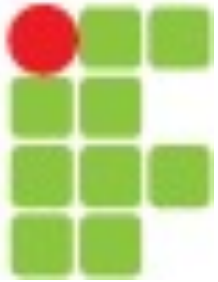


Instituto Federal de Santa Catarina
Campus Florianópolis

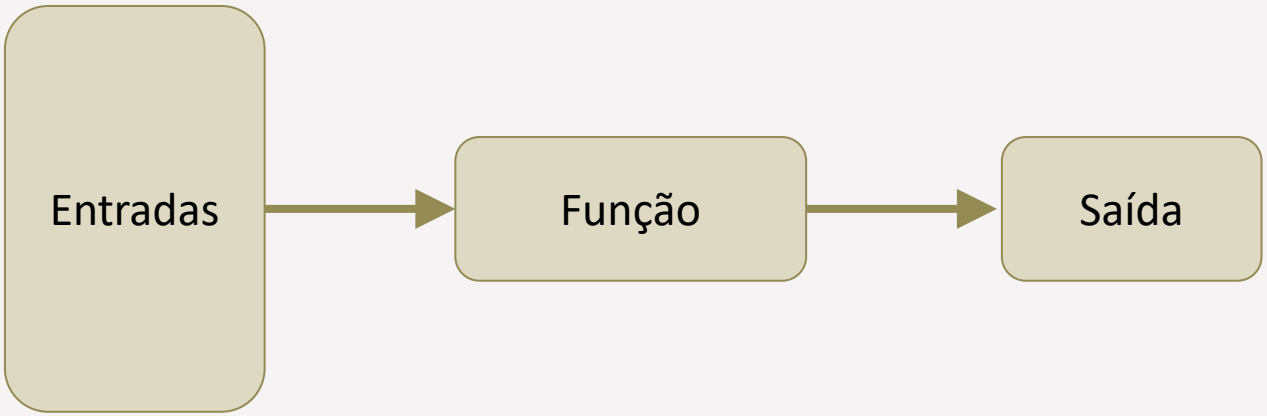
Classificação

Prof. Glauco Cardozo

glauco.cardozo@ifsc.edu.br

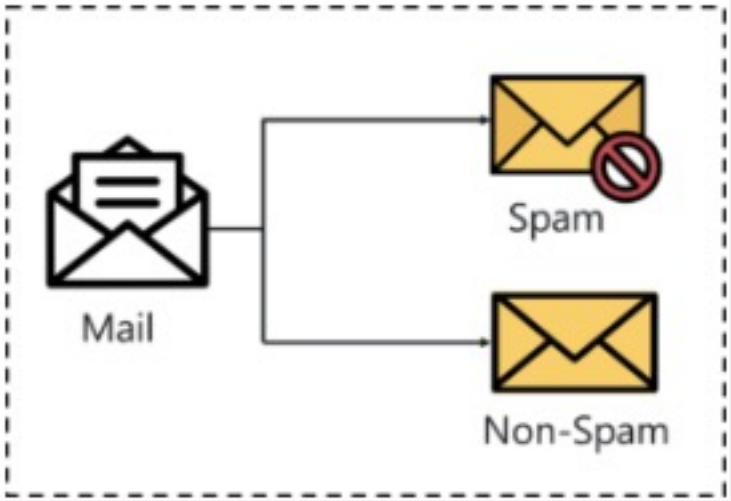
Classificação

A classificação em aprendizado de máquina e estatística é uma abordagem de aprendizado supervisionado na qual o programa de computador aprende com os dados fornecidos e faz novas observações ou classificações.



Classificação

A modelagem preditiva de classificação é a tarefa de aproximar a função de mapeamento de variáveis de entrada para variáveis de saída discretas. O objetivo principal é identificar em qual classe/categoria os novos dados se enquadrarão.



Classificação

A detecção de doenças cardíacas pode ser identificada como um problema de classificação, esta é uma classificação binária, pois pode haver apenas duas classes, ou seja, tem doença cardíaca ou não tem doença cardíaca. O classificador, nesse caso, precisa de dados de treinamento para entender como as variáveis de entrada fornecidas estão relacionadas à classe. E uma vez que o classificador é treinado com precisão, ele pode ser usado para detectar se a doença cardíaca existe ou não para um determinado paciente.

