# MACHINE LEARNING ANSWERS IN GREEN

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.

   Answer:-

   Gini index: p(A)(1-p(A)) + p(B)(1-p(B)) = 40%(1-40%) + 60%(1-60%) = 0.24

   Entropy: -p(A)*log2(p(A)) - p(B)*log2(p(B)) = -40%*log2(40%) - 60%*log2(60%) = 0.97

## MACHINE LEARNING ANSWERS IN GREEN

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

Answer:- The advantages of Random Forests over Decision Tree are:

- Random Forests are less prone to overfitting as compared to decision tree.
- Random Forests are more robust to noise in the dataset.
- Random Forests provide better accuracy compared to decision tree.

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

Answer:- Scaling is the process of standardizing the range of features of a dataset. The need of scaling is to ensure that each feature contributes approximately proportionately to the final distance. Two techniques used for scaling are:

- Min-Max Scaling
- Standardization

# MACHINE LEARNING ANSWERS IN GREEN

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

    Answer:- Scaling provides following advantages in optimization using gradient descent algorithm:

    - It helps to converge faster
    - It helps to find global minima

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?

    Answer:- In case of a highly imbalanced dataset for a classification problem, accuracy is not a good metric to measure the performance of the model because accuracy is computed by dividing the number of correct predictions to total predictions. As the majority class is over-represented, the classifier may predict the majority class most of the time and still have a high accuracy.

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

    Answer:- F-score is a metric that combines precision and recall to provide a single measure of the performance of a classification model. The mathematical formula for f-score is:-
    F-score = (2 * Precision * Recall) / (Precision + Recall).

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

    Answer:- In machine learning, fit(), transform() and fit_transform() are methods of the scikit-learn library used for preprocessing data:

    - fit() method is used to fit the data to the model, it is used to calculate the internal parameters of the model.
    - transform() method is used to transform the data according to the internal parameters calculated during the fit() method.
    - fit_transform() method is used to fit the data to the model and then transform it in one step.