AU-6380 DATA MINING AND DATA WAREHOUSING (MCA-503)

MODEL ANSWER

Q.1

(a) Data mining is extraction of knowledge from large amount of data. NO, data mining is more than a simple transformation of technology developed from databases, satistics and machine learning.

(b) We can apply data mining techniques to the following types of databases.

(i) Relational databases.
(ii) Data warehouse.
(iii) Transactional databases.
(iv) spatial databases.
(v) Time-series databases.
(vi) Text databases.
(vii) multimedia databases.

(c) It stores sequences of values or events obtained over repeated measurements of time (i,e hourly, daily, weekly)

Ex:- stock exchange data, change in temp., inventory control. etc.

(d) Market-basket analysis means to enable us to bundle groups of items together as a strategy for maximizing sales.

Ex The printers are commonly purchased together with computers. We could offer an expensive printer at a discount rate to customers buying computers, in the hope to increase the selling of the expensive printers.

(e) classification predict the class lables of unknown tuples, whereas in case of prediction we develop a model to predict the unknown or ~~future~~ future value.

(f)

Data warehouse	Data mart
(1) It collects information about entire organization.	(1) Data mart contains a subset of corporate wide data that is value to a specific group of users,
(2) It can be implemented using traditional mainframes,	(2) While data mart can be implemented on low-cost servers.
(3) The implementation cycle may take months or years	(3) Its implementation requires few weeks.

(g) DMQL - Data mining query language. SQL is used for writing relational query, where as DMQL can be used to specify data mining tasks.

(h) NO coupling → DM system will not use any function of a DB/DW system. It uses flat files as data sources.

Loose coupling → DM system will use some facilities of a DB/DW system.

Semitight coupling → It means besides linking a DM system to a DB/DW system, efficient implementation of a few essential data mining primitives can be provided in the DB/DW system.

Tight coupling → It means ~~the~~ a DM system is integrated into the DB/DW system.

(i) Data reduction is applied to obtain a reduced representation of data set that is much smaller in volume and maintain the integrity of the original data.

In case of dimensionality reduction, ~~data~~ the no. of attributes are reduced/compressed by appeying data encoding or treansformation techniques. Ex:- PCA, FA, DCT, DWT etc.

(j) Holistic measure must be computed on the entire data set as a whole.
example :- approximate median value for the data.

Q. 3

(a) The snowflake schema and fact constellation are both variants of the star schema model which consists of a fact table and a set of dimension tables, the snowflake schema contains some normalized dimension tables, whereas the fact constellation contains a set of fact tables that share some common dimension tables.

A starnet query model is a query model, which consists each point along the line represents a level of the dimension. Each step away from the center represents the stepping down of a concept hierarchy of the dimensions.

(b) Data cleaning is the process of detecting errors in the data and rectifying them when possible.

Data transformation is the process of converting the data from heterogeneous sources to a unified data warehouse format or semantics.

Refresh is the function propagating the updates from the data sources to the warehouse.

(C)

Enterprise warehouse	Data mart
(1) Enterprise warehouse collects information about entire organization	(1) Data mart contains a subset of corporate wide data that is value to a specific group of users.
(2) It can be implemented using traditional mainframes, computer superservers etc.	(2) While data mart can be implemented on low-cost servers.
(3) The implementation cycle may take months or years	(3) Its implementation requires few weeks.
(4) It contains detailed as well as summarized data	(4) The data in a data mart lead to be summarized.

Q.4

(a) approximate median $= L_1 + \left(\frac{\frac{N}{2} - (\Sigma freq)_L}{freq_{median}}\right) width$

where L_1 = lower boundary of the median interval

N = No. of values in the data set.

$(\Sigma freq)_L$ = sum of frequencies of all intervals that are lower than the median interval.

