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

- In machine learning a model is a kind of equation where we build an equation on the basis of a training data set and we pass the test data set on the built equation to acquire an output from it. So on the built model, we can check with different types of data sets how the equation is predicting the value. We can say that model is a file that has been trained to recognize certain types of patterns. We can train a model over a set of data, providing it an algorithm that it can use to reason over and learn from those data.

Best way to train a model:-

So while training a model at starting we can use an existing data set to train the model.  once the model is trained we can evaluate the model with the help of test data set and check for the predictions. if the predictions are as per our expectation then the model is good to go live if not then the best way to train a model is to tune the hyperparameters in it. So by including hyperparameters before training the model such as the cross-validation approach, Passing the data batch-wise, and defining a parameter in random state can train the model in a much better way. now passing the testing data set on this type of hyper-tuned or re-trained model can give us prediction in a better manner.

the second way is we can include various types of algorithms on one particular single data set to check which model is behaving better. for example, if we are using data to train a model then we can use two algorithms at a time. After training with both of the algorithms we could have checked which algorithm gives us a very good prediction and go with that prediction.

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

* The No Free Lunch Theorem, often called 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 this NFL it stated that within certain constraints, over the space of all possible problems, every optimization technique will perform as well as every other one on average (including Random Search)
* we show that for both static and time dependent optimization problems the average performance of any pair of algorithms across all possible problems is exactly identical.
* If one algorithm performs better than another algorithm on one class of problems, then it will perform worse on another class of problems also.

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

* K fold cross validation is mechanism of splitting a data into K times for training and testing the model.
* K fold CV is basically used when a particular pattern or data isn’t been trained or identified while training the model.
* This Process helps us to identify each and every data point which needs to be covered.
* So basically , It is a method that uses the same data to both train the model and obtain a less biased estimate of prediction error than the direct estimate.
* For example – Lets say we are training a model with 100 number of features and we want to include cross validation technique in it. So , here we are going to divide the dataset into train and test data. Now , when we specify cross\_validation = 5 or cv=5 then the data will be divide into 5 chunks ie 20 each therefore here k = 5 because the number of folds the technique used is 5 which are divided into 20 chunks. So, while training , for the first iteration the first 20 features will be picked and test data will be picked and that will be trained. For the second iteration , the next 20 features will be picked including test data, and similarly for next iteration upto cv=5. So, after training a model using this technique all the data or patterns will be trained and tested for the prediction.
* By this process , the computation time will be reduced as we are repeating process for each iteration.

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

* The bootstrap sampling method Is used to estimate the statistics on a population data By sampling a subset of the population data.
* In this method, a subset of the population data is used to analyse the data and take out insights from it.
* Data insights or summaries include such as mean, standard deviation, mode, and variance present in that data. this data is not included in the training part as it is the pre-processing part before training the model.
* In this sampling method we basically take out a small chunk or a subset of a huge population data analyze and take out insights from that subset or chunk of the data and then go for the model training or preprocessing technique of the data.
* The main aim of this sampling method is to take out summary statistics from the data such as mean , standard deviation etc.

**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.**

- The main significance of calculating the Kappa value is it helps us to understand how much better out classifier is performing on the data by guessing random according to frequency of class that has been occurred. This values occurs in between 0 or 1 and is less than 1. Below 0 or negative is consider as bad classifier with bad kappa value.

Calculating kappa value :-

Kappa value can be calculated with the help of below formula

K = Po – Pe

1 – Pe

Where , Po is the overall accuracy of model , Pe is the  measure of the agreement between the model predictions and the actual class values as if happening by chance.

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

* A Method which uses similar independent models or different models to derive one single best model or base model from it to predict the output in other words a machine learning technique that combines several base models in order to produce one optimal predictive model.
* This model is been compared with all of the other models to derive an output from.
* These models are nothing but called as base estimators.
* The main purpose of using this ensembling technique is one it increases the performance level in predicting the output and second it is robust in predicting the model performance and less in dispersion of model accuracy.
* There are mainly three types of ensembling techniques :- Bagging , Stacking and Boosting.

1. Bagging - fitting many decision trees on different samples of the same dataset and averaging the predictions.

2. Stacking - fitting many different models types on the same data and using another model to learn how to best combine the predictions.

3. Boosting - adding ensemble members sequentially that correct the predictions made by prior models and outputs a weighted average of the predictions.

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

* A descriptive models are those which help us to fetch the output or predictions based on the past data which is been stored in databases or any other form of storage.
* This model basically involves data mining in it.
* A descriptive model will exploit the past data that are stored in databases and provide you with the accurate report.
* This is one of the process of finding useful and important information by analyzing the huge data.
* Real world example :-

1. It is used in fetching sales revenue report.
2. Used to know performance analysis of the organization.

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

- Evaluation of the linear regression model can be done with the help of three types of metrics ie

1. R square / Adjusted R2

2. Mean Square Error(MSE)/Root Mean Square Error(RMSE)

3. Mean Absolute Error(MAE)

**9. Distinguish :**

**1. Descriptive vs. predictive models**

- Descriptive models :-

* A descriptive models are those which help us to fetch the output or predictions based on the past data which is been stored in databases or any other form of storage.
* This model basically involves data mining in it.
* A descriptive model will exploit the past data that are stored in databases and provide you with the accurate report.
* This is one of the process of finding useful and important information by analyzing the huge data.
* Real world example :-

1. It is used in fetching sales revenue report.
2. Used to know performance analysis of the organization.
3. Describe how to evaluate a linear regression model.

- Predictive models :-

* In Predictive model, it helps us to identify and predict the outcomes on the basis of patterns and transactional data.
* Prediction is done on the basis of future intuition of the data.
* This predictive models involves statistical approaches and forecasting techniques in it.
* The results or predictions obtained from these types of model is not accurate or not up to the mark as they are just predictions made out of the given data.
* Real world example :-

1. Sentiment analysis.

2. credit score analysis.

1. **Underfitting vs. overfitting the model**

* Underfitting – In Underfitting, there will be a high biased and low or high variance model. The training error is high is training data set accuracy is not good.
* Overfitting – In Overfitting , there is Low bias and high variance model. The training error is less in this case. The training dataset accuracy is quiet good and test data accuracy is bad.

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

- Bootstrapping - The bootstrap sampling method Is used to estimate the statistics on a population data By sampling a subset of the population data. In this method, a subset of the population data is used to analyze the data and take out insights from it.

- Cross-validation - cross validation is mechanism of splitting a data into K times for training and testing the model.CV is basically used when a particular pattern or data isn’t been trained or identified while training the model. This Process helps us to identify each and every data point which needs to be covered. So basically, It is a method that uses the same data to both train the model and obtain a less biased estimate of prediction error than the direct estimate.

**10. Make quick notes on:**

1. **LOOCV.**

- LOOCV mainly defined with 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. It is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model.

**2. F-measurement**

- F measurement is nothing but called as f-score or f1 score. It is basically a measure of model accuracy on a dataset. It is used to evaluate binary classifications systems. More often it is calculated  as the harmonic mean of precision and recall, giving each the same weighting.

**3. The width of the silhouette**

- The width of silhouette is actually the average distance of to points in the cluster to which it was assigned, and is the average distance of to the points in the nearest cluster to which it was not assigned. Its coefficient is calculated by (b - a) / max(a, b). Here b is the distance between a sample and the nearest cluster and a is the average middle point of cluster.

**4. Receiver operating characteristic curve**

- It is called as ROC curve. The ROC curve is created by plotting the [true positive rate](https://en.wikipedia.org/wiki/True_positive_rate) (TPR) against the [false positive rate](https://en.wikipedia.org/wiki/False_positive_rate) (FPR) at various threshold settings. It is used to show in a graphical way the connection/trade-off between clinical sensitivity and specificity for every possible cut-off for a test or a combination of tests.