

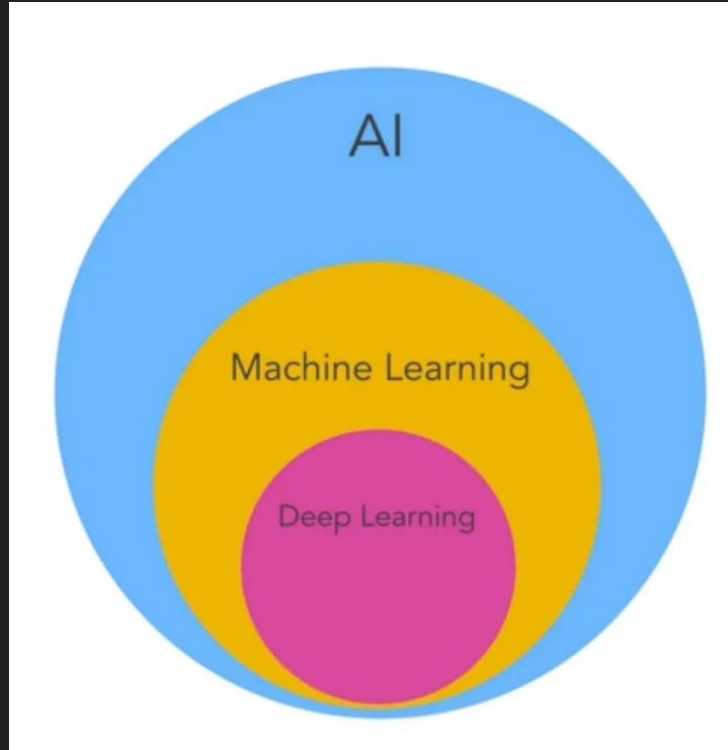
Aprendizado de Máquina

Aula 1

Definição de IA

Chat-GPT: A Inteligência Artificial (IA) é um campo da ciência da computação que se concentra no desenvolvimento de sistemas e máquinas capazes de realizar tarefas que, quando executadas por seres humanos, requerem inteligência. A IA visa criar programas e sistemas que possam aprender, raciocinar, resolver problemas, compreender linguagem natural, reconhecer padrões e tomar decisões de forma autônoma.

IA, Machine Learning e Deep Learning



Introdução à Classificação

Aplicações

- Fraude em compras com cartão de crédito
- Análise de nível de risco de seguro de automóvel
- Prever a evasão de um estudante de um curso
- Identificar câncer de pele em uma imagem
- Reconhecer escrita manual
- Detectar quedas de idosos

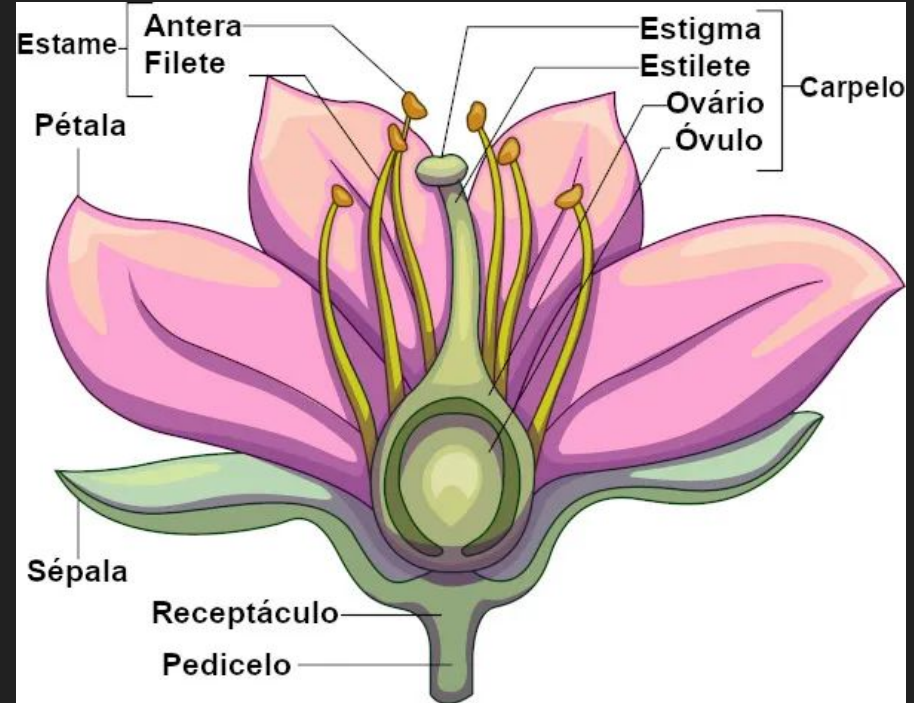
Dataset Iris

- Iris é uma flor
- Pode ser classificada como (espécie):
 - setosa
 - versicolor
 - virginica
- A espécie será a classe
- O dataset possui 150 instâncias, sendo 50 de cada classe

	A	B	C	D	E
1	sepallength	sepalwidth	petallength	petalwidth	class
2	5.1	3.5	1.4	0.2	Iris-setosa
3	4.9	3	1.4	0.2	Iris-setosa
4	4.7	3.2	1.3	0.2	Iris-setosa
5	4.6	3.1	1.5	0.2	Iris-setosa
6	5	3.6	1.4	0.2	Iris-setosa
7	5.4	3.9	1.7	0.4	Iris-setosa
8	4.6	3.4	1.4	0.3	Iris-setosa
9	5	3.4	1.5	0.2	Iris-setosa
10	4.4	2.9	1.4	0.2	Iris-setosa
11	4.9	3.1	1.5	0.1	Iris-setosa
12	5.4	3.7	1.5	0.2	Iris-setosa
13	4.8	3.4	1.6	0.2	Iris-setosa
14	4.8	3	1.4	0.1	Iris-setosa
15	4.3	3	1.1	0.1	Iris-setosa
16	5.8	4	1.2	0.2	Iris-setosa
17	5.7	4.4	1.5	0.4	Iris-setosa
18	5.4	3.9	1.3	0.4	Iris-setosa
19	5.1	3.5	1.4	0.3	Iris-setosa
20	5.7	3.8	1.7	0.3	Iris-setosa
21	5.1	3.8	1.5	0.3	Iris-setosa
22	5.4	3.4	1.7	0.2	Iris-setosa
23	5.1	3.7	1.5	0.4	Iris-setosa

