Prospeção e Análise de Dados

Breast Cancer Wisconsin (Diagnostic) Data Set

Prospeção e Análise de Dados
2022/2023

João Padrão 58288
Riccardo Galarducci 66819

Contents

Abstract

Breast cancer dataset is a widely used dataset in machine learning and data science that contains information 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. The aim of this report is to analyze the result we have obtained while performing four main tasks: *Exploratory Data Analysis*, *Regression Analysis*, *Principal Component Analysis* and *Clustering*.

1 Introduction

Breast cancer is a type of cancer that develops in the cells of the breast tissue. It is the most common type of cancer among women, but it can also affect men. Breast cancer occurs when abnormal cells in the breast begin to grow out of control and form a lump or mass.

There are different types of breast cancer, including ductal carcinoma, lobular carcinoma, inflammatory breast cancer, and others. Each type of breast cancer can have different symptoms, patterns of spread, and treatment options. Breast cancer can be **benign** or **malignant**.

Benign breast tumors are non-cancerous growths that do not spread to other parts of the body. They may cause discomfort or pain, but they are not life-threatening. Benign tumors are usually removed by surgery, and they generally do not recur.

On the other hand, malignant breast tumors are cancerous growths that can invade nearby tissues and organs, and can also spread to other parts of the body through the lymphatic system or bloodstream. Malignant tumors can be life-threatening if not treated early and aggressively. The most common types of malignant breast tumors are invasive ductal carcinoma and invasive lobular carcinoma.

The diagnosis of benign or malignant breast cancer is made by examining the tissue sample obtained through a biopsy. Benign tumors have a well-defined edge, and the cells look normal and are not dividing rapidly. Malignant tumors, however, have an irregular shape, and the cells look abnormal and are dividing rapidly.

Treatment options for breast cancer may include surgery, radiation therapy, chemotherapy, hormone therapy, targeted therapy, or a combination of these approaches. The type of treatment used depends on the stage of the cancer, the type of breast cancer, and other factors such as age and overall health.

For this assignment we have chosen the Breast Cancer Wisconsin (Diagnostic) Data Set because this is an important public health issue that requires attention and research. We want to explore the relation between features and this type of cancer to improve our understanding about this topic.

2 About the Data set

This data set is composed by 546 breast cancer cases with 30 features each and the label of stage of breast cancer Malignant (M) and Benign (B). Columns:

1. ID number of the sample;
2. Diagnosis (M = malignant, B = benign), target variable;
3. radius (mean of distances from center to points on the perimeter) ;
4. texture (standard deviation of gray-scale values);
5. perimeter;
6. area;
7. smoothness (local variation in radius lengths);
8. compactness ($\frac{perimeter^2}{area} - 1.0$);
9. concavity (severity of concave portions of the contour);
10. concave points (number of concave portions of the contour);
11. symmetry;
12. fractal dimension ("coastline approximation" - 1);

The mean, standard error and "worst" (mean of the three largest values) of these features were computed for each of the images, resulting in 30 features. All feature values are recorded with four significant digits.

3 Data Preparation

Before analysing our data, we have to clean the columns that are irrelevant for our analysis. The 32^{th} column is empty so we remove it from the data set. Then we remove the "ID" column because does not provide any meaningful information for the tasks we have to perform.

The next step is to encode the categorical column *diagnosis*, which is the target attribute, into numbers, so we named it "diagnosis_M". We have converted the Malignant (M) diagnosis into 1 and the Benign (B) into 0.

4 Exploratory Data Analysis

We perform exploratory data analysis to gain deeper insight in *Breast cancer data set*

First we have computed the frequency of cancer in this data set. There are 357 Benign and 212 Malignant, we can conclude that there are a higher number of benign stage of cancer which can be cured.

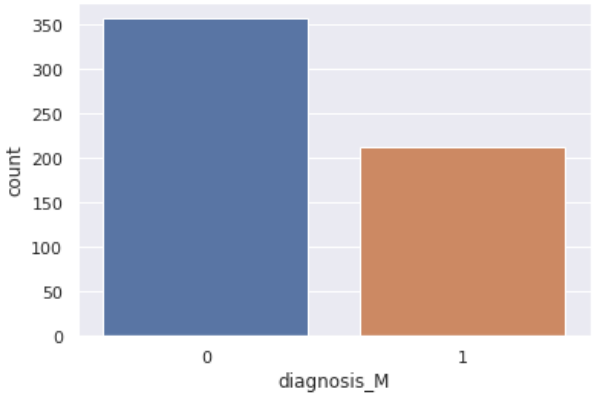


Figure 1: Cancer Frequency

To have a deeper understanding of the dataset, we have to look at the statistics of each column of our dataset.

Table 1: Dataset Mean Statistics

index	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	fractal_dimension_mean
count	569.0000	569.0000	569.0000	569.0000	569.0000	569.0000	569.0000	569.0000	569.0000	569.0000
mean	14.1273	19.2896	91.9690	654.8891	0.0964	0.1043	0.0888	0.0489	0.1812	0.0628
std	3.5240	4.3010	24.2990	351.9141	0.0141	0.0528	0.0797	0.0388	0.0274	0.0071
min	6.9810	9.7100	43.7900	143.5000	0.0526	0.0194	0.0000	0.0000	0.1060	0.0500
max	28.1100	39.2800	188.5000	2501.0000	0.1634	0.3454	0.4268	0.2012	0.3040	0.0974

