Breast Cancer Dataset analysis

The following analysis refers to the *Breast Cancer Dataset,* a public dataset that contains information about the core characteristics of cells present in the image. Its features were calculated from a scanned image of one aspirated by a fine needle of a mammary mass. Download made on the https://www.kaggle.com/uciml/breast-cancer-wisconsin-data. The machine *learning* method - *KMeans was* used in the python programming language for dataset *analysis.*

The *dataset* presents several features *for* the characteristics of breast cells and, as a target, whether the detected breast cancer is benign or malignant. The objective of this analysis is to make a study with *clustering* of the *features* to verify if any of them can indicate the presence of a malignant cancer. It is important to note that in the original *dataset,* the information is not classified, that is, we do not know a priori which one is relevant for cancer detection, so a cluster analysis can indicate which of the features will be *useful* for other models *of supervised machine learning* to perform this prediction.

In the dataset, we have 567 observations, 212 of them indicating malignant cancer. There are, in total, 10 features computed:

- radius(mean distances from the center to points in the perimeter of the cell)
- texture (grayscale standard deviation)
- perimeter
- area
- smoothness (local variation in radius measurements)
- compactness (perimeter^2 / area - 1.0)
- concavity (severity of concavities in contour)
- concave points (number of concavities in the contour)
- symmetry
- fractal dimension ("coastline approximation" - 1)

And the target: diagnosis, being 'M', malignant and, 'B', benign.

The mean, standard deviation and *the largest* or worst (average of the 3 largest values) were also computed, generating a total of 30 *features.*

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean
0	842302	M	17.990	10.38	122.80	1001.0	0.11840	0.27760	0.300100	0.147100	0.2419
1	842517	M	20.570	17.77	132.90	1326.0	0.08474	0.07864	0.086900	0.070170	0.1812
2	84300903	M	19.690	21.25	130.00	1203.0	0.10960	0.15990	0.197400	0.127900	0.2069
3	84348301	M	11.420	20.38	77.58	386.1	0.14250	0.28390	0.241400	0.105200	0.2597
4	84358402	M	20.290	14.34	135.10	1297.0	0.10030	0.13280	0.198000	0.104300	0.1809
5	843786	M	12.450	15.70	82.57	477.1	0.12780	0.17000	0.157800	0.080890	0.2087
6	844359	M	18.250	19.98	119.60	1040.0	0.09463	0.10900	0.112700	0.074000	0.1794
7	84458202	M	13.710	20.83	90.20	577.9	0.11890	0.16450	0.093660	0.059850	0.2196
8	844981	M	13.000	21.82	87.50	519.8	0.12730	0.19320	0.185900	0.093530	0.2350
9	84501001	M	12.460	24.04	83.97	475.9	0.11860	0.23960	0.227300	0.085430	0.2030

Table 1 - Target and part of the *dataset features*

In table 1 we see our target, 'diagnosis' and features, 'radius_mean', 'texture_mean' etc. In the same table, we have a column with the patient ID, and in the end, we have an 'Unnamed: 32' column. Both are not necessary for analysis, so they have been discarded. Then, as 'diagnosis' is a categorical variable, it cannot be used in the KMeans model, so to perform cluster analysis and use the model in the correct way, the map function was used to replace 'M' by 1 and 'B' by 0.