Iris Dataset - atributos

- Sépala
 - comprimento (length)
 - largura (width)
- Pétala
 - comprimento (length)
 - largura (width)



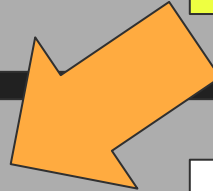
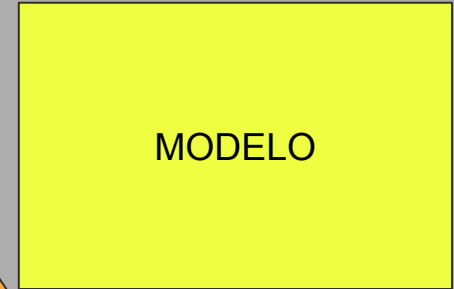
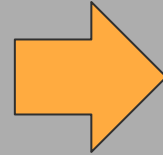
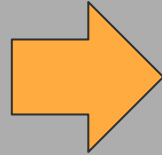
Iris - Modelo

- Baseado nas 150 instâncias (medidas de sépala e pétala) de cada instância de cada espécie
 - → Posso classificar nova instâncias (flores novas) usando o modelo e as medidas da nova flor

Processo de Classificação

	A	B	C	D	E
1	sepalwidth	sepalwidth	petalwidth	petalwidth	class
2	5.1	3.5	1.4	0.2	Iris-setosa
3	4.9	3	1.4	0.2	Iris-setosa
4	4.7	3.2	1.3	0.2	Iris-setosa
5	4.6	3.1	1.5	0.2	Iris-setosa
6	5	3.6	1.4	0.2	Iris-setosa
7	5.4	3.9	1.7	0.4	Iris-setosa
8	4.6	3.4	1.4	0.3	Iris-setosa
9	5	3.4	1.5	0.2	Iris-setosa
10	4.4	2.9	1.4	0.2	Iris-setosa
11	4.9	3.1	1.5	0.1	Iris-setosa
12	5.4	3.7	1.5	0.2	Iris-setosa
13	4.8	3.4	1.6	0.2	Iris-setosa
14	4.8	3	1.4	0.1	Iris-setosa

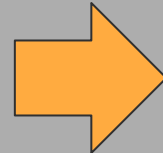
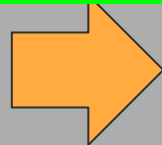
TREINAMENTO



CLASSIFICAÇÃO



medidas da
sépala e da
pétala



O que é o modelo?

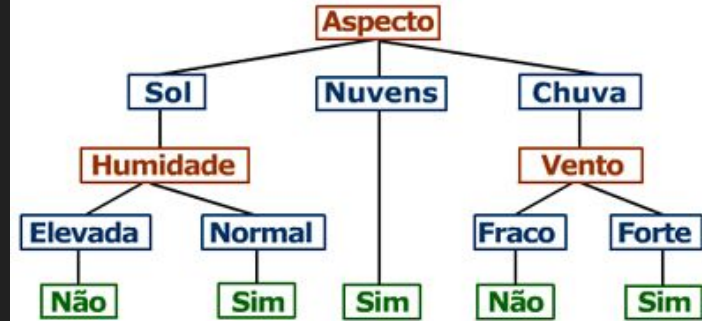
Pode ser:

- Árvore de decisão
- Tabela de probabilidades
- Rede neural
- Entre outros

Exemplos de Treino

Dia	Aspecto	Temp.	Humidade	Vento	Jogar Ténis
D1	Sol	Quente	Elevada	Fraco	Não
D2	Sol	Quente	Elevada	Forte	Não
D3	Nuvens	Quente	Elevada	Fraco	Sim
D4	Chuva	Ameno	Elevada	Fraco	Sim
D5	Chuva	Fresco	Normal	Fraco	Sim
D6	Chuva	Fresco	Normal	Forte	Não
D7	Nuvens	Fresco	Normal	Fraco	Sim
D8	Sol	Ameno	Elevada	Fraco	Não
D9	Sol	Fresco	Normal	Fraco	Sim
D10	Chuva	Ameno	Normal	Forte	Sim
D11	Sol	Ameno	Normal	Forte	Sim
D12	Nuvens	Ameno	Elevada	Forte	Sim
D13	Nuvens	Quente	Normal	Fraco	Sim
D14	Chuva	Ameno	Elevada	Forte	Não

Árvore de Decisão para Jogar Ténis



Mas, como sei que o modelo funciona?

- O modelo pode errar
- É preciso medir o desempenho do modelo antes de usá-lo
- Um modelo com baixo desempenho não é útil

Treinamento e Avaliação de Desempenho

	A	B	C	D	E
1	sepalwidth	sepalwidth	petalwidth	petalwidth	class
2	5.1	3.5	1.4	0.2	Iris-setosa
3	4.9	3	1.4	0.2	Iris-setosa
4	4.7	3.2	1.3	0.2	Iris-setosa
5	4.6	3.1	1.5	0.2	Iris-setosa
6	5	3.6	1.4	0.2	Iris-setosa
7	5.4	3.9	1.7	0.4	Iris-setosa
8	4.6	3.4	1.4	0.3	Iris-setosa
9	5	3.4	1.5	0.2	Iris-setosa
10	4.4	2.9	1.4	0.2	Iris-setosa
11	4.9	3.1	1.5	0.1	Iris-setosa
12	5.4	3.7	1.5	0.2	Iris-setosa
13	4.8	3.4	1.6	0.2	Iris-setosa
14	4.8	3	1.4	0.1	Iris-setosa

