Ficha de Exercícios 03

Gonçalo Rodrigues Pinto - A83732
Universidade do Minho

(19 de Março de 2021)

Resumo

O presente relatório descreve o trabalho de exploração do Weka, avaliação de algoritmos e comparação entre os conceitos de *Database, Data Warehouse e Dataset*. O presente trabalho teve como objetivo perceber as principais diferenças entre uma base de dados, um *data warehouse* e um *dataset*; compreender as questões legais e éticas adjacentes à utilização de dados para processos de *Data Mining*; instalar e atualizar novos recursos no Weka; e ainda avaliar corretamente as médias e desvios padrões dos resultados obtidos. De forma ao seu desenvolvimento foi colocado um conjunto de questões no sentido de alcançar os objetivos definidos.

1 Introdução

No 2º semestre do 1º ano do Mestrado em Engenharia Informática da Universidade do Minho, existe uma unidade curricular enquadrada no perfil de Descoberta do Conhecimento denominada por Engenharia do Conhecimento, que tem como objetivo a introdução ao conceito de descoberta do conhecimento.

A presente ficha enquadra-se nesta unidade curricular e pretende explorar o Weka(*Waikato Environment for Knowledge Analysis*) que possui uma coleção de algoritmos de *Machine Learning* para execução de tarefas de *Data Mining*. É um software que permite pré-processar grandes volumes de dados, aplicar diferentes algoritmos de *Machine Learning* e comparar vários outputs, assim é possível avaliar os seus algoritmos.

Nesta ficha pretendeu-se perceber as principais diferenças entre uma base de dados, um *data warehouse* e um *dataset*, compreender as questões legais e éticas adjacentes à utilização de dados para processos de *Data Mining*, instalar e atualizar novos recursos no Weka e avaliar corretamente as médias e desvios padrões dos resultados obtidos.

2 Questões

1. ***Qual ou quais as diferenças entre uma base de dados um data warehouse e um dataset?***
   Uma base de dados é uma coleção de dados relacionados que representam alguns elementos do mundo real, enquanto um data warehouse é um sistema de informação que armazena dados históricos de fontes únicas ou múltiplas. Uma outra diferença entre a base de dados e um data warehouse é que a primeira é criada de com o objetivo de registar dados, ao contrário da segunda cujo objetivo é analisá-los. A base de dados usa *On-line Transactional Processing* (OLTP) cujo objetivo são operações diárias, por outro lado um data warehouse utiliza *On-line Analytical Processing* (OLAP) cujo objetivo é a análise de dados e tomada de decisões.
   Por último, um dataset é um subset de uma base de dados ou data warehouse para criar um data set é necessário anexar, combinar e simplificar algumas expressões de dados, ou seja, apresentam-se num estado de mais fácil leitura.

2. ***Quais são algumas das limitações do data mining e como podem ser ultrapassadas?***
   Embora o data mining seja muito útil, este enfrenta muitos desafios durante a implementação. Os desafios podem estar relacionados com os dados, durante o processo é necessário extrair informações de grandes volumes de dados, dados em grandes quantidades normalmente encontram-se imprecisos, incompletos ou não são confiáveis, de forma a lidar com este problema poderemos ter que recorrer ao procedimento de Binning consultando os valores ao redor, ou também recorrendo ao Clustering onde os "outliers"podem ser detetados por agrupamento, onde os valores semelhantes são organizados em grupos ou "clusters".
   Uma outra limitação prende-se com o facto os dados do mundo real são realmente heterogéneos e podem ser dados demasiado complexos de analisar (por exemplo, imagens, áudio/vídeo, dados temporais, dados espaciais, texto em linguagem natural), torna-se difícil lidar com tantos tipos diferentes de dados e extrair as informações necessárias. Para isso na maioria das vezes devem ser desenvolvidas novas ferramentas e metodologias para extrair informações relevantes.

3. ***Qual a diferença entre Operational Data e Organizational Data?***
   Dados Operacionais são o tipo mais elementar de dados, vêm de sistemas transacionais que registam as atividades quotidianas, enquanto os Dados Organizacionais ajudam a proteger a privacidade das pessoas, ao mesmo tempo em que são úteis para os dataminers que procurem tendências numa determinada população.

