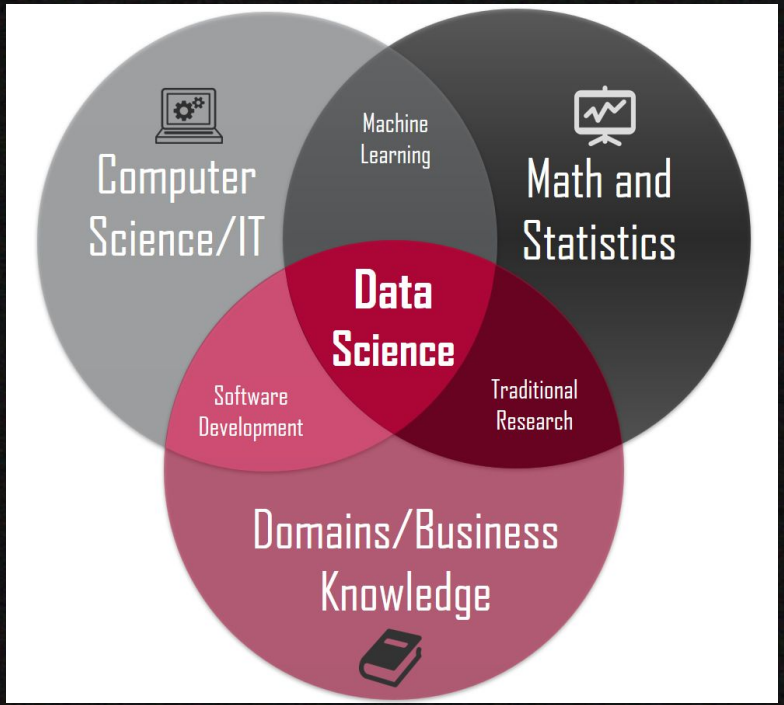
Semana de Data Science

O que é Data Science?

A ciência que tem o objetivo de extrair valor dos dados.



Processo de Data Science

Coleta de Dados

Limpeza e Transformação

Análise e Exploração

Criação de Modelos

Interpretação de Resultados

Previsão do Valor do Imóvel

Preços dos imóveis na cidade de Boston EUA

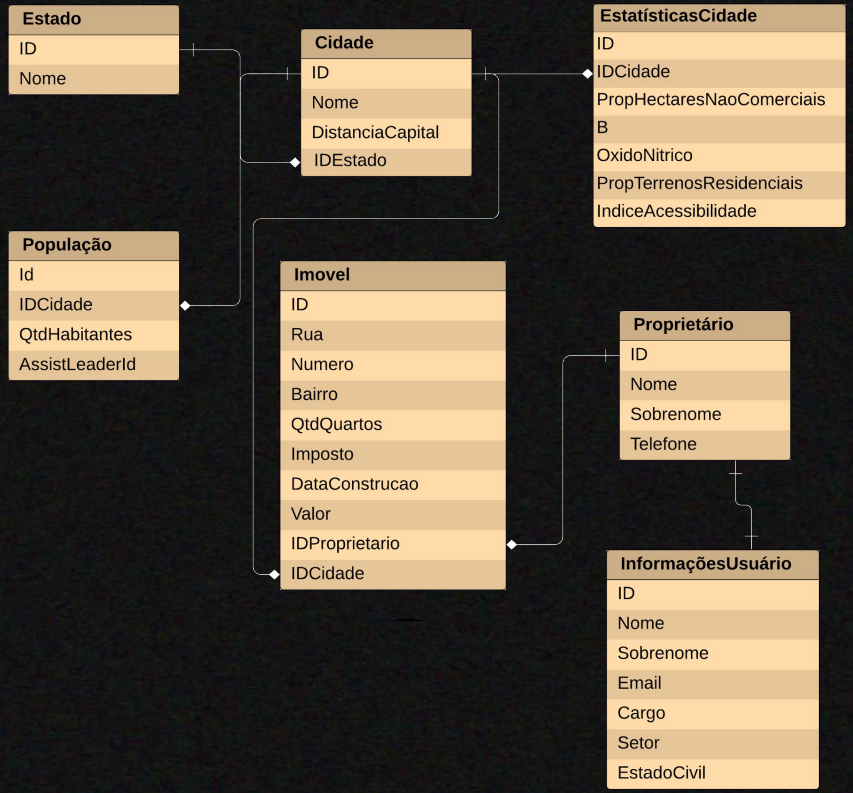
Previsão do Valor do Imóvel

Baseado nas características do imóvel o objetivo é estimar o preço do imóvel.

