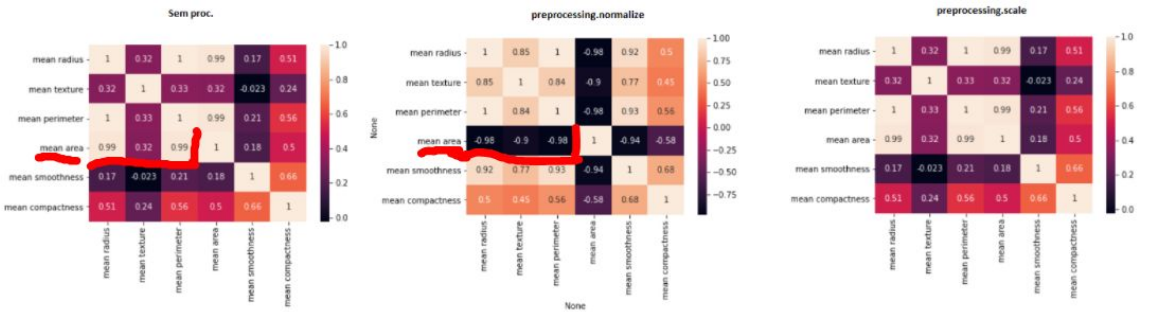
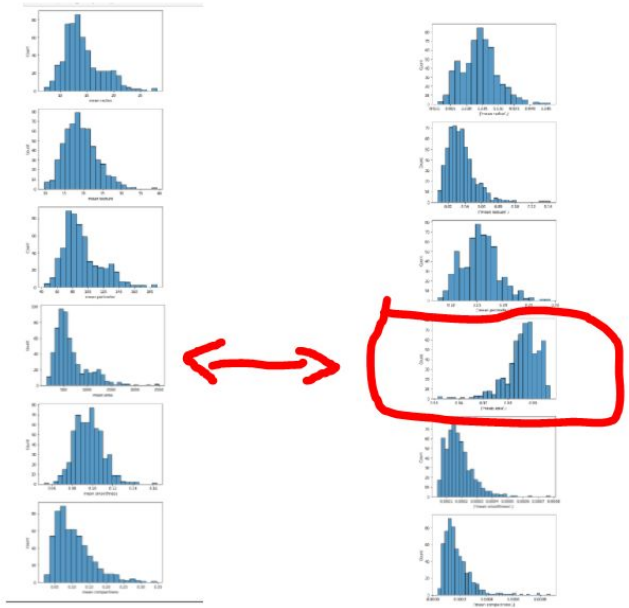
Estudo do efeito da normalização no cálculo da correlação

- Dataset com features sobre câncer de mama
- Dados em diferentes escalas
- Dados com distribuição assimétrica (skewed -> outliers)

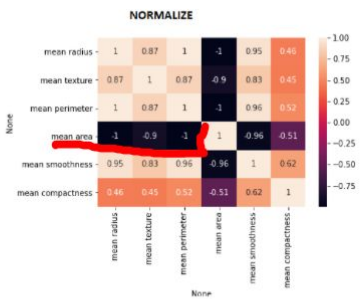
Caso 1: sem processamento x
preprocessing.normalize x
preprocessing.scale


Sem proc.
preprocessing.normalize
preprocessing.scale
mean radius
mean texture
mean perimeter
mean area
mean smoothness
mean compactness
None

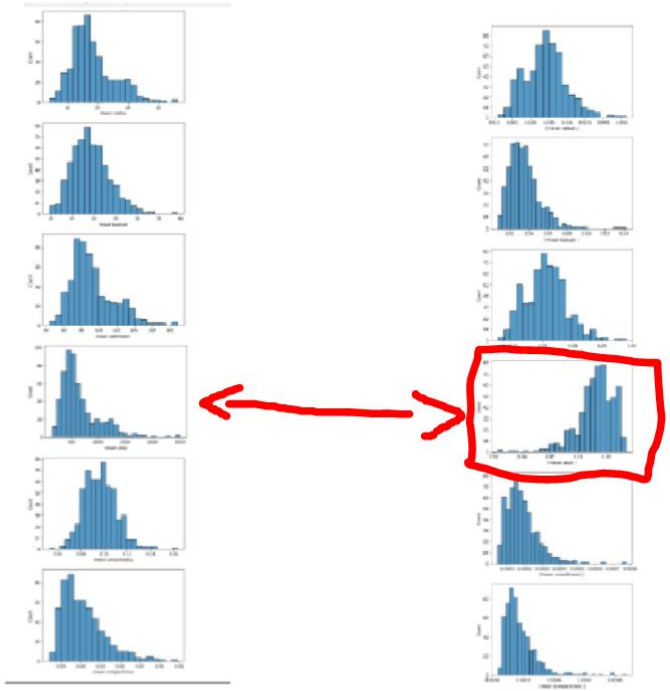


PEARSON


SEM PROC
mean radius
mean texture
mean perimeter
mean area
mean smoothness
mean compactness