4. ***Indique alguns constrangimentos éticos da utilização e aplicação do Data Mining***
   A informação pode ser vista como o novo petróleo, ou como a matéria-prima da nova revolução industrial. Tal como o petróleo, a informação tem de ser extraída, refinada e distribuída. Mas ao contrário de muitas outras matérias-primas, não é um produto escasso. O seu valor pode ser exponencial, dependendo da forma como é utilizado, sendo o custo marginal irrisório.[3]
   Alguns constrangimentos éticos da utilização e aplicação do Data Mining prendem-se, sobretudo, com a utilização de algoritmos, inseridos em sistemas de tratamento automatizado, que passa por mecanismos de data mining. Estes sistemas podem ser de tal forma sofisticados que podem conduzir a efeitos indesejados podendo ser usados para a manipulação de pessoas, para fins comerciais e até políticos, colocando em risco a liberdade e a própria democracia.
   Um outro constrangimento tem a ver a conciliação entre os princípios e regras da proteção de dados com a potencialidade da exploração da informação, sendo determinante para o desenvolvimento e sustentabilidade dos negócios. Esta conciliação passa pela revisão dos processos de negócio, no sentido de garantir um tratamento de dados pessoais mais controlado, evitando riscos de incumprimento.

5. ***O que é a normalização de bases de dados e quais os impactos em sistemas OLTP e OLAP?***
   A normalização de bases de dados é um conjunto de regras que visa, principalmente, a organização da mesma no sentido de reduzir a redundância dos dados, aumentar a integridade destes e o seu desempenho. Para normalizar a bases de dados, deve-se examinar as colunas de uma entidade e as relações entre entidades, com o objetivo de evitar-se anomalias observadas na inclusão, exclusão e alteração dos dados.
   Em relação ao impacto, a normalização em bases de dados implica uma maior capacidade de manter os dados corretos num sistema OLTP, todavia é necessário fazer a "desnormalização"(e respetiva replicação dos dados) quando se preparam bases de dados para sistemas OLAP.

6. ***Desenhe uma base de dados relacional com 3 tabelas. Garanta que cria o número de colunas e as colunas adequadas para estabelecer relações entre as tabelas.***

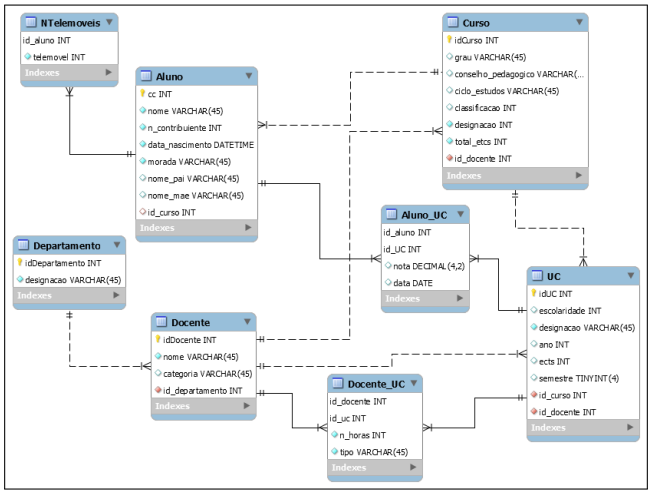


Figura 1: *Base de dados relacional.*

7. ***Desenhe uma tabela datawarehouse com algumas colunas que seriam desnormalizadas. Explique porque faz sentido "desnormalizar" nesta situação.***

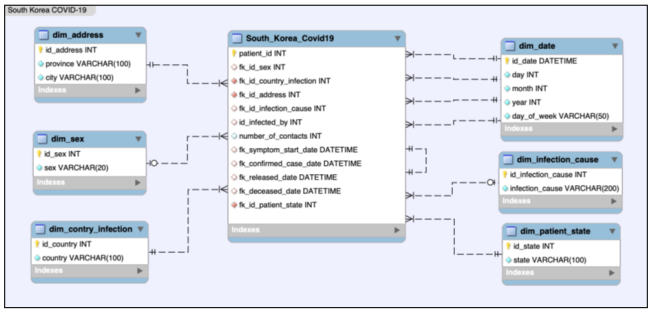


Figura 2: *Datawarehouse.*

Nesta situação faz sentido "desnormalizar"pois possibilita a consulta de forma mais rápida e eficiente, com o menor número possível de junções (operação por si só que coloca imensa carga num sistema de gestão de base de dados).