70%

TREINAMENTO

ALGORITMO

MODELO

ACURÁCIA

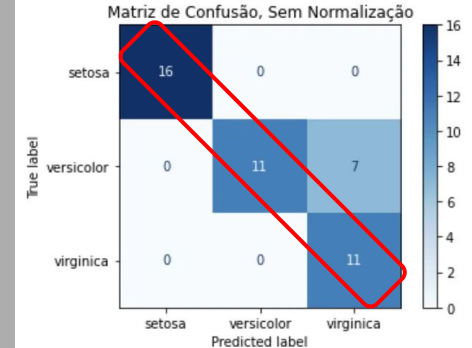
38 corretas/45 = 84,44%

	A	B	C	D	E
1	sepalwidth	sepalwidth	petalwidth	petalwidth	class
2	5.1	3.5	1.4	0.2	Iris-setosa
3	4.9	3	1.4	0.2	Iris-setosa
4	4.7	3.2	1.3	0.2	Iris-setosa
5	4.6	3.1	1.5	0.2	Iris-setosa
6	5	3.6	1.4	0.2	Iris-setosa
7	5.4	3.9	1.7	0.4	Iris-setosa
8	4.6	3.4	1.4	0.3	Iris-setosa
9	5	3.4	1.5	0.2	Iris-setosa
10	4.4	2.9	1.4	0.2	Iris-setosa
11	4.9	3.1	1.5	0.1	Iris-setosa
12	5.4	3.7	1.5	0.2	Iris-setosa
13	4.8	3.4	1.6	0.2	Iris-setosa
14	4.8	3	1.4	0.1	Iris-setosa

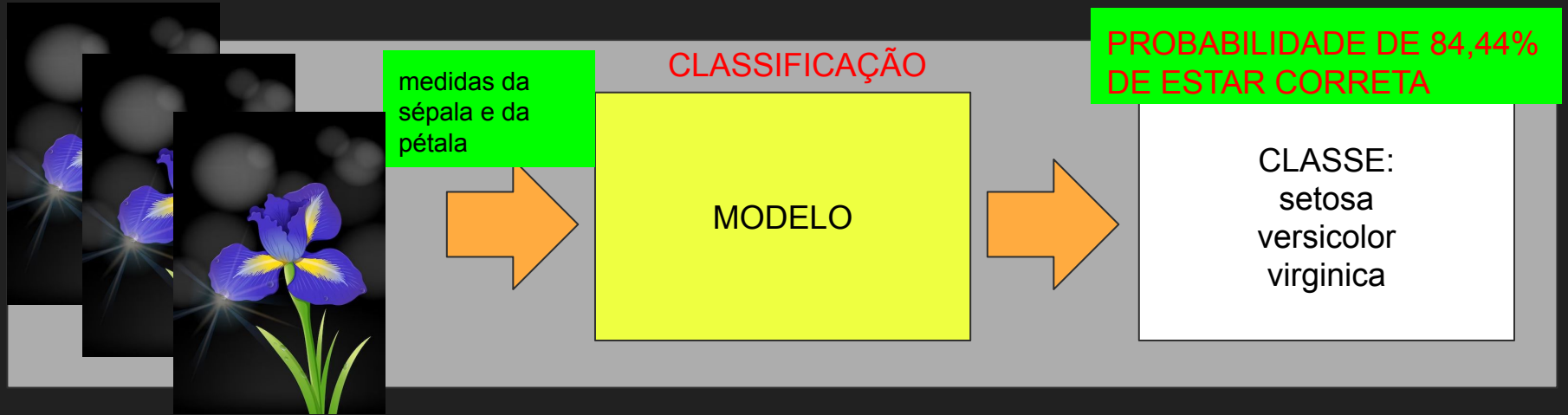
30%

AVALIAÇÃO

MODELO



E agora? Vamos usar o modelo com novas flores



Elementos

- **Classe:** o que se deseja prever
 - espécie da flor, nível de risco de segurado, possibilidade de câncer de pele de uma mancha,...
- **Dataset ou conjunto de dados:** conjunto de ocorrências/instâncias
- **Instância:** é uma ocorrência
 - uma flor, um segurado, uma mancha na pele,...
- **Atributos:** características da instância
 - tamanho da pétala, idade do segurado, tonalidade da mancha,...

	A	B	C	D	E
1	sepal.length	sepal.width	petal.length	petal.width	class
2	5.1	3.5	1.4	0.2	Iris-setosa
3	4.9	3	1.4	0.2	Iris-setosa
4	4.7	3.2	1.3	0.2	Iris-setosa
5	4.6	3.1	1.5	0.2	Iris-setosa
6	5	3.6	1.4	0.2	Iris-setosa
7	5.4	3.9	1.7	0.4	Iris-setosa
8	4.6	3.4	1.4	0.3	Iris-setosa
9	5	3.4	1.5	0.2	Iris-setosa
10	4.4	2.9	1.4	0.2	Iris-setosa
11	4.9	3.1	1.5	0.1	Iris-setosa
12	5.4	3.7	1.5	0.2	Iris-setosa
13	4.8	3.4	1.6	0.2	Iris-setosa
14	4.8	3	1.4	0.1	Iris-setosa
15	4.3	3	1.1	0.1	Iris-setosa
16	5.8	4	1.2	0.2	Iris-setosa
17	5.7	4.4	1.5	0.4	Iris-setosa
18	5.4	3.9	1.3	0.4	Iris-setosa
19	5.1	3.5	1.4	0.3	Iris-setosa
20	5.7	3.8	1.7	0.3	Iris-setosa
21	5.1	3.8	1.5	0.3	Iris-setosa
22	5.4	3.4	1.7	0.2	Iris-setosa
23	5.1	3.7	1.5	0.4	Iris-setosa

