# DETEÇÃO REMOTA PARA A MONITORIZAÇÃO DE OCUPAÇÃO DO SOLO

## MÓDULO 3

CLASSIFICAÇÃO DE IMAGENS COM APRENDIZAGEM AUTOMÁTICA

d·gTerritório
Direção-Geral do Território

Ciências ULisboa

# CONTEÚDOS

# 01

# CLASSIFICAÇÃO DE IMAGENS DE DETEÇÃO REMOTA

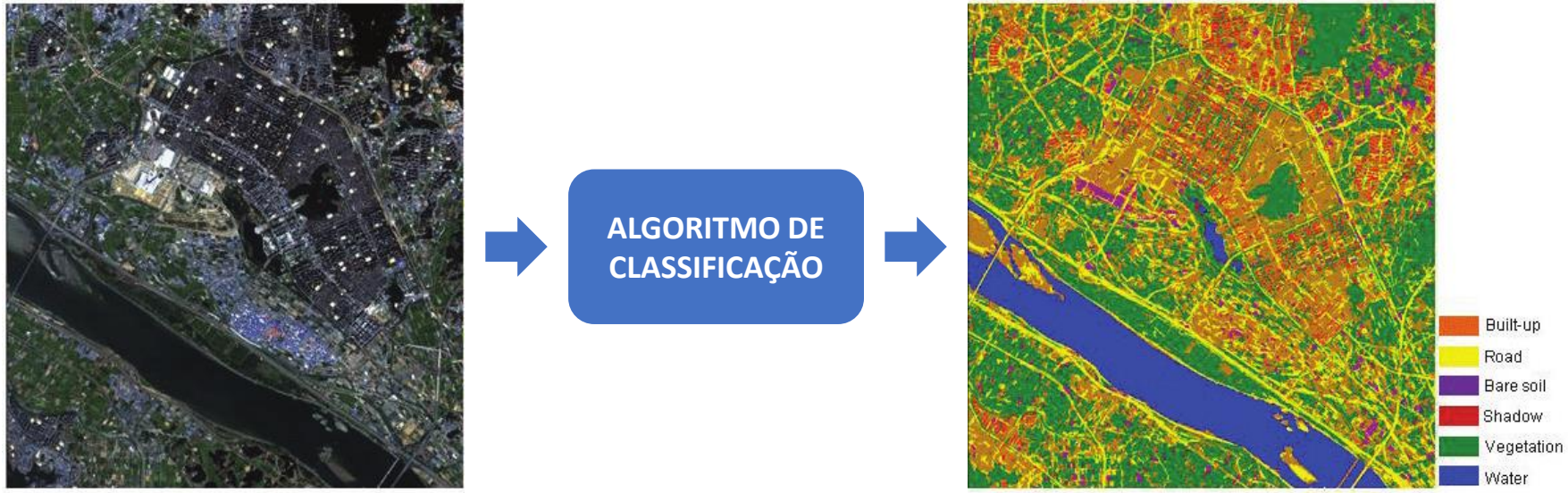
d.gTerritório
Direção-Geral do Território

Ciências
ULisboa

# CLASSIFICAÇÃO DE IMAGENS DE DETEÇÃO REMOTA

## CLASSIFICAÇÃO DE IMAGENS DE SATÉLITE

A classificação automática de imagens de satélite é uma técnica utilizada para o **reconhecimento de padrões espetrais**, que utiliza informação espetral, representada por números digitais numa ou mais bandas espetrais, para atribuir a cada píxel uma dada classe de ocupação do solo. O resultado da classificação consiste num mosaico de píxeis, pertencentes a uma dada classe, correspondendo a uma **mapa temático** da imagem original.



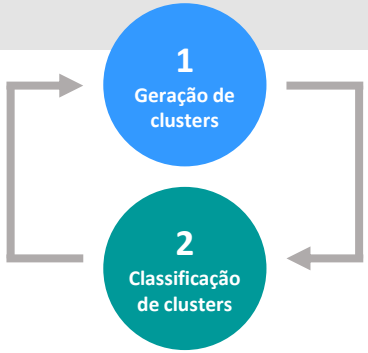

Créditos da imagem: Kim e Yeom, 2012. DOI:10.7780/kjrs.2012.28.2.203

# CLASSIFICAÇÃO DE IMAGENS DE DETEÇÃO REMOTA

## PRINCIPAIS TÉCNICAS DE CLASSIFICAÇÃO DE IMAGENS

**CLASSIFICAÇÃO NÃO-SUPERVISIONADA**

Técnica mais simples pois não requer amostras de treino. Consiste numa forma de segmentação da imagem dado que agrupa os píxeis da imagem em agregados (*clusters*) com base nas suas propriedades espetrais, sendo de seguida atribuída uma classe de ocupação do solo a cada um desses *clusters*.




**CLASSIFICAÇÃO SUPERVISIONADA**

Técnica que requer amostras de treino (amostras representativas de cada classe de ocupação do solo), cuja informação espetral correspondente (assinatura espetral) é utilizada para atribuir uma dada classe de ocupação a todos os píxeis da imagem (classificação ao nível do píxel).


**CLASSIFICAÇÃO ORIENTADA POR OBJETOS**

Técnica que segmenta uma imagem em objetos com diferentes geometrias. A classificação é efetuado com base na forma, textura, caraterísticas espetrais e contexto de cada um dos objetos (classificação por objetos).

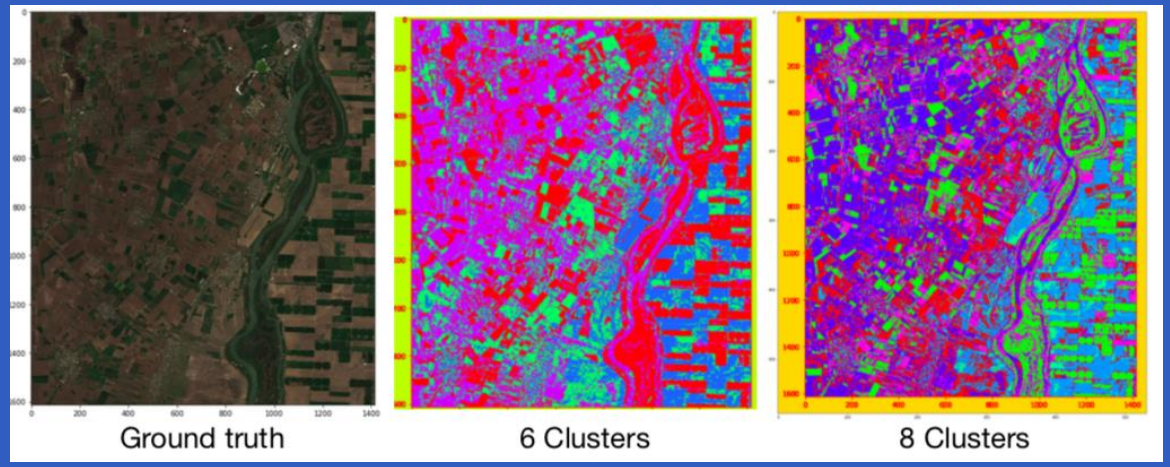

d.gTerritório Direção-Geral do Território

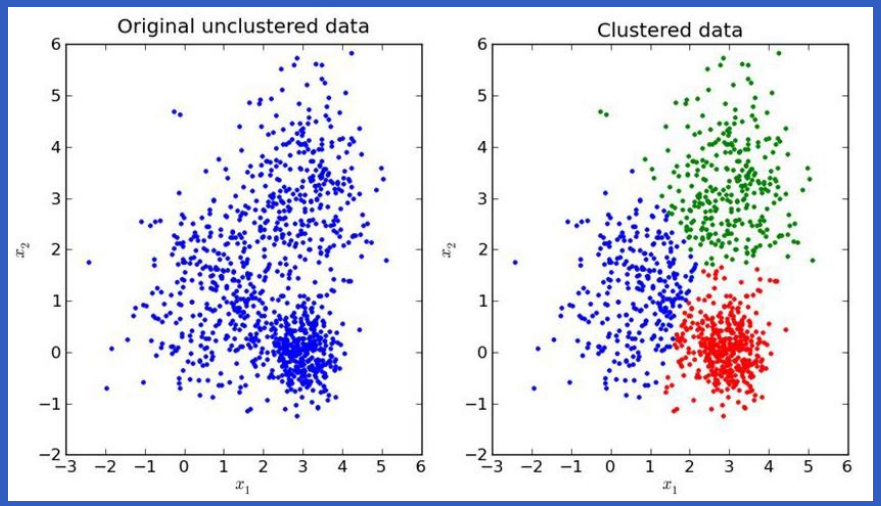
Ciências ULisboa

# CLASSIFICAÇÃO DE IMAGENS DE DETEÇÃO REMOTA

## CLASSIFICAÇÃO NÃO-SUPERVISIONADA

- A informação espetral é agrupada pelo algoritmo de classificação em conjuntos de píxeis (*clusters*) com base unicamente na informação espetral dos dados;
- A única intervenção do operador nesta fase consiste na definição do número de *clusters* a ser identificado pelo algoritmo, e ainda de outros parâmetros tais como a distância entre *clusters* e a variância dentro de cada *cluster*;
- O resultado deste processo de agregação é depois analisado pelo operador para que seja atribuída uma classe da nomenclatura a cada *cluster*, podendo dar-se o caso de haver necessidade de combinar ou de sub-dividir *clusters*, implicando uma nova aplicação do algoritmo de *clustering*.



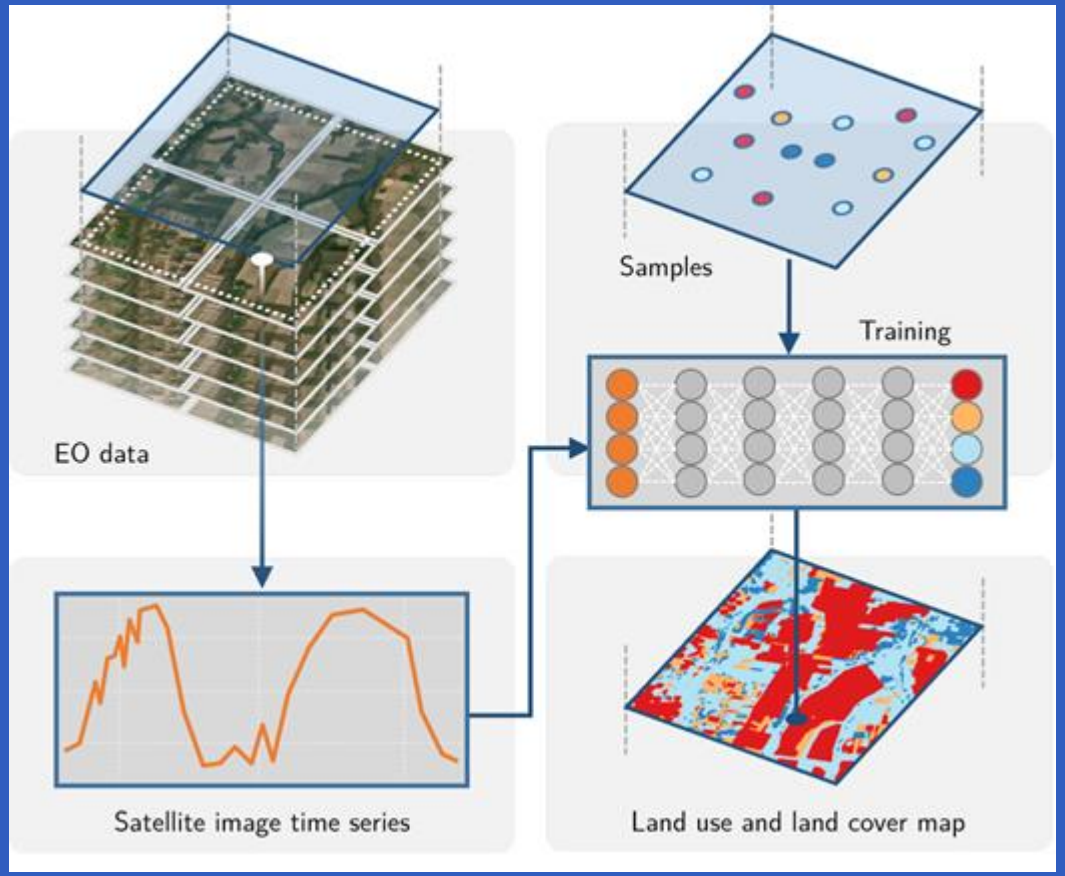




Créditos das imagens: Unsupervised Learning. Clustering for Satellite Imagery | by Daniel Moraite | DataSeries | Medium; A Beginner's Guide to Customer Segmentation with k-Means Clustering | by Bryan Tan | Medium

# CLASSIFICAÇÃO DE IMAGENS DE DETEÇÃO REMOTA

## CLASSIFICAÇÃO SUPERVISIONADA

- Um operador experiente identifica amostras homogéneas representativas de diferentes tipos de cobertura do solo (amostras de treino) nas imagens, e atribui uma dada classe (de um conjunto pré-definido de classes - nomenclatura) a cada amostra;
- A informação espetral extraída a partir de todas as bandas/imagens, para os píxeis dentro dessas áreas de treino, é usada para treinar o algoritmo de classificação;
- O algoritmo determina a assinatura espetral para cada classe e, de seguida, compara cada píxel da imagem com essas assinaturas, atribuindo-lhe a classe com a qual mais se assemelha digitalmente.