Classificação

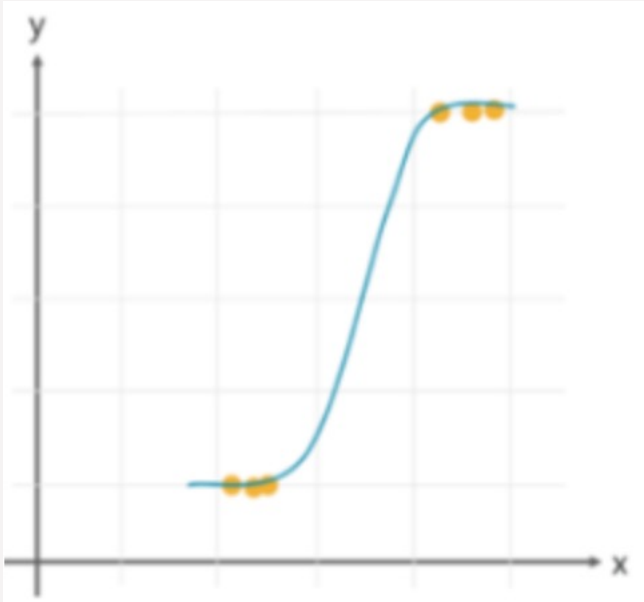
As classificações supervisionadas podem ser do tipo:

- **Classificação Binária** – É um tipo de classificação com apenas dois resultados, por exemplo – verdadeiro ou falso.
- **Classificação Multiclasse** – A classificação com mais de duas classes, na classificação multiclasse cada amostra é atribuída a apenas um rótulo.
- **Classificação Multi-rótulo** – Este é um tipo de classificação em que cada amostra pode ser atribuída a um conjunto de rótulos.

Classificação

Regressão Logística

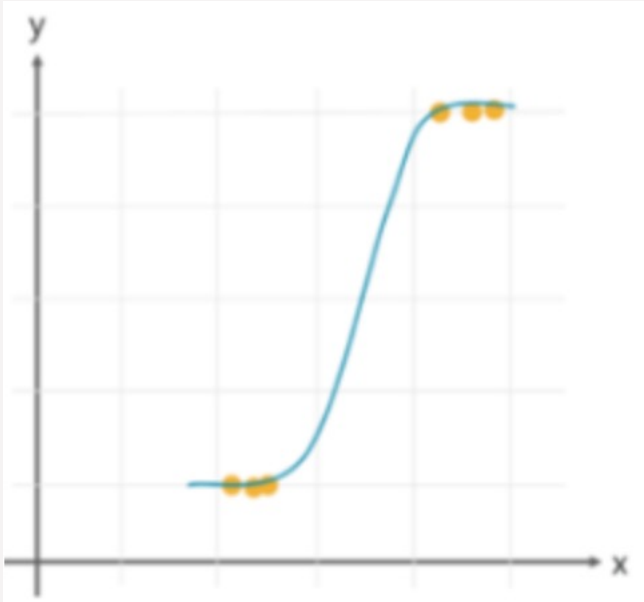
É um algoritmo de classificação em aprendizado de máquina que usa uma ou mais variáveis independentes para determinar um resultado. O resultado é medido com uma variável dicotômica, o que significa **que terá apenas dois resultados possíveis**



Classificação

Regressão Logística

O objetivo da regressão logística é encontrar uma relação de melhor ajuste entre a variável dependente e um conjunto de variáveis independentes.



Classificação

Regressão Logística - **Vantagens e desvantagens**

- A regressão logística destina-se especificamente à classificação, é útil para entender como um conjunto de variáveis independentes afeta o resultado da variável dependente.
- A principal desvantagem do algoritmo de regressão logística é que ele só funciona quando a variável predita é binária, ele assume que os dados estão livres de valores ausentes e assume que os preditores são independentes uns dos outros.

Classificação

Regressão Logística – Exemplo de uso

- Identificando fatores de risco para doenças
- Classificação de palavras
- Previsão do tempo
- Candidaturas de votação

Classificação

Classificador Naïve Bayes

É um algoritmo de classificação baseado no **teorema de Bayes**. Em termos simples, um classificador Naïve Bayes assume que a presença de uma característica particular em uma classe não está relacionada à presença de qualquer outra característica.

$$P(C_i|\, x_1, x_2 \ldots, x_n) = \frac{P(x_1, x_2 \ldots, x_n|C_i).P(C_i)}{P(x_1, x_2 \ldots, x_n)} \; for\; 1 < i < k$$

Classificação

Classificador Naïve Bayes - **Vantagens e desvantagens**

- O classificador Naïve Bayes requer uma pequena quantidade de dados de treinamento para estimar os parâmetros necessários para obter os resultados.
- Eles são extremamente rápidos por natureza em comparação com outros classificadores.
- Mesmo com uma abordagem simplista, Naïve Bayes é conhecido por superar a maioria dos métodos de classificação em aprendizado de máquina.
- A única desvantagem é que eles são conhecidos por serem um mau estimador.

Classificação

Classificador Naïve Bayes - **Vantagens e desvantagens**

- O classificador Naïve Bayes requer uma pequena quantidade de dados de treinamento para estimar os parâmetros necessários para obter os resultados.
- Eles são extremamente rápidos por natureza em comparação com outros classificadores.
- Mesmo com uma abordagem simplista, Naïve Bayes é conhecido por superar a maioria dos métodos de classificação em aprendizado de máquina.
- A única desvantagem é que eles são conhecidos por serem um mau estimador.

Classificação

Classificador Naïve Bayes – Exemplo de uso

- Previsões de doenças
- Classificação de documentos
- Filtros de spam
- Análise de sentimentos

Classificação

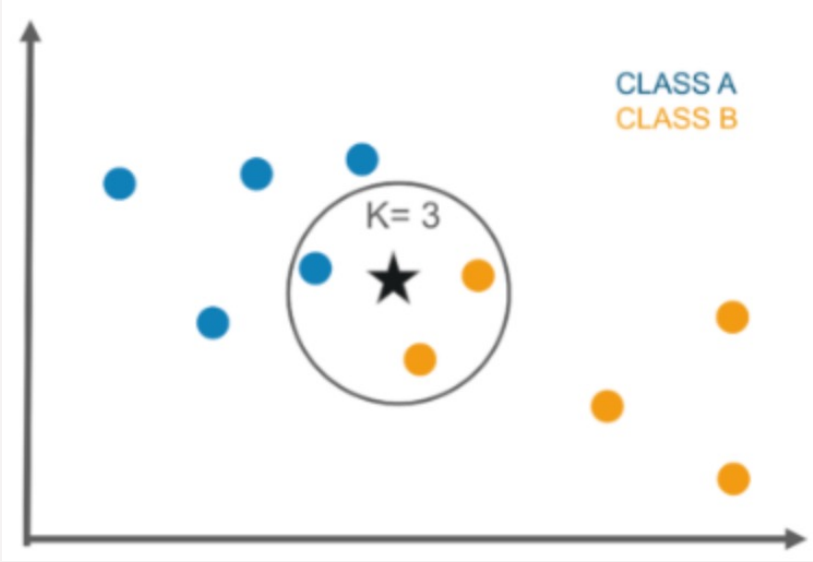
Classificador Naïve Bayes – Exemplo de uso

- Previsões de doenças
- Classificação de documentos
- Filtros de spam
- Análise de sentimentos

Classificação

K-vizinhos mais próximos (KNN)

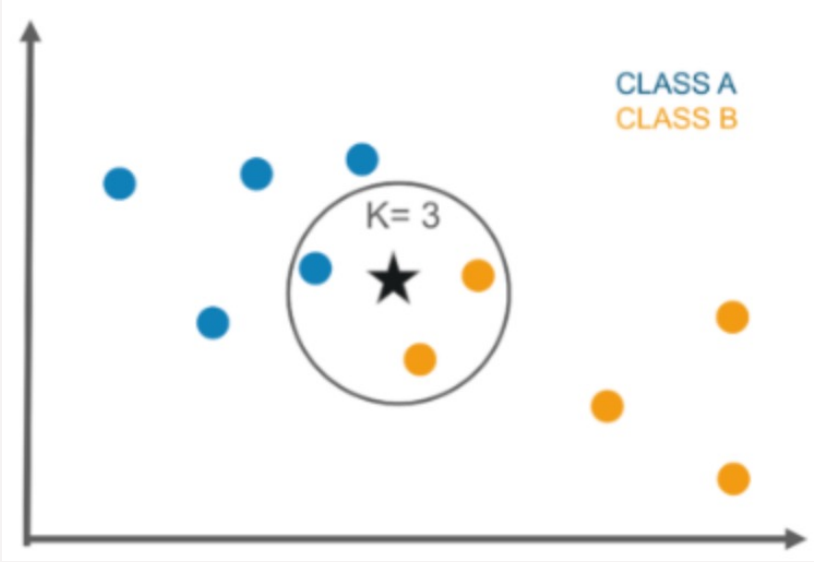
É um algoritmo de aprendizado lento que **armazena todas as instâncias correspondentes aos dados de treinamento no espaço n-dimensional** . É um **algoritmo de aprendizado "preguiçoso"** , pois não se concentra na construção de um modelo interno geral, em vez disso, funciona no armazenamento de instâncias de dados de treinamento.



