

Machine Learning

Prof. Tiago Lima

Sobre o Palestrante

- **Bacharel** em Ciência da Computação (**UNICAP**).
- **Doutorado e Mestrado** em Ciência da Computação (**CIn-UFPE**).
- **Autor** de diversas publicações, além de **revisor** de artigos para conferências relacionadas à **Computação Inteligente**.
- Atualmente **aplica Computação Inteligente** em **problemas de saúde** através de parcerias com o **IMIP** e **HCP**.

Roteiro

- Teoria
 - Introdução
 - Dados
 - Modelos
 - Avaliação
- Prática
 - Orange Data Mining
 - Instalação
 - Widgets e Canvas
 - Carregando seus Dados
 - Fazendo Previsões

Teoria

Introdução

Problemas são resolvidos em computação por meio da escrita de **algoritmos**, que especificam o passo a passo de como um problema pode ser resolvido. No entanto, não é fácil escrever programas que realizem com eficiência algumas tarefas do nosso dia a dia, como reconhecer pessoas pelo rosto ou pela fala.

Introdução

Que **características** dos rostos ou da fala serão **consideradas**?

- Detecção de Face:
 - O que fazer para diferentes expressões faciais de uma pessoa?
 - O que fazer com alterações, como o uso de óculos ou barba?
- Detecção de Fala:
 - O que fazer com mudanças na voz por uma gripe ou estado de espírito?

Introdução

Seres humanos conseguem realizar essas tarefas com relativa facilidade. Fazem isso por meio de **reconhecimento de padrões**, quando aprendem o que deve ser observado em um rosto ou na fala para conseguir identificar pessoas após terem tido vários exemplos de rostos ou falas com identificação clara.

Introdução

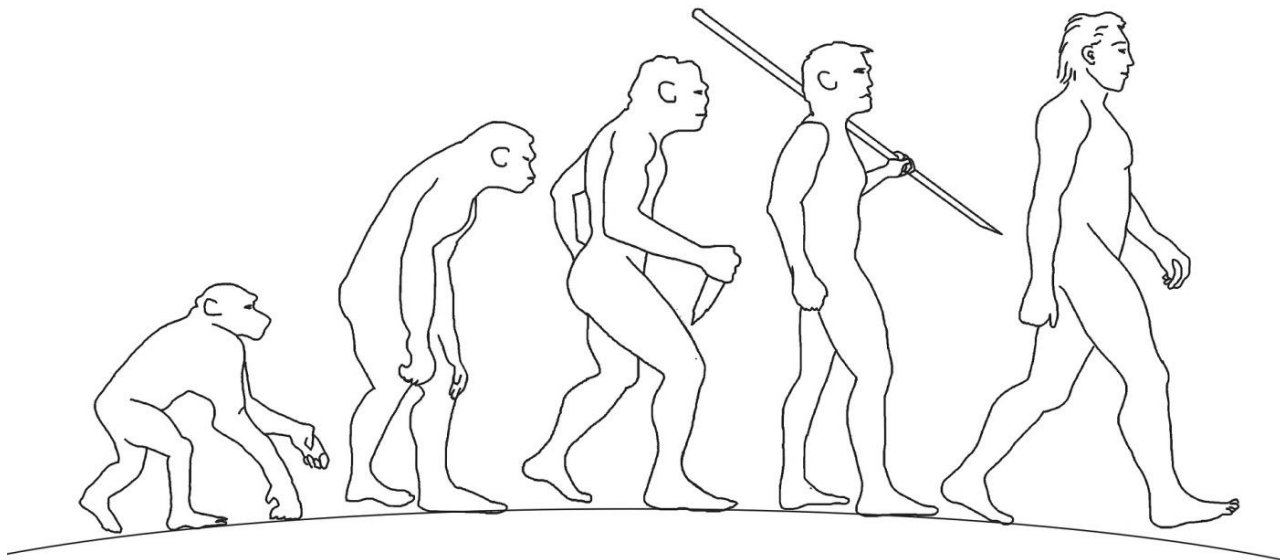
Apesar da **dificuldade** (ou até mesmo impossibilidade) de **escrever um algoritmo** que possa lidar de maneira eficiente com tarefas de **reconhecimento de padrões**, o número de vezes que essas tarefas precisam ser realizadas diariamente é muito grande. Aliado a isso, o volume de informações considerado torna difícil (ou impossível) a realização por seres humanos.

Inteligência Artificial

Técnicas de Inteligência Artificial, em particular Aprendizado de Máquina, têm sido utilizadas com sucesso em um grande número de problemas reais no reconhecimento de padrões



Inteligência



Inteligência Artificial

- **No início**, o processo de aprendizado (ou aquisição do conhecimento) envolvia entrevistas com especialistas para descobrir regras. Esses programas eram conhecidos como **Sistemas Especialistas**.
- Com a **crescente complexidade dos problemas** a serem tratados e do volume de dados, **torna-se clara a necessidade de ferramentas mais autônomas**, reduzindo intervenção humana e dependência de especialistas.
- Para isso, essas técnicas deveriam ser capazes de **criar por si próprias**, a partir da experiência, uma hipótese ou função capaz de resolver o problema.

Aprendizado de Máquina (AM)

“A capacidade de melhorar o desempenho na realização de alguma tarefa por meio da experiência”.

Mitchel, 1997.

Aprendizado Indutivo




Conjunto de Dados

- A cada dia uma quantidade enorme de dados é gerada. Uma estimativa diz que a cada 20 meses dobra a quantidade de dados armazenada nos bancos de dados do mundo [Witten *et al*, 2011]. Esses são oriundos de transações financeiras, monitoramento ambiental, dados clínicos, imagens, etc.
- Conjuntos de dados são formados por informações que podem representar um objeto físico, como uma pessoa, ou uma noção abstrata, como os sintomas apresentados por um paciente.

Conjunto de Dados

- Os dados podem ser representados por uma matriz de objetos: $X_{n \times d}$, em que **n** é o número de objetos e **d** é o número de atributos de cada objeto.
- Cada elemento dessa matriz de objetos $x_{i,j}$, contém o valor da j-ésima característica para o i-ésimo objeto.

Conjunto de Dados

$X_{n \times d}$ 

n	Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
	4201	João	28	M	79	Concentradas	38,0	2	SP	Doente
	3217	Maria	18	F	67	Inexistentes	39,5	4	MG	Doente
	4039	Luiz	49	M	92	Espalhadas	38,0	2	RS	Saudável
	1920	José	18	M	43	Inexistentes	38,5	8	MG	Doente
	4340	Cláudia	21	F	52	Uniformes	37,6	1	PE	Saudável
	2301	Ana	22	F	72	Inexistentes	38,0	3	RJ	Doente
	1322	Marta	19	F	87	Espalhadas	39,0	6	AM	Doente
	3027	Paulo	34	M	67	Uniformes	38,4	2	GO	Saudável

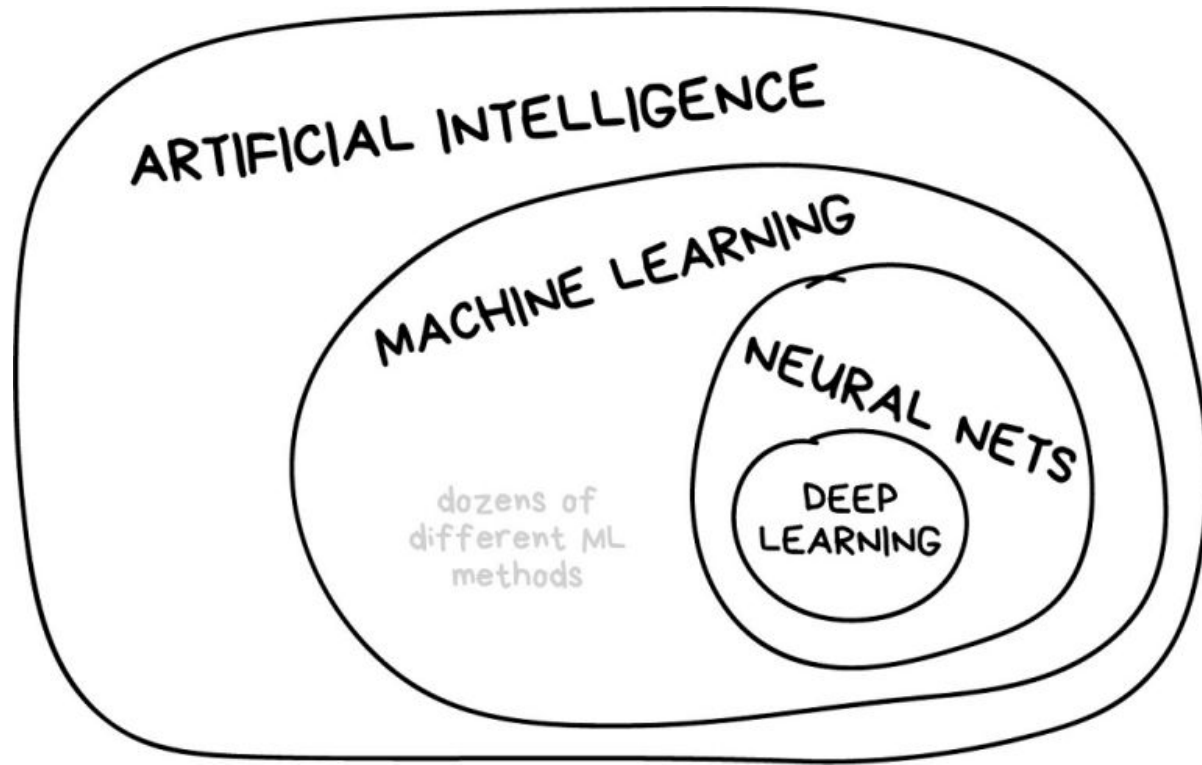
d

Conjunto de Dados

Técnicas de pré-processamento de dados são frequentemente utilizadas para tornar os conjuntos de dados mais adequados para o uso em algoritmos de AM.

- Eliminação manual de atributos
- Integração de dados
- Amostragem de dados
- Balanceamento de dados
- Limpeza de dados
- Redução da dimensionalidade
- Transformação de dados

