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

Solution - Supervised learning, also known as supervised machine learning, is a subcategory of machine learning and artificial intelligence. It is defined by its use of labeled datasets to train algorithms that to classify data or predict outcomes accurately.

Supervised machine learning turns data into real, actionable insights. It enables organizations to use data to understand and prevent unwanted outcomes or boost desired outcomes for their target variable.

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

Solution - Supervised ML models are used in situations when the outcome of interest is specified, and data is explicitly labeled for the outcome. For example, the outcome may be the presence or absence of a disease like diabetes or hypertension.

3. **Give three supervised learning examples.**

Solution - Image and speech recognition, recommendation systems, and fraud detection are all examples of how supervised learning is used

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

Solution - Regression algorithms are used to determine continuous values such as price, income, age, etc. and Classification algorithms are used to forecast or classify the distinct values such as Real or False, Male or Female, Spam or Not Spam, etc.

5. **Give some popular classification algorithms as examples.**

Solution –

- Logistic Regression.
- Naive Bayes.
- K-Nearest Neighbors.
- Decision Tree Classifier.
- Support Vector Machines Classifier

6. **Briefly describe the SVM model.**

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

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. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:

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

Solution - In Support Vector Machines (SVM), the cost of misclassification refers to the penalty or cost associated with misclassifying data points during the model training process. SVM aims to find the optimal decision boundary that maximally separates different classes of data. The cost of misclassification determines the trade-off between achieving a perfect separation of classes and allowing some misclassified data points.

In SVM, there are two types of misclassifications that can occur:

False Positive (Type I Error): This occurs when a data point from the negative class is incorrectly classified as belonging to the positive class.

False Negative (Type II Error): This occurs when a data point from the positive class is incorrectly classified as belonging to the negative class.

8. **In the SVM model, define Support Vectors.**

Solution - Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, we maximize the margin of the classifier. Deleting the support vectors will change the position of the hyperplane. These are the points that help us build our SVM.

9. **In the SVM model, define the kernel.**

Solution - SVM Kernel Functions

The function of kernel is to take data as input and transform it into the required form. Different SVM algorithms use different types of kernel functions. These functions can be different types. For example linear, nonlinear, polynomial, radial basis function (RBF), and sigmoid.

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

Solution - The effectiveness of SVM depends on the selection of kernel, kernel's parameters and soft margin parameter C. Each pair of parameters is checked using cross validation, and the parameters with best cross validation accuracy are picked.

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

Solution - Advantages of SVM Classifier:

- SVM works relatively well when there is a clear margin of separation between classes.
- SVM is more effective in high dimensional spaces and is relatively memory efficient.
- SVM is effective in cases where the dimensions are greater than the number of samples.

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

Solution - Disadvantages of SVM Classifier:

SVM algorithm is not suitable for large data sets. SVM does not perform very well when the data set has more noise i.e. target classes are overlapping. In cases where the number of features for each data point exceeds the number of training data samples, the SVM will underperform

13. **Notes should be written on**
1. **The kNN algorithm has a validation flaw.**

Solution - Just one drawback with k-fold cross-validation is that we are repeating the computations for each value of K (of KNN). So it basically increases the time complexity.

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

Solution - The k value in the k-NN algorithm defines how many neighbors will be checked to determine the classification of a specific query point. For example, if k=1, the instance will be assigned to the same class as its single nearest neighbor.

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

Solution - In the case of decision trees, the depth of the tress is the inductive bias. If the depth of the tree is too low, then there is too much generalisation in the model.

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

Solution - Some Advantages of KNN

- Quick calculation time.
- Simple algorithm – to interpret.
- Versatile – useful for regression and classification.
- High accuracy – you do not need to compare with better-supervised learning models.

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

Solution - Some Disadvantages of KNN

- Accuracy depends on the quality of the data.
- With large data, the prediction stage might be slow.
- Sensitive to the scale of the data and irrelevant features.
- Require high memory – need to store all of the training data.
- Given that it stores all of the training, it can be computationally expensive.

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

Solution - A decision tree algorithm is a machine learning algorithm that uses a decision tree to make predictions. It follows a tree-like model of decisions and their possible consequences. The algorithm works by recursively splitting the data into subsets based on the most significant feature at each node of the tree.

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

Solution - The root node is just the topmost decision node. In other words, it is where you start traversing the classification tree. The leaf nodes (green), also called terminal nodes, are nodes that don't split into more nodes.

18. **What is a decision tree's entropy?**

Solution - In the context of Decision Trees, entropy is a measure of disorder or impurity in a node. Thus, a node with more variable composition, such as 2Pass and 2 Fail would be considered to have higher Entropy than a node which has only pass or only fail.

19. **In a decision tree, define knowledge gain.**

Solution - The information gained in the decision tree can be defined as the amount of information improved in the nodes before splitting them for making further decisions.

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

Solution - Some advantages of decision trees are:

- Simple to understand and to interpret. ...
- Requires little data preparation. ...
- The cost of using the tree (i.e., predicting data) is logarithmic in the number of data points used to train the tree.
- Able to handle both numerical and categorical data. ...
- Able to handle multi-output problems

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

Solution - Some disadvantages of decision trees are:

- Overfitting can occur.
- Decision trees can be sensitive to small changes in the data.
- They can be biased towards certain outcomes.
- Large decision trees can be hard to interpret.
- They may not work well with certain types of data.

22. **Briefly describe the random forest model.**

Solution – Random Forest is a commonly-used machine learning algorithm, which combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems.