# Parte IV — Desenho Físico de *Data Warehouses*

1. **2019.05** Descreva a técnica de compressão de página e justifique se é apropriada para uma dimensão Cliente onde, entre outros, são guardados dados demográficos da região de cada cliente, sendo que existem poucas regiões distintas. Neste contexto, mostre o estado de uma página antes e depois de comprimida.

A técnica de compressão de página é a técnica de compressão mais utilizada em data warehouses. Esta técnica identifica prefixos e valores comuns em linhas na mesma página do disco e guarda esses prefixos e valores comuns numa página dicionário. Os valores originais são substituídos por referências ao dicionário. Para o caso da dimensão cliente descrita no texto é valida a utilização da técnica de compressão da página pois a pouca distinção entre regiões torna ainda mais eficiente a utilização deste método e a possibilidade da substituição por chaves.

Ex.

ID Cliente Região ID Cliente Região

1 João Algarve 1 João (1)

2 Joana Algarve 🡪 2 Joana (1) 🡪 Dicionário : (1) Algarve

3 Manuel Aveiro 3 Manuel Aveiro

1. **2019.05** Complete cada afirmação sobre desenho físico de *data warehouses*.
   * O não suporte de pesquisas por intervalo *(dice)* é uma desvantagem do índice de tipo…função de dispersão (hash)
   * Quando os registos da tabela base seguem a ordem das entradas num índice, este designa-se… agrupado (clustered)
   * Num índice de mapa de *bits* sobre um atributo existem tantas colunas quanto… os valores que o atributo pode tomar
   * Uma interrogação SQL com o seu resultado guardado em disco designa-se… vista materializada
   * Uma desvantagem do nível RAID 0 *(striping)* é… a perda de dados quando ocorrem falhas num disco
2. **2018.05** Indique afirmações verdadeiras (V) ou falsas (F) sobre desenho físico de *data warehouses.* \_\_ Duas partições de uma mesma tabela podem ter índices sobre atributos diferentes.

F\_\_ Os índices de árvore B+ e de função de dispersão são apropriados para pesquisas por intervalo.

F\_\_ A compressão de linha é mais eficaz quando há prefixos iguais nos atributos de um registo.

V\_\_ O resultado de uma interrogação pode ser totalmente obtido por via de uma vista materializada.

F\_\_ Num disco rígido, o atraso de rotação ocupa a maior fatia do tempo de pesquisa de dados.

1. **2018.05** Complete cada afirmação sobre desenho físico de *data warehouses*.
   * Os índices de função de dispersão são mais eficientes que as árvores B+ na operação de…slice (igualdade)
   * O primeiro atributo de um índice multi-atributo de uma tabela de factos deve ser… a chave estrangeira da dimensão data
   * Quando os registos da tabela base seguem a ordem das entradas no índice, este designa-se… agrupado (clustered)
   * Os índices de mapas de *bits* são apropriados para atributos com… poucos valores distintos (ex. estado civil: casado ou solteiro)
   * Um índice oferece tempos de pesquisa de dados mais curtos à custa de… mais espaço utilizado em disco
2. **2017.06** Indique afirmações verdadeiras (V) ou falsas (F) sobre partições de dados e níveis RAID.

V\_\_ Uma tabela muito grande pode ser decomposta em partições, cada uma no seu próprio disco.

V\_\_ Duas estratégias de partição de dados são por intervalo e por lista de valores da *partition key*.

F\_\_ Na arquitetura *shared memory*, os nós do servidor paralelo não partilham o mesmo disco.

F\_\_ O nível RAID 0 oferece menor desempenho e maior confiança face aos outros níveis.

V\_\_ Independentemente dos níveis RAID, o utilizador vê apenas um disco lógico.

1. **2017.06** Indique afirmações verdadeiras (V) ou falsas (F) sobre o desenho físico de *data warehouses*.

F\_\_ Deve ser realizado antes da modelação dimensional.

V\_\_ Procura minimizar a utilização de recursos do sistema informático.

F?\_\_ Duas das técnicas são as vistas materializadas e as bifurcações de tabelas.

F\_\_ Alterações ao desenho físico obrigam a alterações nas aplicações dos decisores.

V\_\_ A sua importância aumenta com a quantidade de dados que precisam de ser processados.

1. **2017.05** Indique afirmações verdadeiras (V) ou falsas (F) sobre compressão de dados.

V?\_\_ A compressão de página costuma poupar mais espaço em disco que a de registo/linha.

V\_\_ As aplicações dos utilizadores não precisam de saber se os dados estão ou não comprimidos.

V\_\_ O dicionário de página guarda prefixos e valores comuns a várias páginas da mesma tabela.