Classificação

K-vizinhos mais próximos (KNN)

A classificação é calculada a partir de uma votação majoritária simples dos k vizinhos mais próximos de cada ponto.



Classificação

K-vizinhos mais próximos (KNN) - Vantagens e desvantagens

- O algoritmo é bastante simples em sua implementação e é robusto a dados de treinamento ruidosos.
- Mesmo que os dados de treinamento sejam grandes, é bastante eficiente.
- Uma desvantagem com o algoritmo KNN é que o custo computacional é bastante alto em comparação com outros algoritmos.

Classificação

K-vizinhos mais próximos (KNN) - Exemplos de uso

- Aplicações industriais para procurar tarefas semelhantes em comparação com outras
- Aplicativos de detecção de manuscrito
- Reconhecimento de imagem
- Reconhecimento de vídeo
- Análise de estoque

Classificação

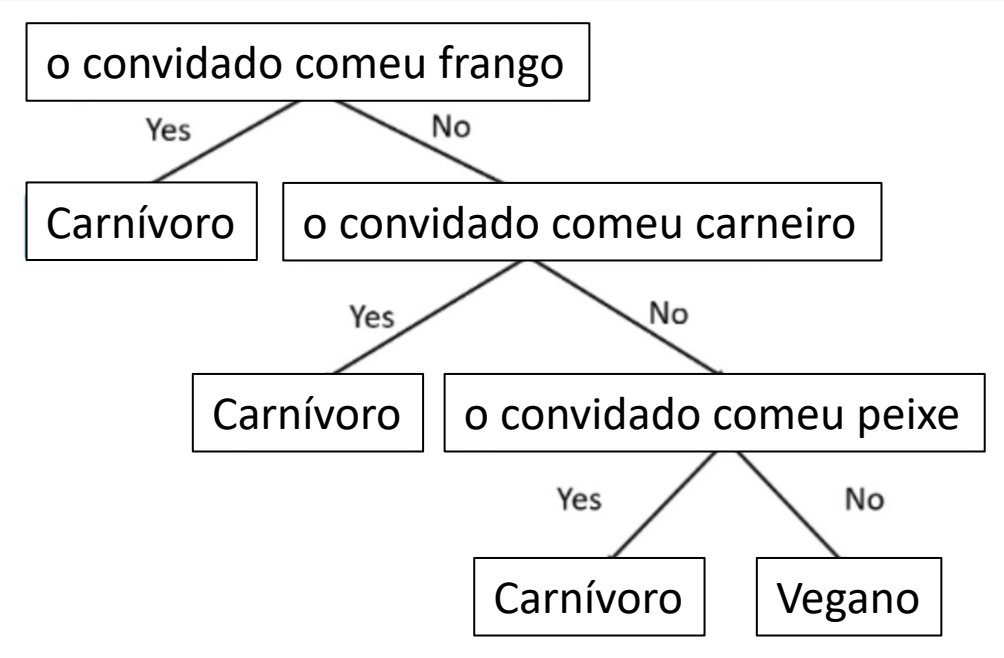
Árvore de decisão

O algoritmo de árvore de decisão constrói o modelo de classificação na forma de uma **estrutura de árvore** . Ele utiliza as regras se-então que são igualmente exaustivas e mutuamente exclusivas na classificação. O processo continua dividindo os dados em estruturas menores e, eventualmente, associando-os a uma árvore de decisão incremental. A estrutura final parece uma árvore com nós e folhas.

Classificação

Árvore de decisão

As **regras são aprendidas sequencialmente** usando os dados de treinamento um de cada vez. Cada vez que uma regra é aprendida, as tuplas que cobrem as regras são removidas. O processo continua no conjunto de treinamento até que o ponto final seja atingido.