Table 2: Dataset Standard Error Statistics

index	radius_se	texture_se	perimeter_se	area_se	smoothness_se	compactness_se	concavity_se	concave points_se	symmetry_se	fractal_dimension_se
count	569.0	569.0	569.0	569.0	569.0	569.0	569.0	569.0	569.0	569.0
mean	0.40517205623901575	1.2168534270650264	2.8660592267135327	40.337079086116	0.007040978910369069	0.025478138840070295	0.03189371634446397	0.011796137082601054	0.020542298769771 53	0.003794903864323374
std	0.2773127329861039	0.5516483926172023	2.0218545540421076	45.491005516131 81	0.003002517943839 0656	0.017908179325677388	0.030186060329884 41	0.006170285174046869	0.008296371528798399	0.002646070967089195
min	0.1115	0.3602	0.757	6.802	0.001713	0.002252	0.0	0.0	0.007882	0.0008948
max	2.873	4.885	21.98	542.2	0.03113	0.1354	0.396	0.05279	0.07895	0.02984

Table 3: Dataset Worst Statistics

index	radius_worst	texture_worst	perimeter_worst	area_worst	smoothness_worst	compactness_worst	concavity_worst	concave points_worst	symmetry_worst	fractal_dimension_worst
count	569.0	569.0	569.0	569.0	569.0	569.0	569.0	569.0	569.0	569.0
mean	16.269189806678387	25.677223198594024	107.26121265377857	880.5831282952548	0.13236859402460457	0.25426504393673116	0.27218848330404216	0.11460622319859401	0.2900755711775044	0.08394581722319 86
std	4.833241580469323	6.146257623038319	33.602542269036356	569.356992669949	0.022832429404835465	0.157336488913742	0.208624280608 1323	0.065732341195942 07	0.061867467537518685	0.018061267348893986
min	7.93	12.02	50.41	185.2	0.07117	0.02729	0.0	0.0	0.1565	0.05504
max	36.04	49.54	251.2	4254.0	0.2226	1.058	1.252	0.291	0.6638	0.2075

From these tables we can see that the mean value of the columns has high and small values which can lead us to think that we need to normalize or standardize our data set, this is due to the values not being in the same units. By doing this we can improve our model performance, reduce computational complexity, increase interpret ability and improve its robustness thus help handling outliers.

Let's keep analysing our data set by using violin plot to have a better view of how our data set is distributed across different groups and categories. By using this plot we can capture the shape of the distribution, highlight outliers and see where the data is more dense. We are now going to look into how the data is spread using Figure 2.

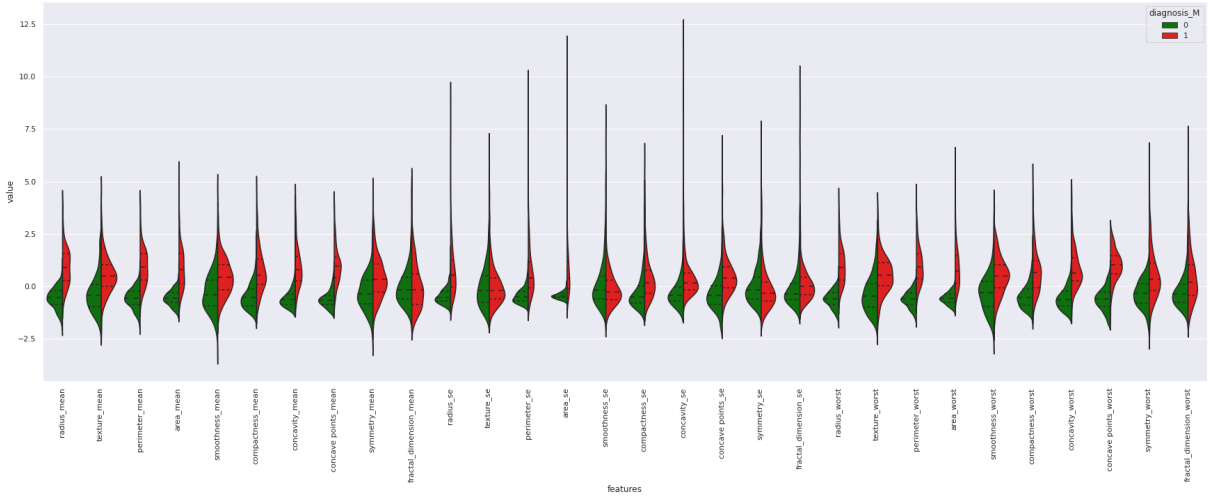


Figure 2: Violin Plot of Features

We want to visualize the scatter matrix to find the existing pairwise relationships in the data set, highlight nonlinear relationships and provide a comprehensive view of the data, thus helping detecting patterns, trends and potential outliers.