V\_\_ A compressão é especialmente indicada quando os dados têm muitos valores repetidos.

F\_\_ Requer a descompressão dos dados lidos do disco, tornando mais lentos os relatórios.

1. **2017.05** Indique afirmações verdadeiras (V) ou falsas (F) sobre vistas materializadas.

V\_\_ Cada vista tem um SELECT associado, cujo resultado fica guardado em disco.

V\_\_ Evitam o cálculo repetitivo de operações muito requisitadas, tais como somas e médias.

F\_\_ As aplicações dos utilizadores precisam de saber os nomes das vistas para as poderem usar.

F\_\_ No comando CREATE MATERIALIZED VIEW, o SELECT não pode incluir GROUP BY.

V\_\_ Quando há atualização de dados, as vistas também são automaticamente atualizadas.

1. **2016.05** Indique uma razão para a taxa de compressão de dados ser habitualmente elevada nos *data warehouses*, e ilustre o uso da técnica de compressão de página, mostrando o estado de uma página antes e depois de comprimida, no contexto de uma dimensão onde a poupança de espaço seja relevante.

A taxa de compressão dos dados é geralmente elevada devido à frequente repetição de atributos que existem num data warehouse.

Ex

Table

Description automatically generated with medium confidence

1. **2016.05** Indique afirmações verdadeiras (V) ou falsas (F) sobre índices.

F\_\_ Num índice multi-atributo a ordem dos atributos a indexar (ex. X,Y ou Y,X) é irrelevante.

V\_\_ Um índice *bitmap* tem tantas colunas quantos os valores distintos do atributo indexado.

V\_\_ Num índice agrupado os registos de dados na tabela seguem a ordem das entradas no índice.

V\_\_ Os índices de dispersão são mais eficientes que os de árvore B+ em pesquisas por igualdade.

F\_\_ Num índice de dispersão cada *bucket* guarda entradas com valores indexados consecutivos.

1. **2014.06** Considere que qualquer dos atributos, A, B, e C, de uma dimensão é frequentemente usado individualmente para aplicar filtros com intervalos de valores em relatórios (por exemplo, A BETWEEN 10 AND 20). Explique se faz mais sentido criar um índice composto que abranja os três atributos, ou três índices separados, um para cada atributo, e justifique qual o tipo de índice que escolheria.

Neste caso, o que faz mais sentido é criar um índice composto que abranja os três atributos, isto porque os índices estão sujeitos a dados redundantes porque permitem o acesso eficiente aos dados relevantes de cada índice que abrange um ou mais atributos de uma tabela base. Como os atributos da dimensão são usados frequentemente com filtros de intervalo, o índice mais apropriado seria o de Árvore B+.

1. **2014.05** Apresente dois motivos para os relatórios analíticos abrangerem grandes volumes de dados. Nestas circunstâncias, identifique o problema que o desenho físico de *data warehouses* procura resolver e descreva como, em traços gerais.

Os relatórios analíticos abrangem grandes volumes de dados porque existe a agregação de muitos valores numéricos e a inclusão de dados históricos. Assim, o desenho físico de data warehouses procura minimizar a utilização de recursos, com menos dados pesquisados por interrogações SQL, menos espaço ocupado em disco aquando de atualizações periódicas e menos tempo para gerar relatórios dinâmicos. A minimização da utilização de recursos passa geralmente pela utilização de índices, de forma a facilitar as pesquisas, e pela utilização de técnicas de compressão, como a compressão por pagina, de forma a minimizar o espaço utilizado em disco.

1. **2014.05** Indique afirmações verdadeiras (V) ou falsas (F) sobre discos rígidos.

F\_\_ O disco rígido é a componente que menos limita o desempenho de um *data warehouse*.

V\_\_ O tempo de posicionamento aumenta com a distância da cabeça de leitura à pista dos dados.

V\_\_ A transferência de dados em posições seguidas é mais rápida do que se estiverem dispersas.

F\_\_ O atraso de rotação é máximo quando os dados estão diametralmente opostos à cabeça.

V\_\_ Numa *solid state drive* (SSD) o tempo de acesso não depende da localização dos dados.

1. **2013.06** Descreva a estrutura e conteúdo de um índice *bitmap* quando aplicado a um atributo com elevada cardinalidade, como os países do mundo, em termos do número de colunas, número de linhas, e valores típicos, e justifique se este cenário é o mais apropriado a este tipo de índice.

No caso descrito, o índice de bitmap ficaria com um elevado número de colunas, de linhas e de valores com dados iguais e redundantes. A solução de um índice bitmap não foi a ideal, uma vez que que os índices bitmap são mais apropriados para atributos com baixa cardinalidade.