Classificação

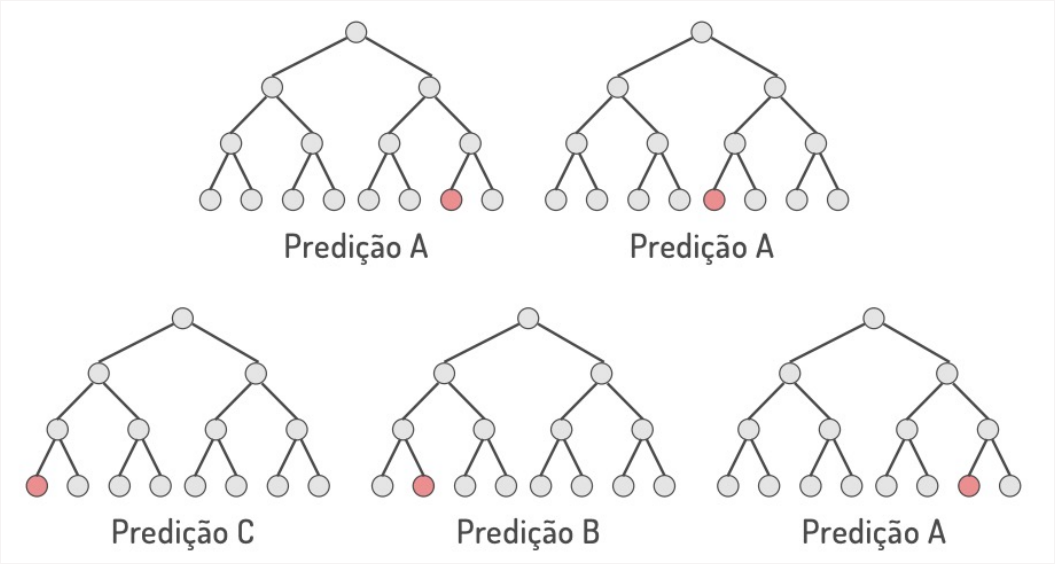
Árvore de decisão - Exemplos de uso

- Exploração de dados
- Reconhecimento de padrões
- Precificação de opções em finanças
- Identificando doenças e ameaças de risco

Classificação

Floresta Aleatória

Floresta aleatória é um **método de aprendizado conjunto** para classificação, regressão, etc. Ele opera construindo uma infinidade de árvores de decisão no tempo de treinamento e gera previsão média das árvores individuais.



Classificação

Floresta Aleatória - Vantagens e desvantagens

- A vantagem da floresta aleatória é que ela é mais precisa do que as árvores de decisão devido à redução do overfitting.
- A única desvantagem dos classificadores de floresta aleatória é que eles são bastante complexos na implementação e ficam muito lentos na previsão em tempo real.

Classificação

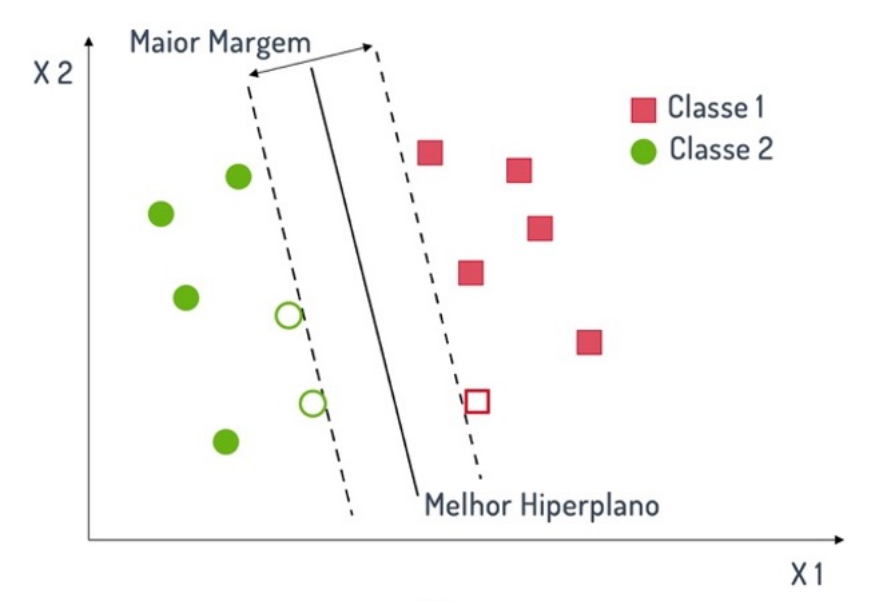
Floresta Aleatória - Exemplos de uso

- Aplicações industriais, como descobrir se um solicitante de empréstimo é de alto ou baixo risco
- Para prever a falha de peças mecânicas em motores de automóveis
- Previsão de pontuações de compartilhamento de mídia social
- Pontuações de desempenho

Classificação

Máquina de vetor de suporte

A máquina de vetores de suporte é um classificador que representa os **dados de treinamento como pontos no espaço** separados em categorias por um intervalo o mais amplo possível. Novos pontos são então adicionados ao espaço, prevendo em qual categoria eles se enquadram e a qual espaço eles pertencerão.



Classificação

Vantagens e desvantagens

- Ele usa um subconjunto de pontos de treinamento na função de decisão, o que o torna eficiente em termos de memória e é altamente eficaz em espaços de alta dimensão.
- A única desvantagem com a máquina de vetores de suporte é que o algoritmo não fornece estimativas de probabilidade diretamente.

