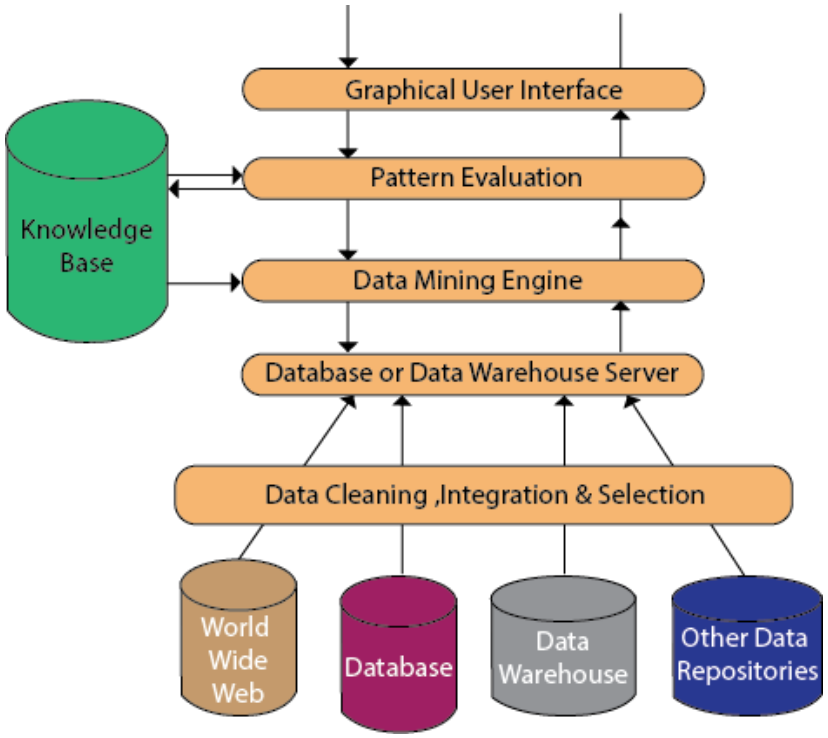
Q) Explain Data Mining Architecture.



Data Source:

The actual source of data is the Database, data warehouse, World Wide Web (WWW), text files, and other documents. You need a huge amount of historical data for data mining to be successful. Organizations typically store data in databases or data warehouses.

Different processes:

Before passing the data to the database or data warehouse server, the data must be cleaned, integrated, and selected. As the information comes from various sources and in different formats, it can't be used directly for the data mining procedure because the data may not be complete and accurate. So, the first data requires to be cleaned and unified.

Database or Data Warehouse Server:

The database or data warehouse server consists of the original data that is ready to be processed. Hence, the server is cause for retrieving the relevant data that is based on data mining as per user request.

Data Mining Engine:

The data mining engine is a major component of any data mining system. In other words, we can say data mining is the root of our data mining architecture. It comprises instruments and software used to obtain insights and knowledge from data collected from various data sources and stored within the data warehouse.

Pattern Evaluation Module:

The Pattern evaluation module is primarily responsible for the measure of investigation of the pattern by using a threshold value. It collaborates with the data mining engine to focus the search on exciting patterns.

Graphical User Interface:

The graphical user interface (GUI) module communicates between the data mining system and the user. This module helps the user to easily and efficiently use the system without knowing the complexity of the process

Knowledge Base:

The knowledge base is helpful in the entire process of data mining. It might be helpful to guide the search or evaluate the stake of the result patterns. The knowledge base may even contain user views and data from user experiences that might be helpful in the data mining process.

Q) Different data Mining Functionalities?