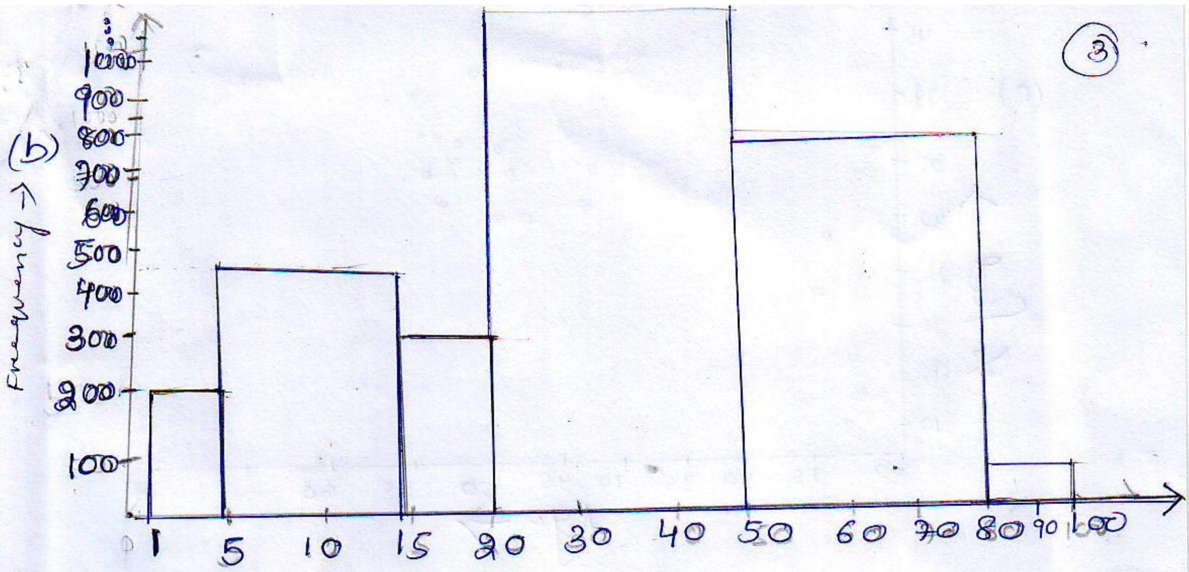
$L = 20$, $(\Sigma freq)_L = 950$ width $= 30$

$N = 3194$ $freq_{median} = 1500$

median interval = 20-50.

median = 32.94 years.

(b)



[Sample Histogram]

Q.5

(a) Age

mean = 46.44

median = 51

Standard deviation = 12.85

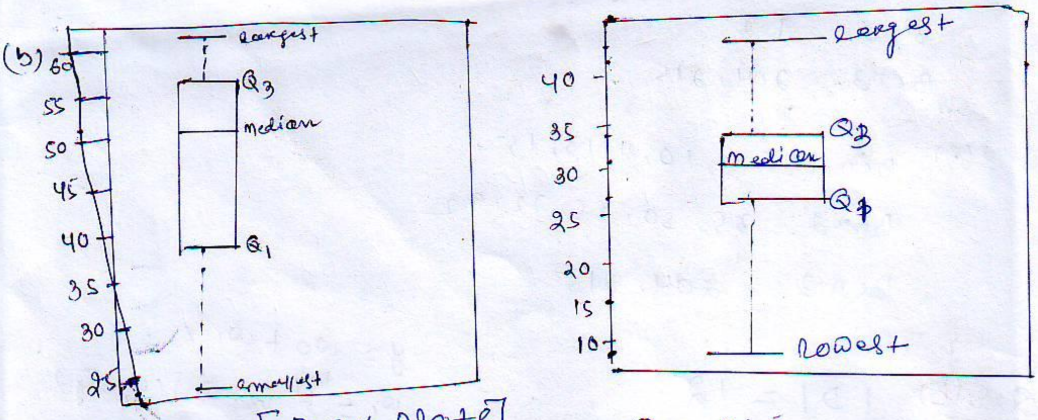
% fat

mean = 28.78

median = 30.7

standard deviation = 8.79

(b)



[BOX Plots]

$Q_1 = 39$

$Q_3 = 57$

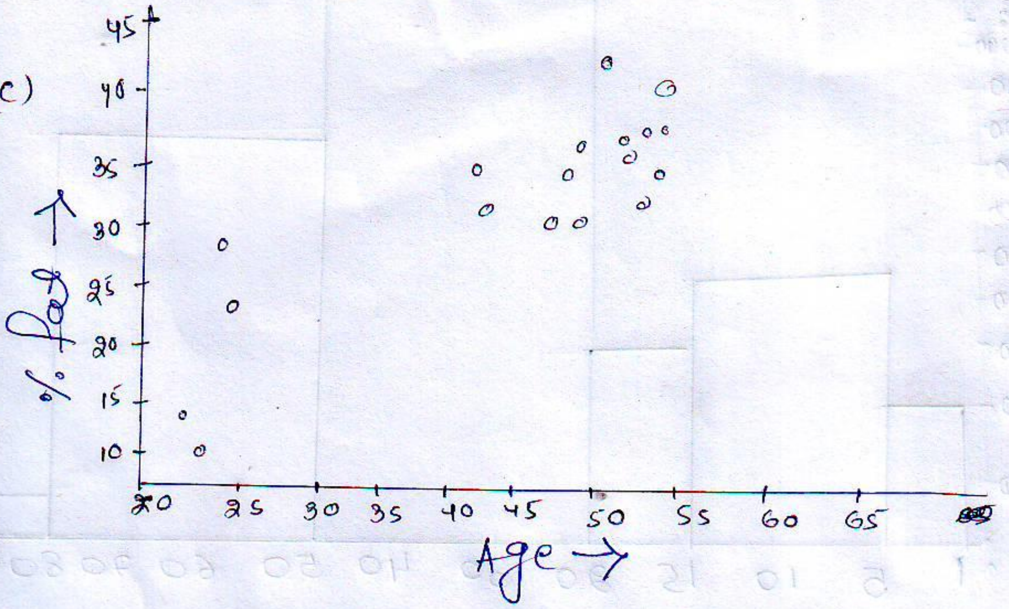
$IQR = 1.5(Q_3 - Q_1) = 27$

$Q_1 = 26.5$

$Q_3 = 34.6$

$IQR = 1.5(Q_3 - Q_1) = 12.15$

(c)



Q.6 (a) Bin 1: 5, 10, 11, 13

Bin 2: 15, 35, 50, 55

Bin 3: 72, 92, 204, 215

(b) width $= \frac{(215-5)}{3} = 70$

Bin1: 5, 10, 11, 13, 15, 35, 50, 55, 72

Bin2: 92

Bin3: 204, 215

(c) Bin 1: 5, 10, 11, 13, 15

Bin2: 35, 50, 55, 72, 92

Bin 3: 204, 215

$y = w_0 + w_1 x$.

Q.8(b) $|D| = 12$.

$$w_1 = \frac{\sum_{i=1}^{|D|} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{|D|} (x_i - \bar{x})^2}$$

$$\bar{x} = \frac{866}{12} = 72.167$$

$$w_0 = \bar{y} - w_1 \bar{x}$$

$$\bar{y} = \frac{888}{12} = 74$$

$w_1 = 0.5816$ $y = 32.028 + 0.5816\,x$

$w_0 = 32.028$ $y = 32.028 + 0.5816(86)$

$= 82.045$

Q.2(a) since 1960's we were basically doing data collection and creation of database for primitive file processing or day-to-day transaction.

The research and development in database systems started since 1970's. During this period relational database systems were developed. users had flexible data access through query language (SQL), user interfaces, optimized query processing and transaction management.

efficient methods for online transaction processing (OLTP), where a query is viewed as a read-only transaction, have developed which helped in efficient storage retrieval and management of large amount of data.

In mid-1980's advanced data models and application oriented database systems, such as spatial, multimedia, active, stream and sensor, scientific and engineering databases, knowledge bases are come into picture.

During late 1980's the technique of data mining and data warehouse come into use.

(b) The steps involved in data mining when viewed as a process of knowledge discovery are

(1) Data cleaning → a process that removes noise and inconsistent data

(2) Data Integration → where multiple data sources may be combined.

(3) Data selection → where data relevant to the analysis task are retrieved from the database.

(4) Data transformation → where data are transformed into forms, appropriate for mining.

(5) Data mining → an essential process where intelligent and efficient methods are applied in order to extract patterns.

(6) pattern evaluation → a process that identifies the truly interesting patterns representing knowledge base.

(7) Knowledge presentation → where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

7(a) Attribute subset selection, reduces the data size by removing irrelevant or redundant attributes or dimensions.

It finds a minimum set of attributes that can be able to represent the original attributes approximately, so that it will be easier to understand a pattern.

mostly heuristic methods are used that reduce the search space for attribute subset selection.

various heuristic methods include the following techniques for attribute subset selection

(1) Stepwise forward selection → It starts with empty attribute sets. when best is determined it is added to attribute set and so on.

(2) Stepwise backward elimination → It starts with the full set of attributes. The worst one is deleted and so on.

(3) combination of backward elimination and forward selection.

(4) Decision tree induction →

(i) In decision tree induction, a tree is constructed from the given data.

(ii) set of attributes not appearing in tree are all irrelevant.

(iii) set of attributes appearing from the reduced subset of attributes.

(b) Curse of dimensionality means, when dimension
increases, database becomes increasingly
sparse. It becomes very complex for applying
various data mining tools on them. Density and
distance between points, which is critical
to clustering, outlier analysis, becomes
less meaningful. The possible combinations
of subspaces will grow exponentially.
~~the more number of~~

In dimensionality reduction, data are reduced
~~or~~ compressed by applying data encoding
or transformation techniques.
Ex:- PCA, FA(Factor analysis), DCT (Discrete
cosine transform), DWT (Discrete wavelet
transform)
~~Brief~~ Write briefly about PCA, DCT and
DWT etc. Explain how these techniques are
used for dimensionality reduction.

8 (a) K- nearest neighbour classification (KNN)
The idea in K-NN method is to identify K-samples
on the training set, which are closest to the testing
sample and then use these K samples to classify
the new sample. ~~into a class~~. It is a Lazy
learner.
Here we require the following steps:
(1) An integer K.
(2) A set of training data. (tuples with class label)
(3) A distance measure to find K-nearest
neighbours to the testing sample, whose
class label is unknown.
(4) After getting K-neighbours using Euclidean
distance, we assign that class label to the
testing sample to which maximum number of
neighbours belong to.