

Mineração de dados e Aprendizado de Máquina

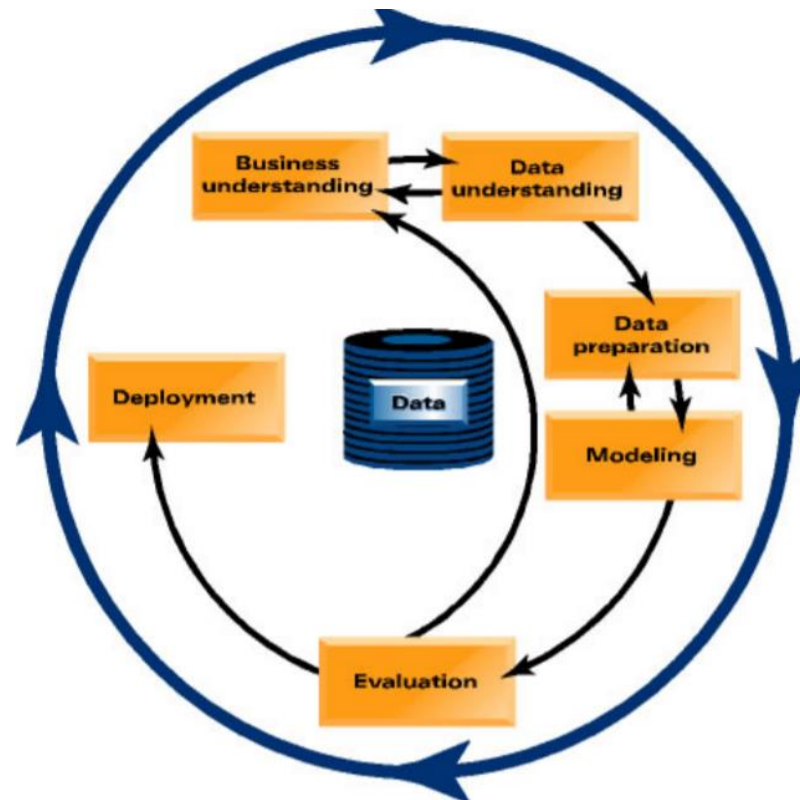
Overview a Descoberta de Conhecimento em Bases de Dados

CONCEITOS FUNDAMENTAIS

- Extrair conhecimento útil dos dados para resolver os negócios problemas podem ser tratados sistematicamente seguindo um processo com etapas razoavelmente bem definidas.
- Formular soluções de mineração de dados e avaliar o resultados envolve pensar cuidadosamente sobre o contexto em quais eles serão usados.”

PROCESSO DE MINERAÇÃO DE DADOS

- Cross Industry Standard Process for Data Mining – CRISP-DM
- Processo padrão entre indústrias para mineração de dados



Por que enfatizar o processo ?

- 1. Rigor - cada estágio do processo geralmente é suportado por princípios teóricos bem pesquisados ou heurísticas bem documentadas
- 2. Confiabilidade - as tarefas de mineração de dados resistem melhor à revisão por pares e gerencial quando as tarefas aderem aos princípios e padrões fundamentais do processo comum
- 3. Reprodutibilidade - com um processo bem definido, podemos replicar melhor os resultados e também automatizar determinadas tarefas de aprendizado

PERGUNTAS COMUNS NAS ORGANIZAÇÕES

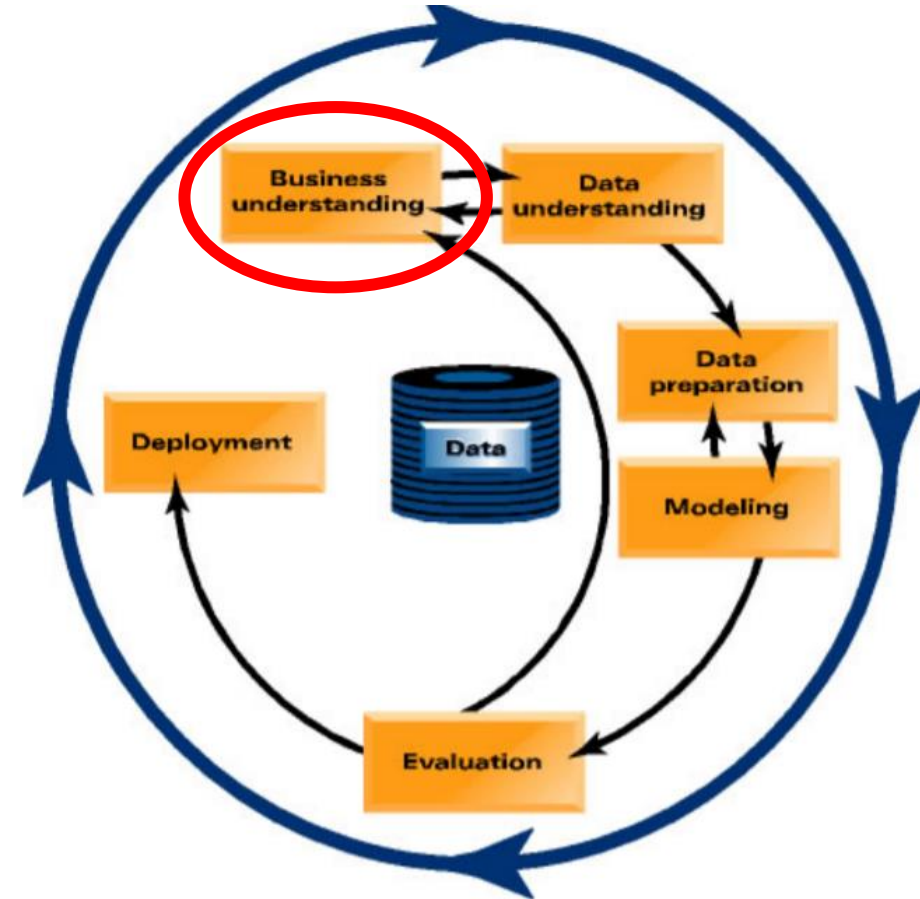
- O cliente X cancelará no próximo mês / inadimplência em seu empréstimo?
- Quanto o cliente em potencial X gastaria se fosse um cliente?
- Quem podem ser bons "amigos" em nosso site de redes sociais?
- X causou Y acontecer?
- O que você deve recomendar ao usuário I.
- Os usuários se enquadram em grupos únicos?
- Esta transação é fraudulenta?

FORMULAÇÃO DE PROBLEMAS - TRADUÇÃO

- Os cientistas de dados falam um idioma diferente e você precisa ser capaz de traduzir. Isso significa formular objetivos de negócios na linguagem da ciência de dados.

COMPREENSÃO DO PROJETO / NEGÓCIOS

- Coloque o problema em contexto
... faça perguntas ... seja criativo!
- Qual é o objetivo da solução?
- Por que precisamos fazer isso?
- Quais dados estão disponíveis?
- Quais restrições existem?
- O que é uma solução aceitável?
- Como medimos?
- O que é sucesso?