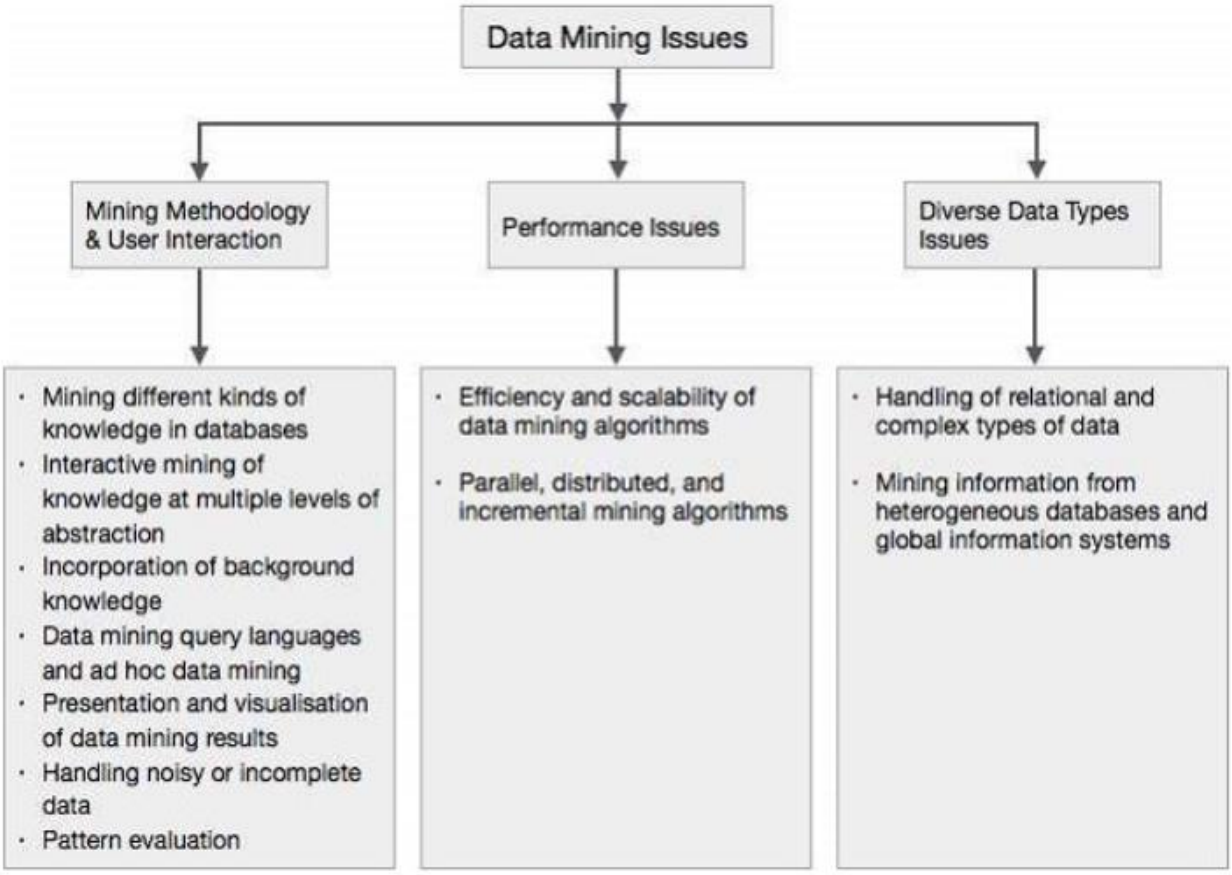
- **Data generalization** − It is a summarization of the general characteristics of an object class of data. The data corresponding to the user-specified class is generally collected by a database query. The output of data characterization can be presented in multiple forms.

- **Association Analysis** − It analyses the set of items that generally occur together in a transactional dataset. There are two parameters that are used for determining the association rules −
    - It provides which identifies the common item set in the database.
    - Confidence is the conditional probability that an item occurs in a transaction when another item occurs.
- **Classification** − Classification is the procedure of discovering a model that represents and distinguishes data classes or concepts, for the objective of being able to use the model to predict the class of objects whose class label is anonymous

- **Clustering** − It is similar to classification but the classes are not predefined. The classes are represented by data attributes. It is unsupervised learning.

- **Outlier analysis** − Outliers are data elements that cannot be grouped in a given class or cluster. These are the data objects which have multiple behavior from the general behavior of other data objects. The analysis of this type of data can be essential to mine the knowledge.

Q) What are the issue in the data mining?



Mining Methodology and User Interaction Issues

- **Mining different kinds of knowledge in databases** − Different users may be interested in different kinds of knowledge. Therefore it is necessary for data mining to cover a broad range of knowledge discovery task.

- **Interactive mining of knowledge at multiple levels of abstraction** − The data mining process needs to be interactive because it allows users to focus the search for patterns, providing and refining data mining requests based on the returned results.