1. **2013.05** Descreva o propósito do desenho físico de um *data warehouse*, justificando se este é efetuado antes ou depois da modelação dimensional, e mencionando o papel desempenhado pelo disco rígido.

O desenho físico de um data warehouse é feito depois da modelação dimensional e pretende resolver a problemática dos relatórios analíticos que envolvem grandes volumes de dados devido à agregação de muitos valores numéricos, inclusão de dados históricos e à construção interativa e incremental de relatórios. Desta forma a geração de relatórios pode ser morosa. Assim, o desenho físico procura minimizar a utilização de recursos através de menos dados pesquisados por interrogações SQL, menos espaço ocupado em disco aquando de atualizações periódicas e menos tempo para gerar relatórios dinâmicos. Neste aspeto o disco rígido desempenha um papel importante na organização e no tempo/custo de acesso aos dados, sendo que numa organização sequencial como num disco rígido, a transferência de dados sequenciais é 10 a 20 vezes mais rápida do que em posições aleatórias.

1. **2013.05** Considere os atributos ID, ÚltimoNome, Sexo, EstadoCivil, e RegiãoDoPaís de uma dimensão Cliente. Apresente um cenário de acesso a clientes em que seja vantajosa a utilização de índices de mapas de *bits*, indicando quais os atributos, de entre os mencionados, mais adequados a este tipo de índice, e descrevendo o conteúdo de um índice *bitmap* sobre um desses atributos.

Os índices de bitmap são frequentemente utilizados porque são compactos (sequências de bits ocupam pouco espaço) e eficientes (disjunções e conjunções realizadas por comparação direta de bits e o processador do computador faz comparação de forma muito eficiente). Para quem se depara com o caso do enunciado, poderia querer saber as mulheres que são casadas, consultando os atributos ID, Sexo e Estado Civil. O atributo Sexo poderia ser transformado em 0 para Masculino e 1 para Feminino e o atributo EstadoCivil transformado em 0 para solteiro e 1 para casado. Assim, o resultado seria o ID que apresentasse 01 01 nos atributos selecionados. Dependendo da granularidade pretendida para a RegiaoDoPais este campo poderia ou não ser transformado em mapa de bits (ex. Regiao = Norte, Centro, Sul vs Regiao = Tras os Montes, Alto Alentejo, Algarve…)

Table

Description automatically generated

1. **2013.05** Os índices nos *data warehouses* de tipo ROLAP costumam ocupar três vezes o espaço dos dados. Indique o propósito de tamanha quantidade de dados indexados, mencionando se os índices incidem sobre atributos só da tabela de factos, só das dimensões, ou de todas as tabelas, e, por fim, justifique se esta abundância de índices seria viável num sistema operacional.
2. **2012.05** Os cubos de dados, incluindo os respetivos dados agregados pré-calculados, podem ser armazenados em partições. Descreva dois cenários de uso de partições que potenciam a geração mais rápida de relatórios, mencionando, se apropriada, a tecnologia de servidores paralelos.

As partições são indicadas para tabelas de factos e dimensões “monstras” e pretendem acelerar a geração mais rápida de relatórios. Dois cenários para gerar relatórios mais rapidamente são as partições por intervalos de valores que é uma estratégia adequada para intervalos de datas e a partição de servidores paralelos através do shared nothing onde duas partições podem ser processadas em simultâneo e originam dois relatórios, focado cada um na sua partição e um outro relatório abrangendo todas as partições, em metade do tempo. Um exemplo seria uma das partições produzir um relatório relativo a dados da primeira metade da década passada e a outra partição da segunda metade da década passada, sendo o resultante um relatório para cada metade da década e um relatório para a década inteira.

1. **2012.05** Uma das tendências da *business intelligence* é os relatórios serem gerados a partir de cubos de dados guardados em memória RAM (volátil) do computador. Indique uma vantagem e uma desvantagem do designado *in-memory* BI face à produção de relatórios a partir do dispositivo tradicional de armazenamento de dados, e justifique se faz sentido construir e usar índices em memória RAM.
2. **2011.06** Considere os atributos RegiãoDoPaís, DonaDeCasa, e DiaDaSemana de um *data warehouse* de tipo ROLAP sobre audiências televisivas. Descreva o conteúdo de um índice de mapas de *bits (bitmap)* aplicado ao atributo que considerar mais adequado (dos apresentados) a este tipo de índice.

Para este caso o conteúdo de um índice de mapas bit poderia ser aplicado ao campo DonaDeCasa, que aparenta ser um simples campo de verificação que toma valores sim ou não o que o torna ideal para representação em modo binário. Dependendo da granularidade da tabela, a RegiaoDoPaís poderia também ser indexada desta forma. O DiaDaSemana pode tomar 7 valores diferentes, o que o torna menos ideal para este tipo de indexação, no entanto é possível fazê-lo caso seja pretendido.