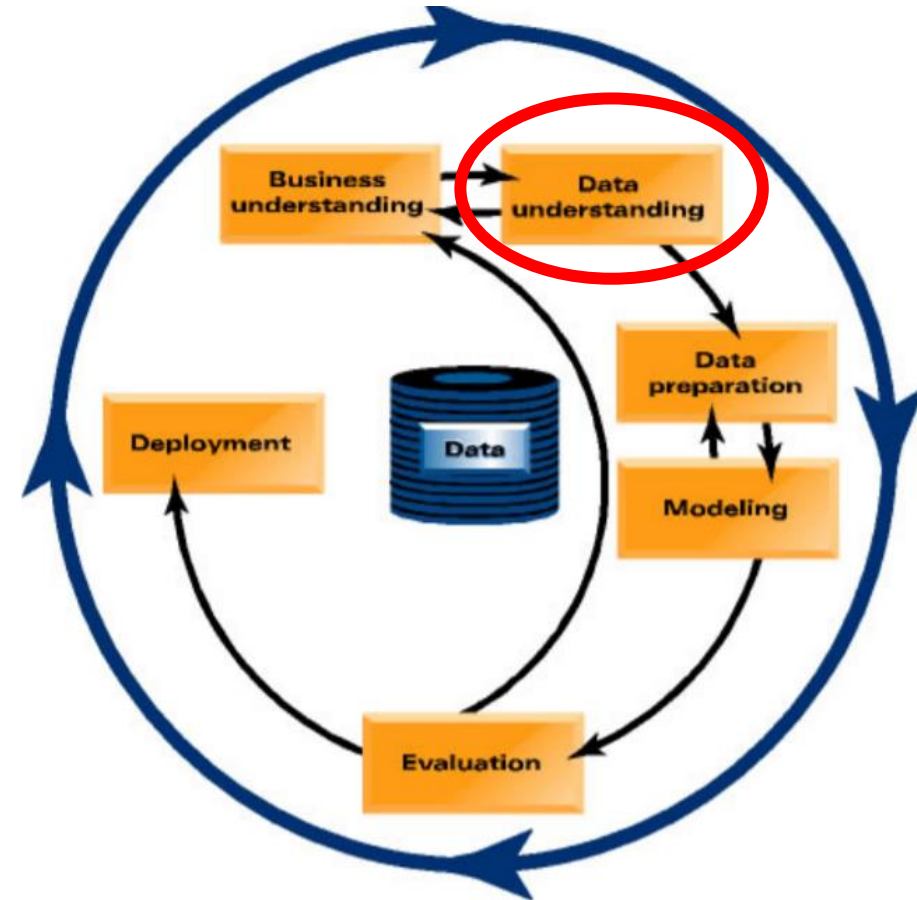


COMPREENSÃO DOS DADOS

Claramente o tópico mais importante até agora...

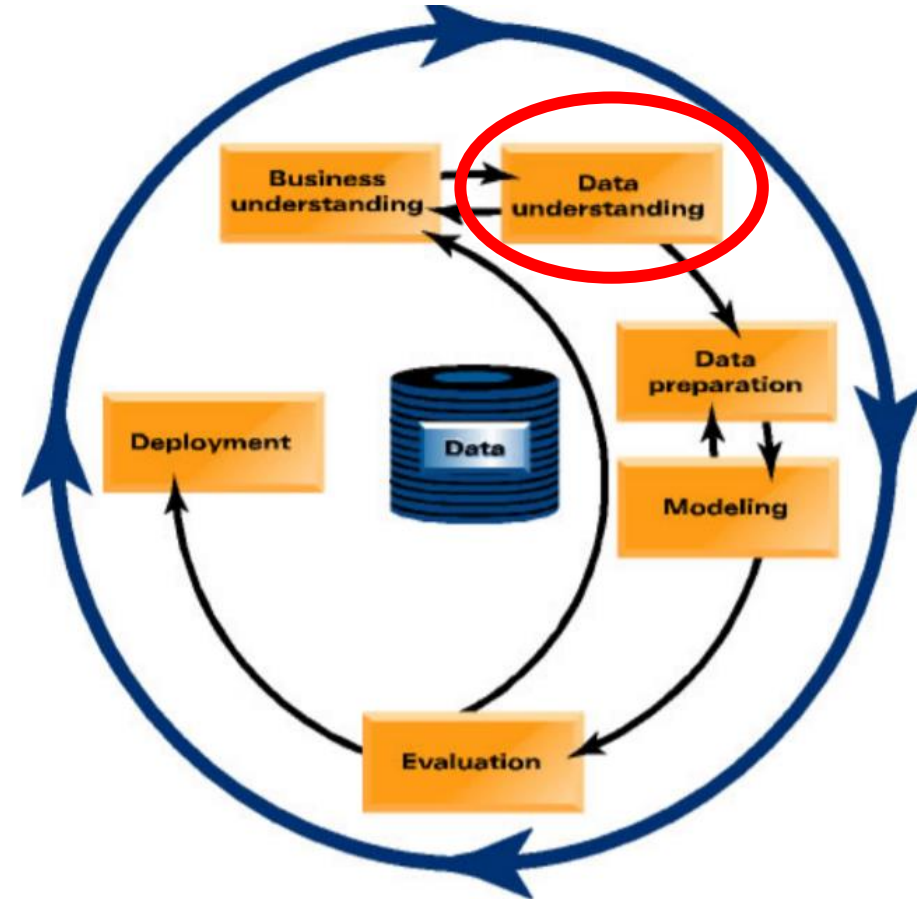
Regras de ouro

- 1.Saiba de onde vêm seus dados.
- 2.Saiba como obter os dados.
- 3.Saiba como são os seus dados.
- 4.Conheça os limites dos seus dados.



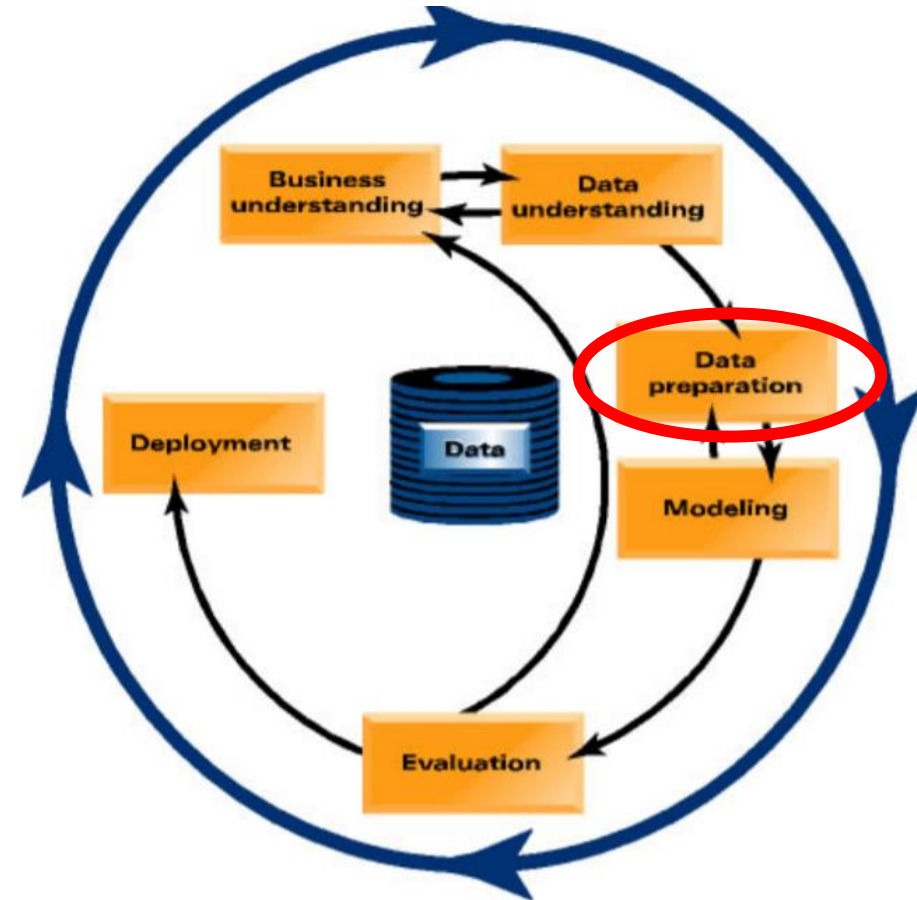
ANÁLISE NO MUNDO DE "GRANDES DADOS"

- A onipresença dos dados vem com um sinal de captura espúrio por toda parte. Ignorar os princípios básicos do método científico pode levar a maus resultados e até prejudicar. Não cometa esse erro ...
- 1. Olhe os dados espalhados
- 2. Faça uma hipótese
- 3. Falsificar / confirmar a hipótese nos mesmos dados
- 4. Tomar decisões que levem a uma generalização



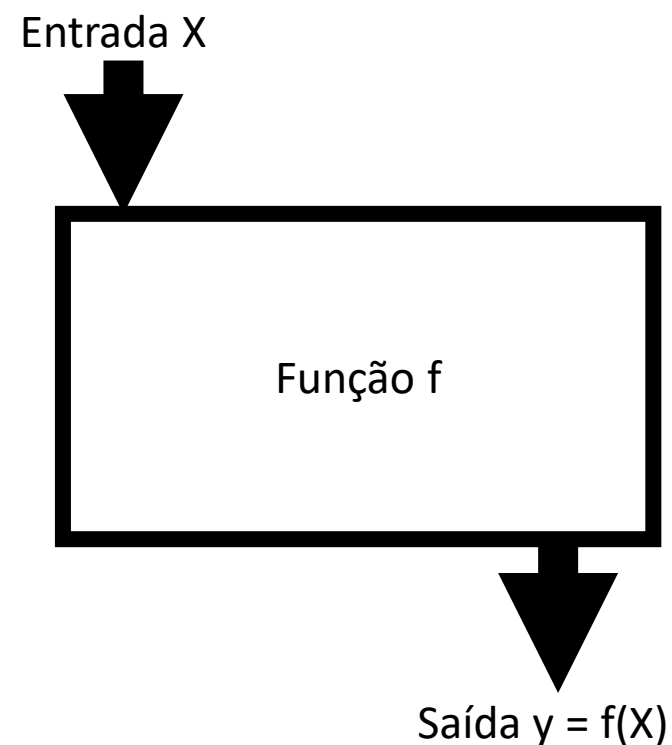
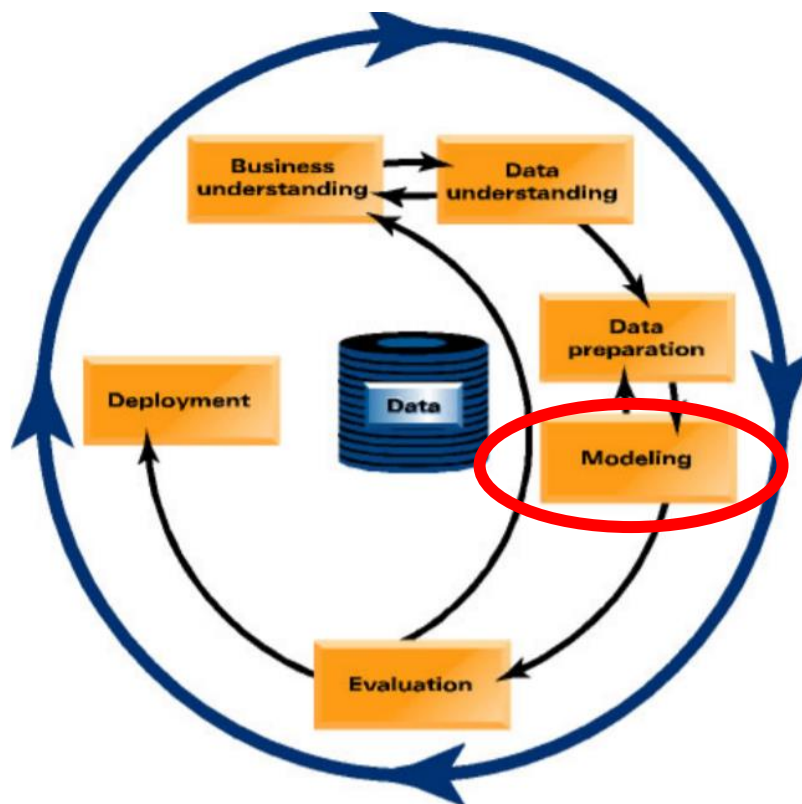
PREPARAÇÃO DOS DADOS

- A onipresença dos dados não quer dizer que eles estejam prontos para serem utilizados
- Problemas como formatação, dados faltantes, preenchimentos errados ou duvidosos podem envolver até 80% do tempo em um projetos de ciência de dados