- **Incorporation of background knowledge** − To guide discovery process and to express the discovered patterns, the background knowledge can be used.

- **Data mining query languages and ad hoc data mining** − Data Mining Query language that allows the user to describe ad hoc mining tasks, should be integrated with a data warehouse for flexible data mining.

- **Presentation and visualization of data mining results** − Once the patterns are discovered it needs to be expressed in high level languages, and visual representations. These representations should be easily understandable.

- **Handling noisy or incomplete data** − The data cleaning methods are required to handle the noise and incomplete objects while mining the data regularities. If the data cleaning methods are not there then the accuracy of the discovered patterns will be poor.

- **Pattern evaluation** − The patterns discovered should be interesting because either they represent common knowledge or lack novelty.

Performance Issues

- **Efficiency and scalability of data mining algorithms** − In order to effectively extract the information from huge amount of data in databases, data mining algorithm must be efficient and scalable.

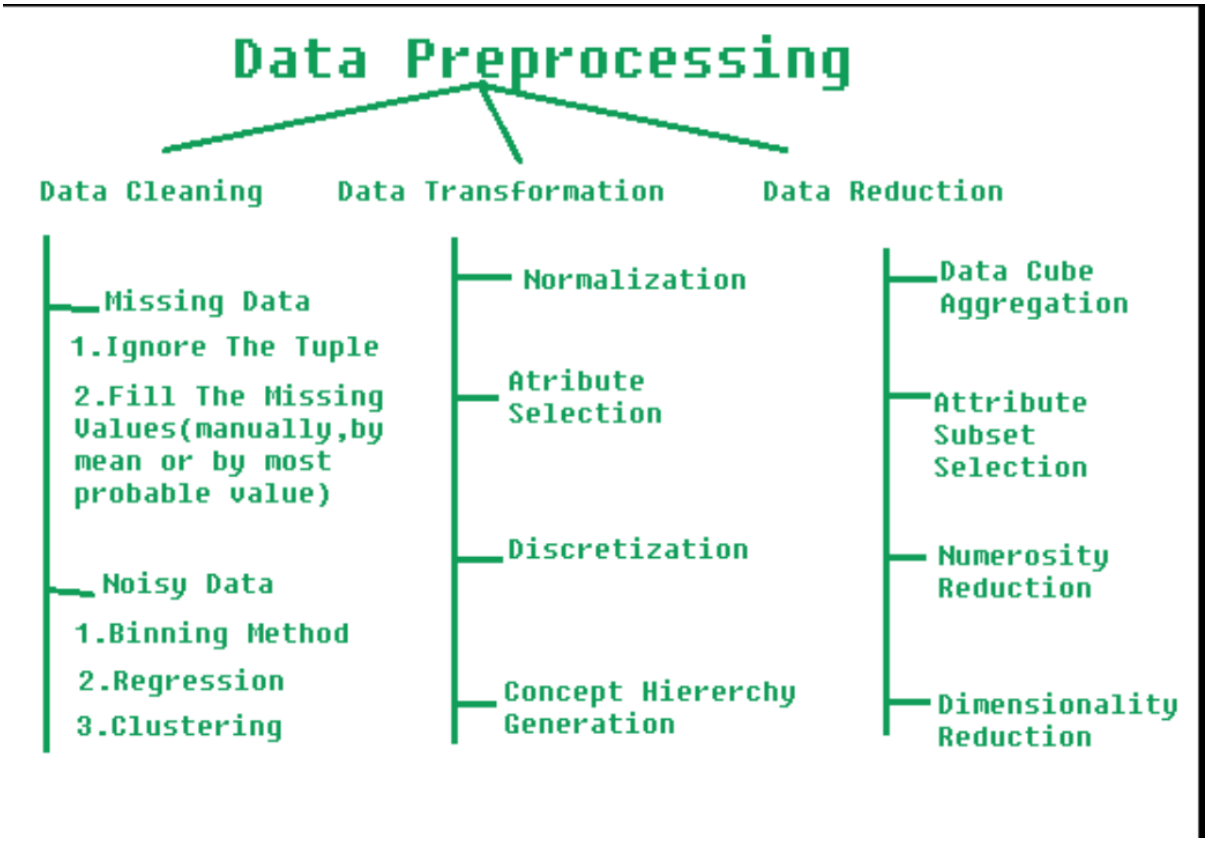
- **Parallel, distributed, and incremental mining algorithms** − The factors such as huge size of databases, and complexity of data mining methods motivate the development of parallel and distributed data mining algorithms.

