uni1500

ANALISTA DE DADOS

| Dez22

CLASSIFICAÇÃO

Modelos de Classificação

- Técnica onde categoriza-se os dados em um determinado número de classes.
- O principal objetivo de um problema de classificação é identificar a categoria/classe na qual um novo dado se enquadrará.
- Os algoritmos de classificação usam dados de treinamento de entrada para prever a probabilidade de que os dados subsequentes caiam em uma das categorias predeterminadas.
- Ex: filtrar e-mails em “spam” ou “não spam”.

CLASSIFICAÇÃO

Terminologias

Classificador

Algoritmo que mapeia os dados de entrada para uma categoria específica.

Modelo de classificação

Tenta tirar conclusões dos valores de entrada fornecidos para treinamento. Irá prever os categorias para os novos dados.

Feature

Propriedade individual mensurável de um fenômeno que está sendo observado.

Classificação Binária

Classificação com dois resultados possíveis.
Ex: Classificação de spam (Spam/Não é spam).

Classificação multiclasse

Mais de duas classes mas apenas um rótulo de destino.
Ex: Classificação de um animal (gato, cachorro ou réptil).

Classificação multi-rótulo

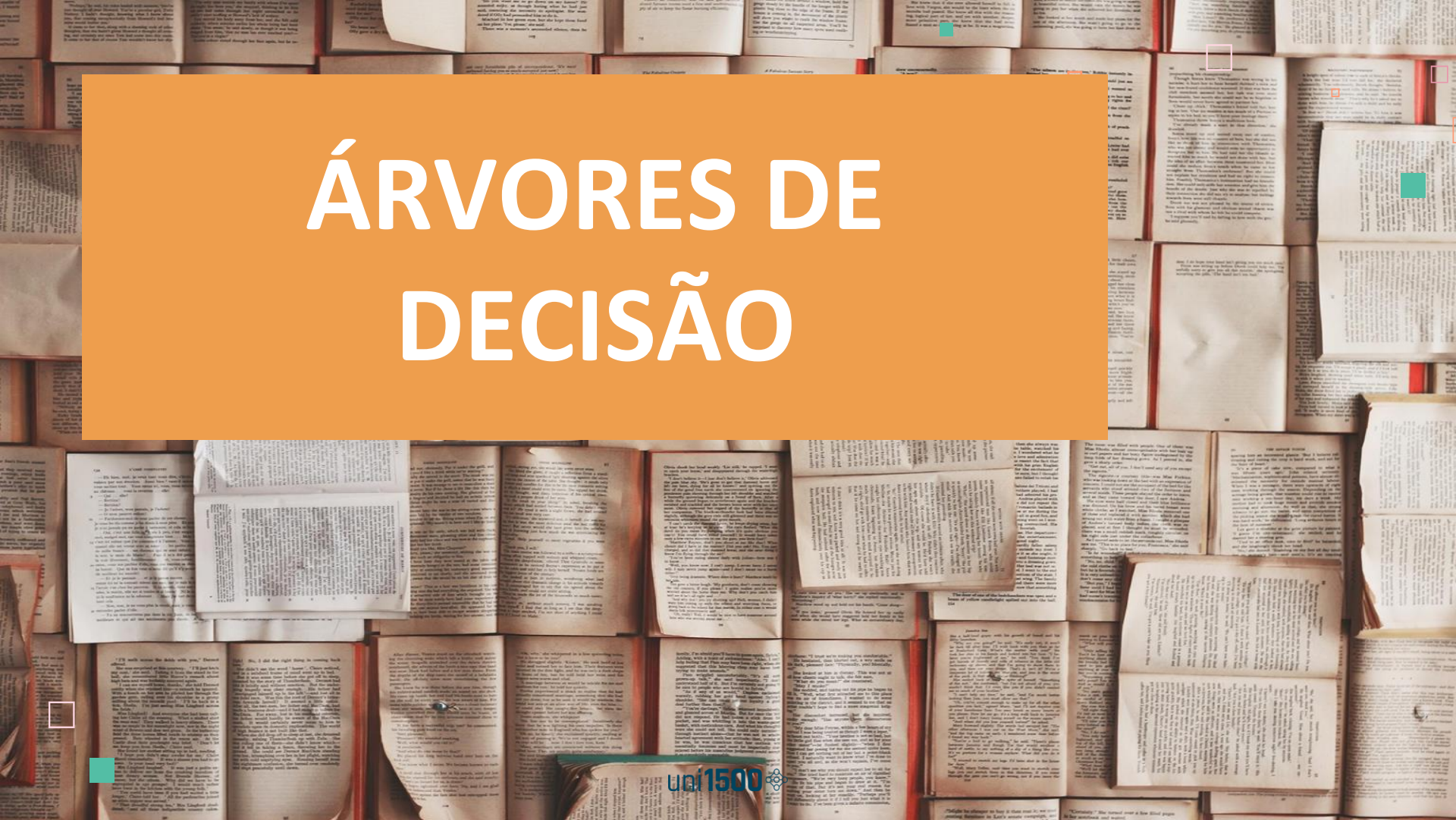
Cada amostra irá para um conjunto de rótulos de destino.
Ex: Artigo de notícias (esportes, pessoa e local).