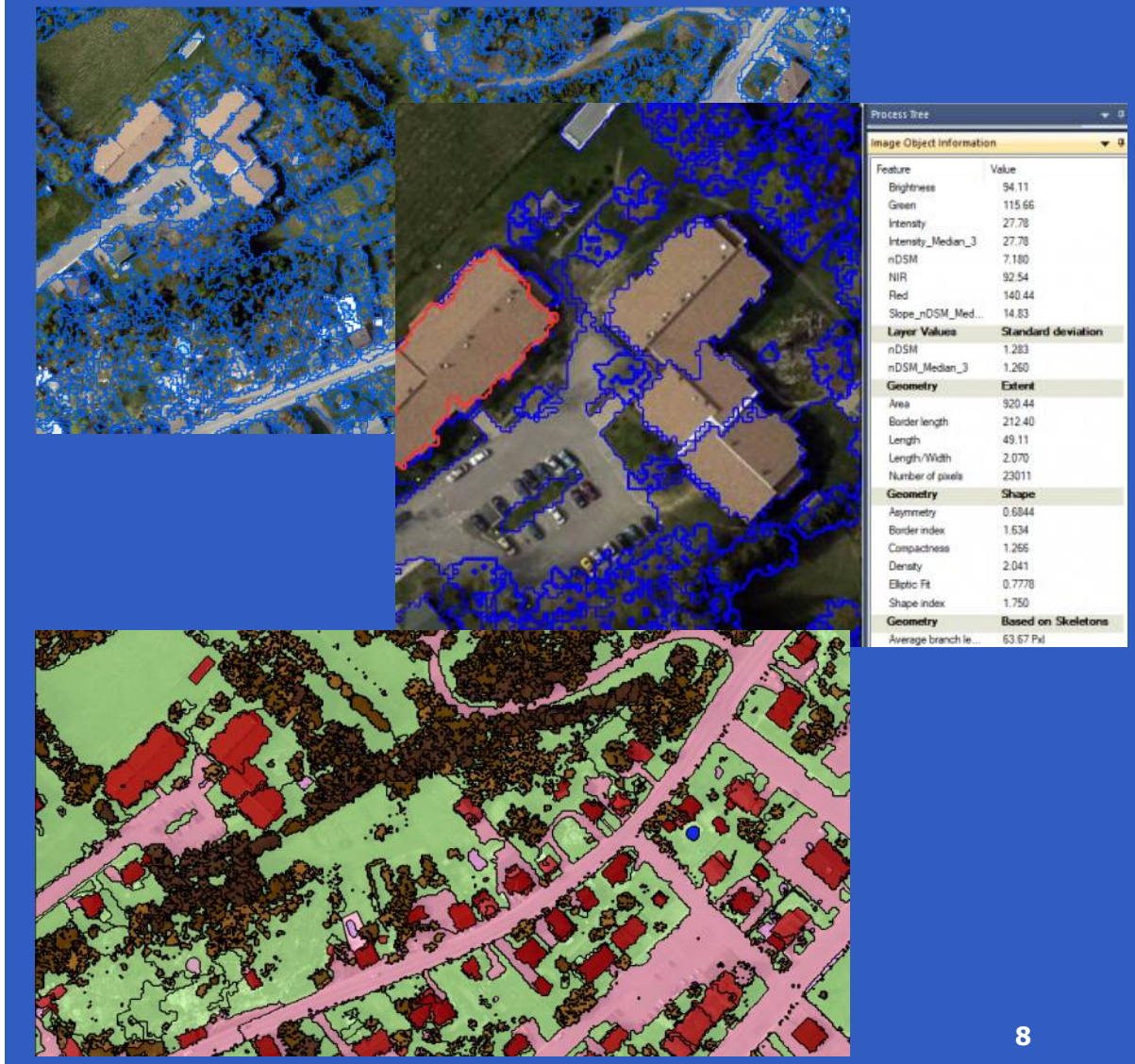



Créditos da imagem: Simões et al., 2021.
DOI:10.3390/rs13132428

# CLASSIFICAÇÃO DE IMAGENS DE DETEÇÃO REMOTA

## CLASSIFICAÇÃO ORIENTADA POR OBJETOS

- A imagem é segmentada em grupos de píxeis (objetos com diferentes geometrias) representativos dos elementos da superfície terrestre, os quais são depois convertidos para formato vetorial;
- O operador seleciona alguns desses objetos para cada classe da nomenclatura (amostras de treino), sendo calculadas diversas estatísticas para cada objeto, tais como a geometria, a área, a cor, a forma, a textura, a adjacência (contexto), entre outras;
- O algoritmo atribui uma dada classe a cada objeto com base na semelhança às estatísticas correspondentes às amostras de treino.




Créditos da imagem: OBIA - Object-Based Image Analysis (GEOBIA) - GIS Geography

# 02

# CLASSIFICAÇÃO DE IMAGENS COM APRENDIZAGEM AUTOMÁTICA (*MACHINE LEARNING*)

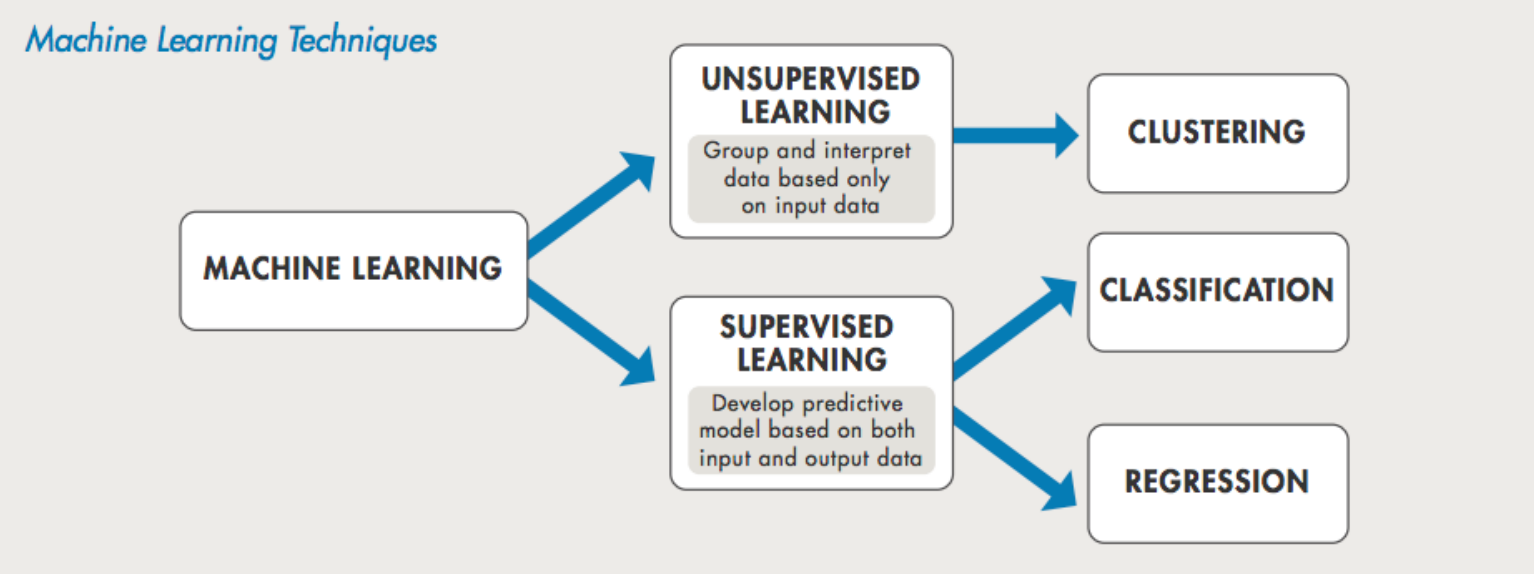
d.gTerritório
Direção-Geral do Território

Ciências
ULisboa

# CLASSIFICAÇÃO DE IMAGENS COM APRENDIZAGEM AUTOMÁTICA

## APRENDIZAGEM AUTOMÁTICA – *Machine Learning* (ML)

Técnica de inteligência artificial (permite aos computadores replicar o comportamento humano) que fornece aos computadores a habilidade de utilizar conhecimento (exposição a um grande volume de dados ao longo do tempo) para melhorar a sua performance.



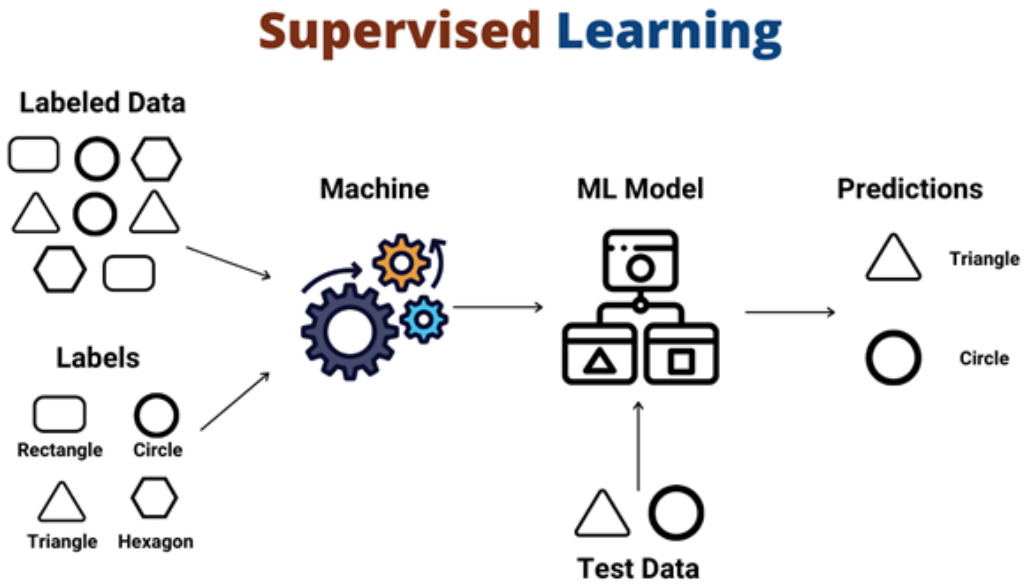

Créditos da imagem: Porntiva Visitsora-at | Medium

# CLASSIFICAÇÃO DE IMAGENS COM APRENDIZAGEM AUTOMÁTICA

## APRENDIZAGEM AUTOMÁTICA SUPERVISIONADA

Consiste numa abordagem em que um algoritmo (programa de computador) aprende a partir de um conjunto de dados de entrada (treino do modelo) e, em seguida, aplica esse conhecimento para atribuir uma dada classe a dados nunca vistos pelo modelo de aprendizagem automática (predição).