Previsão do Valor do Imóvel

Diagrama Entidade Relacionamento

Felipe Santana | May 6, 2020



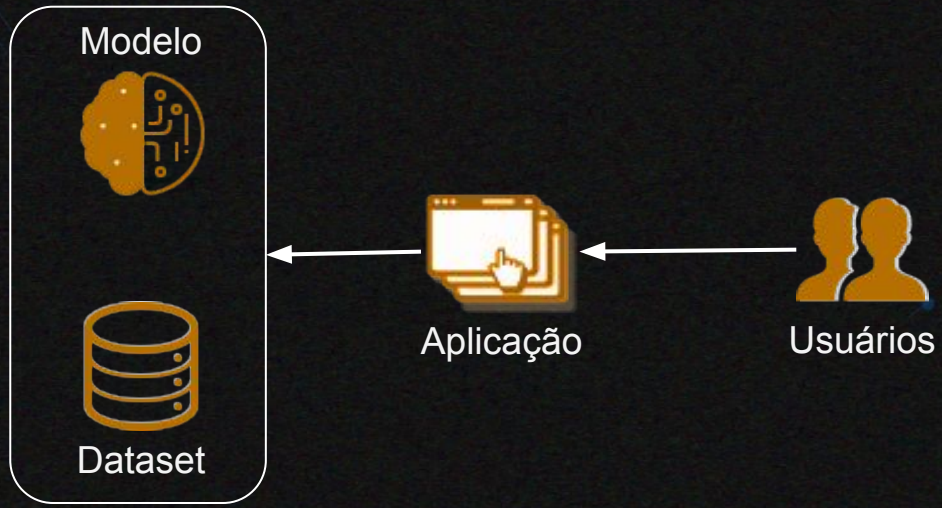
Previsão do Valor do Imóvel

Imoveis
PropHectaresNaoComerciais
PropPessoasDescAfro
OxidoNitrico
PropTerrenosResidenciais
IndiceAcessibilidade
Valor
TempoImovel
LimiteRio
QtdQuartos
Imposto

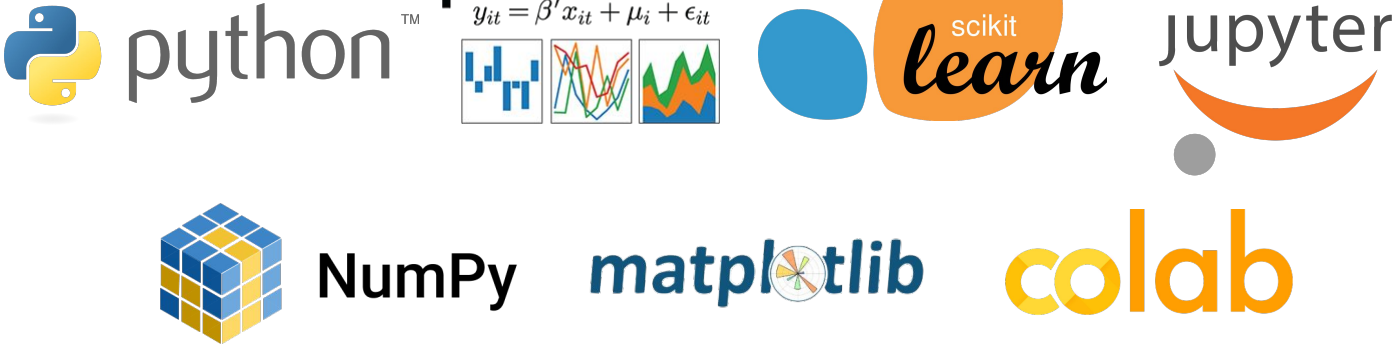
Arquitetura da Solução

Banco de Dados Aplicação Usuários

Arquitetura da Solução



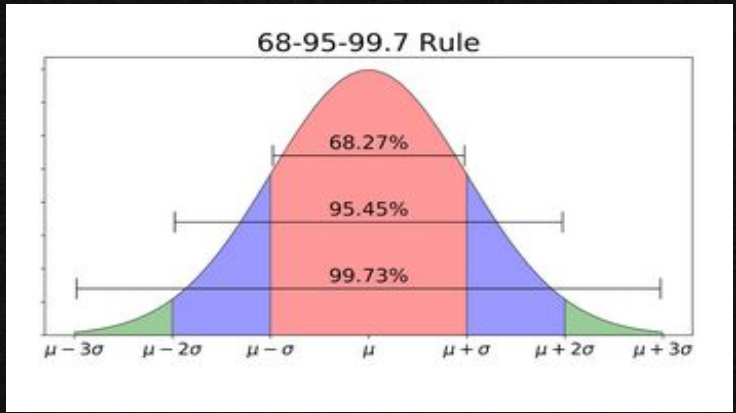
Python e suas bibliotecas



Análise Exploratória de dados

Distribuição Normal

- É simétrica em torno da média.
- A média, a mediana e a moda são todas iguais.
- Todos os dados estão em até 3 desvios padrões.