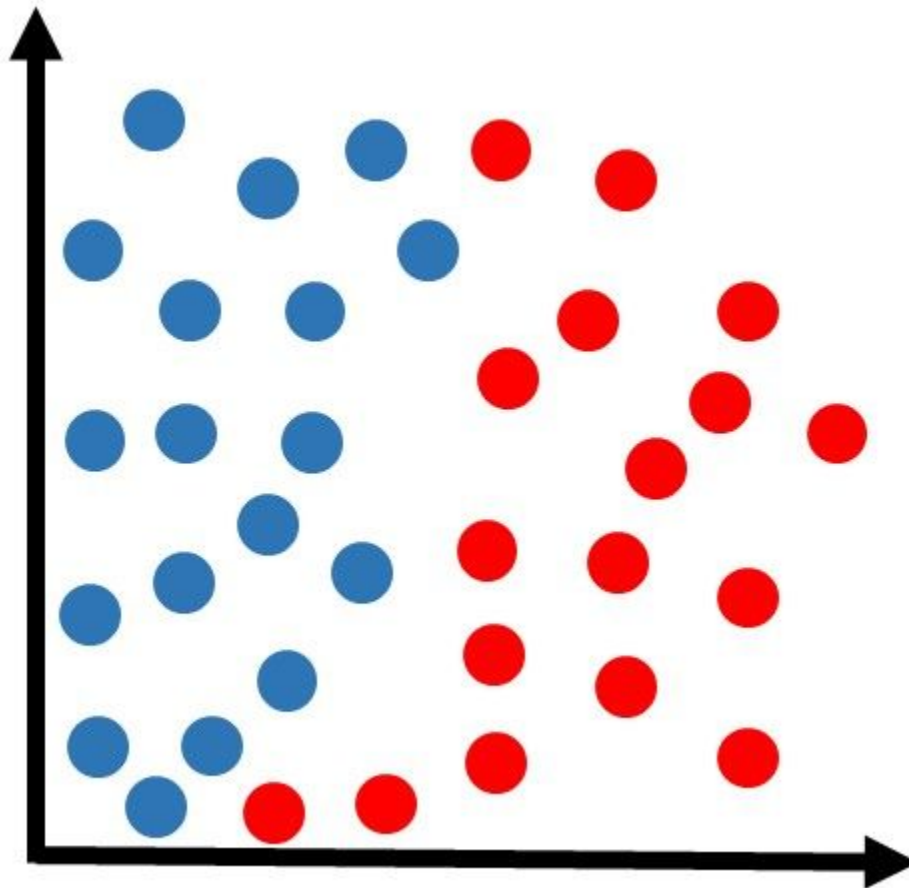
Modelos



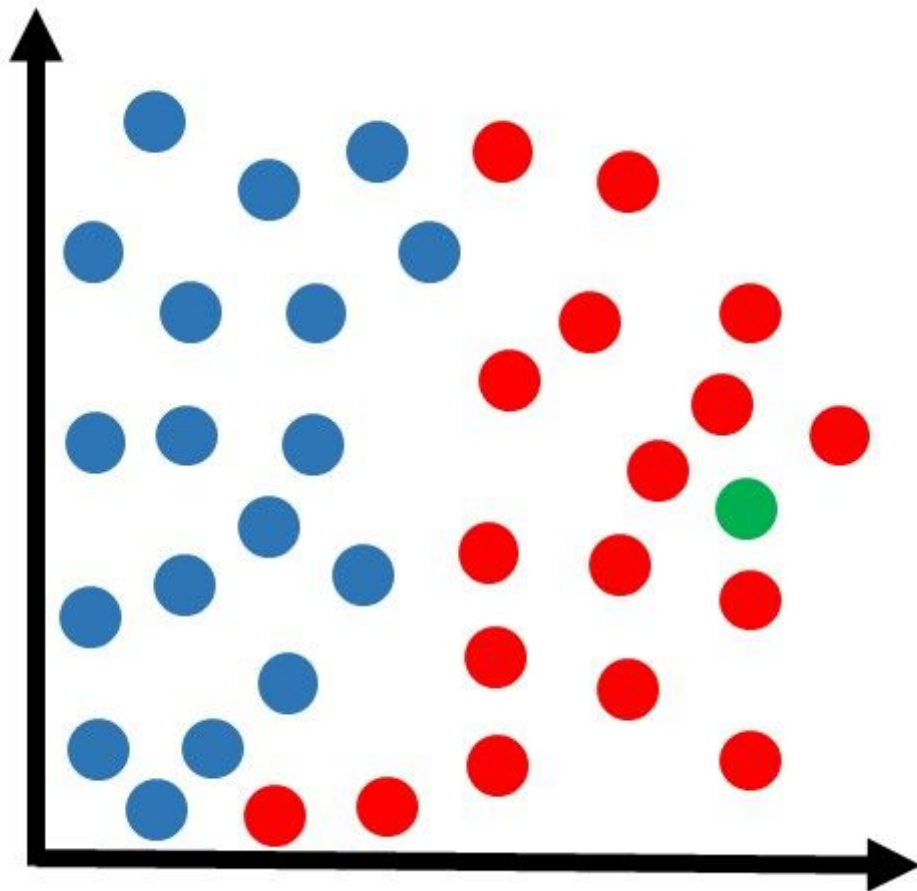
Modelos - KNN

- K-nearest neighbors, ou “K-vizinhos mais próximos”, costuma ser um dos primeiros algoritmos aprendidos por iniciantes no mundo do AM.
- Classificar cada amostra desconhecida avaliando sua distância em relação aos vizinhos mais próximos. Se a vizinhança for majoritariamente de uma classe, a amostra em questão será classificada nesta categoria.