1. **2011.06** As técnicas de compressão permitem reduzir o espaço ocupado pelos dados armazenados em disco. Sabendo que algum tempo extra é gasto pelo processador para descomprimir os dados a fim de poderem ser usados, explique como é que o tempo de geração de relatórios pode, ainda assim, ser inferior relativamente ao não uso de compressão.

As técnicas de compressão, apesar de apresentarem cerca de 10% de acréscimo no tempo de processador, compensam esta utilização extra de processamento pelo menor acesso a dados no disco (30% a 50% inferior), o que em tabelas de factos e dimensões “monstras” são percentagens significativas de dados e naturalmente de tempo e espaço.

1. **2011.05** Considere que pretende particionar os dados de uma tabela de dimensão “monstra,” na qual as mudanças lentas têm vindo a ser registadas através da técnica de tipo 2. Assumindo que adquiriu um novo disco mais rápido que os existentes e que a maioria dos relatórios precisa apenas dos registos que estão em vigor, justifique que estratégia de partição escolheria para esses relatórios serem gerados em menos tempo.

Para particionar estes dados poderia utilizar-se uma técnica de partição por lista, uma partição para os valores atuais guardados no disco mais rápido pois são acedidos mais frequentemente e a outra partição guarda os valores dos registos históricos.

1. **2011.05** Descreva um cenário em que seja vantajosa a utilização de uma vista materializada num *data warehouse* e justifique se seria necessário alterar as interrogações SQL dos relatórios existentes para tirar partido deste tipo de vista.

Num cenário onde existe a geração de um relatório frequente (executado repetidas vezes) como, por exemplo, as somas de quantidades e euros das vendas por clientes e sucursais. Através de uma vista materializada as somas de quantidades e euros ficam guardadas em disco, devidamente agrupadas por cliente e sucursal. Tipicamente as vistas são usadas automaticamente, porque o optimizador de interrogações recorre a vistas se forem opção vantajosa, sendo assim adaptadas automaticamente sem necessidade de alterar as interrogações.

1. **2010.07** Descreva duas caraterísticas das vistas materializadas que permitem reduzir o tempo de geração de relatórios que envolvam o cruzamento de dados provenientes de várias tabelas.

Duas das características das vistas materializadas residem no facto de permitirem o pré-cálculo e armazenamento de valores agregados e o facto de serem utilizadas automaticamente quando são vantajosas sem necessidade do conhecimento do utilizador.

1. **2010.06** Considere uma tabela de dimensão Cliente que, entre outros, guarda um conjunto de atributos demográficos relacionados entre si e atualizados de dois em dois anos. Supondo que é aplicada a técnica de compressão de página a esta tabela, justifique de onde viriam as principais poupanças de espaço em disco.

Perante o cenário descrito, as principais poupanças de espaço em disco residiriam no facto de a compressão de página identificar prefixos e valores comuns em registos da mesma página, guardando esses prefixos e valores comuns num dicionário de página, ou seja, quando houvesse uma atualização aos dados demográficos os prefixos e valores comuns (como por exemplo um ID da freguesia/concelho, o nome da freguesia/concelho, etc) já estariam comprimidos, havendo apenas a necessidade de atualização dos dados variáveis no tempo (por exemplo, número de habitantes)

1. **2010.06** Considere que, no contexto de um *data warehouse* de tipo ROLAP usado numa maternidade, pretende indexar os atributos Idade (em meses), Sexo, e CorDosOlhos, pois estes são frequentemente usados na produção de relatórios dinâmicos.
   1. Assumindo que é usual fazer a distinção entre bebés com menos de 4 meses dos restantes, justifique se seria apropriado utilizar um índice de função de dispersão para o atributo Idade.

O índice de função de dispersão não é indicado para este caso pois não é capaz de realizar operações de pesquisa por intervalos (dicing).

* 1. Admita que os valores possíveis para a cor dos olhos são azul, verde, e castanho. Descreva sucintamente o conteúdo de um índice de tipo *bitmap* aplicado a este atributo.

A aplicação de um índice de bitmap a este caso consistiria em três colunas com descrição “azul”, “verde”, “castanho”, em que cada bebe teria o valor de 1 para a cor de olhos que corresponde à sua e 0 para as outras.

CorDosOlhos

ID Nome Azul Verde Castanho  
1 Maria 0 0 1  
2 Ana 1 0 0

* 1. Considere a interrogação SELECT AVG(F.Peso) FROM tblFactos F, dimBebé B WHERE (F.fkBebé = B.ID) AND (B.Sexo = ’F’ AND B.CorDosOlhos = ’azul’). Justifique se o uso de índices de tipo *bitmap* em sexo e olhos permitiria dar uma resposta eficiente à interrogação.

