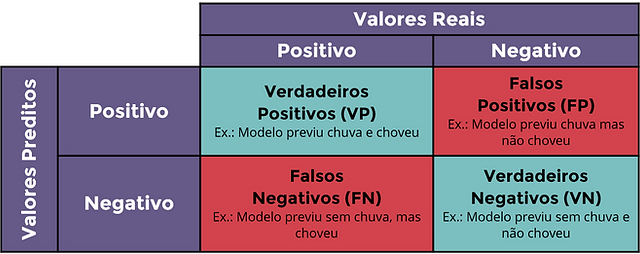
# Matriz de Confusão e Normalização

## Matriz de Confusão

Uma matriz de confusão é uma tabela utilizada para descrever a performance de um classificador binário ou multiclasse. Ela mostra quantas vezes cada classe foi prevista corretamente ou incorretamente. A matriz de confusão é composta por quatro elementos principais: verdadeiros positivos (VP), falsos positivos (FP), verdadeiros negativos (VN) e falsos negativos (FN).



A matriz de confusão é usada para avaliar a precisão de um classificador, pois ela permite calcular métricas como a acurácia, precisão, revocação e F1-score.

A matriz de confusão é preenchida com base nos dados de teste e nas previsões do classificador. O número de VP é o número de instâncias que foram classificadas corretamente como positivas. O número de FP é o número de instâncias que foram classificadas incorretamente como positivas. O número de VN é o número de instâncias que foram classificadas corretamente como negativas. E o número de FN é o número de instâncias que foram classificadas incorretamente como negativas.

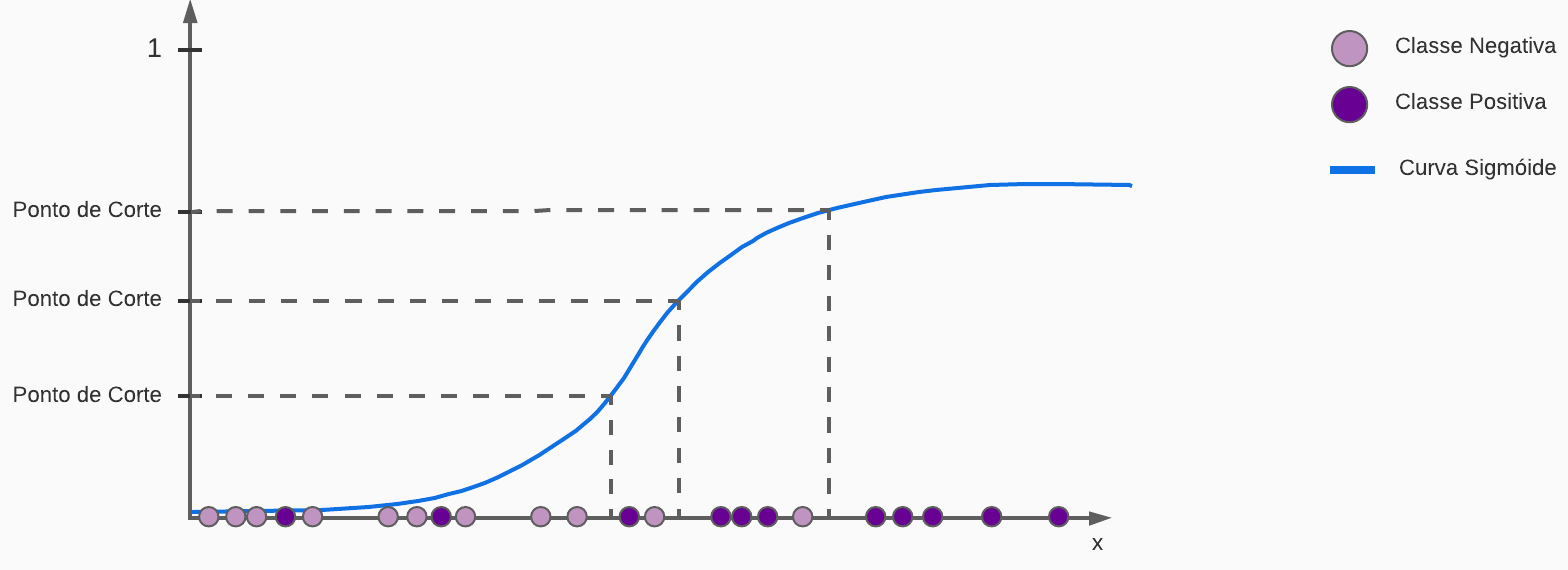
## ROC / AUC

A curva ROC (Receiver Operating Characteristic) é uma técnica utilizada para avaliar a performance de classificadores binários. Ela permite visualizar a relação entre a taxa de verdadeiros positivos (sensibilidade) e a taxa de falsos positivos para diferentes pontos de corte do classificador (no sigmóide). A área sob a curva (AUC) é utilizada como uma medida de desempenho global do classificador.

A ROC é útil principalmente quando há desequilíbrio entre as classes, onde uma classe é muito menor do que a outra. Nestes casos, a acurácia (proporção de acertos) pode não ser uma boa medida de desempenho, pois o classificador pode simplesmente rotular todas as instâncias como pertencentes à classe majoritária e ainda assim obter uma alta taxa de acerto. A ROC permite avaliar o desempenho do classificador considerando tanto a sensibilidade quanto a especificidade, o que é mais indicativo do desempenho real do classificador.

