ANSWERS FOR MACHINE LEARNING ASSIGNMENT

1-In regression, R-squared is a better measure of goodness of fit than Residual Sum of Squares (RSS) because it provides the proportion of the variance in the dependent variable that is predictable from the independent variable(s).

2-TSS (Total Sum of Squares) represents the total variability in the response variable, ESS (Explained Sum of Squares) measures the variability explained by the regression model, and RSS (Residual Sum of Squares) quantifies the unexplained variability. They are related by the equation TSS = ESS + RSS.

3-Regularization in machine learning is needed to prevent overfitting and improve the generalization of the model by adding a penalty to the loss function, discouraging overly complex models.

4-The Gini impurity index is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset.

5-Yes, unregularized decision trees are prone to overfitting because they can keep splitting the nodes until each leaf has only one instance, capturing noise in the data.

6-An ensemble technique in machine learning involves constructing multiple models and combining them to produce a stronger model, which often has better predictive performance than individual models.

7-The main difference between Bagging and Boosting techniques is that Bagging trains each model in the ensemble independently, while Boosting trains models sequentially, with each new model attempting to correct the errors of the previous ones.

8-The out-of-bag error in random forests is the average error for each training sample calculated using predictions from the trees that do not contain the sample in their respective bootstrap sample.

9-K-fold cross-validation is a resampling procedure used to evaluate the performance of a machine learning model. It involves splitting the training dataset into k consecutive folds and using k-1 folds for training and the remaining fold for testing.

10-Hyperparameter tuning in machine learning is the process of choosing a set of optimal hyperparameters for a learning algorithm. It is done to achieve better model performance and generalization on new, unseen data.

11-A large learning rate in Gradient Descent can lead to the algorithm overshooting the minimum and failing to converge, or even diverging, causing unstable and oscillatory behavior.

12-Logistic Regression can be used for classification of non-linear data by using techniques like feature engineering, adding higher-order terms, or using non-linear transformations of the features.

13-Adaboost and Gradient Boosting are both ensemble learning methods, but they differ in how they build the sequential models and how they assign weights to the training instances.

14-The bias-variance trade-off in machine learning refers to the balance between the error due to bias and the error due to variance. A model with high bias will underfit the data, while a model with high variance will overfit the data.

15-Linear kernel: It represents a linear decision boundary. It is the simplest kernel and is used when the data is linearly separable.

RBF (Radial Basis Function) kernel: It is used when the data is not linearly separable. It is able to handle complex relationships between the input and output variables.

Polynomial kernel: It is used when the data has non-linear relationships. It can be of any degree and is specified by its degree and coefficient.