8. ***Faça uma pesquisa online e encontre 3 sites que contenham informação que pode ser aplicada ao processo de Data Mining.***
   https://www.kaggle.com/datasets
   https://www.data.gov/
   https://registry.opendata.aws/
   https://datasetsearch.research.google.com/

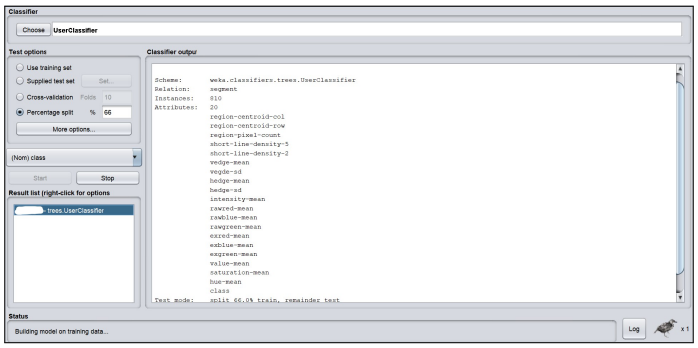
9. ***Faça uma pequena pesquisa online e descubra um dataset disponível para download. Descreva sucintamente o dataset (conteúdo, propósito, tamanho, antiguidade).***
   Dataset : https://www.kaggle.com/kimjihoo/coronavirusdataset
   No dataset em estudo, a Coreia do Sul está dividida em províncias e as suas respetivas cidades, com os diferentes tipos de foco de infeção COVID-19 bem como o seu respetivo país de origem. Cada caso está caracterizado com o seu respetivo range de idades, o seu género e a sua proveniência. Ao nível sintomático, encontra-se registada a data do início dos mesmos, bem como a sua data de cura e/ou óbito. O dataset pretende estudar o impacto e propagação do foco de infeção COVID-19 na Coreia do Sul. Este dataset teve a sua última atualização no passado dia 13/07/2020 e possui 7MB de informação.

   No ecrã inicial do Weka abrir o "package manager" Tools −> Package Manager. Instalar o package "UserClassifier 1.0.3". Após este passo responder às seguintes questões.

10. ***Abrir o Weka / Explorer e carregar o data set "segment-test.arff". Com este dataset carregado responda às seguintes questões:***

    (a) ***No separador Classify utilizar o trees −> UserClassifier;***
    Usar:

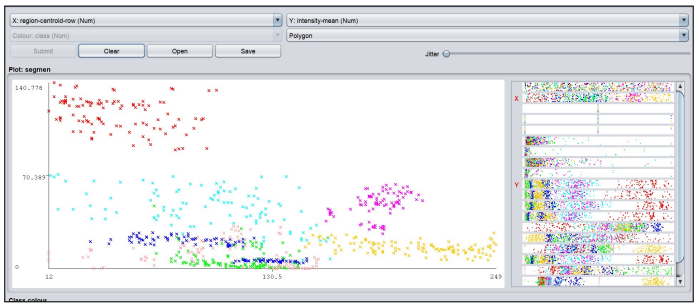
    - *Percentage split % 66*



    Clicar em Start;

Selecionar o separador Data Visualizer; e selecionar as seguintes opções (pode ser utilizado outro valor em vez do retângulo):

- *X: region-centroid-row (Num)*
- *Y: intensity-mean (Num)*



Ir selecionando os grupos possíveis de definir.
Determinar o resultado da classificação.

```
=== Summary ===

Correctly Classified Instances         199               72.3636 %
Incorrectly Classified Instances        76               27.6364 %
Kappa statistic                          0.6782
Mean absolute error                      0.0994
Root mean squared error                  0.2296
Relative absolute error                 40.556  %
Root relative squared error             65.4793 %
Total Number of Instances              275
```

R: Através do gráfico acima apresentado é possível observar a existência de clusters, ou seja, é possível efetuar uma divisão e respetiva diferenciação em classes. Posteriormente foram seleccionados os grupos de treino e teste possíveis de definir com o valor de retângulo, obtendo assim uma percentagem de acerto de 72.3636 % e uma percentagem de erro de 27.6364 % .

