**1. Recognize the differences between supervised, semi-supervised, and unsupervised learning.**Supervised learning is a problem with labeled data, expecting to develop predictive capability. Unsupervised learning is discovering process, diving into unlabeled data to capture hidden information. Semi-supervised learning is a blend of supervised and unsupervised learning.

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

1 - Email Spam. The goal is to predict whether an email is a spam and should be delivered to the Junk folder. ...

2 - Handwritten Digit Recognition. ...

3 - Image segmentation. ...

4 - Speech Recognition. ...

5 - DNA Expression Microarray. ...

6 - DNA Sequence Classification.

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

(1) data acquisition and segmentation

(2) [data preprocessing](https://www.sciencedirect.com/topics/computer-science/data-preprocessing)

(3) feature extraction/dimension reduction

(4) recognition and classification.

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

SVM works by mapping data to a high-dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable. A separator between the categories is found, then the data are transformed in such a way that the separator could be drawn as a hyperplane.

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

* SVM's are very good when we have no idea on the data.
* Works well with even unstructured and semi structured data like text, Images and trees.
* The kernel trick is real strength of SVM. ...
* Unlike in neural networks, SVM is not solved for local optima.

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

KNN is one of the simplest forms of machine learning algorithms mostly used for classification. It classifies the data point on how its neighbor is classified. KNN classifies the new data points based on the similarity measure of the earlier stored data points. For example, if we have a dataset of tomatoes and bananas.

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

Cross-validation: Cross-validation is one way to calculate K. First, divide the data into several folds, then use each fold as a test set while training the algorithm on the rest. After that, the algorithm's effectiveness is tested for each K value, and the one with the best results is picked as the best K value.

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

The test data is the data we use to evaluate a model. For KNN the train data is the data that get's used to vote on the class label of a new data point (KNN doesn't really involve any training).

**9. Create the kNN algorithm.**

The k-nearest neighbor classifier fundamentally relies on a distance metric. The better that metric reflects label similarity, the better the classified will be. The most common choice is the Minkowski distance dist(x,z)=(d∑r=1|xr−zr|p)1/p.

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

A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes.

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

There are 4 popular types of decision tree algorithms: ID3, CART (Classification and Regression Trees), Chi-Square and Reduction in Variance

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

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

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

There are several techniques that can help prevent overfitting in decision trees, such as pruning, regularization, and ensemble methods. Pruning is the process of removing or collapsing branches or nodes that do not contribute much to the accuracy or complexity of the tree.

**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 a type of artificial intelligence that uses trees to represent decisions and their possible outcomes. It can be used for classification problems, where the goal is to categorise data points into different classes.

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

Random Forest is a popular machine learning algorithm used for classification and regression tasks due to its high accuracy, robustness, feature importance, versatility, and scalability. Random Forest reduces overfitting by averaging multiple decision trees and is less sensitive to noise and outliers in the data.

**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. This allows the RandomForestClassifier to be fit and validated whilst being trained