- Requer dados de entrada (treino) e de teste aos quais esteja associada uma determinada classe – *label;*
- Os dados de treino são previamente definidos e categorizados por um supervisor, o que constitui um processo bastante moroso;
- O modelo aprende a relação entre os dados de treino e os de teste (função de mapeamento);
- O modelo treinado é depois utilizado para classificar dados nunca vistos.




Créditos da imagem: Supervised and Unsupervised Learning (an Intuitive Approach) | by Metehan Kozan | Medium

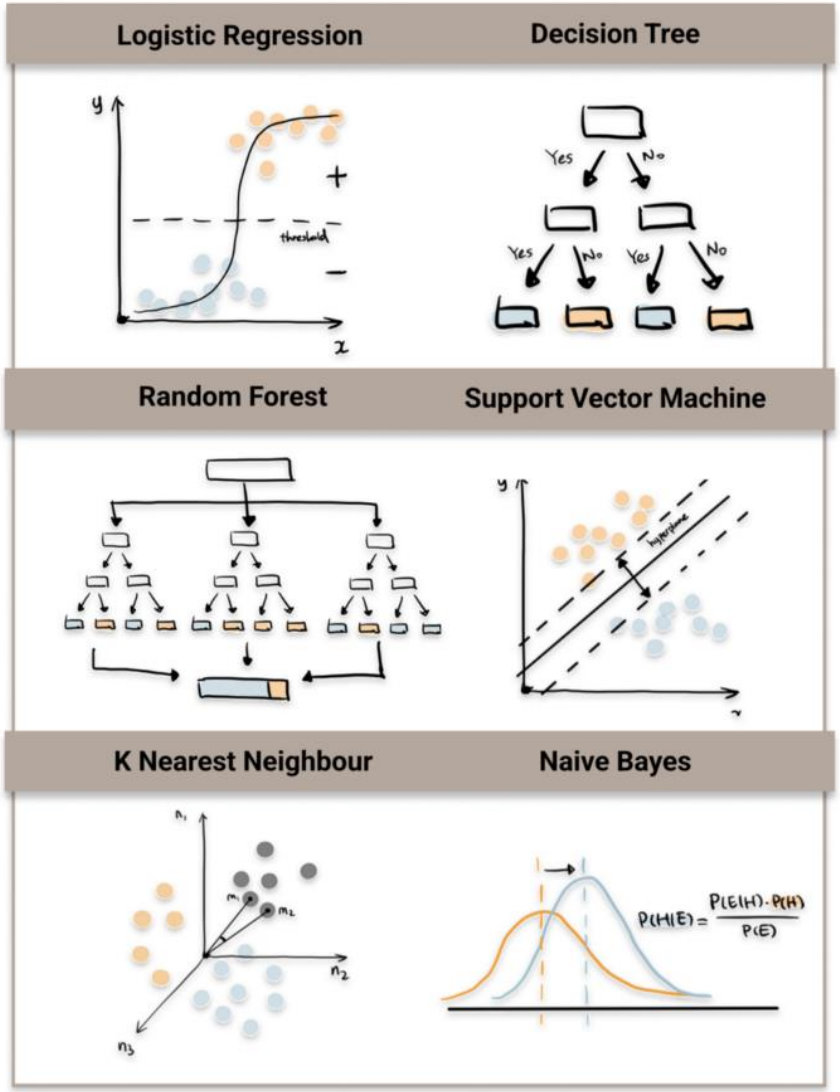
# CLASSIFICAÇÃO DE IMAGENS COM APRENDIZAGEM AUTOMÁTICA

## ALGORITMOS DE APRENDIZAGEM SUPERVISIONADA PARA CLASSIFICAÇÃO

Os algoritmos de aprendizagem supervisionada categorizam um conjunto de dados em classes, podendo ser utilizados em problemas de classificação binária (2 classes) ou de classificação multi-classe (mais de 2 classes). Embora existam diversos algoritmos de aprendizagem automática para classificação, são referidos apenas alguns:

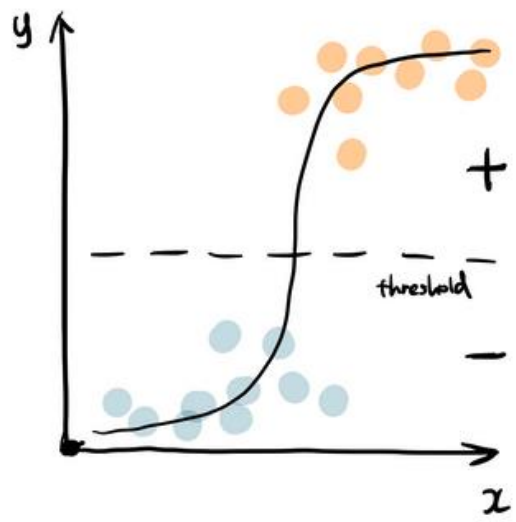
**CLASSIFICAÇÃO**

- **Logistic Regression** (regressão logística)
- **Decision Tree** (árvore de decisão)
- **Random Forest** (floresta aleatória)
- **Support Vector Machines** (máquinas de vetores de suporte)
- **K Nearest Neighbour** (K vizinho mais próximo)
- **Naïve Bayes**




Créditos da imagem: [Top 6 Machine Learning Algorithms for Classification | by Destin Gong | Towards Data Science]

# CLASSIFICAÇÃO DE IMAGENS COM APRENDIZAGEM AUTOMÁTICA

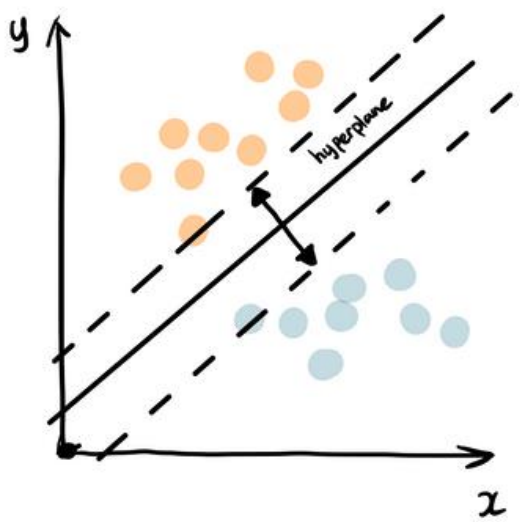



**LOGISTIC REGRESSION**

• Utiliza uma ou mais variáveis independentes para determinar o resultado, o qual só pode ter 2 resultados possíveis;

• Encontra a melhor relação de ajustamento entre a variável dependente e as variáveis independents;

• Melhor que outros algoritmos de classificação binária, tais como o KNN dado que explica quantitativamente os fatores que determinam a classificação.

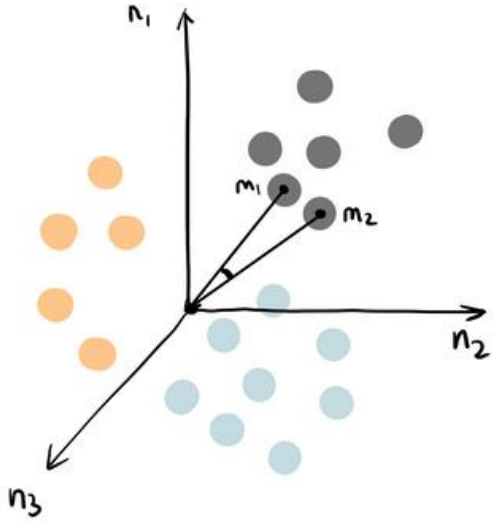



**SUPPORT VECTOR MACHINE**

• Encontra a melhor maneira de classificar os dados com base na sua posição relativamente a uma fronteira (hiperplano) de separação entre 2 classes;

• Esta fronteira maximiza a distância entre os elementos das diferentes classes;

• Assemelha-se às árvores de decisão e às florestas aleatórias, podendo também ser usado para regressão.




**K NEAREST NEIGHBOUR**

• Representa cada elemento da amostra de treino num espaço n dimensional, definido pelas n variáveis;

• Calcula para cada elemento a distância (euclideana, entre outras) aos k elementos mais próximos (na sua vizinhança);

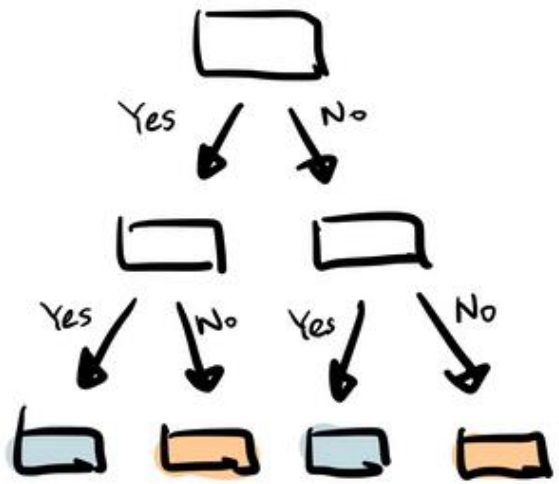
• Atribui uma classe a elementos não vistos pelo algoritmo com base na moda das classes que se encontram na sua vizinhança.

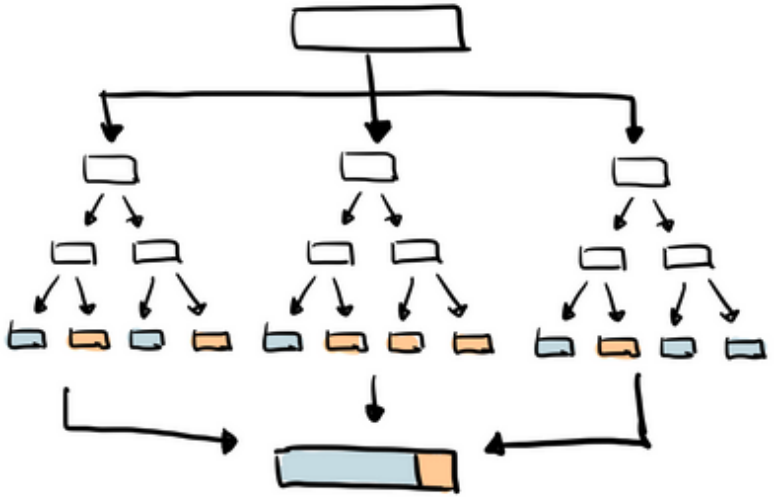
d.gTerritório Direção-Geral do Território

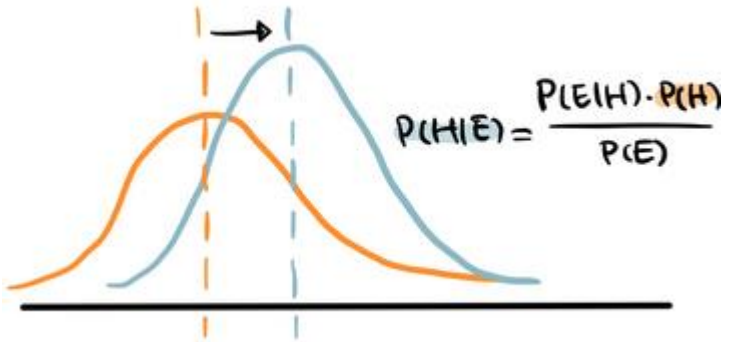
Ciências ULisboa

Créditos da imagem: Top 6 Machine Learning Algorithms for Classification | by Destin Gong | Towards Data Science

# CLASSIFICAÇÃO DE IMAGENS COM APRENDIZAGEM AUTOMÁTICA






**DECISION TREE**

• Constroi uma árvore com vários ramos de uma forma hierárquica, correspondendo cada ramo a uma condição *if-else*;

• Os ramos vão sendo ramificados através da partição do conjunto de dados com base nas variáveis mais relevantes;

• A classe a atribuir a cada píxel corresponde à classe nas folhas da árvore (quando terminam as ramificações).