Tipos de Algoritmos

Algoritmos

- Vamos entender como alguns algoritmos funcionam
- Não precisamos saber como implementar esses algoritmos
- Mas é importante conhecer o que eles fazem para chegar em uma classificação

Árvores de Decisão



Naive Bayes

- Analisa os atributos individualmente, calculando a probabilidade para cada classe a partir do valor de cada atributo
- No final calcula a probabilidade total

Play-tennis example: estimating $P(x_i|C)$

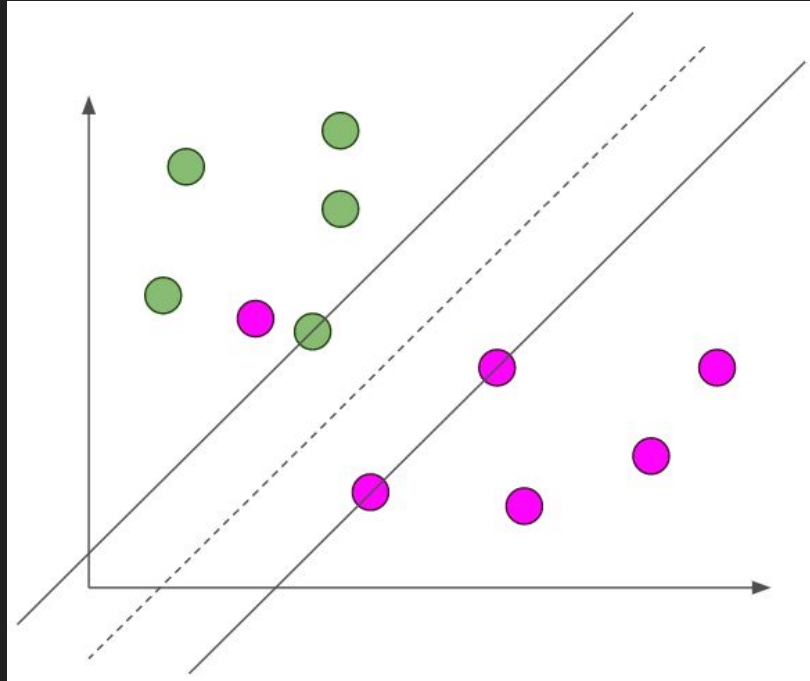
Outlook	Temperature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N

$$P(p) = 9/14$$

$$P(n) = 5/14$$

outlook	
$P(\text{sunny} p) = 2/9$	$P(\text{sunny} n) = 3/5$
$P(\text{overcast} p) = 4/9$	$P(\text{overcast} n) = 0$
$P(\text{rain} p) = 3/9$	$P(\text{rain} n) = 2/5$
temperature	
$P(\text{hot} p) = 2/9$	$P(\text{hot} n) = 2/5$
$P(\text{mild} p) = 4/9$	$P(\text{mild} n) = 2/5$
$P(\text{cool} p) = 3/9$	$P(\text{cool} n) = 1/5$
humidity	
$P(\text{high} p) = 3/9$	$P(\text{high} n) = 4/5$
$P(\text{normal} p) = 6/9$	$P(\text{normal} n) = 2/5$
windy	
$P(\text{true} p) = 3/9$	$P(\text{true} n) = 3/5$
$P(\text{false} p) = 6/9$	$P(\text{false} n) = 2/5$

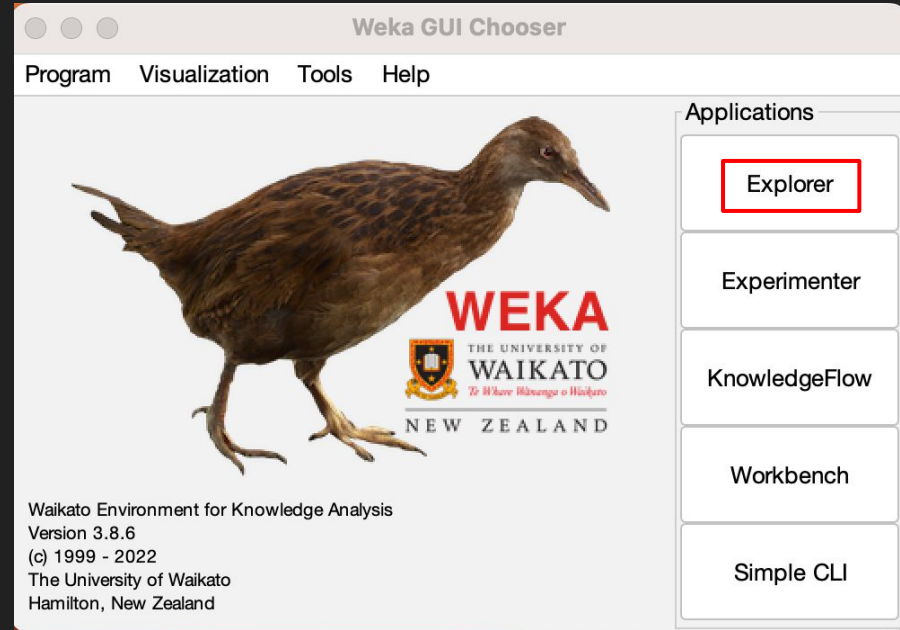
Máquinas de Vetor de Suporte (SVM)



