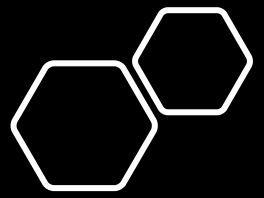
Machine Learning

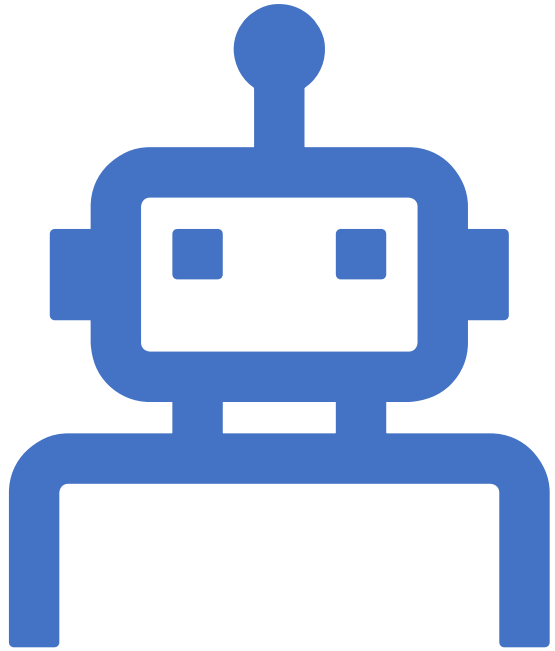


Roteiro...

- Neste tutorial: Curso rápido de Machine Learning
- Próximo tutorial: Características de Machine Learning no Spark

Machine Learning: Aplicações

- Prever fraude
- Saber se um candidato a empréstimo será bom pagador
- Classificar uma doença
- Prever se aluno vai abandonar o curso
- Anteceder a ocorrência de uma epidemia
- Reconhecer um caractere manuscrito
- Reconhecer uma imagem ou uma música



Churn

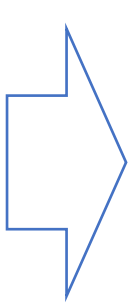
CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
619	France	Female	42	2	0	1	1	1	10134888	1
608	Spain	Female	41	1	8380786	1	0	1	11254258	0
502	France	Female	42	8	1596608	3	1	0	11393157	1
699	France	Female	39	1	0	2	0	0	9382663	0
850	Spain	Female	43	2	12551082	1	1	1	790841	0
645	Spain	Male	44	8	11375578	2	1	0	14975671	1
822	France	Male	50	7	0	2	1	1	100628	0
376	Germany	Female	29	4	11504674	4	1	0	11934688	1
501	France	Male	44	4	14205107	2	0	1	749405	0
684	France	Male	27	2	13460388	1	1	1	7172573	0
528	France	Male	31	6	10201672	2	0	0	8018112	0
497	Spain	Male	24	3	0	2	1	0	7639001	0
476	France	Female	34	10	0	2	1	0	2626098	0
549	France	Female	25	5	0	2	0	0	19085779	0
635	Spain	Female	35	7	0	2	1	1	6595165	0
616	Germany	Male	45	3	14312941	2	0	1	6432726	0
653	Germany	Male	58	1	13260288	1	1	0	509767	1
549	Spain	Female	24	9	0	2	1	1	1440641	0
587	Spain	Male	45	6	0	1	0	0	15868481	0
726	France	Female	24	6	0	2	1	1	5472403	0
732	France	Male	41	8	0	2	1	1	17088617	0
636	Spain	Female	32	8	0	2	1	0	13855546	0

Como funciona?

Treinamento do Modelo

CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
619	France	Female	42	2	0	1	1	1	10134888	1
608	Spain	Female	41	1	8380786	1	0	1	11254258	0
502	France	Female	42	8	1596608	3	1	0	11393157	1
699	France	Female	39	1	0	2	0	0	9382663	0
850	Spain	Female	43	2	12551082	1	1	1	790841	0
645	Spain	Male	44	8	11375578	2	1	0	14975671	1
822	France	Male	50	7	0	2	1	1	100628	0
376	Germany	Female	29	4	11504674	4	1	0	11934688	1
501	France	Male	44	4	14205107	2	0	1	749405	0
684	France	Male	27	2	13460388	1	1	1	7172573	0
528	France	Male	31	6	10201672	2	0	0	8018112	0
497	Spain	Male	24	3	0	2	1	0	7639001	0
476	France	Female	34	10	0	2	1	0	2626098	0
549	France	Female	25	5	0	2	0	0	19085779	0
635	Spain	Female	35	7	0	2	1	1	6595165	0
616	Germany	Male	45	3	14312941	2	0	1	6432726	0
653	Germany	Male	58	1	13260288	1	1	0	509767	1
549	Spain	Female	24	9	0	2	1	1	1440641	0
587	Spain	Male	45	6	0	1	0	0	15868481	0
726	France	Female	24	6	0	2	1	1	5472403	0
732	France	Male	41	8	0	2	1	1	17088617	0
636	Spain	Female	32	8	0	2	1	0	13855546	0