**RANDOM FOREST**

• Consiste num conjunto de várias árvores de decisão, agregando o resultado de múltiplas classificações;

• Utiliza uma técnica designada por *bagging* que permite que cada árvore seja treinada com uma amostra aleatória dos dados originais;

• A atribuição de uma dada classe a cada píxel resulta de uma votação por maioria do resultado obtido por cada árvore;

• Comparativamente às árvores de decisão, permite uma melhor generalização do modelo.

**NAÏVE BAYES**

• Calcula a probabilidade condicional de cada variável com base em conhecimento adquirido *a priori* e na assunção de Naïve de que cada variável é independente das restantes;

• Efetua a predição tendo em conta a classe com a maior probabilidade;

• Apresenta a vantagem, face a outros algoritmos, de não requerer uma grande quantidade de dados de treino.



Créditos da imagem: Top 6 Machine Learning Algorithms for Classification | by Destin Gong | Towards Data Science

# 03

# CLASSIFICAÇÃO SUPERVISIONADA: ALGORITMO RANDOM FOREST

d.gTerritório Direção-Geral do Território

Ciências ULisboa

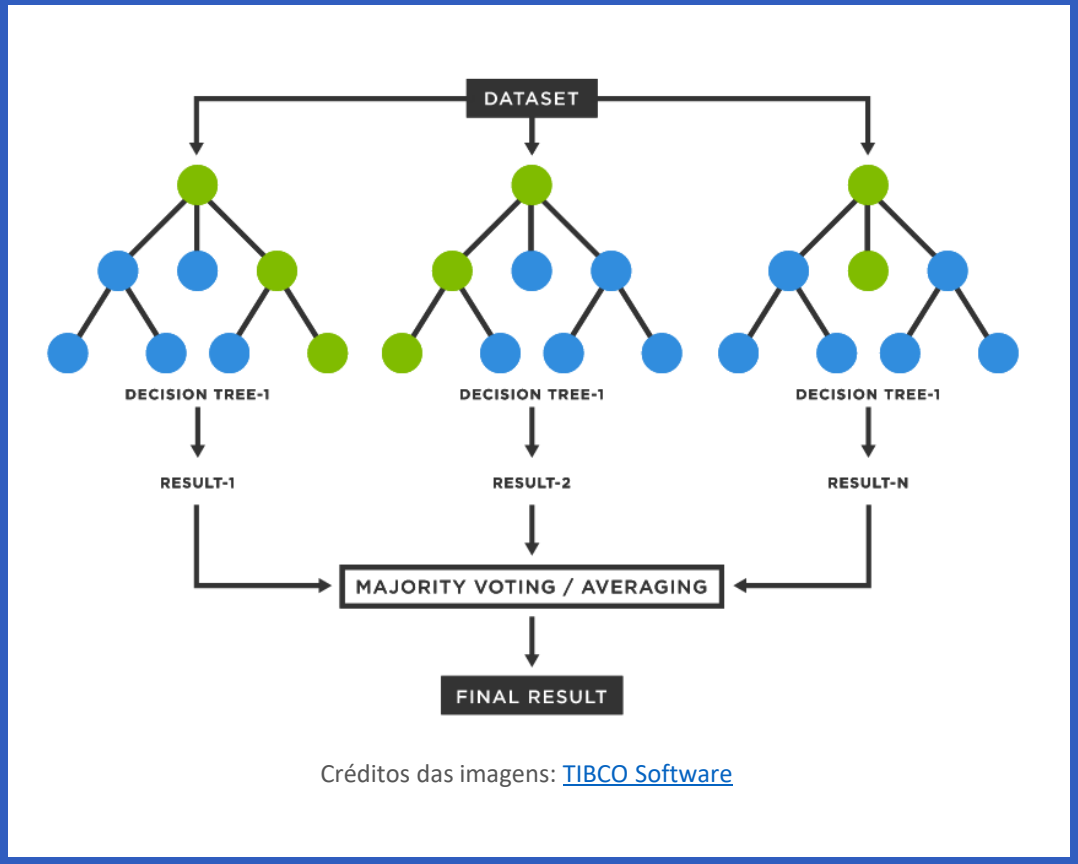
# CLASSIFICAÇÃO SUPERVISIONADA: ALGORITMO RANDOM FOREST

## RANDOM FOREST

A floresta aleatória (*Random Forest*) é um algoritmo de **aprendizagem automática supervisionada** amplamente utilizado em tarefas de **classificação** e **regressão**, embora sejam obtidos melhores resultados em tarefas de classificação.

O algoritmo constrói diversas árvores de decisão com base em diferentes amostras do conjunto de dados e considera o **voto por maioria**, no caso de uma classificação, e a **média**, no caso de uma regressão.

O termo *Random Forest* foi proposto pela primeira vez em 1995 por Tim Kam Ho. Em 2006, Leo Breiman e Adele Cutler desenvolveram o algoritmo e criaram as atuais florestas aleatórias.

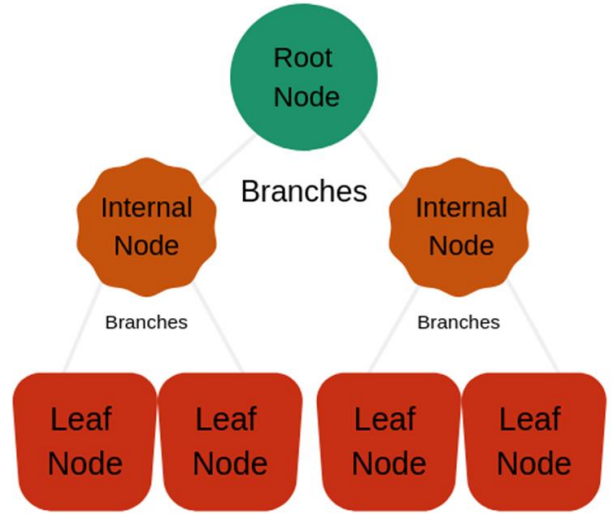


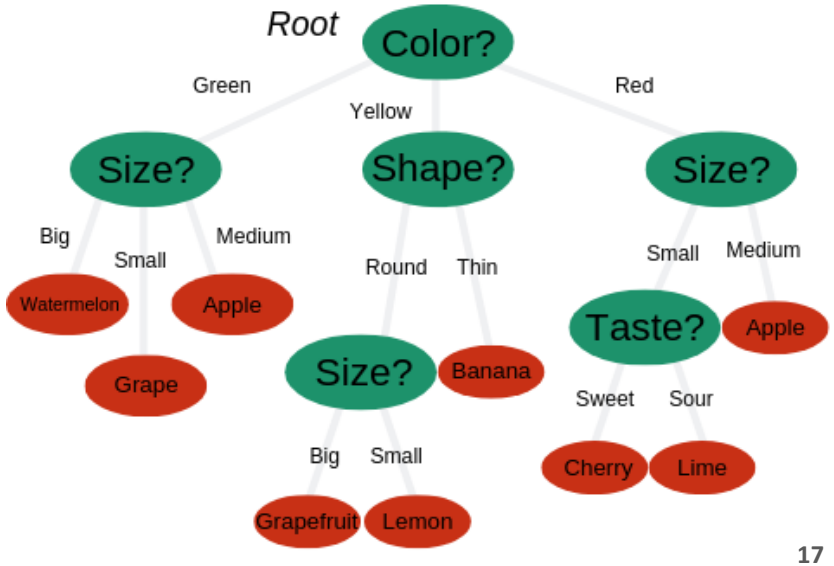

Créditos das imagens: TIBCO Software

# CLASSIFICAÇÃO SUPERVISIONADA: ALGORITMO RANDOM FOREST

## ÁRVORES DE DECISÃO

- Uma árvore de decisão consiste num **processo de tomada de decisão**;
- Este processo é iniciado a partir de um **nodo raíz** (*root node*), que gera **ramos** (*branches*) dando origem a **nodos internos** (*internal nodes*), e assim sucessivamente, até se obter uma **folha** (*leaf node*), que corresponde a um nodo que não tem descendentes (ou seja, ramos);
- Em cada nodo é colocada uma questão de forma a permitir a classificação dos dados, sendo que os seus ramos representam as diferentes possibilidades a que esse nodo pode conduzir. Uma folha corresponde ao final de uma árvore de decisão.



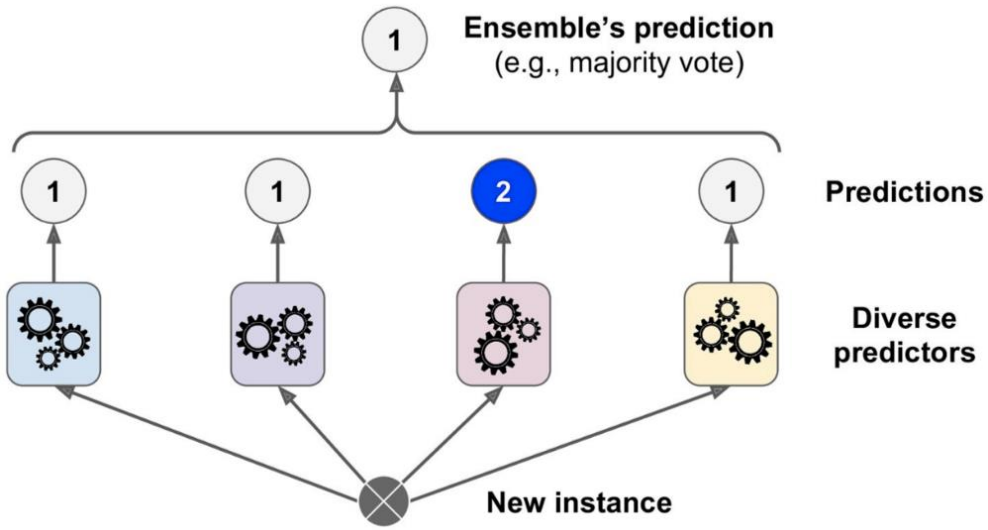




Créditos das imagens: Random Forest Algorithm for Machine Learning | by Madison Schott | Capital One Tech | Medium

# CLASSIFICAÇÃO SUPERVISIONADA: ALGORITMO RANDOM FOREST

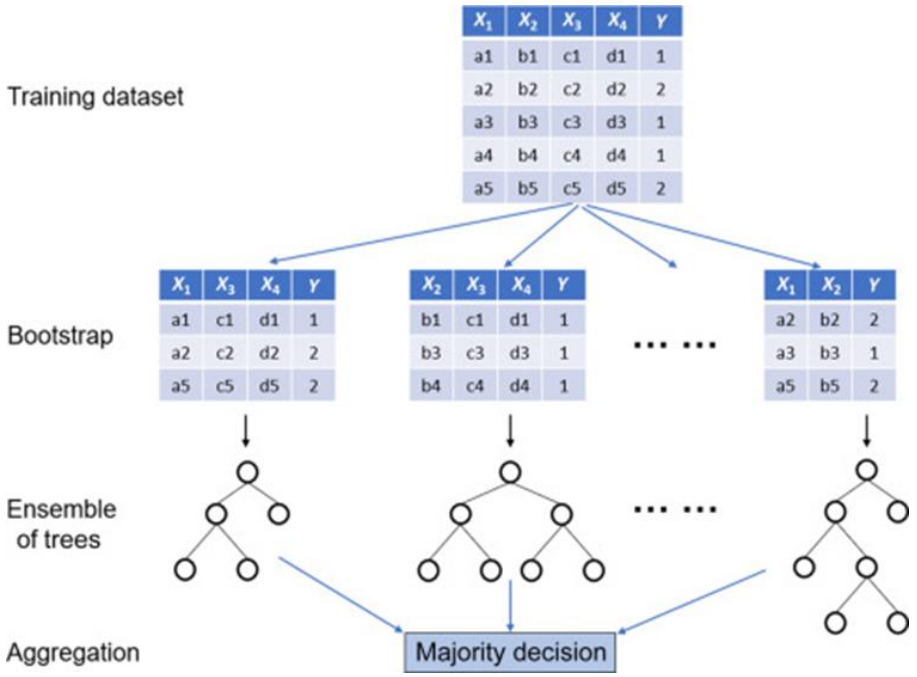
O algoritmo *Random Forest* utiliza um método de **aprendizagem em conjunto** (*ensemble learning*), designado por ***bagging*** (combinação das técnicas de *bootstrap* e de *aggregation*), que consiste em combinar múltiplas árvores de decisão individuais por forma a melhorar a performance do modelo, uma vez que desta forma é introduzida aleatoriedade no modelo.

Em vez de depender de uma única árvore de decisão, o algoritmo considera a predição de cada árvore e, com base no **voto por maioria** das diversas predições, prevê o resultado final.




Créditos das imagens: StackExchange

# CLASSIFICAÇÃO SUPERVISIONADA: ALGORITMO RANDOM FOREST




Créditos das imagens: Misra & Lin, 2020

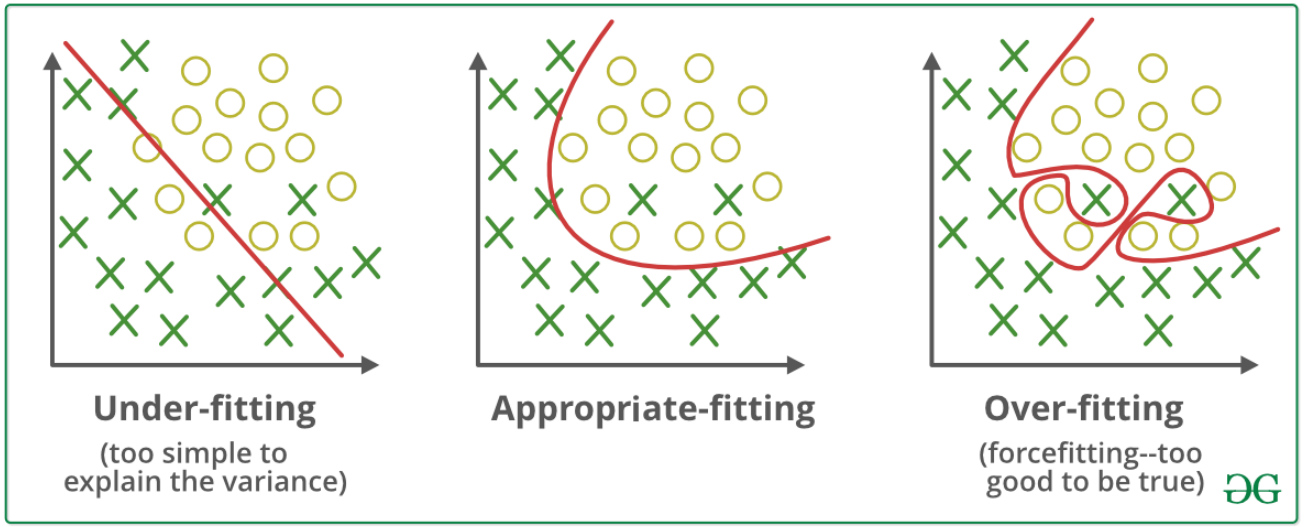
**BOOTSTRAPPING:** As várias árvores de decisão são treinadas em paralelo com diferentes subconjuntos do conjunto de dados de treino (subconjuntos de variáveis e/ou de amostras), assegurando a unicidade de cada árvore de decisão e, consequentemente, a **redução da variância global do modelo**.

**AGGREGATION:** A agregação dos diferentes modelos (árvores de decisão) permite a generalização do modelo final, sendo obtida uma predição com base numa votação por maioria. O algoritmo RF tende a mostrar uma melhor performance comparativamente a outros algoritmos, por permitir o aumento da exatidão da predição e a **redução do sobre-ajustamento (*over-fitting*) aos dados**.

# CLASSIFICAÇÃO SUPERVISIONADA: ALGORITMO RANDOM FOREST

O efeito de **sobre-ajustamento** (*over-fitting*) ocorre quando um qualquer modelo de aprendizagem automática está demasiadamente dependente dos dados (sobre-ajustado), apresentando uma correspondência quase total a um subconjunto específico de dados e falhando quando testado com outros subconjuntos (dados nunca vistos).

Por outro lado, o **sub-ajustamento** (*under-fitting*) ocorre quando um modelo não consegue expressar a maior parte da variabilidade dos dados, não permitindo predições corretas com os dados com que foi treinado.




Créditos das imagens: Underfitting and Overfitting – GeeksforGeeks

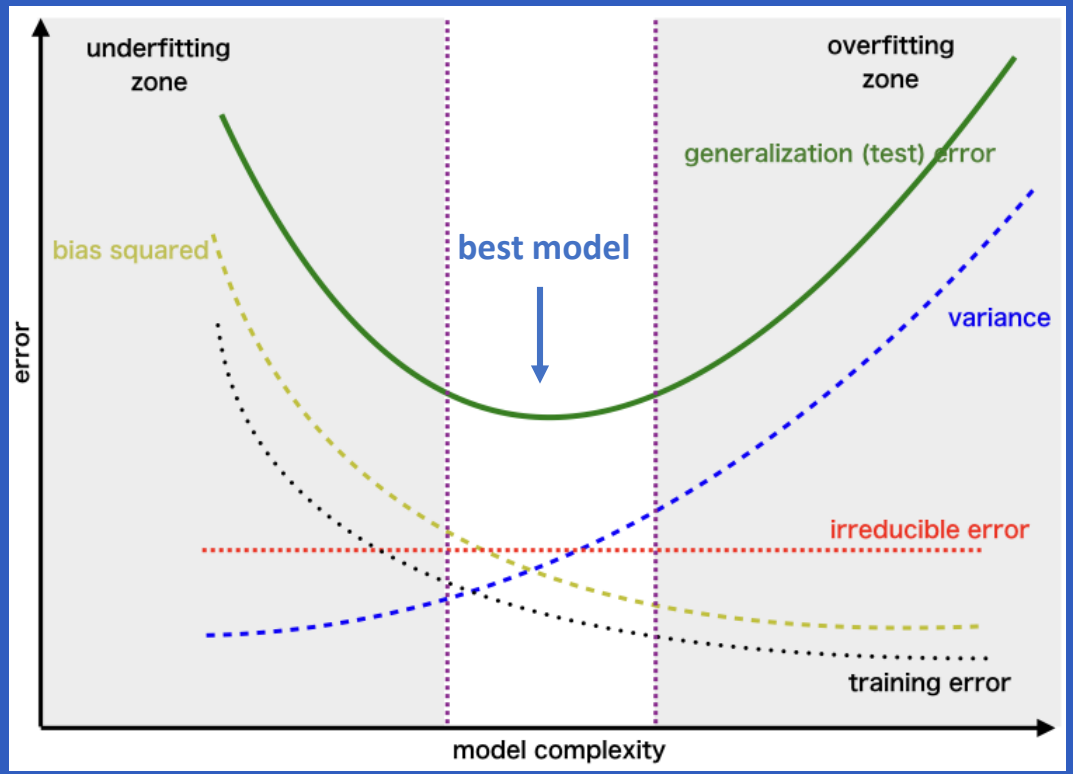
# CLASSIFICAÇÃO SUPERVISIONADA: ALGORITMO RANDOM FOREST

Se o modelo é simples (por exemplo, uma equação linear), este vai apresentar um enviesamento elevado (elevado erro nos dados de treino) e uma variância baixa (com pouca variação relativamente ao valor médio).

Se o modelo é complexo (por exemplo, uma equação de grau elevado), este vai apresentar uma variância elevada (com grande variação relativamente ao valor médio) e um enviesamento baixo (erro nos dados de treino reduzido), fazendo com que o modelo funcione bem nos dados de treino, mas não seja capaz de generalizar no caso de dados nunca vistos (erro nos dados de teste elevado).

Créditos das imagens: Bias-Variance Trade off - Machine Learning - GeeksforGeeks

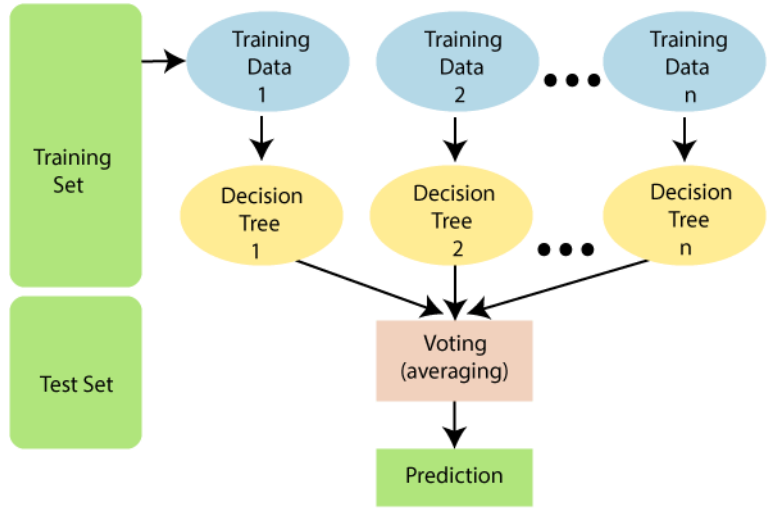
**RELAÇÃO CUSTO-BENEFÍCIO ENTRE O ENVIESAMENTO (*BIAS*) E A VARIÂNCIA (*VARIANCE*)**

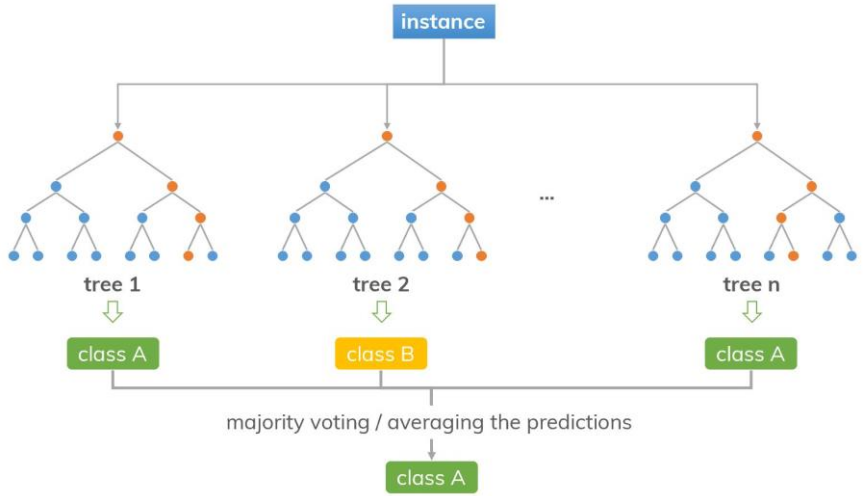


# CLASSIFICAÇÃO SUPERVISIONADA: ALGORITMO RANDOM FOREST

## ETAPAS DE UMA CLASSIFICAÇÃO COM O RANDOM FOREST

1. Seleção aleatória de k amostras a partir do conjunto de dados de treino;
2. Construção das árvores de decisão associadas aos subconjuntos de dados de treino;
3. Seleção das N árvores de decisão que vão gerar os modelos;
4. Repetição dos passos 1 & 2;
5. Realização da predição de cada árvore de decisão para amostras novas (nunca vistas pelo modelo), e atribuição de uma categoria (classe), resultante de um votação, a cada uma dessas amostras.




