1. What is the concept of supervised learning? What is the significance of the name?

Supervised learning involves training a model on labeled data, where the output is known. The model learns to predict outputs from inputs. The term "supervised" indicates that the learning process is guided by known outcomes.

2. In the hospital sector, offer an example of supervised learning.

Example: Predicting patient disease based on medical history and test results, where the model is trained on labeled datasets of past patient records.

3. Give three supervised learning examples.

1. Email spam detection.

2. Predicting house prices.

3. Diagnosing diseases from medical images.

4. In supervised learning, what are classification and regression?

Classification assigns inputs to discrete categories (e.g., spam or not spam). Regression predicts continuous values (e.g., house prices).

5. Give some popular classification algorithms as examples.

- Logistic Regression

- Decision Trees

- Support Vector Machines (SVM)

- k-Nearest Neighbors (kNN)

6. Briefly describe the SVM model.

Support Vector Machine (SVM) is a supervised learning algorithm that finds the optimal hyperplane which maximizes the margin between different classes.

7. In SVM, what is the cost of misclassification?

Cost of misclassification refers to the penalty imposed on the model for incorrectly classifying data points, influencing the trade-off between a large margin and classification errors.

8. In the SVM model, define Support Vectors.

Support Vectors are data points that lie closest to the hyperplane and influence its position and orientation, critical for defining the optimal boundary.

9. In the SVM model, define the kernel.

Kernel is a function that transforms data into a higher dimension, allowing SVM to find a linear separator in transformed space for non-linear problems.

10. What are the factors that influence SVM's effectiveness?

- Choice of kernel

- Regularization parameter (C)

- Margin width

- Data preprocessing

11. What are the benefits of using the SVM model?

- Effective in high-dimensional spaces

- Robust to overfitting (with proper regularization)

- Works well with a clear margin of separation

12. What are the drawbacks of using the SVM model?

- Inefficient with large datasets

- Sensitive to the choice of kernel and parameters

- Computationally intensive for training

13. Notes should be written on:

1. The kNN algorithm has a validation flaw: kNN is sensitive to the choice of k and distance metric; performance can degrade with irrelevant features.

2. In the kNN algorithm, the k value is chosen: k is usually selected via cross-validation to balance bias-variance trade-off.

3. A decision tree with inductive bias: Decision trees assume that data can be split based on feature values, preferring shorter, simpler trees.

14. What are some of the benefits of the kNN algorithm?

- Simple and easy to implement

- No training phase

- Flexible to different distance metrics

15. What are some of the kNN algorithm's drawbacks?

- Computationally expensive during prediction

- Sensitive to irrelevant features and noise

- Requires large memory for storing data

16. Explain the decision tree algorithm in a few words.

Decision tree uses a tree-like model of decisions and their possible consequences, splitting data into subsets based on feature values.

17. What is the difference between a node and a leaf in a decision tree?

Node represents a feature or attribute decision point. Leaf represents the outcome or class label after all decisions are made.

18. What is a decision tree's entropy?

Entropy measures the impurity or disorder in a dataset, used to decide the best feature splits to minimize uncertainty.

19. In a decision tree, define knowledge gain.

Knowledge gain (information gain) is the reduction in entropy after a dataset is split on an attribute, indicating the attribute's effectiveness in classifying data.

20. Choose three advantages of the decision tree approach and write them down.

1. Easy to understand and interpret.

2. Handles both numerical and categorical data.

3. Requires little data preprocessing.

21. Make a list of three flaws in the decision tree process.

1. Prone to overfitting.

2. Sensitive to noisy data.

3. Can create biased trees if some classes dominate.

22. Briefly describe the random forest model.

Random forest is an ensemble method that builds multiple decision trees using random subsets of data and features, then aggregates their predictions to improve accuracy and control overfitting.