Machine Learning (Answers)

1. R-squared is a better measure of the goodness of fit model in regression, than the residual sum of squares (RSS); because R-squared is a statistical measure that represents the proportion of the variance for the dependent variable which is determined by the independent variables in the model.

2. TSS: It is a variation of the values of a dependent variable from the sample mean of the dependent variable.

   ESS: It is the sum of the squares of the deviations of the predicted values from the mean value of a response variable, in a standard regression model.

   RSS: It measures the level of variance in the error term, or residuals, of a regression model.

   Equation: TSS = ESS + RSS

3. Regularization is importantly used in machine learning to prevent overfitting and improve the generalization performance of models.

4. The Gini-impurity Index is a way of quantifying how messy or clean a dataset is, especially when we use decision trees to classify it.

5. Yes, unregularized decision trees are prone to overfitting. It happens when they learn a training set to a point of high granularity that makes them easily overfit. Allowing a decision tree to split to a granular degree, is the behaviour of this model that makes it prone

to learning every point extremely well — to the point of perfect classification.

6. Ensemble technique in machine learning, is an approach where several models are trained to address a common problem, and their predictions are combined to enhance the overall performance.

7. Bagging reduces variance by averaging predictions from models trained on different subsets of data, while boosting reduces bias by sequentially training models that focus on errors of previous models.

8. The out-of-bag (OOB) error is a performance metric that estimates the performance of the Random Forest model using samples not included in the bootstrap sample.

9. K-fold cross-validation is an approach which divides the input dataset into K groups of samples of equal sizes. These samples are called as folds.

10. Hyperparameter tuning is the process of selecting the optimal values for a machine learning model's hyperparameters. It is crucially done, as they control the overall behaviour of a machine learning model.

11. If we have a large learning rate in a gradient descent, it can cause the model to converge too quickly to a suboptimal solution.

12. Logistic regression cannot be used for classification of non-linear data as –
    - It cannot capture the complexity and non-linearity of the data which means, it assumes a linear relationship between the input features and the output.

- It is sensitive to outliers and noise, which can affect the accuracy and stability of the model.

- Logistic regression also has a limited capacity to learn from multiple features, as it can only combine them linearly.

13. Gradient boosting is more robust to outliers and noise since it equally considers all training instances when optimizing the loss function' whereas AdaBoost is faster but more impacted by dirty data since it fixates on hard examples.

14. The bias-variance trade-off is about finding the right balance between simplicity and complexity in a machine learning model. High bias means the model is too simple and consistently misses the target, while high variance means the model is too complex and shoots all over the place.

15. Linear Kernel is used when the data is Linearly separable, that is, it can be separated using a single Line.

    RBF SVM works by mapping the input data into a higher-dimensional feature space, where the classes can be separated by a hyperplane. The algorithm uses Radial Basis Function, to measure the similarity between pairs of data points in the feature space.

    The polynomial kernel represents the similarity of vectors (training samples) in a feature space over polynomials of the original variables, allowing learning of non-linear models.