Créditos das imagens: Machine Learning Tutorial - Javatpoint

# CLASSIFICAÇÃO SUPERVISIONADA: ALGORITMO RANDOM FOREST

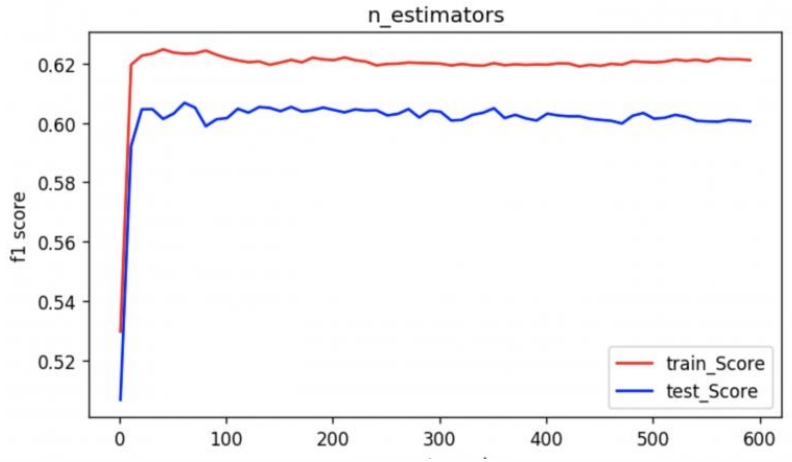
## HIPERPARÂMETROS DO ALGORITMO RANDOM FOREST

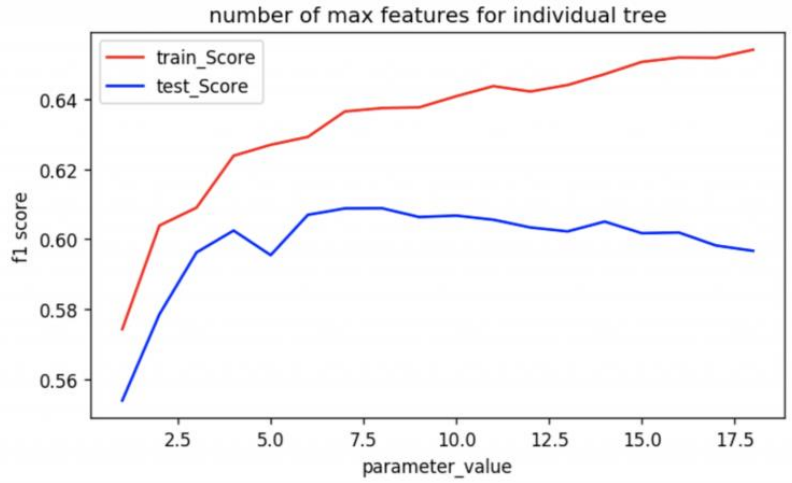
Utilizados para aumentar o poder preditivo do modelo ou para o tornar mais rápido, sendo os seguintes os mais relevantes:

| Nº DE ÁRVORES NA FLORESTA (n_estimators) |
|---|
| Número de árvores construídas pelo algoritmo, sendo que um maior número de árvores não conduz necessariamente a um melhor ajustamento do modelo (embora também não o degrade), podendo, no entanto, conduzir a um aumento da sua complexidade (tempo de processamento). |
| **100**[1] |

[1]Todos os valores padrão (default) apresentados são os propostos na biblioteca scikit-learn (machine learning in Python) da liguagem de programação Python.

| NÚMERO MÁXIMO DE VARIÁVEIS (max_features) |
|---|
| Número máximo de variáveis a serem utilizadas na construção de uma dada árvore. A performance do modelo aumenta inicialmente com o aumento de variáveis, mas após determinado ponto, o modelo começa a sofrer de sobre-ajustamento. |
| **"auto"**<br>(corresponde à raiz quadrada do número total de variáveis) |




Créditos das imagens: Random Forest Hyperparameter Tuning in Python | Machine learning

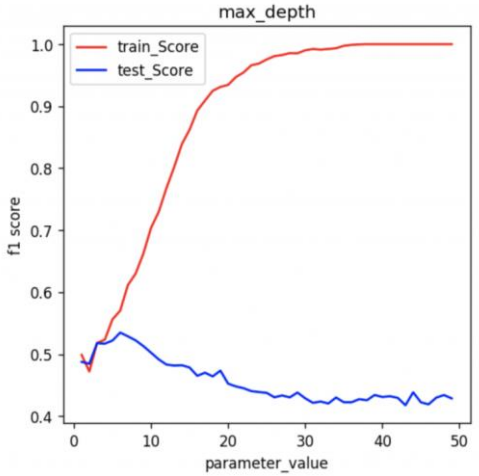
# CLASSIFICAÇÃO SUPERVISIONADA: ALGORITMO RANDOM FOREST

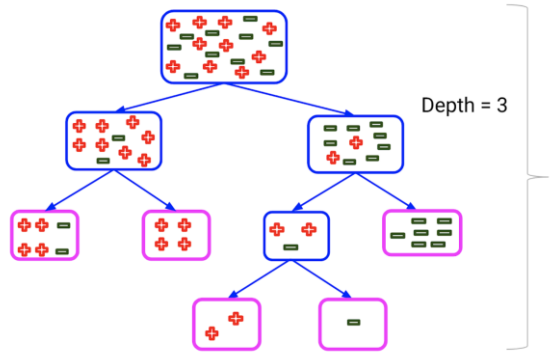
**Profundidade máxima de uma árvore (max_depth)**

Profundidade máxima até à qual as árvores podem crescer, sendo dos hiperparâmetros mais relevantes quando se pretende aumentar a exatidão do modelo. Uma maior profundidade conduz a um aumento da exatidão do modelo, mas apenas até determinado limite, a partir do qual o modelo fica sobre-ajustado.

**None**

(nodos continuam a crescer até que todas as folhas se tornem puras ou todas as folhas contenham menos que o valor atribuído a outro hiperparâmetro, o **min_samples_split**)

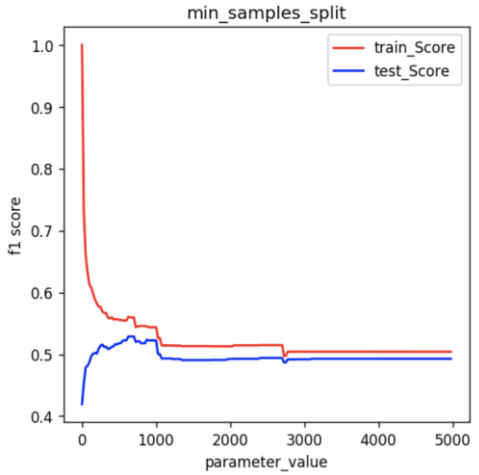


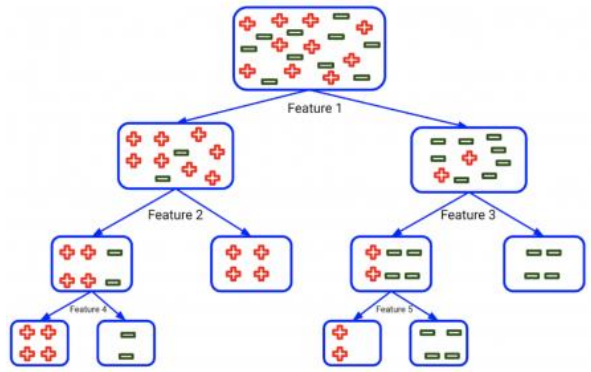




**Nº mínimo de amostras requerido para a divisão de um nodo interno (min_samples_split)**

Número mínimo de amostras que um nodo interno deve conter para se dividir em outros nós. Se o valor adotado for baixo, a árvore vai continuar a crescer e a dar origem a um sobre-ajustamento do modelo. Caso contrário, se for elevado, o número total de divisões é reduzido, o que limita o número de parâmetros do modelo e, consequentemente, reduz o sobre-ajustamento. No entanto, este número não deve ser muito elevado para que o número de parâmetro não se reduza de forma abrupta tornando o modelo insuficiente.

**2**



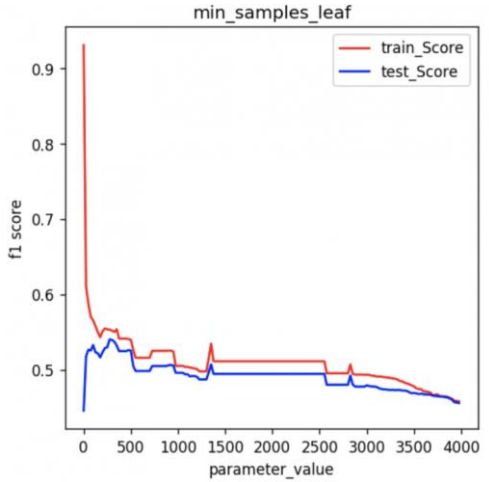

Créditos das imagens: [Random Forest Hyperparameter Tuning in Python | Machine learning]

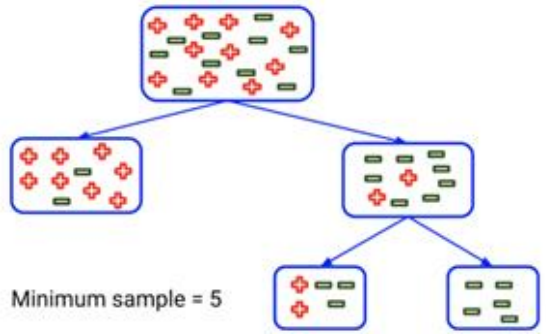
# CLASSIFICAÇÃO SUPERVISIONADA: ALGORITMO RANDOM FOREST

## Nº MÍNIMO DE AMOSTRAS REQUERIDO NUM NODO FOLHA (min_samples_leaf)

Número mínimo de amostras que um nodo deve conter após ser dividido . Tal como no caso do hiperparâmetro (min_samples_split), um menor valor pode dar origem ao sobre-ajustamento do modelo, e um maior valor pode tornar o modelo insuficiente.

**1**
(corresponde a uma amostra para o ramo da esquerda e outra para o ramo da direita)

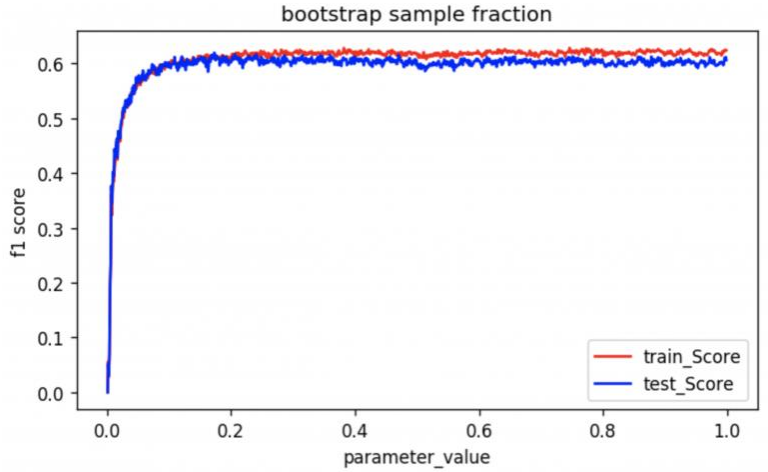



## NÚMERO MÁXIMO DE AMOSTRAS (max_samples)

Número máximo de amostras do conjunto de dados de treino que é utilizado para treinar cada árvore individual. Embora se pensa que quanto mais dados melhor, a performeance do modelo sobe abruptamente, saturando logo de seguida. No exemplo, a performance do modelo atinge um máximo apenas com 20% dos dados de treino originais.

