1.

Uma base de dados apresenta as informações de forma normalizada, distribuindo os dados a reter por diversas tabelas. É otimizada para sofrer reads e writes constantes, com imensas transações. Por sua vez, em prol da rapidez e facilidade de consulta, num data warehouse é feita a desnormalização de alguns destes dados, passando os mesmos a constar numa só tabela. Por último, num dataset, os dados apresentam-se em estado de mais fácil leitura, como um refinamento que facilita o processo analítico.

A partir de uma base de dados (OLTP) é possível construir um data warehouse, onde se reúnem as informações relevantes a análise (OLAP), de onde se extrai um subconjunto específico para análise com os formatos desejáveis.

2.

Algumas das limitações do data mining estão relacionadas com a violação de privacidade, na medida em que nem tudo pode ser incluído como dado analisável. Adicionalmente, pode ser incluída informação irrelevante e que apenas atrase ou complique o processo de análise. De acordo com os dados disponíveis, pode acontecer o uso de informação de forma incorreta, atribuindo-lhe importância errada. Por fim, a qualidade dos dados também é muito importante e, caso estes sejam falíveis e pouco corretos, o processo de data mining torna-se bastante mais falível.

3.

Os dados operacionais relacionam-se com o fluxo de operações diárias de uma qualquer entidade possuidora de base de dados. Dão suporte a todas as necessidades operacionais e de serviço. Por outro lado, os dados organizacionais são referentes a dados que podem ser importantes para o uso e reconhecimento de informação, quer por processos de data mining, quer por ferramentas de exploração OLAP.

4.

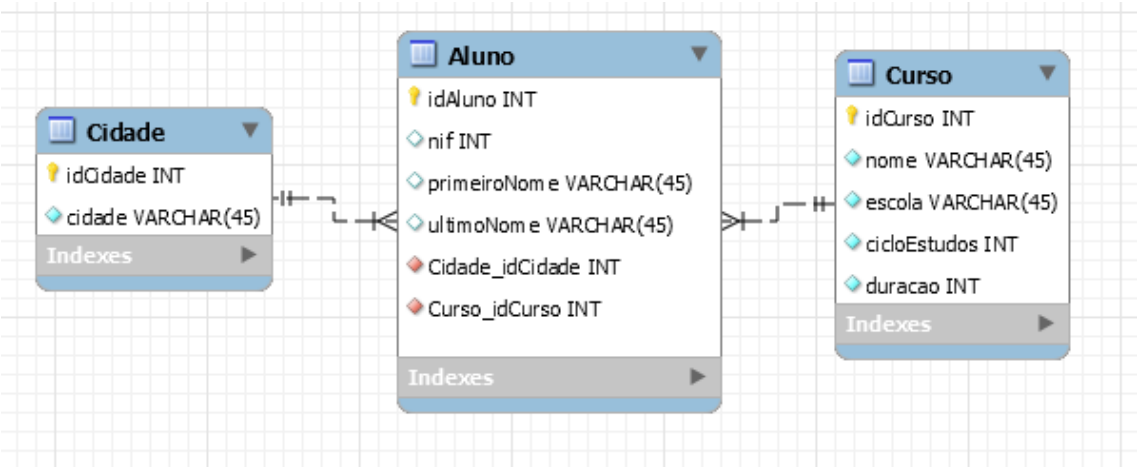
No processo de data mining é possível o uso de diversos dados, entre os quais dados privados. Quando se reúnem um grande conjunto de informações sobre determinada pessoa, é possível começar a perceber padrões e ocorrências que, dependendo da situação, podem penetrar na vida privada do indivíduo. Para além do uso de dados pessoais, há também a possibilidade de serem reconhecidas informações acerca de determinado indivíduo com recurso a dados perfeitamente gerais mas que, sob análise de uma ferramenta de processamento de dados, dê reporte informações privadas. Quando tal acontece, entra-se num campo muito suscetível da ética.

5.

