# Análise do Breast Cancer Dataset

A análise a seguir se refere ao *Breast Cancer Dataset,* um *dataset* público que contém informações sobre características do núcleo de células presentes em imagem. As suas *features* foram calculadas a partir de uma imagem digitalizada de um aspirado por uma agulha fina de uma massa mamária. Download feito no site https://www.kaggle.com/uciml/breast-cancer-wisconsin-data. Utilizou-se o método de *machine learning – KMeans* feito na linguagem de programação *python* para a análise do *dataset*.

O *dataset* apresenta diversas *features* para as características das células mamarias e, como target, se o câncer de mama detectado é benigno ou maligno. O objetivo dessa análise é fazer um estudo com *clustering* das *features* para verificar se alguma delas pode indicar a presença de um câncer tipo maligno. É importante ressaltar que no *dataset* original, as informações não estão classificadas, ou seja, não sabemos a priori qual delas é relevante para a detecção de câncer, portanto, uma análise por clusters pode indicar qual das *features* serão uteis para outros modelos de *supervised machine learning* para realizar essa previsão.

No dataset, temos 567 observações, sendo 212 delas indicando câncer maligno. São, no total, 10 features computadas:

- radius (média das distancias do centro para pontos no perímetro da célula)
- texture (desvio padrão da escala de cinza)
- perimeter
- area
- smoothness (variação local em medidas de raio)
- compactness (perímetro^2 / área - 1.0)
- concavity (severidade das concavidades no contorno)
- concave points (número de porções concavas no contorno)
- symmetry
- fractal dimension ("coastline approximation" - 1)

E o target: diagnosis, sendo ‘M’, maligno e, ‘B’, benigno.

A média, desvio padrão e o *largest* ou *worst* (média dos 3 maiores valores) também foram computados, gerando num total de 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 |

Tabela 1 – Target e parte das *features* do *dataset*

Na tabela 1, vemos o nosso target, ‘diagnosis’ e as features, ‘radius_mean’, ‘texture_mean’ etc. Na mesma tabela, temos uma coluna com o ID dos pacientes e, no final, temos uma coluna ‘Unnamed: 32’. Ambas não são necessárias para a análise, então foram descartadas. Em seguida, como ‘diagnosis’ é uma categorical variable, ela não pode ser usada no modelo KMeans, então para realizar a análise por clusters e utilizar o modelo da maneira correta, utilizou-se a função map para substituir ‘M’ por 1 e ‘B’ por 0.

