MACHINE LEARNING

Com conteúdos adaptados da tese PhD “Desenvolvimento de Modelos Analíticos de Apoio à Gestão de Instituições do Ensino Superior, com Recurso a Data Mining”, de Maria Martins, 2020

Inteligência Artificial

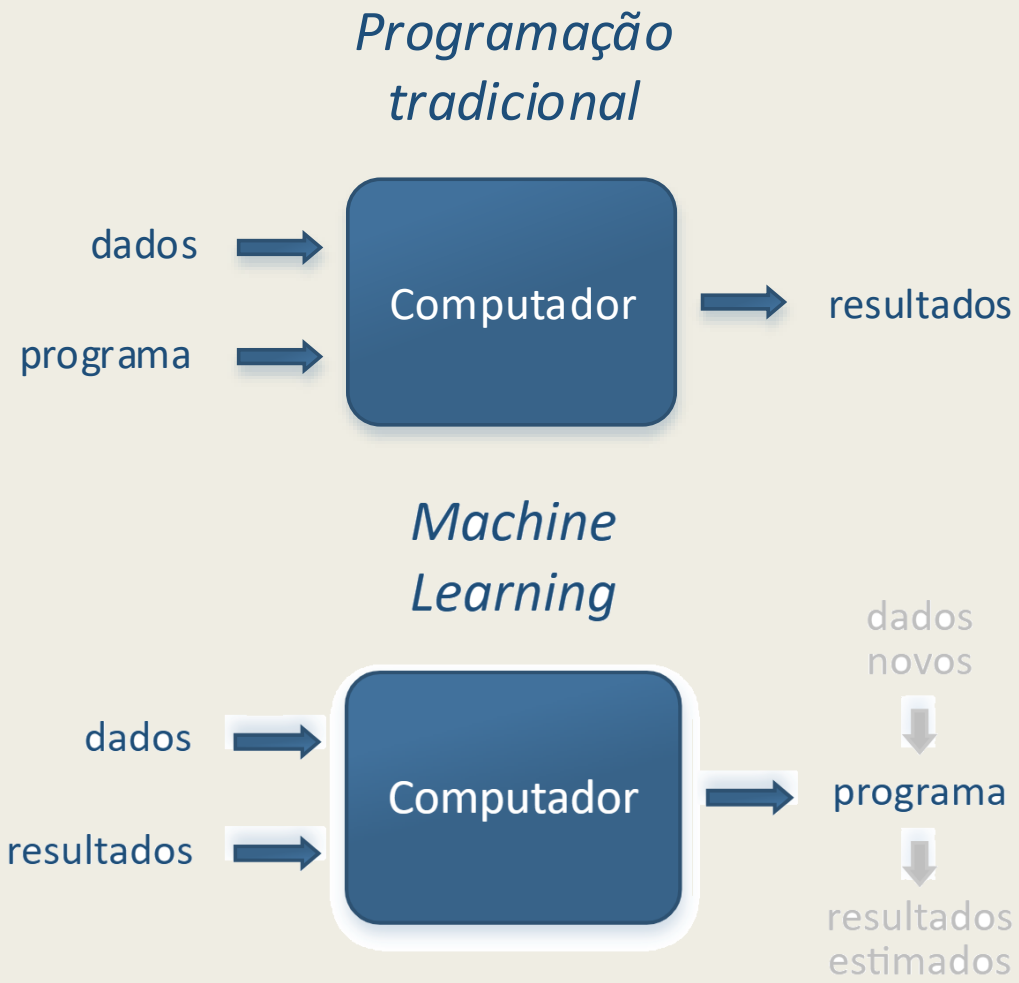
- Numa tentativa de lidar com problemas cada vez mais complexos, já há muito que se tenta incorporar nos programas de computadores e noutros sistemas informáticos capacidades de **compreensão** e **aprendizagem**, aptidões que normalmente estão associadas ao ser humano
  - *Continua a ser esse, portanto, o objetivo central da Inteligência Artificial (IA)*
- Por sua vez, a aprendizagem automática (Machine Learning – ML) sempre se apresentou como um dos principais conceitos e desafios da IA
  - *mas tem sido nos tempos mais atuais que se tem imposto como uma importante área, de interesse crescente, quer ao nível da investigação, quer no domínio da sua aplicabilidade*
- O âmbito da presente Unidade Curricular tem como foco principal, precisamente, essa importante subárea da IA
  - *a Machine Learning*

