MACHINE LEARNING ASSIGNMENT

1) R-squared is generally considered a better measure of the goodness of fit in regression compared to the Residual Sum of Squares (RSS) because R-squared provides a normalized measure of how well the independent variables explain the variability of the dependent variable. Specifically, R-squared indicates the proportion of the variance in the dependent variable that is predictable from the independent variables, making it easier to interpret and compare across different models. RSS, on the other hand, is a raw measure of the sum of squared residuals (differences between observed and predicted values), which can be more difficult to interpret on its own, especially when comparing models with different scales.

2) Measures the total variance in the dependent variable.

ESS (Explained Sum of Squares): Measures the variance explained by the regression model.

RSS (Residual Sum of Squares): Measures the variance not explained by the regression model (the residuals).

The relationship between these metrics is:

TSS=ESS+RSS\text{TSS} = \text{ESS} + \text{RSS}TSS=ESS+RSS

3) Regularization is needed in machine learning to prevent overfitting, which occurs when a model is too complex and fits the training data too closely, capturing noise as if it were a true pattern.

4) The Gini-impurity index is a measure of the probability of a randomly chosen element being incorrectly classified if it was randomly labeled according to the distribution of labels in the subset. It is used in decision trees to determine the best split.

5) Yes, unregularized decision trees are prone to overfitting because they can create overly complex models that fit the training data too closely. They can split the data into very fine partitions, capturing noise and small fluctuations as if they were significant patterns.

6) An ensemble technique in machine learning involves combining the predictions of multiple models to improve accuracy and robustness. The idea is that by aggregating the results of diverse models, the ensemble can mitigate the errors of individual models and improve overall performance.

7) • **Bagging (Bootstrap Aggregating)**:

- Involves training multiple models independently using random subsets of the training data (with replacement).
- Aggregates the predictions by averaging (for regression) or voting (for classification).
- Reduces variance and helps prevent overfitting.

- **Boosting**:

- Involves training models sequentially, where each model attempts to correct the errors of the previous one.
- Aggregates the predictions by weighted voting or summing.
- Reduces bias and can achieve high accuracy, but can be prone to overfitting if not regularized.

8) Out-of-bag (OOB) error in random forests is an estimate of the model's prediction error obtained by using the data that was not sampled (left out) during the bootstrap sampling process for training individual trees. It provides an unbiased evaluation of the model's performance without the need for a separate validation set.

9) K-fold cross-validation is a technique used to assess the generalizability of a machine learning model. The data is divided into k subsets (folds). The model is trained kkk times, each time using a different fold as the validation set and the remaining k−1 folds as the training set. The performance metrics are averaged over the K iterations to provide a more robust estimate of model performance.

10) Hyperparameter tuning involves finding the optimal set of hyperparameters (parameters that are not learned from data but set prior to training) for a machine learning model. It is done to improve model performance by finding the best configuration for the model's learning process.

11) A large learning rate in gradient descent can cause the algorithm to overshoot the optimal solution, leading to divergence or oscillation around the minimum rather than converging to it. This results in poor performance and an inability to find the best parameters.

12) Logistic regression is inherently a linear classifier and is not suitable for non-linear data. It can only model a linear decision boundary. For non-linear data, techniques like polynomial features, kernel methods, or non-linear classifiers like SVM with RBF kernel or neural networks are needed.

13) Adaptive boosting Sequentially adjusts the weights of misclassified instances, giving higher importance to misclassified points.

Combines weak classifiers to form a strong classifier.

Gradient boosting Sequentially builds models by optimizing a loss function.

Each model attempts to correct the errors of the previous one by fitting the residuals.

14) The bias-variance trade-off is the balance between the model's ability to generalize to new data (variance) and its accuracy on the training data (bias). High bias can lead to underfitting, while high variance can lead to overfitting. The goal is to find a balance where the model has low bias and low variance for optimal performance.

15) • **Linear Kernel**:

- Used for linearly separable data.
- Simple and computationally efficient.
- The decision boundary is a straight line (or hyperplane).

- **RBF (Radial Basis Function) Kernel**:

- Used for non-linear data.
- Transforms data into a higher-dimensional space.
- The decision boundary can be non-linear.

- **Polynomial Kernel**:

- Used for non-linear data.
- Maps input features into a polynomial feature space.
- The decision boundary can be curved, depending on the degree of the polynomial