DATA SCIENCE

AULA 3 - Pré-Processamento dos dados

Prof. Gabriel Resende Machado

gabrielmachado@unifeso.edu.com

https://www.linkedin.com/in/machadogabriel

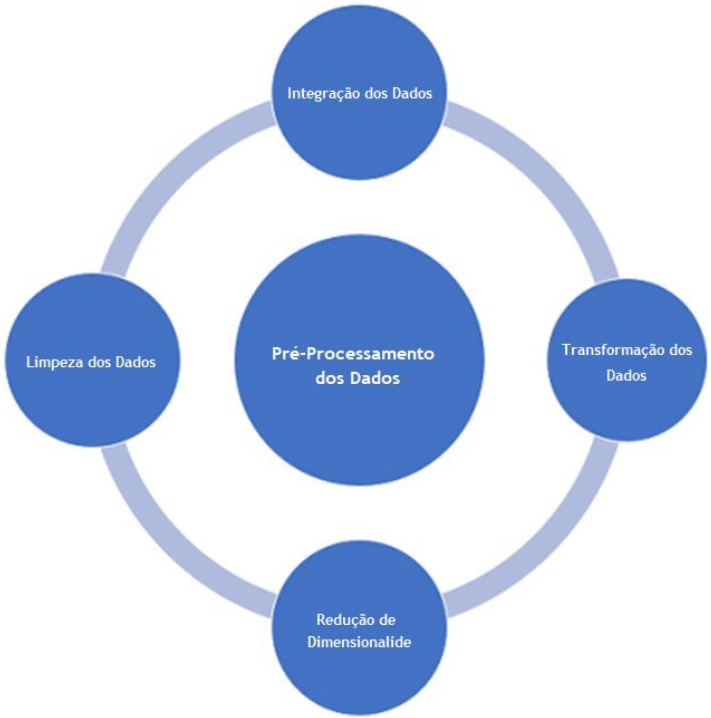
https://github.com/UNIFESO-Gabriel/data-science

O que é o pré-processamento dos dados?

- Conjunto de atividades referentes à (i) integração, (ii) limpeza, (iii) transformação e (iv) redução de dimensionalidade de **dados brutos** em **dados úteis**;

- Etapa **que consome até 80% do tempo** do *pipeline* de uma projeto de Ciência de Dados;

- Tem **grande importância** para a geração de dados de qualidade que servirão de insumo para modelos preditivos e analíticos.



Integração dos Dados

- Os dados podem estar organizados de três maneiras diferentes:
    - **Dados estruturados:** dados que seguem um esquema pré-definido e estão organizados de acordo com um modelo relacional. Geralmente são armazenados e administrados por um Sistema Gerenciador de Banco de Dados (SGBD);
    - **Dados semi-estruturados:** são dados que, geralmente, possuem uma estrutura hierárquica não compatível com os tradicionais modelos relacionais. Podem ser representados por arquivos de extensões .XML, .JSON e .HTML;
    - **Dados não-estruturados:** não possuem uma estrutura definida *a priori.* Podem ser representados por arquivos de texto, áudio e vídeo.
- Os principais meios de integração com os dados incluem: (i) *data warehouses*; (ii) *data lakes* e (iii) arquivos *.csv* e *.xlsx.*

Limpeza dos Dados

- Um *dataset* pode conter **vários registros ruidosos, irrelevantes ou ausentes**. A limpeza dos dados envolve o tratamento de **dados ausentes e *outliers***, como também a correção de **valores discrepantes, duplicados ou inconsistentes**.

- Algumas opções para tratamento desses valores incluem:

  - exclusão dos registros com valores anômalos;

  - substituição do valores anômalos pela média, mediana ou moda;

  - substituição do valores anômalos a partir do cálculo de um valor via regressão ou agregação.

Transformação dos Dados

- Esta etapa do pré-processamento dos dados tem como objetivo transformar os valores originais do *dataset* em valores mais adequados para o processo de mineração de dados.

- Algumas opções para tratamento incluem:

    - **Normalização dos valores** em intervalos entre 0 e 1 (normalização *min/max*) ou -1 a 1 (padronização)**;**

    - **seleção de atributos mais relevantes;**

    - criação de novos atributos via métodos de **engenharia de atributos**;

    - exclusão de atributos menos relevantes ou mais correlacionados via métodos de **redução de dimensionalidade**.