Machine Learning

- A ML tem como objetivo central dar aos computadores a capacidade de aprender sem serem explicitamente programados
  - *Na ML tentam-se desenvolver algoritmos com a capacidade de se ajustarem automaticamente (melhorando o seu desempenho) a partir de conjuntos de dados que lhes sejam fornecidos*
  - *A aprendizagem é conseguida através de modelos com parâmetros de afinação que são ajustados automaticamente de acordo com diferentes critérios de desempenho*
- A ML pode então ser entendida como sendo o conjunto de algoritmos e técnicas usadas no desenvolvimento de sistemas que têm a capacidade de **aprender com os dados**,
  - *conseguindo, de forma automática, realizar previsões ou reconhecer padrões em conjuntos de dados que lhes sejam fornecidos*

ML vs Programação tradicional

- A ML altera o paradigma de Programação a que estávamos habituados



Tipos de ML

- A ML, por sua vez, é, normalmente, divida em 3 grandes subáreas:
  1. ***Aprendizagem supervisionada****: os algoritmos aprendem a partir de um conjunto de exemplos que incluem a respetiva resposta (também designados exemplos rotulados), para depois generalizar o seu comportamento a todas as possíveis entradas – serão essencialmente estas técnicas o alvo do nosso estudo.*
  2. ***Aprendizagem não supervisionada****: os algoritmos aprendem a partir de exemplos não rotulados, tentando classificar esses mesmos exemplos de acordo com critérios de similaridade – estudaremos só um exemplo destas técnicas.*
  3. ***Aprendizagem por reforço****: os algoritmos aprendem a partir do feedback que recebem sobre a qualidade das soluções que vão produzindo (o desempenho vai sendo melhorado explorando iterativamente o espaço de soluções) – técnica não abordada no nosso estudo.*
- As técnicas ML podem ser usadas para extrair conhecimento em grandes bases de dados, integradas num processo mais abrangente, conhecido por KDD (Knowledge Discovery in Databases) e que envolve várias fases de processamento.

