Hype

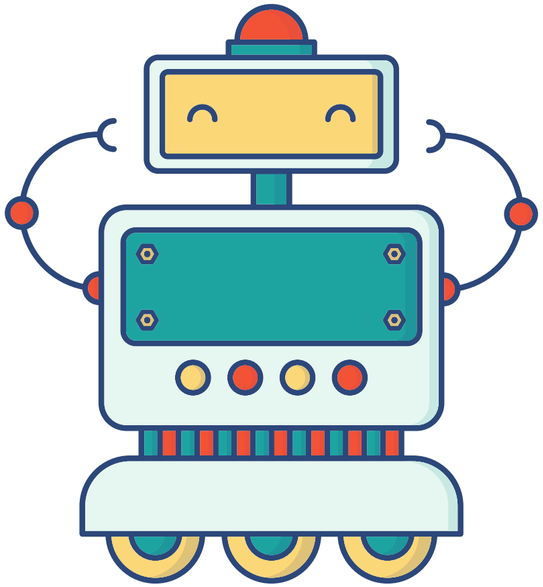
APRENDIZADO DE MÁQUINA

Vinicius Alves Matias

Aprendizado de Máquina

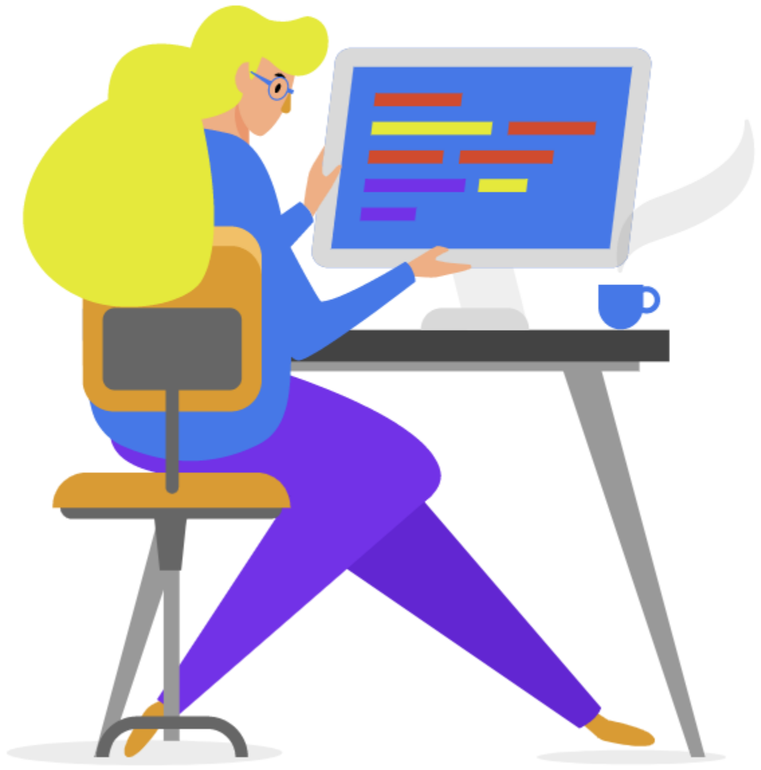
COM SCIKIT-LEARN

1. O que é Machine Learning
2. Problemas e Ferramentas
3. Modelos lineares
4. Modelos baseados em probabilidade e similaridade
5. Árvores e Florestas aleatórias
6. Recapitulação



Árvores de Decisão

Visão Geral



Decision Tree

01

Conceitos Iniciais

Discutiremos o que são árvores de decisão, como representá-las, por que utilizá-las e alguns exemplos