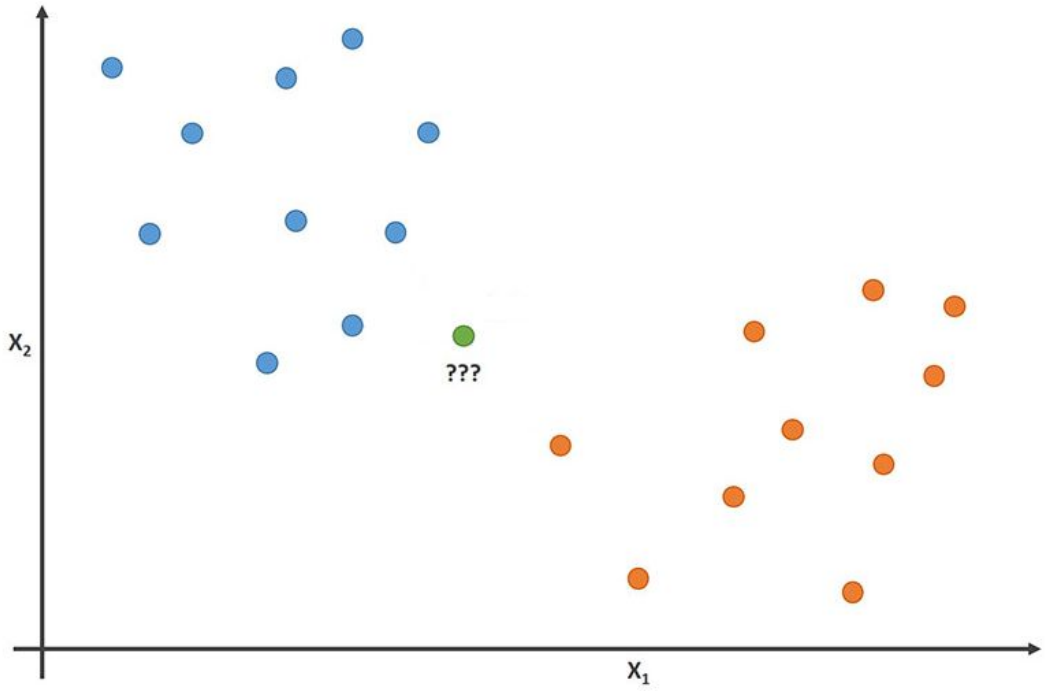
Tipos de Conjuntos de Dados

- Um conjunto de dados pode **ser** ou **não ser rotulado**.
  - conjunto de dados rotulados representam a categoria de **problemas supervisionados**;
  - conjunto de dados não rotulados representam a categoria de **problemas não-supervisionados**.
- Problemas supervisionados podem ser de **classificação** ou **regressão**.
  - Problemas de classificação visam gerar um modelo preditivo que, após a etapa de treinamento, atribui um rótulo (ou classe) a uma determinada amostra. **São de natureza discreta**;
  - classes com uma quantidade predominante de registros são conhecidas como **classes desbalanceadas;**
  - Problemas de regressão visam gerar um modelo preditivo que atribui um valor real a uma nova amostra. **São de natureza contínua**.
- Problemas não-supervisionados podem ser do tipo **agregação, clusterização** e **associação**.