Classificação

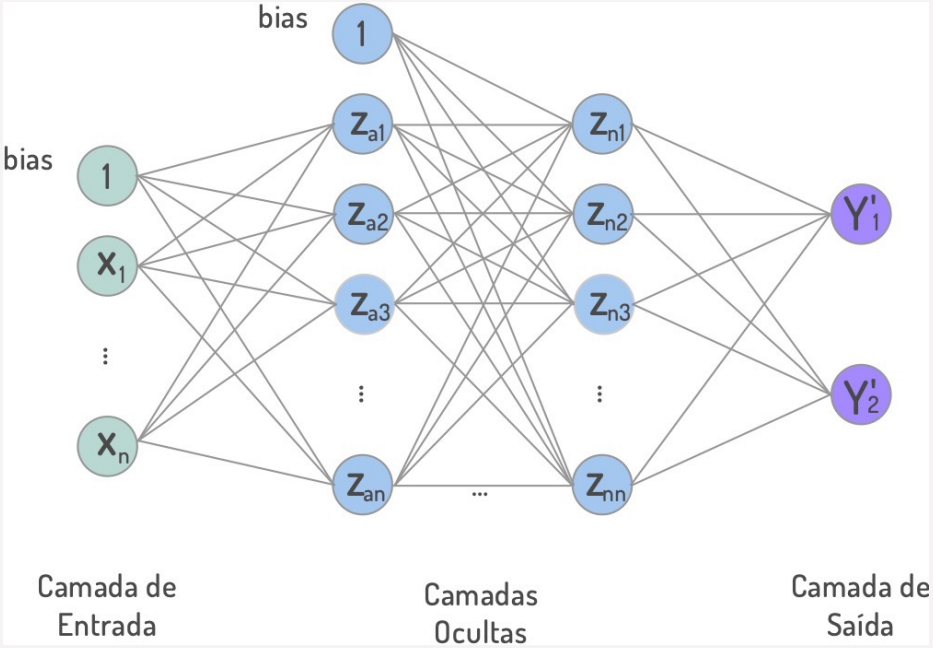
Vantagens e desvantagens

- Ele usa um subconjunto de pontos de treinamento na função de decisão, o que o torna eficiente em termos de memória e é altamente eficaz em espaços de alta dimensão.
- A única desvantagem com a máquina de vetores de suporte é que o algoritmo não fornece estimativas de probabilidade diretamente.

Classificação

Redes neurais artificiais

Uma rede neural consiste em neurônios que são **organizados em camadas** , elas usam um vetor de entrada e o convertem em uma saída. O processo envolve cada neurônio recebendo dados da entrada e aplicando uma função e em seguida passa a saída para a próxima camada.



Classificação

Redes neurais artificiais - Vantagens e desvantagens

- Ele tem uma alta tolerância a dados ruidosos e capaz de classificar padrões não treinados, tem melhor desempenho com entradas e saídas de valor contínuo.
- A desvantagem com as redes neurais artificiais é que tem uma interpretação pobre em comparação com outros modelos.

Classificação

Redes neurais artificiais - Scikit-learn

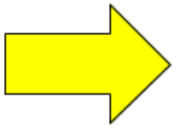
```
from sklearn.neural_network import MLPClassifier

model = MLPClassifier(hidden_layer_sizes=[10, 10], activation='relu',
                      solver='adam', alpha=0.001, learning_rate_init=0.001,
                      max_iter=300)
model.fit(X_train, y_train)
```

Classificação

Hot Encoding

One **Hot encoding** é uma transformação que fazemos nos dados para representarmos uma variável categórica de forma binária

