MACHINE LEARNING

1. Which of the following in sk-learn library is used for hyper parameter tuning?
   A) GridSearchCV()
   B) RandomizedCV()
   C) K-fold Cross Validation
   D) All of the above
   ANS- RandomizedCV()

2. In which of the below ensemble techniques trees are trained in parallel?
   A) Random forest
   B) Adaboost
   C) Gradient Boosting
   D) All of the above
   ANS- All of the above

3. In machine learning, if in the below line of code:
   *sklearn.svm.**SVC** (C=1.0, kernel='rbf', degree=3)*
   we increasing the C hyper parameter, what will happen?
   A) The regularization will increase
   B) The regularization will decrease
   C) No effect on regularization
   D) kernel will be changed to linear
   ANS- kernel will be changed to linear

4. Check the below line of code and answer the following questions:
   *sklearn.tree.**DecisionTreeClassifier**(*criterion='gini',splitter='best',max_depth=None, min_samples_split=2)*
   Which of the following is true regarding max_depth hyper parameter?
   A) It regularizes the decision tree by limiting the maximum depth up to which a tree can be grown.
   B) It denotes the number of children a node can have.
   C) both A & B
   D) None of the above
   E) ANS- both A & B

5. Which of the following is true regarding Random Forests?
   A) It's an ensemble of weak learners.
   B) The component trees are trained in series
   C) In case of classification problem, the prediction is made by taking mode of the class labels predicted by the component trees.
   D) None of the above

6. What can be the disadvantage if the learning rate is very high in gradient descent?
   A) Gradient Descent algorithm can diverge from the optimal solution.
   B) Gradient Descent algorithm can keep oscillating around the optimal solution and may not settle.
   C) Both of them
   D) None of them
   ANS- Both of them

7. As the model complexity increases, what will happen?
   A) Bias will increase, Variance decrease
   B) Bias will decreaseincrease Variance increase
   C)both bias and variance increase D) Both bias and variance decrease.

ANS- B) Bias will decreaseincrease Variance increase

8. Suppose I have a linear regression model which is performing as follows:
   Train accuracy=0.95 and Test accuracy=0.75

MACHINE LEARNING

Which of the following is true regarding the model?

A) model is underfitting
B) model is overfitting
C) model is performing good
D) None of the above

ANS= model is underfitting

Q9 to Q15 are subjective answer type questions, Answer them briefly.

9. Suppose we have a dataset which have two classes A and B. The percentage of class A is 40% and percentage of class B is 60%. Calculate the Gini index and entropy of the dataset.

   ANS-

10. What are the advantages of Random Forests over Decision Tree?

    ANS- The random forest has complex visualization and accurate predictions, but the decision tree has simple visualization and less accurate predictions. The advantages of Random Forest are that it prevents overfitting and is more accurate in predictions.

11. What is the need of scaling all numerical features in a dataset? Name any two techniques used for scaling.

ANS- Feature scaling is an important preprocessing step in machine learning that helps to ensure that all features are on a similar scale, which can improve the performance of algorithms .
The two most popular techniques for scaling numerical data prior to modeling are normalization and standardization. Normalization scales each input variable separately to the range 0-1, which is the range for floating-point values where we have the most precision.

MACHINE LEARNING

12. Write down some advantages which scaling provides in optimization using gradient descentalgorithm.

ANS- **The main advantages:**

- We can use fixed learning rate during training without worrying about learning rate decay.
- It has straight trajectory towards the minimum and it is guaranteed to converge in theory to the global minimum if the loss function is convex and to a local minimum if the loss function is not convex.

13. In case of a highly imbalanced dataset for a classification problem, is accuracy a good metric to measure the performance of the model. If not, why?

ANS- Accuracy is not a good metric for imbalanced datasets.

This model would receive a very good accuracy score as it predicted correctly for the majority of observations, but this hides the true performance of the model which is objectively not good as it only predicts for one class

14. What is “f-score" metric? Write its mathematical formula.

ANS- What does F-score measure?

The F-score, also called the F1-score, is a measure of **a model's accuracy on a dataset**. It is used to evaluate binary classification systems, which classify examples into 'positive' or 'negative'.

2 x [(Precision x Recall) / (Precision + Recall)]

15. What is the difference between fit(), transform() and fit_transform()?

ANS- the fit() method will allow us to get the parameters of the scaling function. The transform() method will transform the dataset to proceed with further data analysis steps. The fit_transform() method will determine the parameters and transform the dataset.