MACHINE LEARNING

1. **R-squared or Residual Sum of Squares (RSS) which one of these two is a better measure of goodness of fit model in regression and why?**

   R-squared vs R-squared is a more accurate indicator of goodness of fit because it offers a normalized estimate (between 0 and 1) of how well the regression model explains the variability of the response data.
   Residual Sum of Squares (RSS)is an absolute measure of the difference between the data and the model, which makes it more difficult to comprehend, particularly when contrasting models with varying sizes.

2. **What are TSS (Total Sum of Squares), ESS (Explained Sum of Squares) and RSS (Residual Sum of Squares) in Regression. Also mention the equation relating these three metrics with each other.**

   TSS, ESS, and RSS in Regression are explained as follows

   - **Total Sum of Squares (TSS)**: This measures how much the response variable varies overall. It shows the total spread of the observed data points and is used to see how much of this variation can be explained by the model.
   - **Explained Sum of Squares (ESS)**: This shows how much of the total variation in the response variable is explained by the model. It indicates how well the model's predictors capture the data's variability.
   - **Residual Sum of Squares (RSS)**: This measures the variation in the response variable that the model doesn't explain. It represents the errors and shows the difference between the actual data points and the model's predictions..

   The Equation relating them is: TSS=ESS+RSSTSS = ESS + RSSTSS=ESS+RSS

3. **What is the need of regularization in machine learning?**

   Regularization is needed to prevent overfitting by penalizing larger coefficients in the model. It enhances the model's ability to generalize to fresh, untested data.

4. **What is Gini–impurity index?**

   Gini-impurity Index a metric used to assess a decision tree node's purity. It measures how likely it is that a random sample would be incorrectly classified if it were randomly assigned labels based on the node's label distribution.

5. **Are Unregularized decision-trees prone to overfitting? If yes, why?**

   Yes, **Unregularized decision-trees** can produce extremely complex trees that precisely fit the training data, capturing both noise and the underlying patterns, which makes them susceptible

to overfitting

.

6. **What is an ensemble technique in machine learning?**

   An ensemble technique **in Machine Learning** integrates several different ML models to increase the overall performance, The theory goes that weak learners can combine to become strong learners.

7. **What is the difference between Bagging and Boosting techniques?**

- **Bagging Techniques**: Lowers variance by averaging the predictions of several models trained on various subsets of the data.
- **Boosting Techniques** is a technique that lessens bias by training models in a sequential manner, with each model attempting to fix the mistakes of its predecessors.

8. **What is out-of-bag error in random forests?**

   Out-of-Bag Error in Random Forests: The term "out-of-bag error" refers to an estimate of the prediction error for random forests that is derived from the samples (out-of-bag samples) that were not utilized in the training of each tree.

9. **What is K-fold cross-validation?**

   The **K-fold Cross-Validation** technique divides the data into K folds, or subsets. Using a separate fold as the validation set and the remaining K-1 folds as the training set, the model is trained K times. The results are then averaged to estimate model performance.

10. **What is hyper parameter tuning in machine learning and why it is done?**

    Hyperparameter Tuning is the process of finding the best set of hyperparameters for a machine learning model. It is done to improve model performance and generalization to new data.

11. **What issues can occur if we have a large learning rate in Gradient Descent?**

    If we have a large learning rate in Gradient Descent, it can cause the model to converge too quickly to a suboptimal solution or even diverge, as it may overshoot the optimal solution.

12. **Can we use Logistic Regression for classification of Non-Linear Data? If not, why?**

    Since logistic regression is by nature a linear classifier, it has trouble with non-linear data unless it is modified using polynomial features or another method that can capture the non-linearity.

13. **Differentiate between Adaboost and Gradient Boosting.**

- **Adaboost:** , This technique focuses on cases that were incorrectly classified by changing the instances' weights in later models.
- **Gradient Boosting:** This method creates models one after the other by optimizing a loss function. Usually, gradient descent is applied on the residual errors of earlier models.

14. **What is bias-variance trade off in machine learning?**

The bias-variance trade-off is the balance between the error introduced by bias (error from overly simplistic models) and variance (error from overly complex models). A good model needs to find the optimal balance between the two.

15. Give short description each of Linear, RBF, Polynomial kernels used in SVM.

- **Linear Kernel**: Ideal for data that can be separated linearly, meaning classes can be divided by a straight line (or hyperplane in higher dimensions). It's straightforward, efficient in computation, and effective for large datasets.
- **RBF Kernel (Radial Basis Function)**: Best for non-linear data as it transforms the data into higher dimensions, capturing complex relationships. Requires careful parameter tuning to balance between overfitting and underfitting.
- **Polynomial Kernel**: Suitable for non-linear data that follows polynomial relationships. It generates intricate decision boundaries, with the degree of the polynomial affecting flexibility—higher degrees may overfit, while lower degrees may underfit.