CLASSIFICAÇÃO

Principais Algoritmos

- Regressão Logística
- Naive Bayes
- K-ésimo vizinho mais próximo
- Árvore de decisão
- Support Vector Machines

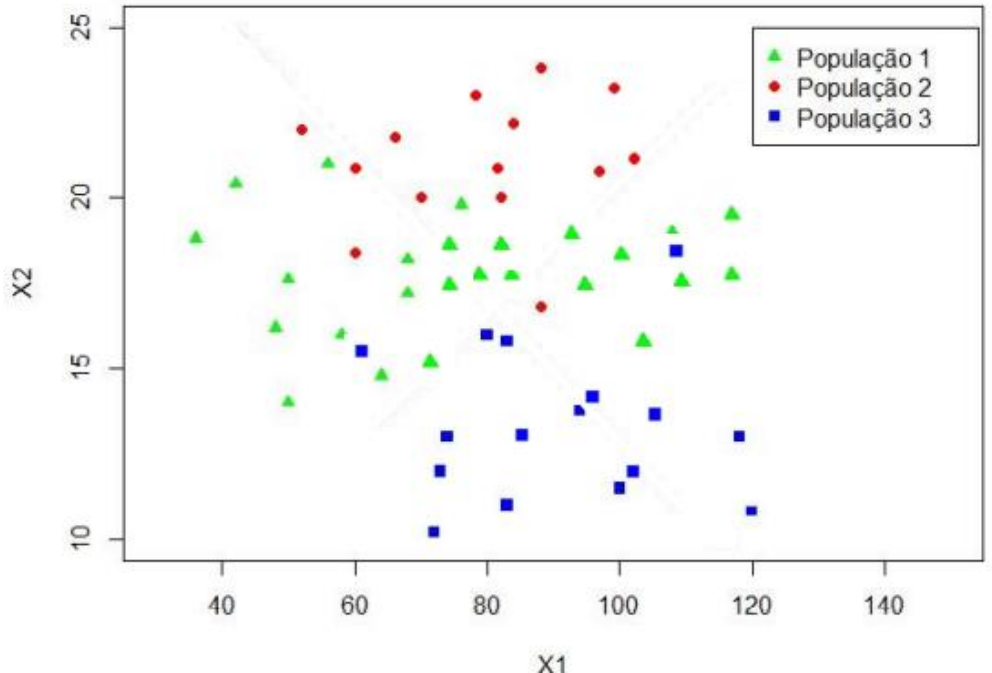
ÁRVORES DE DECISÃO



CLASSIFICAÇÃO

Árvore de decisão

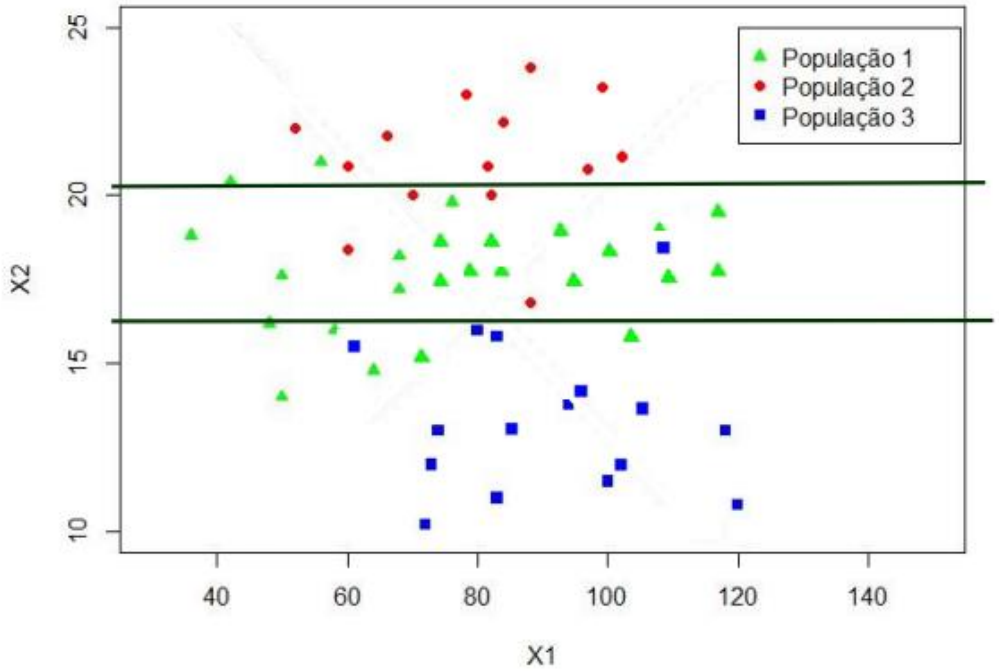
Regra para identificar 3 populações