Diverse Data Types Issues

- **Handling of relational and complex types of data** − The database may contain complex data objects. It is not possible for one system to mine all these kind of data.

- **Mining information from heterogeneous databases and global information systems** − The data is available at different data sources on LAN or WAN. Therefore mining the knowledge from them adds challenges to data mining.

Q) what is Data Preprocessing?



Q) Explain various methods to handle noisy and ,missing values in data mining

Noisy data is a meaningless data that can't be interpreted by machines. It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways:

Binning Method:

This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.

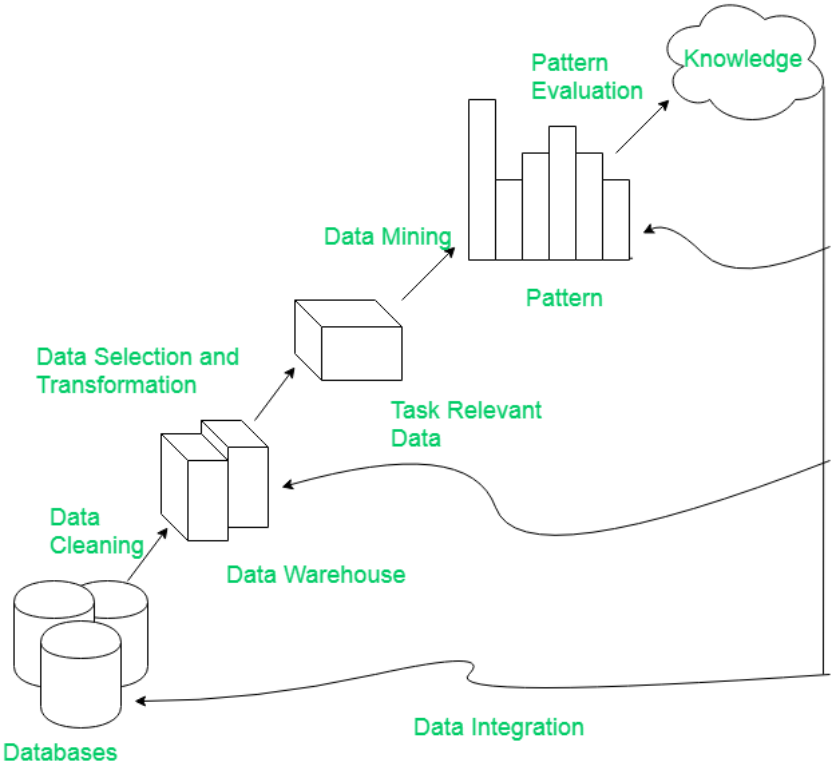
Regression:

Here data can be made smooth by fitting it to a regression function. The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).

Clustering:

This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

Q) Knowledge Discovery Process (KDD Process)



1. **Data Cleaning**: Data cleaning is defined as removal of noisy and irrelevant data from collection.
2. **Data Integration**: Data integration is defined as heterogeneous data from multiple sources combined in a common source (DataWarehouse).
3. **Data Selection**: Data selection is defined as the process where data relevant to the analysis is decided and retrieved from the data collection.
4. **Data Transformation**: Data Transformation is defined as the process of transforming data into appropriate form required by mining procedure.
    i. Data Transformation is a two step process:
    b. **Data Mapping**: Assigning elements from source base to destination to capture transformations.
    c. **Code generation**: Creation of the actual transformation program.
5. **Data Mining**: Data mining is defined as clever techniques that are applied to extract patterns potentially useful.

6. **Pattern Evaluation**: Pattern Evaluation is defined as identifying strictly increasing patterns representing knowledge based on given measures.
7. **Knowledge representation**: Knowledge representation is defined as technique which utilizes visualization tools to represent data mining results.

Disclaimer: - Read at your own risk

Q1) Explain Support Vector Machine

- Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems.

- The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

- SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

Types of SVM

SVM can be of two types:

- **Linear SVM:** Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.

