ASSIGNMENT – 5

MACHINE LEARNING

Answer-1:

R-squared is considered better measure of goodness of fit than RSS because it provides the score between 0 - 1 and also explains the variance in the data.

Answer-2:

TSS is the total variability in the dependent variable. It is the sum of squared differences between each observed value and the mean of that value.

ESS is the variability in the dependent variable that is explained by the independent variables. It is the sum of the squared differences between the predicted values and the mean of the values.

RSS is the unexplained variability in the dependent variable that is attributed to the residuals or errors of the regression model.

TSS = ESS + RSS

Answer-3:

Regularisation is done to prevent the data from overfitting and to avoid the problem of multi-collinearity.

Answer-4:

The Gini impurity index is a measure of impurity used in decision tree algorithms. It ranges from 0-1.

Answer-5:

Yes, unregularized decision tress can lead to overfitting. This is due to the high variance in the data which further leads to max depth of the tree being very high.

Answer-6:

Ensemble approach in machine learning is a technique where we build multiple models to improve the performance and increased accuracy.

Answer-7:

Bagging involves training multiple instances of the same learning algorithm on different subsets of the training data. The individual models in the bagging are trained independently and in parallel.

Boosting involves training a series of weak learners sequentially. The training of subsequent models in boosting is influenced by the performance of previous models.

Answer-8:

Out-of-bag (OOB) error in Random Forest is a way to estimate the model's performance without the need for a separate validation set. During the construction of each decision tree in a Random Forest, a bootstrap sample is drawn from the original dataset. The bootstrap samples are drawn with replacement, some data points are not included in the sample for each tree. Each tree in the Random Forest, the OOB instances that were not used in training that particular tree can be used to estimate the performance of the tree. The tree's predictions on the OOB instances are compared to the true labels, and the error is computed.

Answer-9:

K-fold cross-validation is a technique used in machine learning to assess the performance and generalization ability of a model. It involves dividing the dataset into K subsets, called folds, and then training the model K times, each time using K-1 folds for training and the remaining fold for validation.

Answer-10:

Hyperparameter tuning is the process of selecting the optimal hyperparameter values for a machine learning model. Hyperparameters are external configuration settings that are not learned from the data but are set before the training process begins. They influence the behaviour of the

model during training and can significantly impact the model's performance and increase the accuracy.

Answer-11:

Having a large learning rate in gradient descent can lead to several issues:

The most significant issue with a large learning rate is that the optimization process may fail to converge.

A large learning rate may cause the algorithm to overshoot the minimum of the loss function.

Large learning rates can introduce instability into the training process.

Answer-12:

Logistic Regression is a linear model used for binary classification, and it makes decisions based on a linear combination of input features. While Logistic Regression can effectively handle linearly separable data, it may struggle with non-linearly separable data. If the decision boundary between classes is non-linear, Logistic Regression may not capture the underlying patterns well.

Answer-13:

AdaBoost trains weak learners sequentially, and each learner tries to correct the mistakes made by its predecessor. The

final prediction is obtained by a weighted sum of all weak learners.

Gradient Boosting also trains weak learners sequentially, but each learner is trained to minimize the gradient of the loss function with respect to the model's predictions.

Answer-14:

Bias-Variance Trade-off is a concept in machine learning that helps to enhance the performance of the model by balancing these two sources of error: bias and variance.

Answer-15:

Linear kernel represents the original feature space and is suitable for linearly separable data.

RBF kernel is commonly used for capturing complex, nonlinear relationships in the data.

Polynomial kernel allows SVM to capture polynomial relationships between features, introducing additional complexity to the decision boundary.