Weka

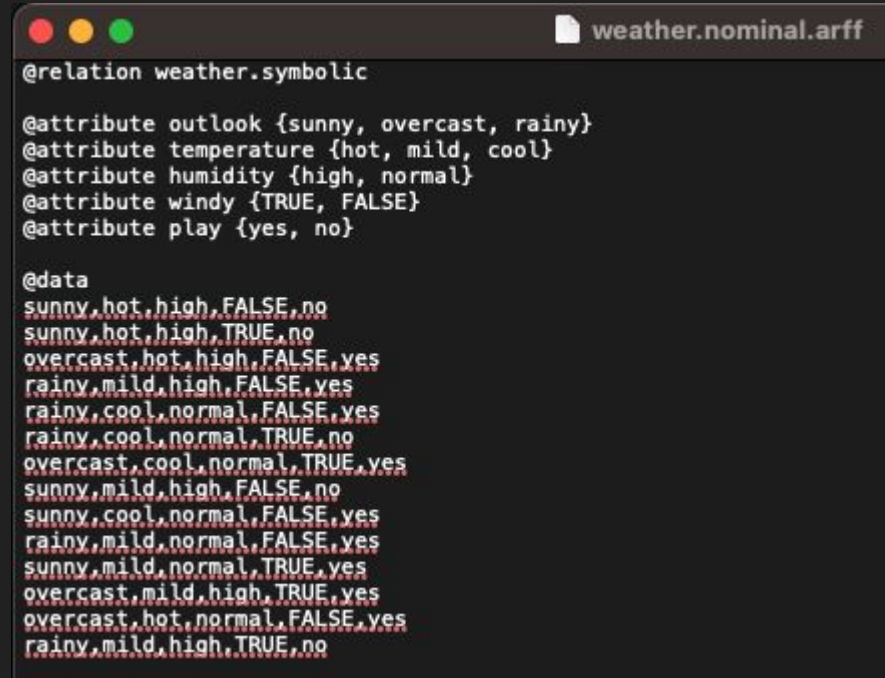
Weka

- Open source
- Fácil de instalar e usar (Interface Gráfica)
- Disponibiliza biblioteca Java
- Requer a JVM (Java)
- Desenvolvido pela Universidade de Waikato (Nova Zelândia)
- Disponível para Windows, Linux e Mac



Arquivos exemplo

- São encontradas na pasta “data” dentro da pasta do weka que fica dentro de Arquivos de Programas no Windows
- Tem e extensão arff (formato próprio do weka)
- Lista de atributos e valores possíveis para cada atributo
- Instâncias (ocorrências) abaixo de @data
- É possível usar csv



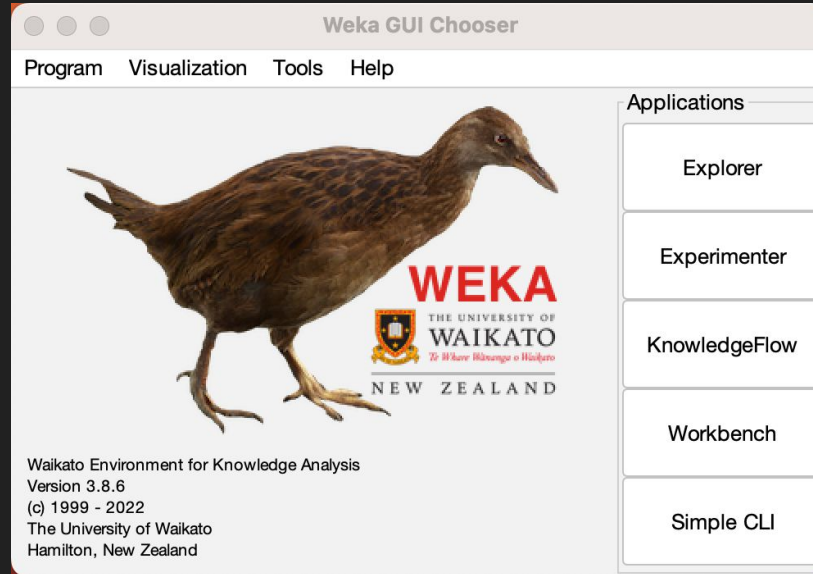
```
@relation weather.symbolic

@attribute outlook {sunny, overcast, rainy}
@attribute temperature {hot, mild, cool}
@attribute humidity {high, normal}
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

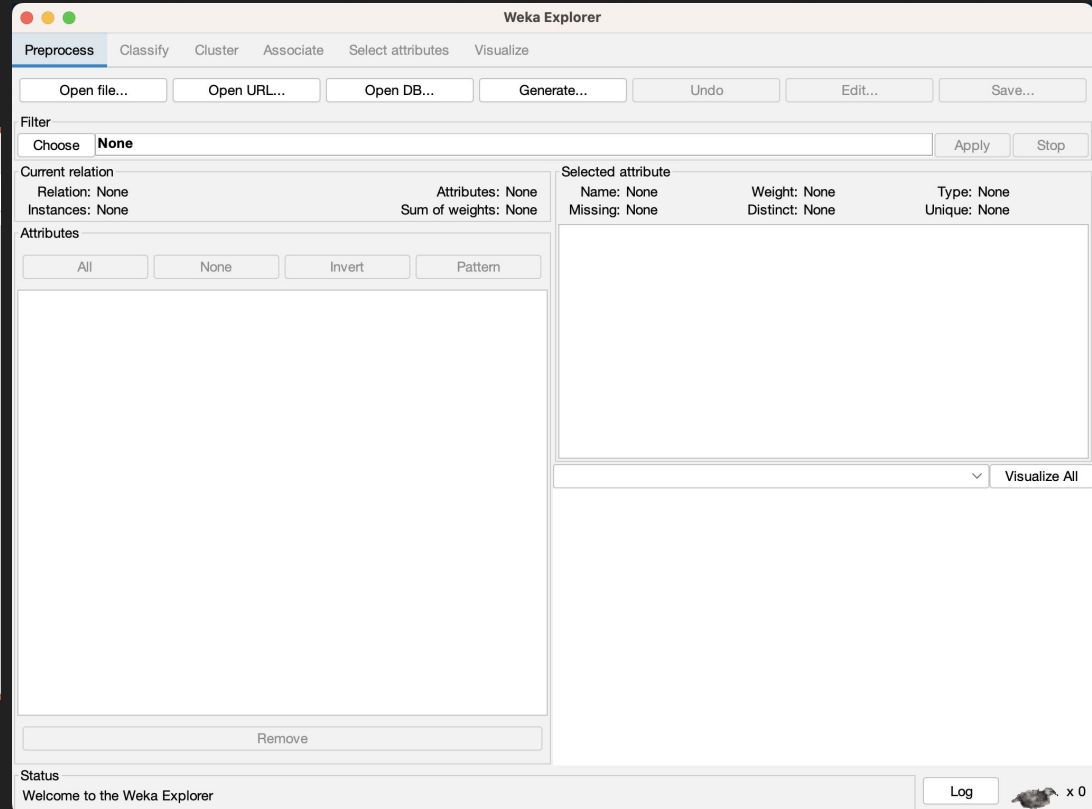
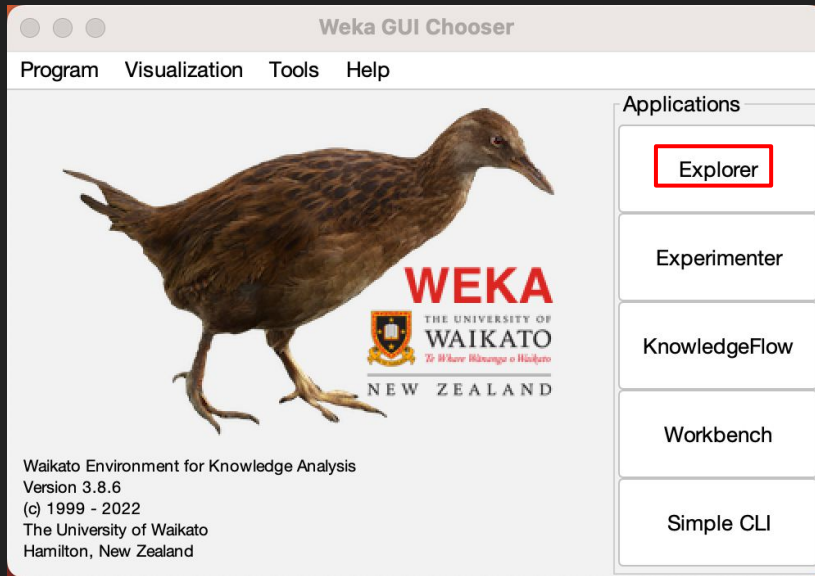
@data
sunny.hot.high.FALSE.no
sunny.hot.high.TRUE.no
overcast.hot.high.FALSE.yes
rainy.mild.high.FALSE.yes
rainy.cool.normal.FALSE.yes
rainy.cool.normal.TRUE.no
overcast.cool.normal.TRUE.yes
sunny.mild.high.FALSE.no
sunny.cool.normal.FALSE.yes
rainy.mild.normal.FALSE.yes
sunny.mild.normal.TRUE.yes
overcast.mild.high.TRUE.yes
overcast.hot.normal.FALSE.yes
rainy.mild.high.TRUE.no
```