NORMALIZE
mean radius
mean texture
mean perimeter
mean area
mean smoothness
mean compactness

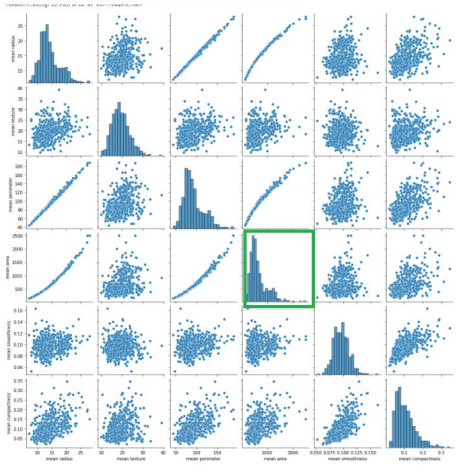

SCALER/STANDARDIZE
mean radius
mean texture
mean perimeter
mean area
mean smoothness
mean compactness



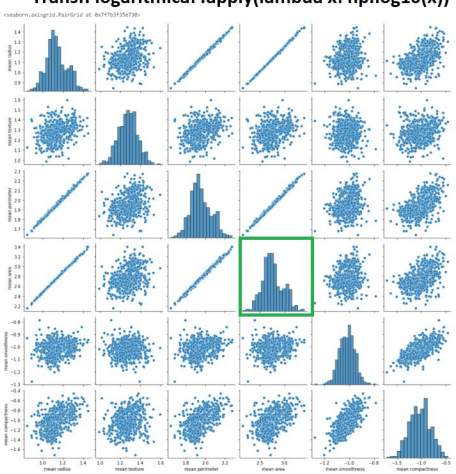
SPEARMAN

Caso 2: Sem processamento x
Transformação Logarítmica para normalização

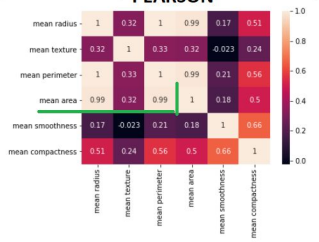
SEM PRE PROCESSAMENTO



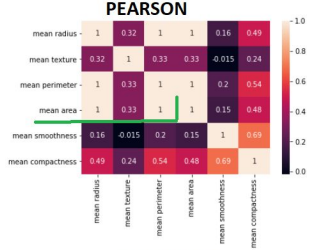
Transf. logarítmica: .apply(lambda x: np.log10(x))