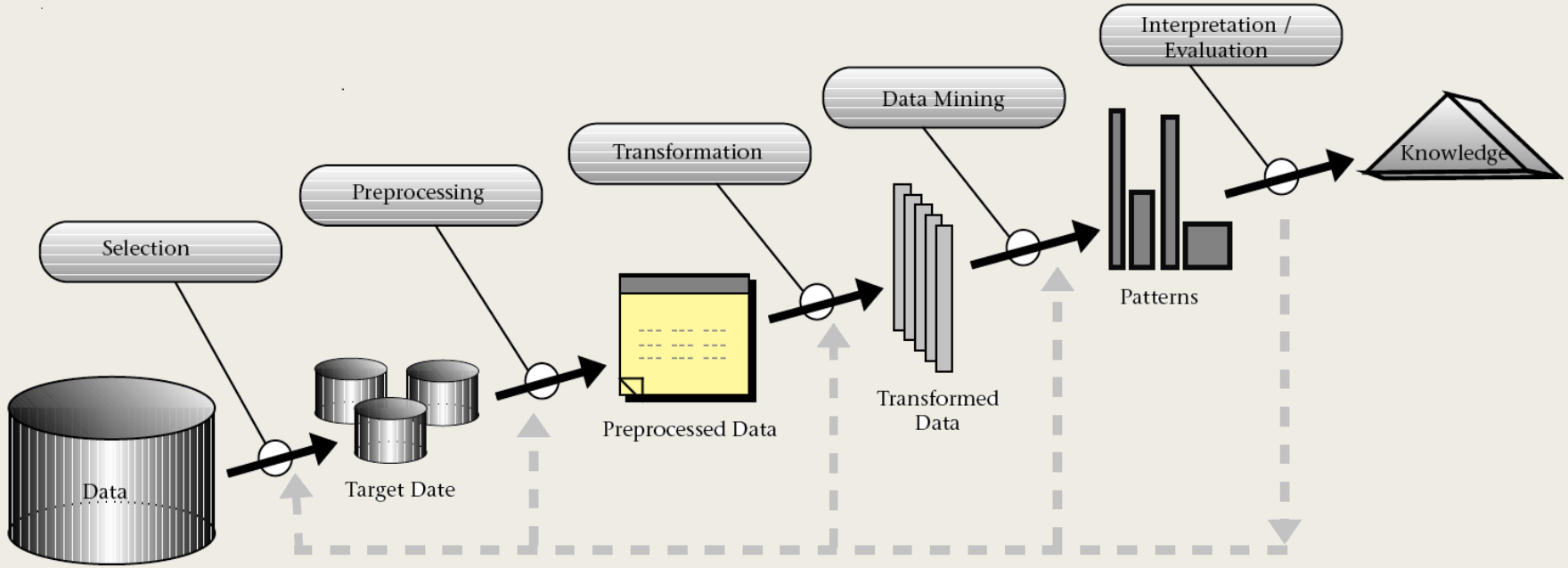
O processo KDD

- A constante evolução dos sistemas informáticos propiciou a que as organizações procedessem à recolha, processamento, e depósito de grandes volumes de informação, oriunda das suas atividades diárias
- **Essa grande e diversificada volumetria de dados** gerados, procedentes de múltiplas fontes e disponíveis em configurações heterogéneas, **torna quase impossível a sua análise**
  - *será ínfima a proporção de dados que são analisados, em comparação com os que são gerados*
  - *muita informação, potencialmente útil, presente nesses grandes volumes de dados, não é devidamente aproveitada.*
- Foi devido à necessidade premente de conceber e desenvolver **uma metodologia que permitisse a extração de conhecimento útil a partir de grandes volumes de dados** que emergiu a KDD (*Descoberta de Conhecimento em Base de Dados*)
- Segundo Fayyad et al.[1], considerados os principais percursores desta área de investigação,
  - *"A KDD é o processo iterativo e interativo não trivial, de identificação válida, original e potencialmente útil de padrões compreensíveis nos dados".*
  - *"Envolvendo um grande relacionamento interdisciplinar, o objetivo unificador da KDD é extrair conhecimento de dados de alto nível a partir de dados de baixo nível no contexto de grandes conjuntos de dados".*

[1] *Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. AI magazine, 17(3):37, 1996*

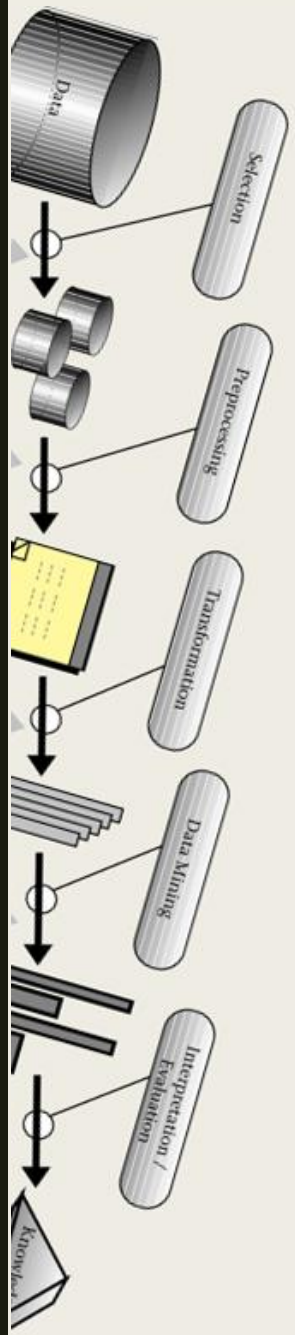
O ciclo de vida da KDD

- O ciclo de vida que caracteriza a KDD compreende uma sequência de fases e tarefas associadas entre si
  - *ainda que apresentem uma ordem tendencialmente sequencial, pode-se voltar atrás para qualquer uma das etapas, dependendo da qualidade dos resultados que se forem obtendo*



(Fayyad et al.)

Fases da KDD



- Seleção

  esta fase compreende a seleção dos dados corretos a analisar, os atributos mais relevantes e o período de tempo mais apropriado

- Pré-processamento

  preparar adequadamente os dados, removendo o ruído e os outliers (valores atípicos) presentes nos dados e decidindo o que fazer com valores em falta

- Transformação

  após a seleção e pré-processamento dos dados, convertem-se os dados para o formato que melhor se adeque à posterior aplicação dos algoritmos de data mining, aplicando-lhes transformações de normalização, agregação ou de discretização, com vista à redução do número efetivo de variáveis ou valores, à procura de representações invariantes para os dados ou de outras características úteis para o conhecimento

- Data Mining (Prospeção de Dados)