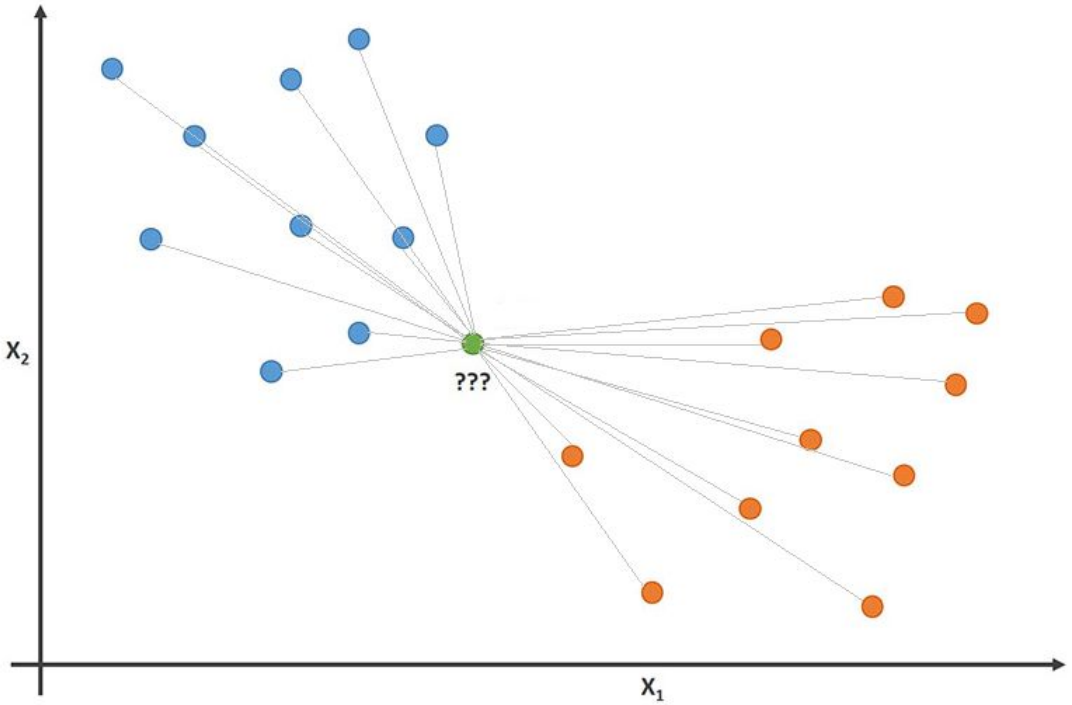
k-Nearest Neighbors

- Também conhecido como **k-NN**, é um algoritmo de aprendizagem supervisionado simples que utiliza cálculos de proximidade para classificar um novo registro;
- Pode ser utilizado em tarefas de **classificação** e **regressão**, sendo mais utilizado em **tarefas de classificação**;
- **Em tarefas de classificação**, é atribuído uma classe a um novo registro a partir de **voto majoritário**. São analisadas ***k* amostras** e a classe majoritariamente presente ao redor do registro em análise é atribuída ao novo registro;
- **Em tarefas de regressão**, é calculado o valor médio dos vizinhos e atribuído à nova amostra;
- A medida de distância mais comumente utilizada é a **distância euclidiana;**
- Recomenda-se utilizar um **valor ímpar** para *k* de modo a evitar empates no processo de votação.

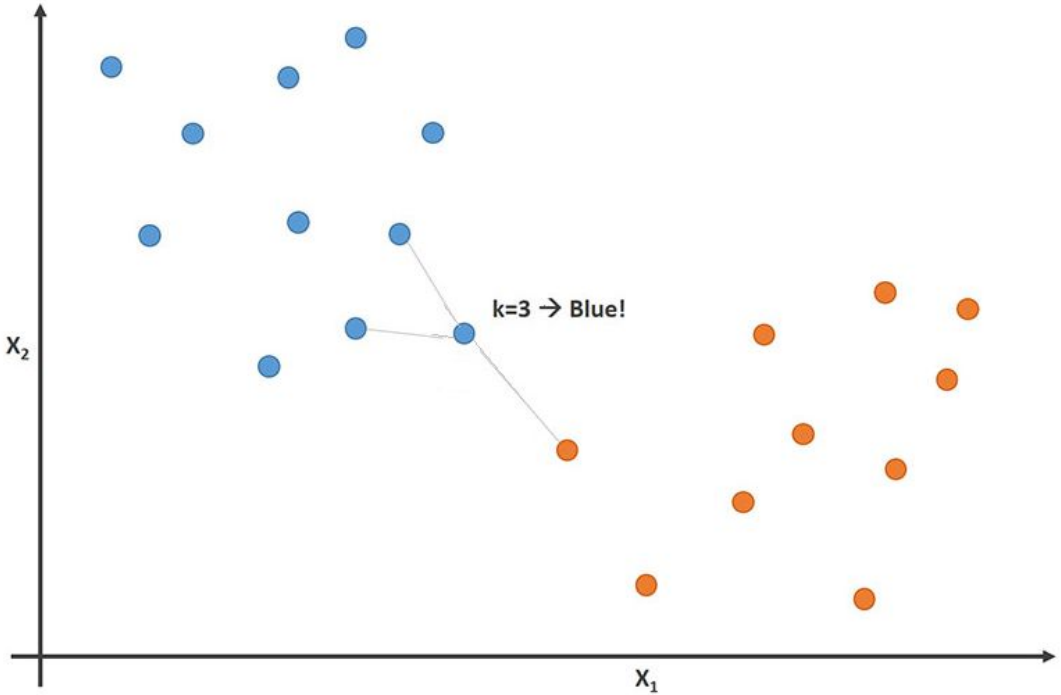
k-Nearest Neighbors (k = 3)



