Student Number: S5132483 Last Name: Perry First Name: Sebastian

1)

a. Data mining is the extraction, extrapolation, and inference of knowledge and patterns from large amounts of data, ranging in sources from databases and physical media to videos and images.

b. Different tasks achieved by data mining include:

Web page analysis – useful in categorising websites and tracking their usage

Collaborative analysis – useful for inferring good recommendations of content to users based on their previously viewed content.

Basket data analysis – useful for generating targeted advertisement based on previous purchase history.

Medical data analysis – allowing for the classification and statistical analysis of conditions, and possible predictions of complications based on medical history.

Database insight tools – such as MySQL and such

c. The Major issues in Data Mining include:

Mining method – Finding a method which works for the given situation.

User interaction – finding ways to allow users to effectively gain the insights they want and displaying said insights in a way which is useful.

Efficiency and scalability – ensuring that data mining methods stay efficient and scalable.

Data type diversity – finding ways to extract useful insights from diverse types of data.

Social impact of Data mining – data mining sometimes being seen as a bad practice.

d. To attain a higher level of understanding and skills which improve my ability to extract meaning insights from large amounts of data.

2)

a. The attribute types are as follows:
   Age = Interval-Scaled
   Gender = Nominal
   Postcode = Nominal
   Weight = Interval-Scaled
   Admission date = Interval-Scaled

b. The attribute types are as follows:
   Age = Continuous
   Gender = Discrete
   Postcode = Discrete
   Weight = Continuous
   Admission date = continuous

c. The basic statistical descriptions includes the central tendency of the data (this could be seen with things such as the mean, median, mid-range) and the dispersion of the data (which can be seen with something like the five number summary).

d. $given\ the\ set\ of\ body\ weights\ (in\ kg) = v$

$$v = \{41, 41, 42, 43, 45, 46, 49, 50, 55, 58, 61, 66, 73, 79, 80, 85, 87, 89, 92, 98\}$$
$$therefore:$$
$$vMinimum = 41$$
$$vFirstQuartile = 45.5$$
$$vMedian = 59.5$$
$$vThirdQuartile = 82.5$$
$$vMaximum = 98$$

e. $let\ body\ weight\ data = v$

$v = \{41, 41, 42, 43, 45, 46, 49, 50, 55, 58, 61, 66, 73, 79, 80, 85, 87, 89, 92, 98\}$
$which\ can\ be\ devided\ into\ 3\ groups\ (v \leq 55, \quad 56 \leq v \geq 80, \quad 81 \leq v)$

$Interval$	$Frequency$
$v \leq 55$	9
$56 \leq v \geq 80$	6
$81 \leq v$	5

$Therefore\ the\ median\ interval\ is\ 56 \leq v \geq 80\ which\ has\ a\ frequency\ of\ 6$

$As\ such: L_1 = 56, \quad N = 20, \quad \left(\sum freq\right)_l = 9, \quad freq_{median} = 6, \quad width = 24$

$$median\ approximation = L_1 + \left(\frac{\frac{N}{2} * (\sum freq)_l}{freq_{median}}\right) * width$$
$$= 56 + \left(\frac{\frac{20}{2} - 9}{6}\right) * 24$$
$$= 60kg$$

3)

a. Three other dimensions of data quality include:

Interpretability – how easy it is to relevantly understand the data.

Timeliness – how up to date the data is.

Believability – how trustable is the data.

b. While incomplete data refers to situations in which data has not been input at all or is otherwise missing, noisy data refers to situations in which attributes have values ranging from blatant error to outliers. Incomplete data can be remedied by either removing the record out auto filling a value such as unknown; while noisy data on the other hand can also be handled by disregarding the record, it can alternatively be smoothed by fitting the data to a regression function.

