1. **In the sense of machine learning, what is a model? What is the best way to train a model?**

A machine learning model is a program that can find patterns or make decisions from a previously unseen dataset. For example, in natural language processing, machine learning models can parse and correctly recognize the intent behind previously unheard sentences or combinations of words.

Ways to train model:

Step 1: Prepare Your Data.

Step 2: Create a Training Datasource.

Step 3: Create an ML Model.

Step 4: Review the ML Model's Predictive Performance and Set a Score Threshold.

Step 5: Use the ML Model to Generate Predictions.

Step 6: Clean Up.

1. **In the sense of machine learning, explain the "No Free Lunch" theorem.**

The No Free Lunch Theorem, often abbreviated as NFL or NFLT, is a theoretical finding that suggests all optimization algorithms perform equally well when their performance is averaged over all possible objective functions. In computational complexity and optimization the no free lunch theorem is a result that states that for certain types of mathematical problems, the computational cost of finding a solution, averaged over all problems in the class, is the same for any solution method.

In Simple Words “No Free Lunch” theorem means we can’t rely on one model to be best of all models. We have to understand data properly and make use of ML understanding and make use of models to find best out of it.

**3. Describe the K-fold cross-validation mechanism in detail.**

In K-fold cross validation, data D is subset into k subsets randomly. Let us assume S1...Sk are the subsets where Sk is the kth randomly split subset of data D. In the first iteration, D-S1 is used for training and S1 for testing the model. When the model has been trained and tested, evaluation can be done, score is noted elsewhere and the trained model is discarded.

These k-iterations go on where 1/k subset of D is always set aside for testing the data and D-1/k subsets are used for training, evaluating and discarding the model. At the end of all the iterations, average of all the evaluation scores is taken and used as output.

**4. Describe the bootstrap sampling method. What is the aim of it?**

The bootstrap method is a statistical technique for estimating quantities about a population by averaging estimates from multiple small data samples.  
Importantly, samples are constructed by drawing observations from a large data sample one at a time and returning them to the data sample after they have been chosen. This allows a given observation to be included in a given small sample more than once. This approach to sampling is called sampling with replacement.

**5. What is the significance of calculating the Kappa value for a classification model? Demonstrate** **how to measure the Kappa value of a classification model using a sample collection of results.**

Kappa value or Cohen's Kappa coefficient is an evaluation metric for classification models. Its significance as an evaluation metric is that it can be used to evaluate multi class classification models and also works on models trained on imbalanced datasets(scores like accuracy scores fail for imbalanced datasets).

In simpler words It basically tells you how much better your classifier is performing over the performance of a classifier that simply guesses at random according to the frequency of each class. Cohen's kappa is always less than or equal to 1. Values of 0 or less, indicate that the classifier is useless Cohen suggested the Kappa result be interpreted as follows: values ≤ 0 as indicating no agreement and 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41– 0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement.

**6. Describe the model ensemble method. In machine learning, what part does it play?**

Ensemble modeling is a process where multiple diverse models are created to predict an outcome, either by using many different modeling algorithms or using different training data sets. The ensemble model then aggregates the prediction of each base model and results in once final prediction for the unseen data.

**7. What is a descriptive model's main purpose? Give examples of real-world problems that** descriptive models were used to solve.

Models that are primarily used for understanding, predicting and communicating are referred to as descriptive models

Descriptive models use data aggregation and data mining to uncover patterns in past or current events. A familiar example of descriptive modeling is business reporting in the form of graphs, charts, and dashboards.

**8. Describe how to evaluate a linear regression model.**

Evaluation of a linear regression model can be done using R-square. R square is calculated as the sum of squared errors in predictions made, divided by summation of all sum of squares. R square measures how much of the change in target variable can be explained by the linear regressor. Its value ranges from 0 to 1 where 0 means poor performance and 1 means good. Some other techniques which can be used to evaluate a linear regression model are:

* Mean Square Error(MSE)/Root Mean Square Error(RMSE)
* Mean Absolute Error(MAE)

**9. Distinguish :**

1. Descriptive vs. predictive models:

* Descriptive models are built to identify trends and underlying patterns.
* Predictive models are built to predict a dependent variable value.
* Most of descriptive models are built using unsupervised machine learning.
* Most of predictive models are built using classification and regression models.
* Example for descriptive model: Finding why consumers are engaging more with a social media post.
* Example for predictive model: Predicting the chances of cancer in a patient.

**2. Underfitting vs. overfitting the model**

* Underfitting is a situation arising when the hypothesis is way too simple, or when the machine learning model is way too simple to produce good results.
* Overfitting is a situation arising when the hypothesis is way too complex, or when the machine learning model is way too complex to produce good results.
* Underfitting causes a model to produce poor results due to heavily simplified algorithm reacting lightly to changes in the unseen data for independent variables from the training data.
* Overfitting makes a model produce poor results due to slightest variations in the unseen data for independent variables from the training data
* Underfitting is also called High Bias.
* Overfitting is also called High variance

1. **Bootstrapping vs. cross-validation**

* Boostrap sampling is a method of sampling in which the repeated sampling is done with replacement using a data D in random draws over which machine learning models are trained for better performance.
* Cross validation is a method used to check the efficacy of the machine learning model on test data.
* End goal of bootstrapping is to reduce overfitting and increase performance.
* End goal of cross validation is only to produce test scores to check efficacy of model
* Bootstrapping is best employed in Random Forest Classifier.
* Cross Validation is best employed using K-fold cross validation technique.

**10. Make quick notes on:**

1. **LOOCV.**

* LOOCV(Leave One Out Cross-Validation) is a type of cross-validation approach in which each observation is considered as the validation set and the rest (N-1) observations are considered as the training set. In LOOCV, fitting of the model is done and predicting using one observation validation set.

**2. F-measurement**

* The F-measure is the harmonic mean of the precision and recall

**3. The width of the silhouette**

* A clustering with an average silhouette width of over 0.7 is considered to be "strong", a value over 0.5 "reasonable" and over 0.25 "weak", but with increasing dimensionality of the data, it becomes difficult to achieve such high values because of the curse of dimensionality, as the distances become more similar.

**4. Receiver operating characteristic curve**

* An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: True Positive Rate. False Positive Rate.