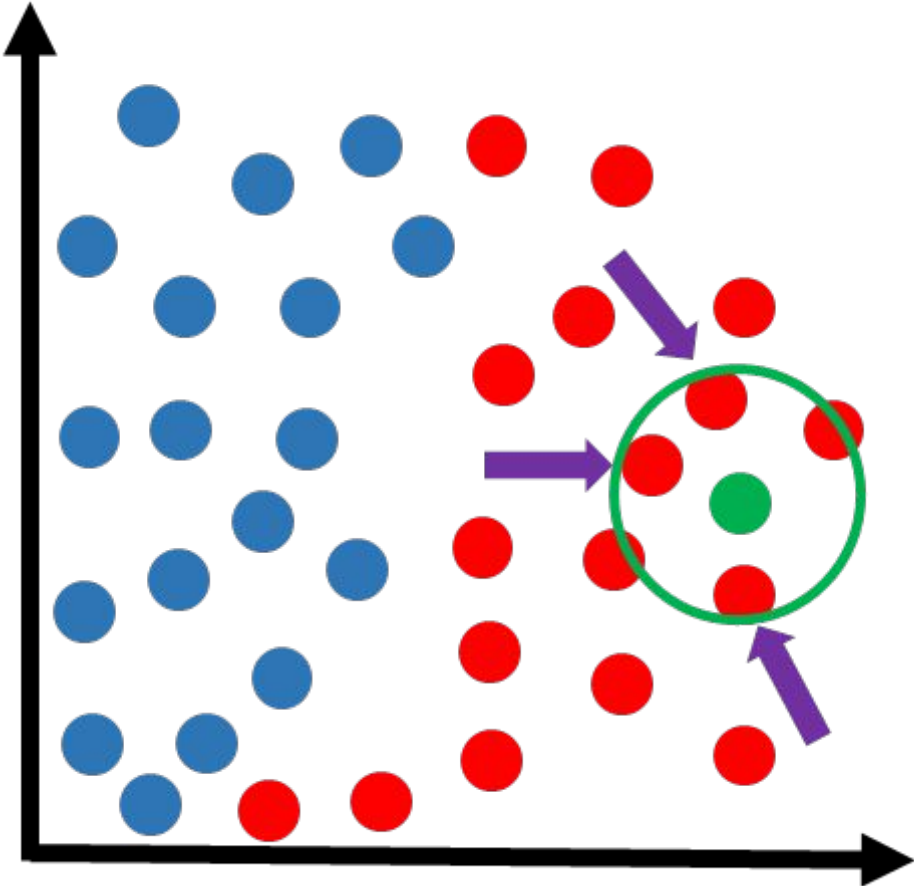
Modelos - KNN



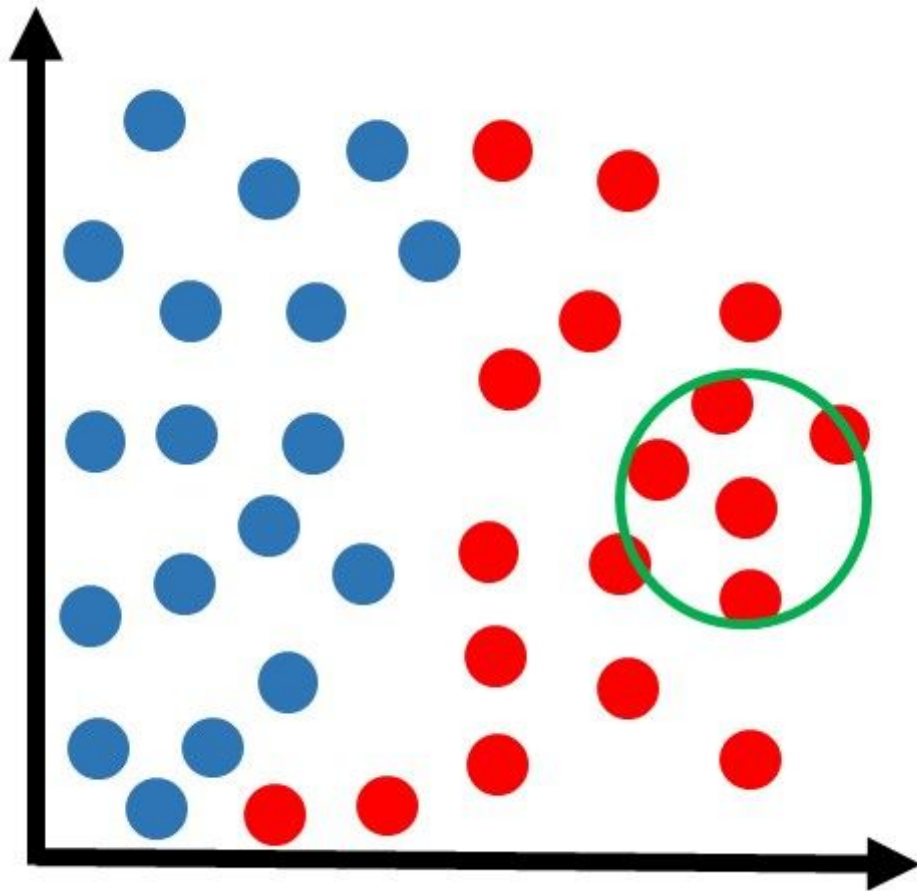
Modelos - KNN



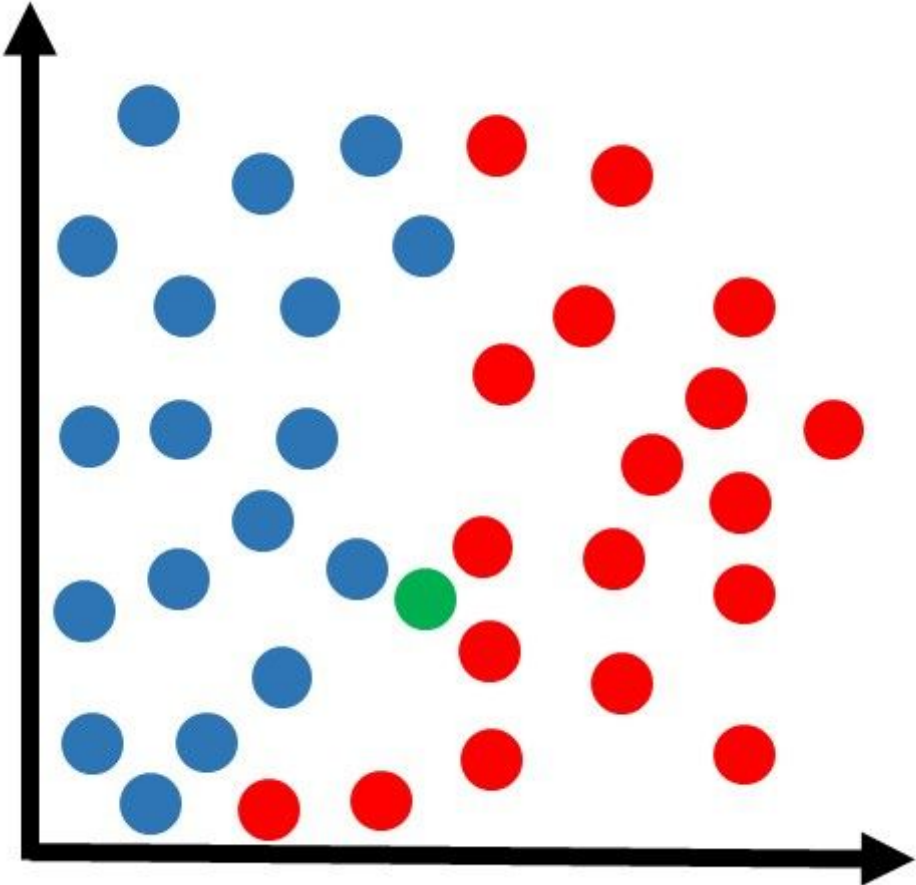
Modelos - KNN



Modelos - KNN

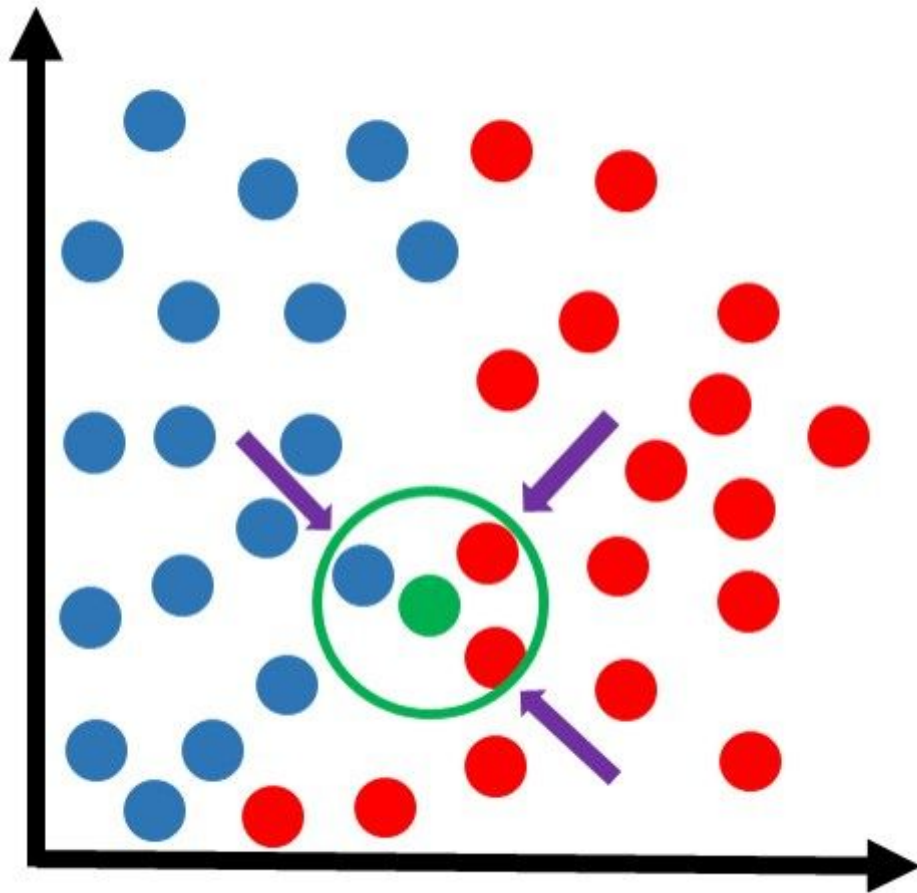


Modelos - KNN



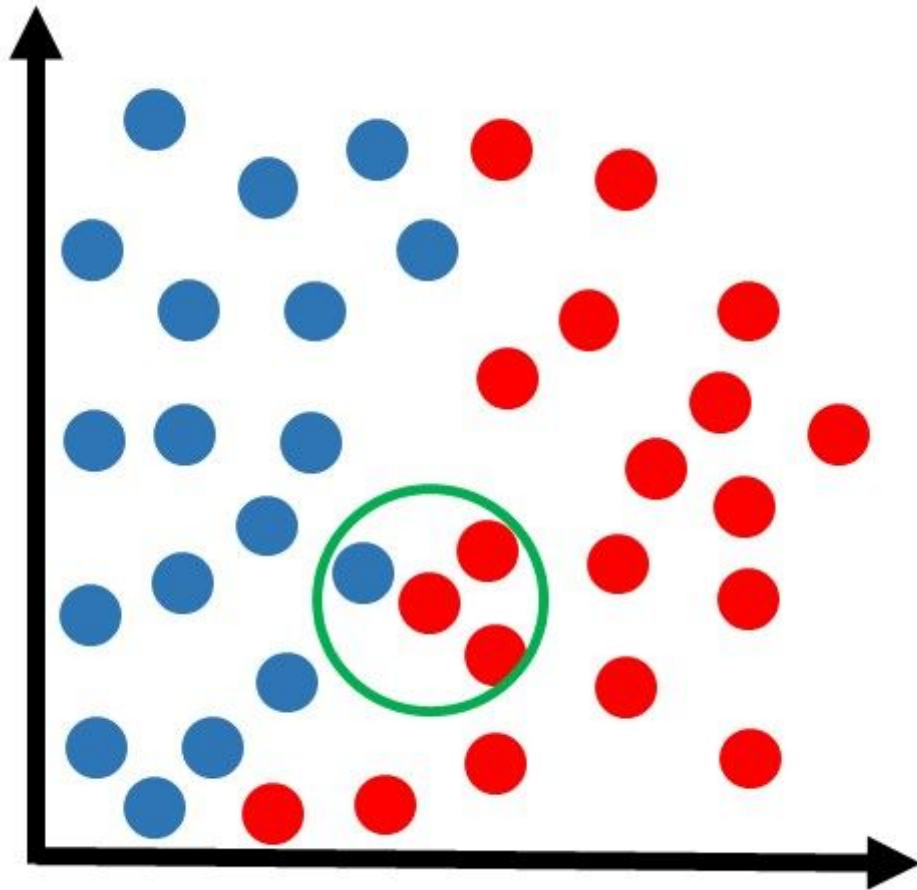
Modelos - KNN

K=3



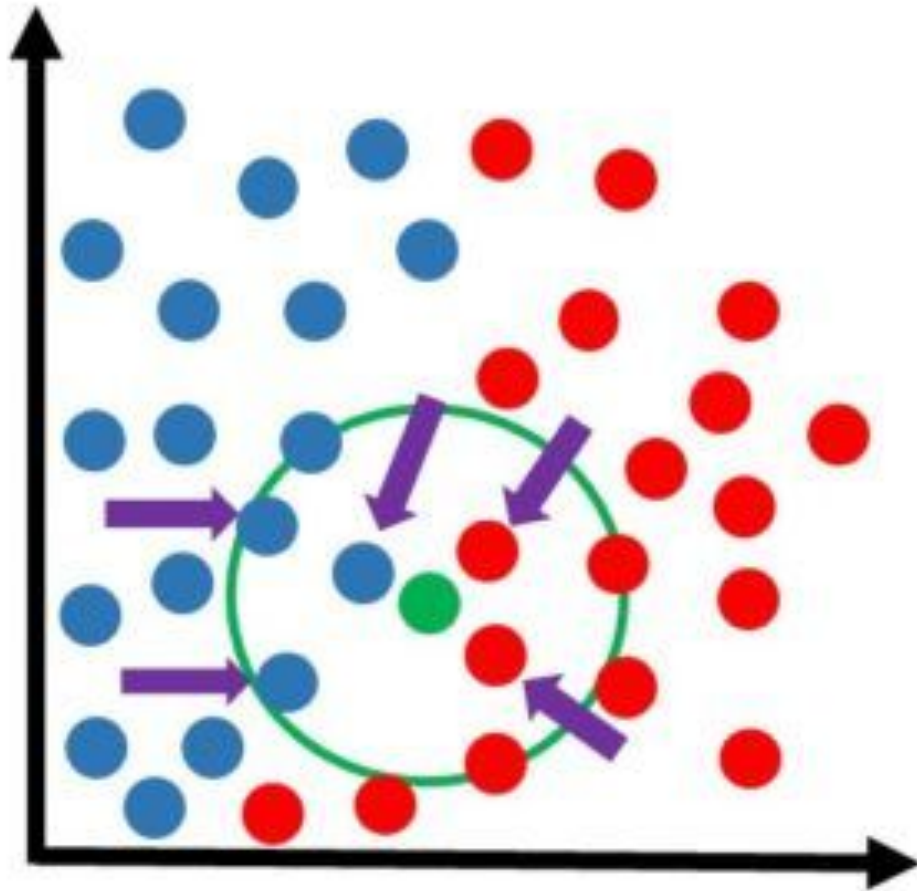
Modelos - KNN

K=3



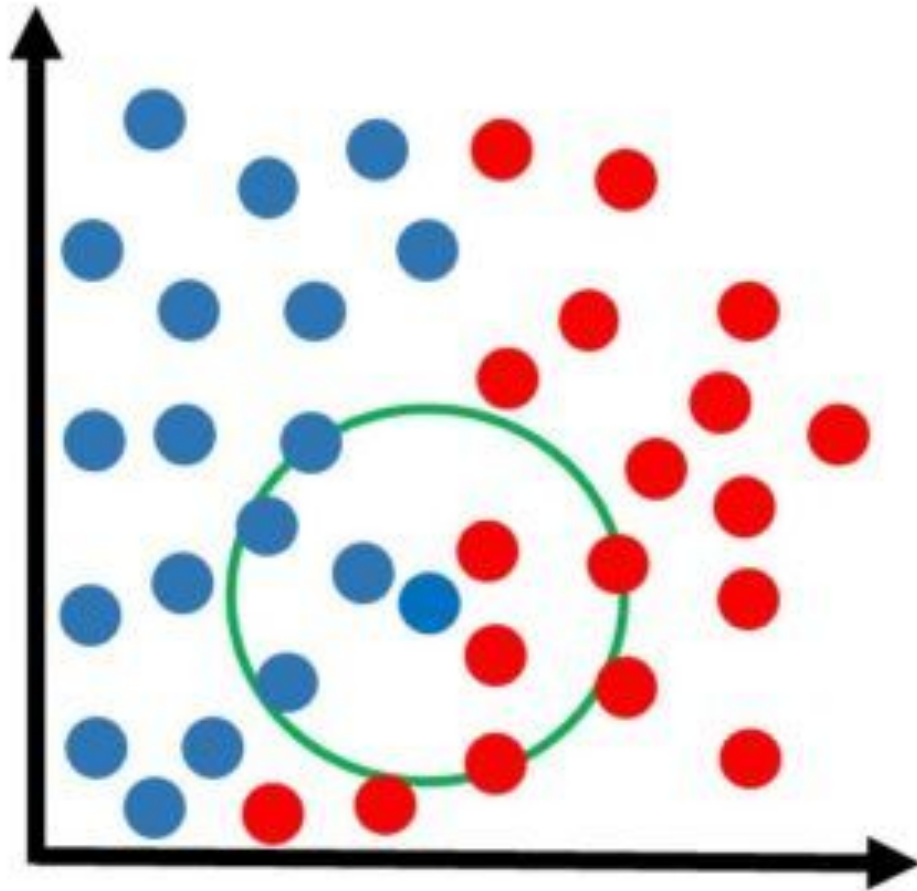
Modelos - KNN

K=5



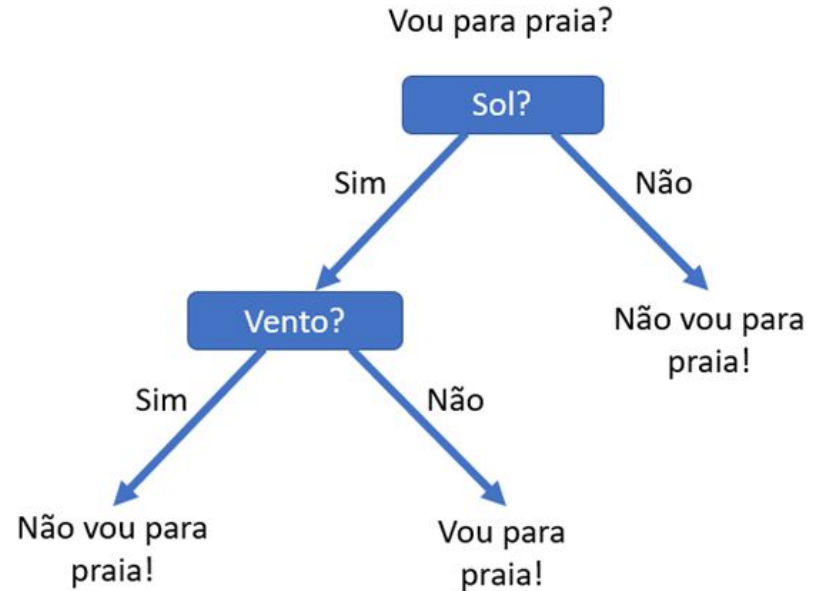
Modelos - KNN

K=5



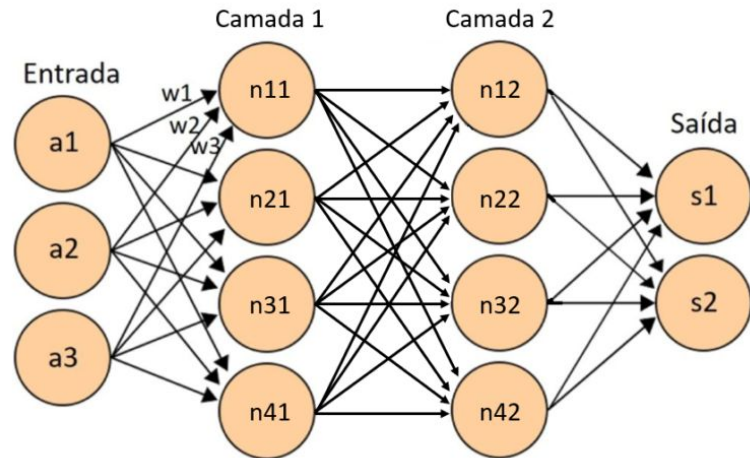
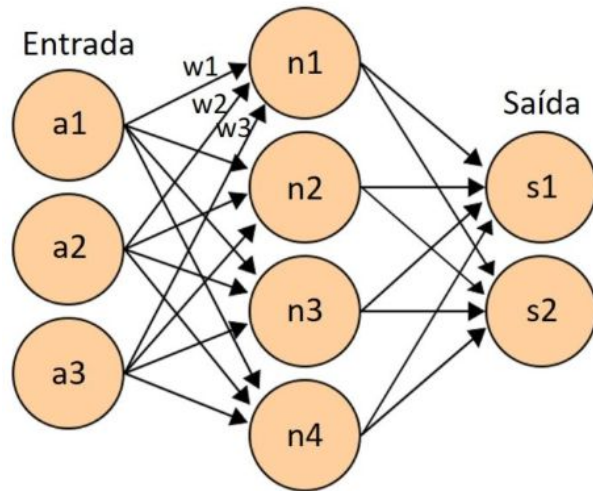
Modelos - Tree