c. $where\ d_1 = \{4,1,5,6,8,2,4,6,1\}\ and\ d_2 = \{0,1,3,3,9,0,4,2,5\}$

$$Cosine\ Similarity = \cos(d_1, d_2) = \frac{d_1 . d_2}{||d_1|| * ||d_2||}$$

$$= \frac{(4*0+1*1+5*3+6*3+8*9+2*0+4*4+6*2+1*5)}{(4*4+1*1+5*5+6*6+8*8+2*2+4*4+6*6+1*1)^{0.5} * (0*0+1*1+3*3+3*3+9*9+0*0+4*4+2*2+5*5)^{0.5}}$$

$$= \frac{(0+1+15+18+72+0+16+12+5)}{(16+1+25+36+64+4+16+36+1)^{0.5} * (0+1+9+9+81+0+16+4+25)^{0.5}}$$

$$= \frac{139}{199^{0.5} * 145^{0.5}}$$

$$= 0.818$$

4)

a. Data reduction is the act of reducing down a representation of data while still attempting to preserve the same analytical results. Some strategies of data reduction include:

Dimensionality reduction – a method in which an attempt is made to remove less important attributes (including but not limited to attributes which provide irrelevant data which can be calculated given the presence of other data).

Clustering – a process in which data can be portioned into groups based on similarities and simply represent these groups.

Sampling – a method in which a representative subset of the data is chosen to represent the data as a whole.

b.

i. $Let\ range\ 200\ to\ 1000\ normalised\ to\ [0.0, 1.0]\ then:$

$$Using\ the\ equation: v' = \frac{v - min}{max - min}(new\_max\ - new\_min) + new\_min$$

$$v' = \frac{v - 200}{1000 - 200}(1\ - 0) + 0$$

$$v = \{200,300,400,600,1000\}$$

$$v' = \{0, 0.125, 0.25, 0.5, 1\}$$

ii. $Let\ Average = 500\ and\ standard\ deviation = 282.842$
$(calculated\ on\ a\ spreadsheet)$

$$Using\ the\ equation: v' = \frac{v - average}{standard\ deviation}$$

$$v' = \frac{v - 500}{282.8427}$$

$$v = \{200,300,400,600,1000\}$$

$$v' = \{-1.060, -0.707, -0.353, 0.353, 1.767\}$$

c. $given\ the\ set\ of\ body\ weights\ (in\ kg) = v$

$$v = \{41, 41, 42, 43, 45, 46, 49, 50, 55, 58, 61, 66, 73, 79, 80, 85, 87, 89, 92, 98\}$$

i. Equal depth partitioning method

Bin#	Records
Bin 1	$41, 41, 42, 43, 45$
Bin 2	$46, 49, 50, 55, 58$
Bin 3	$61, 66, 73, 79, 80$
Bin 4	$85, 87, 89, 92, 98$

ii. Equal width partitioning method

$$width = (max - \min)/N$$
$$= \frac{(98 - 41)}{4}$$
$$= 14.25$$

Bin#	Records
Bin 1	$41, 41, 42, 43, 45, 46, 49, 50, 55$
Bin 2	$58, 61, 66$
Bin 3	$73, 79, 80$
Bin 4	$85, 87, 89, 92, 98$

5)

a. The four characteristics of data warehouses is that they are:

subject orientated – are focused on important decision altering subjects

integrated – possibly from many different sources in such a way that is clear, consistent, concise , and streamlined

time variant – a source for a historical perspective far into the past, far more so then what may be supplied by the current operational database

non-volatile – is able to first emplace and find information, but never update.

b. The main components that make up the bottom tier of a data warehouse will more often than not consist of a relational database, in addition to a repository containing information on the warehouse and its contents.

c. Below will be a brief description of star, snowflake, and fact constellation schemas:

Star schema – the most basic form (off which the other 2 are based), it consists of a single fact table connected to a set of dimension tables.

Snowflake schema – is a refinement of the star schema in that some of the dimension tables have been normalised into a set of smaller dimension tables.

Fact constellation schema – slightly similar to the star scheme, it instead has multiple fact tables sharing the sets of dimension tables.

d. There are 4 main OLAP operations which include:
Roll up
Roll down
Slice and dice
Pivot

e. If a data cube has 5 dimensions and 3 levels per dimension, then the total number of cuboids that can be generated are:

$$total\ number\ of\ cuboids = 3^5$$