11. ***Comparar este método de criação de árvore de decisão com o algoritmo J48.***
Utilizando o método de criação de árvore de decisão com o algoritmo J48 (com Percentage split 66%) este classificou 257 instâncias corretamente, equivalendo a 93.4545% do total, por outro lado apenas classificou 18 instâncias como incorretas, equivalendo apenas 6.5455% do total. Os resultados obtidos podem ser observados na imagem abaixo apresentada. Os valores obtidos com este algoritmo foram consideravelmente melhores que o algoritmo UserClassifier contudo este último dependeu do utilizador para definir os grupos de treino e teste, assim é óbvio que algoritmo torne-se menos preciso.

```
=== Summary ===

Correctly Classified Instances         257               93.4545 %
Incorrectly Classified Instances        18                6.5455 %
Kappa statistic                          0.9233
Mean absolute error                      0.0203
Root mean squared error                  0.1246
Relative absolute error                  8.2765 %
Root relative squared error             35.526  %
Total Number of Instances              275
```

Figura 3: Método de criação de árvore de decisão com o algoritmo J48.

12. ***Abrir o Weka / Explorer e carregar o data set "segment-challenge.arff". Com este dataset carregado responda às seguintes questões:***

    (a) ***Usar o algoritmo J48 como classificador; Usar o data set "segment-test.arff". Qual o valor da classificação?***
    Utilizando o algoritmo J48 como classificador e usando o data set "segment-test.arff", este classificou 779 instâncias corretamente, equivalendo a 96.1728% do total, por outro lado apenas classificou 31 instâncias como incorretas, equivalendo apenas 3.8272% do total. Os resultados obtidos podem ser observados na imagem abaixo apresentada.

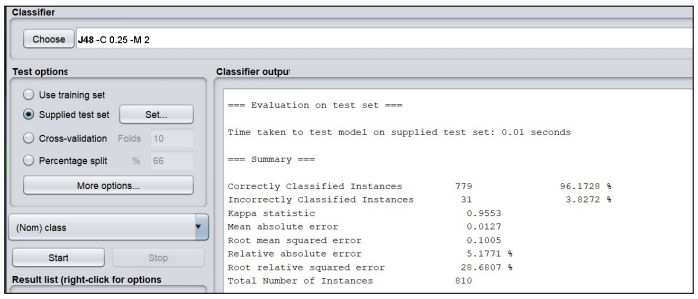


Figura 4: Algoritmo J48 como classificador utilizando o data set "segment-test.arff".

    (b) ***Usando a opção "Use training set" determine o valor da classificação. Por que não deve ser usada esta opção para determinar a qualidade e a aplicabilidade dos algoritmos aos dados?***
    Utilizando o algoritmo J48 como classificador e usando a opção "Use training set", este classificou 1485 instâncias corretamente, equivalendo a 99% do total, por outro lado apenas classificou 15 instâncias como incorretas, equivalendo apenas 1% do total. Contudo não deve ser usada esta opção pois significa que encontramos a testar conhecimento com os mesmos dados que foram aprendidos previamente.

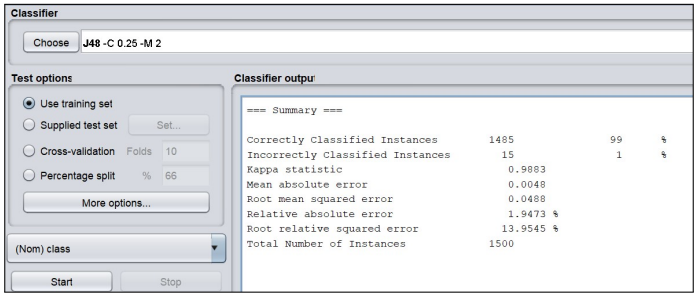


Figura 5: Algoritmo J48 como classificador utilizando a opção "Use training set".

13. ***Abrir o Weka / Explorer e carregar o data set "segment-challenge.arff". Com este dataset carregado responda às seguintes questões:***