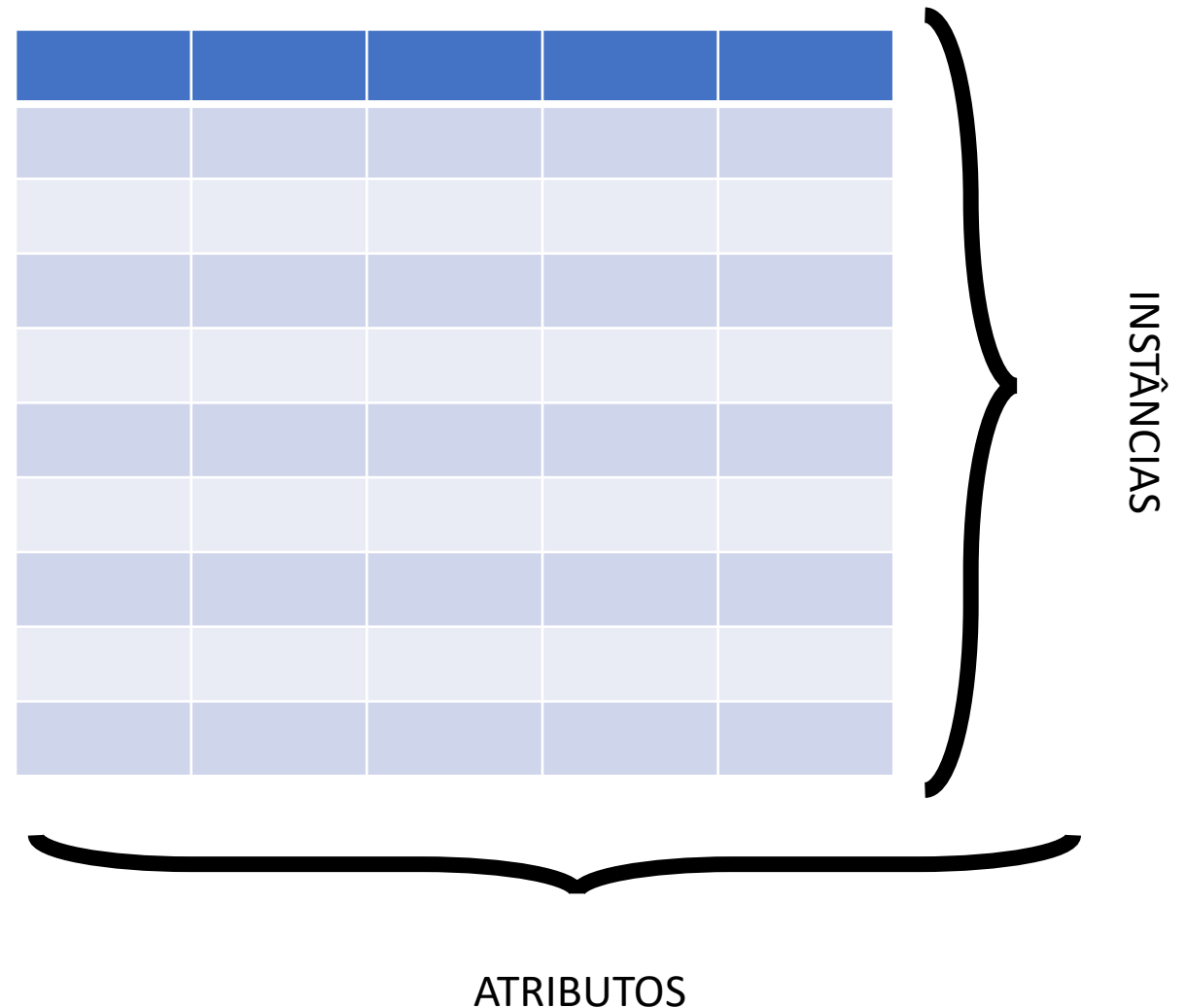
MODELAGEM

- O mecanismo da ciência de dados.
- Modelagem é como você obtém dados, insights e tomada de decisão.
- Abordaremos como isso é feito extensivamente neste curso.



A Estrutura de Dados de um Modelo

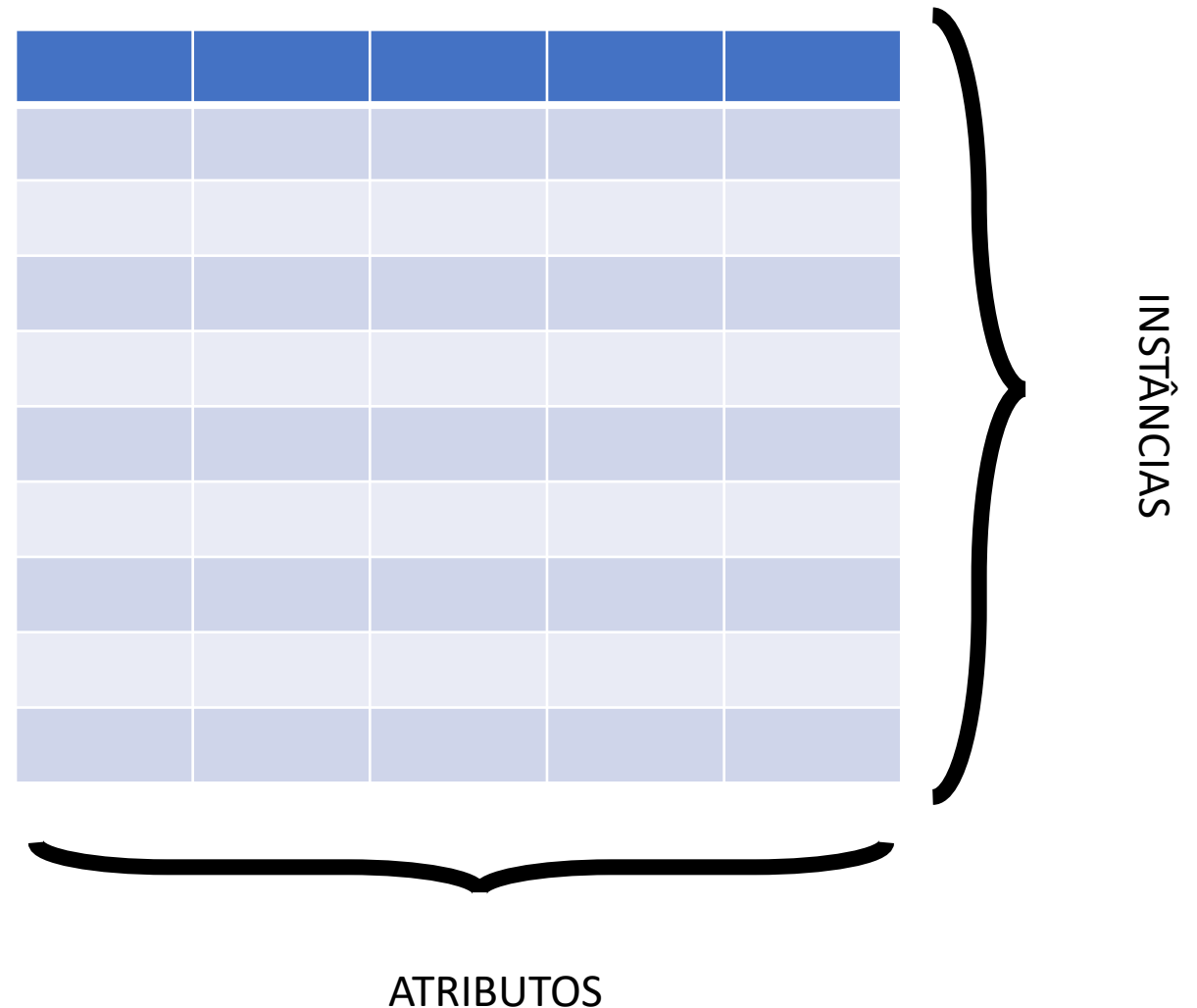
- Em geral a estrutura de dados de um modelo é entendida na forma de uma matriz (uma tabela)
- As linhas são chamadas instâncias
- As colunas são chamadas atributos