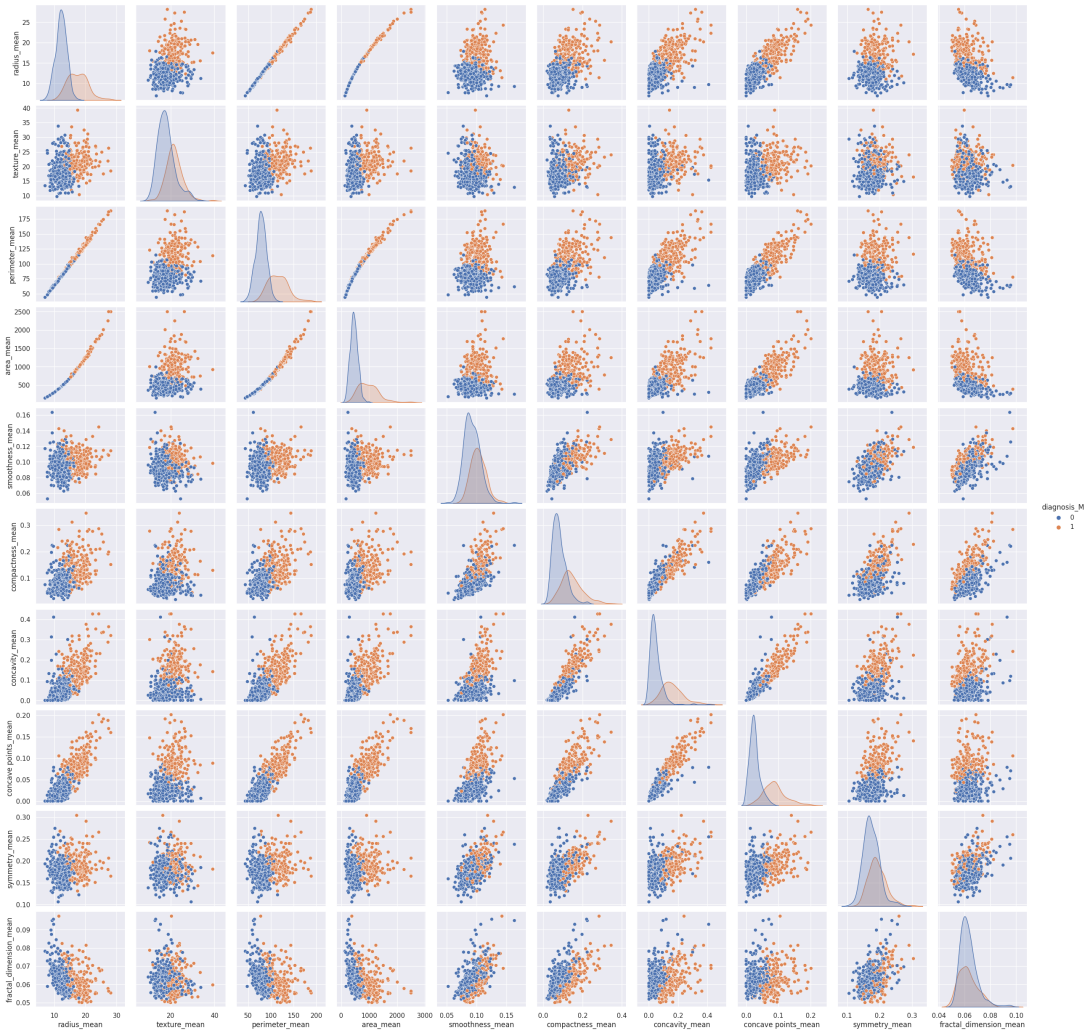


Figure 3: Scatter Matrix of Mean Features

To have a clear understanding of how features correlate, we will analyse the Correlation Matrix as well. This can help us quickly identify strong and weak correlations between features, they can be positively correlated or negatively or close to zero if not correlated at all. With this matrix we will be able to measure complex relationships, identify outliers and errors and choose features that correlate thus guiding our feature analysis.

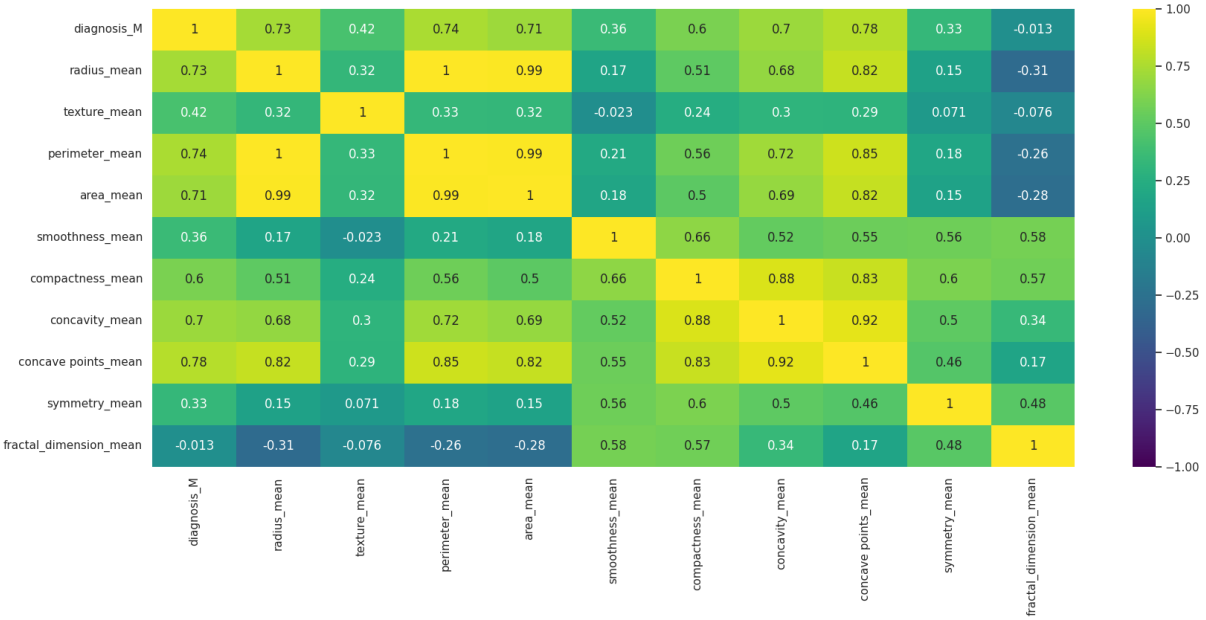


Figure 4: Correlation Matrix of Mean Features

As we can see in figure 3 and figure 4, there are several features that are highly correlated with each other. For instance, the radius mean and the perimeter mean. However, this is intuitive since these measurements are all related to the size of the breast. Moreover, there is a positive correlation between compactness mean and concavity mean, which points out that these might be measuring similar aspects of the shape of the breast. We can also notice that there are outliers in the data set, which are data points that are far away from the majority of the data. These potential outliers could be due to measurement errors, data entry errors, anomalous or rare events, etc.