As árvores de Decisão, como o próprio nome sugere, nesse modelo vários pontos de decisão serão criados. Estes pontos são os “nós” da árvore e em cada um deles o resultado da decisão será seguir por um caminho, ou por outro. Os caminhos existentes são os “ramos”.



Modelos - Neural Networks

De maneira simplificada, podemos enxergar uma rede neural como uma estrutura que conecta pequenas unidades, os neurônios, de forma organizada. Através desta organização, a combinação das operações unitárias simples realizadas por cada neurônio levará a soluções de problemas complexos.



Avaliação - Treino e Teste



Avaliação - KFOLD Cross Validation



Avaliação

K = 5

Testing	Training	Training	Training	Training
Training	Testing	Training	Training	Training
Training	Training	Testing	Training	Training
Training	Training	Training	Testing	Training
Training	Training	Training	Training	Testing
$R^2 = 0.73$	$R^2 = 0.71$	$R^2 = 0.75$	$R^2 = 0.79$	$R^2 = 0.70$





$$\bar{R}^2 = \frac{\sum_{i=1}^k R_i^2}{k}$$

$$\frac{0.73 + 0.71 + 0.75 + 0.79 + 0.70}{5} = 0.736$$

Avaliação - Matriz de Confusão

		P R E D I T O	
R E A L		 POSITIVO	 NEGATIVO
	 POSITIVO	  TP verdadeiro positivo	  FN falso negativo
	 NEGATIVO	  FP falso positivo	  TN verdadeiro negativo

Avaliação - Matriz de Confusão

		P R E D I T O	
		 POSITIVO	 NEGATIVO
R E A L	 POSITIVO	231 [80%]	57 [20%]
	 NEGATIVO	129 [09%]	1329 [91%]

Avaliação - Acurácia

A acurácia, ou *accuracy* em inglês, nos diz quantos de nossos exemplos foram de fato classificados corretamente, independente da classe. Por exemplo, se temos 100 observações e 90 delas foram classificadas corretamente, nosso modelo possui uma acurácia de 90%. A acurácia é definida pela fórmula abaixo:

$$\text{Acurácia} = \frac{\checkmark \text{👍 TP} + \checkmark \text{👎 TN}}{\checkmark \text{👍 TP} + \checkmark \text{👎 TN} + \text{✗} \text{👍 FP} + \text{✗} \text{👎 FN}}$$

Avaliação - Acurácia

Em alguns problemas a **acurácia pode ser elevada mas, ainda assim, o modelo pode ter uma performance inadequada**. Por exemplo, considere o modelo que classifica exames de câncer entre positivo ou negativo para a doença, e em nosso conjunto de dados temos 1000 exemplos, sendo 990 de pacientes sem câncer e 10 de pacientes com câncer. Caso nosso modelo seja ingênuo e sempre classifique todos os exemplos com negativo (sem câncer), ele ainda obteria uma acurácia de 99%. O que parece uma excelente métrica, mas na verdade não estamos avaliando nosso modelo de forma adequada. Para melhor avaliar modelos que lidam com conjuntos de dados desbalanceados como este, outras métricas que serão apresentadas em seguida devem ser utilizadas.

Avaliação - Precisão

A precisão, ou *precision* em inglês, também é uma das métricas mais comuns para avaliar modelos de classificação. Esta métrica é definida pela razão entre a quantidade de exemplos classificados corretamente como positivos e o total de exemplos classificados como positivos, conforme a fórmula abaixo:

$$\text{Precisão} = \frac{\text{✓👍 TP}}{\text{✓👍 TP} + \text{✗👍 FP}}$$

Avaliação - Precisão

A precisão dá ênfase maior para os erros por falso positivo. Podemos entender a precisão como sendo a expressão matemática para a pergunta: dos exemplos classificados como positivos, quantos realmente são positivos? Voltando ao exemplo do modelo de câncer, se o valor para a precisão fosse de 90%, isto indicaria que a cada 100 pacientes classificados como positivo, é esperado que apenas 90 tenham de fato a doença.

Avaliação - Revocação

Ao contrário da precisão, a revocação, ou *recall* em inglês e também conhecida como sensibilidade ou taxa de verdadeiro positivo (TPR), dá maior ênfase para os erros por falso negativo. Esta métrica é definida pela razão entre a quantidade de exemplos classificados corretamente como positivos e a quantidade de exemplos que são de fato positivos, conforme a fórmula abaixo:

$$\text{Revocação} = \frac{\text{✅👍 TP}}{\text{✅👍 TP} + \text{❌👎 FN}}$$

Avaliação - Revocação

A revocação busca responder a seguinte pergunta: de todos os exemplos que são positivos, quantos foram classificados corretamente como positivos? Considerando o exemplo do modelo de câncer, se o valor para a revocação fosse de 95%, isto indicaria que a cada 100 pacientes que são de fato positivos, é esperado que apenas 95 sejam corretamente identificados como doentes.

Prática

Orange Data Mining



Orange Data Mining - Instalação



Windows



macOS



Linux / Source

Download the latest version for Windows

[Download Orange 3.33.0](#)

Standalone installer (default)

[Orange3-3.33.0-Miniconda-x86_64.exe \(64 bit\)](#)

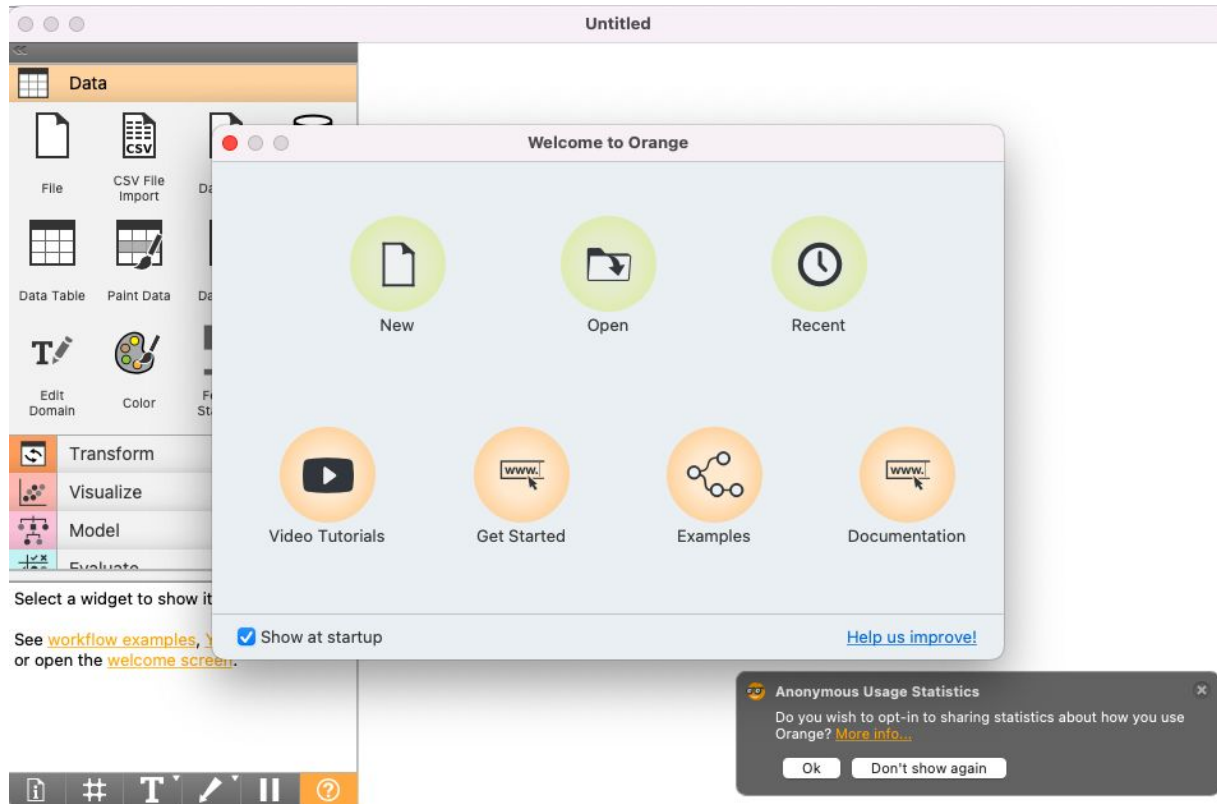
Can be used without administrative privileges.

Portable Orange

[Orange3-3.33.0.zip](#)

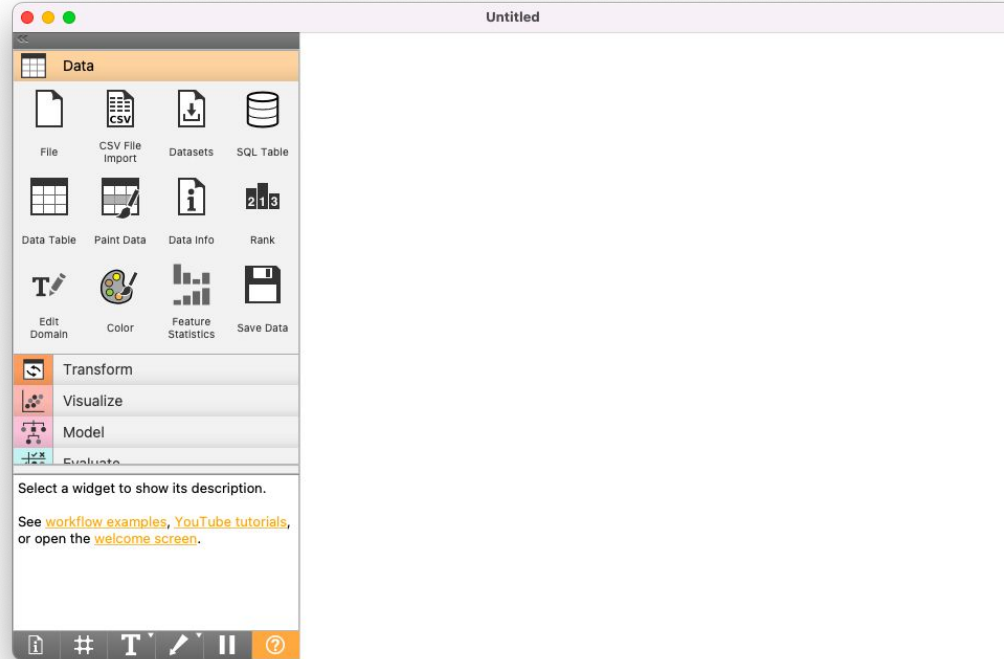
No installation needed. Just extract the archive and open the shortcut in the extracted folder.