A Estrutura de Dados de um Modelo

ATRIBUTOS

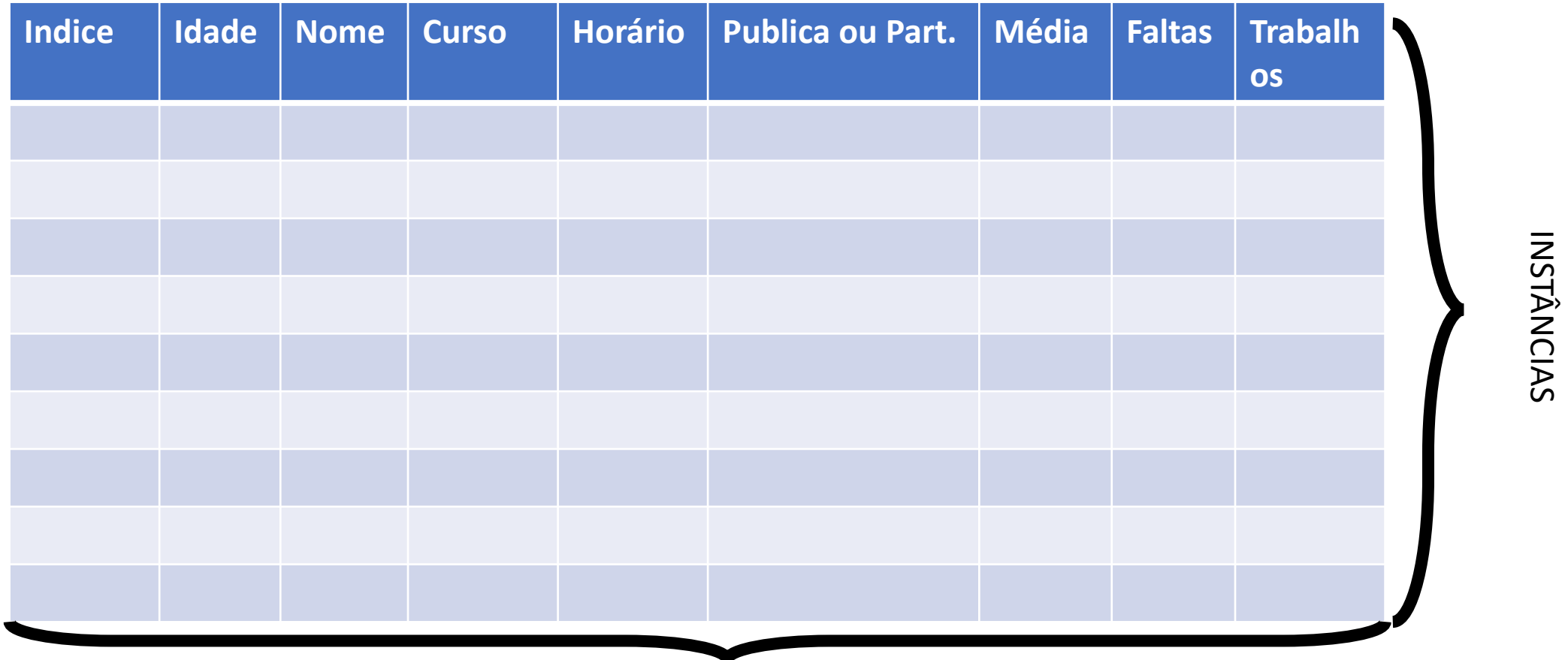
- Os atributos podem ser entendidos como as características que temos disponíveis para entender cada um dos dados de nosso conjunto de dados ou dataset
- Por exemplo num conjunto de dados ALUNOS suas características seriam:
 - índice
 - Idade
 - Nome
 - Local de nascimento
 - Curso
 - Horário
 - Se veio de uma escola publica ou particular
 - Média
 - Faltas
 - Número de trabalhos entregues
 - Etc..



A Estrutura de Dados de um Modelo

ATRIBUTOS

Indice	Idade	Nome	Curso	Horário	Publica ou Part.	Média	Faltas	Trabalhos



ATRIBUTOS

INSTÂNCIAS

A Estrutura de Dados de um Modelo

INSTÂNCIAS

- Os atributos podem ser entendidos como as características que temos disponíveis para entender cada um dos dados de nosso conjunto de dados ou dataset
- Por exemplo num conjunto de dados PARA CADA ALUNO teríamos as instâncias:
 - Índice [0,1,2,3,4,5,6,7,8,...]
 - Idade [21,22,31,27, 40, 49, 31, 22, ETC..]
 - Nome [JOÃO, JOSÉ, MARIA, CLAUDIO, TOMÁS, JACÓ, IRINEU,..]
 - Local de nascimento [São Paulo, Presidente Prudente, Osasco, Guarulhos, São Paulo, São Bernado, etc..]
 - Curso [CC, TADS, SI, CC, TADS, SI...]
 - Horário [D, N, D, N, N, N , N, etc...]
 - Se veio de uma escola publica ou particular [Pb, Pa, Pb, Pb, PB, Pa, Pb, etc...]
 - Média [10, 9, 9.5, 7, 7.5, 6.6, 8.1, etc..]
 - Faltas [2, 3 ,4 ,2 ,2 3 ,3 . 10, 3, 5, etc..]
 - Número de trabalhos entregues [10,8,10,9, 6, 5, 6, 8,etc..]
 - Etc..

The diagram illustrates a data matrix with 5 columns and 10 rows. The columns are labeled 'ATRIBUTOS' and the rows are labeled 'INSTÂNCIAS'. The matrix is divided into two groups of 5 rows each, each group labeled 'INSTÂNCIAS'.

Este exemplo completo

Indice	Idade	Nome	Curso	Horário	Publica ou Part.	Média	Faltas	Trabalho s
0	27	Artur	CC	D	PB	6,5	3	10
1	20	Maria	TADS	D	PT	8,9	2	7
2	35	Antonio	TADS	N	PB	7,9	4	9
3	47	Cleber	CC	N	PT	6,2	4	10
4	46	Silas	SI	N	PB	6,3	2	6
5	30	Mateus	TADS	N	PB	6,4	4	7
6	40	Rafael	TADS	N	PB	7,0	5	8
7	20	Edson	SI	N	PT	5,9	2	5
8	41	Ana	SI	N	PB	5,9	3	6
9	28	Paula	SI	N	PB	9,9	2	8

