Q1. What is the concept of supervised learning? What is the significance of the name?

Supervised learning is a machine learning paradigm where the algorithm learns from labeled training data. It involves training the model on a dataset where both input features and corresponding target labels are provided.

The significance of the name "supervised" comes from the fact that the process is guided by a supervisor or teacher in the form of the provided labeled data.

Q2. In the hospital sector, offer an example of supervised learning.

In the hospital sector, an example of supervised learning could be predicting whether a patient has a certain medical condition based on their symptoms, medical history, and test results. The historical data with labeled outcomes (presence or absence of the condition) would be used to train a supervised learning algorithm.

Q3. Give three supervised learning examples.

**Examples of Supervised Learning:**

1. Predicting house prices based on features like location, size, and amenities.
2. Email spam detection, where the algorithm classifies emails as spam or not spam based on their content.
3. Handwritten digit recognition, where the algorithm identifies the digits written in images.

Q4. In supervised learning, what are classification and regression?

**Classification and Regression:** In supervised learning, classification is the task of categorizing input data into predefined classes or categories. Regression, on the other hand, involves predicting a continuous numerical output based on input features.

Q5. Give some popular classification algorithms as examples.

**Popular Classification Algorithms:**

* Decision Trees
* Random Forest
* Support Vector Machines (SVM)
* k-Nearest Neighbors (kNN)
* Naïve Bayes
* Logistic Regression

Q6. Briefly describe the SVM model.

**Support Vector Machine (SVM) Model:** SVM is a powerful supervised learning algorithm used for classification and regression tasks. It finds a hyperplane that best separates different classes in feature space while maximizing the margin between them.

Q7. In SVM, what is the cost of misclassification?

**Cost of Misclassification in SVM:** The cost of misclassification in SVM refers to the penalty assigned to the misclassification of data points. It determines the trade-off between achieving a wider margin and correctly classifying data points. Different types of misclassifications can have different costs associated with them.

Q8. In the SVM model, define Support Vectors.

**Support Vectors in SVM:** Support vectors are the data points closest to the hyperplane that determine its position and orientation. They are the critical data points that have the most influence on the separation of classes.

Q9. In the SVM model, define the kernel.

**Kernel in SVM:** In SVM, a kernel is a function that transforms the input data into a higher-dimensional space. This transformation allows SVM to find a hyperplane that might not be linear in the original feature space but can effectively separate classes in the transformed space.

Q10. What are the factors that influence SVM's effectiveness?

**Factors Influencing SVM's Effectiveness:**

* Choice of kernel function
* Selection of hyperparameters (e.g., regularization parameter)
* Handling imbalanced data
* Proper scaling of features
* Handling noisy data

Q11. What are the benefits of using the SVM model?

**Benefits of Using SVM:**

* Effective in high-dimensional spaces.
* Works well with both linear and non-linear data.
* Robust against overfitting, especially with proper tuning.

Q12. What are the drawbacks of using the SVM model?

**Drawbacks of Using SVM:**

* Computationally intensive, especially for large datasets.
* Requires careful selection of hyperparameters.
* Interpreting the model can be complex.

Q13. Notes should be written on

* The kNN algorithm has a validation flaw.
* In the kNN algorithm, the k value is chosen.
* A decision tree with inductive bias

kNN algorithm's validation flaw: It's sensitive to the choice of k and the distance metric, which can impact its performance.

Choosing k value: The value of k affects the algorithm's sensitivity to noise and data distribution.

Decision tree with inductive bias: Decision trees inherently have a bias towards selecting features that offer the best split to separate classes.

Q14. What are some of the benefits of the kNN algorithm?

**Benefits of kNN Algorithm:**

* Simple to understand and implement.
* Can handle multi-class classification.
* Robust against noisy data.

Q15. What are some of the kNN algorithm's drawbacks?

**Drawbacks of kNN Algorithm:**

* Sensitive to feature scaling and irrelevant features.
* Computationally expensive for large datasets.
* Decision boundary can be sensitive to small fluctuations in data.

Q16. Explain the decision tree algorithm in a few words.

**Decision Tree Algorithm:** A decision tree algorithm recursively splits the dataset into subsets based on features, aiming to create subsets with more homogeneous target labels. This process results in a tree-like structure where each internal node represents a decision based on a feature, and each leaf node represents a class label.

Q17. What is the difference between a node and a leaf in a decision tree?

**Difference between Node and Leaf in Decision Tree:**

* Node: Represents a decision point where data is split based on a feature.
* Leaf: Represents the final predicted class label.

Q18. What is a decision tree's entropy?

**Entropy in Decision Tree:** Entropy is a measure of impurity in a node. It's used to determine the quality of a split in a decision tree. A node with lower entropy is more pure and contains more homogeneous class labels.

Q19. In a decision tree, define knowledge gain.

**Knowledge Gain in Decision Tree:** Knowledge gain is the reduction in entropy achieved after splitting a node. It measures how well a split separates classes.

Q20. Choose three advantages of the decision tree approach and write them down.

**Advantages of Decision Tree Approach:**

* Easy to understand and visualize.
* Can handle both categorical and numerical data.
* Requires minimal data preprocessing.

Q21. Make a list of three flaws in the decision tree process.

**Flaws in Decision Tree Approach:**

* Prone to overfitting, especially with deep trees.
* Sensitive to small variations in data.
* Not suitable for capturing complex relationships in data.

Q22. Briefly describe the random forest model.

**Random Forest Model:** Random Forest is an ensemble learning technique that combines multiple decision trees to improve predictive accuracy and control overfitting. It randomly selects subsets of features and data for each tree and then combines their predictions for better results. It's widely used for classification and regression tasks.