02

Construção

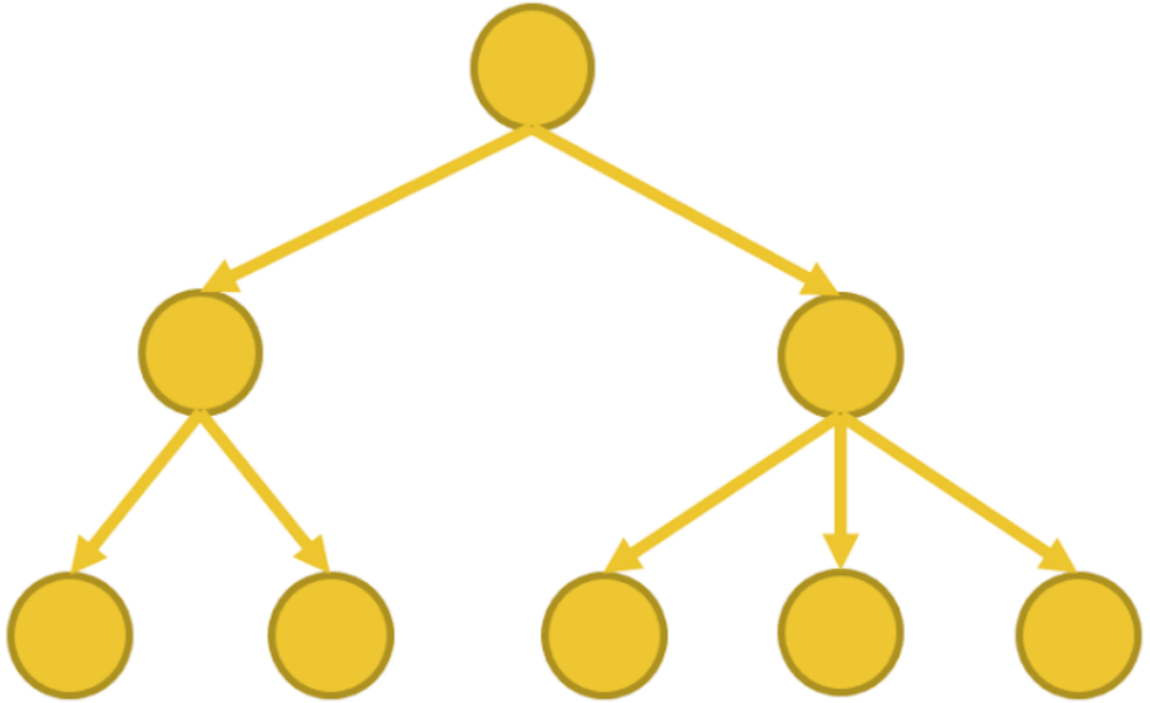
Aqui veremos resumidamente como é o treinamento de uma DT, além de algumas medidas como Entropia e Ganho de Informação

03

Prática

Após um pouco de teoria, vamos ver um exemplo prático com Python

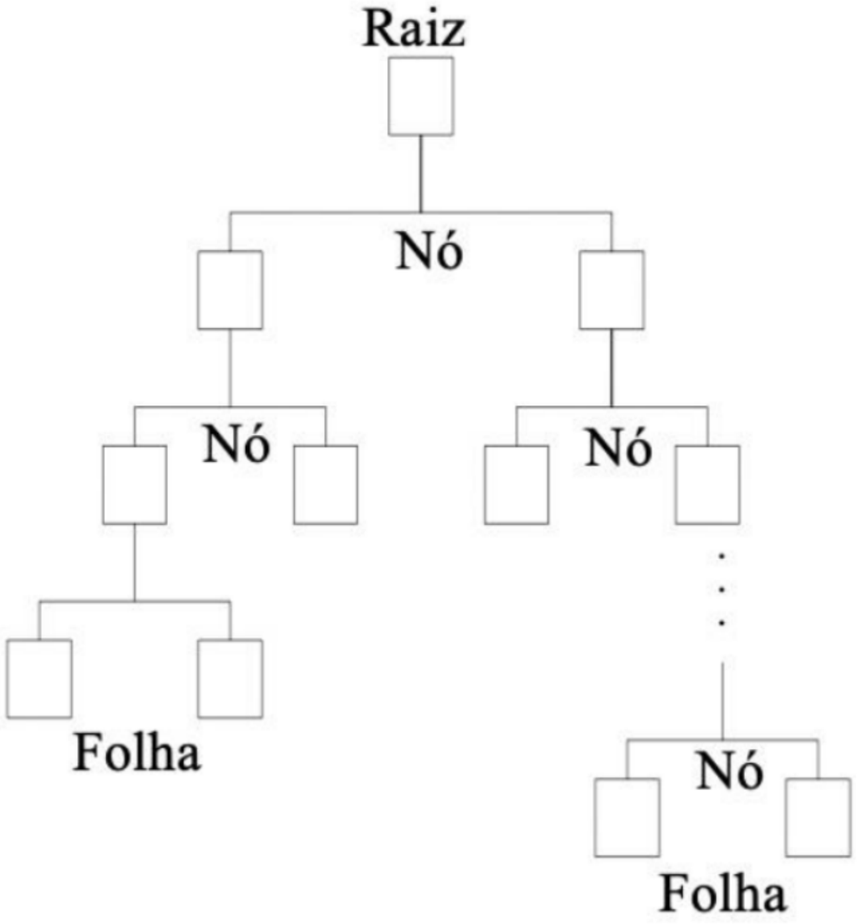
Decision Tree



Decision Tree - Easy Mode

Conceitos Iniciais

- Uma árvore de decisão é uma estrutura hierárquica semelhante a um fluxograma, formada por nós.
- Cada nó interno testa um atributo
- Cada ramo corresponde a um valor do atributo
- Cada folha representa uma classe