A utilização dos índices de tipo bitmap permitiria dar uma resposta eficiente nos campos de sexo e olhos pois estes são atributos que possuem uma quantidade bastante limitada de valores.

1. **2009.06** Indique uma vantagem e uma desvantagem dos índices *bitmap* e descreva uma situação em que os usaria.

Vantagem: Compactos. As sequências de bits ocupam relativamente pouco espaço.

Desvantagem: Pouco apropriados para atributos com alta cardinalidade

Poderiam ser usados quando se pretende saber, por exemplo, as clientes do sexo feminino que são solteiras.

1. **2009.06** Na estruturação de um *cluster* de discos para *data warehousing*, justifique que tipo de distribuição de dados pelos discos, ou nível RAID, escolheria.

O nível RAID mais indicado para um data warehouse é o nível 5, data parity. Neste caso os dados são segmentados e distribuídos por vários discos. É gerada a parity data que permite recuperar os dados de qualquer disco que falhe sem necessitar da duplicação de dados. Este é o método RAID que melhor utiliza o espaço em disco.

1. **2009.06** Comente a seguinte afirmação: “o maior fator condicionante do desempenho de sistemas de *data warehousing* é o custo de *input/output*.”

A maior limitação do desempenho dos sistemas de data warehouse é o acesso ao disco rígido, onde são feitos os inputs e outputs do sistema. Isto deve-se ao facto de um disco rígido ter várias partes moveis, que necessitam de tempo para mover as cabeças até á linha do prato onde se encontram os dados, atraso na rotação do prato e tempo de transferência dos dados.

1. **2008.05** Indique as vantagens e desvantagens de guardar resultados agregados em disco no contexto da sua utilização em *data warehousing*.

# Parte V — Prospeção de Dados

1. **2019.06** Complete cada afirmação sobre prospeção de dados.
   * O atributo no topo de uma árvore de decisão é… uma classe possível e predeterminada
   * A geração de regras de associação tem em conta um valor mínimo de… suporte e confiança
   * No ciclo virtuoso da prospeção de dados, depois da geração de modelos vem a etapa de… decisão
   * Um subconjunto dos dados em que a classe (ou decisão) é sempre a mesma tem entropia… 0

• A proporção de verdadeiros positivos nos positivos detetados pelo modelo é dada pela métrica de…precisão

1. **2019.05** Complete cada afirmação sobre prospeção de dados.
   * Duas desvantagens das redes neuronais são serem caixas opacas e… de aprendizagem lenta
   * Os bons agrupamentos *(clusters)* devem ter simultaneamente vários dados e… estarem distantes uns dos outros
   * A proporção de dados relevantes obtidos face a todos os dados relevantes é dada pela métrica de… recall
   * O conjunto de dados usado para estimar a taxa de erro de um modelo designa-se… test dataset
   * No método *k-means*, existem tantos agrupamentos iniciais quanto… o número de k definido
2. **2019.05** O algoritmo Apriori permite reduzir o espaço de descoberta de conjuntos frequentes de itens em cestos de compras. Indique a propriedade em que se baseia este algoritmo e descreva como são obtidos os conjuntos e qual a condição de paragem. Por último, explique como são geradas as regras de associação a partir dos conjuntos frequentes de itens.

O algoritmo apriori é um algoritmo bottom up baseado na propriedade dos conjuntos de itens frequentes que constata que para um conjunto de itens frequente, qualquer subset desse conjunto deve também ser frequente. Este algoritmo segue os seguintes passos:

1. 1ª iteração – calculo do suporte dos itens individuais, remoção dos itens com menos suporte do que o limite mínimo estabelecido
2. Formação de pares com os itens que sobreviveram, calculo do suporte de todos os pares, remover os pares com suporte menor que o mínimo
3. E assim sucessivamente.

A condição de paragem é o quando deixarem de existir grupos que satisfaçam o critério de suporte definido inicialmente.

As regras de associação são geradas com base nos conjuntos frequentes que sobreviveram ao processo do algoritmo apriori e que gera regras com confiança superior ao mínimo. Podem existir duas regras possíveis de associação para um par {a,b}: {a}🡪 {b} ou {b}🡪 {a}. Sobrevive apenas a regra com maior confiança.

1. **2018.06** Descreva um cenário exemplificativo dos quatro passos do ciclo virtuoso da prospeção de dados e indique em quais dos passos participam o *data warehouse* e os sistemas operacionais.

Analisar – Modelar – Decidir – Avaliar – Analisar

SO é utilizado apenas para entrada dos dados na organização e não é tipicamente utilizado na prospeção de dados.

DW é utilizado em todas as fases de data mining descritas acima.

1. **2018.06** No âmbito das regras de associação, justifique se haveria ou não vantagem para a tomada de decisão em passar a aplicar o método a cestos contendo os produtos adquiridos em cada mês, em vez dos cestos de compras de cada visita ao supermercado. Indique também se essa mudança faria aumentar ou diminuir o suporte de cada produto, mencionando explicitamente os casos de produtos comprados frequentemente, como o pão, comprados rotineiramente, mas menos vezes, como o detergente de lavar a loiça, e comprados raramente, como um forno micro-ondas.

No âmbito das regras de associação existiria uma vantagem para a tomada de decisão aplicando o método a cestos para os produtos adquiridos em cada mês pois é uma forma de avaliar padrões em produtos que são comprados menos frequentemente e que numa análise de cestos feita para todas as visitas não terão suporte suficiente para serem geradas regras de associação. Neste caso, o suporte de produtos comprados frequentemente e dos produtos comprados rotineiramente subiria, mas o caso dos produtos comprados raramente provavelmente continuaria a não ter suporte suficiente para gerar regras de associação.

1. **2017.06** Indique afirmações verdadeiras (V) ou falsas (F) sobre o ciclo virtuoso da prospeção de dados.

F\_\_ A identificação do problema de negócio é feita usando os dados dos sistemas operacionais.

V\_\_ A decisão de alterar o negócio tem por base os modelos gerados pela prospeção de dados.

F\_\_ O *data warehouse* só participa na etapa da geração dos modelos dos dados.

F\_\_ A medição dos resultados da decisão é feita consultando os sistemas operacionais.

V\_\_ Pode ser necessário transformar os dados para poderem ser usados em métodos de prospeção.

1. **2017.05** Indique afirmações verdadeiras (V) ou falsas (F) sobre métodos de agrupamento.

V\_\_ É necessário, mas não suficiente, um bom *cluster* ter vários elementos.

F\_\_ O método *k-means* gera *clusters* a partir de pares com os *subclusters* mais próximos entre si. F\_\_ São supervisionados, pois o decisor fornece as caraterísticas que deseja encontrar nos dados.

V\_\_ No método hierárquico aglomerativo existem inicialmente tantos *clusters* quantos os dados.

F\_\_ O dendrograma pode ser usado para mostrar os *clusters* gerados pelo método *k-means*.

1. **2017.05** Indique afirmações verdadeiras (V) ou falsas (F) sobre árvores de decisão.

F\_\_ As folhas são decisões possíveis.

F\_\_ Os atributos que melhor diferenciam os dados ficam junto das folhas.

V\_\_ O atributo colocado no topo da árvore é que tiver maior ganho de informação.

V\_\_ A entropia de um conjunto de dados é zero quando todos os dados são iguais.

F\_\_ As árvores maiores costumam ser preferíveis, pois estão menos sujeitas a sobre-ajustamento.

1. **2017.05** Indique para que tipo de método de prospeção de dados fazem sentido os conjuntos de teste, treino, e validação, descrevendo o propósito de cada conjunto e algumas das suas caraterísticas desejáveis, como o tamanho relativo e os seus elementos. Nota: mostre qual a sequência lógica de uso destes conjuntos.

O método de classificação através de arvores de decisão ou redes neuronais, necessita da utilização de conjuntos de teste, treino e validação. O conjunto de treino serve para treinar o algoritmo para reconhecer as classes que estão a ser classificadas e corresponde a cerca de 2/3 ou 70% dos dados. O conjunto de validação é utilizado para melhorar o treino e para evitar o overfitting, sendo de tamanho de amostras muito reduzido. O conjunto de teste é utilizado para testar a classificação obtida e corresponde a 1/3 ou 30% dos dados.

Conjunto de treino 🡪 Método Classificação 🡪 Conjunto de validação 🡪 conjunto de teste

1. **2016.07** Indique afirmações verdadeiras (V) ou falsas (F) sobre métodos de prospeção de dados.

F\_\_ Os métodos de agrupamento servem para descobrir conjuntos frequentes de itens.

V\_\_ No método *k-means* são considerados k *clusters* iniciais de forma aleatória.

V\_\_ Se um conjunto de itens for frequente então os seus subconjuntos também são frequentes.

F\_\_ As redes neuronais agrupam dados com caraterísticas idênticas desconhecidas à partida.

V\_\_ O atributo no topo de uma árvore de decisão é o que tem maior ganho de informação.

