MACHINE LEARNING

Q.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?

Ans= R-squared is more effective method to measure the goodness of fit model in regression, because when we use R-squared method it provides normalized data so that we can understand better.

Q.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.

Ans= TSS is the total variance in dependent variable. It is the sum of the square of differences between the observed value and the mean of the observed value. Here is the equation.

-TSS=$\sum_{i=1}^{n}(yi-\overline{y})^2$, Here yi is observed value and $\overline{y}$ is mean of observed value.

ESS is the sum of the square of difference between the predicted value and the mean of the observed value. Here is the equation.

-ESS=$\sum_{i=1}^{n}(\widehat{yi}-\overline{y})^2$, Here $\widehat{yi}$ is the predicted value and $\overline{y}$ is mean of observed value.

RSS is the sum of the squared of difference between observed value and predicted value. Here is the equation.

-RSS=$\sum_{i=1}^{n}(yi-\widehat{yi})^2$, Here yi is observed value and $\widehat{yi}$ is the predicted value.

TSS=ESS+RSS

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

Ans= In machine learning regularization is needed because it is

-Prevent Overfitting

-Enhance Generalization

-Reduce Variance

-Improve Interpretability

Q.4. What is Gini–impurity index?

Ans=The Gini impurity index is a measure used in decision tree algorithms to evaluate the quality of a split or node. It measures the likelihood of an incorrect classification of a randomly chosen element if it was randomly labeled according to the distribution of labels in the subset.

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

Ans= Yes, unregularized decision trees are prone to overfitting because they can grow very deep and create many branches, each capturing noise and specific details from the training data. This excessive complexity leads to a model that fits the training data very well but performs poorly on new, unseen data due to its inability to generalize.

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

Ans= An ensemble technique in machine learning involves combining multiple models to improve the overall performance and accuracy of predictions.

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

Ans= Bagging Trains multiple models independently on random subsets of data while boosting Trains models sequentially, each correcting errors of the previous ones. As we know Bagging techniques Reduces overfitting by averaging predictions and on other hand Boosting techniques.

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

Ans= Out-of-bag (OOB) error in random forests is an estimate of the model's prediction error based on the observations that are not included in the bootstrap sample for each decision tree.

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

Ans= K-fold cross-validation is a technique used to evaluate the performance and generalizability of a machine learning model.

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

Ans= Hyperparameter tuning is the process of finding the optimal set of hyperparameters for a machine learning model. Hyperparameters are the settings that you configure before training a model, unlike parameters which are learned during training. And Hyperparameter tuning is crucial because the right hyperparameters can significantly improve a model's predictive performance and efficiency.

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

Ans= If we have a large learning rate in Gradient Descent, the following issues can occur Overshooting,Non-Convergence,Instability,Poor Performance,Using an appropriate learning rate is crucial for the stability and efficiency of Gradient Descent.

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

Ans= No, Logistic Regression is not suitable for classifying non-linear data because Logistic Regression models a linear relationship between the input features and the log-odds of the target class. It assumes that the data can be separated by a straight line and It cannot capture complex non-linear relationships between features and the target variable.

Q.13 Differentiate between Adaboost and Gradient Boosting.

Ans= Adaboost is Emphasizes difficult cases and sequentially corrects errors using weighted samples while Gradient Boosting is Minimizes residual errors through gradient descent optimization, often using decision trees as weak learners.Both techniques aim to build a strong learner from a sequence of weak learners, but they differ in their approach to error correction and optimization.

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

Ans= The bias-variance tradeoff is a fundamental concept in machine learning that relates to the performance of a model.

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

Ans= The linear kernel calculates the dot product between two feature vectors. The RBF kernel computes the similarity between two feature vectors based on the Euclidean distance between them. The polynomial kernel calculates the similarity between two feature vectors using polynomial functions.