CLASSIFICAÇÃO

Árvore de decisão

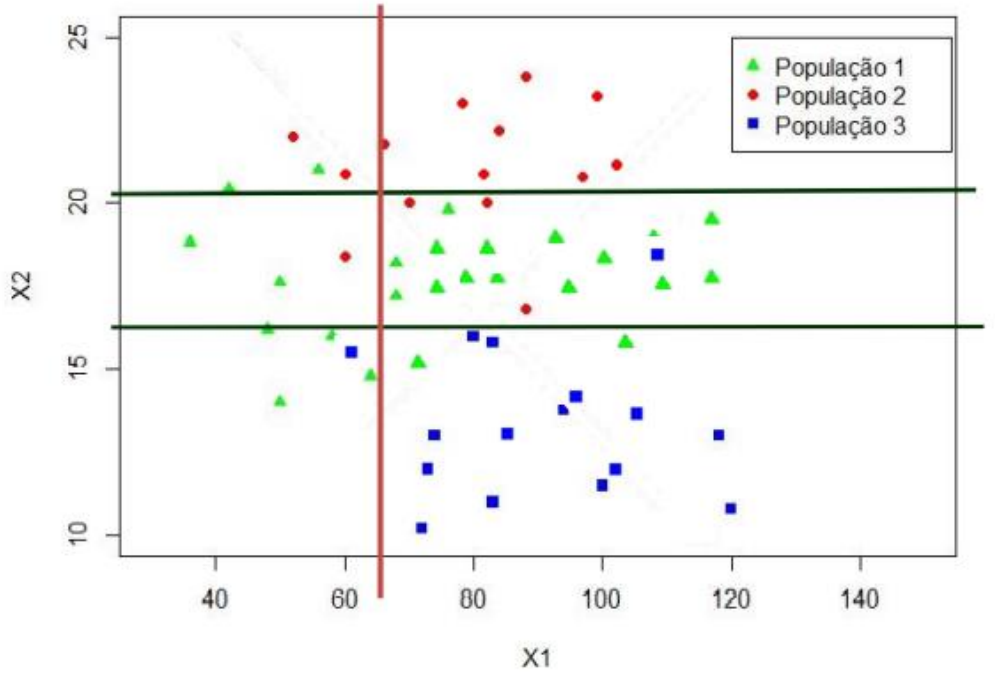
Regra para identificar 3 populações



CLASSIFICAÇÃO

Árvore de decisão

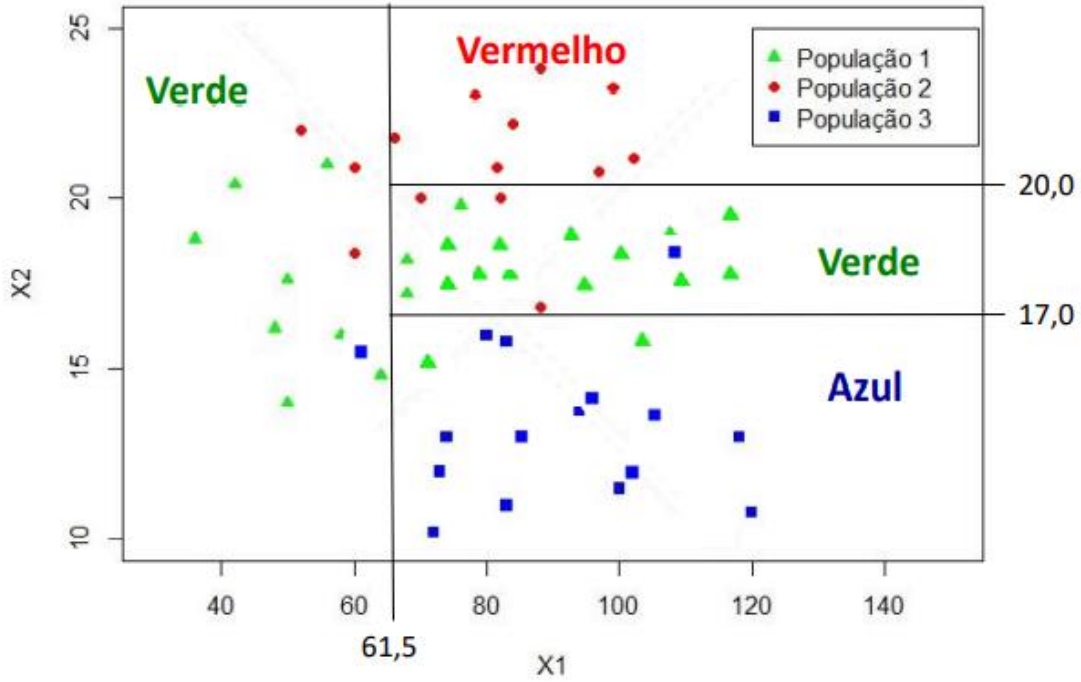
Regra para identificar 3 populações



CLASSIFICAÇÃO

Árvore de decisão

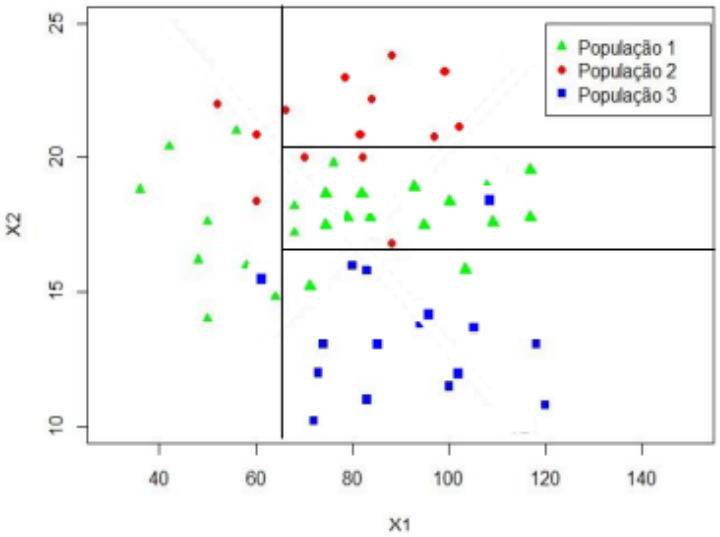