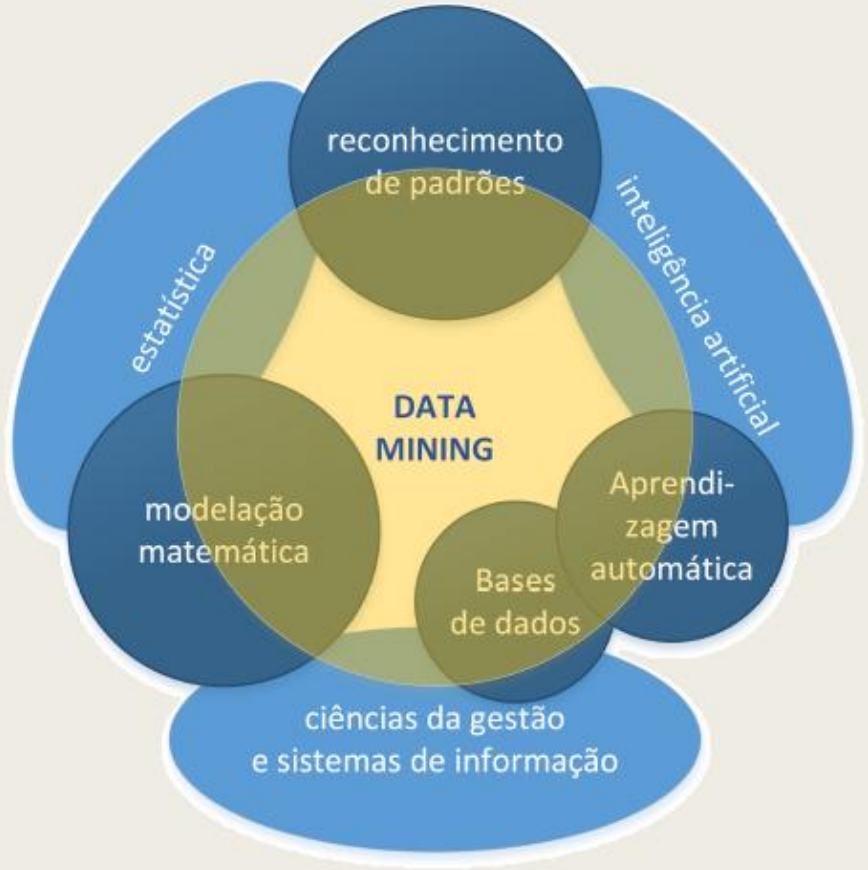
  aplicação de poderosos métodos e algoritmos de análise automática (normalmente algoritmos de Machine Learning) que sirvam o propósito da prospeção — a descoberta de padrões e relacionamentos que existam nos dados

- Interpretação/Avaliação

  interpretação dos padrões descobertos e a avaliação da sua utilidade para a aplicação pretendida — é nesta derradeira fase que se consolida, ou não, o conhecimento descoberto e que se avalia a necessidade de novas iterações, regressando-se, se necessário, a alguma das fases anteriores, para novos aperfeiçoamentos

A multidisciplinaridade do Data Mining

- É através do processo KDD que ocorre a transformação de dados brutos em conhecimento útil e compreensível,
  - *esta informação, "escondida" em bases de dados de grande volumetria, é então recuperada através de algoritmos computacionais inerentes à fase do Data Mining (DM)*
- O *DM* é a fase mais importante de todo o processo de KDD
  - ***Tal como o próprio KDD, também o DM é uma área interdisciplinar***



(adaptada de Turban et al.[1])

[1] *Efraim Turban, Ramesh Sharda, and Dursun Delen. Decision support and business intelli-gence systems. Pearson Education India, ninth edition, 2011*

Data Mining vs Machine Learning

- Ainda que se trate de uma metodologia que envolva conhecimento interdisciplinar, **o Data Mining está intimamente ligado com a Machine Learning**:
  - *"O DM é uma etapa no processo global de descoberta de conhecimento, em que se extrai, através da aplicação de algoritmos de **aprendizagem automática** um conjunto de padrões e relacionamentos dos dados, que permitem revelar, de forma automática ou semiautomática, informação implícita que esteja presente em grandes base de dados", Fayyad et al.*
  - *O DM é um campo interdisciplinar que junta técnicas de **aprendizagem automática**, reconhecimento de padrões, estatística, bases de dados e visualização, no intuito de conseguir extrair informação útil de grandes bases de dados", Cabena et al.*[1]
  - *"O DM é uma poderosa ferramenta de **inteligência artificial**, com capacidade para descobrir informações úteis, através da análise de dados de muitas proveniências ou dimensões, categorizar essas informações e resumir as relações identificadas em bases de dados", Algarni.*[2]

[1] Peter Cabena, Pablo Hadjinian, Rolf Stadler, Jaap Verhees, and Alessandro Zanasi. Discovering data mining: from concept to implementation. Prentice-Hall, Inc., 1998

[2] Abdulmohsen Algarni. Data mining in education. International Journal of Advanced Computer Science and Applications, 7:456–461, 2016

Modelos de Data Mining vs Machine Learning

- No fundo, o *data mining* tem como objetivo a análise exploratória dos dados, criando modelos analíticos ou computacionais que permitam extrair padrões e informações úteis a partir de grandes conjuntos de dados
  - *Já a ML, enquadrando-se na DM, põe mais o seu foco na construção de modelos de previsão*
- Dependendo do domínio do problema que representem, os modelos podem então ser categorizados em dois grandes grupos:
  - *modelos **descritivos***
    - o objetivo principal é encontrar padrões frequentes, que possam explicar, ou generalizar, a estrutura intrínseca dos dados, incluindo os seus relacionamentos.