1. **2016.05** Descreva os passos do método de agrupamento hierárquico aglomerativo, indicando quantos são e onde estão os agrupamentos iniciais, como são criados novos agrupamentos, e quais as condições de paragem.

O método de agrupamento hierárquico aglomerativo é do tipo bottom up e começa por assumir que existem tantos agrupamentos iniciais quantos os dados existentes, começando depois a juntar iterativamente os dados em clusters de forma hierárquica, primeiro os dois mais próximos, depois o par mais próximo deste etc. A condição de paragem deste método é a existência de um número suficiente de clusters ou a distância elevada entre clusters.

1. **2015.06** Indique afirmações verdadeiras (V) ou falsas (F) sobre prospeção de dados.

F\_\_ Os conjuntos de treino e de teste são essenciais com métodos não supervisionados.

F\_\_ Para aceitar um bom *cluster* é suficiente verificar se inclui vários pontos de dados.

F\_\_ O atributo no topo de uma árvore de decisão é o que melhor diferencia os dados.

F\_\_ Uma regra de associação com confiança elevada tem necessariamente suporte elevado.

V\_\_ As redes neuronais aprendem lentamente e são rápidas a classificar.

1. **2014.06** Na avaliação de modelos de classificação podem ser usadas as métricas de precisão e rechamada. Defina estas métricas e justifique se devem ambas ser consideradas em simultâneo ou se, regra geral, é suficiente calcular apenas uma delas.

Precisão = verdadeiros positivos / (verdadeiros positivos + falsos positivos ) – serve para identificar dos valores positivos quais ‘e q são verdadeiramente positivos

Recall = verdadeiros positivos / (VP+ falso negativos) - serve para identificar quanta da amostra foi classificada corretamente

Tem que ser ambos considerados para avaliar o modelo de classificação

1. **2014.06** Descreva o significado de entropia dos dados e explique, genericamente, de que forma pode ser usada na construção de árvores de decisão. Nota: tenha o cuidado de indicar se a entropia é aplicada a um atributo ou separadamente a cada valor de um atributo.

Entropia é a desorganização dos dados. A entropia é máxima na raiz da arvore e 0 nas folhas. De cada vez que os dados passam um no e é identificado um dado atributo a entropia diminui. A entropia é aplicada a cada valor de um atributo pois o tamanho da arvore pode não ter todos os ramos do mesmo tamanho.

1. **2014.05** Indique afirmações verdadeiras (V) ou falsas (F) sobre prospeção de dados.

F\_\_ O sobre-ajustamento *(overfitting)* causa más classificações nos dados de teste/avaliação.

V\_\_ Um conjunto de treino deve incluir dados representativos de todas as classes existentes.

F\_\_ O uso de vários métodos aumenta a confiança e dispensa a comparação com a realidade.

F\_\_ Se existir causalidade entre duas variáveis então estão correlacionadas, e vice-versa.

V\_\_ Os *outliers* podem ser causados por ruído ou más medições dos dados.

1. **2014.05** Identifique os papéis dos sistemas operacionais, *data warehouse*, e prospeção de dados num contexto organizacional e faça o seu enquadramento nas etapas do ciclo virtuoso da prospeção de dados.

SO – olhos e ouvidos – permitem a entrada de dados na organização

DW – memoria – guarda dados conformados das várias formas

DM – inteligência – permite analises sobre os dados, identificar padrões e extrair informação