Análise Exploratória de dados

Distribuições enviesadas

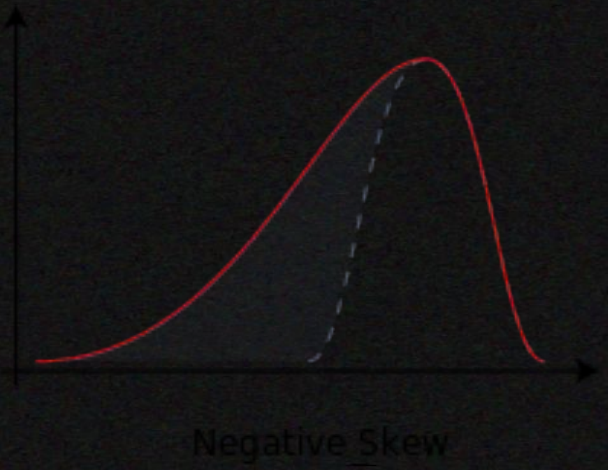
- Assimétrica positiva / à direita.

  Média > Mediana > Moda

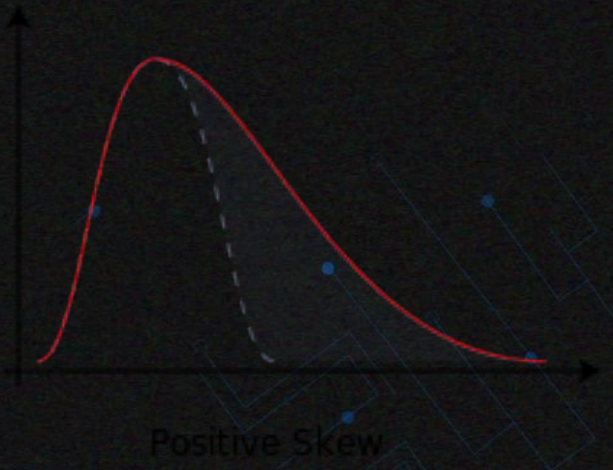
- Assimétrica negativa / à esquerda.

  Média < Mediana < Moda

Assimetria Negativa



Assimetria Positiva



Erro Médio Quadrático

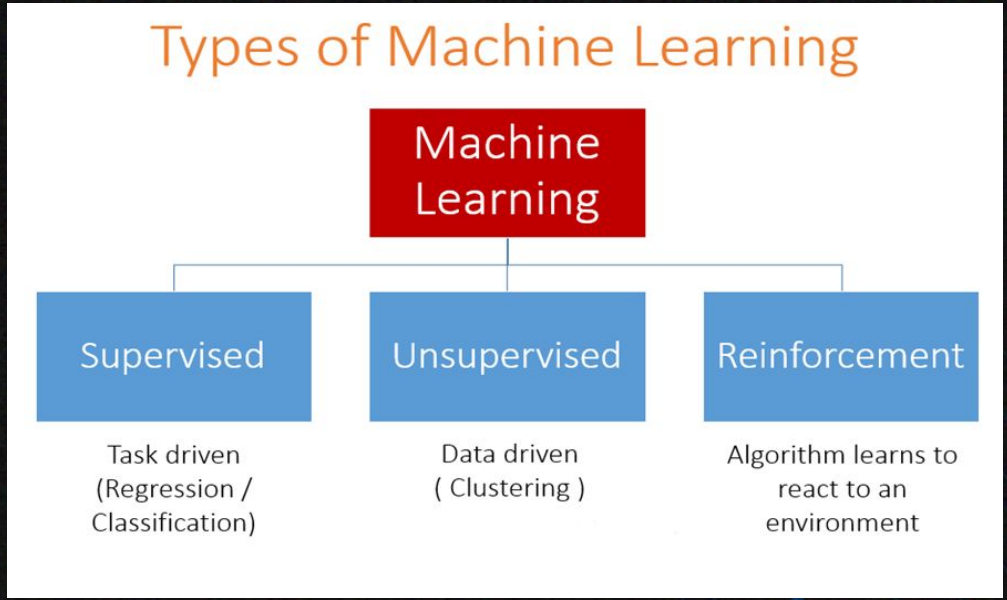
- Erro médio quadrático - Somatório da diferença entre os valores preditos e os valores reais.
- Utiliza a mesma unidade dependente.