5 Regression Analysis

Based on our analysis of the correlation of the features, we have selected two features with more or less "linear-like" scatter-plot, *Smoothness Mean* and *Fractal Dimension Mean*. These have a positive correlation of 0,58. The relationship between these two features can provide us valuable insights about our data.

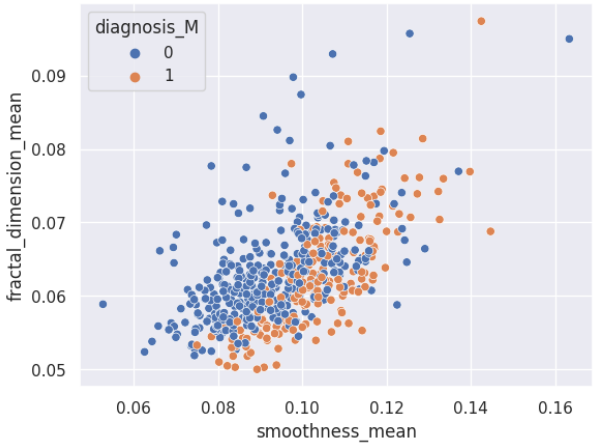


Figure 5: Scatter Matrix of Smoothness Mean and Fractal Dimension Mean

This scatter plot in figure 5 confirms the correlation between these two features, there is a general trend of increasing fractal dimension with the increase of the smoothness. However, there is also a lot of variation in the distribution of the data, with many data points agglomerated in the lower left corner of the plot with some point spread on the rest of the plot. These isolated data points significantly far away from the majority are potential outliers in the data, which might indicate measurement errors or anomalies.

Overall, the scatter plot of *smoothness mean* and *fractal dimension mean* may be helpful understanding the distribution of the data and identifying outliers.

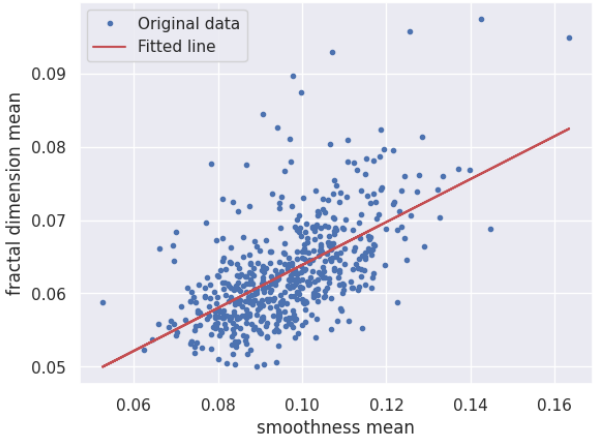


Figure 6: Linear Regression of Smoothness Mean and Fractal Dimension Mean

A linear regression model such as the one in the figure 6 can be used to visualize the relationship between features in more detail. In this linear regression, the ERE (estimated regression equation) of *smoothness mean* on *fractal dimension mean* have a slope of approximately 0.294 and intercept of the line that best fits the data is approximately 0.035. This means that for every unit increase in *smoothness mean*, *fractal dimension mean* increases by 0.294 and the intercept of approximately 0.035 represents the expected value of *fractal dimension mean* when *smoothness mean* is equal to zero.

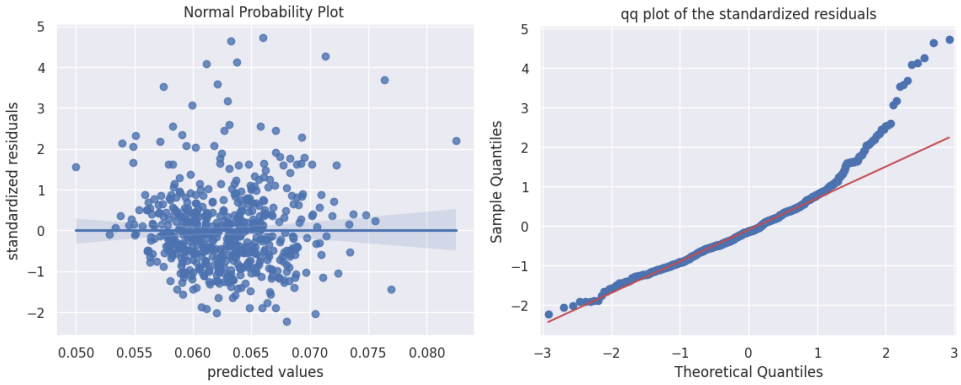


Figure 7: Normal Probability Plot of the Standardized Residuals

By analysing the standard residuals and the quantile-quantile (QQ plot) of standardized residuals of the linear regression model present in figure 7 we can clearly see that there is an acceptable normality distribution with a skew on the right of the plot which might indicate the presence of outliers. This positive skewness can lead to biased estimates and wrong confidence intervals in our analysis.