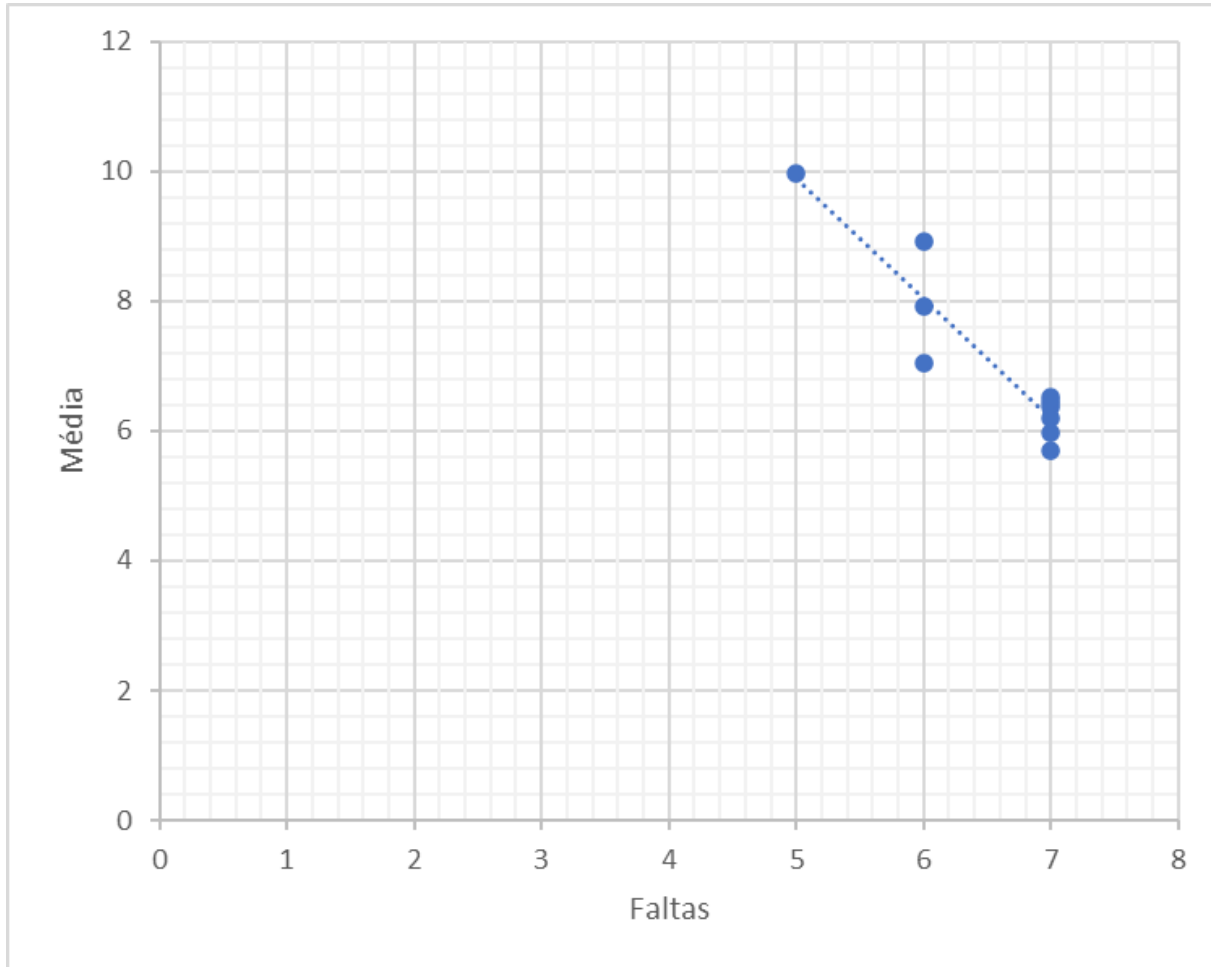
Colunas como variáveis para os modelos

- Por exemplo posso estar interessado se existe uma relação entre Média e Faltas
- Então eu seleciono estas duas colunas e desenvolvo um modelo que vai avaliara esta relação

Indice	Média
0	6,5
1	8,9
2	7,9
3	6,2
4	6,3
5	6,4
6	7,0
7	5,9
8	5,9
9	9,9

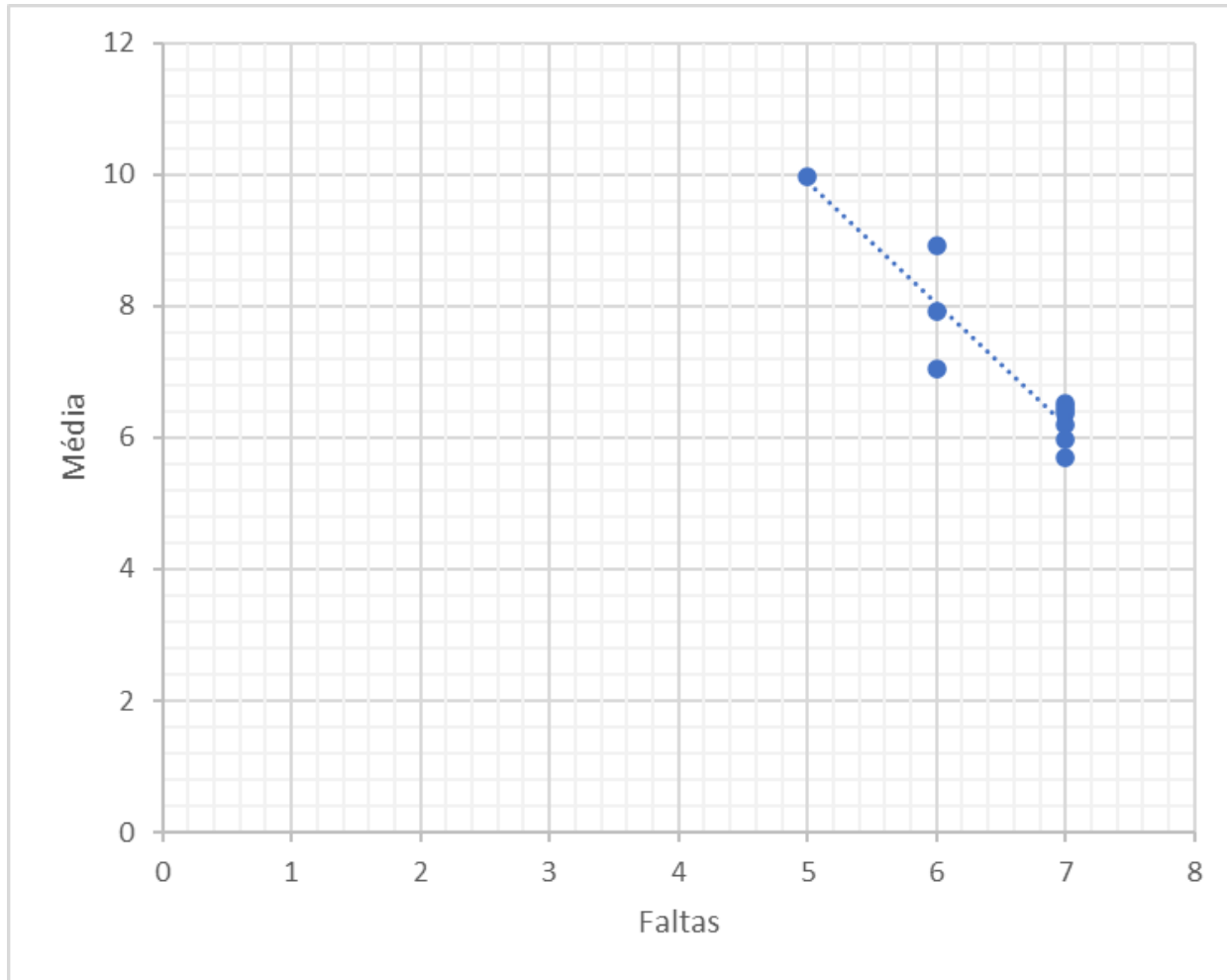
Indice	Faltas
0	3
1	2
2	4
3	4
4	2
5	4
6	5
7	2
8	3
9	2

Um primeiro modelo



- Temos um Gráfico onde X é o número de faltas dos alunos e y é a média
- O que podemos dizer ?