$$\mathrm{RMSE} = \sqrt{\sum \frac{(y_{pred} - y_{ref})^2}{N}}$$

Tipos de Aprendizado

Machine Learning

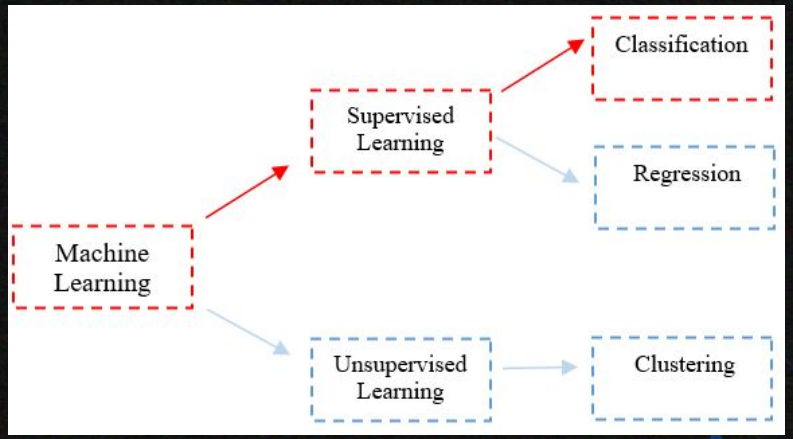
- Aprendizado Supervisionado.
- Aprendizado Não Supervisionado.
- Aprendizado por reforço.



Tarefas de Machine Learning

Machine Learning

- Classificação.
- Regressão.
- Agrupamento.

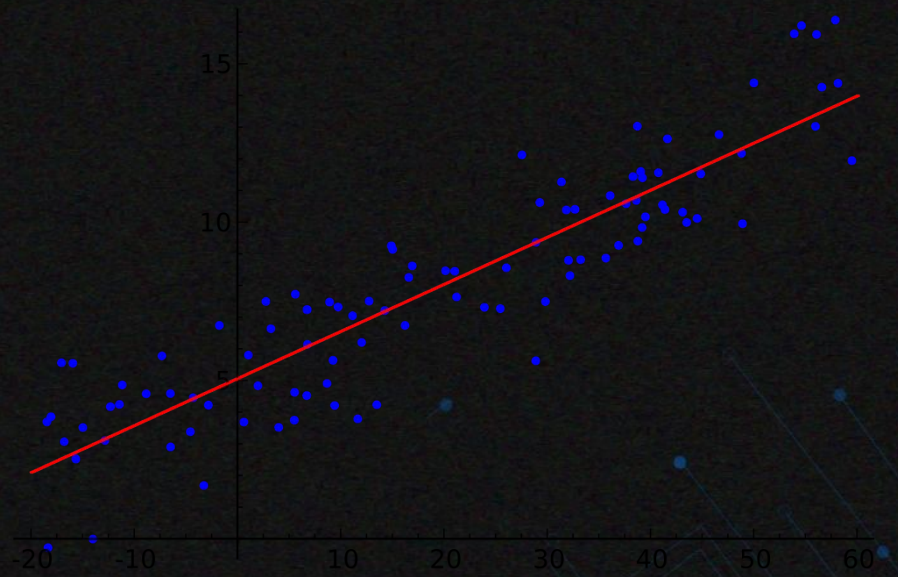


Algoritmos de Machine Learning

Regressão Linear

Regressão linear

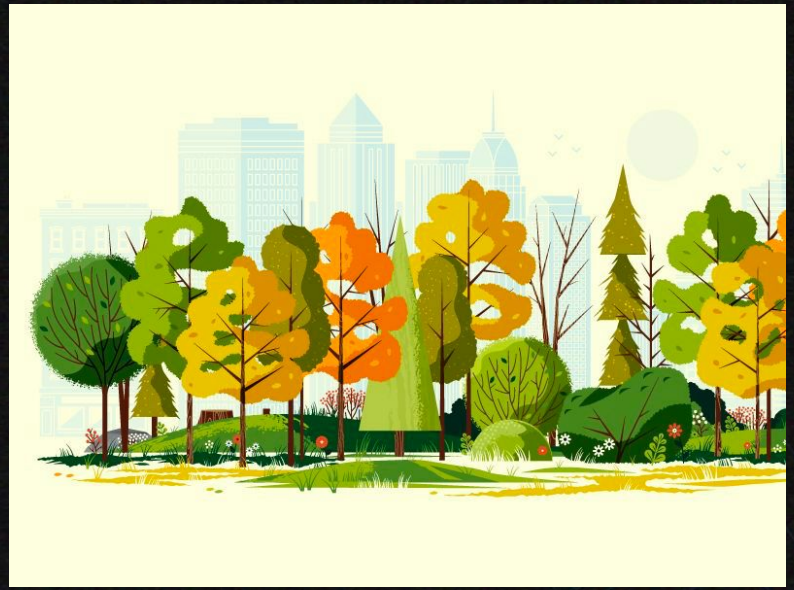
- Algoritmo supervisionado.
- Utiliza equação linear que usa os valores de entrada para predizer as saídas.
- Trabalha apenas com dados numéricos.
- Os pesos são atualizados conforme a função que minimiza erros.