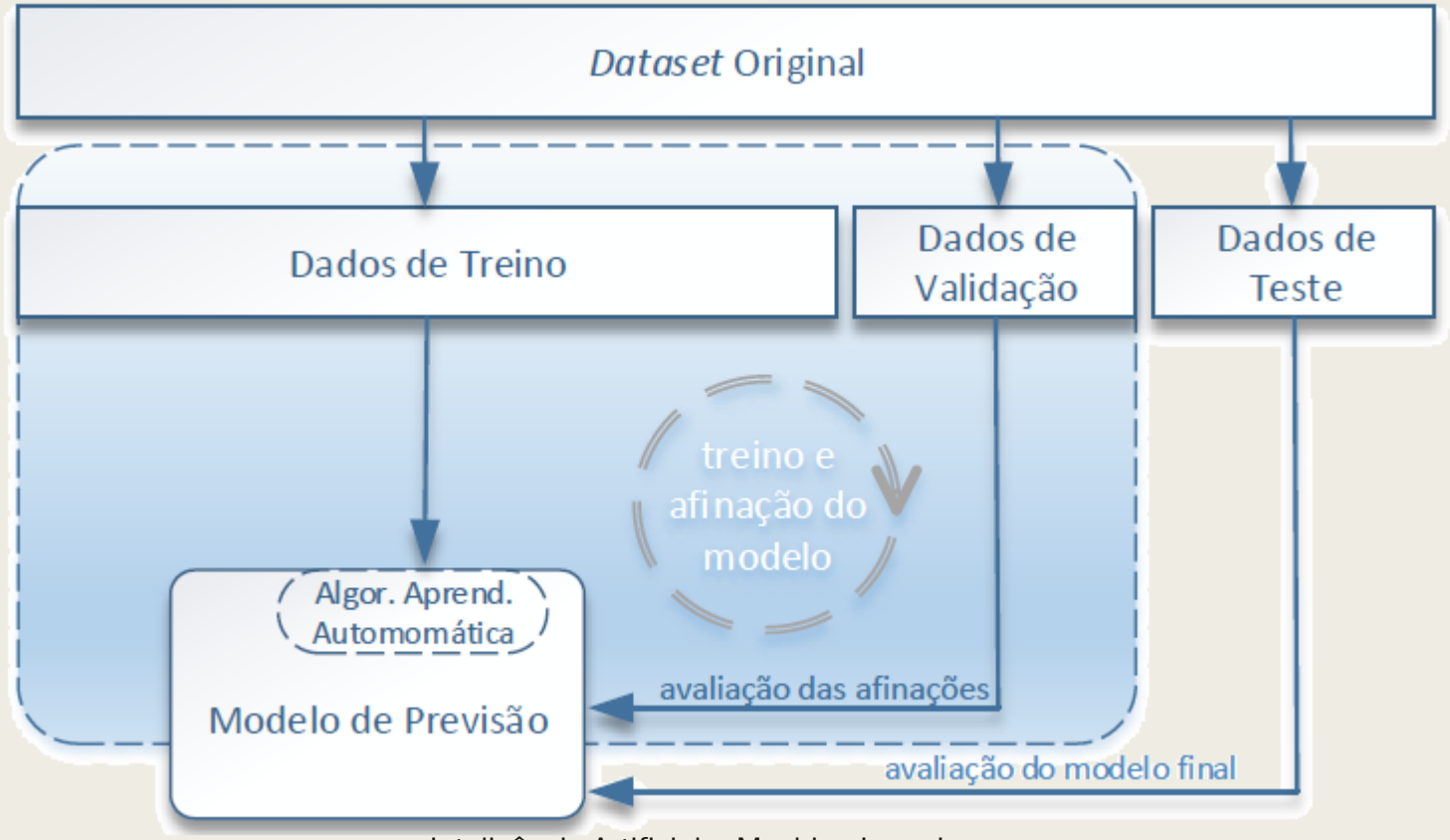
      (os padrões podem ser, por exemplo, presença de anomalias, tendências, agrupamentos entre objetos, associações e correlações entre variáveis)
  - *modelos **preditivos** (principal missão da ML)*
    - o objetivo principal é estimar valores, desconhecidos ou futuros, de uma ou mais variáveis alvo de interesse, a partir de alguma combinação de outras características presentes nos dados.

