absolute correlation because they remark the range of p-value between 0 - 0.001.

Accordingly, we have *radius_mean, perimeter_mean, area_man, compactness_mean, concavity_mean, concave.points_mean, fractal_dimension_mean, radius_se, perimeter_se, perimeter_se, area_se, compactness_se, concavity_se, concave.points_se, concavity_worst and concave.points_worst* columns correlated with each other. What is interesting is that following features do not correlate with any of columns: *smoothness_se, texture_worst*. Finally, *symmetry_se* with *symmetry_mean, fractal_dimension_se* with *fractal_dimension_worst* correlate only with each other.

4. An overview to the object relationships

Now, we are aiming to search the objects. The amount of object is relatively high. I tested HC with dendrograms and noticed that does not help in our case. Neither I liked the results of PCA. Cumulative Proportion of first loading was equal to 44.27, followed by 18,97% and 9%. The sum of PC1 and PC2 and even the third dimensions' was far away being complete to hundred that might be resulted with a big loss of information. Thus, we will be using *k-means*. k-means is an unsupervised clustering method which uses the Mean linkage method, meaning that we are concerned about distance between possible clusters' centers. We will be also use Euclidean as it is also the default parameter.

Figure 3

K-mean Graph of the Dataset

By that, we also emphasize the outliers and the distribution. We prefer k-means because other methods failed to process my device

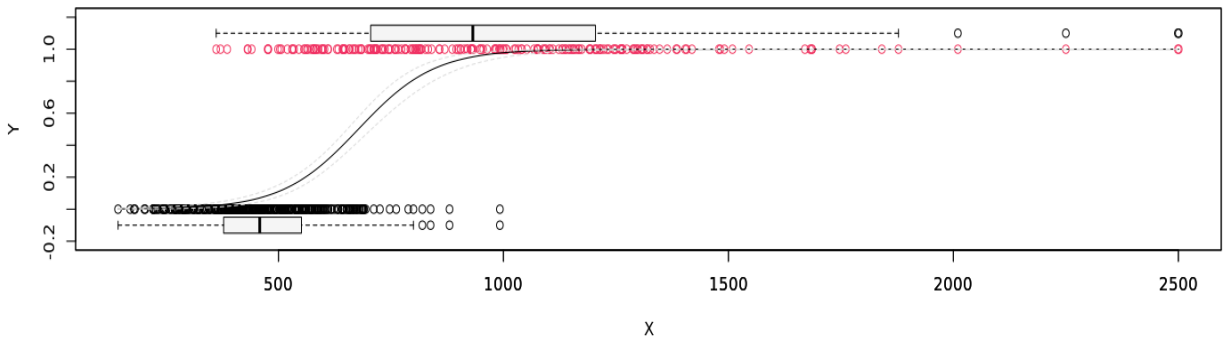
because they are extremely expensive. k-means is usually more than enough for illustrations. They are hard to implement to large datasets. However, we are aware that this method's drawback is that it is highly sensitive to the outliers and noise. The other drawback is that one has to know the optimal *k* value. Therefore, I tested our dataset with various and some of them include Huber index with D index. In conclusion, 8 tests proposed two, 6 tests offer three for the number of clusters. I choose to illustrate with two clusters. Subsequently, when we divided them to the clusters and look for the whole dataset including *diagnosis*, we can easily see an increase of the so-called correlated columns that were placed in the beginning.

5. Analysis of relationships between features and clusters

Later, noting the binary nature of diagnosis, we compared it with compactness_mean variables and used Logistic Regression in order to accomplish. p-value is almost 0.

Figure 4

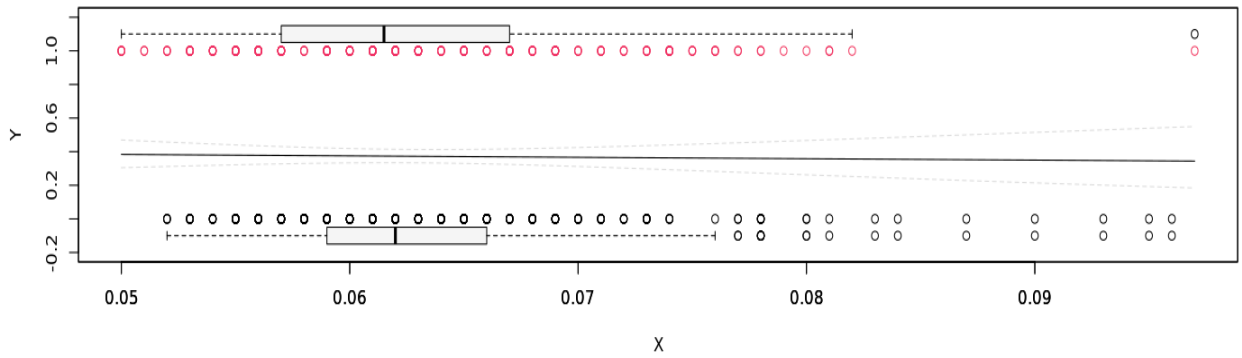
Logistic Regression Graph of compactness_mean and diagnosis variables



For the second try, I compared the diagnosis dataset with radius_se and the results were surprizing. Despite that radius_se and compactness mean were correlating, so I am expected to see the same result. However, it was different for this regression.

Figure 5

Logistic Regression Graph of radius_se and diagnosis variables



6. Conclusion

We compared the dataset and explored that there were two clusters. Nevertheless, many columns were indeed correlating, our application of logistic regression has proved that regression and the correlation were two different concepts. We know that many columns correlate with each other. Compactness_mean is easily readable with its regression based on diagnosis.