Árvores de Decisão

Árvores de Decisão

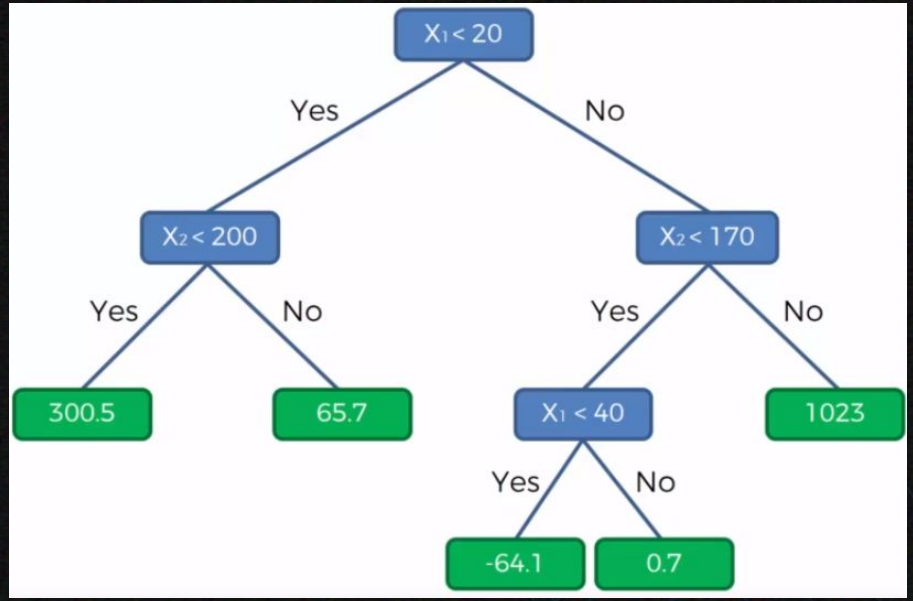
- Algoritmo supervisionado.
- Pode ser usado para classificação ou regressão.
- Consiste na representação em forma de árvore.
- Ao percorrer cada nó o algoritmo toma decisões.



Árvores de Decisão

Árvores de Decisão

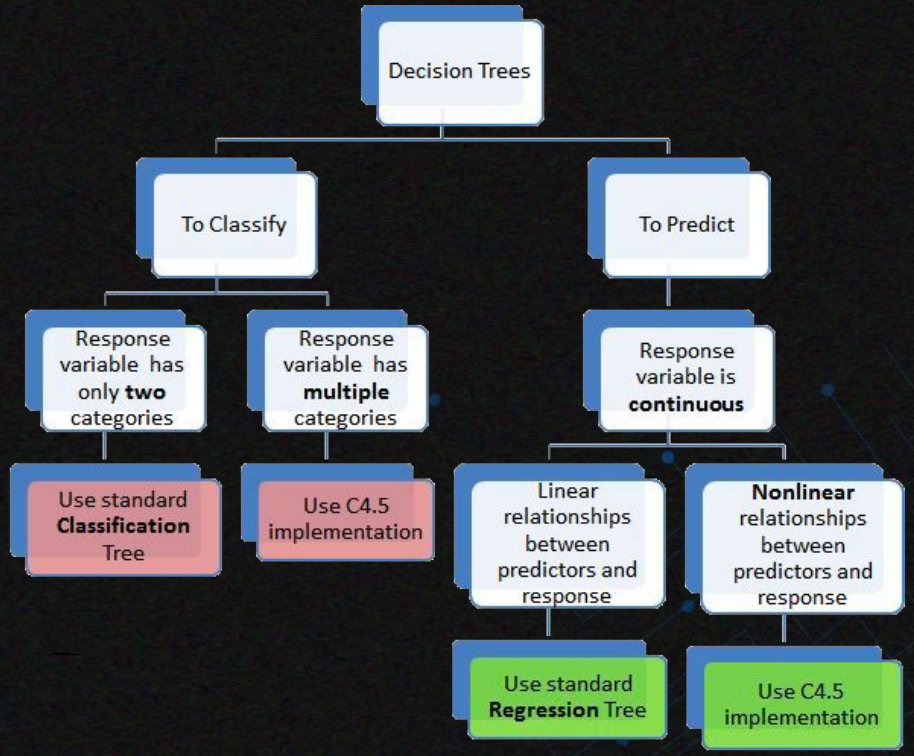
- As árvores são construídas a partir da indução de regras.
- Para cada regra são feitas decisões que ditam a estrutura da árvore.
- Veja no exemplo as raízes, ramos e folhas da árvore a seguir.
- Perceba que os valores dos atributos são decisões a serem tomadas.



