1. **In the sense of machine learning, what is a model?**
   A machine learning model is a file that has been trained to recognize certain types of patterns. You train a model over a set of data, providing it an algorithm that it can use to reason over and learn from those data.
   What is the best way to train a model?

   Step 1: Begin with existing data.

   Step 2: Analyze data to identify patterns.

   Step 3: Make predictions.

2. **In the sense of machine learning, explain the "No Free Lunch" theorem.**
   The No Free Lunch theorem states that all optimization algorithms perform equally well when their performance is averaged across all possible problems.
3. **Describe the K-fold cross-validation mechanism in detail.**
   In k-fold cross-validation, the original sample is randomly partitioned into k equal sized subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k − 1 subsamples are used as training data.
4. **Describe the bootstrap sampling method.**

   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.

   What is the aim of it?
   It is a resampling technique used to estimate statistics on a population by sampling a dataset 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.

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

5. **Describe the model ensemble method.**
   Ensemble methods are techniques that aim at improving the accuracy of results in models by combining multiple models instead of using a single model. The combined models increase the accuracy of the results significantly

In machine learning, what part does it play?

Ensemble methods are techniques that create multiple models and then combine them to produce improved results.

6. **What is a descriptive model's main purpose?**
   Descriptive modeling is a mathematical process that describes real-world events and the relationships between factors responsible for them.
   Give examples of real-world problems that descriptive models were used to solve.
   It is used by consumer-driven organizations for example to help them target their marketing and advertising efforts

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

There are 3 main metrics for model evaluation in regression:

- R Square/Adjusted R Square.
- Mean Square Error(MSE)/Root Mean Square Error(RMSE)
- Mean Absolute Error(MAE)

9. Distinguish :

1. **Descriptive vs. predictive models**

Models that are primarily used for understanding, predicting and communicating are referred to as descriptive models, whereas models mainly used for implementation are called prescriptive models.

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

- ☒ Underfitting means that your model makes accurate, but initially incorrect predictions. In this case, train error is large and val/test error is large too.
- ☒ Overfitting means that your model makes not accurate predictions. In this case, train error is very small and val/test error is large.

3. **Bootstrapping vs. cross-validation**

- ☒ Bootstrapping method uses the original dataset to create multiple datasets after resampling with replacement.
- ☒ Cross validation splits the available dataset to create multiple datasets.

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.

2. F-measurement

It tells us how precise our classifier is (how many instances it classifies correctly), as well as how robust it is (it does not miss a significant number of instances)

3. The width of the silhouette

The silhouette width, s(i), is defined as: s(i) ranges between −1 and 1. Values near 1 indicate that object i is much closer to the other objects in the same cluster than to objects of the closest other cluster, implying a correct classification

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.