Orange Data Mining - Boas Vindas



Orange Data Mining - New

Widgets

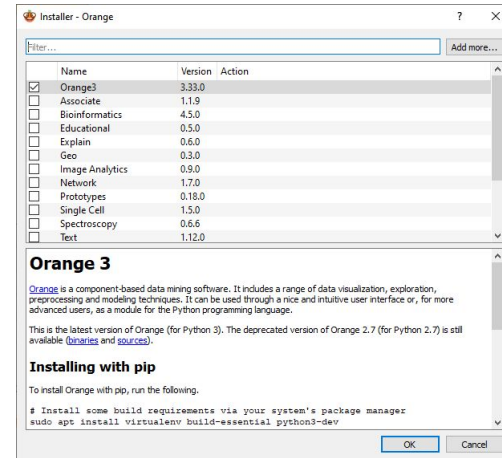
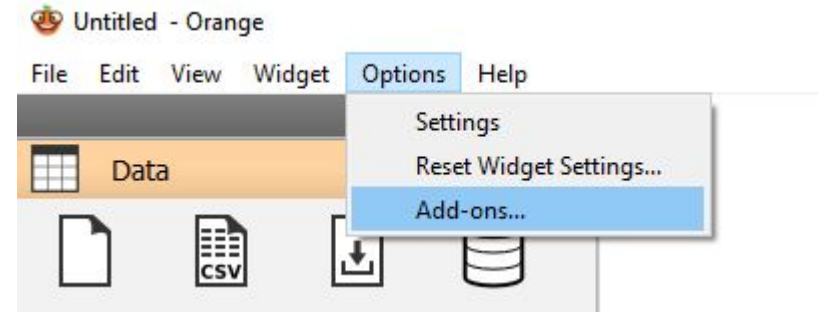
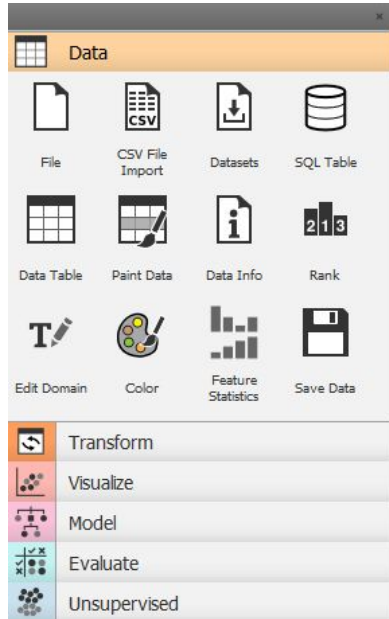


Canvas



Orange Data Mining - Widgets

São as **unidades computacionais** do Orange:
Lêem dados, processam-nos, visualizam-nos, fazem clustering, constroem modelos preditivos, etc.



Orange Data Mining - Widget “File” (Arquivo)

The screenshot displays the Orange Data Mining software interface. On the left is a widget palette with categories: Data, Transform, Visualize, Model, Evaluate, and Unsupervised. The 'Data' category is selected, showing various data source widgets. The 'File' widget is highlighted with a mouse cursor. The main workspace contains a 'File - Orange' widget configuration window. This window shows the 'Source' set to 'File: iris.tab', 'File Type' set to 'Automatically detect type', and 'Info' details for the 'Iris flower dataset'. Below the info, a table lists the dataset's columns with their names, types, roles, and values.

File - Orange

Source

☒ File: iris.tab

☐ URL:

File Type

Automatically detect type

Info

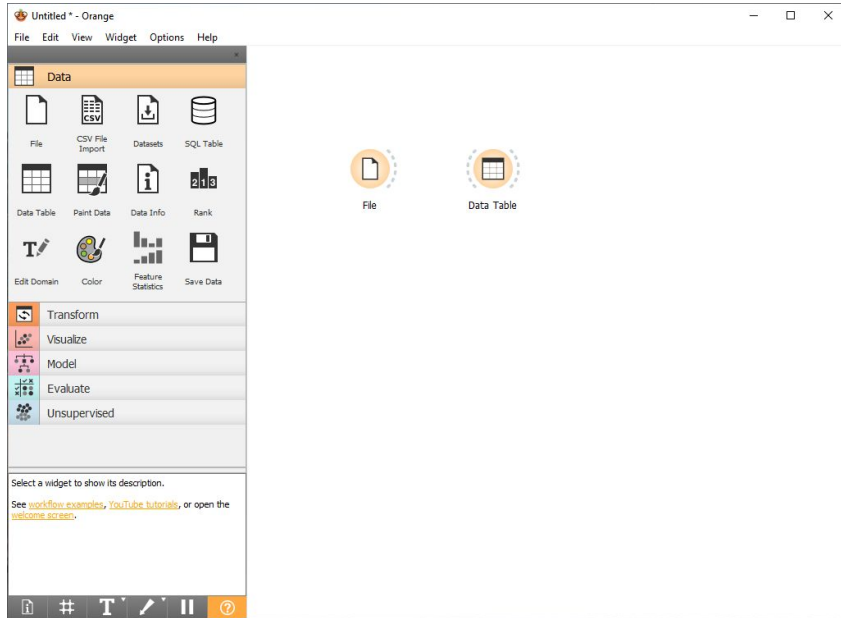
Iris flower dataset
Classical dataset with 150 instances of Iris setosa, Iris virginica and Iris versicolor.
150 instance(s)
4 feature(s) (no missing values)
Classification; categorical class with 3 values (no missing values)
0 meta attribute(s)

Columns (Double click to edit)

	Name	Type	Role	Values
1	sepal length	N numeric	feature	
2	sepal width	N numeric	feature	
3	petal length	N numeric	feature	
4	petal width	N numeric	feature	
5	iris	C categorical	target	Iris-setosa, Iris-versicolor, Iris-virginica

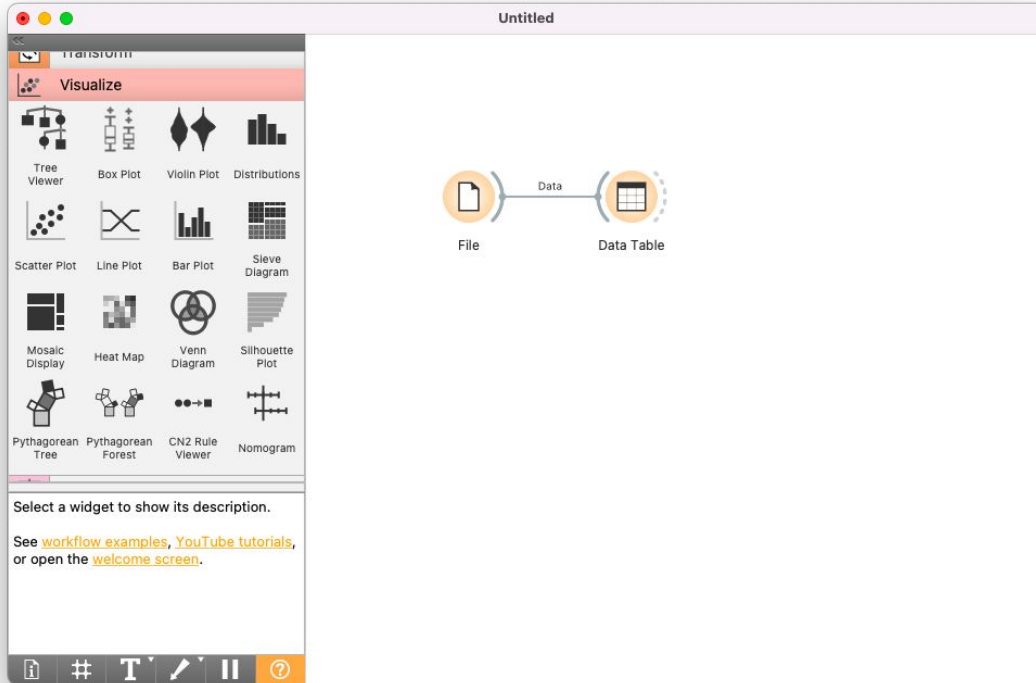
? | 150

Orange Data Mining - Widget “Data Table”



- Perceba que Widgets podem ter um canal de entrada, um canal de saída, ou ambos.
- Para alimentar os dados do Widget File para o Widget Data Table, arraste uma linha a partir do lado da saída do primeiro à entrada do segundo.

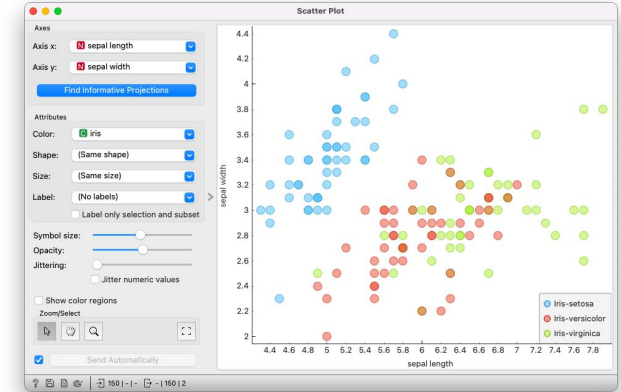
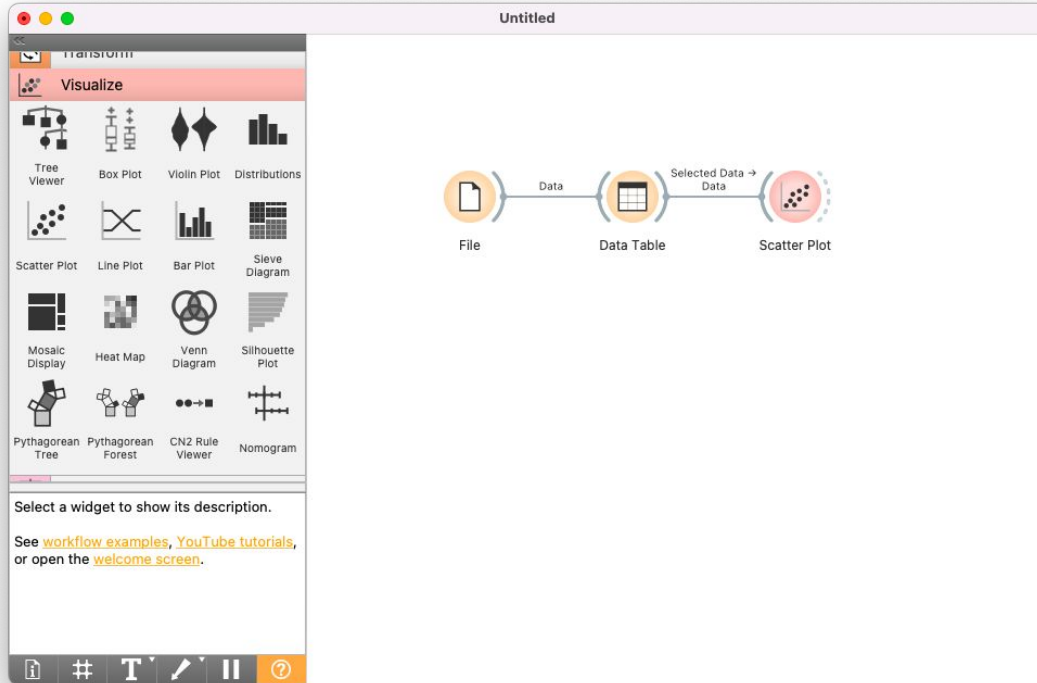
Orange Data Mining - Widget “Data Table”