Árvores de Decisão

Algumas vantagens

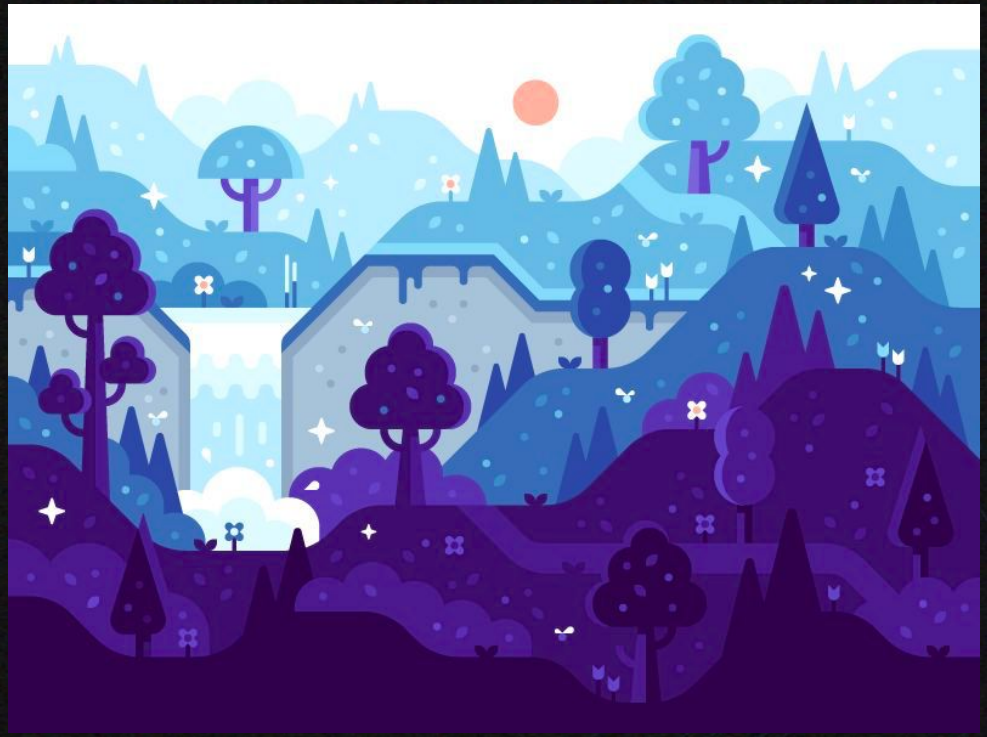
- Fácil entendimento.
- Viabiliza a exploração dos dados.
- Lidam bem com dados não lineares.



Random Forest

Random Forest

- Algoritmo supervisionado.
- Pode ser utilizado para classificação ou regressão.
- Dezenas de árvores combinadas para predizer o melhor resultado.
- Aleatoriedade na seleção de atributos ao invés da seleção a partir do cálculo de impureza.
- Resolve o problema de overfitting da árvore de decisão.



Random Forest

Random Forest

- Primeiro passo, criação do bootstrap dataset.

Dor no peito	Boa Circulação Sanguínea	Arterias Bloqueadas	Peso	Doença Cardiaca
Sim	Não	Sim	125	Sim
Não	Sim	Não	180	Não
Não	Não	Sim	210	Não
Sim	Não	Sim	130	Sim

Random Forest

- A partir do conjunto original .. selecione um número N de features aleatoriamente

Dor no peito	Boa Circulação Sanguínea	Arterias Bloqueadas	Peso	Doença Cardiaca
Sim	Não	Sim	125	Sim
Não	Sim	Não	180	Não
Não	Não	Sim	210	Não
Sim	Não	Sim	130	Sim

Dor no peito	Boa circ Sanguínea	Arterias Bloq.	Peso	Doença Cardiaca
Não	Sim	Não	180	Não
Sim	Não	Sim	130	Sim
Sim	Não	Sim	130	Sim

Bootstrap Dataset

A partir do conjunto original .. Selecione um número N de features aleatoriamente

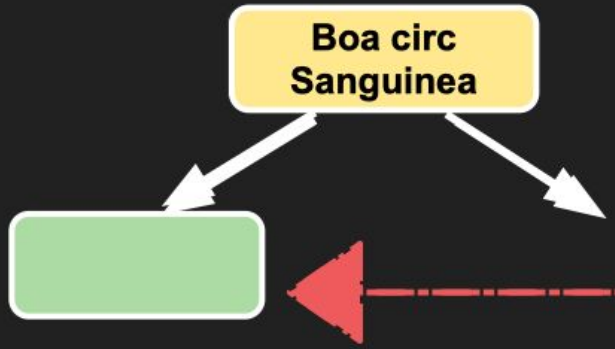
Boa circ Sanguínea	Arterias Bloq.
Sim	Não
Não	Sim
Não	Sim

Dor no peito	Boa circ Sanguínea	Arterias Bloq.	Peso	Doença Cardiaca
Não	Sim	Não	180	Não
Sim	Não	Sim	130	Sim
Sim	Não	Sim	130	Sim

Bootstrap Dataset

Random Forest

A partir do subconjunto selecionado é feita a verificação do atributo que melhor separa os dados..