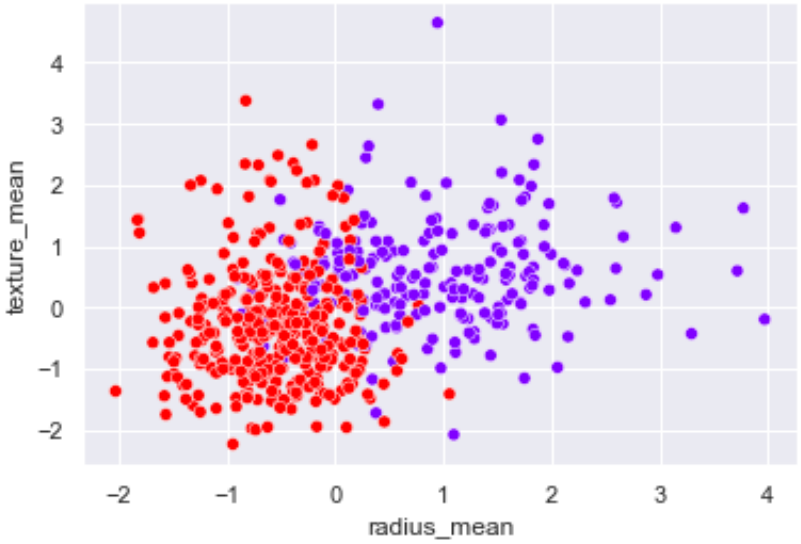

Figure 1 – Dispersion of radius_mean and texture_mean

All *features were* normalized for better analysis and performance of the algorithm. In Figure 1, the purple dots represent malignant cancer (1 or 'M') and benign reds (0 or 'B'). With this, it was verified that the larger the cell radius, the more malignant cancer diagnoses were identified. Therefore, 'radius_mean' is an important feature to be analyzed in prediction algorithms. 'Texture_mean,' even with its increase, was not determinant for the type of cancer to be malignant. Although there are purple dots when the radius is not very large (less than 0), most observations are when the radius has a high value (greater than 0), so when the cell has a high radius value, the cancer is likely to be malignant.

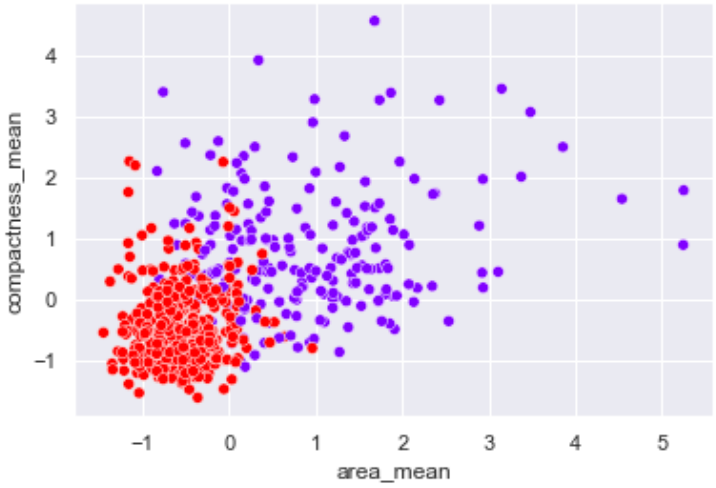

Figure 2 - Dispersion of area_mean e compactness _mean

The other *features that* have radius in their formula, for example, area and perimeter, are expected to behave in the same way as what was discovered with 'radius_mean', as can be seen in Figure 2.

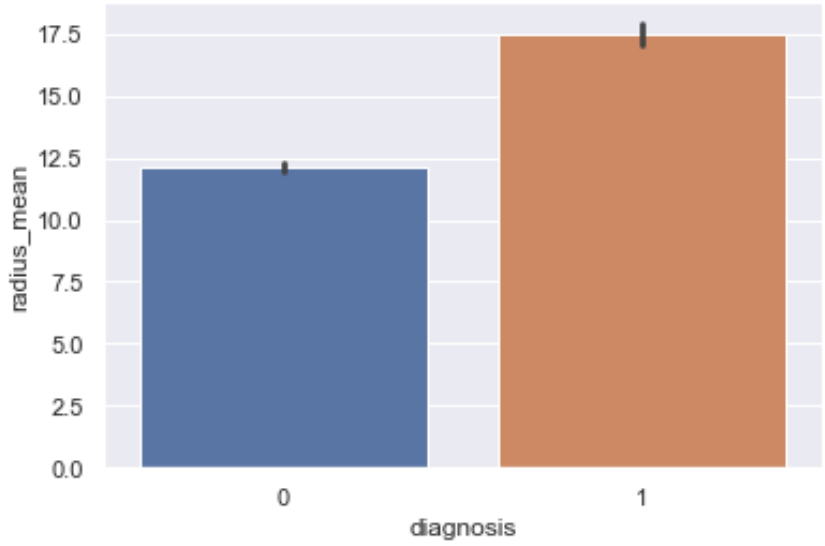

Figure 3 - Boxplot radius_mean and diagnosis

To corroborate the information that 'radius_mean' is an extremely relevant feature, we have Figure 3, which shows that the larger the cell radius, the greater the chance of the cancer being malignant.

