Machine Learning Assignment

1. **Which of the following in sk-learn library is used for hyper parameter tuning?**

Answer: A) GridSearchCV()

2. **In which of the below ensemble techniques trees are trained in parallel?**

Answer: A) Random forest

3. I**n 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?

Answer: B) The regularisation will decrease

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?

Answer: A) It regularises 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?**

Answer: A) It's an ensemble of weak learners.

6. **What can be the disadvantage if the learning rate is very high in gradient descent?**

Answer: C) Both of them

7. **As the model complexity increases, what will happen?**

Answer: 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?

Answer: 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.**

To calculate the Gini index of a dataset with two classes A and B where the percentage of class A is 40% and the percentage of class B is 60%, the formula is: Gini index = 1 - (pA^2 + pB^2), where pA and pB are the probabilities of class A and class B respectively.
Gini index = 1 - (0.4^2 + 0.6^2) = 1 - (0.16 + 0.36) = 1 - 0.52 = 0.48

To calculate the entropy of the dataset, the formula is: Entropy = - (pA log2(pA) + pB log2(pB)), where pA and pB are the probabilities of class A and class B respectively.

Entropy = - (0.4 log2(0.4) + 0.6 log2(0.6)) = - (0.6309 + 0.5185) = - 1.1494

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

Advantages of Random Forests over Decision Tree:

- Random Forests reduce overfitting by averaging the results of multiple decision trees.
- Random Forests handle missing values and outliers better than Decision Trees.
- Random Forests can be used for both classification and regression tasks, whereas Decision Trees are mainly used for classification tasks.
- Random Forests are more robust to noise and can handle large datasets with higher dimensionality.

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

Scaling all numerical features in a dataset is necessary to ensure that no one feature is disproportionately weighted in the analysis and that all features are on the same scale. Two common techniques used for scaling are normalisation and standardisation.

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

Scaling provides several advantages in optimization using gradient descent algorithm:

- It helps to converge faster as the optimizer can make bigger steps in the direction of the optimum.
- It helps to avoid local minima and saddle points as the optimizer is less likely to be trapped in such regions.
- It helps to achieve better generalisation as the optimizer is less likely to overfit the training data.

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?

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 it does not take into account the imbalance. A model that simply classifies all examples as belonging to the majority class will have a high accuracy, but this does not reflect its ability to correctly classify the minority class.

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

F-score is a metric that combines precision and recall to provide a single measure of a model's performance. 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()?

The difference between fit(), transform() and fit_transform() is:

- fit() is used to fit the model on the training data and compute any internal parameters needed for the transformation.
- transform() is used to apply the transformation to the training data after the model has been fit.
- fit_transform() is a convenience function that combines the fit() and transform() steps in one function call.