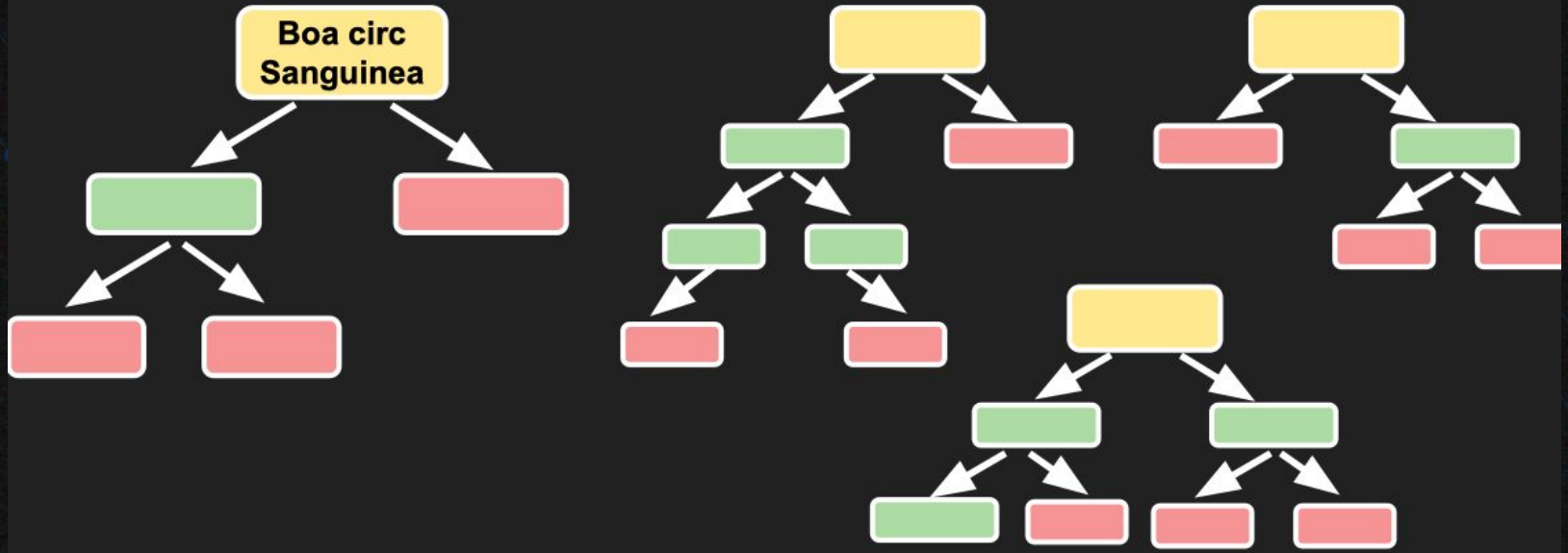
Dor no peito	Boa circ Sanguínea	Arterias Bloq.	Peso	Doença Cardiaca
Não	Sim	Não	180	Não
Sim	Não	Sim	130	Sim
Sim	Não	Sim	130	Sim

Agora é preciso separar mais 2 atributos a partir dos três resultantes para separar os dados

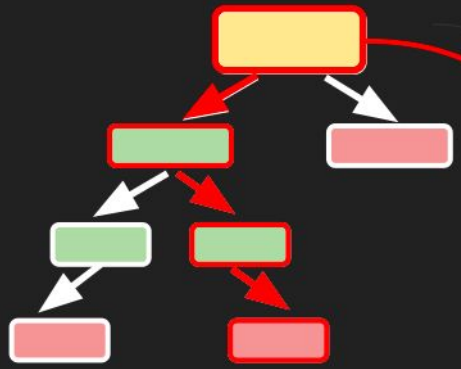
Bootstrap Dataset

Random Forest

As árvores são construídas considerando apenas os **subconjuntos de atributos** selecionados.



