1. **Recognize the differences between supervised, semi-supervised, and unsupervised learning.**

**Supervised machine learning** relies on labelled input and output training data, whereas unsupervised learning processes unlabelled or raw data. **semi-supervised** involves a small portion of labeled examples and a large number of unlabeled examples from which a model must learn and make predictions on new examples.

1. **Describe in detail any five examples of classification problems.**

* Email Spam.
* Handwritten Digit Recognition.
* Image segmentation.
* Speech Recognition.
* DNA Sequence Classification.

**3. Describe each phase of the classification process in detail.**

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

1. **Go through the SVM model in depth using various scenarios.**

Support vector machines are a set of supervised learning methods used for classification, regression, and outliers detection. All of these are common tasks in machine learning.

You can use them to detect cancerous cells based on millions of images or you can use them to predict future driving routes with a well-fitted regression model.

There are specific types of SVMs you can use for particular machine learning problems, like support vector regression (SVR) which is an extension of support vector classification (SVC).

The main thing to keep in mind here is that these are just math equations tuned to give you the most accurate answer possible as quickly as possible.

SVMs are different from other classification algorithms because of the way they choose the decision boundary that maximizes the distance from the nearest data points of all the classes. The decision boundary created by SVMs is called the maximum margin classifier or the maximum margin hyper plane.

1. **What are some of the benefits and drawbacks of SVM?**

* The advantages of SVM and support vector regression include that they can be used to avoid the difficulties of using linear functions in the high-dimensional feature space, and the optimization problem is transformed into dual convex quadratic programs.
* It does not execute very well when the data set has more sound i.e. target classes are overlapping. In cases where the number of properties for each data point outstrips the number of training data specimens, the support vector machine will underperform.

1. **Go over the kNN model in depth.**

The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other.

1. **Discuss the kNN algorithm's error rate and validation error.**

Hence, error rate initially decreases and reaches a minima. After the minima point, it then increase with increasing K. To get the optimal value of K, you can segregate the training and validation from the initial dataset. Now plot the validation error curve to get the optimal value of K.

8. **For kNN, talk about how to measure the difference between the test and training results.**

For evaluating the performance of k-NN we use both accuracy (A) and F-score (F) metric.

9. **Create the kNN algorithm.**

from sklearn.neighbors import KNeighborsClassifier

model = KNeighborsClassifier(n\_neighbors=3)

# Train the model using the training sets

model.fit(features,label)

#Predict Output

predicted= model.predict([[0,2]]) # 0:Overcast, 2:Mild

print(predicted)

**10.What is a decision tree, exactly? What are the various kinds of nodes? Explain all in depth.**

A decision tree typically starts with a single node, which branches into possible outcomes. Each of those outcomes leads to additional nodes, which branch off into other possibilities. This gives it a treelike shape. There are three different types of nodes: chance nodes, decision nodes, and end nodes.

11. **Describe the different ways to scan a decision tree.**

Decision tree analysis can deal with missing data in two ways: it can either classify missing values as a separate category that can be analyzed with the other categories or use a built decision tree model which set the variable with lots of missing value as a target variable to make prediction and replace these.

12. **Describe in depth the decision tree algorithm.**

Decision trees use multiple algorithms to decide to split a node into two or more sub-nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. In other words, we can say that the purity of the node increases with respect to the target variable.

13. **In a decision tree, what is inductive bias? What would you do to stop overfitting?**

Before learning a model given a data and a learning algorithm, there are a few assumptions a learner makes about the algorithm. These assumptions are called the inductive bias. It is like the property of the algorithm.

14.**Explain advantages and disadvantages of using a decision tree?**

* Advantages of Decision Trees. Interpretability. Less Data Preparation. Non-Parametric. Versatility. Non-Linearity.
* Disadvantages of Decision Tree. Overfitting. Feature Reduction & Data Resampling. Optimization.

15. Describe in depth the problems that are suitable for decision tree learning.

Decision tree learning is generally best suited to problems with the following characteristics: Instances are represented by attribute-value pairs. There is a finite list of attributes (e.g. hair colour) and each instance stores a value for that attribute (e.g. blonde).

16. **Describe in depth the random forest model. What distinguishes a random forest?**

A decision tree combines some decisions, whereas a random forest combines several decision trees. Thus, it is a long process, yet slow. Whereas, a decision tree is fast and operates easily on large data sets, especially the linear one. The random forest model needs rigorous training.

17. **In a random forest, talk about OOB error and variable value.**

The out-of-bag (OOB) error is the average error for each calculated using predictions from the trees that do not contain in their respective bootstrap sample.