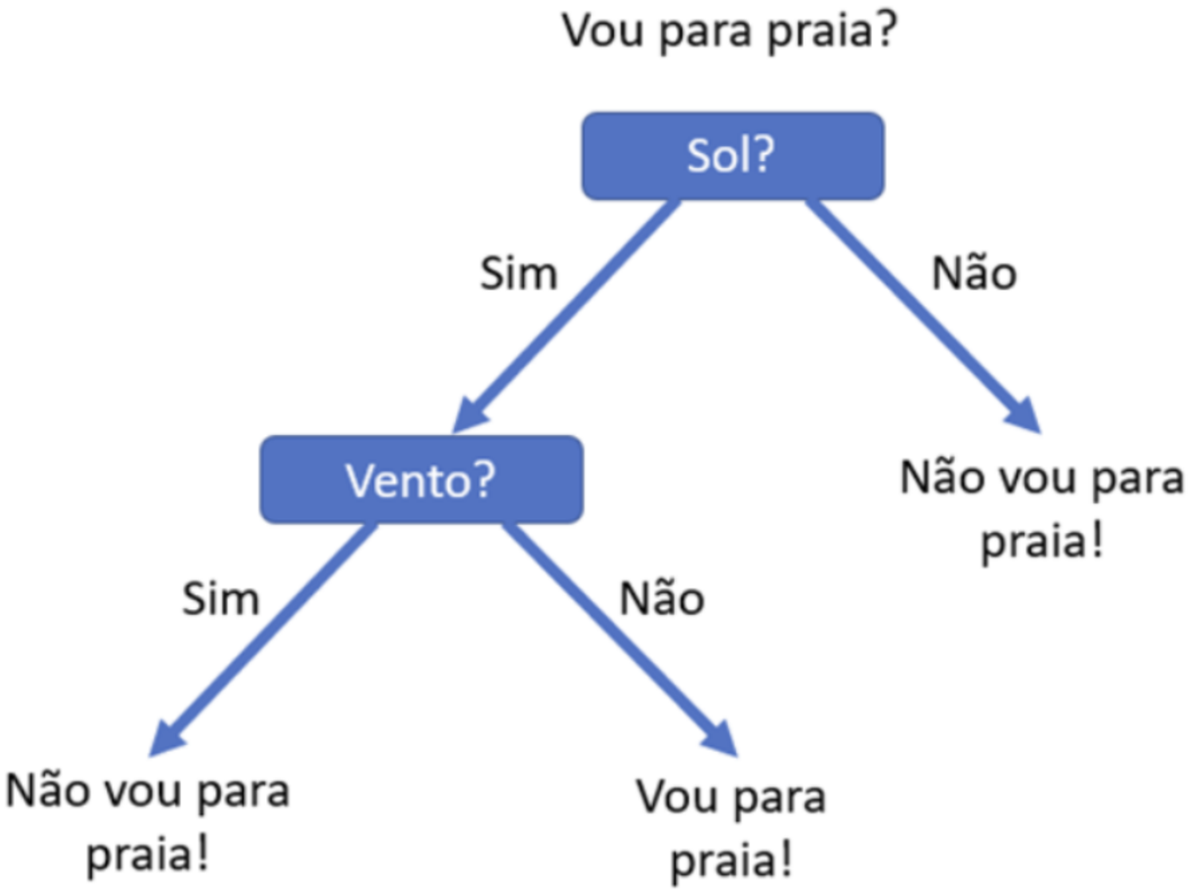


Conceitos Iniciais

- Em termos técnicos, uma árvore de decisão é um **algoritmo de aprendizado de máquina supervisionado** utilizado para tarefas de classificação e regressão

- Isto indica que ela pode ser usada tanto para prever categorias discretas (sim ou não, por exemplo) quanto para prever valores contínuos (valor do lucro em reais, por exemplo)

- Um caminho da raiz da árvore até uma de suas folhas pode ser transformado em uma **regra de classificação** (condições).



Por que árvores de decisão são tão utilizadas ?

01

Simplicidade
Fáceis de entender, visualizar e interpretar

02

Adaptabilidade
Lida bem com dados numéricos e categóricos

03

Versatilidade
Funciona tanto para classificação quanto para regressão

04

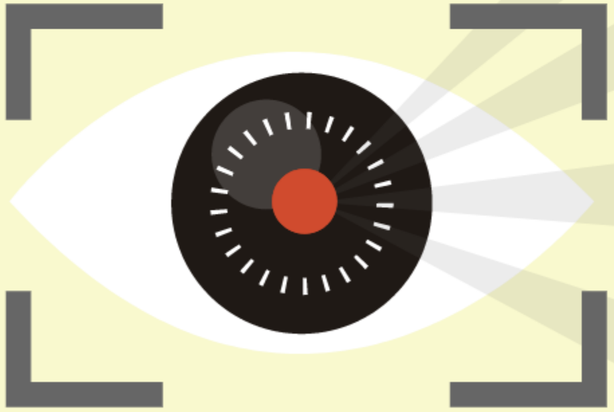
Facilidade
Requer pouca ou nenhuma preparação dos dados

05

Robustez
Trabalha com problemas que possuem múltiplos rótulos

Exemplos

Alguns usos clássicos de árvore de decisão



Diagnóstico de doenças

Classificar se um paciente pode ou não ser acometido de determinada doença

Previsão de empréstimo

Prever um valor de empréstimo que pode ser concedido a um cliente

Análise de sentimentos

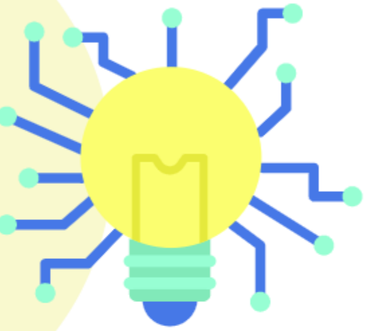
Categorizar um texto como positivo ou negativo

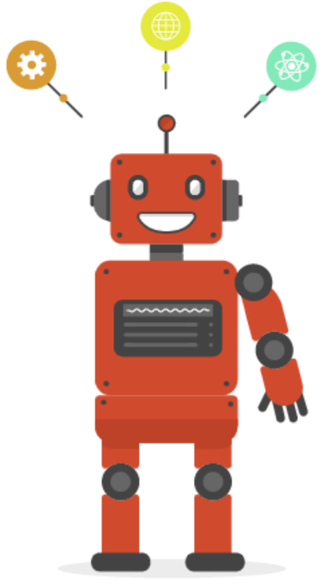
Análise de crédito

Decidir se deve ou não liberar crédito para um cliente com base no score do Serasa

Construção

O processo de construção da árvore (treinamento do modelo) se chama **indução**. O propósito da DT é fazer diversas divisões dos dados em subconjuntos, de tal forma que os subconjuntos vão ficando cada vez mais “puros”, ou seja, a medida em que eles contém menos classes (ou apenas uma) da variável *target*





Uma forma de trabalhar matematicamente com a pureza é por meio da análise da **entropia** e do **ganho de informação**.

Entropia

Basicamente a medida que nos diz o quanto nossos dados estão desorganizados e misturados

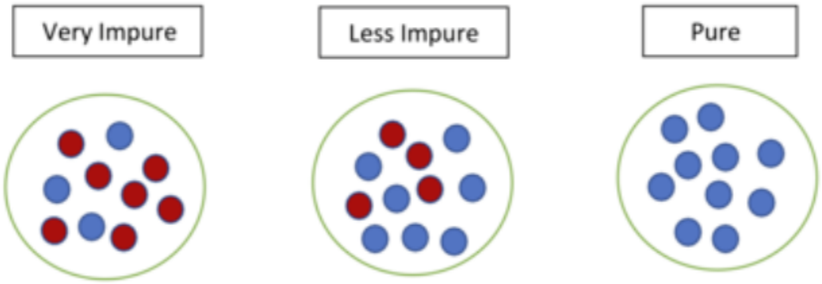
A entropia é um valor que varia de 0 a 1, sendo que o zero indica um conjunto totalmente puro