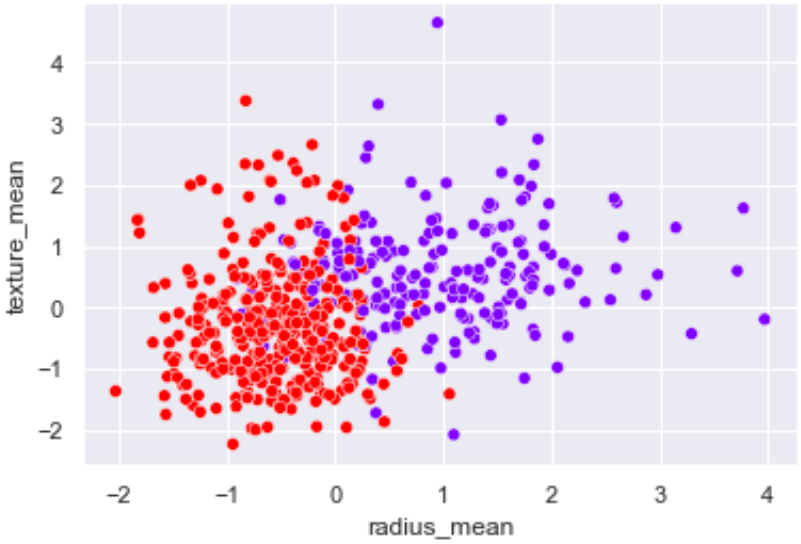


*Figura 1 – Dispersão de radius_mean e texture_mean*

Todas as *features* foram normalizadas para uma melhor análise e performance do algoritmo. Na figura 1, os pontos roxos representam câncer maligno (1 ou ‘M’) e vermelhos benignos (0 ou ‘B’). Com isso, foi verificado que quanto maior o raio da célula, mais diagnósticos de câncer malignos foram identificados. Portanto, ‘radius_mean’ é um feature importante para ser analisada em algoritmos de previsão. Já ‘texture_mean’, mesmo com seu aumento, não foi determinante para o tipo de câncer ser maligno. Embora existam pontos roxos quando o raio não é muito grande (menor que 0), a maioria das observações são quando o raio possui valor alto (maior que 0), portanto, quando a célula apresentar um valor alto de raio, é provável que o câncer seja do tipo maligno.

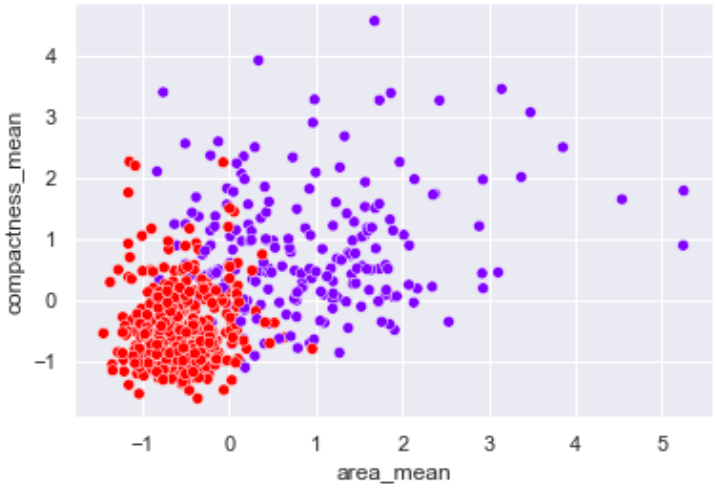


*Figura 2 - Dispersão de area_mean e compactness _mean*

As outras *features* que apresentam raio na sua formula, por exemplo, área e perímetro, é de se esperar que se comportem da mesma maneira que o que foi descoberto com 'radius_mean', como pode ser visto na figura 2.

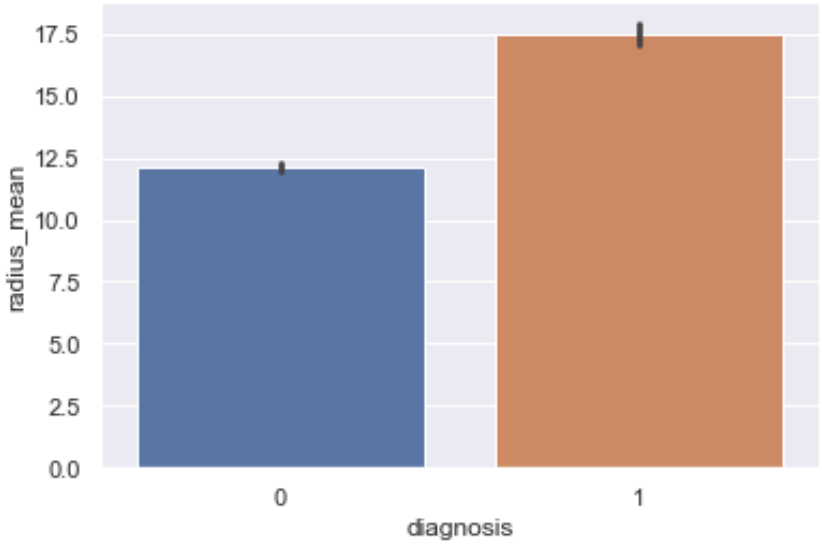


*Figura 3 - Boxplot radius_mean and diagnosis*

Para corroborar a informação de que 'radius_mean' é uma feature extremamente relevante, temos a figura 3, que mostra que, quanto maior o raio da célula, maior a chance de o câncer ser do tipo maligno.