Algoritmo



Modelo

Classificar

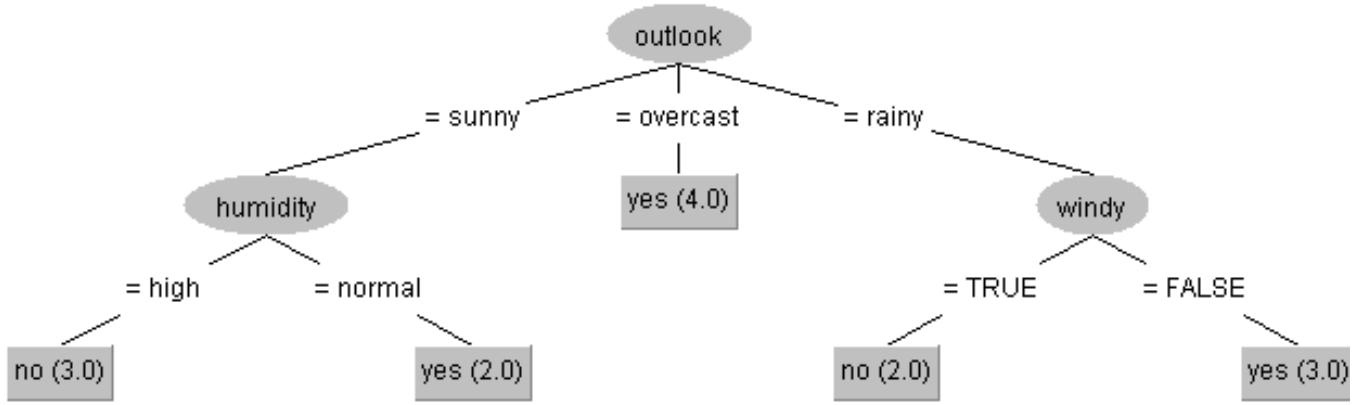
Modelo

Classificação:
Não

PROF. FERNANDO AMARAL
www.datascientist.com.br

APACHE Spark

O que é um Modelo?



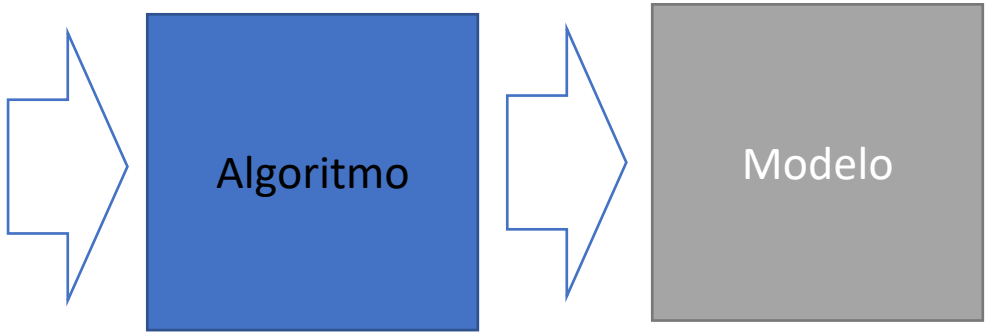
Como sei que a classificação está correta?

- Não vai acertar sempre!
- Podemos medir a performance do modelo e assim prever como vai se comportar no mundo real
- Além de treinar o modelo, precisamos testa-lo antes de aplica-lo na vida real

Como funciona?

Treinamento do Modelo

Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*
4.7	3.2	1.3	0.2	*I. setosa*
6.4	3.2	4.5	1.5	*I. versicolor*
6.9	3.1	4.9	1.5	*I. versicolor*
5.5	2.3	4.0	1.3	*I. versicolor*
6.5	2.8	4.6	1.5	*I. versicolor*
5.7	2.8	4.5	1.3	*I. versicolor*



Teste do Modelo

Sepal length	Sepal width	Petal length	Petal width	Species
6.3	3.3	4.7	1.6	*I. versicolor*
6.3	2.5	5.0	1.9	*I. virginica*
6.5	3.0	5.2	2.0	*I. virginica*
6.2	3.4	5.4	2.3	*I. virginica*
5.9	3.0	5.1	1.8	*I. virginica*