Color
Red
Red
Yellow
Green
Yellow



Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1

Classificação

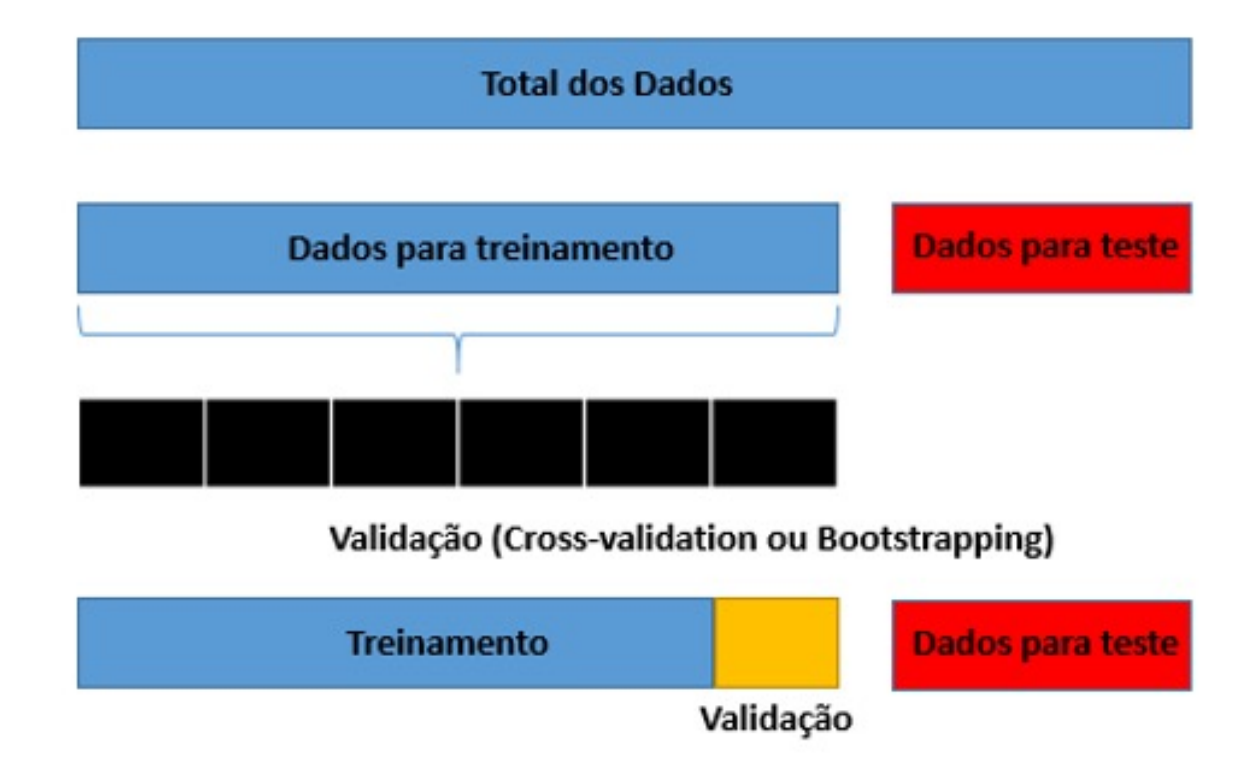
Divisão dos Dados - **Método Holdout**

Este é o método mais comum para avaliar um classificador. Neste método, o conjunto de dados fornecido é dividido em duas partes como um teste e um conjunto de treinamento 20% e 80%, respectivamente.

O conjunto de treinamento é usado para treinar os dados e o conjunto de teste não visto é usado para testar seu poder preditivo.

Classificação

Divisão dos Dados - Método Holdout



Classificação

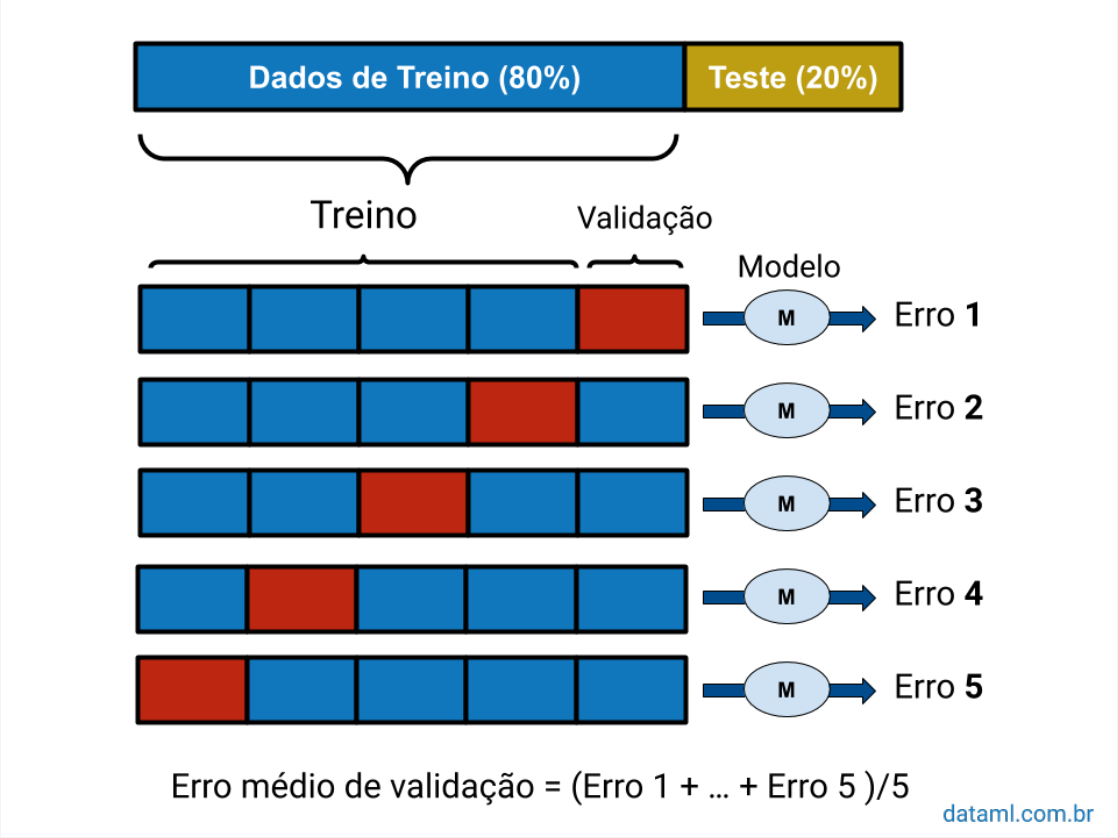
Divisão dos Dados - **Validação cruzada**

