MACHINE LEARNING
Worksheet 7

1. Which of the following in sk-learn library is used for hyper parameter tuning?
Ans: Option D, all of the above

2. In which of the below ensemble techniques trees are trained in parallel?
Ans: Option A, Random Forest

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?
Ans: Option A, The regularization will increase

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?
Ans: Option A, It regularizes the decision tree by limiting the maximum depth up to which a tree can be grown.

5. Which of the following is true regarding Random Forests?
Ans: Option A and C, It's an ensemble of weak learners. And In case of classification problem, the prediction is made by taking mode of the class labels
predicted by the component trees.

6. What can be the disadvantage if the learning rate is very high in gradient descent?
Ans: Option C, both of them

7. . As the model complexity increases, what will happen?
Ans: Option B, Bias will decrease, Variance increase

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?
Ans: Option B, model is overfitting

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: A=40 and B=60
Calculating Gini Index

$$Gini = 1 - \sum_{i} p_j^2$$

Gini = 1 - (0.40^2 + 0.60^2)
= 1 - (0.16+0.36)
= 1 – (0.52)
=0.48

Calculating Entropy

$$Entropy = -\sum_j p_j \log_2 p_j$$

Entropy = - [0.4 * log2(0.4) + 0.6 * log2(0.6)]
= - [0.4 * -1.32192809489 + 0.6 * -0.736965594166]
= 0.97

10. What are the advantages of Random Forests over Decision Tree?
Ans: Random Forests have higher accuracy over decision trees along with lower overfitting problem. Also Random Forests is less sensitive to outliers compared to decision trees. Also Ran dom Forests are trained in parallel making them faster to train on large datasets.

11. What is the need of scaling all numerical features in a dataset? Name any two techniques used for scaling.
Ans: Scaling is performed so that all the features in dataset are at an equal scale. This means that model should not give higher weightage to some features and less weigtage to other features. Scaling ensures that all the features values are within certain range that are readable by machine learning algorithm making an unbiased model.
Two commonly used techniques are standardisation and min-max scaling.

12. Write down some advantages which scaling provides in optimization using gradient descent algorithm.
Ans: Scaling can ensure that the gradients computed for each feature are similar in magnitude, which can help to provide a more accurate estimate of the true gradient. Also, Large differences in feature scales can lead to numerical overflow or underflow issues, which can make it difficult to compute the gradient accurately. Scaling can prevent such issues by ensuring that the numerical computations are within a reasonable range.

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: In the case of a highly imbalanced dataset for a classification problem, accuracy may not be a good metric to measure the performance of the model. This is because accuracy can be misleading in the presence of imbalanced classes, where one class is significantly more prevalent than the other.In such cases, alternative metrics such as precision, recall, F1-score, and Area Under the Receiver Operating Characteristic curve (AUC-ROC) are recommended.

14. What is “f-score" metric? Write its mathematical formula.
Ans: It is the harmonic mean of precision and recall, and provides a balance between the two metrics. The formula for F-score is:

F1-score = 2 * (precision * recall) / (precision + recall)
The F1-score ranges between 0 and 1, where 1 is the best possible score. A higher F1-score indicates a better balance between precision and recall, indicating a better-performing model. The F1-score is often used in cases where both precision and recall are equally important, such as in medical diagnosis or fraud detection.

15. What is the difference between fit(), transform() and fit_transform()?
Ans: fit(): This method is used to compute the mean and standard deviation, or other statistics, of the input data. It is used to "fit" the preprocessing model to the data.

transform(): This method applies the preprocessing model to new data to transform it in the same way as the original data. This method is used after the fit() method has been applied to the original data.

fit_transform(): This method combines the fit() and transform() methods into a single step. It is used to fit the preprocessing model to the data and transform it in one step.