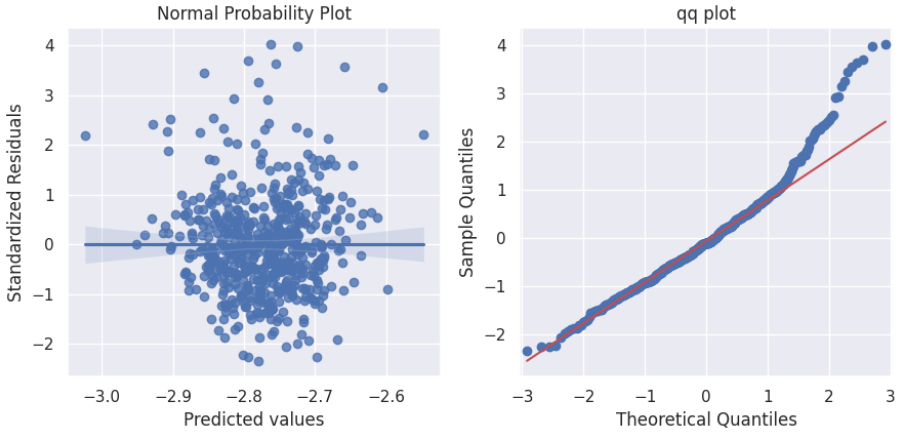


Figure 8: Logaritmic Normal Probability Plot of the Standardized Residuals

In order to minimize the skewed distribution, we have applied the log transformation to our data as shown in Figure 8. This transformation improves the linearization of the data, equalize the variance across the range of the data and help normalize the data. However, when we applied this transformation, we saw little difference in the skew despite still being an acceptable normality.

The population regression equation for the linear model with the natural log of *smoothness mean* and *fractal dimension mean* as predictor variables can be written as:

$$log(y) = \beta 0 + \beta 1 \times log(x)$$

$$log(y) = -1.784 + 0.421 \times log(x)$$

Where y corresponds to the *fractal dimension mean* and x to *smoothness mean*. The coefficient β1 with value 0.421 represents the change in log(y) per one unit increase in the variable log(x). The intercept β0 with value -1.784 represents the log of the *fractal dimension mean* when the log of *smoothness mean* is equal to zero. However, we cannot interpret this value since log of *smoothness mean* cannot take a value of zero.

These features have a **Pearson Correlation Coefficient** of 0.5848, this means that they have a modest positive linear relationship. The Coefficient of Determination (R^2) has a value of 0.3420, which means that approximately 34.20% of the variability in *fractal dimension mean* can be explained by *smoothness mean*. This leads us to conclude that this model is not able to explain a large portion of the variability in the data, yet it still significant. This small coefficient may be due to various factors such as the presence of the outliers or a poor linear relationship between these features.

Because we got a poor result in the coefficient of determination, we proceeded to analyse the linear relationship between *smoothness mean* and *fractal dimension mean* variables in the data set. This relationship is measured through F-statistic and is associated probability value (p-value). The results of the t-test show us F-statistic=294.7 with a very low p-value=$1.69e^{-53}$ for the *smoothness mean* coefficient is very close to 0. This leads us to conclude that there is a significant linear relationship between these two features.

We then constructed a 95% **confidence interval for the unknown true slope of the regression line** and from its output, we can see that the slope of the regression line is between 0.259982 and 0.327163. This means that there is a positive linear relationship where the *fractal dimension mean* increases approximately 0.26 to 0.33 for every unit increase in the *smoothness mean*.

Next, we constructed the 95% **confidence interval for the population correlation coefficient**. The coefficient lays between these values [0.5179, 0.6517]. This means that there is 95% confidence that the true population correlation coefficient between the two variables falls within this interval. Since the interval does not include zero, we can conclude that there is a statistically significant correlation between the two variables. Moreover, since the interval is positive, we can conclude that the correlation is positive.

We proceeded to construct a 95% **confidence interval for the mean** of *fractal dimension mean* at a random fixed value of *smoothness mean*. The x value is 0.1072. We then calculated the predicted *fractal dimension mean*:

$$\hat{y} = \beta 0 + \beta 1 \times x = -1.7843 + 0.4209(0.1072) = -1.7364$$

Using the t-critical value of 1.964 with 567 degrees of freedom (n-2), the 95% confidence interval for the predicted mean value of y is -1.7364 $\pm$ 1.964 = [-0.0380, 0.1700]. This means that we are 95% sure that the true mean value of *fractal dimension mean* at a fixed value of 0.1072 for *smoothness mean* falls between -0.0380 and 0.1700. However, this interval is not very useful since it contains positive and negative values, making it difficult to reach any conclusion.

Finally, we constructed a 95% **confidence interval for the mean** of *fractal dimension mean* but this time at a random value of *smoothness mean*. The x value is 0.09081. The confidence interval for the predicted value of the *fractal dimension mean* is [-1.9050, 2.0273]. This means that we are 95% certain that if we were to choose a new observation with a *smoothness mean* of 0.09081, we can be 95% confident that its *fractal dimension mean* value falls within this range. This prediction interval has the same problem as the previous one. Not only is it wide but it includes the value 0 which may not be very useful for making accurate predictions.

6 Principal Component Analysis

Principal Component Analysis (PCA) is a statistical technique used for dimensionality reduction and data compression. It is commonly used in data preprocessing and exploratory data analysis to identify patterns and reduce the number of variables in a data set.

Before PCA we performed a feature selection task to reduce the number of attributes and focus on six attributes. The choice was made by trying to keep as many different semantic aspects among the attributes as possible. We decide to discard highly correlated attributes because they provide redundant information.
The selected features are: *radius mean*, *texture mean*, *smoothness mean*, *compactness mean*, *concavity mean* and *fractal dimension mean*.

After this first step we proceed with PCA with two main goals in mind:

1. Plot the data using two and three principal components.
2. Determine which normalization technique, either *by range* or *by standard deviation*, is more suitable and useful for the breast cancer dataset.

We normalize data:

- by standard deviation : $\frac{x-\mu}{\sigma}$
- by range: $\frac{x-\mu}{x_{max}-x_{min}}$

and we obtain two data set on which we applied PCA.

We select the components associated to the three highest eigenvalues. The information on the variance captured by a principal component is contained in its eigenvalue, the higher the more variance of the data explained by that component.