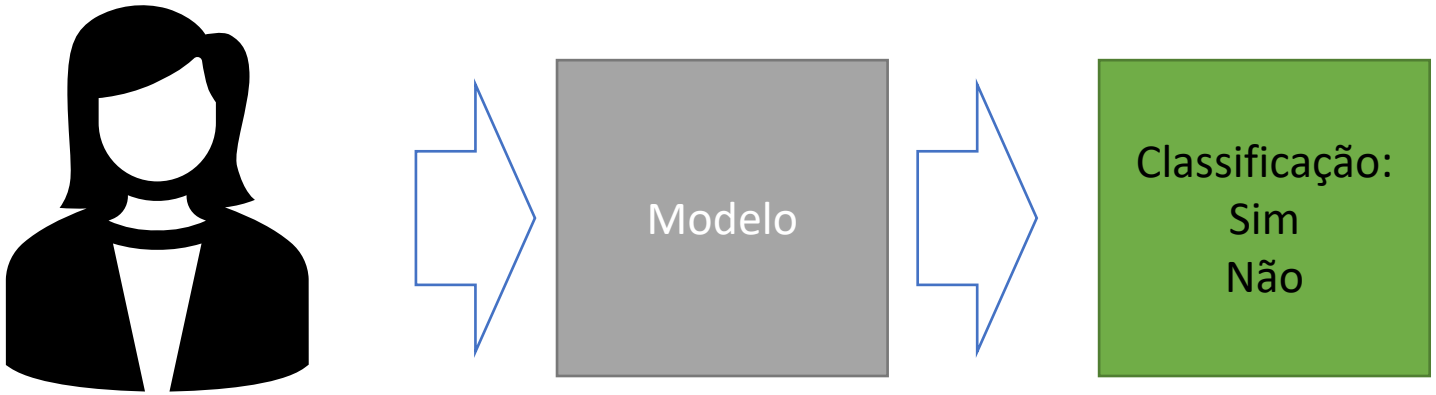
	Sim	Não
Sim	50	10
Não	5	45

Acurácia: percentual de acertos

PROF. FERNANDO AMARAL
www.datascientist.com.br

APACHE Spark

Classificar Outras Clientes



Conceitos

- Classe: é o que se busca prever ou classificar
  - Ex: Espécie de planta, doença do paciente, se o cliente é bom pagador
- Dimensão ou Atributo: são as características usadas como parâmetros para classificar
- Instância: é uma observação
- Relação: conjunto de dados

Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*
4.7	3.2	1.3	0.2	*I. setosa*
6.4	3.2	4.5	1.5	*I. versicolor*
6.9	3.1	4.9	1.5	*I. versicolor*
5.5	2.3	4.0	1.3	*I. versicolor*
6.5	2.8	4.6	1.5	*I. versicolor*
5.7	2.8	4.5	1.3	*I. versicolor*

Conceitos

- Classe é também um atributo
- Atributos nominais ou numéricos
- Classificação binária, multiclasse e multilabel

Considerações Sobre o Modelo

- Depende do Algoritmo utilizado
- Pode perder a eficiência
- Muito específico do negócio

Regressão

- Quando a classe é contínua
- Exemplo: prever a potência de um motor

Consumo	Cilindros	Cilindradas	RelEixoTraseiro	Peso	Tempo	TipoMotor	Transmissao	Marchas	Carburadors	HP
21	6	160	39	262	1646	0	1	4	4	110
21	6	160	39	2875	1702	0	1	4	4	110
228	4	108	385	232	1861	1	1	4	1	93
214	6	258	308	3215	1944	1	0	3	1	110
187	8	360	315	344	1702	0	0	3	2	175
181	6	225	276	346	2022	1	0	3	1	105
143	8	360	321	357	1584	0	0	3	4	245
244	4	1467	369	319	20	1	0	4	2	62
228	4	1408	392	315	229	1	0	4	2	95
192	6	1676	392	344	183	1	0	4	4	123
178	6	1676	392	344	189	1	0	4	4	123
164	8	2758	307	407	174	0	0	3	3	180

Métricas de Erros

- Previsão de valores numéricos (reais, inteiros)
- Métricas diferentes da previsão de categorias

- Uso:
  - Regressão clássica
  - Regressão ML
  - Series Temporais
  - Etc.

Root Mean Squared Error (RMSE)

Independente de Escala

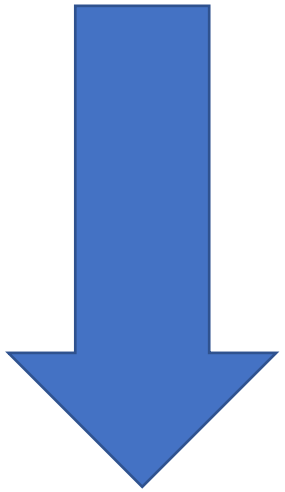
- O desvio padrão da amostra da diferença entre o previsto e o teste

Previsto	Realizado	Dif. ao Quad.
3,34	3,00	0,1156
4,18	4,00	0,0324
3,00	3,00	0
2,99	3,00	1E-04
4,51	4,50	1E-04
5,18	4,00	1,3924
8,18	4,50	13,5424

$$\text{RMSE} = \sqrt{\frac{\sum_{I=1}^{N}(p_i - t_i)^2}{N}}$$

$$\text{RMSE} = \sqrt{\frac{15{,}083}{7}}$$

$$\text{RMSE} = 1{,}46$$



PROF. FERNANDO AMARAL
www.datascientist.com.br

APACHE Spark