Random Forest

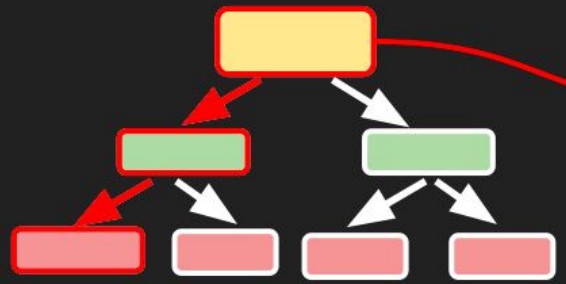


NÃO

Dor no peito	Boa circ Sanguínea	Arterias Bloq.	Peso	Doença Cardiaca
Não	Sim	Não	180	

Doença Cardiaca	
SIM	NÃO
0	1

Random Forest

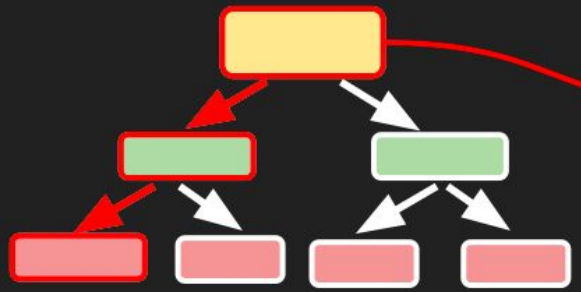


SIM

Dor no peito	Boa circ Sanguínea	Arterias Bloq.	Peso	Doença Cardiaca
Não	Sim	Não	180	

Doença Cardiaca	
SIM	NÃO
0	1

Random Forest

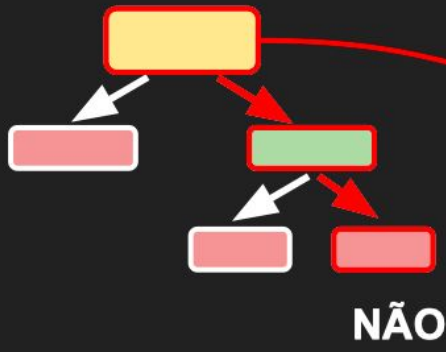


SIM

Dor no peito	Boa circ Sanguínea	Arterias Bloq.	Peso	Doença Cardiaca
Não	Sim	Não	180	

Doença Cardiaca	
SIM	NÃO
0	1

Random Forest

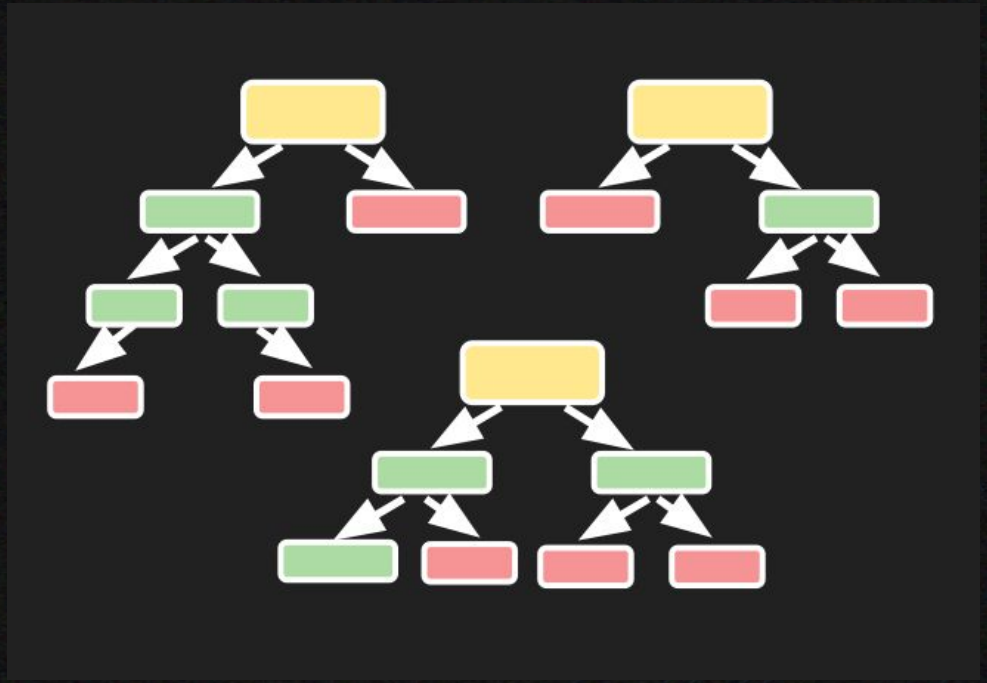


Dor no peito	Boa circ Sanguínea	Arterias Bloq.	Peso	Doença Cardiaca
Não	Sim	Não	180	

Doença Cardiaca	
SIM	NÃO
1	3

Random Forest

- Algumas vantagens
  - Maior robustez
  - Menos propenso a sofrer Overfitting em comparação com uma única Árvore de Decisão
  - Permite a descoberta de conhecimento.
  - Poucos parametros para ajustes.



Random Forest

- Algumas desvantagens
  - Exige um maior poder de processamento
  - Pode ser lento o processo de classificação de novas amostras.

Random Forest

Hands on!