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

Solution - 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**.**

Training a machine learning model involves several steps and considerations. While the specific approach can vary depending on the problem domain and available resources, here is a general outline of the best practices for training a machine learning model:

- Define the problem
- Gather and preprocess data
- Choose an appropriate algorithm
- Feature engineering
- Model selection and training
- Evaluate model performance
- Fine-tuning and regularization
- Testing and deployment
- Monitor and iterate

The best approach may vary depending on the specific problem, dataset characteristics, and available resources. It's essential to experiment, iterate, and adapt your approach based on empirical results and domain expertise.

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

Solution - The “no free lunch” (NFL) theorem for supervised machine learning is a theorem that essentially implies that no single machine learning algorithm is universally the best-performing algorithm for all problems.

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

Solution - K-fold cross-validation approach divides the input dataset into K groups of samples of equal sizes. These samples are called folds. For each learning set, the prediction function uses k-1 folds, and the rest of the folds are used for the test set.

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

Solution - In statistics, Bootstrap Sampling is a method that involves drawing of sample data repeatedly with replacement from a data source to estimate a population parameter.

Aim of the bootstrap sampling method is to estimating quantities about a population by averaging estimates from multiple small data samples.

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.**

Solution - 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.

Assume there are two raters r1 and r2 and they are rating ‘yes’ and ‘no’. Their choices are as follows:

r1=[‘yes’,’no’,’yes’,’no’,’yes’,’no’,’yes’,’no’,’yes’]

r2=[‘yes’,’yes’,’yes’,’no’,’no’,’no’,’yes’,’yes’,’yes’]

Now I will make a grid that can calculate the number of yeses and no’s.

	R2 yes	R2 no
R1 yes	4	1
R1 no	2	2

Now, let us make the calculations.First, let us calculate the total possibilities in which both parties agree. That is the diagonal of the above matrix.

Agreement= sum of agreements / total number of instances

Agreement = (4+2)/9 = 0.66

Now we need to consider the cases where raters are not in agreement. We will do this calculating probability of yes and no.

p(yes)= ((4+1)/9)*((4+2)/9)=0.37

p(no)=((2+2)/9)*((2+1)/9)=0.14

Total non disagreement= 0.37+0.14= 0.51

To calculate the Kappa coefficient we will take the probability of agreement minus the probability of disagreement divided by 1 minus the probability of disagreement.

cohens kappa → K= 1-(0.34/0.49) = 0.31

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

Solution - 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.

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

Solution - A descriptive model describes a system or other entity and its relationship to its environment. It is generally used to help specify and/or understand what the system is, what it does, and how it does it.

The steps used by the students to calculate the number of pizza kits to sell for their class trip represent an example of a descriptive model.

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

Solution - Linear Regression Analysis consists of more than just fitting a linear line through a cloud of data points. It consists of 3 stages –

- analysing the correlation and directionality of the data
- estimating the model, i.e.fitting the line
- evaluating the validity and usefulness of the model.

9. Distinguish :

1. **Descriptive vs. predictive models**

Solution - A descriptive model will exploit the past data that are stored in databases and provide you with an accurate report. A Predictive model, identifies patterns found in past and transactional data to find risks and future outcomes.

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

Solution - 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 no accurate predictions. In this case, train error is very small and val/test error is large.

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

Solution - Cross validation splits the available dataset to create multiple datasets, and Bootstrapping method uses the original dataset to create multiple datasets after resampling with replacement. Bootstrapping it is not as strong as Cross validation when it is used for model validation.

10. Make quick notes on:

1. **LOOCV.**

Solution - LOOCV Model Evaluation

Cross-validation, or k-fold cross-validation, is a procedure used to estimate the performance of a machine learning algorithm when making predictions on data not used during the training of the model.

2. **F-measurement**

Solution - An F-score is the harmonic mean of a system's precision and recall values. It can be calculated by the following formula: 2 x [(Precision x Recall) / (Precision + Recall)].

3. **The width of the silhouette**

Solution - The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.

4. **Receiver operating characteristic curve**

Solution - An ROC curve is a plot of sensitivity on the y axis against (1−specificity) on the x axis for varying values of the threshold t. The 45° diagonal line connecting (0,0) to (1,1) is the ROC curve corresponding to random chance. The ROC curve for the gold standard is the line connecting (0,0) to (0,1) and (0,1) to (1,1).