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

A model in ML terms is a kind of mathematical equation. This equation is derived by going through the training dataset. Equation consists of a set of dependent & independent features. Once model is approved & deployed, the input features are passed through it for which an output is derived. Train- Test splits the best way to train a model. This helps us in overcoming the data leakage problem.

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

No Free Lunch theoram states that all optimisation algorithm [perform equally well when their performance are averaged across all possible situations. This means that directly or indirectly a free lunch will incur its cost. In machine learning, No Free Lunch means that there don’t exist a single algorithm which can perform equally well in all the given circumstances. The assumption which are made for one algorithm might not hold true for other algorithm in a given situation.

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

Cross validation is a technique in which Train-Test split is done with equal representation of each and every class of the dataset so that maximum accuracy can be achieved. K fold CV takes a fixed number of data as train set and the rest as test set. Based on this accuracy is calculated. For example – We are having 1000 datasets. Now we divide it by number of iterations. If suppose we want it to be done in 5 experiments then 200 will be the result. Suppose for 1st experiment, we take 1st 200 records as test set and rest 800 as training set. Now accuracy is calculated based on this. Next for 2nd experiment, next 200 record is taken as test dataset and the rest of the 800 record is taken as training dataset. Similarly for this step accuracy is calculated. In the same manner rest of the 3 experiments will be calculated. The average mean of accuracy is calculated by summing all the accuracies and dividing by the number of iterations. This can be communicated to the stakeholders. The problem of equal representation of each class may remain intact which is a drawback with this kind of process.

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

Bootstrap sampling is selecting random rows from a dataset with replacement. This kind of experiments are conducted for N number of experiments. It is quite possible that a single row may get repeated many number of times.The main aim is to estimate summary statistics like mean and standard deviation. Also used in ML to estimate the skills of ML model.

**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.**

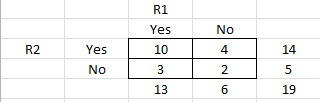
Kappa value for a classification model is a metric often used to assess the agreement between two raters. It can also be used to measure the performance of classification model. The formula for this is given by:

K = P(o) – P(e)/1 - P(e) (i)

Where P(o) is Observed agreement

and P(e) is Agreement by chance

Consider a case in which two bankers are given the responsibility of assessing the creditworthiness of those people willing to take loans. The outcome came to us like this



Observed agreement P(o) = 12/19 = 0.63

Agreement by chance = P(e) = (6/19)\*(5/19)+(13/19)\*(14/19) = 212/361 = 0.58

Now as per eqn. (i)

K= 0.12

It shows there is a slight agreement between the two bankers.

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

Ensemble technique are the process of combining multiple models in the prediction process. Multiple weak learners are combined together in order to form a strong learners. These weak learners are known as base estimators.

In ML, it plays the part of overcoming the situation of overfitting. Multiple weak learners are combine together to reduce overfitting and give accurate predictions.

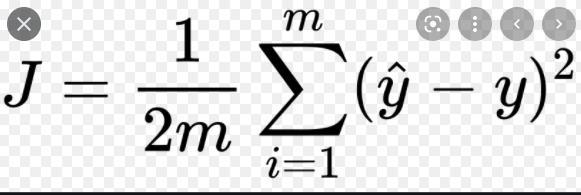
**7. What is a descriptive model's main purpose? Give examples of real-world problems that descriptive models were used to solve.**

Descriptive model main purpose is to look into what happened in the past. It would give you the past analytics using the data that are stored. For a company, it is necessary to know the past events that help them to make decisions based on the statistics using historical data.

The process is used by consumer-driven organizations to help them target their marketing and advertising efforts.

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

A linear Regression model can be evaluated based on the cost or loss function. The formula for which is given by

****

where

m is the number of terms

y is the original datapoint and y^ is the perpendicular point on the LR equation line.

**9. Distinguish :**

**1. Descriptive vs. predictive models**

* Descriptive Analytics will give you a vision into the past and tells you: what has happened? Whereas the Predictive Analytics will recognize the future and tells you: What might happen in future?
* Descriptive Analytics uses Data Aggregation and [Data Mining techniques](https://www.educba.com/data-mining-techniques/) to give you knowledge about past but Predictive Analytics uses Statistical analysis and Forecast techniques to know the future.
* Descriptive Analytics is used when you need to analyze and explain different aspects of your organization whereas Predictive Analytics is used when you need to know anything about the future and fill the information that you do not know.
* A descriptive model will exploit the past data that are stored in databases and provide you with the accurate report. In a [Predictive model](https://www.educba.com/predictive-modeling/), it identifies patterns found in past and transactional data to find risks and future outcomes.
* Descriptive analytics will help an organization to know where they stand in the market, present facts and figures. Whereas predictive analytics will help an organization to know, how they will stand in the market in future and forecasts the facts and figures about the company.
* Reports generated by Descriptive analysis are accurate but the reports generated by Predictive analysis are not 100% accurate it may or may not happen in future.

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

Overfitting is the situation in which our model will perform better in case of training dataset but will perform poorly in case of testing dataset. Thus, Overfitting will lead to low bias but high variance.

Underfitting is the situation in which our model will perform poorly in both training and testing dataset. This leads to high bias as well as variance.

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

Bootstrapping involves random selection with replacement of dataset. This kind of experiments are conducted for N number of experiments. It is quite possible that a single row may get repeated many number of times.

Cross validation is the method in which Train test split of dataset is done with equal representation of each class in both.

**10. Make quick notes on:**

**1. LOOCV (Leave One Out Cross Validation)**

Cross validation is a technique in which Train-Test split is done with equal representation of each and every class of the dataset so that maximum accuracy can be achieved. Suppose we have thousand records. In LOOCV, firstly 1st row will be selected as a test data & the rest as training dataset. In second experiment, only 2nd row will be test data & the rest will be training data. So on the next experiment will be conducted. Problem with LOOCV, is that 1000 iterations need to be performed for building the model. Also, the bias will be quite low but the accuracy will surely fall in case new data comes in leading to overfitting condition.

**2. F-measurement**

F Beta Score is where FN and FN scores both ate important. In such a case, precision and recall both are used.

1+(b^2) (Precision \* Recall)/ b\*(Precision + Recall)

When FN and FP both are important, b=1 , then F score is called F1 score

When FP>FN, b value should be increased and can be taken as 2 and when FP<FN, b value should be decreased and can be taken as 0.5.

**3. The width of the silhouette**

Silhouette Width or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1.

1: Means clusters are well apart from each other and clearly distinguished.

0: Means clusters are indifferent, or we can say that the distance between clusters is not significant.

-1: Means clusters are assigned in the wrong way.

Silhouette Score = (b-a)/max(a,b)

where

a = average intra-cluster distance i.e the average distance between each point within a cluster.

b = average inter-cluster distance i.e the average distance between all clusters.

**4. Receiver operating characteristic curve**

Receiver Operating Characteristic(ROC) curve is a curve drawn between TPR (True Positive Rate) and FPR(False Positive Rate) at various threshold level. A good model should have more than 50% area under its curve. Hence, The more the area under the curve the better the model is.