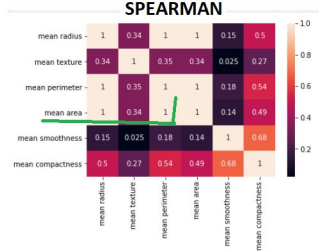
PEARSON



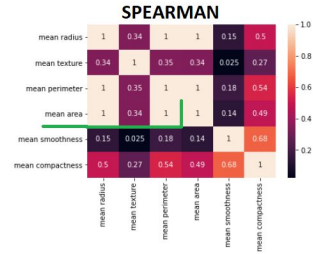
PEARSON



SPEARMAN

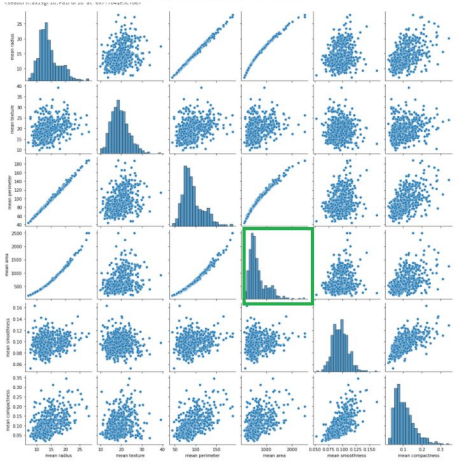


SPEARMAN

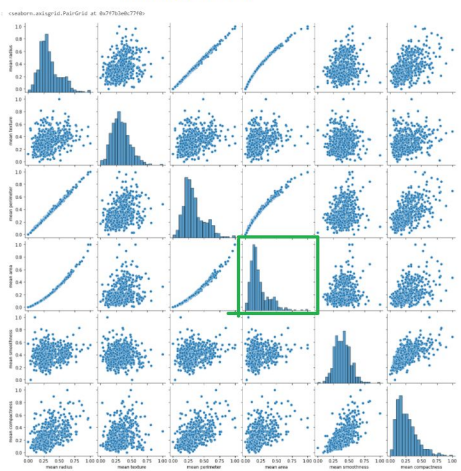


Caso 3: Sem processamento x MinMax Scaler para normalização

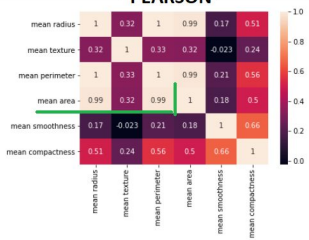
SEM PRE PROCESSAMENTO



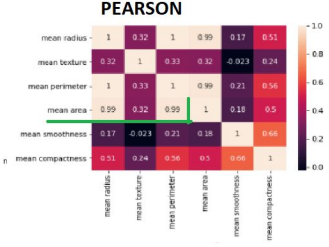
MinMax Scaler



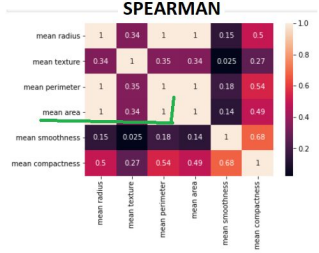
PEARSON



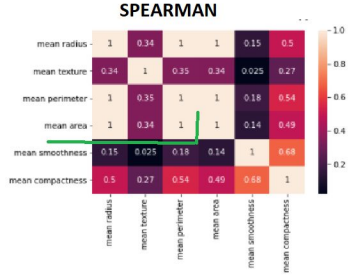
PEARSON



SPEARMAN



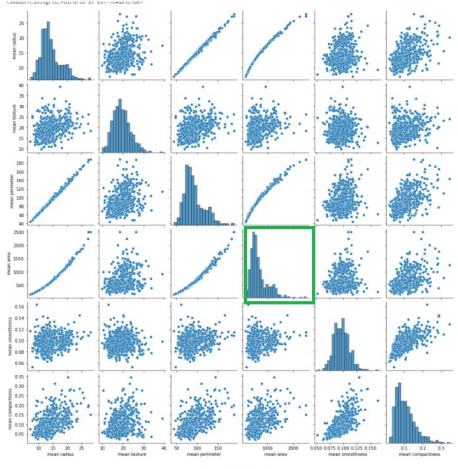
SPEARMAN



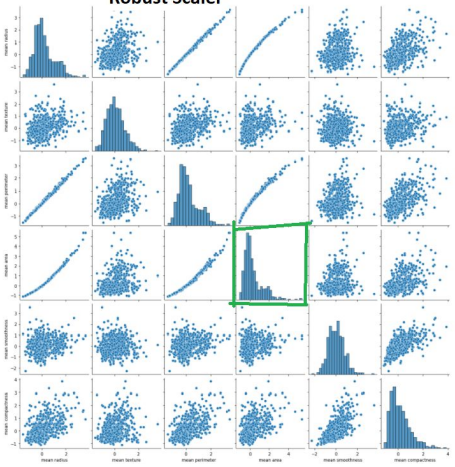
Caso 4: Sem processamento x Robust Scaler