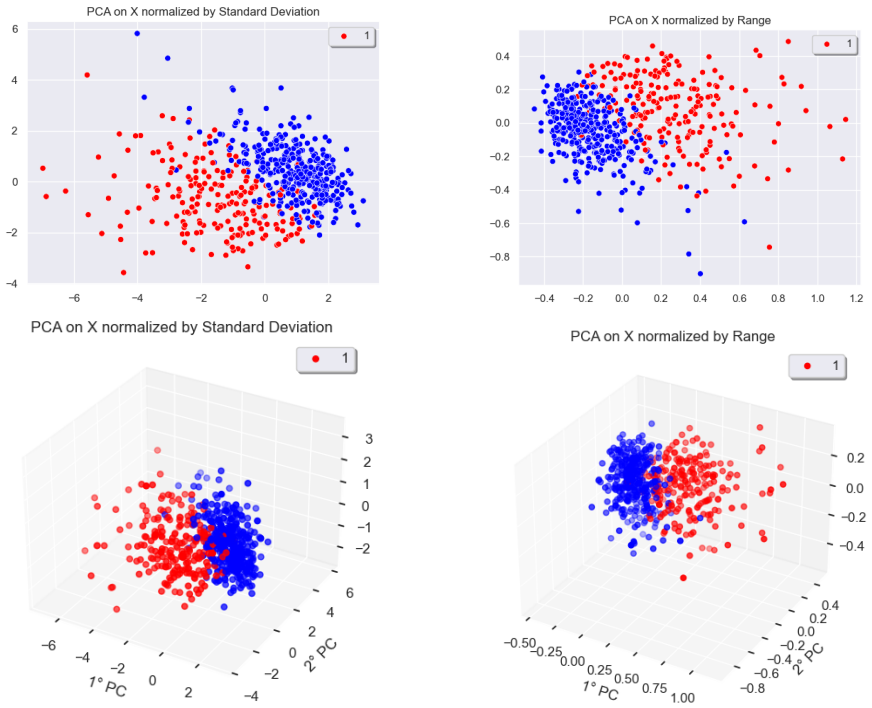


Figure 9: Plot in 2 and 3 PCs

Figure 9 shows plot in 2PCs and 3PCs plane, on the left side we have data normalized by standard deviation while on the right side by range.
We can see how clearly-separated the two classes are in the reduced space (malignant diagnosis=1).

We assess PCA quality through a **cumulative plot**. Figures 11 and 10 show that the first three components explained more than 90% of the total variance of data in both normalization techniques. However cumulative variance captured by the first three PCs is higher with data normalized by range (92.3%) than standard deviation (91%). This suggest that normalization by range is slightly preferable for *breast cancer data set*.

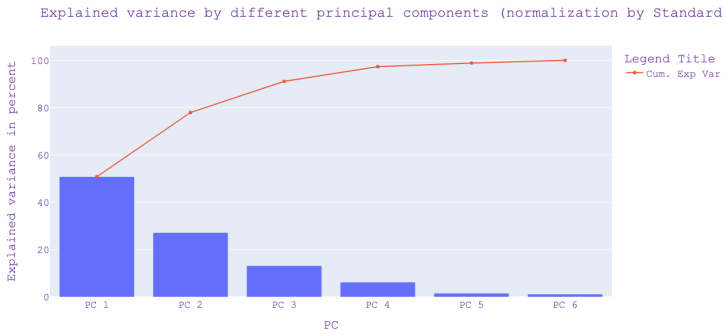


Figure 10: Cumulative plot PCA - data normalized by std

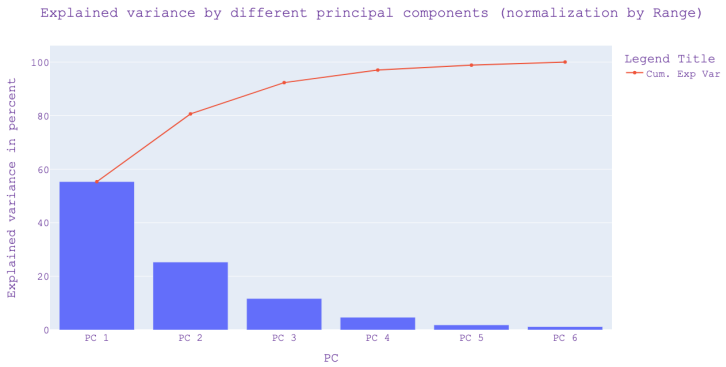


Figure 11: Cumulative plot PCA - data normalized by range

7 Clustering

In this section, we will explore **Fuzzy c-means** clustering algorithm that allows for soft clustering where a data point can belong to multiple clusters to varying degrees. It can be useful in the situation of breast cancer diagnosis because there may be ambiguity in assigning a data points to a single cluster. We will also discuss how we have evaluate the quality of clustering results and how to choose the appropriate number of clusters.

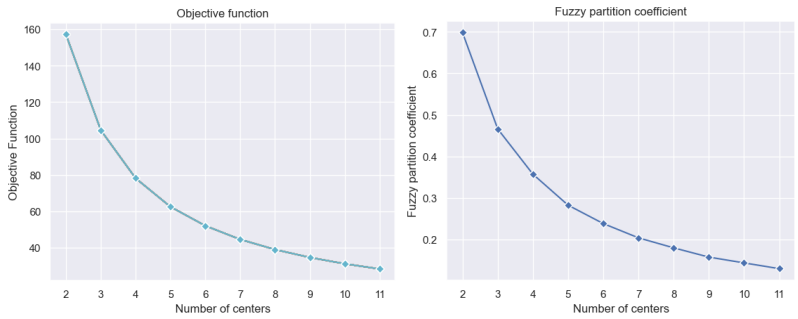
FCM results are dependent on the initialization of the centroids of the clusters. We will exploit **Anomalous patterns** clustering algorithm to try to improve the accuracy and effectiveness of FCM clustering results through a smart initialization of the centroids. Anomalous patterns are data points that are significantly different from the majority of the data and can provide a better representation of the overall data distribution. By including these anomalous patterns as initial centroids in FCM, it can help avoid suboptimal solutions and lead to more accurate and robust clustering results.