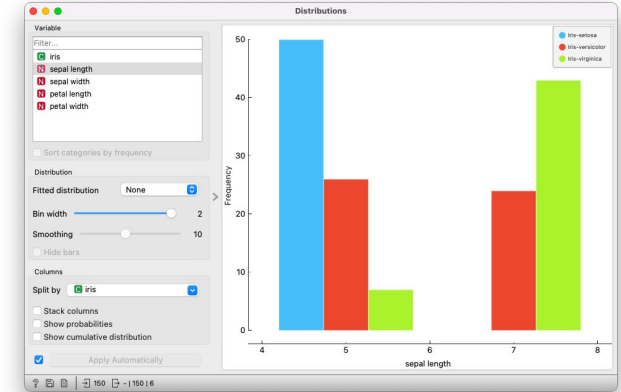
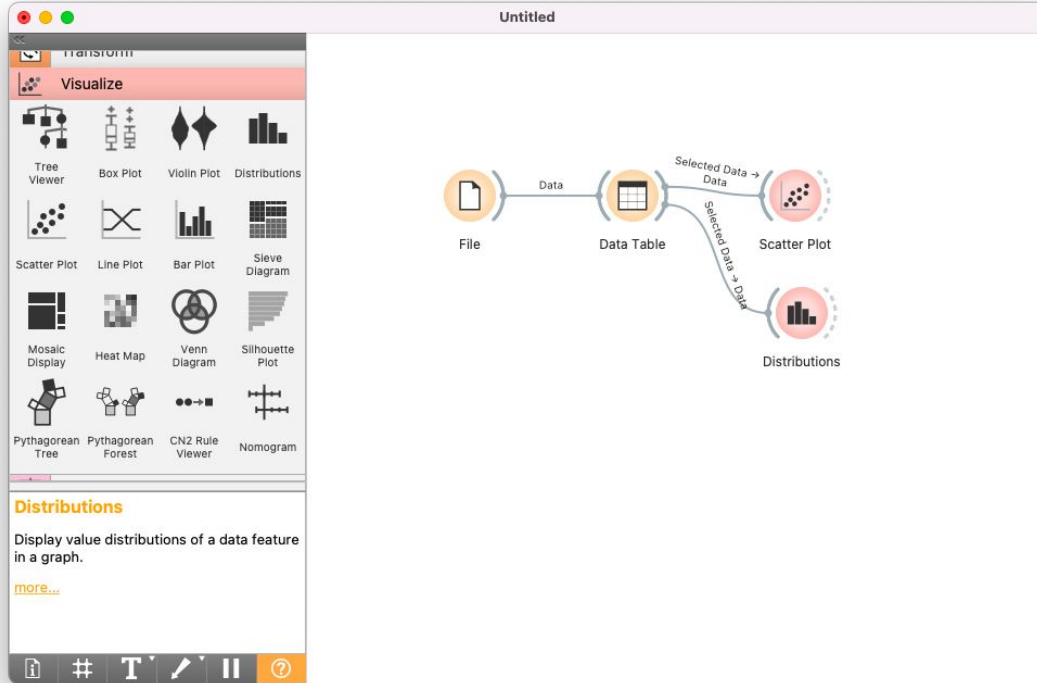
The screenshot shows the 'Data Table' widget displaying the Iris dataset. The left sidebar contains information about the data: 150 instances (no missing data), 4 features, Target with 3 values, and No meta attributes. It also shows options for variables (Show variable labels, Visualize numeric values, Color by instance classes) and selection (Select full rows). The main area displays a table with 20 rows of data.

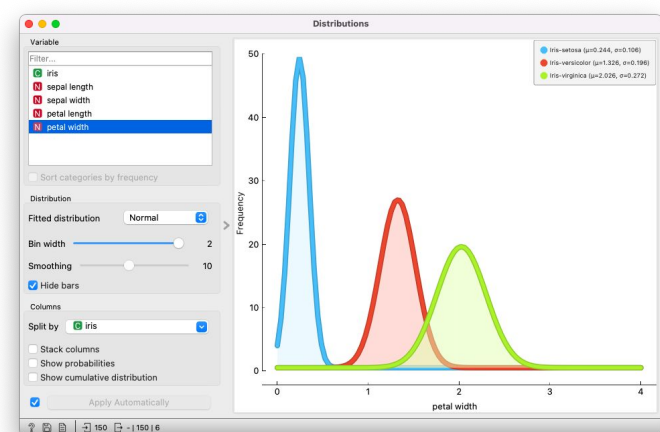
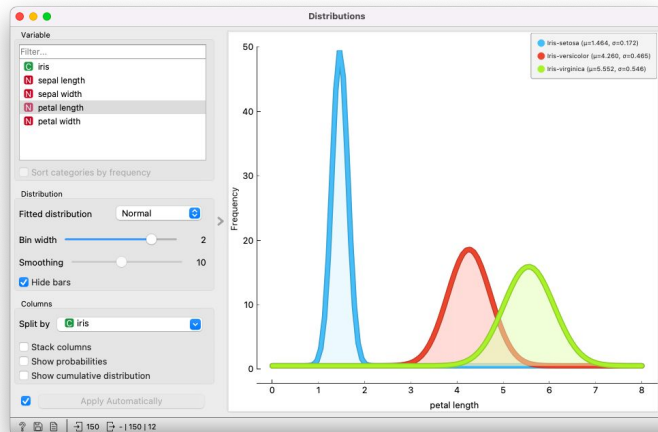
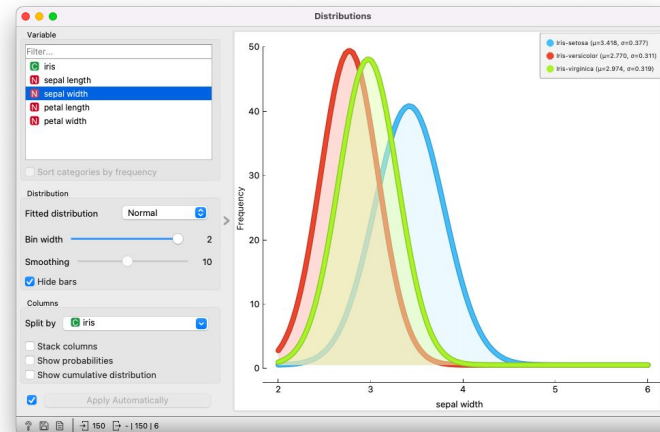
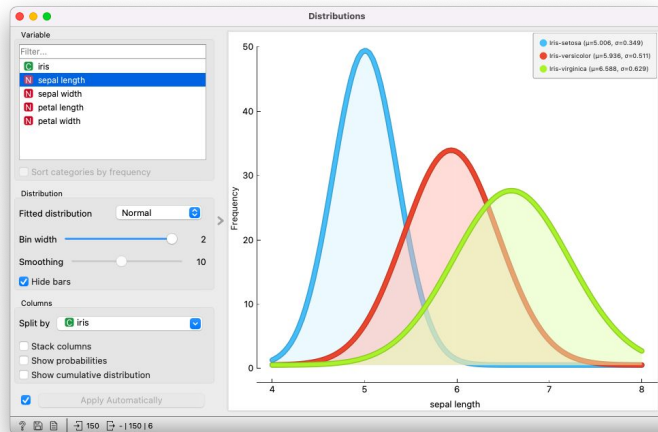
	Iris	sepal length	sepal width	petal length	petal width
1	Iris-setosa	5.1	3.5	1.4	0.2
2	Iris-setosa	4.9	3.0	1.4	0.2
3	Iris-setosa	4.7	3.2	1.3	0.2
4	Iris-setosa	4.6	3.1	1.5	0.2
5	Iris-setosa	5.0	3.6	1.4	0.2
6	Iris-setosa	5.4	3.9	1.7	0.4
7	Iris-setosa	4.6	3.4	1.4	0.3
8	Iris-setosa	5.0	3.4	1.5	0.2
9	Iris-setosa	4.4	2.9	1.4	0.2
10	Iris-setosa	4.9	3.1	1.5	0.1
11	Iris-setosa	5.4	3.7	1.5	0.2
12	Iris-setosa	4.8	3.4	1.6	0.2
13	Iris-setosa	4.8	3.0	1.4	0.1
14	Iris-setosa	4.3	3.0	1.1	0.1
15	Iris-setosa	5.8	4.0	1.2	0.2
16	Iris-setosa	5.7	4.4	1.5	0.4
17	Iris-setosa	5.4	3.9	1.3	0.4
18	Iris-setosa	5.1	3.5	1.4	0.3
19	Iris-setosa	5.7	3.8	1.7	0.3
20	Iris-setosa	5.1	3.8	1.5	0.3

Orange Data Mining - Widget “Scatter Plot”

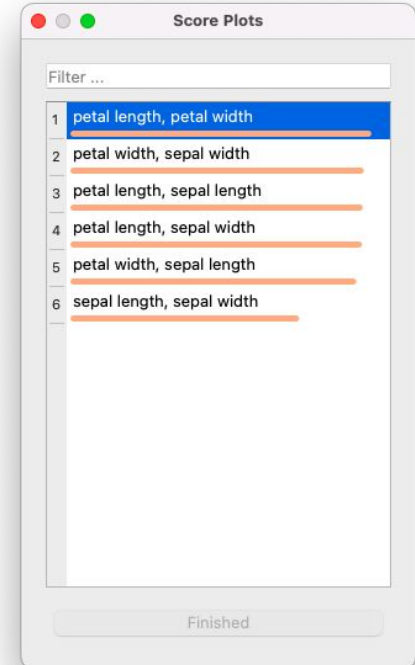
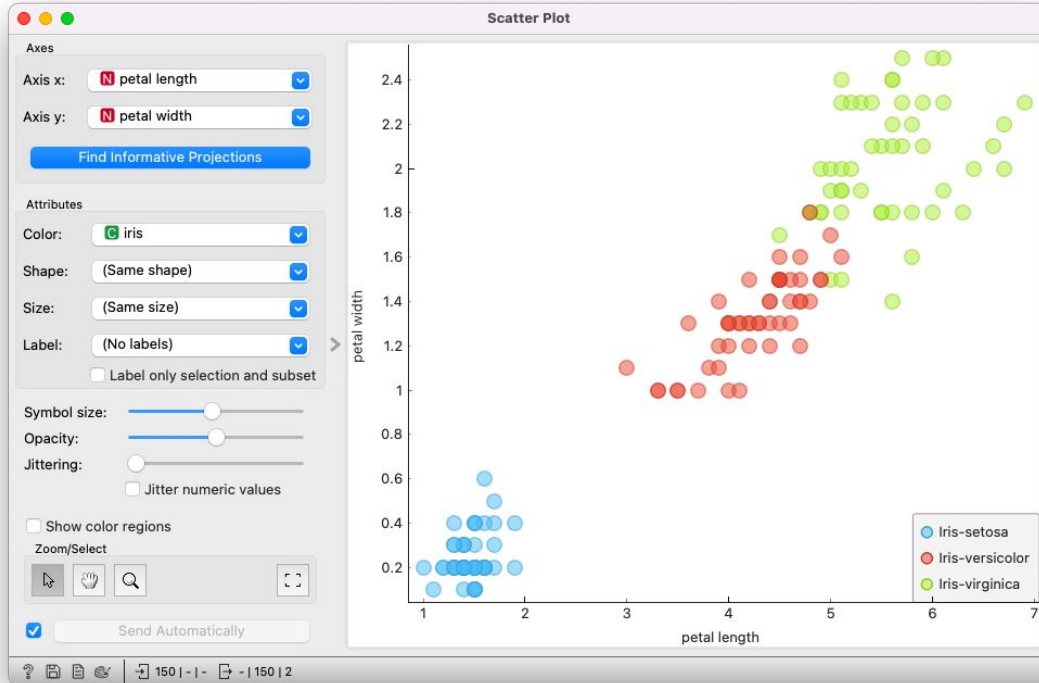


Orange Data Mining - Widget “Distribution”



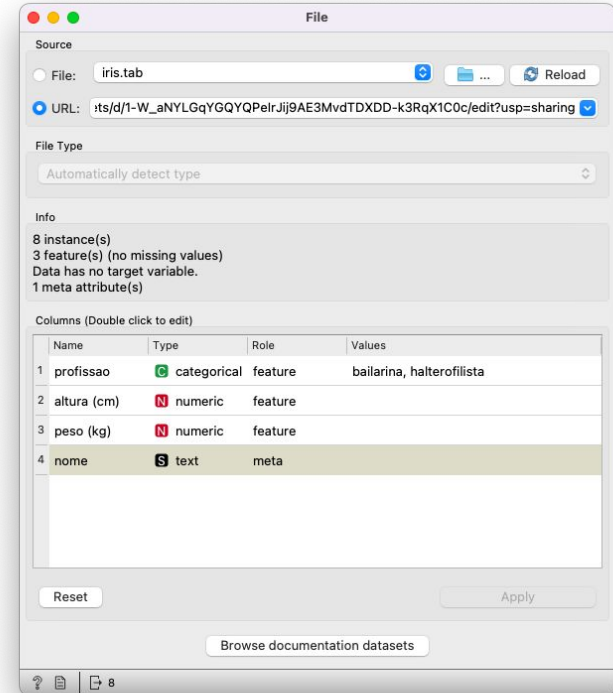


Orange Data Mining - Widget “Scatter Plot”



