Machine learning assignment

1. R-squared is a better measure of goodness of fit in regression. It represents the proportion of the variance in the dependent variable that is predictable from the independent variables. Higher R-squared values indicate a better fit.
   RSS measures the total variance of the residuals (the differences between actual and predicted values). While it gives an idea of how much the model is off in absolute terms, R-squared provides a normalized measure of model performance relative to the total variance.

2. TSS (Total Sum of Squares): It measures the total variance in the dependent variable.
   ESS (Explained Sum of Squares): It measures the variance explained by the regression model.
   RSS (Residual Sum of Squares): It measures the unexplained variance, the sum of squared residuals.
   The equation relating these three metrics is: TSS = ESS + RSS.
3. Regularization is needed to prevent overfitting in machine learning models. It adds a penalty term to the model's objective function, discouraging overly complex models that may perform well on training data but generalize poorly to new, unseen data.
4. Gini impurity is a measure of how often a randomly chosen element would be incorrectly labelled if it was randomly labelled according to the distribution of labels in the set. It is commonly used in decision tree algorithms for classification.
5. Yes, unregularized decision trees are prone to overfitting. They can become too complex, capturing noise in the training data and fitting it as if it were a real pattern. This leads to poor generalization to new data.
6. An ensemble technique combines multiple individual models to create a stronger, more robust model. It helps improve predictive performance and reduce overfitting.
7. Difference between Bagging and Boosting:
   - Bagging (Bootstrap Aggregating) builds multiple models independently and combines them by averaging or voting.
   - Boosting builds models sequentially, with each model correcting the errors of the previous one, resulting in a strong overall model.
8. Out-of-bag error is the error rate of a model on the instances that were not used in training but left out during the bootstrap sampling. It serves as an unbiased estimate of the model's performance.
9. K-fold cross-validation involves partitioning the dataset into K equally sized folds, using K-1 folds for training, and the remaining fold for testing. This process is repeated K times, with each fold used as the test set exactly once.
10. Hyperparameter tuning involves finding the optimal values for the hyperparameters of a machine learning model. It is done to improve the model's performance and generalization to new data.
11. Large learning rates can lead to overshooting the minimum of the loss function, causing the algorithm to fail to converge or converge to a suboptimal solution. It may result in unstable and oscillatory behaviour.

12. Logistic Regression is a linear model and may not perform well on highly non-linear data. It relies on a linear decision boundary, so for non-linear data, more complex models like SVM or decision trees might be more appropriate.
13. Difference between Adaboost and gradientboosting
    Adaboost focuses on correcting misclassifications by assigning higher weights to misclassified instances.
    Gradient Boosting builds sequential models, with each model correcting the errors of the previous one using gradient descent.
14. The bias-variance trade-off is the balance between the model's ability to capture the underlying patterns in the data (low bias) and its sensitivity to noise and fluctuations in the training data (low variance). Achieving a good balance is crucial for model generalization.
15. Linear Kernel: Suitable for linearly separable data, assumes a linear decision boundary.
    RBF (Radial Basis Function) Kernel: Handles non-linear data by transforming it into higher-dimensional space, allowing complex decision boundaries.
    Polynomial Kernel: Similar to the RBF kernel but uses a polynomial function for transformation, providing flexibility in capturing non-linear relationships.