SO – não entra no ciclo virtuoso mas entra no data mining

DW – Entra em todas as fases

DM – entra na geração e avaliação dos modelos

1. **2013.05** Explique o algoritmo *k-means*, mencionando os pontos iniciais (quantos são e onde estão), como vão sendo criados os agrupamentos, qual a condição de paragem, e se é ou não um método supervisionado.
2. **2013.05** Descreva o significado de uma regra de associação ter confiança muito elevada e explique se do ponto de vista da tomada de decisão isso se traduz sempre em algo de útil. Por fim, justifique se uma regra nessas circunstâncias pode ser ignorada pelo algoritmo Apriori.
3. **2013.05** Explique de que forma as redes neuronais podem auxiliar a tomada de decisão e indique se estas são adequadas a casos em que se pretende que o modelo de prospeção de dados vá evoluindo com a chegada de dados novos em vez de ter de ser criado um modelo de raiz.
4. **2012.05** Suponha o início de construção de uma árvore de decisão, estando em análise dois atributos, A e B, os quais admitem os valores “sim” ou “não”. Explique o que significa Entropia(Asim) = 0, e, assumindo que Entropia(Bsim) = 0, indique em que circunstâncias o atributo A seria o principal diferenciador, e ainda, justifique abaixo de que ramo da árvore, Asim ou Anão, seria colocado o atributo B.
5. **2011.06** Os métodos de prospeção de dados podem ser supervisionados ou não supervisionados. Indique em que circunstâncias um decisor pode ser levado a recorrer a cada tipo de método, ilustrando cada caso com um cenário plausível.
6. **2011.06** Comente a validade da seguinte afirmação, justificando: “as redes neuronais tendem a ser lentas durante a fase de treino/aprendizagem, mas rápidas na classificação de dados.”
7. **2011.06** Explique o propósito dos conjuntos de teste, treino, e validação na classificação de dados, indicando explicitamente a sequência lógica de uso destes conjuntos, bem como a percentagem da totalidade dos dados pré-classificados que é recomendada em cada.
8. **2011.06** Descreva uma vantagem e uma desvantagem do algoritmo de *k-means* face ao de agrupamento hierárquico aglomerativo, indicando também se ambos os algoritmos seguem ou não uma abordagem de tipo *bottom-up* (dos pequenos para os grandes *clusters*).
9. **2011.05** Explique o propósito das tarefas de agrupamento e classificação, indicando um método apropriado a cada uma das tarefas, e justificando se esses métodos são ou não supervisionados.
10. **2011.05** Considere uma loja onde diariamente são feitas milhares de compras pelos clientes, e onde se detetou que “quem leva pão também costuma levar leite.” Mais precisamente, em 1000 cestos, 600 continham pão e leite, 800 continham pão, e 700 continham leite.
    1. Calcule o suporte dos cestos com pão, leite, e pão e leite simultaneamente.
    2. Mostre qual das duas regras tem maior confiança: pão → leite ou leite → pão.
11. **2011.05** Justifique se uma regra de associação com confiança muito elevada (no caso limite, próxima ou até mesmo igual a 100%) pode ser ignorada pelo algoritmo Apriori. Nota: imagine um hipermercado onde também se vende ouro, prata, e diamantes.
12. **2011.05** Explique sucintamente o algoritmo de aprendizagem assente em *back-propagation*, no contexto das redes neuronais. Nota: pode considerar o caso em que o resultado obtido no neurónio de saída é superior ao resultado esperado.
13. **2010.07** Considere uma loja onde se detetou, com confiança de 100%, que “quem leva joias também leva caviar.” Sabendo que em 1000 cestos, 2 continham caviar, estime o número de pessoas que levaram joias. Nota: explique o seu raciocínio.
14. **2010.06** Os métodos de prospeção de dados podem ser supervisionados ou não supervisionados. Descreva uma diferença fundamental entre estas duas abordagens e dê um exemplo de método de cada tipo.
15. **2009.07** “Correlação não é causalidade” é um dito habitual em prospeção de dados. Discuta a validade deste dito no contexto da avaliação de modelos, e indique que cuidados se devem ter para discriminar os dois conceitos.
16. **2009.06** Comente a seguinte afirmação “o *overfitting* é um problema típico em modelos de prospeção de dados mal ajustados.”
17. **2009.06** Descreva os passos habituais na preparação de um conjunto de dados numéricos e alfanuméricos para a sua utilização num problema de classificação usando redes neuronais, salientando os principais cuidados a ter.
18. **2009.06** Considere que numa base de dados com cerca de 5 milhões de infrações e acidentes registados pela brigada de trânsito, encontra-se informação sobre condutores, viaturas envolvidas, e informação sobre o acidente/infração (por exemplo, local e tipo de infração, álcool e/ou drogas no sangue dos condutores envolvidos, comportamento, agentes envolvidos, entre outros). Indique como faria para procurar relações de correlação/causalidade entre os vários fatores e o tipo de acidente, referindo ainda as principais dificuldades que encontraria e que técnicas utilizaria para as ultrapassar.
19. **2008.09** Depois do ajustamento de um modelo de *data mining* para classificação, a sua utilização está sujeita a condicionantes. Refira algumas das principais, justificando a sua resposta.
20. **2008.05** Indique como faria para analisar os dados seguintes, referindo ainda quais as principais dificuldades que encontraria e que estratégias utilizaria para as ultrapassar.

A DAS-28 é uma medida empírica com valores contínuos da severidade de sintomas de artrite reumatoide, atribuída por análise visual de pacientes. Para um estudo desta patologia, fez-se o mapeamento genético de 8 caraterísticas, que controlam a expressão de uma proteína identificada em todos os casos desta doença (mas também presente em pessoas sem a doença) para um conjunto de 300 pacientes. Para cada paciente conhece-se ainda a medicação que toma, bem como um conjunto de cerca de 200 atributos qualitativos e quantitativos referentes às suas caraterísticas, ao seu modo de vida, e aspetos relacionados com a doença. Nem todos os atributos têm valores conhecidos.