Orange Data Mining - Carregando seus Dados

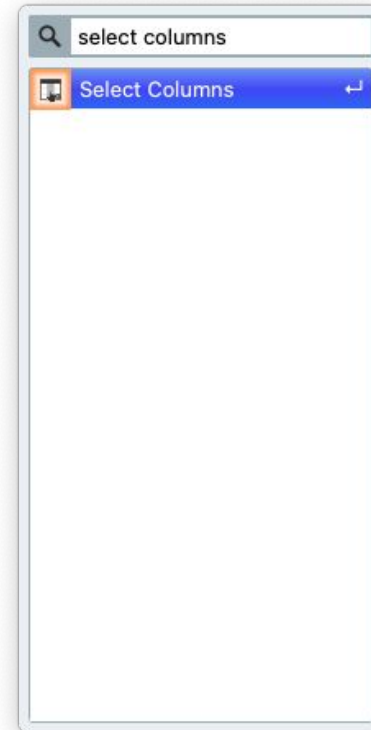
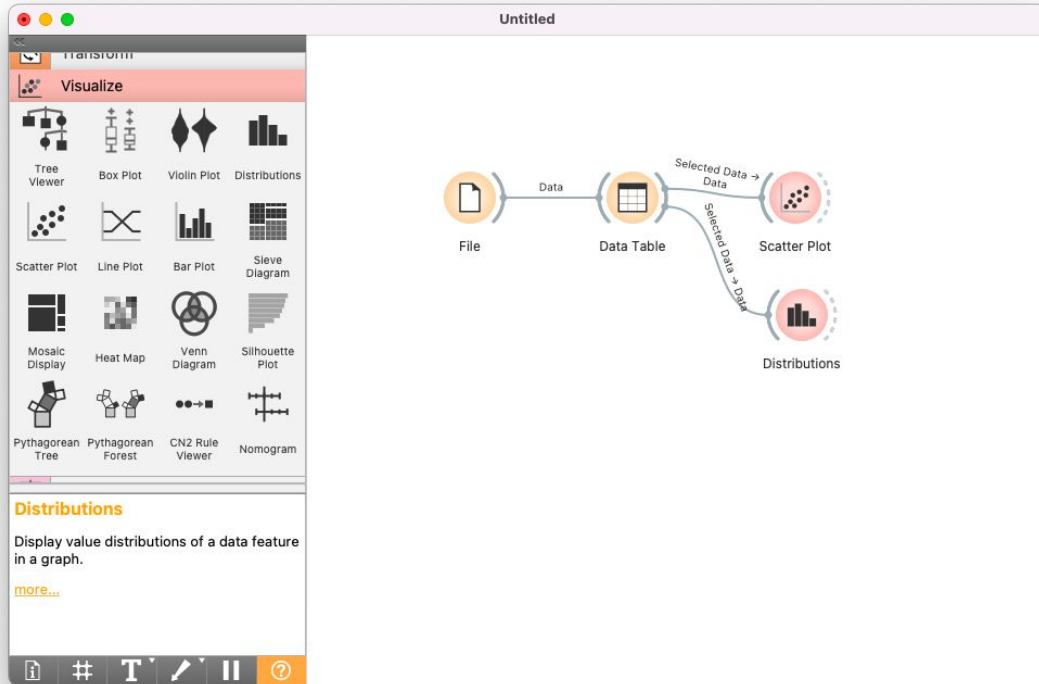
- Orange pode ler vários formatos, como xlsx, tab, csv, etc.
- Os dados são dispostos em uma tabela na qual os registros (ou objetos) estão em linhas e os atributos (ou características) estão em colunas.



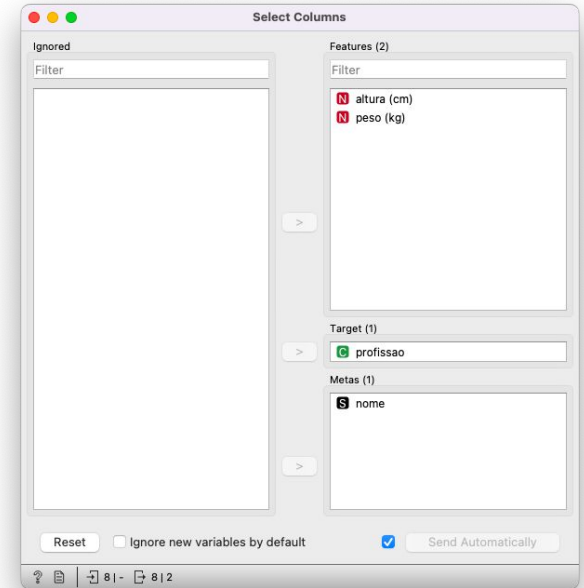
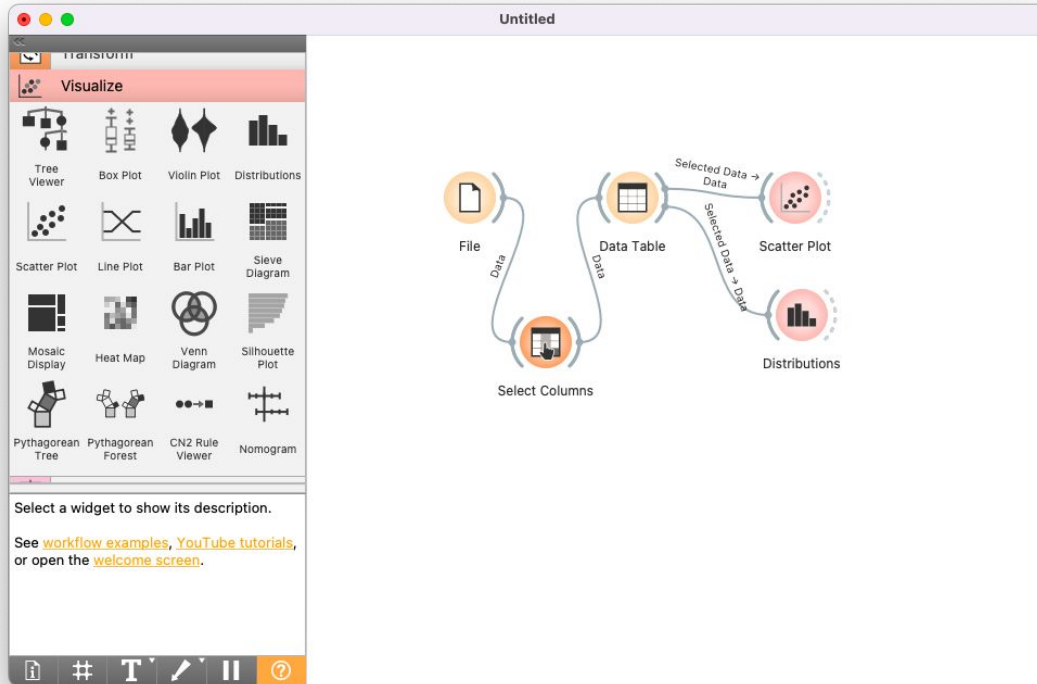
Exercício

1. Crie uma tabela com as colunas: nome, peso, altura, profissão.
2. Preencha a tabela com informações de bailarinas e halterofilistas.
3. Gerar gráficos de dispersão e distribuição da sua tabela.

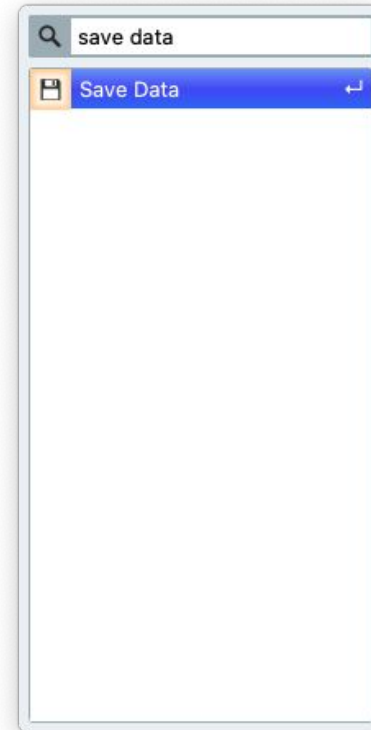
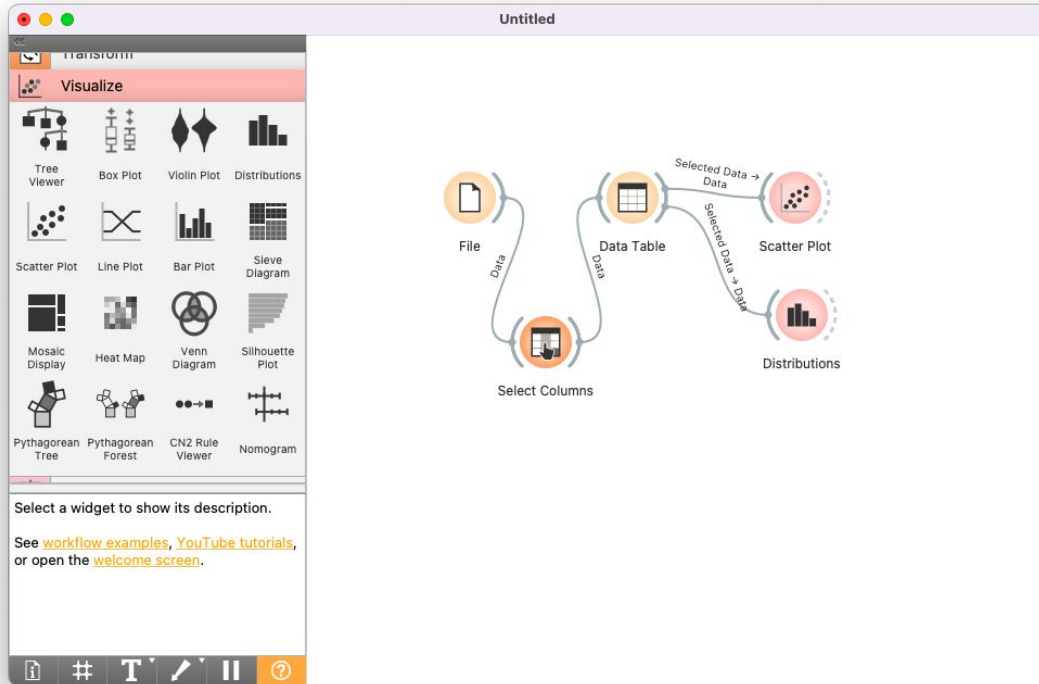
Orange Data Mining - Widget “Select Columns”



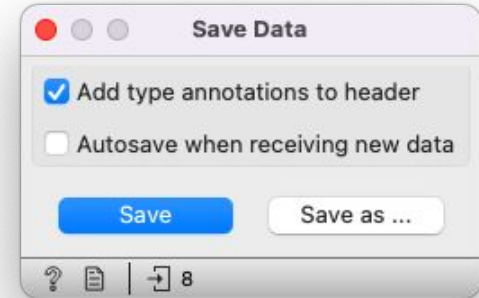
Orange Data Mining - Widget “Select Columns”



Orange Data Mining - Widget “Save Data”



Orange Data Mining - Widget “Save Data”

