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

**Ans1:-** Model is an algorithmic function that has been trained on some data. This model can be used for solving prediction and clustering related problems by using input in the form of independent variables the model used to establish patterns and output results using the same patterns.

There are multiple ways to build models and can be filtered down by gauging the advantages and disadvantages against our problem statement. But to put in the most general terms, if the key steps in a machine learning model building are done with utmost care, the model would be the best. This includes good data pre-processing practices, carefully choosing an algorithm over others and more.

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

**Ans2:-** In the sense of machine learning, No Free Lunch theorem means that under some constraints, performance of all optimization algorithms is similar or identical. In addition to that, no machine learning algorithm is the best. Since no single model is the best than the rest of the model always according to the theorem, we should rely on our knowledge of multiple machine learning algorithms and try a wide variety of algorithms to flush out the most optimal one

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

**Ans3:-** In K-fold cross validation, data D is subset into k subsets randomly. Let us assume S1...Sk are the subsets where k 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?**

**Ans4:-** Bootstrap sampling is a method of sampling in which the repeated sampling is done with replacement using a data D in random draws. Bootstrap sampling is often used to simulate sampling distributions instead of relying on theorems like the Central Limit Theorem.

In Machine Learning, it is primarily used in ensemble models like Random Forest classifier. Its aim in Machine Learning use case is to avoid overfitting and improve stability of machine learning algorithms while improving performance.

**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.**

**Ans5:-** 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).

from sklearn.metrics import cohen\_kappa\_score

Observed=['0','1','0','1','0','1','0','1','0','1']

Predicted=['0','0','1','1','0','0','1','1','0','0']

cohen\_kappa\_score(Observed,Predicted) #True Value

output – 0.0

agreed=(3+2)/10

pyes=((3+2)/10)\*((3+2)/10)

pno=((3+2)/10)\*((3+2)/10)

print(agreed,pyes,pno)

output – 0.5 0.25 0.25

disagreed= pyes+pno

print(1-disagreed)

print(1-agreed)

output

0.5

0.5

K= 1-(0.5/0.5)

K=0.9

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

Ans6:- Ensemble methods or ensemble machine learning models are models where more than one models are being used spontaneously to produce better results than individually trained models. Ensemble models are of 4 types

Bagging which is also called Bootstrap aggregation. Multiple base learners learn on subsets of data. A majority vote based on the results of the individual base learners is then used to output the final result. Example: Random Forest

Boosting is mainly used to reduce bias in model. Multiple base learners are trained in a sequence in which each learner covers up the weakness of the previous learner to ultimately produce a powerful model. Example: Adaptive Boosting or Adaboost

Stacking is an ensemble model that combines multiple base learner models constructed parallel to each other and independent. All the models produce predictions that are used creating a meta classifier which is trained on the predictions made by base models. The resultant prediction is then taken as the final prediction. Space and Time complexity issue is huge for stacking models.

Cascading models are built when the cost of making a mistake in prediction is high. A threshold is chosen for multiple base learners. If one base learner is sure about its prediction(Probability>0.99), then the solution of this learner is taken as the final solution. Else, the problem is sent to a more sophisticated, complex model that does the same as the previous model. The problem either gets solved with probability>0.99 or it keeps getting transferred to a more complicated model that is bound to solve it.

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

Ans7:- Main purpose of descriptive models is to find patterns and underlying trends. In machine learning, this is done mostly using unsupervised machine learning algorithms.

Market analysis based on consumer's purchase data, social media post engagement analysis and sales data are real world applications of descriptive models.

8. Describe how to evaluate a linear regression model.

Ans8:- 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.

9. Distinguish :

i. Descriptive vs. predictive models

ii. Underfitting vs. overfitting the model

iii. Bootstrapping vs. cross-validation

Ans9:-

i. Descriptive vs. predictive models

Descriptive models are built to identify trends and underlying patterns. Predictive models are built to predict a dependent variable value.

a. Most of descriptive models are built using unsupervised machine learning. Most of predictive models are built using classification and regression models.

b. 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.

ii. 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.

iii. Bootstrapping vs. cross-validation

Bootstrap 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.

A. 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

B. Bootstrapping is best employed in Random Forest Classifier. Cross Validation is best employed using K-fold cross validation technique.

10. Make quick notes on:

Ans10.

i. LOOCV LOOCV or Leave One Out Cross Validation is a form of K-fold cross validation where only one observation is left out for validation purpose while the rest of the data is used for model training each iteration. It is computationally taxing and should only be used for data with low dimensionality.

ii. F-measurement Harmonic mean of Precision score and recall score is called F-measurement or F-score. It is formulated as 2 (pr re)/pr +re where pr is precision score and re is recall score.

iii. The width of the silhouette Estimate of average inter cluster distance to give efficacy/performance of cluster algorithms is called width of the silhouette. It can also be defined as how identical/similar a data point 'x' is to the data points inside the cluster to which x is assigned. Its value ranges from -1 to 1 where 1 means good and -1 means bad.

iv. Receiver operating characteristic curve Curve plotted between True Positive Rate and False Positive Rate is Receiver Operating Characteristics curve and is used to find the area under the curve for ROC-AUC score for binary classification evaluation. True Positive Rate and False Positive Rate are calculated for different thresholds values where thresholds take values starting from the highest probability scores assigned to data points and goes up to the lowest probability score. The curve is impacted by presence of outliers, and simple models. Extensions can be made to this curve to suit multiclass classification evaluation requirements.