    (a) ***Escolha o J48 como classificador e vá alternando as percentagens de divisão ("Percentage Split") dos grupos de treino e de teste em: 10%, 20%, 40%, 60% e 80%. O que observa?***
    Utilizando o método de criação de árvore de decisão com o algoritmo J48, com Percentage split inicial de 10% este classificou 1202 instâncias corretamente, equivalendo a 89.037% das 1350 instâncias totais, por outro lado apenas classificou 148 instâncias como incorretas, equivalendo apenas 10.963% das 1350 instâncias totais. Alternado as percentagens de divisão dos grupos de treino e de teste em 20%, 40%, 60% foi possível observar uma maior percentagem de acerto e consequentemente um menor número de instâncias classificadas como incorretas. Utilizou-se ainda uma percentagem e divisão dos grupos de treino e de teste de 80%, onde observou-se que foram apenas classificadas 300 instâncias das quais 290 foram corretas e apenas 10 incorretas. Estes resultados podem ser explicados pelo erro de *overfitting* que consiste num erro de modelagem que ocorre quando uma função se ajusta muito a um conjunto limitado de dados. Ao alterar a percentagens de divisão dos grupos de treino e de teste levou a que o modelo ficasse muito próximo de dados imprecisos infetando o modelo com erros substanciais e assim reduzir a capacidade de previsão.

(b) ***Repetir a questão anterior usando 90%, 95%, 98% e 99%. O que acontece ao número de instâncias corretamente classificadas? E o que acontece à percentagem de instâncias corretamente classificadas? Explicar esta variação.***
Utilizando o método de criação de árvore de decisão com o algoritmo J48, com Percentage split inicial de 90% este classificou 145 instâncias corretamente, equivalendo a 96.6667% das 150 instâncias totais, por outro lado apenas classificou 5 instâncias como incorretas, equivalendo apenas 3.3333% das 150 instâncias totais. Alternado a percentagem de divisão do grupo de treino e de teste para 95% foi possível observar uma diminuição da percentagem de instâncias corretamente classificadas de 0.6667% e em relação ao número de instâncias corretamente classificadas foram 72 das 75 instâncias totais. Alterando a percentagem de divisão do grupo de treino e de teste para 98% observou-se a mesma percentagem de instâncias corretamente classificadas obtida com Percentage split inicial 90% e a propósito do número de instâncias corretamente classificadas foram 29 das 30 instâncias totais. Por fim, alternado a percentagem de divisão do grupo de treino e de teste para 99% foi possível perceber que o modelo conhece já todo o dataset, portanto, já não precisa de efetuar cálculos apenas apresentar o resultado final direto dos valores que possui. Esta variação pode ser explicada pelo erro acima apresentado, *overfitting*, ao alterar as percentagens de divisão do grupo de treino e de teste o modelo apenas encontra-se a verificar dados que já lhe são próximos, assim algo que é diferente do que já conhecido não é processado com tanta flexibilidade, e consequentemente, é afetado a capacidade de previsão.

(c) ***Apesar de com uma percentagem de 98% para o treino e 2% para o teste dar uma classificação de 100% isto quer dizer que o modelo construído é o mais indicado para o problema apresentado?***
Tal como foi observado previamente com uma percentagem de 98% para o treino e 2%, este modelo construído não é o mais indicado para o problema apresentado apesar de para o teste dar uma classificação de 100%.

(d) ***Tendo em conta as experiências acima realizadas qual será a percentagem de classificações corretas do algoritmo J48 neste data set?***
Tendo em conta as experiências acima realizadas a percentagem de classificações corretas do algoritmo J48 neste data set será aproximadamente de 96%.

14. ***Abrir o Weka / Explorer e carregar o dataset "diabetes.arff". Com este dataset carregado responda às seguintes questões:***

    (a) ***Selecionando "Percentage Split" a 80% quantas instâncias serão usadas para treino e quantas serão usadas para teste? (O Weka arredonda ao número inteiro mais próximo).***
    Utilizando o algoritmo J48 com Percentage split de 80% são usadas 154 instâncias para teste das 768 instâncias totais, assim 614 instâncias são usadas para treino.

    (b) ***Mudando o "Random seed" entre 1,2,3,4 e 5, mantendo o "Percentage Split" a 80% indique o valor mínimo e máximo de instâncias incorretamente classificadas.***
    Utilizando o método de criação de árvore de decisão com o algoritmo J48, com Percentage split de 80% mas variando o "Random seed" entre 1 e 5 foi obtido o máximo de 44 instâncias incorretas com o "Random seed" de 5 e o mínimo de 31 instâncias incorretas com o "Random seed" de 4.

    (c) ***Qual a média da percentagem de instâncias corretamente classificadas?***
    A média da percentagem de instâncias corretamente classificadas foi de 75.97402 % $[\frac{75{,}974+77{,}2727+75{,}3247+79{,}8701+71{,}4286}{5}]$.

