PYTHON – NÍVEL INICIANTE

POWERED BY MULHERES EM DADOS

PROJETO SATISFAÇÃO DO CONSUMIDOR

AULA 04 – AVALIAÇÃO E MELHORIAS NO MODELO DE MACHINE LEARNING

06 DE JULHO DE 2022

CRONOGRAMA

Acesse **aqui** o notebook do projeto

JUNHO-JULHO 2022

DOM	SEG	TER	QUA	QUI	SEX	SAB
12	13	14	15 **Entendimento**	16	17	18
19	20	21	22 **Análise Exploratoria**	23	24	25
26	27	28	29 **Criação de modelo**	30	1	2
3	4	5	6 **Avaliação**	7	8	9

Entendimento, carregamento e pré-processamento dos dados

Entendimento do negócio e configurar o colab para pre-processar os dados

Análise exploratória de dados (EDA)

Analisaremos os dados resultando nos principais insights sobre o negócio

Feature Engineering e criação de modelo de ML

Criaremos o nosso modelo de machine learning a partir das variáveis que mais fazem sentido para o negócio

Avaliação e melhorias no modelo de ML

Entenderemos melhor sobre como avaliar e melhorar a performance do nosso modelo de machine learning

AULA 04 – AVALIAÇÃO E MELHORIAS NO MODELO DE MACHINE LEARNING

SUMÁRIO

MATRIZ DE CONFUSÃO

CLASSE PREDITIVA

CLASSE ORIGINAL

	positive	negative
positive	**T**rue **P**ositive	**F**alse **N**egative
negative	**F**alse **P**ositive	**T**rue **N**egative

Usa-se a matriz como representação visual. As métricas de classificação, ao lado, são as funções desta matriz e que verificam as taxas de acertos e erros das predições.