- with_centering = True, with_scaling = True

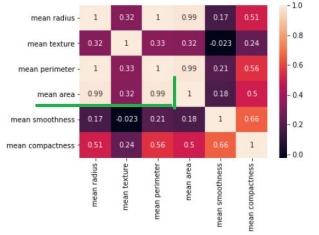
SEM PRE PROCESSAMENTO



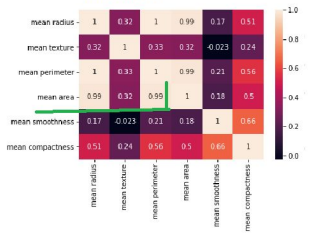
Robust Scaler



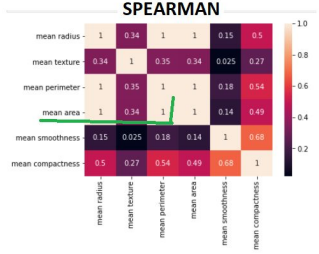
PEARSON



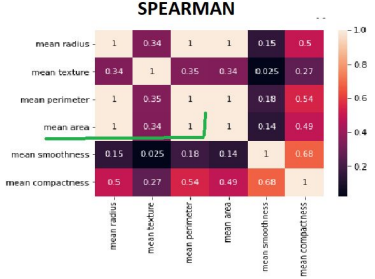
PEARSON



SPEARMAN

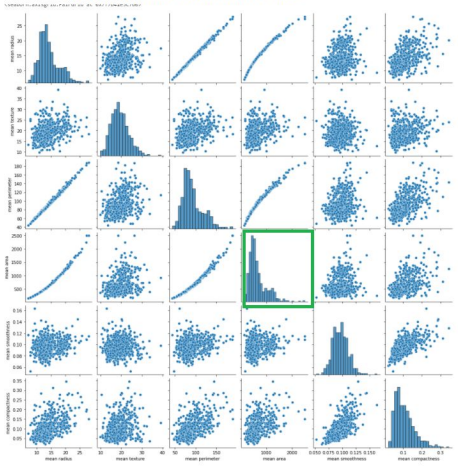


SPEARMAN

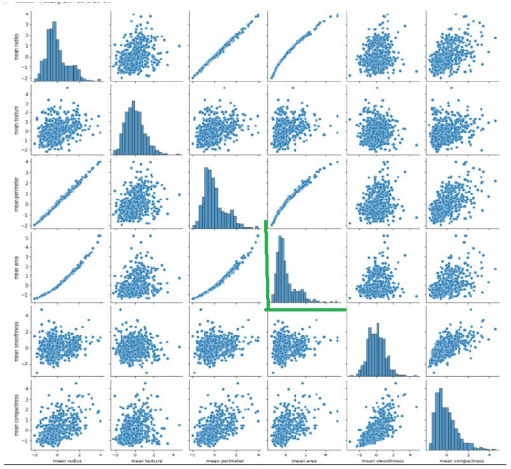


Caso 5: Sem processamento x Standard Scaler

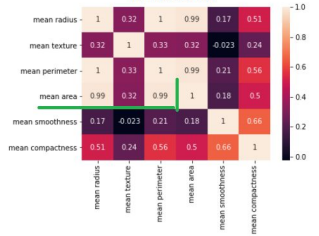
SEM PRE PROCESSAMENTO



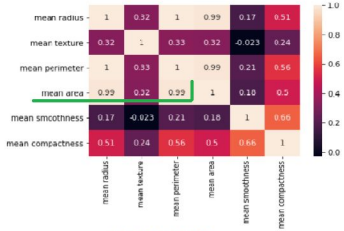
Standard Scaler



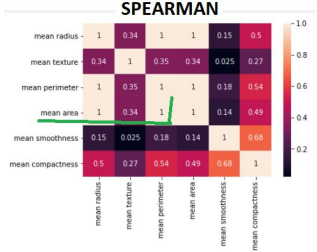
PEARSON



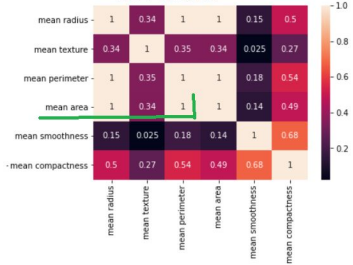
PEARSON



SPEARMAN



SPEARMAN



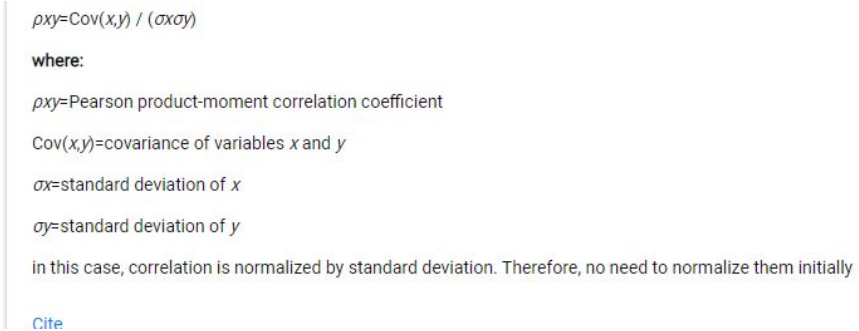
Conclusões

- Tanto normalização quanto escalonamento não parecem influenciar na correlação.

- A função preprocessing.normalize deve ser utilizada apenas para vetores : https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.normalize.html

- Discussão sobre o assunto:

https://www.researchgate.net/post/Do_we_need_to_standardize_variables_with_different_scales_before_doing_correlation_analysis



ρxy=Cov(x,y) / ($\sigma x \sigma y$)

where:

ρxy=Pearson product-moment correlation coefficient

Cov(x,y)=covariance of variables x and y

σx=standard deviation of x

σy=standard deviation of y

in this case, correlation is normalized by standard deviation. Therefore, no need to normalize them initially

Cite

Links utilizados

- https://www.researchgate.net/post/Do_we_need_to_standardize_variables_with_different_scales_before_doing_correlation_analysis
- https://machinelearningmastery.com/standardscaler-and-minmaxscaler-transforms-in-python/
- https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.normalize.html
- https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html?highlight=robust%20scaler#sklearn.preprocessing.RobustScaler
- https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.minmax_scale.html?highlight=min%20max#sklearn.preprocessing.minmax_scale