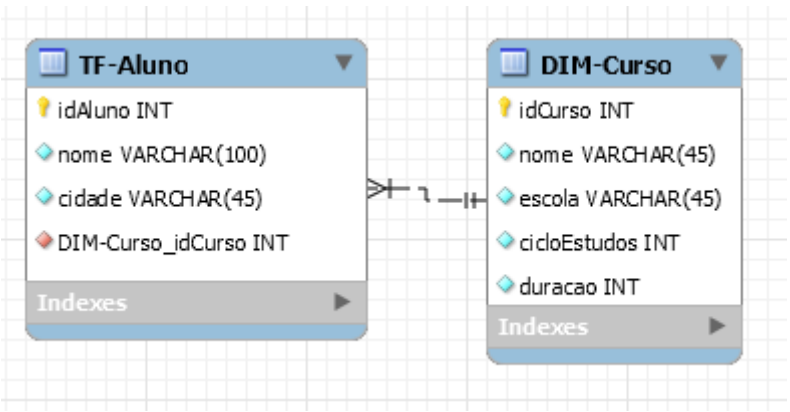
A normalização em bases de dados implica uma maior capacidade de manter os dados corretos num sistema OLTP. Por exemplo, no caso da inserção de uma cidade numa base de dados, com "Guimaraes", "guimaraes", "Guimarães" provavelmente queria apenas referir-se a mesma cidade. O uso de chaves possibilita a não inserção de elementos que devam ser os mesmos sob formatos diferentes. Adicionalmente, é poupado espaço.

Todavia, é necessário fazer a desnormalização (e respetiva replicação dos dados) quando se preparam bases de dados para sistemas OLAP, de forma a possibilitar a consulta de forma mais rápida e eficiente, com o menor número possível de JOINS, operações que colocam o SGBD sob imensa carga.

6.



7.



8.

https://www.kaggle.com/

https://registry.opendata.aws/

https://cloud.google.com/bigquery/public-data/

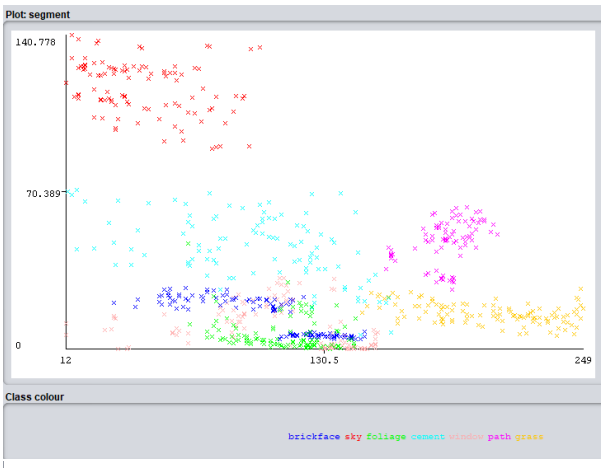
9.

Mais de 13 milhões de moradas australianas com geocodes (coordenadas de latitude e longitude). Este dataset, referido como G-NAF, é um dos mais ambíguos e poderosos datasets espaciais, não contendo qualquer nome ou outras informações pessoais. O G-NAF é produzido pelo PSMA Australia Limited (PSMA), uma empresa formada pelos nove governos da austrália, com o propósito de colecionar, standardizar e agregar localizações.

https://data.gov.au/search

10.

No caso especificado e em outros gráficos é possível perceber a existência de clusters, ou seja, é possível fazer uma divisão e diferenciação de classes. Com essa divisão criou-se o modelo da última imagem. Todavia, não foi possível apresentar o nível de acerto (por motivo que não consegui identificar).




```
=== Classifier model (full training set) ===


Split on region-centroid-row AND intensity-mean (In Set): N1 sky(109.0)
Split on region-centroid-row AND intensity-mean (Not in Set)
|   Split on region-centroid-row AND intensity-mean (In Set): N3 path(93.0)
|   Split on region-centroid-row AND intensity-mean (Not in Set)
|   |   Split on region-centroid-row AND intensity-mean (In Set): N5 cement(97.0/4.0)
|   |   Split on region-centroid-row AND intensity-mean (Not in Set)
|   |   |   Split on region-centroid-row AND intensity-mean (In Set): N7 grass(117.0/2.0)
|   |   |   Split on region-centroid-row AND intensity-mean (Not in Set)
|   |   |   |   Split on hue-mean AND intensity-mean (In Set): N9 brickface(129.0/6.0)
|   |   |   |   Split on hue-mean AND intensity-mean (Not in Set)
|   |   |   |   |   Split on hue-mean AND intensity-mean (In Set): N11 window(96.0/23.0)
|   |   |   |   |   Split on hue-mean AND intensity-mean (Not in Set)
|   |   |   |   |   |   Split on hue-mean AND intensity-mean (In Set): N13 foliage(95.0/1.0)
|   |   |   |   |   |   Split on hue-mean AND intensity-mean (Not in Set)
|   |   |   |   |   |   |   Split on hue-mean AND intensity-mean (In Set): N15 window(48.0/14.0)
|   |   |   |   |   |   |   Split on hue-mean AND intensity-mean (Not in Set): N16 window(26.0/12.

Time taken to build model: 138.34 seconds
```

11.

No entanto, o método J48 foi agradavelmente correto na classificação, como se pode ver pela imagem.

```
=== Summary ===

Correctly Classified Instances         757               93.4568 %
Incorrectly Classified Instances        53                6.5432 %
Kappa statistic                          0.9235
Mean absolute error                      0.02
Root mean squared error                  0.1312
Relative absolute error                  8.1735 %
Root relative squared error             37.5168 %
Total Number of Instances              810

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Cla
               0,960     0,009     0,952       0,960   0,956      0,948    0,990     0,917     bri
               1,000     0,001     0,991       1,000   0,995      0,995    0,999     0,991     sky
               0,844     0,022     0,873       0,844   0,858      0,834    0,934     0,823     fol
               0,900     0,010     0,934       0,900   0,917      0,904    0,954     0,842     cem
               0,881     0,031     0,841       0,881   0,860      0,834    0,928     0,809     win
               0,989     0,001     0,989       0,989   0,989      0,988    0,995     0,989     pat
               0,984     0,003     0,984       0,984   0,984      0,981    0,994     0,976     gra
Weighted Avg.  0,935     0,012     0,935       0,935   0,935      0,923    0,970     0,903

=== Confusion Matrix ===

   a   b   c   d   e   f   g   <-- classified as
 120   0   3   0   2   0   0 |   a = brickface
   0 110   0   0   0   0   0 |   b = sky
   4   0 103   1  14   0   0 |   c = foliage
   0   1   2  99   5   1   2 |   d = cement
   2   0  10   3 111   0   0 |   e = window
   0   0   0   1   0  93   0 |   f = path
```

12.

a.

```
=== Summary ===

Correctly Classified Instances         779               96.1728 %
Incorrectly Classified Instances        31                3.8272 %
```

b. Este método usa apenas os dados existentes e corre o risco de fazer over-fitting, sendo preciso para os dados disponíveis mas não flexível o suficiente para outros.

```
=== Summary ===

Correctly Classified Instances        1485               99      %
Incorrectly Classified Instances        15                1      %
```

13.

a. Observa-se um cada vez maior acerto. No entanto, este é derivado a um over-fitting, pelo que não será o melhor método a utilizar.

b. Vê-se um decréscimo ligeiro de 90% para 95%. Como sugerido anteriormente, o modelo está apenas a verificar dados que já lhe são próximos e algo que seja diferente do que é conhecido não é recebido com a mesma flexibilidade e, consequentemente, precisão. Por fim, nos 98% para 99%, percebe-se que o modelo conhece já todo o dataset e, portanto, já não precisa de adivinhar mas sim fazer o output direto dos valores que tem nos neurons.

c. Como sugerido anteriormente, definitivamente que não.

d. Seria cerca de 95%, usando o método de cross-validation com 10 folds. Pode ser aprimorado mas dificilmente ficará muito melhor, pois com 50 folds fica a 96,6% e com 100 fica a 96,8%, sendo que com 10 é um modelo mais flexível.

14.

a. Serão usadas 614 instâncias para treino e 154 para teste.

b. O número mínimo foi 20 (com random seed = 4) e o máximo foi 37 (com random seed = 1).

c. 75.9740%.

d. Não estou certo sobre como seria suposto fazer este cálculo. Se for com recurso a “Root mean squared error”, seria de cerca de 0.41.

e. A média de acerto subiria uma vez que o valor com random seed = 10 é melhor que com a mesma = 5. O desvio padrão seria menor pois a taxa de acerto seria superior.

15.

a. Apenas 33,33%.

b. Apenas 29,41%.

16.

a. Apenas 27,40%.

b. 57,53%.

c. 0,589.

17.

a. 11.6049%.

b. 95.8025%.

c. 95.5556%.