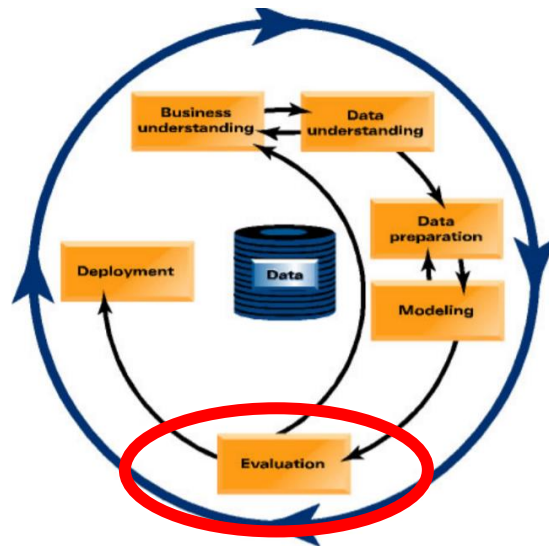
Um primeiro modelo



- Temos um Gráfico onde X é o número de faltas dos alunos e y é a média
- O que podemos dizer ?
- Que quando as faltas aumentam a média diminui
- Portanto nosso primeiro modelo $y = f(X)$ é:
- $Média = f(Faltas)$

Avaliação Técnicas de Avaliação de Modelos

- Diversas questões podem ser tratadas avaliando o desempenho de um modelo de aprendizado de máquina, que é um componente integrante de qualquer projeto de ciência de dados.
- A avaliação do modelo visa estimar a precisão da generalização de um modelo em dados futuros (invisíveis / fora da amostra).



- Os métodos para avaliar o desempenho de um modelo são divididos em 2 categorias: validação cruzada e validação.
- Ambos os métodos usam um conjunto de testes (ou seja, dados não vistos pelo modelo) para avaliar o desempenho do modelo.
- Não é recomendável usar os dados que usamos para criar o modelo para avaliá-lo.
- Isso ocorre porque nosso modelo simplesmente se lembra de todo o conjunto de treinamento e, portanto, sempre prediz o rótulo correto para qualquer ponto do conjunto de treinamento.