O over-fitting é o problema mais comum prevalente na maioria dos modelos de aprendizado de máquina. A validação cruzada K-fold pode ser realizada para verificar se o modelo está “superajustado”.

Neste método, o conjunto de dados é dividido aleatoriamente em **k** subconjuntos mutuamente exclusivos, cada um dos quais do mesmo tamanho. Destes, um é mantido para teste e outros são usados para treinar o modelo. O mesmo processo ocorre para todas as k dobras.

Classificação

Divisão dos Dados - Validação cruzada



Classificação

Métricas de Classificação

A matriz de confusão é uma apresentação útil da precisão de um modelo com duas ou mais classes, fornecendo uma matriz como saída e descrevendo o desempenho completo do modelo.

		Valor Verdadeiro	
		Doente	Saudável
Valor Predito	Doente	Verdadeiro Positivo - VP	Falso Positivo - FP
	Saudável	Falso Negativo - FN	Verdadeiro Negativo - VN

Classificação

Métricas de Classificação

A tabela apresenta previsões no eixo x e resultados de precisão no eixo y. As células da tabela são o número de previsões feitas por um algoritmo de aprendizado de máquina.

		Valor Verdadeiro	
		Doente	Saudável
Valor Predito	Doente	Verdadeiro Positivo - VP	Falso Positivo - FP
	Saudável	Falso Negativo - FN	Verdadeiro Negativo - VN

Classificação

Métricas de Classificação

$$\text{Sensibilidade ou Taxa de Verdadeiro Positivo} = \frac{VP}{VP + FN}$$

$$\text{Especificidade ou Taxa de Verdadeiro Negativo} = \frac{VN}{VN + FP}$$

$$\text{Precisão ou Valor Preditivo Positivo} = \frac{VP}{VP + VN}$$

$$\text{Valor Preditivo Negativo} = \frac{VN}{VN + FN}$$

Classificação

Métricas de Classificação - **Acurácia**

A acurácia é a métrica de avaliação mais utilizada em problemas de classificação; sendo muitas vezes utilizada de maneira incorreta. Ela é recomendada apenas quando há um número igual de observações em cada classe, ou seja, a base esteja balanceada, e que todas as previsões e erros de previsão sejam igualmente importantes

$$Acurácia = \frac{Número\ de\ predições\ corretas}{Número\ total\ de\ predições}$$

Classificação

Métricas de Classificação - **Escore F1**

Usado para medir a precisão de um teste, o escore F1 é a média harmônica entre precisão e sensibilidade. O intervalo para a pontuação F1 é [0, 1]. Ele mostra o quão preciso é o seu classificador (quantas instâncias ele classifica corretamente) e também o quão robusto é (não perde um número significativo de instâncias).

$$F1 = 2 * \frac{1}{\left(1/Sensibilidade\right) + \left(1/Precisão\right)}$$

Classificação

Métricas de Classificação – Área sob a curva ROC

A área sob a curva ROC (ou AUC para abreviar) é uma métrica de desempenho para problemas de classificação binária.

A AUC representa a capacidade de um modelo de discriminar entre classes positivas e negativas. Uma área de 1,0 representa um modelo que fez todas as previsões perfeitamente. Uma área de 0,5 representa um modelo tão bom quanto aleatório.

Classificação

Métricas de Classificação – Área sob a curva ROC