k-Nearest Neighbors (*k* = 3)



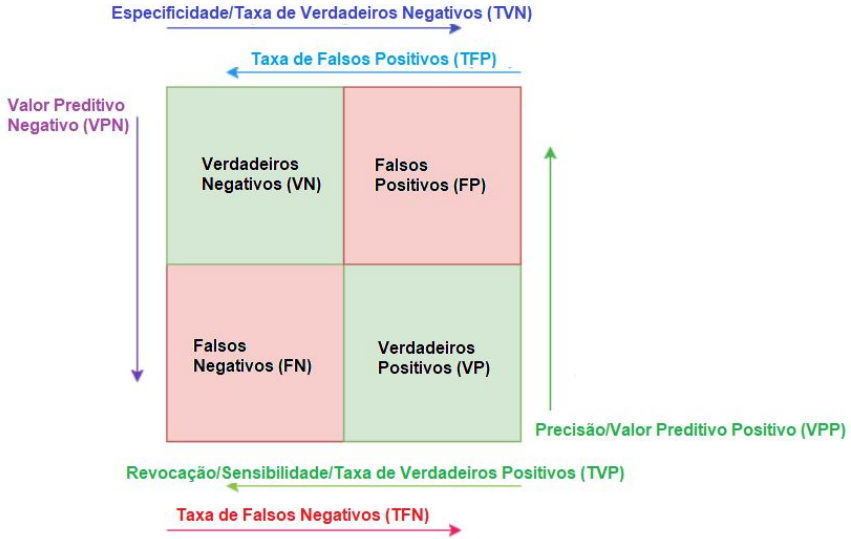
k-Nearest Neighbors (k = 3)



Métricas de Desempenho

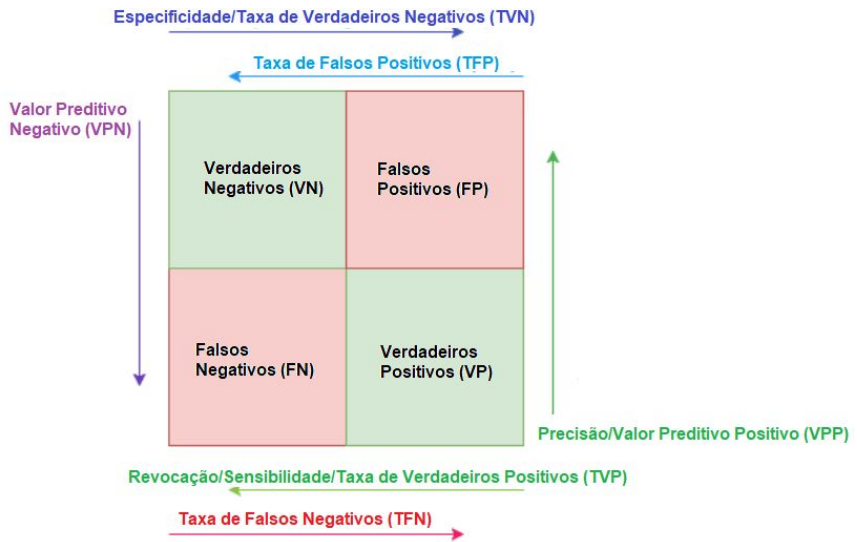
- Em tarefas de classificação, é necessário verificar o desempenho dos modelos em prever as classes de novas amostras;

- A **matriz de confusão** é comumente utilizada para avaliar modelos de classificação a partir das saídas dos modelos de amostras rotuladas.