Conjuntos capazes de representar apenas uma classe do modelo

⬇

dados menos entrópicos

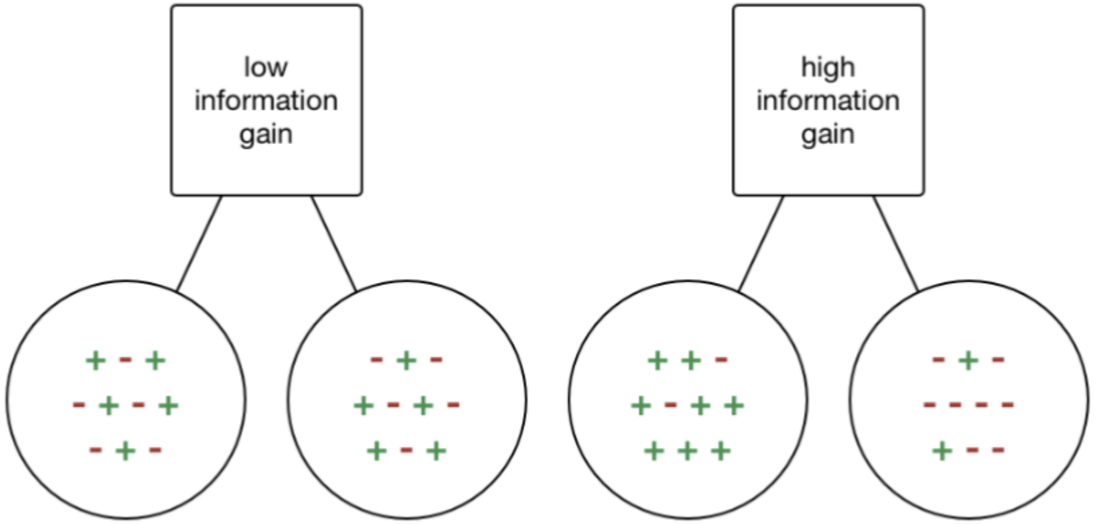
A construção da árvore de decisão é pautada na criação de ramificações baseadas em condições que minimizem a entropia



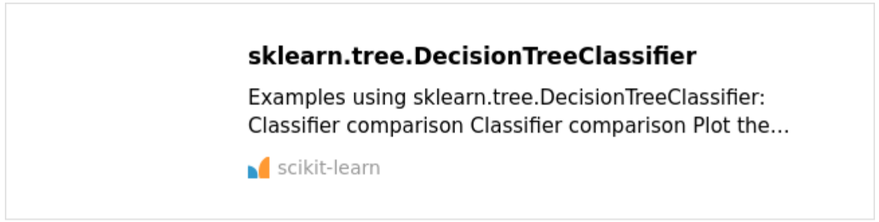
Ganho de Informação

Uma medida que nos diz o quão bem uma feature do conjunto de dados separa os registros conforme as suas classes

Basicamente nos diz o quanto ganharíamos de pureza ao se dividir um conjunto segundo um atributo

