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

Ans: **supervised learning has the labelled data, unsupervised learning doesn’t have the labelled data and semi-supervised learning is combination of supervised and unsupervised where it learns from the labelled data and tries to label unlabelled data.**

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

Ans: **Example 1: finding the a requested loan for a person is accepted or rejected based on their financial records or history.**

**Example2: predict the weather if it is going to be sunny, cloudy or outcast**

**Example 3: predicting if the person is diabetic or not**

**Example 4: predicting if a person is smoker or not**

**Example 5: detecting spam emails.**

1. Describe each phase of the classification process in detail.

Ans: **classification follows the same life cycle of a data science project except the algorithms which could be different from regression problems:**

**Phase 1: data gathering:**

**Phase 2: EDA**

**Phase 3: Feature Engineering**

**Phase 4: Feature selection**

**Phase 5: Model Training**

**Phase 6: Model Evaluation**

**Now the model training parts consists of splitting the data, applying cross validation, applying a range of classification algortithms and compare their efficiency to see which algorithms fits the data the best.**

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

Ans: **SVM algorithm is used for regression and classification both. It is used for both linear and non linear problems. This algorithm creates a hyperplane which separate the data into classes. Support vectors are the data points nearest to the hyperplane. The further the data points lie from the hyperplane, the more confident we are that they have been correctly classified. The seperability of the data is defined in terms of soft margin and hard margin. Hard margin could cause overfitting as the model is very strict in terms of classifying points.**

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

Ans: **SVM is more effective in high dimensional spaces and it works well when there is a clear margin of separation between classes but SVM doesn’t perform well when target classes are overlapping.**

1. Go over the kNN model in depth.

Ans: **KNN is a distance-based algorithm and it is also a lazy learner which doesn’t learn from the train data set and rather stores the dataset and finds the distance between new and stored data point to identify its class. We use distance formula to find the distance between K nearest neighbour and a new point. The two most common formulas for distance are Euclidean and Manhattan distance.**

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

Ans: **in KNN, K is regarded as a hyperparameter and with the help of hyperparameter tuning, we try to find the best value of K for which the error can be minimised. We can plot the graph between error rate and K and see at what value of K the error rate is not increasing or remains the same, that value of K can be chosen.**

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

Ans: **we use confusion matrix to measure the performance of the KNN model. We can check and compare the accuracy values to find how model is performing.**

1. Create the kNN algorithm.

Ans: **KNN algorithm is available in Scikit Learn library where an object can be created of the KNN class called KNeighbourCLassifier. The hyperparameter tuning can be done using GridSearchCV which could help us find the best value of K for the given dataset.**

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

Ans: **decision tree works on the concept of if-else conditions. There are three nodes: root node which is the first node in the tree, second is decision node which is responsible for splitting the data, and last node is the leaf node where further splitting can not be done.**

1. Describe the different ways to scan a decision tree.

AnS: **there are mainly 2 techniques for decision trees: IDE3 and CART**

**CART is used for binary splitting and IDE3 doesn’t lead to binary splitting. To determine if the node is a leaf node or not, we use Gini or Entropy and to decide which feature should be selected for splitting as root node can be determined by information gain.**

1. Describe in depth the decision tree algorithm.

Ans: **there are mainly 2 techniques for decision trees: IDE3 and CART**

**CART is used for binary splitting and IDE3 doesn’t lead to binary splitting. To determine if the node is a leaf node or not, we use Gini or Entropy and to decide which feature should be selected for splitting as root node can be determined by information gain.**

**Decision tree can be used for both regression and classification problems and hyperparameter tuning is used to find the max\_depth or min\_split\_leaf**

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

Ans: **inductive bias is the assumptions we make before training the model. Prepruning or postpruning can be used to avoid overfitting. Prepuning is good for large dataset and post pruning is good for small dataset.**

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

Ans: **Decision tree can cause overfitting If it constructs strict rules to split the data and the advantage is interpretability as it is easy to interpret the decision tree,**

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

Ans: **it is good for breaking down complex data into more manageable parts and also non linear data set can be with decision tree for classification or regression.**

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

Ans: **random forest is an ensemble technique where we use multiple decision trees to find the one which gives the highest accuracy.**

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

Ans: **Out of bag error in random forest is used to measure the prediction error, where bagging is used, meaning subsampling is used with replacement. OOB error is done on the test data where unseen data is fed to the model and if the data is misclassified then its contributes to the error**.