- **Non-linear SVM:** Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

Q2) Explain Backpropagation ?

Backpropagation, short for “backward propagation of errors”, is a mechanism used to update the weights using gradient descent. It calculates the gradient of the error function with respect to the neural network’s weights. The calculation proceeds backwards through the network.

Backpropagation is “backpropagation of errors” and is very useful for training neural networks. It’s fast, easy to implement, and simple. Backpropagation does not require any parameters to be set, except the number of inputs. Backpropagation is a flexible method because no prior knowledge of the network is required.

Backpropagation Algorithm:

Step 1: Inputs X, arrive through the preconnected path.

Step 2: The input is modeled using true weights W. Weights are usually chosen randomly.

Step 3: Calculate the output of each neuron from the input layer to the hidden layer to the output layer.

Step 4: Calculate the error in the outputs

Backpropagation Error= Actual Output – Desired Output

Step 5: From the output layer, go back to the hidden layer to adjust the weights to reduce the error.

Step 6: Repeat the process until the desired output is achieved.

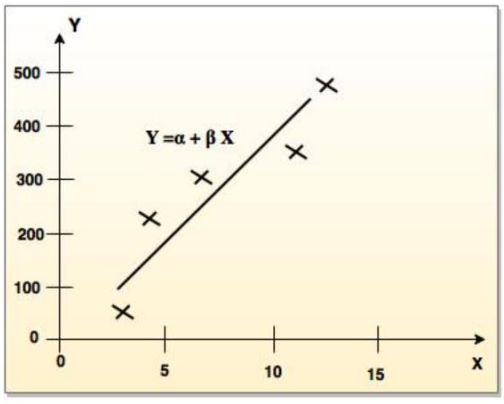
Q3) Explain Linear and Non-linear Regression in detail.

Regression refers to a type of supervised machine learning technique that is used to predict any continuous-valued attribute. Regression helps any business organization to analyze the target variable and predictor variable relationships. It is a most significant tool to analyze the data that can be used for financial forecasting and time series modeling.

1. Linear regression

- It is simplest form of regression. Linear regression attempts to model the relationship between two variables by fitting a linear equation to observe the data.
- Linear regression attempts to find the mathematical relationship between variables.
- If outcome is straight line then it is considered as linear model and if it is curved line, then it is a non linear model.
- The relationship between dependent variable is given by straight line and it has only one independent variable.

  Y = α + B X
- Model **'Y'**, is a linear function of **'X'**.
- The value of 'Y' increases or decreases in linear manner according to which the value of 'X' also changes.



Linear Regression

2.Non-Linear Regression

The term “nonlinear” refers to the parameters in the model, as opposed to the independent variables. Unlimited possibilities exist for describing the deterministic part of the model. Such flexibility provides a good ground on which to make statistical inferences.

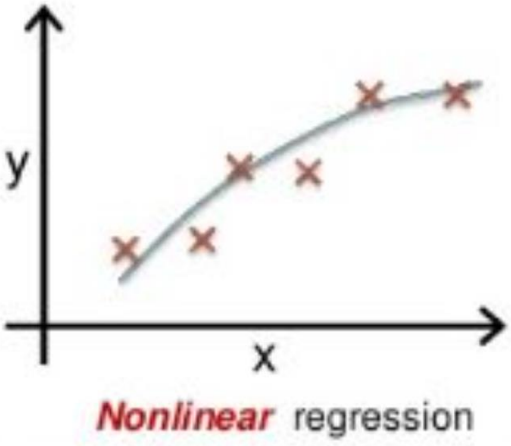
The goal of the model is to minimize the sum of the squares as least as possible using iterative numeric procedures.

A simple nonlinear regression model is expressed as follows:

$$Y = f(X,\beta) + \epsilon$$

Where:

- **X** is a vector of P predictors
- **β** is a vector of k parameters
- **F** (-) is the known regression function
- **ϵ** is the error term



Q) What is Classification?

Classification: It is a data analysis task, i.e. the process of finding a model that describes and distinguishes data classes and concepts.

Classification is the problem of identifying to which of a set of categories (subpopulations), a new observation belongs to, on the basis of a training set of data containing observations and whose categories membership is known.

Following is the examples of Classification –

- A bank loan officer wants to analyze the data in order to know which customer (loan applicant) are risky or which are safe.