7.1 Fuzzy c-means

The first step consists in run **fuzzy c-means** with different value of c, which is the number of cluster returned by the algorithm. Because different initialization can lead to different FCM clustering solutions, ten different random initialization are performed for each value of c.
Assessing the results with multiple initialization can help identify the stability of the clustering results, as a robust solution should be consistent across multiple initialization.

The obtained results of FCM can be evaluated by plotting the value assumed by the **objective function** at the last iteration wrt c hyperparameter. Figure 12a shows surprisingly consistent values regardless the initialization as the objective function always converge to very similar values.

Figure 12b plot the **partition coefficient** wrt c hyperparameter. It shows high jumps until c equal to five, after that the increase in partition coefficient index start to be smaller, this would suggest that the optimal number of cluster is equal to five.



(a) FCM Objective Function (b) FCM Partition Coefficient

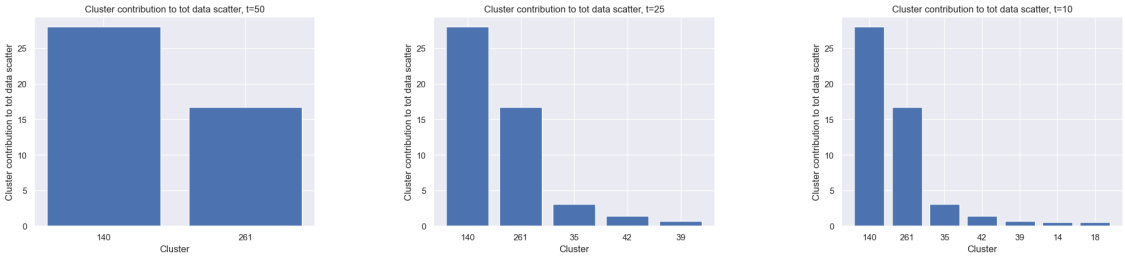
Figure 12: FCM results wrt different initializations on c values

7.2 Anomalous Pattern - Fuzzy c-means

FCM clustering results can be improved through smart initialization exploiting **Anomalous pattern** algorithm which do not require to set apriori the number of cluster. The goal of AP is to find populated anomalous clusters. The algorithm have a threshold parameter for the minimum number of points in a cluster. It can be tuned by trying different threshold values (e.g. 50, 25 and 10) and then evaluate the size of the clusters returned and their contributions to the total data scatter. Looking at figure 13 the best setting corresponds to threshold at 25, because smaller values return clusters with very few points compared to the other.

Figure 13b show that the first two clusters overcome the others both in terms of contribution to the total data scatter and size, indeed we've decided to perform FCM with c equal to two. Then we have computed the membership matrix of data wrt to AP's centroids coordinates to initialize the algorithm.

To check the clustering of AP-FCM we proceed by doing the defuzzification of the membership matrix returned by AP-FCM by *maximum membership degree*.

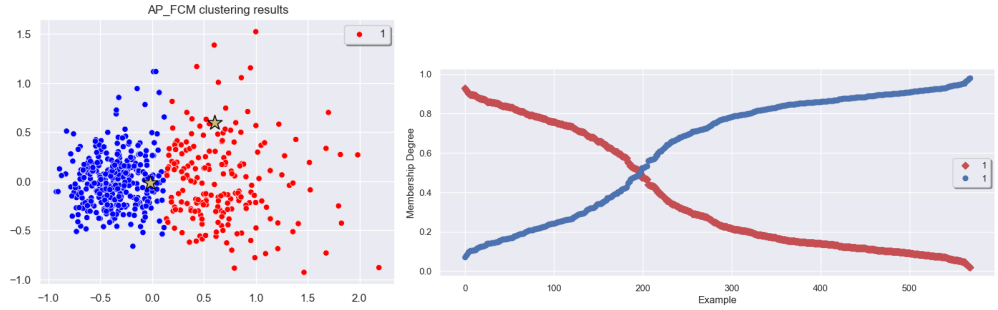


(a) t = 50 (b) t = 25 (c) t = 10

Figure 13: Anomalous Pattern

We exploit PCA to be able to plot the crisp partition we obtained in two PCs (figure 14a)

In **membership degree plot** in figure 14b the points are sorted according to their membership degree wrt to the two clusters. The plot show that there are approximately 100 data points with membership degree between 0.4 and 0.6, meaning that they are difficult to fit into one cluster or the other.

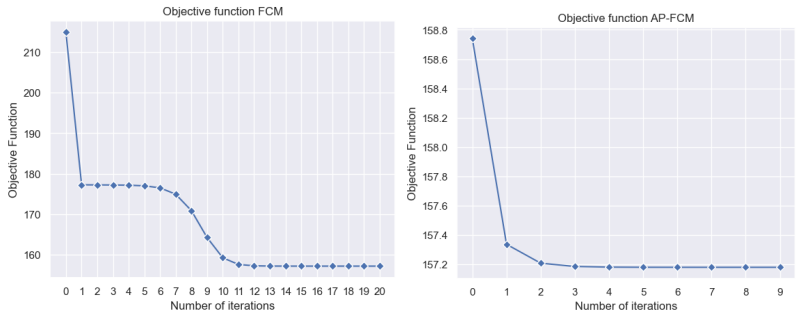


(a) AP-FCM clustering results in 2PC (b) Membership Degree Plot

Figure 14: AP-FCM clustering results

7.3 AP-FCM vs FCM

To assess the outcome of AP-FCM we can compare its results with a random initialized FCM (both with c equal to two). In such a way we can evaluate if the smart initialization really lead to improvements. Figure 15 show the value of the objective function according to the iteration. The starting value of the objective function is considerable slower with smart initialization and the objective function curve is also smoother. Furthermore, it takes a lower number of iterations to reach the minimum (9 wrt to 20), and the value drop almost to the minimum after just two iterations. In terms of efficiency AP-FCM is definitely better than randomly initialized FCM.