    (d) ***Qual o desvio padrão da taxa de acerto para os resultados acima calculados?***
    O desvio padrão da taxa de acerto para os resultados acima calculados foi de 3.08011, este valor foi calculado através da seguinte expressão $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$.

    (e) ***Se repetisse o exercício [14/b] com 10 "random seed" em vez de 5 qual seria o efeito na média e desvio padrão?***
    Utilizando "random seed" com 10 em vez de 5 cuja taxa de acerto é 75.3247 % provocaria um ligeiro aumento na média para 76.75324 % e um decréscimo no desvio padrão para 1.91528.

COMPARAR COM "BASE LINE"

15. ***Abrir o Weka / Explorer e carregar o data set "iris.arff". Com este dataset carregado responda às seguintes questões:***

    (a) ***Este dataset caracteriza 3 classes com 50 instâncias cada uma. Qual será a percentagem de acerto do algoritmo ZeroR quando aplicado ao training set?***
    A percentagem de acerto do algoritmo ZeroR quando aplicado ao training set é de 33.3333 %.

(b) ***Qual é o resultado da classificação baseline quando é usado o método "Percentage Split" em 66%?***
Quando é usado o método "Percentage Split" em 66% com o algoritmo ZeroR foi obtido o seguinte resultado da classificação baseline: 29.4118 % de percentagem de acerto (15 instâncias das 51 totais) e 70.5882 % de percentagem de erro (36 instâncias das 51 totais).

16. ***Abrir o Weka / Explorer e carregar o data set "glass.arff". Com este dataset carregado responda às seguintes questões:***

(a) ***Qual é a percentagem de acerto do algoritmo ZeroR com 66% de "Percentage Split"?***
A percentagem de acerto do algoritmo ZeroR com 66% de "Percentage Split" é de 27.3973 % (20 instâncias corretas das 73).

(b) ***Qual o valor usando o J48 e os restantes parâmetros por defeito?***
O valor usando o J48 e os restantes parâmetros por defeito é de 57.5342 % (42 instâncias corretas das 73).

(c) ***Qual a precisão (accuracy) do algoritmo NaiveBayes' usando os parâmetros por defeito?***
A precisão (accuracy) do algoritmo NaiveBayes' usando os parâmetros por defeito foi de 0,589.

17. ***Abrir o Weka / Explorer e carregar o data set "segment-challenge.arff". Utilize o data set "segment-test.arff" para dataset de avaliação (teste). Com estes datasets carregados responda às seguintes questões:***

(a) ***Qual a precisão do algoritmo ZeroR?***
A precisão do algoritmo ZeroR foi de 11.6049 % (94 instâncias das 810).

(b) ***Qual a precisão do algoritmo Ibk's, com todos os parâmetros por defeito?***
A precisão do algoritmo Ibk's, com todos os parâmetros por defeito foi de 95.8025 % (776 instâncias das 810).

(c) ***Qual a precisão do algoritmo PART, com todos os parâmetros por defeito?***
A precisão do algoritmo PART, com todos os parâmetros por defeito foi de 95.5556 % (774 instâncias das 810).

3 Conclusão

O presente relatório descreveu, de forma sucinta, o trabalho de exploração do Weka e avaliação dos seus algoritmos.

Após a realização deste trabalho, percebi as principais diferenças entre uma base de dados, um *data warehouse* e um *dataset*, compreendi as questões legais e éticas adjacentes à utilização de dados para processos de *Data Mining*, bem como a instalação e atualização de novos recursos no Weka permitiu avaliar corretamente as médias e desvios padrões dos resultados obtidos.

Por fim, espero que os conhecimentos obtidos e consolidados sejam de enorme utilidade tendo uma perspetiva futura.

Referências

[1] Database vs Data Warehouse: Key Differences. Guru99. 2021. URL: https://www.guru99.com/database-vs-data-warehouse.html (acedido em 12 de março de 2021).

[2] 08 - Challenges in Data Mining. Wideskills. 2015. URL: https://www.wideskills.com/data-mining/challenges-in-data-mining (acedido em 12 de março de 2021).

[3] A privacidade dos dados: uma questão ética. Ver. 2018. URL: https://www.ver.pt/a-privacidade-dos-dados-uma-questao-etica/ (acedido em 12 de março de 2021).