In above examples, a model or classifier is constructed to predict the categorical labels. These labels are risky or safe for loan application data .

Q) What is Prediction?

To find a numerical output, prediction is used. The training dataset contains the inputs and numerical output values. According to the training dataset, the algorithm generates a model or predictor. When fresh data is provided, the model should find a numerical output. This approach, unlike classification, does not have a class label. A continuous-valued function or ordered value is predicted by the model.

Consider the following scenario: A marketing manager needs to forecast how much a specific consumer will spend during a sale. In this scenario, we are bothered to forecast a numerical value. In this situation, a model or predictor that forecasts a continuous or ordered value function will be built.

Q) Issue of Prediction?

Data Cleaning: Cleaning data include reducing noise and treating missing values. Smoothing techniques remove noise, and the problem of missing values is solved by replacing a missing value with the most often occurring value for that characteristic.

Relevance Analysis: The irrelevant attributes may also be present in the database. The correlation analysis method is used to determine whether two attributes are connected.

Data Transformation and Reduction: Any of the methods listed below can be used to transform the data.

- **Normalization:** Normalization is used to transform the data. Normalization is the process of scaling all values for a given attribute so that they lie within a narrow range. When neural networks or methods requiring measurements are utilized in the learning process, normalization is performed.
- **Generalization:** The data can also be modified by applying a higher idea to it. We can use the concept of hierarchies for this.

Q) Difference between Classification and Prediction?

Classification	Prediction
Classification is the process of identifying which category a new observation belongs to based on a training data set containing observations whose category membership is known.	Predication is the process of identifying the missing or unavailable numerical data for a new observation.
In classification, the accuracy depends on finding the class label correctly.	In prediction, the accuracy depends on how well a given predictor can guess the value of a predicated attribute for new data.
In classification, the model can be known as the classifier.	In prediction, the model can be known as the predictor.
A model or the classifier is constructed to find the categorical labels.	A model or a predictor will be constructed that predicts a continuous-valued function or ordered value.
For example, the grouping of patients based on their medical records can be considered a classification.	**For example**, We can think of prediction as predicting the correct treatment for a particular disease for a person.

Q) Explain Ensemble methods.

• Construct a set of base classifiers learned from the training data

• Predict class label of test records by combining the predictions made by multiple classifiers (e.g., by taking majority vote)

Bagging and Boosting are the two different types of ensemble method

1. Bagging: It is a homogeneous weak learners' model that learns from each other independently in parallel and combines them for determining the model average.

→ Bootstrap Aggregating, also known as bagging, is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It decreases the variance and helps to avoid overfitting. It is usually applied to decision tree methods. Bagging is a special case of the model averaging approach.

2. Boosting: It is also a homogeneous weak learners' model but works differently from Bagging. In this model, learners learn sequentially and adaptively to improve model predictions of a learning algorithm.

→Boosting is an ensemble modeling technique that attempts to build a strong classifier from the number of weak classifiers. It is done by building a model by using weak models in series. Firstly, a model is built from the training data. Then the second model is built which tries to correct the errors

present in the first model. This procedure is continued and models are added until either the complete training data set is predicted correctly or the maximum number of models is added

Q) Bagging vs Boosting

S.NO	Bagging	Boosting
1.	The simplest way of combining predictions that belong to the same type.	A way of combining predictions that belong to the different types.
2.	Aim to decrease variance, not bias.	Aim to decrease bias, not variance.
3.	Each model receives equal weight.	Models are weighted according to their performance.
4.	Each model is built independently.	New models are influenced by the performance of previously built models.
5.	Different training data subsets are selected using row sampling with replacement and random sampling methods from the entire training dataset.	Every new subset contains the elements that were misclassified by previous models.
6.	Bagging tries to solve the over-fitting problem.	Boosting tries to reduce bias.
7.	If the classifier is unstable (high variance), then apply bagging.	If the classifier is stable and simple (high bias) the apply boosting.
8.	In this base classifiers are trained parallelly.	In this base classifiers are trained sequentially.
9	Example: The Random forest model uses Bagging.	Example: The AdaBoost uses Boosting techniques

Q) Decision Tree algo

Day	Outlook	Temp	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Q) Backpropagation method

Q) Rue Base classification