(a) FCM (b) AP-FCM

Figure 15: Objective Function comparison

Validation indices allow the evaluation of two clustering. We compute two validation indices to assess differences between FCM and AP-FCM:

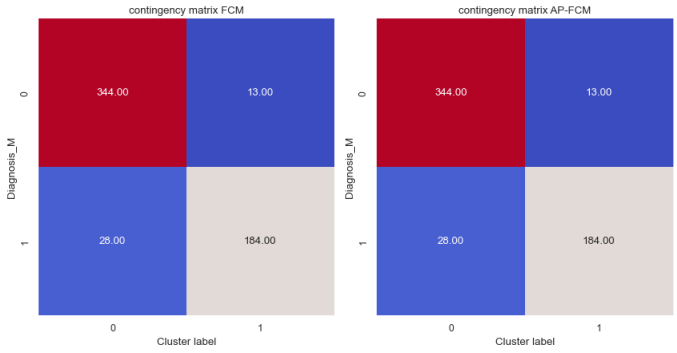
- **Fuzzy Partition Coefficient**, defined on the range from 0 to 1, with 1 being best. It is a metric which tells us how cleanly our data is described by a certain model.
- **ARI - Adjusted Rand Index**, is a basic measure of similarity between two clusterings by considering all pairs of samples and counting pairs that are assigned in the same or different clusters in the predicted and true clusterings. It is an adjustment of the Rand Index which does not take into account the fact that some agreement between two clusterings can occur by chance. The adjusted Rand index is thus ensured to have a value close to 0.0 for random labeling independently of the number of clusters and samples and exactly 1.0 when the clusterings are identical. The true labels considered in ARI are the labels of the target attribute *diagnosis M*.

Table 4 show that the results are the same for both random initialization and smart initialization.

	Fuzzy Partition Coefficient	**ARI**
FCM	0.698	0.730
AP-FCM	0.698	0.730

Table 4: Validation Indices

By comparing AP-FCM and FCM clustering results with the ground truth partition of benignant and malignant diagnoses, can be determine the extent to which the clustering adhere to the diagnosis partition. **Contingency matrix** in figure 16 show that the results for FCM and AP-FCM are the same, this support the same ARIs' value. The total number of mislabeled entities is 41 over a total of 569 records in the dataset, which is a great results. The clustering output obtained from FCM accurately reflects the distinction between benign and malignant diagnoses of the data points.



(a) FCM (b) AP-FCM

Figure 16: Contingency matrices FCM vs AP-FCM

8 Conclusions

Breast Cancer data set is an high quality data set, with very interesting properties to explore but also with possibly many direct machine learning applications to develop in reality.

- **Regression Analysis** of fractal dimension mean on smoothness mean show a positive correlation between the two attributes and significant coefficients of the slope and intercept. We cannot be completely confident to exploit ERE to make inference because some of the regression assumptions are slightly violated, such as normality assumption. This is due to the behavior of few data points which can be discarded through an outlier detection step, which was out of the scope of our analysis.
- **Principal Component Analysis** point out that the distinction between benignant and malignant could be carried out more easily in the lower dimensional space where the two classes where highly separated. We also reach the conclusion that normalization by range can be preferred to normalization by standard deviation on *breast cancer data set* because it allow to retain more variance in the lower dimensional space.
- In **Clustering** section we exploited fuzzy c-means that allows us to get a fuzzy partition of the data, a very useful property in breast cancer diagnosis context as it allow to figure out which records were difficult to fit into one cluster or another. We performed an analysis to see if initializing FCM with centroids returned by Anomalous pattern increase the performance of the algorithm compared to a random initialization. We concluded that the smart initialization increased the efficiency of FCM. Concerning the clustering partition obtained with AP-FCM, we saw through the contingency matrix that it almost reflects the partition between benignant and malignant data points.

9 Appendix

9.1 Part III - Principal Component Analysis

```
def PCA(X, n_components):
    """
    Principal Component Analysis

    input
    - X: data matrix
    - n_components: number of components returned

    output
    - PCA_X: projection of X in the reduced space
    - eigenvalues
    - eigenvectors
    """
```

```
def SVD(X, n):
    """
    Singular Value Decomposition.

    input
    - X: data matrix
    - n: number of dimension in the reduced space

    output:
    - SVD_X: reduced space
    - s: vectors with the singular values
    - vh.T: singular vectors

    """
```

```
def norm_by_range(X):
    """
    Normalization by Range

    input
    X: data matrix

    output
    norm_by_range_X: data normalized by range
    """
```

```
def norm_by_std(X):
    """
    Normalization by Standard deviation

    input
    X: data matrix

    output
    norm_by_range_X: data normalized by std
    """
```

9.2 Part IV - Clustering

```
def membership_matrix(data, centroids):
    """
    Compute the membership matrix

    input:
    data: data matrix (normalized)
    centroids: list of centroids coordinate

    return:
    U: membership matrix
    """
```