      (permitem, portanto, antever circunstâncias futuras, tais como certas tendências e certos comportamentos)
      - *a variável objeto da previsão é designada variável alvo, variável resposta ou variável dependente (VD),*
      - *os atributos usados para a previsão são designados caraterísticas (features), variáveis independentes, preditivas ou explicativas.*

Modelo preditivo

- Os modelos preditivos também poderão facultar alguma descrição do problema e os modelos descritivos revelar previsões de eventos futuros
  - *A distinção entre uns e o outros é muito ténue*

    (tendencialmente, os modelos cuja representação não seja facilmente interpretável serão usados para previsão; por outro lado, os modelos preditivos cuja representação surja em forma de regras, ou nalguma estrutura interpretável podem também ser usados para descrever os dados)
- Segue-se um esquema ilustrativo de todo o processo de indução de um modelo preditivo



Indução do modelo preditivo

- Para o modelo ser induzido,
  - *o dataset inicial é normalmente dividido em três subconjuntos: de **treino**, de **validação** e de **teste**;*
  - *o conjunto de treino é processado por um algoritmo de Machine Learning, servindo o segundo subconjunto para ajustar os hiperparâmetros responsáveis pela afinação (configuração) do algoritmo, num processo normalmente iterativo (e interativo) em que o modelo vai sendo avaliado com os dados de validação;*
  - *por fim, concluído que estão o treino e a afinação do modelo, os dados do subconjunto de teste são mostrados pela primeira vez (e última) ao modelo encontrado, de forma a avaliar a sua verdadeira capacidade preditiva e em especial a sua capacidade de generalização*
- Se o modelo evidenciar elevada precisão no conjunto de dados deixado para teste, assume-se que o modelo tem efetivamente boa capacidade de **generalização**, ou seja, será expectável que apresente bom desempenho com dados futuros e desconhecidos
  - *por outro lado, se o modelo apresentar um fraco desempenho com os dados do conjunto de teste, então o modelo criado não será adequado para efetuar a previsão*

    (perante esta segunda situação, é usual retornar-se à fase de pré-processamento para aprimorar os dados, ou, simplesmente, recorrer-se a outra técnica de aprendizagem)

Entendendo melhor os tipos de aprendizagem

- Como já referido anteriormente, dependendo do tipo de informação que é posta à disposição do modelo, o mesmo terá por base um algoritmo de Aprendizagem Supervisionada ou Não Supervisionada
  - *Na aprendizagem **supervisionada** o modelo é treinado com dados que incluem a variável resposta, tentando perceber (aprender) a relação entre essa variável e as restantes variáveis (variáveis preditivas)*
  - *Na aprendizagem **não supervisionada**, as instâncias do conjunto de dados só estão caracterizadas por atributos de entrada (variáveis preditivas) e não existe informação sobre o valor da variável resposta associada a cada exemplo*
    - a aprendizagem do modelo é efetuada descobrindo similaridades nos dados, formando-se agrupamentos de dados com características semelhantes
- Em qualquer tipo de aprendizagem, seja ou não supervisionada, um dos objetivos principais é a criação de modelos com **capacidade de generalização**
  - *trata-se de um conceito que traduz a capacidade de um modelo prever com precisão novos exemplos, ainda não observados, após ter sido construído unicamente com base num conjunto de dados de aprendizagem*
  - *considera-se haver **overfitting** (sobreajuste) quando o modelo se ajusta (se vicia) demasiado aos dados de treino, comprometendo, dessa forma, a sua capacidade de generalização*
    - e ***underfitting*** quando nem com os dados de treino o modelo consegue um bom desempenho

Classificação vs Regressão

- Os métodos de previsão podem ser classificados, essencialmente, em duas tipologias diferentes: os de classificação e os de regressão.
  - *A principal diferença entre eles reside na tipologia da variável alvo da previsão:*
    - na classificação a varável alvo é **categórica**,
    - na regressão é **numérica**, assumindo valores contínuos.
- Métodos de **Regressão**
  - *Os modelos têm como objetivo prever os valores futuros, ou desconhecidos, de uma ou mais variáveis numéricas contínuas, a partir de outros atributos presentes no conjunto de dados.*
- Métodos de **Classificação**
  - *Classificar é a ação de categorizar um determinado objeto/instância de acordo com a sua tipologia*
    - pretende-se que o modelo gerado atribua a cada um dos objetos/instâncias uma das categorias (rótulos) predefinidas, em função das suas variáveis explicativas

Métricas de avaliação

Qualquer que seja o modelo preditivo, é sempre necessário avaliar o quão preciso é o seu desempenho