**None**
(corresponde a utilizarem-se todas as amostras de treino)



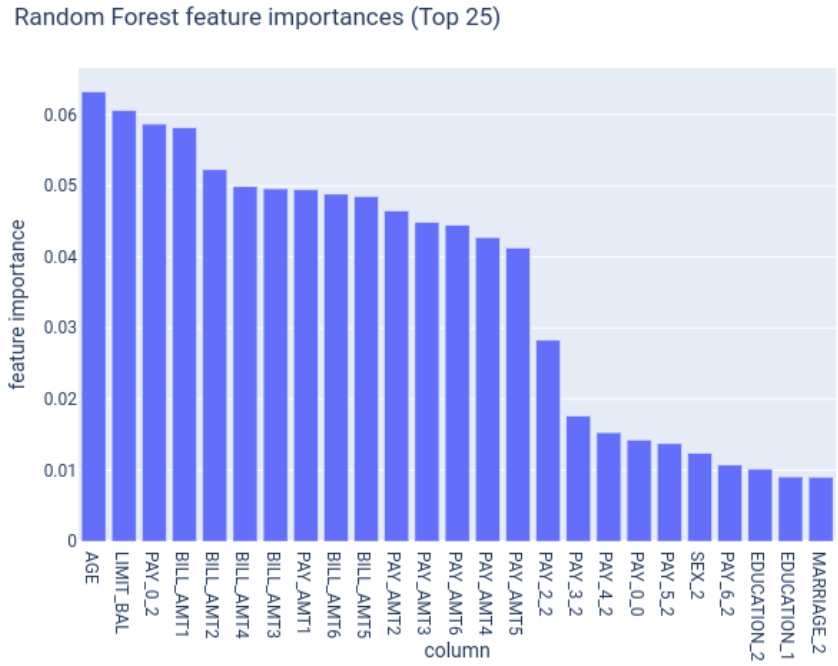

Créditos das imagens: Random Forest Hyperparameter Tuning in Python | Machine learning

# CLASSIFICAÇÃO SUPERVISIONADA: ALGORITMO RANDOM FOREST

## IMPORTÂNCIA DAS VARIÁVEIS – *Feature Importance*

O algoritmo RF permite avaliar a **importância relativa de cada uma das variáveis** (features) para a predição. Esta relevância é avaliada pela análise do número de nodos das árvores, que utilizam uma determinada variável, e que reduzem a impureza geral do modelo. Este valor é calculado automaticamente para cada variável após o treino do modelo e normaliza os resultados para que a soma de todas as importâncias seja igual a 1.

Dado que as variáveis menos relevantes não contribuem o suficiente para o processo de predição, estas podem ser **removidas do treino do modelo**, uma vez que quanto mais variáveis existirem maior será a probabilidade de ocorrer um sobre-ajustamento do modelo e vice-versa.

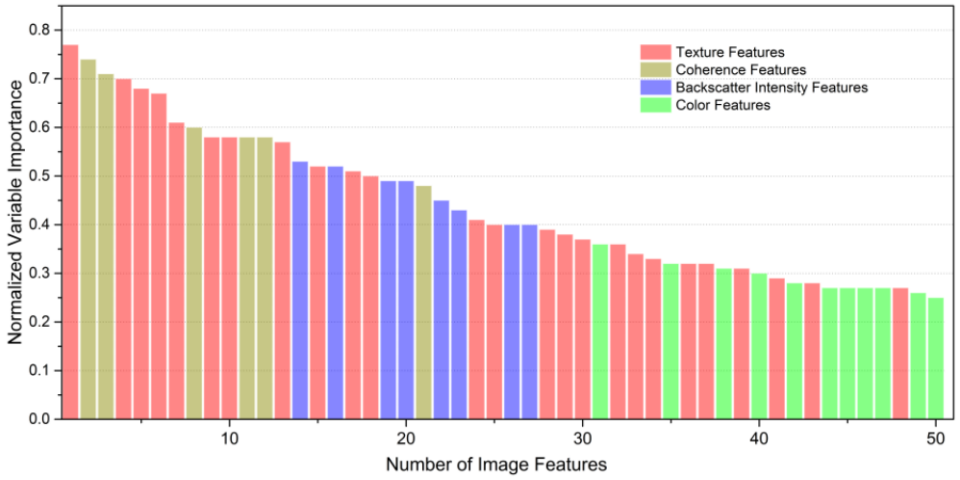


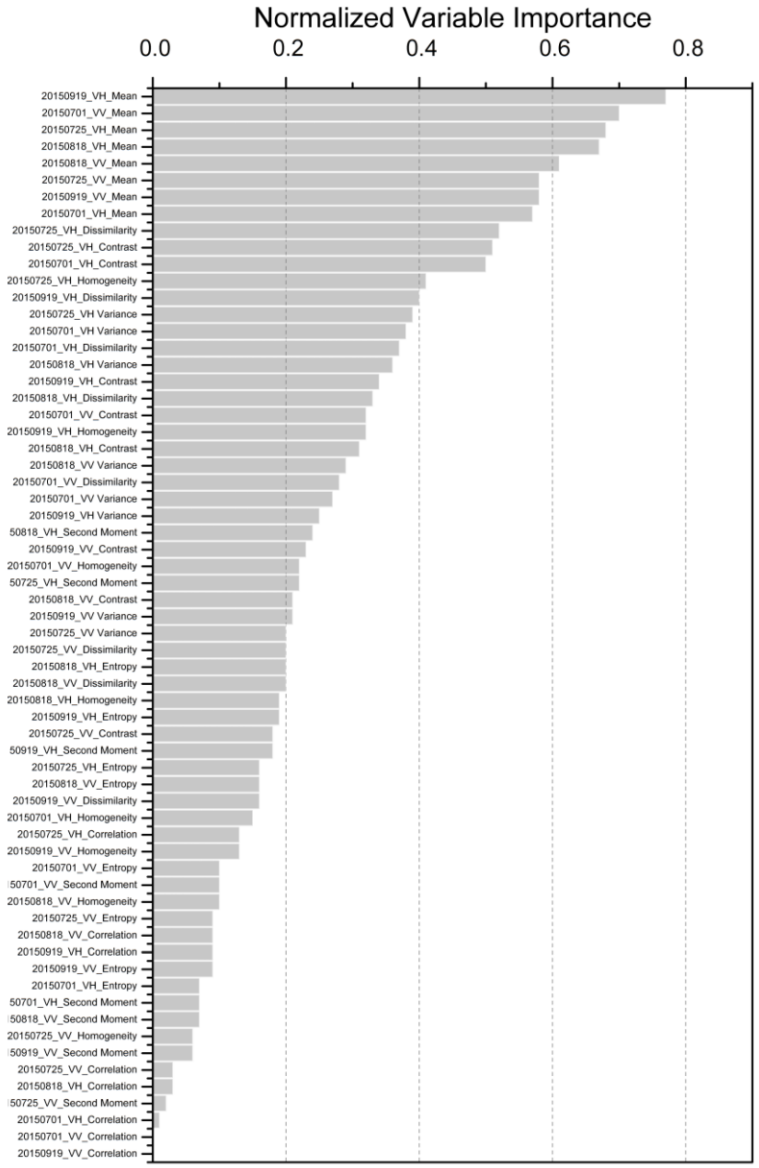

Créditos das imagens: [Explain your machine learning with feature importance | by Naren Santhanam | Towards Data Science]

# CLASSIFICAÇÃO SUPERVISIONADA: ALGORITMO RANDOM FOREST

A figura em baixo apresenta as primeiras 50 variáveis derivadas de dados do satélite Sentinel-1 ordenadas por importância utilizadas para classificação de cobertura do solo em meio urbano por Zhou *et al*. (2018).

De acordo com a ordenação (*ranking*), os elementos de textura foram as variáveis que mais contribuiram para a classificação, seguidas pela coerência interferométrica e a retrodispersão; sendo os elementos de cor (composições falsa-cor obtidas com a dupla polarização) os que menos contribuiram.



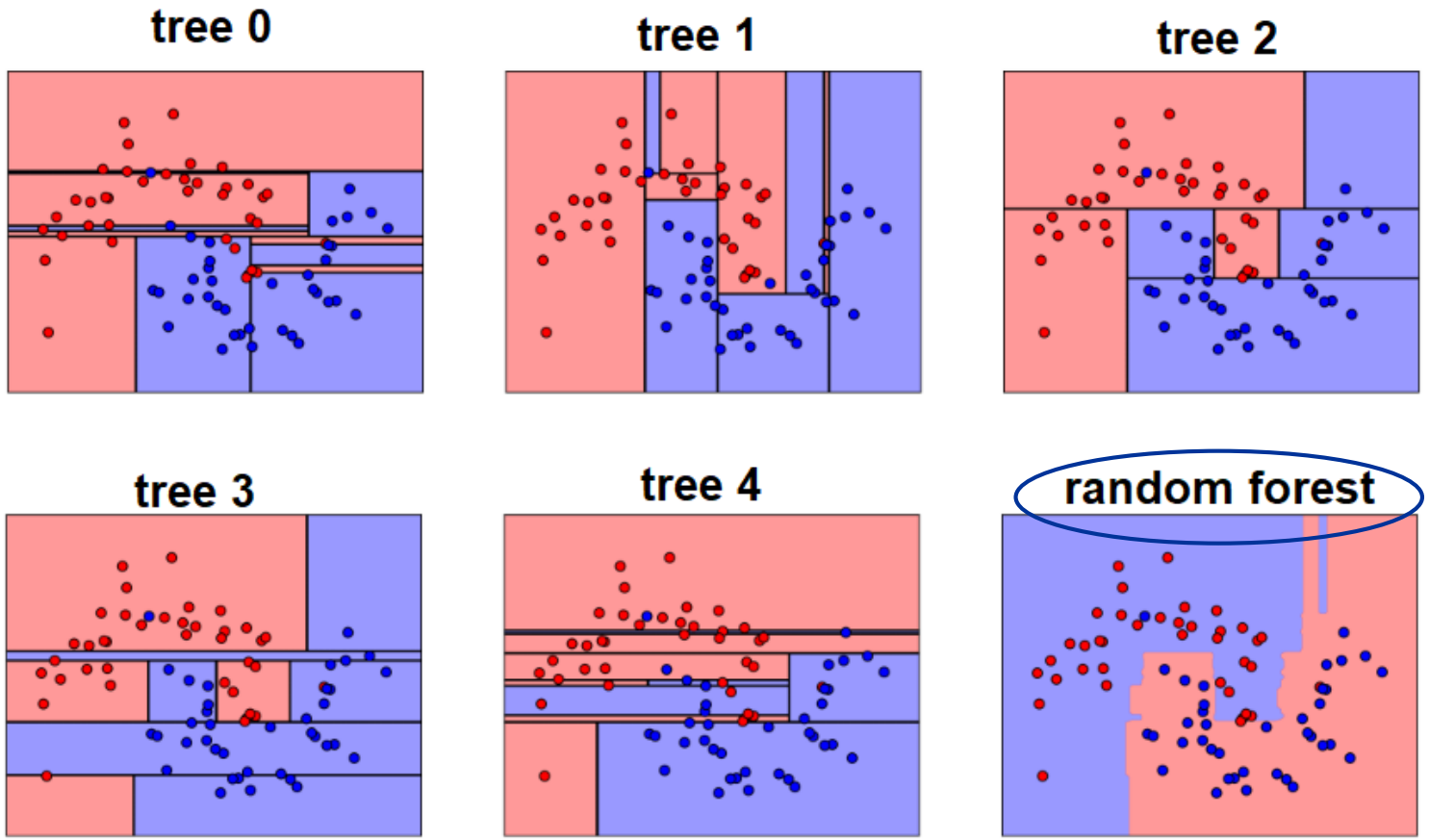

A figura à esquerda apresenta todos os elementos de textura ordenados por importância no mesmo estudo.

De acordo com a ordenação, o elemento "média" foi o que mais contribuiu para a classificação, seguido pelos elementos "dissimilaridade", "contraste", e "variância"; sendo os elementos "correlação" e "segundo momento angular" os que menos contribuiram.

