(Q.3) Explain KDD process & Different Data mining tech.

→ - Data mining is a synonym for another popularly used term 'KDD (Knowledge Discovery from Data', while other view Data mining as merely as an essential step in process of knowledge Discovery.

- KDD Process —

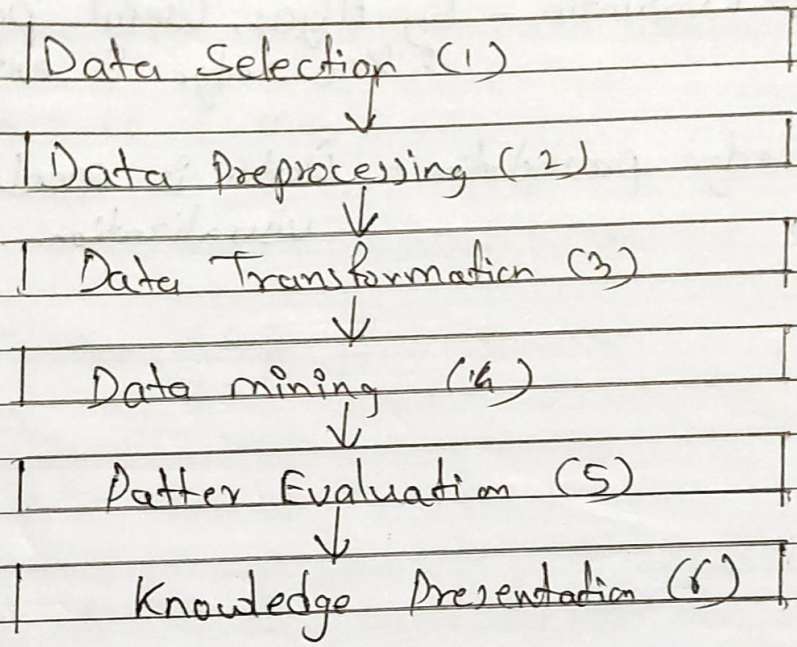

Fig. KDD Process

- Data Selection : Relevant Data is selected from various Sources.

- Data Preprocessing : Data is made into a consistent state.
Unnecessary information is Removed.

- Data Transformation - We convert the Data into Req. format.

- Data mining - We use various Data mining Algo for pattern finding.

- Patter Evaluation - By Algos, Useful pattern & knowledge is extracted.

- knowledge presentation - Data is made for visualization.

Q.4) Suppose that the Data for Analysis includes the attribute age. Age values for the Data tuples are: 13, 15, 16, 16, 19, 20, 23, 29, 35, 41, 44, 53, 62, 69, 72. Use min-max normalization to transform the value 45 for age onto the range 0 to 1.

⇒

$$V_i' = \frac{V_i - \min_A}{\max_A - \min_A (new\_max_A - new\_min_A) + new\_min_A}$$

$min_A = 13$ & $Max_A = 72$

$new\_min_A = 0$ & $new\_max_A = 1$ – (Given)

$V_i = 45$ .. (Given age to be Transformed)

$$V_i' = \frac{(45-13)}{(72-13) \times (1-0) + 0}$$

$$V_i' = 0.542$$

∴ Age # 45 is transformed to 0.542, within the range 0 to 1.

Q.5) Diff betn Classification & Prediction.

→

Classification	Prediction.
i) Classification is a major type of prediction problem where classification is used to predict or nominal values.	i) Prediction can be viewed as the construction & use of a model to assess the class of an unlabeled sample.
ii) Classification is the use of prediction to predict class.	ii) It is used to access the values or value ranges of an attribute that a given sample is likely to have.
iii) Eg. Group patients based on their known medical data & treatment outcome then it's a classification.	iii) Eg. If a classification model is used to predict the treatment for a new patient, then it would be a prediction.
iv) Eg use cases are like Spam email Detection, Disease Diagnosis.	v) Eg. use cases are like Stock prediction, Sales forcasting.

Q.6) Explain Accuracy & error measure for classifiers.

→ - Accuracy of a classifier M, acc(M) is the percentage of test set tuples that are correctly classified By model M.

A) Partition the Data randomly into 3 set — Training set, validation set & test set.

- Training set is the subset of data used to train / build the model.
- Test set is a set of instances that have not been used in the training process. The model's performance is evaluated on unseen Data. Testing just estimates the probability of success on ~~unknown~~ unknown data.
- Validation data is used for parameter tuning but it can't be the test Data. Validation data can be the training data, or a subset of training data.
- General Error : Model error on the test Data.

B) Success :- Instance (record) class is predicted correctly.
C) Error :- Instance class is predicted correctly.
D) The confusion matrix :- It is a useful tool for analyzing how well your classifier can recognize tuples of different classes.

If we have only 2 way ~~da~~ classification then only four classification outcomes are possible —

		Predicted Class		
	Class Label	C_1	C_2	Total
Actual class	C_1	True Positives (TP)	False Negatives (FN)	P
	C_2	False Positives (FP)	True Negatives (TN)	N
	Total	P'	N'	All

TP - Class members which are classified as class member.

TN - Class non-members which are classified as non-members

FP - Class non-members which are classified as class members

FN - class members which are classified as class non-members.

P - No. of positive tuples

N - No. of Negative tuples

P' - No. of tuples that were labelled as positive.

N' - No. of tuples that were labelled as negative.

All - Total no. of tuples.

Sensitivity : TP/P

Specificity : TN/N

Classifier Accuracy $= (TP+TN)/(P+N)$

Error Rate $= \frac{FP+FN}{P+N}$

Precision $= |TP|/(|TP|+|FP|)$

Recall $= |TP|/(|TP|+|FN|)$

F_1/F-Score $= \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$