Algumas das métricas de avaliação de desempenho preditivo mais usadas

- *em problemas de **regressão:***
  - Erro quadrático médio (MSE) ou raiz desse valor (RMSE)
  - Coeficiente de determinação (R^2)
- *em problemas de **classificação:***
  - Matriz de Confusão
  - Um conjunto diverso de medidas que se obtêm a partir dos valores que compõem a Matriz de Confusão
    - *Taxas de verdadeiros positivos, de verdadeiros negativos, de falsos positivos, de falsos negativos, precisão…*
  - Curva ROC e valor AUC

Métricas para regressão

- R^2 (coeficiente de determinação)

  mede a correlação entre os valores observados (reais) e os valores preditos (estimados) – varia entre 0 e 1

- MSE (*mean square error* — erro quadrático médio)

  mede o desvio das previsões em relação ao valor efetivo, calculando a média dos quadrados das distâncias entre os valores y_i conhecidos e o valores $\hat{f}(x_i)$ preditos pelo modelo:

$$\text{MSE}(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

- RMSE (*root mean square error*)

  raiz quadrada do MSE (medida de erro mais usada que a anterior, uma vez que dá uma indicação mais direta do verdadeiro desvio)

Matriz de Confusão – para avaliar classificadores

Quando representando um classificador binário, trata-se de uma tabela que ilustra o número de previsões corretas e incorretas em cada classe (categoria), positiva (+) ou negativa (-)

- Para um determinado conjunto de dados, as linhas da matriz representam as classes (categorias) verdadeiras, e as colunas representam as classes estimadas pelo classificador

		classe predita	
		+	-
classe verdadeira	+	VP	FN
	-	FP	VN

- Verdadeiros Positivos (VP): representa o número de previsões positivas que estão corretas;
- Falsos Positivos (FP): representa o número de previsões positivas que estão incorretas;
- Falsos Negativos (FN): é o número de previsões negativas que estão incorretas;
- Verdadeiros Negativos (VN): é o número de previsões negativas que estão corretas.

- Logo, cada elemento m_{ij} de uma matriz de confusão apresenta o número de exemplos da classe i classificados como pertencentes à classe j.
- Para k classes a matriz de confusão teria a dimensão k × k.
- A diagonal apresenta os casos corretamente classificados pelo modelo, enquanto os outros elementos correspondem aos erros cometidos nas suas previsões

Outras métricas para classificação

A partir dos valores da Matriz de Confusão retiram-se várias outras medidas

- Taxa de Falsos Negativos (TFN)

  representa a taxa de erro na classe positiva. É uma medida da proporção de exemplos da classe positiva incorretamente classificados pelo preditor

$$\text{TFN} = \frac{\text{FN}}{\text{VP} + \text{FN}}$$

- Taxa de Falsos Positivos (TFP)

  representa a taxa de erro na classe negativa. É uma medida da proporção de exemplos da classe negativa incorretamente classificados pelo preditor

$$\text{TFP} = \frac{\text{FP}}{\text{FP} + \text{VN}}$$

- Taxa de erro

  representa a percentagem de classificações incorretas do total de exemplos n, independentemente da direção do erro

$$\text{erro} = \frac{\text{FP} + \text{FN}}{n}$$

Outras métricas para classificação

- Taxa de acerto ou acurácia

  representa a percentagem de classificações corretas do total de exemplos n, independentemente da direção do acerto. É calculada pela soma dos valores da diagonal principal da matriz, dividida pela soma dos valores de todos os elementos da matriz (n)

$$\text{acurácia} = \frac{VP + VN}{n}$$

- **Precisão**

  representa a taxa de acerto entre os exemplos classificados pelo preditor como positivos

$$\text{precisão} = \frac{VP}{VP + FP}$$

- Taxa de Verdadeiros Positivos (TVP) ou **Sensibilidade** (Recall)

  representa a proporção dos exemplos positivos que foram classificados corretamente pelo preditor

$$\text{sensibilidade} = \text{recall} = TVP = \frac{VP}{VP + FN}$$

Outras métricas para classificação

- Taxa de Verdadeiros Negativos (TVN) ou Especificidade

  corresponde à taxa de acerto na classe negativa, ou seja, a proporção dos exemplos negativos que foram classificados corretamente pelo preditor. O complementar corresponde à taxa TFP

$$\text{especificidade} = 1 - \text{TFP} = \frac{\text{VN}}{\text{VN} + \text{FP}}$$

- F-medida

  é a média harmónica ponderada da precisão e sensibilidade. Trata-se de uma medida única que valoriza os erros cometidos em qualquer dos sentidos (FP e FN)

$$\text{Fmedida} = \frac{(w+1) \times \text{precisao} \times \text{sensibilidade}}{w \times \text{precisao} + \text{sensibilidade}}$$