Classificação

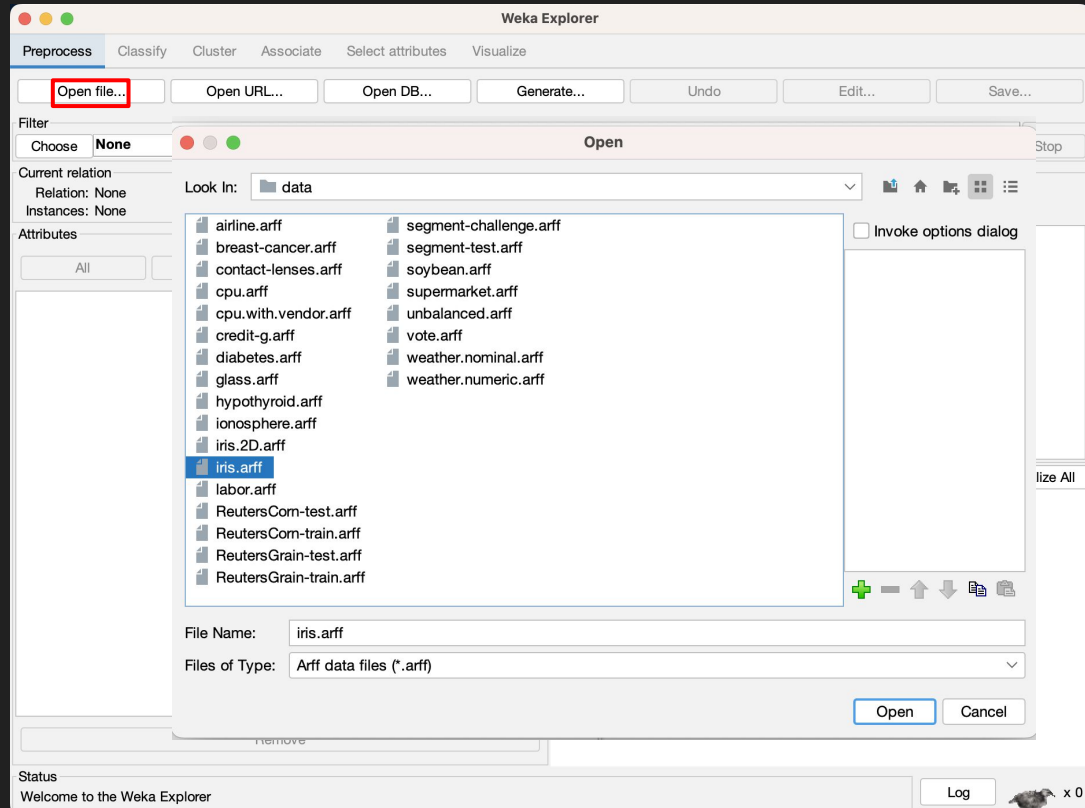
Abrir o Weka



Abrir o Weka Explorer

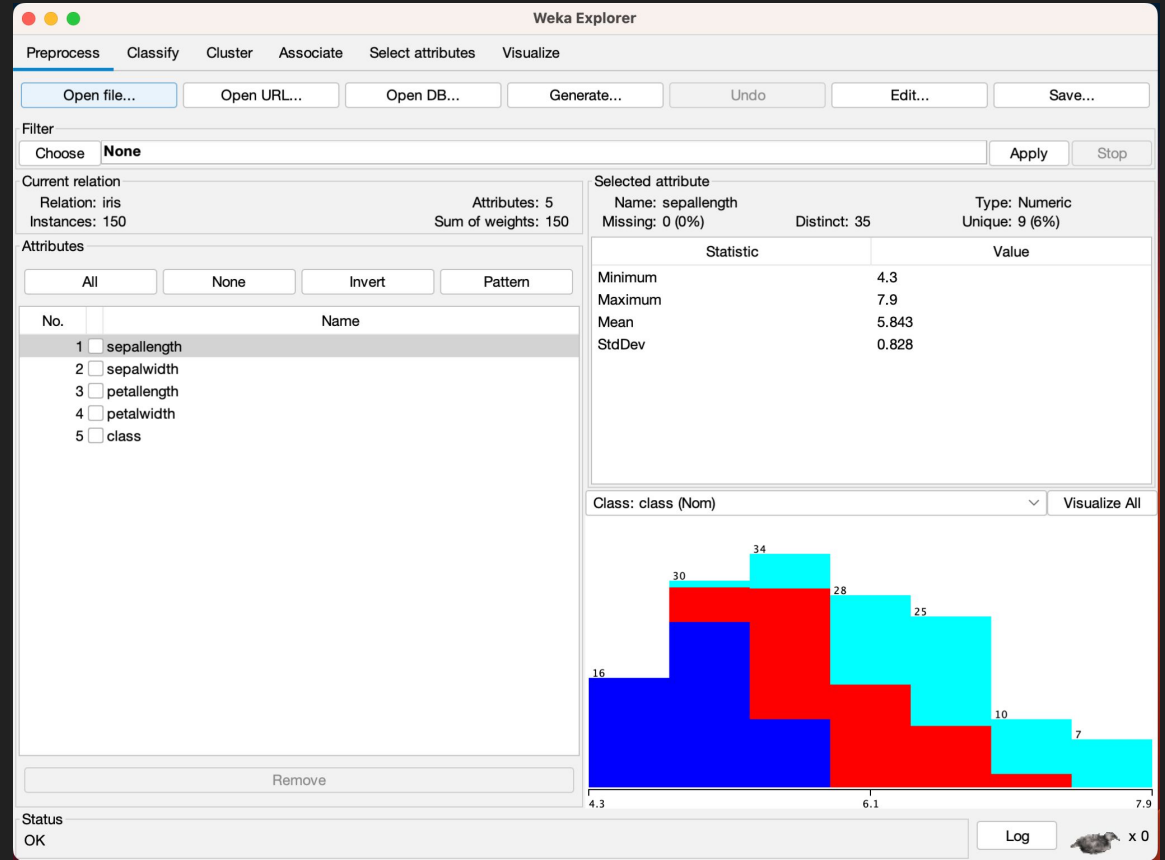


Abrir o dataset - Open file...



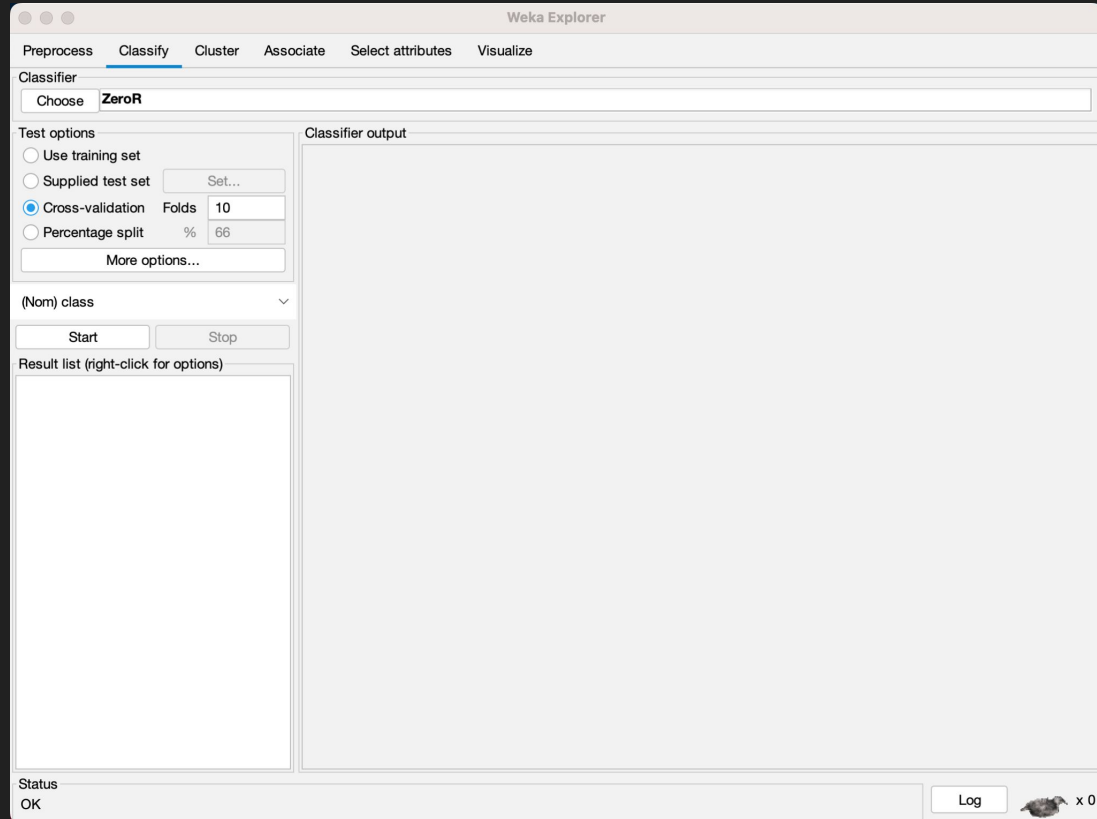
Dataset aberto no Weka em Preprocess

- Dataset: Iris
- Instâncias: 150
- Atributos: 5 (contando com a classe)
- O Weka considera o último atributo como a classe



Aba Classificar - Classify

- Escolher o algoritmo
(Choose): Naive Bayes
- Tipos de teste: Percentage Split
- Clicar em Start



Resultado da Classificação

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier
Choose **NaiveBayes**

Test options
☐ Use training set
☐ Supplied test set Set...
☐ Cross-validation Folds 10
☒ Percentage split % 66
More options...

