MACHINE LEARNING

Q1 to Q15 are subjective answer type questions, Answer them briefly.

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?
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.
3. What is the need of regularization in machine learning?
4. What is Gini–impurity index?
5. Are unregularized decision-trees prone to overfitting? If yes, why?
6. What is an ensemble technique in machine learning?
7. What is the difference between Bagging and Boosting techniques?
8. What is out-of-bag error in random forests?
9. What is K-fold cross-validation?
10. What is hyper parameter tuning in machine learning and why it is done?
11. What issues can occur if we have a large learning rate in Gradient Descent?
12. Can we use Logistic Regression for classification of Non-Linear Data? If not, why?
13. Differentiate between Adaboost and Gradient Boosting.
14. What is bias-variance trade off in machine learning?
15. Give short description each of Linear, RBF, Polynomial kernels used in SVM.

Answers for the following above Questions: -

Ans-1 R-Squared is generally a better measure of goodness of fit because it gives a clear, Standardized understanding of how well the model explains the variation. R-squared is a relative measure that is easy to interpret across different models.

Ans-2 **TSS Definition: -** TSS (Total Sum of Squares) is a measure of the total variation in the dependent variable. It is a sum of ESS & RSS.

Equation of TSS-

$$TSS = \sum(Yi - \bar{Y})^2$$

Where, Y= Yi is the actual value of the dependent variable for observation i,
$\bar{y}$ is the mean of the actual values of the dependent variable,

ESS Definition: - ESS (Explained Sum of Squares), also known as **Regression Sum of Squares**, measures the portion of the total variation in the dependent variable that is explained by the regression model.

Equation of ESS-

$$ESS = \sum_{i=1}^{n}(yi-\bar{y})2$$

Where, Y= Yi is the predicted value of the dependent variable for observation i,
$\bar{y}$ is the mean of the actual values of the dependent variable,

RSS Definition: - **RSS (Residual Sum of Squares)**, also called the **Sum of Squared Residuals**, measures the amount of variation in the dependent variable that is not explained by the regression model.

Equation of RSS-

$$RSS= \sum_{i=1}^{n}(yi-y\hat{}i)2$$

Where, Y= Yi is the actual value of the dependent variable for observation i,
Y^i= is the predicted value of the dependent variable for observation i,

Ans-3 Regularization is a technique used in machine learning to prevent overfitting, improve model generalization, and ensure that models perform well on unseen data.

Ans-4 The Gini Impurity Index is a metric used in decision trees (specifically in classification tasks) to measure the degree of impurity or disorder in a dataset. It is a key criterion for determining how to split the data at each node in a decision tree, especially in algorithms like CART (Classification and Regression Trees).

Ans-5 Yes, unregularized decision trees are highly prone to overfitting. This is because decision trees can become overly complex, fitting to the noise or irrelevant details in the training data rather than capturing the general patterns.

Ans-6 An **ensemble technique** in machine learning refers to methods that combine the predictions of multiple models to improve the overall performance and robustness of the system.

Ans-7 ***Bagging***: - 1. The main goal of bagging is to **reduce variance** and prevent **overfitting**.
2. Bagging aims to improve the accuracy and stability of machine learning algorithms by averaging multiple models trained on different random subsets of the data.

Boosting: - 1. The primary objective of boosting is to **reduce bias** and improve the accuracy of the model by sequentially building models that focus on correcting the errors of the previous ones.
2. Boosting iteratively adjusts the weights of misclassified samples, emphasizing difficult cases for the next model to handle better.

Ans-8 Out-of-bag (OOB) error is a key concept in Random Forests and other bagging-based models. It provides an internal, unbiased estimate of the model's accuracy without the need for a separate validation set or cross-validation.

Ans-9 **K-fold cross-validation** is a robust method for assessing the performance of a machine learning model and ensuring that it generalizes well to unseen data. It is commonly used to evaluate a model's effectiveness and prevent overfitting by making efficient use of the available dataset.

Ans-10 Hyperparameters are configuration settings used to control the learning process and the structure of a machine learning model. It is done to improve Model Performance, Prevent Overfitting and Underfitting, Optimize Computational Resources, Adapt to Different Datasets.

Ans-11 Following issues may occur if we have a large learning rate in Gradient Descent: -

1. Divergence
2. Oscillation
3. Exploding Gradients
4. Slow Convergence
5. Inability to Find a Good Local Minimum

Ans-12 **Logistic Regression** is primarily designed for linear classification problems and may not perform well with non-linear data due to its linear decision boundary.

Ans-14 The **bias-variance trade-off** is a fundamental concept in machine learning that helps in understanding and managing the balance between two sources of error that affect a model's performance: **bias** and **variance**. These two sources of error are inversely related, and finding the right balance between them is crucial for building models that generalize well to unseen data.

Ans-15 ***Linear: -*** In Support Vector Machines (SVM), the term **"linear"** refers to the type of decision boundary that the algorithm constructs in the feature space to separate different classes

RBF: - In Support Vector Machines (SVM), the **Radial Basis Function (RBF)** kernel is a popular and powerful tool used to handle non-linear classification problems.

Polynomial kernels The **Polynomial kernel** is another type of kernel function used in Support Vector Machines (SVM) for handling non-linear classification and regression problems.