- F1

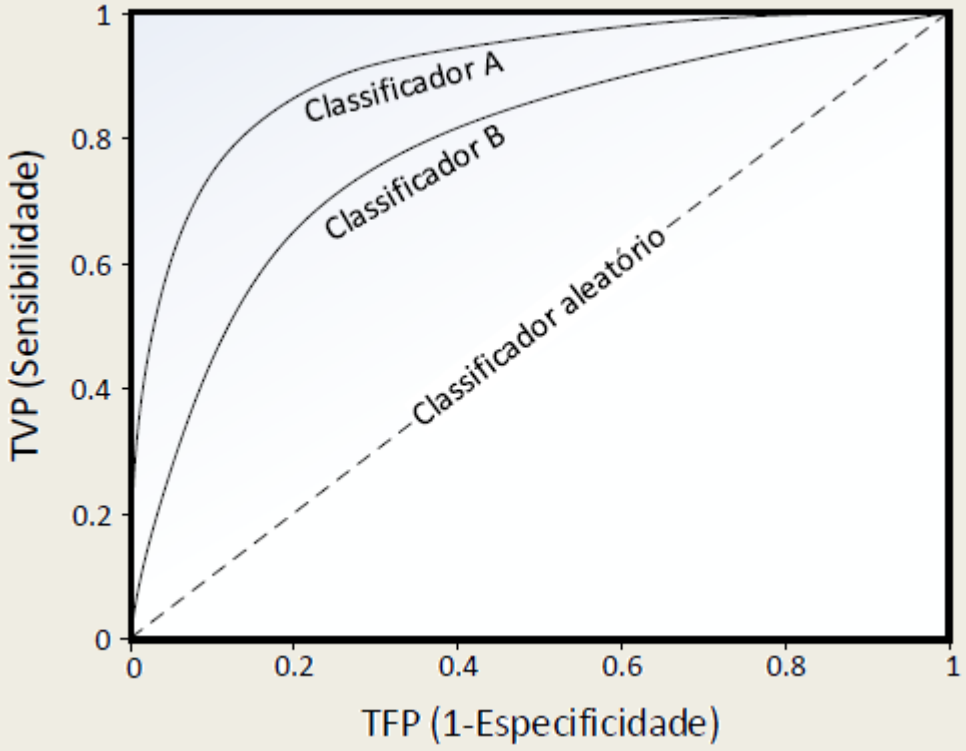
  Caso os erros sejam valorados de igual modo, a média harmónica F-medida deixa de ser ponderada (w = 1), passando a assumir a forma a que normalmente se designa F1

$$\text{F1} = \frac{2 \times \text{precisao} \times \text{sensibilidade}}{\text{precisao} + \text{sensibilidade}}$$

Curvas ROC

- Ainda uma outra forma de avaliação dos algoritmos é através das curvas ROC (*Receiver Operating Characteristic)*
- Uma curva ROC é um gráfico que ilustra o desempenho de um modelo de classificação binário através da variação do limiar de discriminação entre elementos positivos e negativos

- Considera-se que um classificador é melhor que um outro se a sua curva no espaço ROC se posiciona acima e à esquerda da curva correspondente ao segundo classificador
- Quando se comparam duas ou mais curvas, se estas não se intercetarem, aquela que mais se aproxima do ponto (0,1) corresponde ao melhor desempenho.
    - *No caso de ocorrerem interseções, cada algoritmo tem uma região com melhor desempenho.*



Valor AUC

- Também é usual mensurar o desempenho dum classificador em termos de uma medida única extraída da sua curva ROC:

  *a **área abaixo da curva ROC**, designada AUC (Area Under Curve).*

- A medida AUC produz valores entre 0 e 1.
  - *Valores mais próximos de 1 são considerados melhores, ou seja, quanto maior for a área sob a curva ROC, maior será a precisão do algoritmo.*

Concluindo…

- Cada uma das métricas apresentadas avalia de forma quantitativa (com exceção da curva ROC, que é qualitativa) um modelo de classificação, providenciando informação sobre a eficácia do método de aprendizagem.
  - *Nenhuma delas substitui por completo todas as outras, pelo que, é habitual usarem-se várias em simultâneo*

  Por exemplo, a taxa de erro e a acurácia são medidas simples que, perante problemas desbalanceados, não permitem evidenciar a diferença entre falsos positivos e falsos negativos, levando a resultados ilusórios ou enganadores.

  Nesse caso, quando o número de exemplos de cada classe é muito diferente, é recomendável usar métricas que enfatizem ambas as medida de erro, FP e FN, combinando-as, como por exemplo a F-medida e o próprio valor AUC.

- O valor **AUC** é uma das métricas mais apreciadas e usadas pelos analistas de *data mining*, em parte por lidar relativamente bem, precisamente, com *datasets* desbalanceados.