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?

   R-squared is a better measure because it is a normalized value between 0 and 1, indicating how well the model explains the variance in the data. RSS measures the total squared difference between actual and predicted values, but it lacks the intuitive interpretation that R-squared offers. R-squared provides a proportion of variance explained, making it easier to interpret.

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 (Total Sum of Squares): Total variance in the data.

ESS (Explained Sum of Squares): Variance explained by the model.

RSS (Residual Sum of Squares): Variance not explained by the model (error term).

Equation: TSS=ESS+RSSTSS = ESS + RSSTSS=ESS+RSS

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

Regularization helps prevent **overfitting** by penalizing large coefficients in models like linear regression. It encourages simpler models that generalize better to unseen data by adding a penalty to the cost function (e.g., L1 or L2 regularization).

4. What is Gini–impurity index?

Gini impurity measures the likelihood of an incorrect classification of a randomly chosen element

from the dataset. It is used in decision trees to evaluate how well a split separates classes.

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

Yes, because decision trees can grow too complex by splitting until they perfectly classify the training data. This leads to overfitting, where the tree performs well on training data but poorly on new data.

6. What is an ensemble technique in machine learning?

Ensemble methods combine multiple models to improve prediction accuracy. They aggregate the predictions of multiple models (e.g., decision trees) to produce a stronger final prediction, reducing overfitting and variance.

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

Bagging (e.g., Random Forest): Creates multiple independent models by training each on random subsets of the data and then averages their predictions.

Boosting (e.g., AdaBoost): Builds models sequentially, where each new model focuses on correcting the errors of the previous models, resulting in a strong final model.

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

The out-of-bag error is an estimate of the model's performance on unseen data. It is calculated using data not included in each bootstrap sample during random forest training and serves as a cross-validation method.

9. What is K-fold cross-validation?

K-fold cross-validation splits the dataset into K equal-sized subsets (folds). The model is trained on K-1 folds and tested on the remaining fold. This process is repeated K times, and the average performance is reported to reduce bias and variance.

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

Hyperparameter tuning is the process of selecting the best hyperparameters for a model (e.g., learning rate, regularization strength) to optimize its performance. It is essential because the wrong hyperparameters can lead to poor model performance.

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

A large learning rate can cause the model to overshoot the optimal solution, leading to divergence or oscillation instead of convergence. This results in poor model performance and failure to find the minimum of the cost function.

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

Logistic regression cannot directly classify non-linear data because it assumes a linear decision boundary. However, it can handle non-linear data if transformed with non-linear features or by using techniques like polynomial features or kernel methods.

13. Differentiate between Adaboost and Gradient Boosting.

AdaBoost: Weights misclassified data points more heavily in each subsequent model.

Gradient Boosting: Sequentially minimizes the error by fitting models on the residual errors of the previous model, using gradient descent.

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

The bias-variance trade-off refers to the balance between **bias** (error due to assumptions in the model) and variance (error due to sensitivity to small fluctuations in the training data). High bias leads to underfitting, while high variance leads to overfitting.

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

Linear kernel: Assumes a linear relationship between the features and the target variable.

RBF (Radial Basis Function) kernel: Measures similarity between two points based on their distance, allowing it to capture non-linear relationships.

Polynomial kernel: Computes the similarity between points as a polynomial function of their feature values, useful for non-linear data.