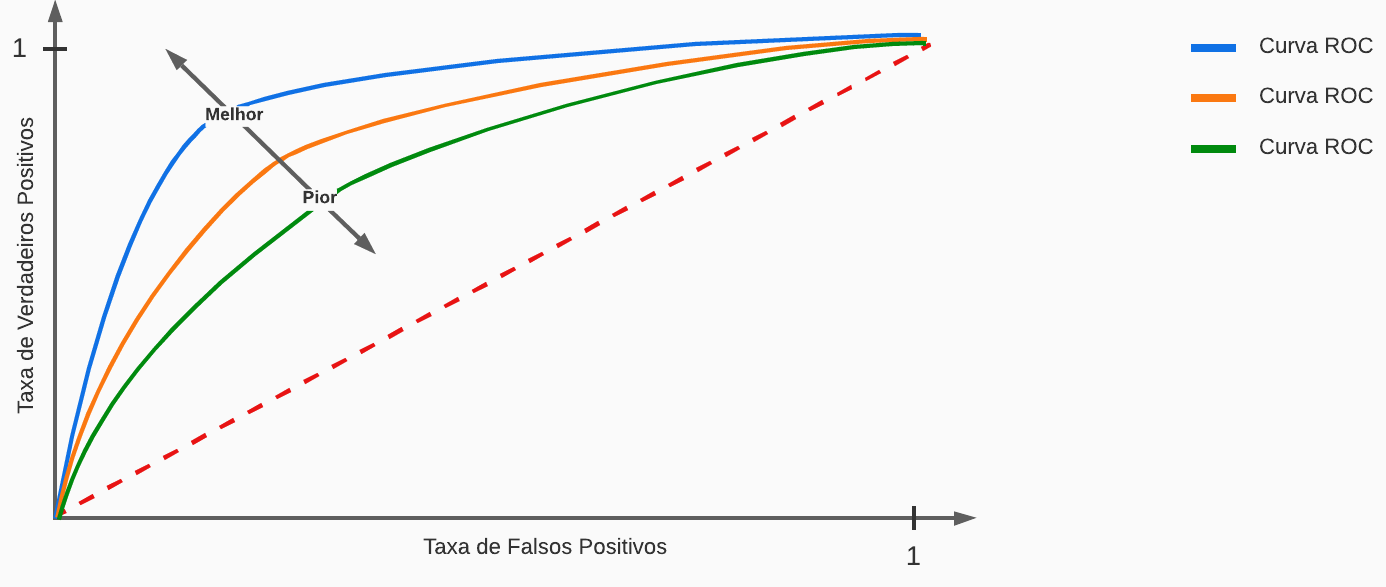
A função sigmóide é utilizada para calcular os valores de sensibilidade e especificidade através do processo de escolha de um ponto de corte para a probabilidade de uma instância pertencer à classe positiva. O ponto de corte é escolhido de acordo com a necessidade específica do problema e pode ser variado para obter diferentes pontos na curva ROC.

Uma vez que a saída do classificador é transformada em uma probabilidade através da função sigmóide, o valor é comparado com o ponto de corte escolhido. Se a probabilidade for maior que o ponto de corte, a instância é classificada como pertencente à classe positiva, caso contrário, é classificada como pertencente à classe negativa.



A sensibilidade é calculada como a proporção de verdadeiros positivos (VP) entre o total de positivos reais (VP + FN), enquanto que a especificidade é calculada como a proporção de verdadeiros negativos (VN) entre o total de negativos reais (VN + FP). Esses cálculos são feitos para diferentes pontos de corte do classificador, e o resultado é plotado no gráfico ROC, com a sensibilidade no eixo y e a especificidade no eixo x.

A AUC (Área Sob a Curva) é uma medida de desempenho global do classificador, que varia entre 0 e 1, sendo que quanto maior a AUC, melhor é a performance do classificador. A AUC é calculada como a área sob a curva ROC. Valores próximos a 1 indicam que o classificador é capaz de distinguir com sucesso entre as classes, enquanto valores próximos a 0 indicam que o classificador não é capaz de distinguir entre as classes.



Em resumo, a curva ROC é uma técnica utilizada para avaliar a performance de classificadores binários, permitindo visualizar a relação entre a sensibilidade e a especificidade para diferentes pontos de corte do classificador. A AUC é uma medida global de desempenho, calculada como a área sob a curva ROC, e varia entre 0 e 1, sendo que quanto maior a AUC, melhor é a performance do classificador. A função sigmoide é uma ferramenta comumente utilizada para transformar a saída do classificador em uma probabilidade, e é usada para calcular a sensibilidade e a especificidade necessárias para construir a curva ROC.

## Normalização

A normalização de dados é o processo de ajustar os valores de um conjunto de dados para que eles estejam dentro de um determinado intervalo. Isso é comumente feito para preparar os dados para algoritmos de aprendizado de máquina que se beneficiam de uma escala comum.

Existem várias técnicas de normalização, incluindo:

**MinMaxScaler**: Essa função padroniza os dados entre dois parâmetros estipulados, da seguinte forma:

, onde é o valor a ser normalizado, e são os valores mínimo e máximo da determinada coluna do dataset e os valores e definem os extremos dos parâmetros estipulados.

**StandartScaler**: Normaliza os dados a partir da fórmula:

, onde é o valor a ser normalizado, é a média e o desvio padrão da determinada coluna.

**MaxAbsScaler**: Normaliza os dados a partir da fórmula:

, onde é o valor a ser normalizado e é o valor máximo da determinada coluna.

**Normalize**: Realiza a normalização de cada linha da matriz (o cálculo é feito, por padrão, linha por linha em vez de coluna por coluna). possui três parâmetros possíveis: ‘l1’, ‘l2’ ou ‘max’.

* L1:
* L2:
* Max:

, onde é o valor a ser normalizado e z é um dos parâmetros acima.