Regra para identificar 3 populações



CLASSIFICAÇÃO

Árvore de decisão

Regra para identificar 3 populações



Regra 1: Se X1>61,5 e X2<17, classifico como População 3

Regra 2: Se X1 < 61,5, classifico como População 1

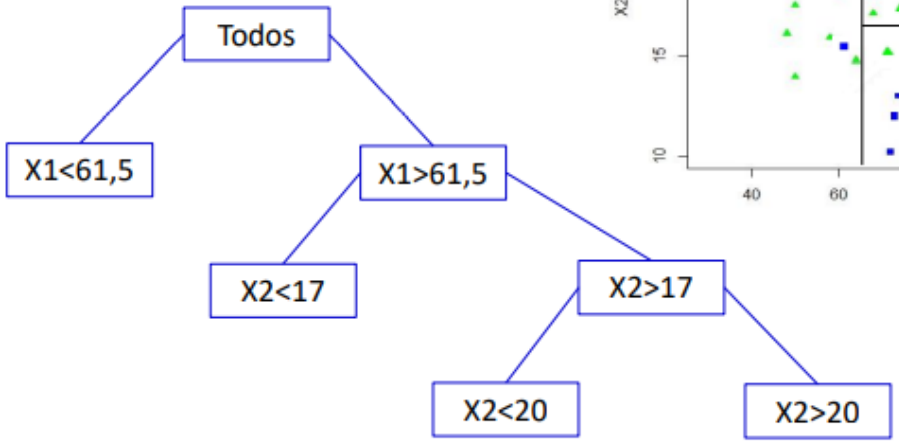
Regra 3: Se X1>61,5 e 17<X2<20, classifico como População 1

