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

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

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

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

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

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

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

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.

MACHINE LEARNING

10. What are the advantages of Random Forests over Decision Tree?
    - A decision tree is more simple and interpretable but prone to overfitting, but a random forest is complex and prevents the risk of overfitting.
    - Random forest is a more robust and generalized performance on new data, widely used in various domains such as finance, healthcare, and deep learning

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

    Feature Scaling is done to **normalize** the features in the dataset into a finite range.

    1. Standard Scaler
    2. Power Transform.

12. Write down some advantages which scaling provides in optimization using gradient descent algorithm.
    - Flexibility: Gradient Descent can be used with various cost functions and can handle non-linear regression problems.
    - Scalability: Gradient Descent is scalable to large datasets since it updates the parameters for each training example one at a time.
    - Convergence: Gradient Descent can converge to the global minimum of the cost function, provided that the learning rate is set appropriately.

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?

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

    Accuracy in making positive predictions is measured by a recall, while identifying all positive occurrences in the data is quantified by precision. The F-score ranges from 0 to 1, with higher values indicating better performance.

    F-score = 2 * (precision * recall) / (precision + recall)

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

    Fit(): Method calculates the parameters μ and σ and saves them as internal objects.
    Transform(): Method using these calculated parameters apply the transformation to a particular dataset.
    Fit_transform(): joins the fit() and transform() method for transformation of dataset.