QUESTION 1: What is the concept of supervised learning? What is the significance of the name?

ANSWER: Supervised learning is a machine learning paradigm where the model is trained on a labeled dataset, meaning the input data is paired with the correct output. The significance of the name "supervised" comes from the fact that during training, the algorithm is supervised or guided by providing it with the correct answers.

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

ANSWER : An example of supervised learning in the hospital sector could be predicting whether a patient has a certain disease based on their symptoms and medical history. The model is trained on past data where patients' symptoms and diagnoses are known, allowing it to learn patterns and make predictions for new patients.

3. Give three supervised learning examples.

ANSWER: **Three Supervised Learning Examples:**

* Email spam detection
* Handwritten digit recognition
* Stock price prediction

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

ANSWER:

* **Classification:** Involves predicting a categorical label. For example, classifying emails as spam or not spam.
* **Regression:** Involves predicting a continuous value. For example, predicting the price of a house based on its features.

5. Give some popular classification algorithms as examples.

ANSWER: **Popular Classification Algorithms:**

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

6. Briefly describe the SVM model.

ANSWER: SVM is a supervised learning algorithm used for classification and regression tasks. It finds the optimal hyperplane that best separates the classes in the feature space.

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

ANSWER: The cost of misclassification in SVM refers to the penalty associated with misclassifying data points. It is typically a parameter that can be adjusted to control the trade-off between maximizing the margin and minimizing misclassification.

8. In the SVM model, define Support Vectors.

ANSWER: Support vectors are the data points that lie closest to the decision boundary (hyperplane) between classes. These points are crucial in defining the decision boundary and maximizing the margin in SVM.

9. In the SVM model, define the kernel.

ANSWER: A kernel in SVM is a function that takes low-dimensional input space and transforms it into a higher-dimensional space, making it easier to classify the data nonlinearly.

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

ANSWER: **Factors Influencing SVM's Effectiveness:**

* Choice of kernel
* Regularization parameter (C)
* Kernel parameters (e.g., gamma for Gaussian kernel)
* Handling of imbalanced data
* Feature scaling

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

ANSWER: **Benefits of SVM Model:**

* Effective in high-dimensional spaces
* Versatile - can handle linear and non-linear decision boundaries
* Robust against overfitting, especially in high-dimensional spaces
* Memory efficient, as it only uses a subset of training points as support vectors

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

ANSWER: **Drawbacks of SVM Model:**

* Choosing an appropriate kernel and parameters can be challenging
* Computationally intensive, especially for large datasets
* Can be sensitive to noisy or overlapping data
* Interpretability can be limited compared to simpler models like logistic regression or decision trees

13. Notes should be written on

ANSWER:

**1. The KNN algorithm has a validation flaw.**

The KNN algorithm has a validation flaw: It tends to perform poorly with high-dimensional data due to the curse of dimensionality.

**2. In the KNN algorithm, the k value is chosen**.

In the KNN algorithm, the k value is chosen: The choice of k significantly impacts the algorithm's performance and is often determined using cross-validation or other validation techniques.

**3. A decision tree with inductive bias**

A decision tree with inductive bias: A decision tree algorithm has an inductive bias towards simpler decision trees, meaning it prefers shorter trees with fewer nodes to avoid overfitting.

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

ANSWER: **Benefits of KNN Algorithm:**

* Simple to understand and implement
* No training period involved
* Naturally handles multiclass problems
* Robust to noisy training data

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

ANSWER: **Drawbacks of KNN Algorithm:**

* Computationally expensive during testing phase
* Sensitive to irrelevant features and the choice of distance metric
* Requires a large amount of memory to store the entire dataset
* Performance can degrade with high-dimensional data

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

ANSWER: Decision tree algorithm recursively splits the dataset into subsets based on the most significant attribute, creating a tree-like structure where each internal node represents a feature, each branch represents a decision based on that feature, and each leaf node represents the outcome or class label.

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

ANSWER:

* **Node:** Represents a decision point based on a feature. It splits the dataset into smaller subsets.
* **Leaf:** Represents the final decision or outcome. It does not have any child nodes.

18. What is a decision tree's entropy?

ANSWER: Entropy measures the impurity or randomness in a dataset. In a decision tree, entropy is used to calculate the information gain when splitting a node. Nodes with low entropy are preferred as they contain more homogeneous classes.

19. In a decision tree, define knowledge gain.

ANSWER: Knowledge gain, also known as information gain, measures the effectiveness of a feature in reducing uncertainty (entropy) about the class labels. It helps decide which feature to split on at each node.

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

ANSWER: **Advantages of Decision Tree Approach:**

* Easy to understand and interpret, suitable for visual representation
* Handles both numerical and categorical data
* Requires minimal data preprocessing (e.g., normalization or scaling)

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

ANSWER: **Flaws in Decision Tree Process:**

* Prone to overfitting, especially with deep trees
* Can be biased towards features with more levels or categories
* Instability - small variations in the data can lead to significantly different trees

22. Briefly describe the random forest model.

ANSWER: Random forest is an ensemble learning method based on decision trees. It builds multiple decision trees during training and outputs the class that is the mode of the classes (classification) or the mean prediction (regression) of the individual trees. Random forest introduces randomness during tree construction to improve generalization and reduce overfitting.