A Estrutura de Dados de um Modelo

Indice	Idade	Nome	Curso	Horário	Publica ou Part.	Média	Faltas	Trabalhos

INSTÂNCIAS PARA TREINAMENTO

INSTÂNCIAS PARA TESTES

ATRIBUTOS SELECIONADOS

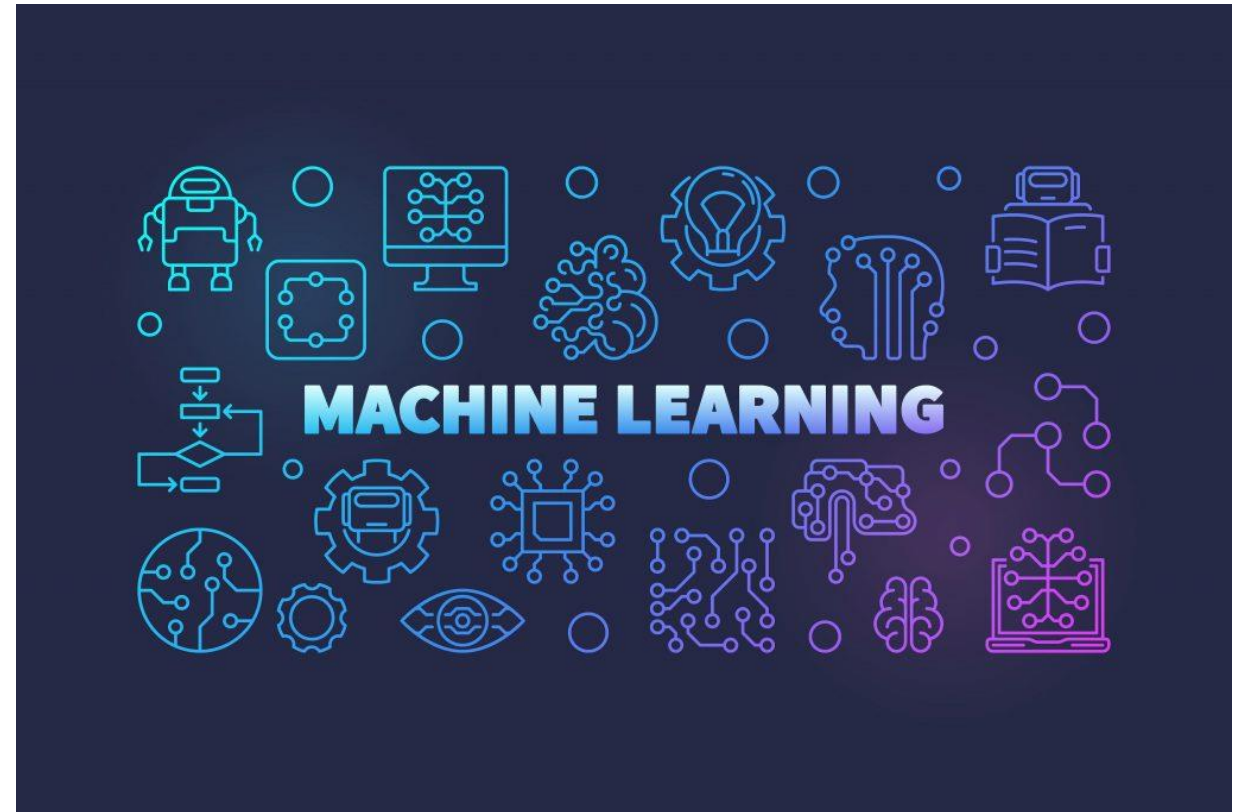
Implementação



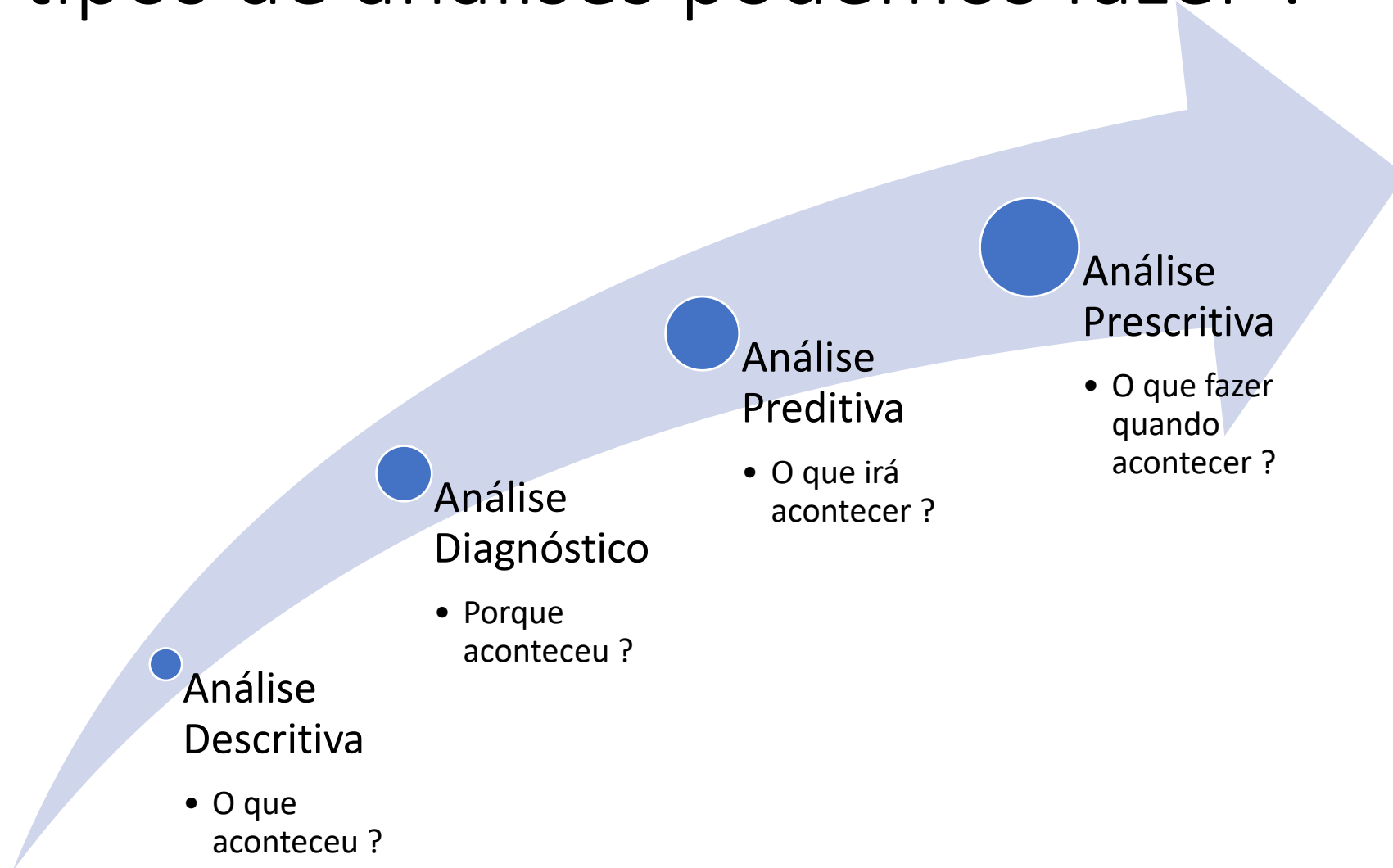
- Muitas vezes o modelo é implementado em linguagens diferentes de seu desenvolvimento
- Você pode atuar como desenvolvedor ou como um “consultor” para ajudar os engenheiros de machine learning a implementar o modelo em outra linguagem

Vamos entender os tipos de modelos

- Para cada projeto podem existir diversos tipos de modelos necessários
- É importante então entendermos as tarefas que podem ser realizadas com os algoritmos de machine learning
- Ao longo das aulas vamos ver diferentes tipo de modelos e sua aplicação a cada uma destas tarefas
- Também vamos estudar mais a fundo as tarefas



Que tipos de análises podemos fazer ?



KDD x Data Mining

- Mineração de dados é o passo do processo de KDD que produz um conjunto de padrões sob um custo computacional aceitável;
- KDD utiliza algoritmos de *data mining* para extrair padrões classificados como “conhecimento”. Incorpora também tarefas como escolha do algoritmo adequado, processamento e amostragem de dados e interpretação de resultados;

Descrição (Description)

- É a tarefa utilizada para descrever os padrões e tendências revelados pelos dados.
- A descrição geralmente oferece uma possível interpretação para os resultados obtidos.
- A tarefa de descrição é muito utilizada em conjunto com as técnicas de análise exploratória de dados, para comprovar a influência de certas variáveis no resultado obtido.

Classificação (Classification)

- Uma das tarefas mais comuns, a Classificação, visa identificar a qual classe um determinado registro pertence.
- Nesta tarefa, o modelo analisa o conjunto de registros fornecidos, com cada registro já contendo a indicação à qual classe pertence, a fim de 'aprender' como classificar um novo registro (aprendizado supervisionado).
- Por exemplo, categorizamos cada registro de um conjunto de dados contendo as informações sobre os colaboradores de uma empresa:
 - Perfil Técnico,
 - Perfil Negocial e
 - Perfil Gerencial.
- O modelo analisa os registros e então é capaz de dizer em qual categoria um novo colaborador se encaixa.
- A tarefa de classificação pode ser usada por exemplo para:
 - Determinar quando uma transação de cartão de crédito pode ser uma fraude;
 - Identificar em uma escola, qual a turma mais indicada para um determinado aluno;
 - Diagnosticar onde uma determinada doença pode estar presente;
 - Identificar quando uma pessoa pode ser uma ameaça para a segurança.

Estimação (Estimation) ou Regressão (Regression)

- A estimação é similar à classificação, porém é usada quando o registro é identificado por um valor numérico e não um categórico.
- Assim, pode-se estimar o valor de uma determinada variável analisando-se os valores das demais.
- Por exemplo, um conjunto de registros contendo os valores mensais gastos por diversos tipos de consumidores e de acordo com os hábitos de cada um.
- Após ter analisado os dados, o modelo é capaz de dizer qual será o valor gasto por um novo consumidor.
- A tarefa de estimação pode ser usada por exemplo para:
 - Estimar a quantia a ser gasta por uma família de quatro pessoas durante a volta às aulas;
 - Estimar a pressão ideal de um paciente baseando-se na idade, sexo e massa corporal.

Predição (Prediction)

- A tarefa de predição é similar às tarefas de classificação e estimação, porém ela visa descobrir o valor futuro de um determinado atributo. Exemplos:
- Predizer o valor de uma ação três meses adiante;
- Predizer o percentual que será aumentado de tráfego na rede se a velocidade aumentar;
- Predizer o vencedor do campeonato baseando-se na comparação das estatísticas dos times.
- Alguns métodos de classificação e regressão podem ser usados para predição, com as devidas considerações.

Agrupamento (Clustering)

- A tarefa de agrupamento visa identificar e aproximar os registros similares.
- Um agrupamento (ou cluster) é uma coleção de registros similares entre si, porém diferentes dos outros registros nos demais agrupamentos.
- Esta tarefa difere da classificação pois não necessita que os registros sejam previamente categorizados (aprendizado não-supervisionado).
- Além disso, ela não tem a pretensão de classificar, estimar ou prever o valor de uma variável, ela apenas identifica os grupos de dados similares.

Exemplos:

- Segmentação de mercado para um nicho de produtos;
- Para auditoria, separando comportamentos suspeitos;

Associação (Association)

- A tarefa de associação consiste em identificar quais atributos estão relacionados. Apresentam a forma: SE atributo X ENTÃO atributo Y.
- É uma das tarefas mais conhecidas devido aos bons resultados obtidos, principalmente nas análises da "Cestas de Compras"(Market Basket), onde identificamos quais produtos são levados juntos pelos consumidores.
- Alguns exemplos:
- Determinar os casos onde um novo medicamento pode apresentar efeitos colaterais;
- Identificar os usuários de planos que respondem bem a oferta de novos serviços.

O modelo CRISP DM