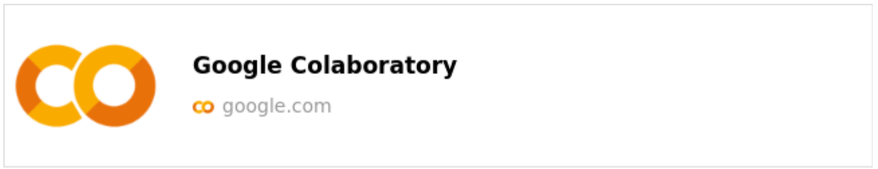


Documentação



Documentação Decision Tree

Código



Random Forest

Random Forest

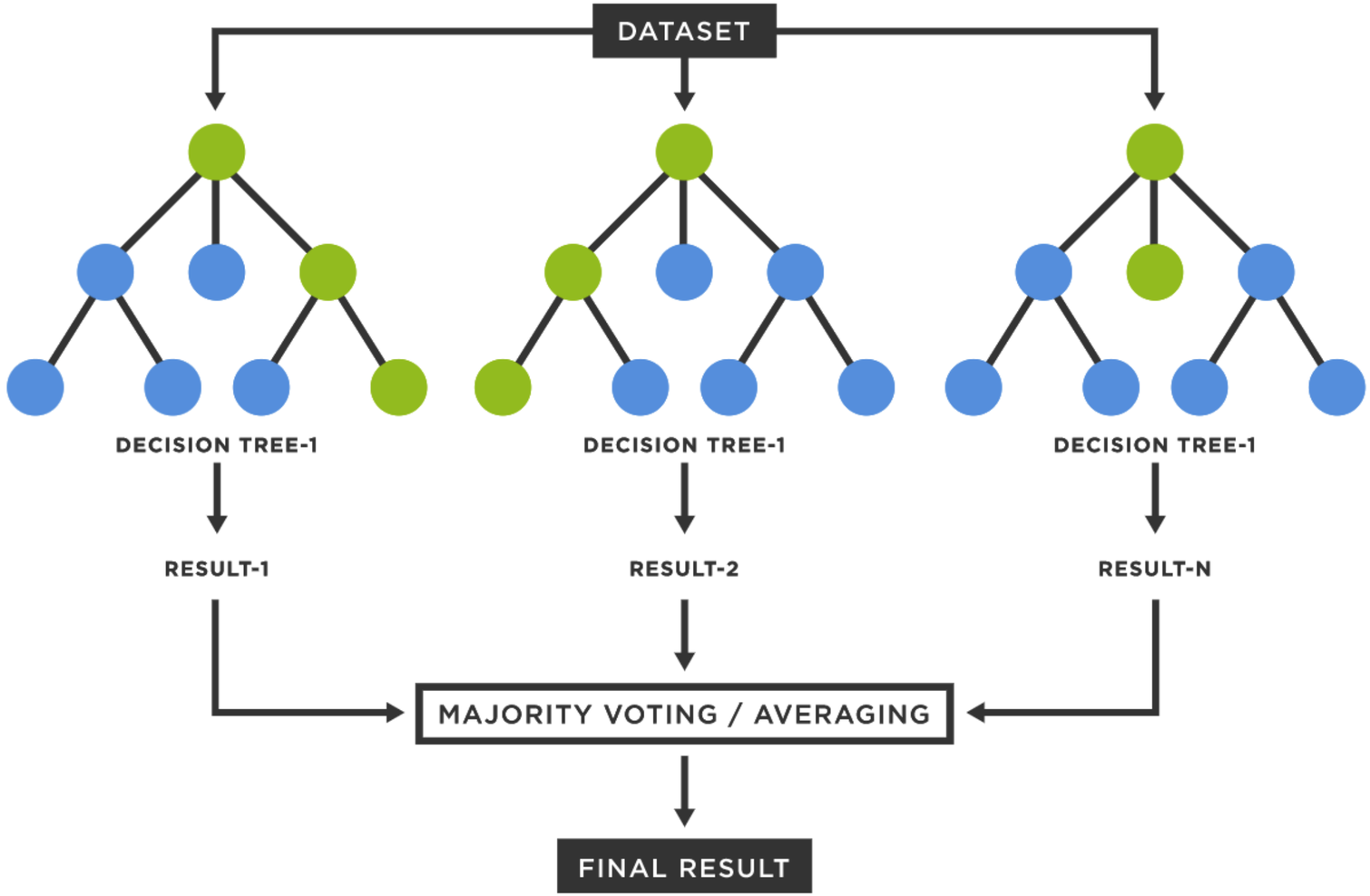


Imagem tibco

Random Forest

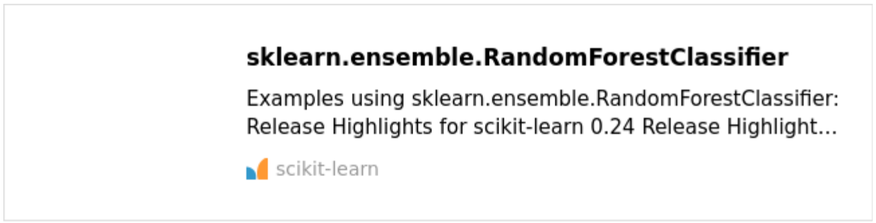
Métodos ensemble (Bagging) - Junção de classificadores