(Nom) class
Start Stop

Result list (right-click for options)
13:34:39 - bayes.NaiveBayes

Classifier output

mean	3.4815	2.7687	2.9629
std. dev.	0.3925	0.3038	0.3088
weight sum	50	50	50
precision	0.1091	0.1091	0.1091
petallength			
mean	1.4694	4.2452	5.5516
std. dev.	0.1782	0.4712	0.5529
weight sum	50	50	50
precision	0.1405	0.1405	0.1405
petalwidth			
mean	0.2743	1.3097	2.0343
std. dev.	0.1096	0.1915	0.2646
weight sum	50	50	50
precision	0.1143	0.1143	0.1143

Time taken to build model: 0 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances	48	94.1176 %
Incorrectly Classified Instances	3	5.8824 %
Kappa statistic	0.9115	
Mean absolute error	0.0447	
Root mean squared error	0.1722	
Relative absolute error	10.0365 %	
Root relative squared error	36.4196 %	
Total Number of Instances	51	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Iris-setosa
	0.947	0.063	0.900	0.947	0.923	0.876	0.988	0.980	Iris-versicolor
	0.882	0.029	0.938	0.882	0.909	0.867	0.988	0.980	Iris-virginica
Weighted Avg.	0.941	0.033	0.942	0.941	0.941	0.909	0.992	0.986	

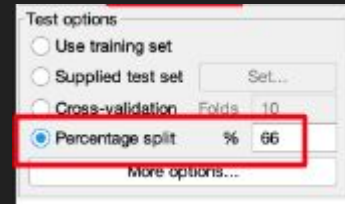
=== Confusion Matrix ===

a	b	c	<-- classified as
15	0	0	a = Iris-setosa
0	18	1	b = Iris-versicolor
0	2	15	c = Iris-virginica

Status

Test Options - Percentage Split

- Separa 66% do dataset para treino e 34% para teste (avaliação de desempenho)
- As instâncias são selecionadas aleatoriamente por padrão



Resultados

- Acurácia de 94,11%
- Incorretos 5,89%

- Acurácia = Previsões corretas /
TOTAL

$$(15 + 18 + 15) / ((15 + 18 + 15) + 2 + 1)$$

$$48 / 51 = 94,11\%$$

```
=== Summary ===
```

Correctly Classified Instances	48	94.1176 %
Incorrectly Classified Instances	3	5.8824 %
Kappa statistic	0.9113	
Mean absolute error	0.0447	
Root mean squared error	0.1722	
Relative absolute error	10.0365 %	
Root relative squared error	36.4196 %	
Total Number of Instances	51	

```
=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Iris-setosa
	0.947	0.063	0.900	0.947	0.923	0.876	0.988	0.980	Iris-versicolor
	0.882	0.029	0.938	0.882	0.909	0.867	0.988	0.980	Iris-virginica
Weighted Avg.	0.941	0.033	0.942	0.941	0.941	0.909	0.992	0.986	

```
=== Confusion Matrix ===
```

a	b	c	<-- classified as
15	0	0	a = Iris-setosa
0	18	1	b = Iris-versicolor
0	2	15	c = Iris-virginica

Olhando melhor os resultados por classe

- Analisando os resultados por classe
- 100% de precisão e recall na classe setosa
- Precisão: previsões corretas / previsões para a classe

Ex: versicolor: $\text{Prec: } 18/20 = 0.9$ (90%)

- Recall: previsões corretas / total de ocorrências da classe

Ex: versicolor: $\text{Recall } 18/19 = 0.947$ (94.7%)

```
=== Summary ===
```

Correctly Classified Instances	48	94.1176 %
Incorrectly Classified Instances	3	5.8824 %
Kappa statistic	0.9113	
Mean absolute error	0.0447	
Root mean squared error	0.1722	
Relative absolute error	10.0365 %	
Root relative squared error	36.4196 %	
Total Number of Instances	51	

```
=== Detailed Accuracy By Class ===
```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	Iris-setosa
	0.947	0.063	0.900	0.947	0.923	0.876	0.988	0.980	Iris-versicolor
	0.882	0.029	0.938	0.882	0.909	0.867	0.988	0.980	Iris-virginica
Weighted Avg.	0.941	0.033	0.942	0.941	0.941	0.909	0.992	0.986	

```
=== Confusion Matrix ===
```

a	b	c	<-- classified as
15	0	0	a = Iris-setosa
0	18	1	b = Iris-versicolor
0	2	15	c = Iris-virginica

Porque o recall é importante?

- Exemplo previsão de câncer de pele com base em imagens
- É mais importante um recall alto, ele indica que estou conseguindo identificar todos ou a maioria das pessoas com câncer
- Tudo bem se a precisão for menor, vamos ter alguns falsos-positivos que serão verificados posteriormente em um exame mais detalhado
- Importante: não deixar de detectar os verdadeiros-positivos

Mas e a precisão?

- Em um sistema de prevenção a fraude em um e-commerce, é importante prever as fraudes com precisão alta
- falsos-positivos serão vendas não realizadas e clientes insatisfeitos
- Por isso, nesse caso, se deseja uma precisão alta

Weka Datasets

<https://github.com/Waikato/weka-3.8/tree/master/wekadocs/data>