MÉTRICAS DE CLASSIFICAÇÃO

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

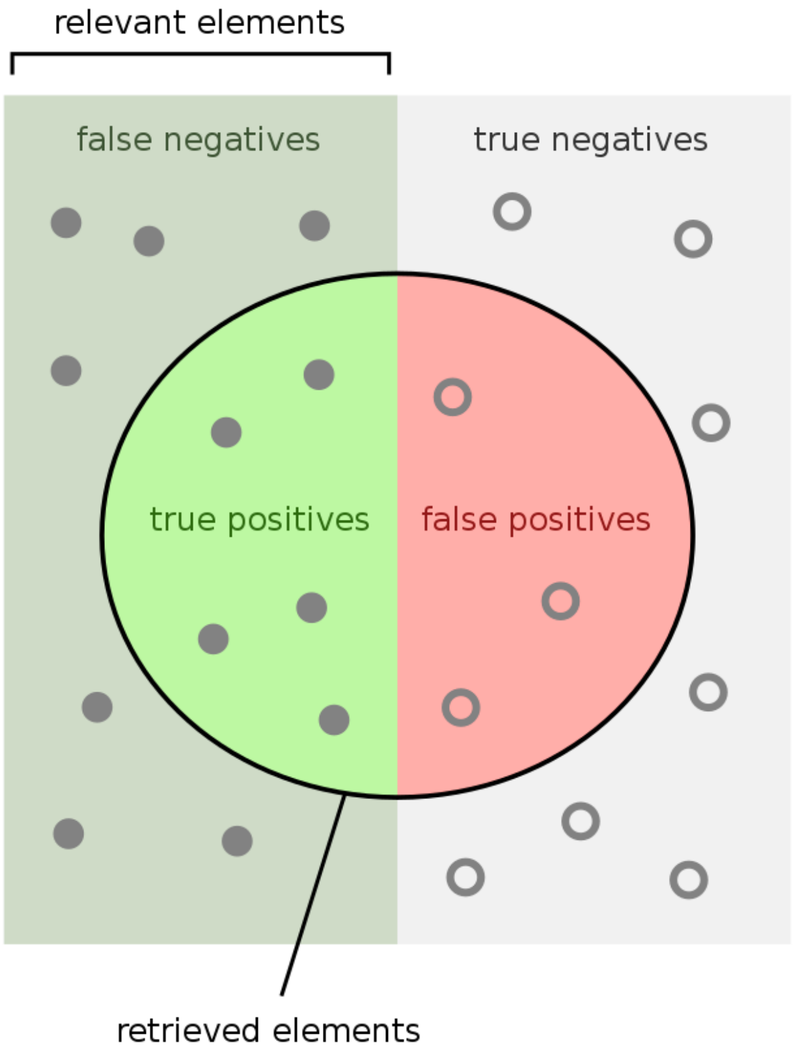
$$Precision = \frac{TP}{TP + FP}$$

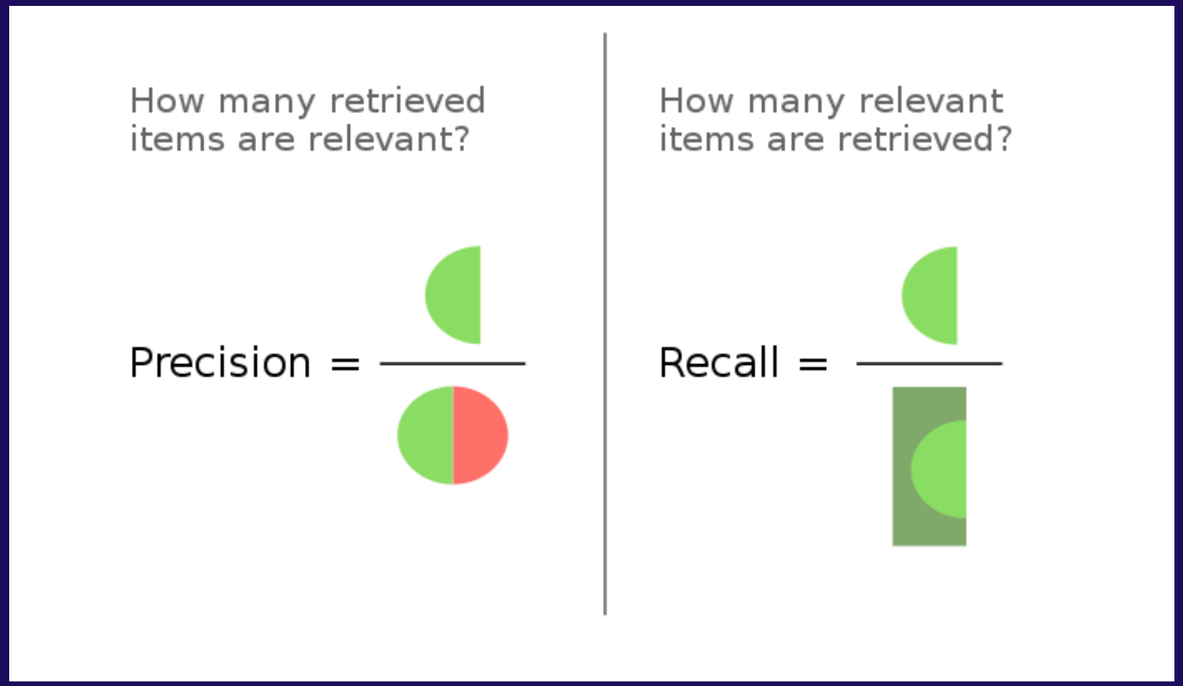
$$Recall = \frac{TP}{TP + FN}$$

$$F1\text{-}score = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

MÉTRICAS DE CLASSIFICAÇÃO





Precision : número de resultados positivos dividido pelo número de resultados positivos preditos

Recall: número de resultados positivos dividido pelo número de amostras que realmente eram positivas

A precisão é necessária para reduzir o número de falsos positivos e o recall é necessário para reduzir o número de falsos negativos.

BALANCEAMENTO DOS DADOS

- **Alta discrepância** entre quantidade de dados de duas classes

Exemplo: em detecção de fraude, a fraude é rara, então teremos muito mais valores apontando como 'não fraude'

- **SMOTE - Synthetic Minority Over-sampling Technique**: gera novas amostras artificialmente da classe minoritária utilizando o algoritmo KNN. Dessa forma, conseguimos balancear os tamanhos das amostras para treinamento

ESCALONAMENTO DOS DADOS

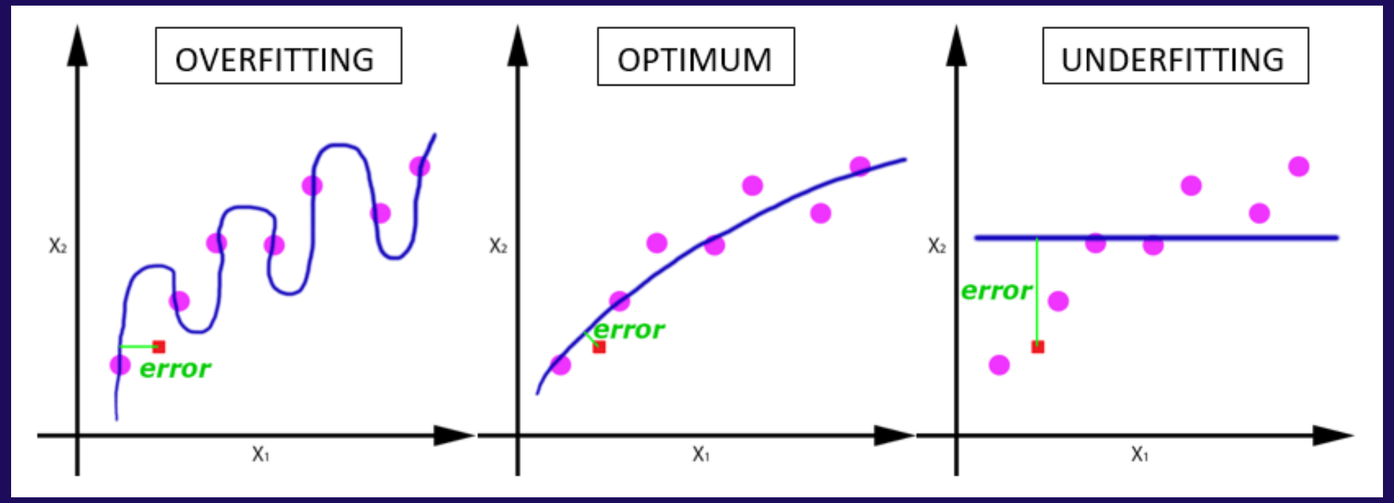
- Se o modelo possui diferentes ordens de grandeza
Exemplo: idades entre 0-100 e salários de 0-milhares
- Alguns modelos podem não funcionar com propriedades com diferentes escalas
- Pode resultar em algumas propriedades tendo mais peso que outras