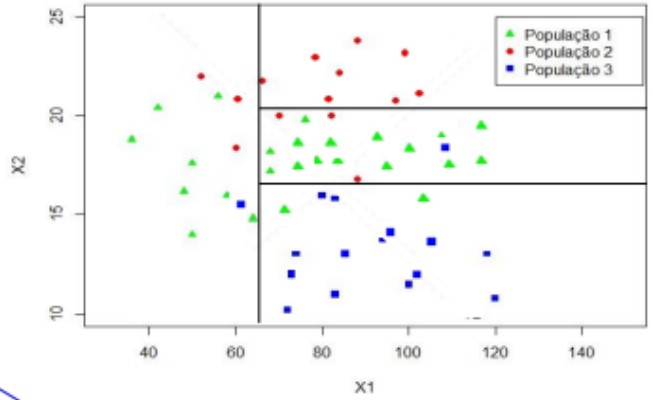
Regra 4: Se X1>61,5 e X2>20, classifico como População 2

CLASSIFICAÇÃO

Árvore de decisão

Regra para identificar 3 pop.

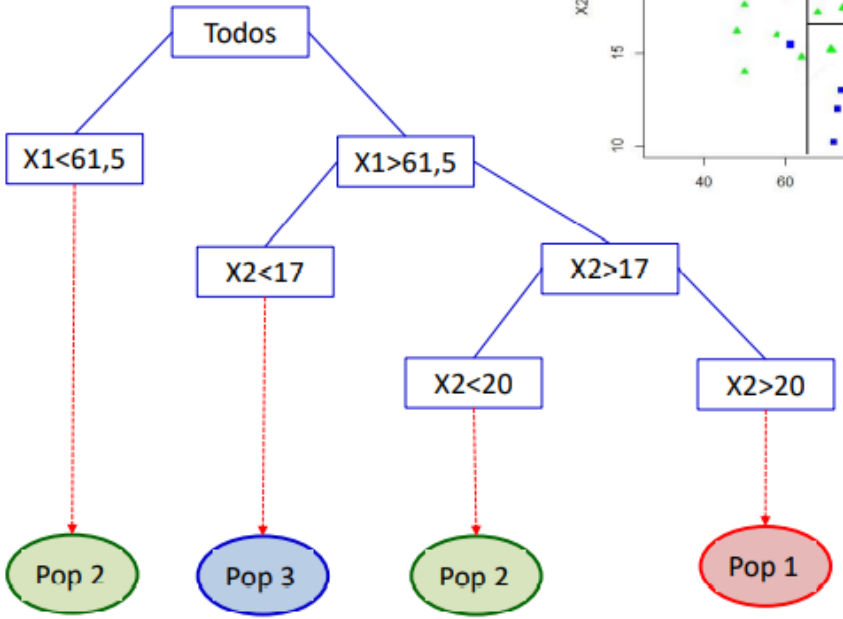


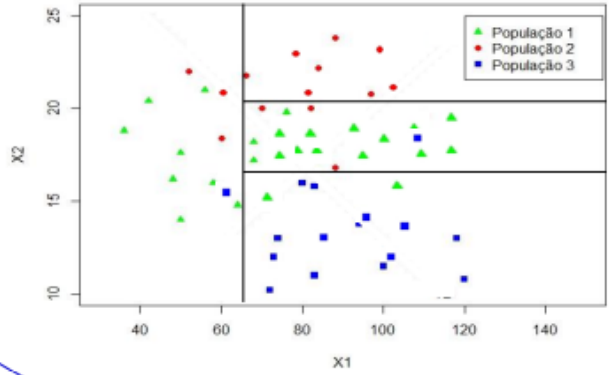


CLASSIFICAÇÃO

Árvore de decisão

Regra para identificar 3 pop.