Métricas de Desempenho

- **Acurácia:** quantidade de amostras classificadas corretamente: **(VN + VP) / (VN + VP + FN + FP).** Suscetível a problemas de classificação desbalanceados;
- **Precisão:** a razão dos verdadeiros positivos pelo total de saídas preditas como positivas: **VP / (VP + FP);**
- **Revocação:** a razão da quantidade de registros preditos como positivos pela quantidade de registros positivos: **VP / (VP + FN);**
- **Métrica F1:** média harmônica entre a precisão e revocação: **(2 * precisão * revocação) / (precisão + revocação).**

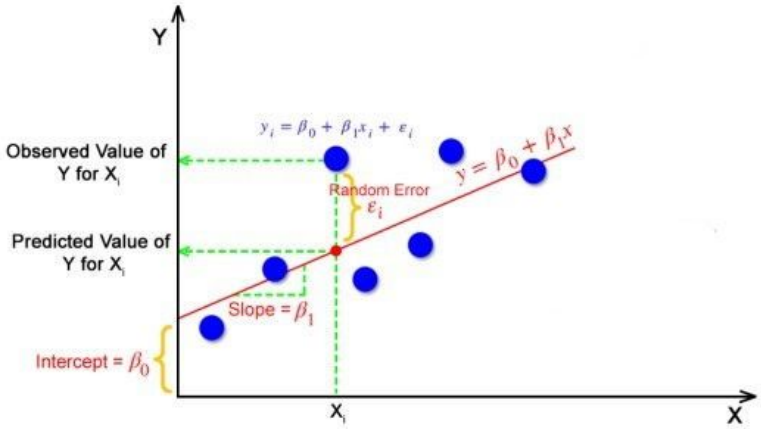


Regressão Linear

- Modelo de aprendizagem supervisionado que assume um relacionamento linear entre a variável independente e a variável dependente;
- Busca ajustar a melhor linha que **minimize a soma do quadrado das diferenças** entre os valores preditos e os valores reais;
- A regressão linear calcula os **coeficientes** e um **interceptador**. Os coeficientes determinam a inclinação da reta enquanto que o interceptador representa o valor predito para a variável dependente quando a variável independente é zero.

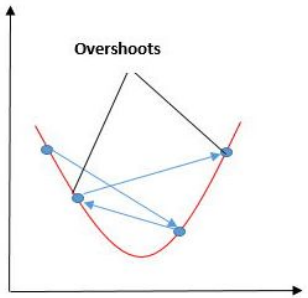
$$y = \beta_0 + \beta_1 x$$

  - y: variável dependente;
  - β_0: interceptador;
  - $\beta_1 x$: inclinação da reta/variável independente.

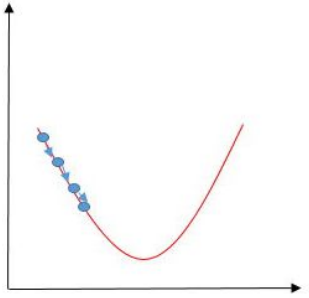


Aprendizado da Regressão Linear

- **Objetivo:** encontrar valores para β_0 e β_1 de modo que os erros residuais sejam minimizados;
- Para isso, é utilizada uma **função de custo**. A função de custo ajuda a identificar quanto o modelo está aprendendo a partir do cômputo dos parâmetros β_0 e β_1.
- A função de custo mais utilizada no aprendizado de uma regressão linear é o **Erro Quadrático Médio (MSE)**: $MSE = \frac{1}{2m}\sum_{i=1}^{m}(\hat{y}^i - y^i)^2$
- Os parâmetros β_0 e β_1 são atualizados a partir do cálculo do **gradiente descendente.**
- **O parâmetro α representa a taxa de aprendizado.**
  - $\beta_0 = \beta_0 - \alpha * \frac{2}{m}(\sum_{i=1}^{m}\hat{y}^i - y^i)$
  - $\beta_1 = \beta_1 - \alpha * \frac{2}{m}(\sum_{i=1}^{m}\hat{y}^i - y^i) * x^i$



Taxa de aprendizado alta


Taxa de aprendizado baixa

DATA SCIENCE

AULA 3 - Pré-Processamento dos dados

Dúvidas e/ou perguntas?

gabriel.rmachado10@gmail.com

https://www.linkedin.com/in/machadogabriel

https://github.com/UNIFESO-Gabriel/data-science