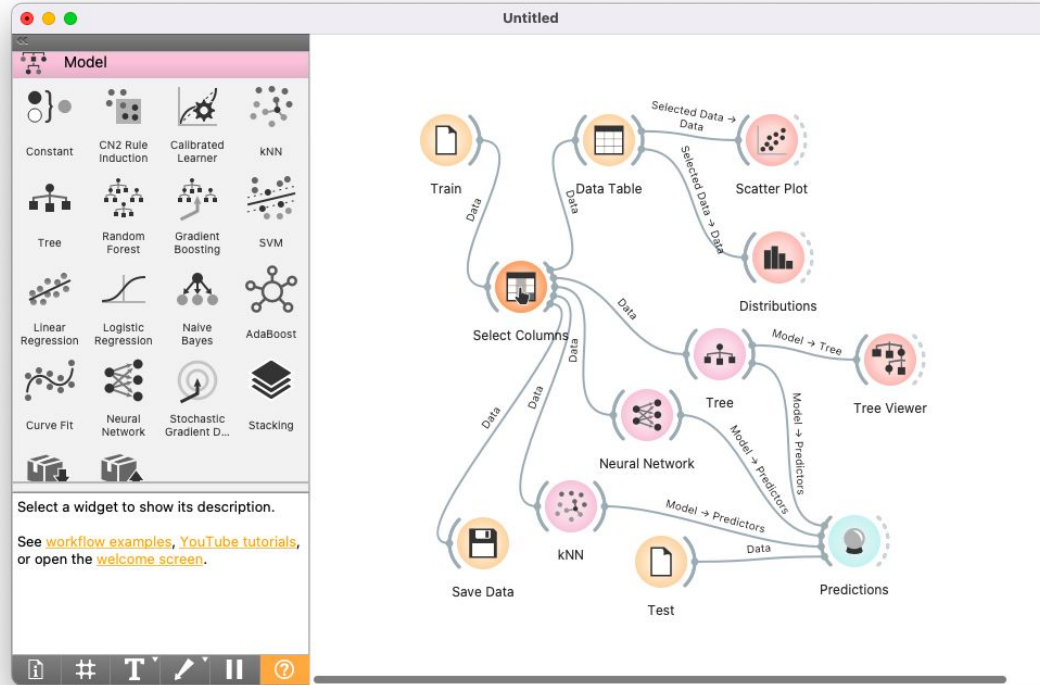


Orange Data Mining - Fazendo Previsões

Uma das principais características de um cientista de dados é prever o futuro.



Orange Data Mining - Widget “Predictions”



Orange Data Mining - Widget “Predictions”

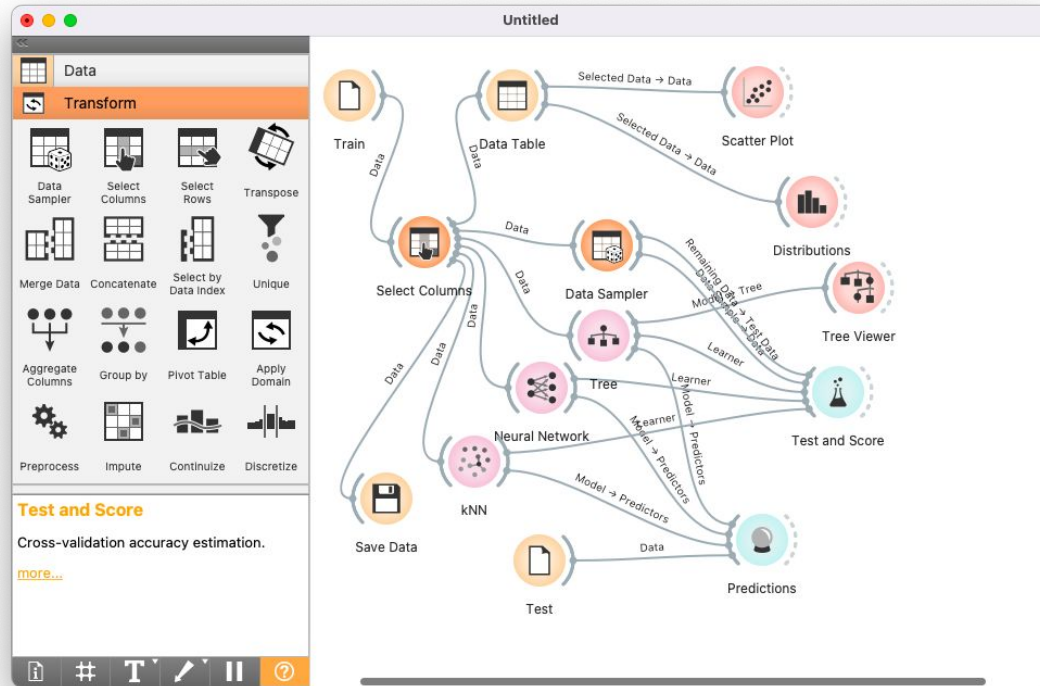
Predictions

Show probabilities for Classes known to the model Restore Original Order

	Tree	Neural Network	kNN	name	vitamin A %	vitamin C %	calcium %	
1	1.00 : 0.00 → fruit	0.88 : 0.12 → fruit	0.80 : 0.20 → fruit	?	1.0	154.0	3.0	1
2	1.00 : 0.00 → fruit	0.00 : 1.00 → vegetable	0.60 : 0.40 → fruit	?	15.0	300.0	2.0	1
3	1.00 : 0.00 → fruit	0.99 : 0.01 → fruit	1.00 : 0.00 → fruit	?	0.0	43.0	2.0	3

3 | 3 | -

Orange Data Mining - Widget “Test and Score”



Orange Data Mining - Widget “Test and Score”

The screenshot shows the 'Test and Score' widget in the Orange Data Mining software. The interface is divided into several sections:

- Left Panel (Settings):**
 - Cross validation:** Selected. Number of folds: 10. ☒ Stratified. Cross validation by feature: (empty dropdown).
 - Random sampling:** Not selected. Repeat train/test: 10. Training set size: 66 %.
 - Test on:** ☒ Stratified. ☐ Leave one out. ☐ Test on train data. ☐ Test on test data.
- Top Right (Evaluation results):**

Evaluation results for target: (None, show average over classes)

Model	AUC	CA	F1	Precision	Recall
kNN	0.520	0.680	0.607	0.791	0.680
Tree	0.550	0.600	0.574	0.579	0.600
Neural Network	0.793	0.760	0.744	0.775	0.760
- Bottom Right (Compare models):**

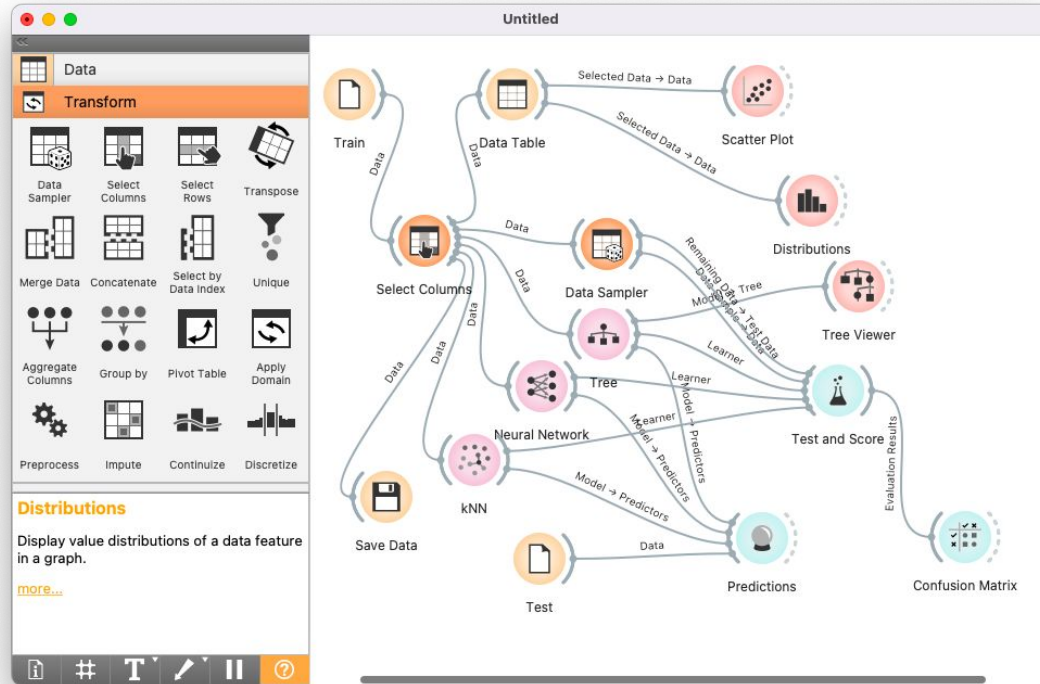
Compare models by: Area under ROC curve. Negligible diff.: 0.1

	kNN	Tree	Neural Network
kNN		0.379	0.268
Tree	0.621		0.462
Neural Network	0.732	0.538	

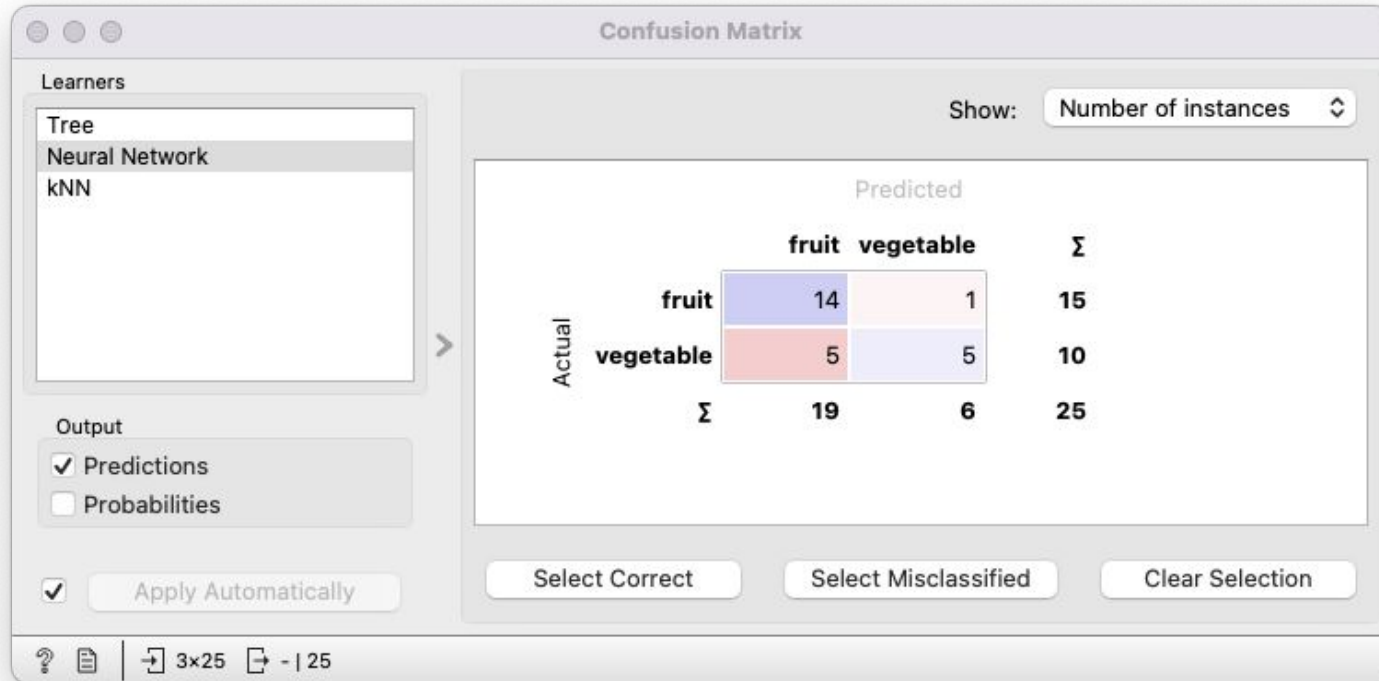
Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

The bottom status bar shows: ? | 25 | 10 | 10 | 10 | 10 | 25 | 3x25

Orange Data Mining - Widget “Confusion Matrix”



Orange Data Mining - Widget “Confusion Matrix”



Considerações Finais

- scikit-learn: <https://scikit-learn.org/stable/>