TIPOS DE ESCALONAMENTO DOS DADOS

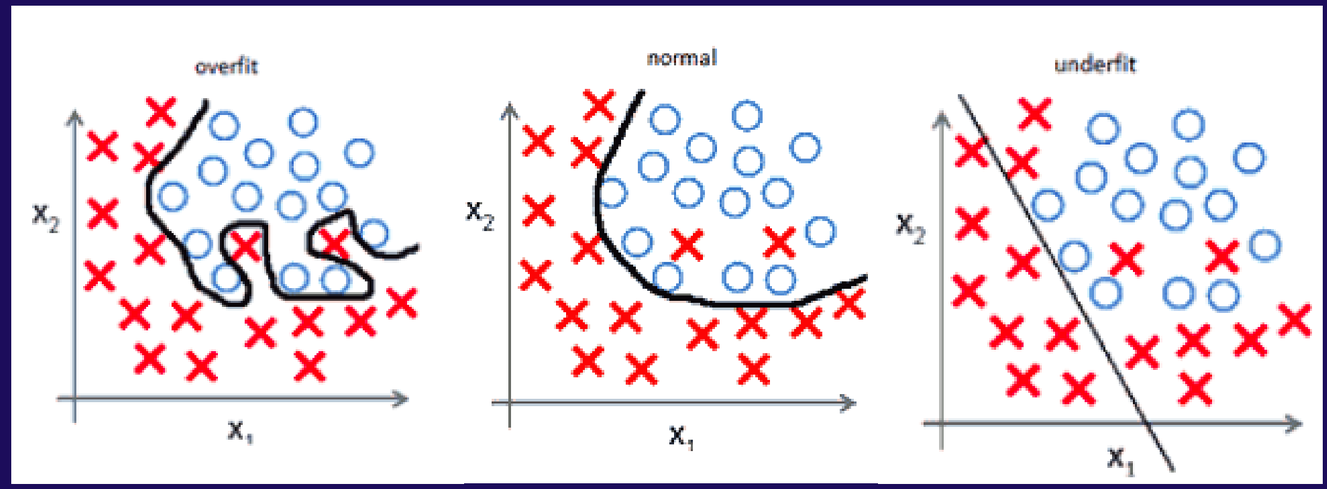
- **StandardScaler** segue a distribuição normal. Portanto, transforma os dados em média = 0 e variância unitária
- **MinMaxScaler** escala todas as variáveis no intervalo [0,1] ou [-1, 1], se tiver valores negativos

OVERFITTING (SOBREAJUSTE) E UNDERFITTING (SUBAJUSTE)

EXEMPLO NUMÉRICO (REGRESSÃO)



EXEMPLO CATEGÓRICO (CLASSIFICAÇÃO)



OVERFITTING (SOBREAJUSTE) E UNDERFITTING (SUB-AJUSTE)

O sobreajuste acontece quando apenas uma parte da amostra está disponível no treinamento, levando à sobregeneralização do todo. Com isso, ele aprende muito bem sobre os dados já coletados, mas não é eficaz em prever novos resultados.

Já o sub-ajuste ocorre quando há um erro de representação [também conhecido como **viés** (**bias**, em inglês)], em que tem-se os dados para o treinamento, porém o modelo não é adequado/tem baixo ajuste para avaliá-los.

REFERÊNCIAS

(CLIQUE NO LINK)

COELHO, Caíque. **Um guia completo para o pré-processamento de dados em machine learning**

GRUS, JOEL. Aprendizado de Máquina. In.: **Data Science do Zero: noções fundamentais com Python**. 2ª edição. Rio de Janeiro: Alta Books, 2021.

PARA SABER MAIS:

(CLIQUE NOS LINKS)

Entendendo de vez a diferença entre normalização e padronização dos dados @ Canal Ciência dos Dados

Trilha de Estudos para Cientista de Dados @ GitHub das Mulheres em Dados

OBRIGADA PELA PARTICIPAÇÃO!

POWERED BY MULHERES EM DADOS

- Obrigada por estar conosco neste projeto. Até o próximo!
- Não esqueça de responder o feedback **aqui**

Equipe Python:
Andressa Apio, Crislane Maria, Érika Santos e Joice Oliveira