Seleção de atributos

Overfitting e diferença entre as árvores

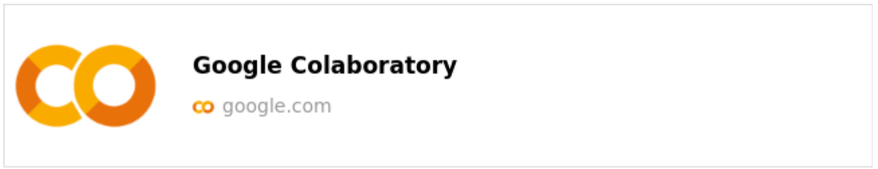
Quantidade de árvores e atributos

Documentação



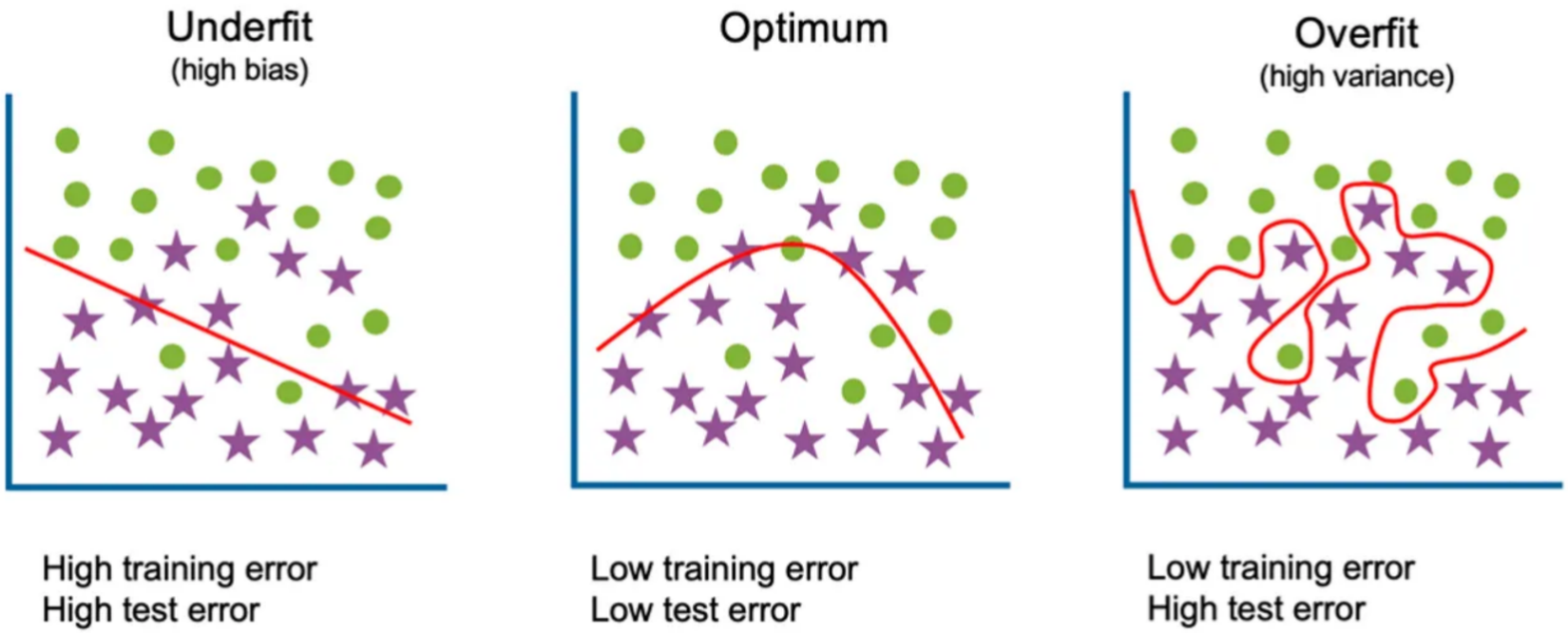
Documentação Random Forest

Código



Revisitando alguns tópicos

Overfitting e Underfitting



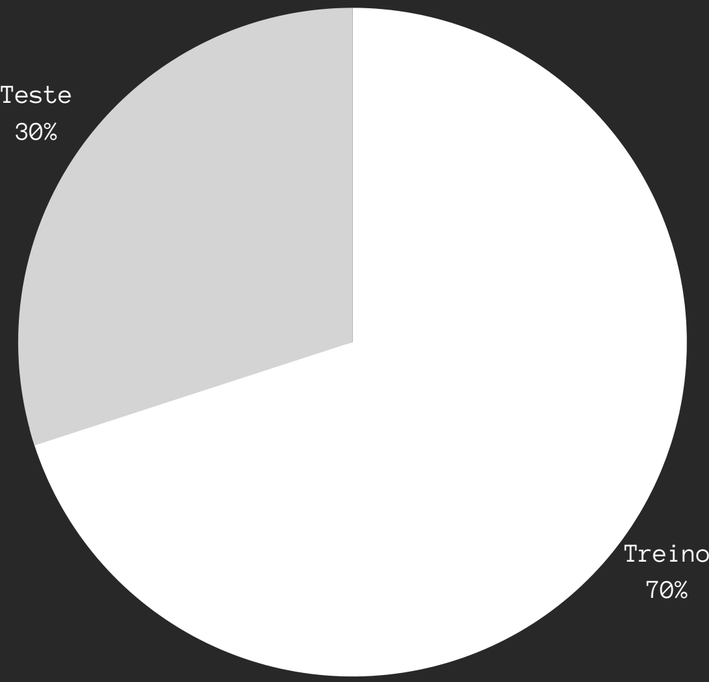
Sofisticado e simples
Parábola da prova

Imagem: IBM

Separação em treino e teste

NUNCA SE MEXE EM TESTE, APÓS SEPARADO

- Amostragem aleatória
- Eventuais pré-processamentos precisam acontecer em treino e teste
- 70% treino e 30% teste - Questionável (quantidade de dados)