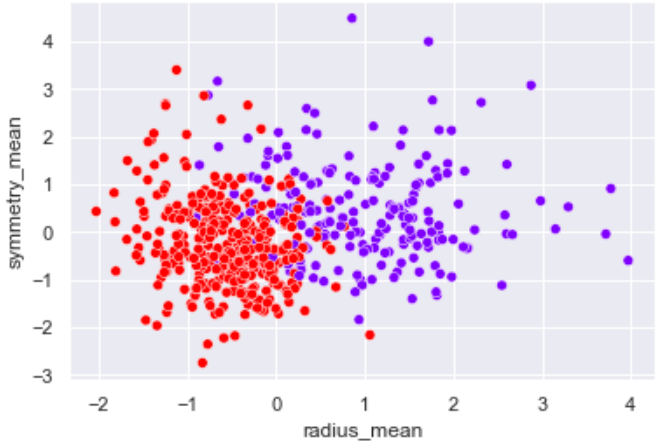

Figure 4 - Dispersion of radius_mean e symmetry _mean

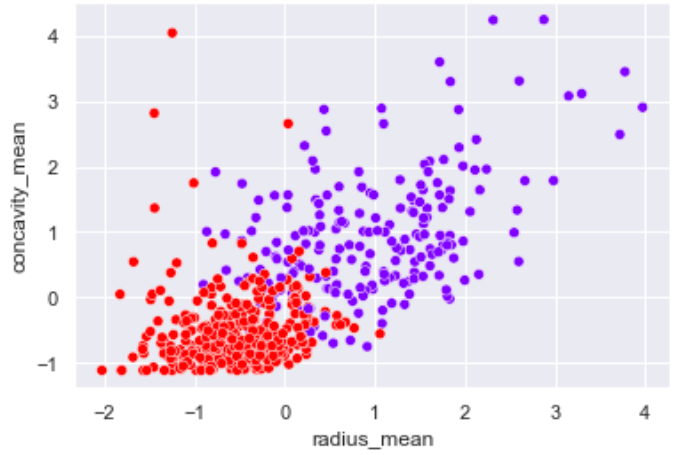

Figure 5 – Dispersion of radius_mean e concavity _mean

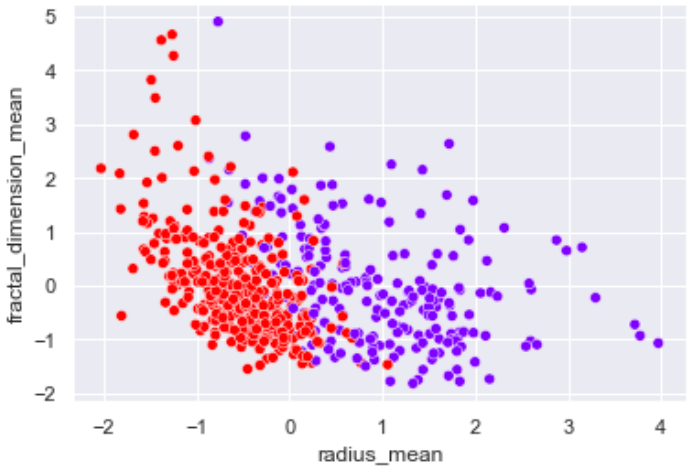

Figure 6 - Dispersion de radius_mean e fractal_dimension_mean

As was verified with 'texture_mean', 'symmetry', 'concavity' and 'fractal_dimension', they do not influence as much as the ray for malignant cancer, as seen in figures 4, 5 and 6, although there is an occurrence of the malignant form of cancer when the ray does not have a high value. Therefore, the *most* important feature for future forecastmodels is 'radius_mean'.