CLASSIFICAÇÃO

Árvore de decisão

Algoritmo de aprendizado supervisionado.

Divide a população em dois ou mais conjuntos homogêneos com base nos atributos mais significativos, tornando os grupos tão distintos quanto possível.

Vantagens

- Simples de entender e visualizar;
- Requer pouca preparação de dados;
- Pode lidar com dados numéricos e categóricos.

Desvantagens

- Pode criar árvores complexas;
- Árvores podem ser instáveis (pequenas variações nos dados podem resultar árvore diferente).

ENTENDENDO O PROBLEMA

CLASSIFICAÇÃO

Entendendo o Problema

O banco de dados a ser analisado é composto por 12 variáveis e busca prever quais pessoas são mais propensas a sobreviver depois da colisão do navio Titanic com o iceberg. As variáveis possuem informações como:

- Nome do passageiro;
- Idade do passageiro;
- Classe do passageiro;
- Cabine em que viajava;
- Preço da passagem;
- Se sobreviveu ou não.

a) Qual seria a variável de interesse?

b) Quais seriam as possíveis variáveis preditoras?

c) Por que deveríamos usar classificação?

ANÁLISE PRELIMINAR

CLASSIFICAÇÃO

Análise Preliminar

Algumas análises que podemos fazer nesta etapa são:

- Descartar as variáveis que julgar desnecessárias ou irrelevantes;
- Verificar se todas as variáveis categóricas estão identificadas como *factor*;
- Descartar os valores ausentes.

EXECUÇÃO DO MODELO

CLASSIFICAÇÃO

Execução do Modelo

Medidas que necessitam ser realizadas nesta etapa:

- Separação entre treino e teste;
- Verificar se as proporções estão adequadas;
- Executar o modelo.

PREDIÇÕES

VALIDAÇÃO DO MODELO

CLASSIFICAÇÃO

Métricas de qualidade do ajuste

Para verificar a qualidade do ajuste de um modelo, utiliza-se **quatro métricas**:

- Matriz de Confusão;
- Acurácia;
- Precisão;
- Recall.

CLASSIFICAÇÃO

Avaliação do modelo: Matriz de confusão

Tabela que indica os erros e acertos do modelo, comparando com o resultado esperado (ou etiquetas/labels).

A imagem abaixo demonstra um exemplo de uma matriz de confusão.

		Valor Predito	
		Sim	**Não**
Real	**Sim**	Verdadeiro Positivo (TP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (TN)

CLASSIFICAÇÃO

Avaliação do modelo: Acurácia

- Mede a frequência com que o classificador prevê corretamente (range 0 - 100), indicando a performance geral do modelo.
- Não deve-se usar precisão em problemas desbalanceados, pois é fácil obter uma alta acurácia simplesmente classificando todas as observações como a classe majoritária. Por exemplo, em um modelo para detecção de fraudes classe que costuma ter baixa frequência o algoritmo provavelmente classificará todas as transações como não fraudulentas, podendo obter uma precisão superior a 90%.

Define-se Acurácia como a razão entre o número de previsões corretas e o número total de previsões: $\frac{TP + TN}{TP + FP + TN + FN}$.

CLASSIFICAÇÃO

Avaliação do modelo: Precisão

- Dentre todas as classificações de classe positivo que o modelo fez, quantas estão corretas?
- Quantifica o número de previsões positivas de classe que realmente pertencem à classe positiva. Ajuda quando os custos de falsos positivos são altos. Responde a pergunta: Que proporção de identificações positivas foi realmente correta?

Define-se Precisão como a razão entre o número de verdadeiros positivos dividido pelo número de verdadeiros positivos mais o número de falsos positivos : $\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$.

CLASSIFICAÇÃO

Avaliação do modelo: Recall

- Medida do modelo que identifica corretamente os verdadeiros positivos.
- Responde a pergunta: Que proporção de positivos reais foi identificada corretamente? Ou seja, dentre todas as situações de classe positivo como valor esperado, quantas estão corretas?

Define-se Recall como a razão entre o número de verdadeiros positivos dividido pelo número de verdadeiros positivos mais o número de falsos negativos: $\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$.