Classificação binária

MÉTRICAS

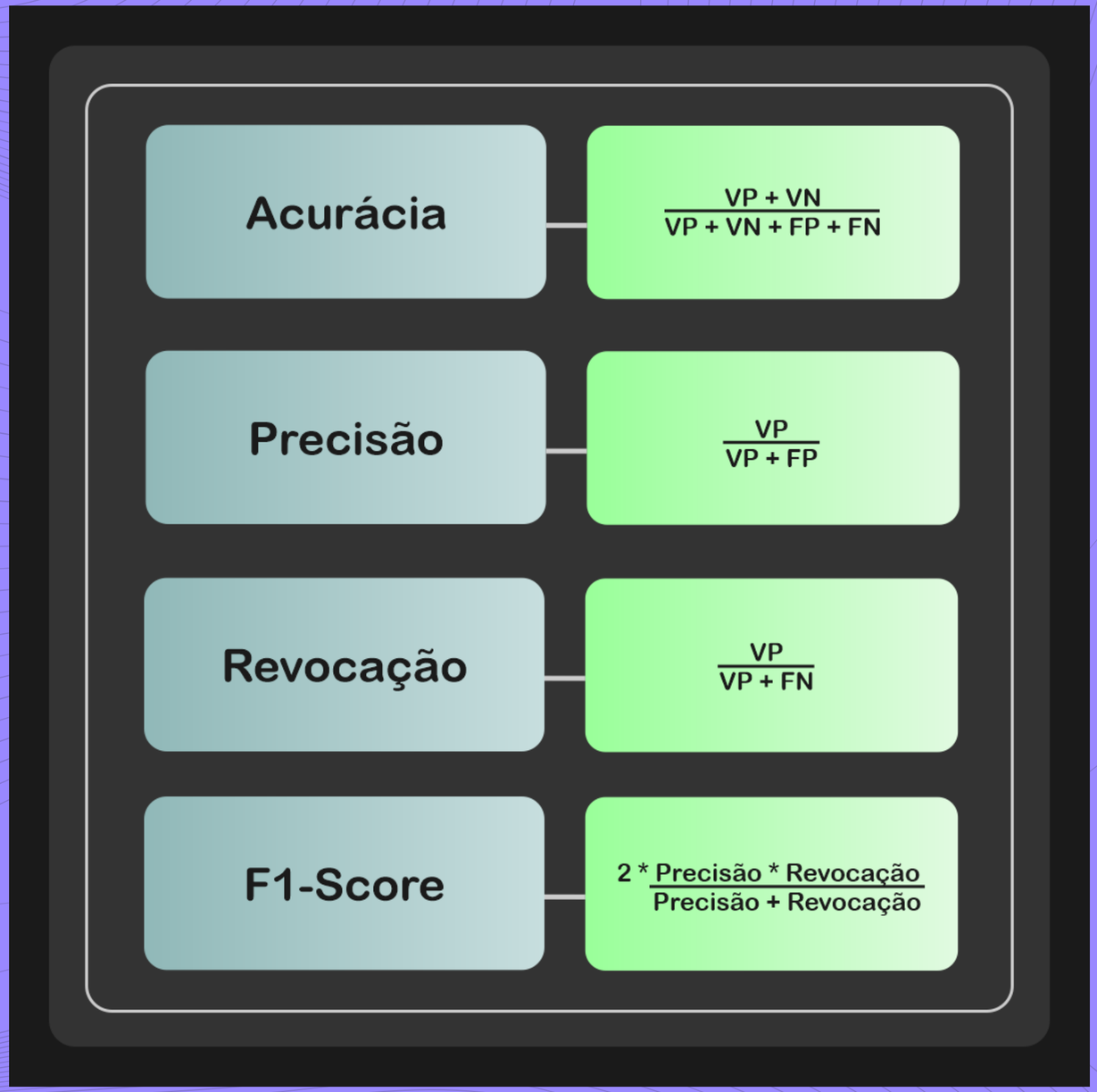


Imagem: vitorborbarodrigues.medium

Métricas multiclasse

Class	TP	FP	FN	Precision	Recall
A	5	2	1	0.71	0.83
B	10	90	7	0.1	0.58
C	15	11	2	0.57	0.88

Precisão micro

$$\frac{TP_A+TP_B+TP_C}{TP_A+TP_B+TP_C+FP_A+FP_B+FP_C} \quad \frac{5+10+15}{5+10+15+2+90+11} = 0.22$$

Precisão macro

$$\frac{Pre_A+Pre_B+Pre_C}{3} \quad \frac{.71+0.1+.57}{3} = 0.46$$

Imagens: androidkt

Outros tópicos interessantes

Pipeline

GridSearchCV

PCA para extração de features

SelectKBest para seleção de features

Hype

APRENDIZADO DE MÁQUINA

Vinicius Alves Matias