Comparando as polarizações VV e VH, a polarização cruzada (VH) contribuiu mais para a classificação do a polarização igual (VV).

Créditos das imagens: Zhou et al., 2018. DOI: 10.3390/s18020373

# CLASSIFICAÇÃO SUPERVISIONADA: ALGORITMO RANDOM FOREST




Fronteiras de decisão para 5 árvores de decisão aleatórias e fronteira de decisão resultante da junção de todos os modelos, sendo que a floresta aleatória sofre menos do efeito de sobre-ajustamento do que qualquer uma das árvores individuais.

Créditos das imagens: Machine Learning Algorithms : Ensemble methods, Bagging, Boosting and Random Forests | by Nadir Tariverdiyev | Medium

# CLASSIFICAÇÃO SUPERVISIONADA: ALGORITMO RANDOM FOREST

## VANTAGENS E DESVANTAGENS DO ALGORTIMO RF

- Menos propenso a sobre-ajustamento do que as árvores de decisão ou outros algoritmos;
- Efetua o cálculo da importância das variáveis, que é bastante útil para a redução da dimensionalidade dos dados;
- É eficiente mesmo que existam lacunas ou valores anómalos (*outliers*) nos dados;
- Não requer a normalização dos dados, uma vez que utiliza uma abordagem baseada em regras (rule-based);
- Tem a capacidade de lidar com grandes volumes de dados com elevada dimensionalidade.

- O modelo pode mudar consideravelmente apenas com pequenas alterações nos dados;
- Envolve um cálculo computacional superior relativamente a outros algoritmos dada a construção de um elevado número de árvores de decisão;
- Implica um treino mais prolongado do modelo à medida que o número de árvores aumenta;
- Adapta-se menos bem a problemas de regressão do que a problemas de classificação;
- É mais difícil de interpretar do que as árvores de decisão ou outros algoritmos.

# 04

# AVALIAÇÃO DA EXATIDÃO DE UMA CLASSIFICAÇÃO: MATRIZ DE CONFUSÃO E MÉTRICAS DE EXATIDÃO

d.gTerritório
Direção-Geral do Território
Ciências ULisboa

# AVALIAÇÃO DA EXATIDÃO DE UMA CLASSIFICAÇÃO

## MATRIZ DE CONFUSÃO

Depois de treinado o modelo é necessário avaliar a performance da classificação. Tal pode ser efetuado através de uma matriz de confusão, ou matriz de erro, que compara os resultados da predição com a verdade do terreno (dados de referência utilizados como amostras de validação). Esta matriz é quadrada, podendo apresentar duas ou mais classes. No caso de uma classificação multi-classe, a matriz terá tantas linhas e colunas quantas as classes existentes na predição.

**Predição correta: quando para uma dada amostra a classe atribuída pelo classificador é consistente com a classe real.**

**Predição incorreta: quando para uma dada amostra a classe atribuída pelo classificador não é consistente com a classe real.**

| REFERÊNCIA | PREDIÇÃO |
|---|---|
| A | A |
| A | A |
| B | C |
| A | B |
| A | A |
| C | C |
| D | C |
| D | B |
| ... | ... |
| B | A |
| D | D |

| | | PREDIÇÃO | | | | |
|---|---|---|---|---|---|---|
| | CLASSES | A | B | C | D | Total |
| REFERÊNCIA | A | **50** | 37 | 24 | 39 | 150 |
| | B | 10 | **480** | 5 | 3 | 498 |
| | C | 14 | 10 | **765** | 1 | 790 |
| | D | 0 | 2 | 9 | **101** | 112 |
| | Total | 74 | 529 | 803 | 144 | 1550 |

d.gTerritório Direção-Geral do Território

Ciências ULisboa

# AVALIAÇÃO DA EXATIDÃO DE UMA CLASSIFICAÇÃO

- Independentemente de apresentar duas ou mais classes, estas são listadas pela mesma ordem nas linhas (classe real) e nas colunas (predição do modelo), pelo que os elementos da amostra corretamente classificados se encontram na diagonal principal da matriz, correspondendo ao número de vezes em que a predição da classe corresponde à classe real.

- As matrizes de confusão permitem uma representação efetiva da exatidão da classificação, dado que a exatidão individual de cada classe é também representada através do erros de inclusão (**erros de comissão**) e dos erros de exclusão (**erros de omissão**).

| | |
|---|---|
| **ERRO DE OMISSÃO** | Ocorre quando uma amostra é excluída da classe à qual pertence. Na matriz à direita, das 150 amostras de referência para a classe A, 100 foram incorretamente atribuídas (exclusão) pelo classificador a outras classes. |
| **ERRO DE COMISSÃO** | Ocorre quando uma amostra é incluída numa classe à qual não pertence. Na matriz à direita, das 74 amostras atribuídas pelo classificador à classe A, na realidade apenas 50 estão corretamente classificadas (inclusão). |

| | | PREDIÇÃO | | | | |
|---|---|---|---|---|---|---|
| | CLASSES | A | B | C | D | Total |
| REFERÊNCIA | A | **50** | 37 | 24 | 39 | 150 |
| | B | 10 | **480** | 5 | 3 | 498 |
| | C | 14 | 10 | **765** | 1 | 790 |
| | D | 0 | 2 | 9 | **101** | 112 |
| | Total | 74 | 529 | 803 | 144 | 1550 |

# AVALIAÇÃO DA EXATIDÃO DE UMA CLASSIFICAÇÃO

## MÉTRICAS DE EXATIDÃO

### EXATIDÃO DO PRODUTOR OU REVOCAÇÃO (*Recall*)

Percentagem de elementos corretamente classificados pelo modelo relativamente ao número total real de elementos dessa classe.

Corresponde ao **complementar do erro de omissão** e é calculada pela razão entre o número de amostras corretamente atribuídas a uma dada classe e o número total de amostras de referência dessa mesma classe (soma de todos os elementos de uma linha).

### EXATIDÃO DO UTILIZADOR OU PRECISÃO (*Precision*)

Percentagem de elementos corretamente classificados pelo modelo relativamente ao número total de elementos classificados como pertencentes a essa classe.

Corresponde ao **complementar do erro de comissão** e é calculada pela razão entre o número de amostras corretamente atribuídas a uma dada classe e o número total de elementos classificados pelo modelo como pertencentes a essa classe (soma de todos os elementos de uma coluna).

### F1-SCORE

Contribuição relativa da precisão e da revocação.

Média harmónica (média pesada) dos valores da revocação e da precisão para cada classe.

$$\text{Revocação } i = \frac{n_{ii}}{\sum_{j=1}^{k} n_{ij}}$$

$$\text{Precisão j} = \frac{n_{jj}}{\sum_{i=1}^{k} n_{ij}}$$

$$\text{F1-score} = 2 * \frac{(Revocação * Precisão)}{(Revocação + Precisão)}$$

| | | j = colunas PREDIÇÃO | | | |
|---|---|---|---|---|---|
| | **CLASSES** | **1** | **2** | **k** | **Total linhas ($n_{i+}$)** |
| **i = linhas REFERÊNCIA** | **1** | $n_{11}$ | $n_{12}$ | $n_{1k}$ | $\mathbf{n_{1+}}$ |
| | **2** | $n_{21}$ | $n_{22}$ | $n_{11}$ | $\mathbf{n_{2+}}$ |
| | **k** | $n_{k1}$ | $n_{k2}$ | $n_{kk}$ | $\mathbf{n_{k+}}$ |
| | **Total colunas ($n_{+j}$)** | $\mathbf{n_{+1}}$ | $\mathbf{n_{+2}}$ | $\mathbf{n_{+k}}$ | **n** |

# AVALIAÇÃO DA EXATIDÃO DE UMA CLASSIFICAÇÃO

## MÉTRICAS DE EXATIDÃO

### EXATIDÃO GLOBAL (*Overall Accuracy*)

Percentagem de elementos corretamente classificados pelo modelo relativamente ao número total de elementos da amostra.

Soma dos elementos da diagonal principal da matriz a dividir pelo número total de elementos da amostra $n$.

### COEFICIENTE KAPPA (**Cohen's Kappa**)

Medida de concordância entre os resultados da predição e os dados de referência para os elementos da amostra, considerando-os como variáveis aleatórias independentes.

Resulta do cálculo da expressão à direita, podendo ser obtido um valor compreendido entre -1 e 1.

$$\text{Exatidão global} = \frac{\sum_{i=1}^{k} n_{ii}}{n}$$

$$\text{Kappa} = \frac{n \cdot \sum_{i=1}^{k} n_{ii} - \sum_{i=1}^{k} n_{i+} \cdot n_{+i}}{n^2 - \sum_{i=1}^{k} n_{i+} \cdot n_{+i}}$$

| Concordância | Kappa |
|---|---|
| Excelente | > 0.81 |
| Boa | 0.80 – 0.61 |
| Moderada | 0.60 – 0.41 |
| Fraca | 0.40 – 0.21 |
| Má | 0.20 – 0.0 |
| Muito má | < 0 |

Inglada *et al.*, 2017

| | | j = colunas PREDIÇÃO | | | |
|---|---|---|---|---|---|
| | CLASSES | 1 | 2 | k | Total linhas ($n_{i+}$) |
| i = linhas REFERÊNCIA | 1 | $n_{11}$ | $n_{12}$ | $n_{1k}$ | $n_{1+}$ |
| | 2 | $n_{21}$ | $n_{22}$ | $n_{11}$ | $n_{2+}$ |
| | k | $n_{k1}$ | $n_{k2}$ | $n_{kk}$ | $n_{k+}$ |
| | Total colunas ($n_{+j}$) | $n_{+1}$ | $n_{+2}$ | $n_{+k}$ | $n$ |

d.gTerritório Direção-Geral do Território

Ciências ULisboa

# AVALIAÇÃO DA EXATIDÃO DE UMA CLASSIFICAÇÃO

## MÉTRICAS DE EXATIDÃO

### MACRO F1-SCORE

Corresponde à média dos valores de F1-score para cada classe, pelo que atribui o mesmo peso a todas as classes. Havendo classes minoritárias mal classificadas, estas vão provocar uma descida no valor da métrica, dado que estas têm a mesma importância que as classes maioritárias.

### WEIGHTED F1-SCORE

Corresponde à soma dos valores de F1-score para cada classe multiplicados pela frequência dessa classe, pelo que é atribuído um maior peso às classes maioritárias.

**Exatidão global= 0.90**

**Coeficiente Kappa= 0.84**

**Macro F1-score= 0.78**

**Weighted F1-score= 0.89**

| | | PREDIÇÃO | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | CLASSES | A | B | C | D | Total | REVOCAÇÃO | F1-SCORE | Frequência | Frequência* F1-score |
| REFERÊNCIA | A | **50** | 37 | 24 | 39 | 150 | 0.33 | 0.45 | 0.10 | 0.04 |
| | B | 10 | **480** | 5 | 3 | 498 | 0.96 | 0.93 | 0.32 | 0.30 |
| | C | 14 | 10 | **765** | 1 | 790 | 0.97 | 0.96 | 0.51 | 0.49 |
| | D | 0 | 2 | 9 | **101** | 112 | 0.90 | 0.79 | 0.07 | 0.06 |
| | Total | 74 | 529 | 803 | 144 | 1550 | | | | |
| | PRECISÃO | 0.68 | 0.91 | 0.95 | 0.70 | | | | | |

d.gTerritório Direção-Geral do Território

Ciências ULisboa

# BIBLIOGRAFIA


- https://www.geeksforgeeks.org/getting-started-machine-learning/
- https://towardsdatascience.com/machine-learning-basics-part-1-a36d38c7916
- https://www.javatpoint.com/machine-learning
- Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for Multi-Class Classification: an Overview. Retrieved from http://arxiv.org/abs/2008.05756
- https://www.mygreatlearning.com/blog/random-forest-algorithm/