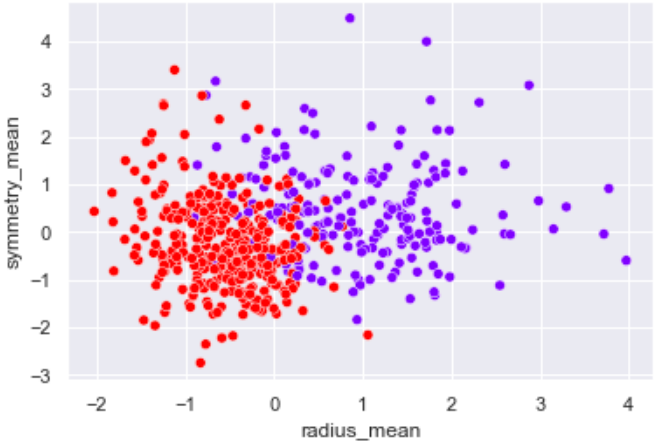


*Figura 4 - Dispersão de radius_mean e symmetry _mean*

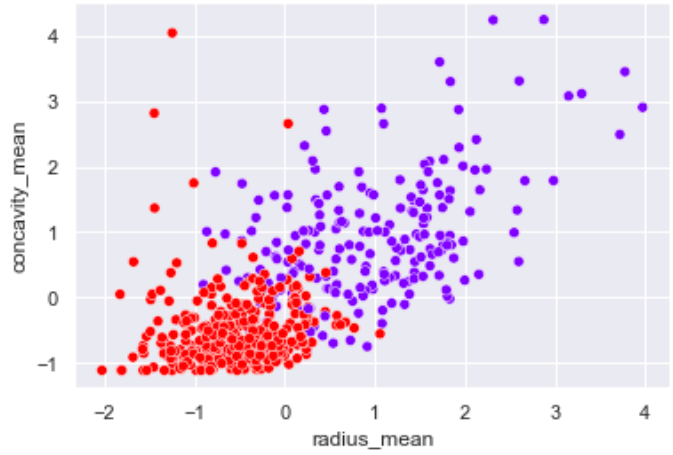


*Figura 5 – Dispersão de radius_mean e concavity _mean*

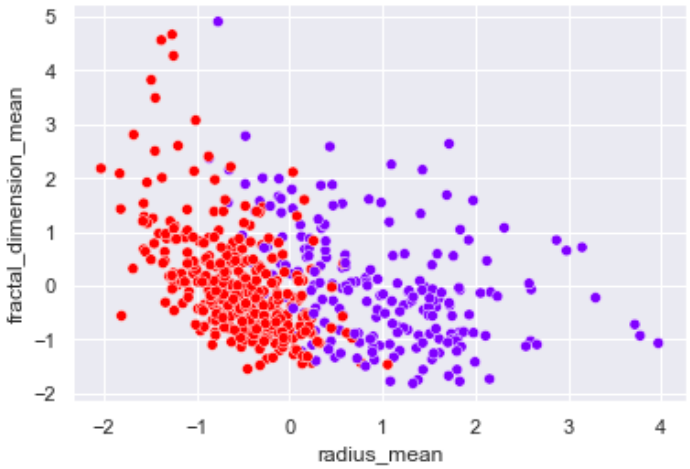


*Figura 6 - Dispersão de radius_mean e fractal_dimension_mean*

Assim como foi verificado com ‘texture_mean’, ‘symmetry’,’concavity’ e ‘fractal_dimension’, não influenciam tanto como o raio para o câncer maligno, como visto nas figuras 4, 5 e 6, embora haja ocorrência da forma maligna do câncer quando o raio não possui um valor alto. Portanto, a *feature* mais importante para futuros modelos de previsão é ‘radius_mean’.