MACHINE LEARNING -ASSIGNMENT – 5

1. R-squared (Coefficient of Determination) is generally considered a better measure of the goodness of fit in regression models compared to the Residual Sum of Squares.
   because ,
   Interpretability:
   R-squared: R-squared provides a clear interpretation of the proportion of the variance in the dependent variable that is explained by the independent variables in the model. A value of 1 indicates a perfect fit, while 0 indicates no explanatory power.
   RSS: RSS is a raw sum of squared differences between the observed and predicted values, and its magnitude is less interpretable in terms of explaining variability.

   Normalization:
   R-squared: R-squared is normalized, scaling between 0 and 1. Normalization allows for easier comparison across different models and datasets.
   RSS: RSS is not normalized and can vary based on the scale of the dependent variable, making it less suitable for comparisons.

   Comparison Across Models:
   R-squared: R-squared facilitates model comparison. An increase in R-squared generally suggests improved model performance, but adjusted R-squared can be used to penalize the inclusion of irrelevant variables.
   RSS: RSS alone does not consider the number of predictors and may not be suitable for comparing models with different numbers of variables.

2. **total sum of squares(TSS)**
   In statistical data analysis the total sum of squares (TSS or SST) is a quantity that appears as part of a standard way of presenting results of such analyses.
   explained sum of squares(ESS)
   The explained sum of squares (ESS) is the sum of the squares of the deviations of the predicted values from the mean value of a response variable,
   Residual Sumof Squares(RSS)
   The residual sum of squares (RSS) is a statistical technique used to measure the variance in a data set that is not explained by the regression model

   relation between these are , **TSS= ESS + RSS**

3. While training a machine learning model, the model can easily be overfitted or under fitted. To avoid this, we use regularization in machine learning to properly fit a model onto our test set. Regularization techniques help reduce the chance of overfitting and help us get an optimal model.

4. Gini Impurity is the probability of incorrectly classifying a randomly chosen element in the dataset if it were randomly labeled according to the class distribution in the dataset. It's calculated as. G = ∑ i = 1 C p (i) ∗ (1 − p (i)) G = \sum_{i=1}^C p(i) * (1 - p(i)) G=i=1∑Cp(i)∗(1−p(i))

5. Yes, unregularized decision trees are prone to overfitting. Overfitting occurs when a model learns the training data too well, capturing noise or random fluctuations in the data rather than the underlying patterns. Decision trees, by their nature, are highly flexible and have the capacity to create complex, deep tree structures that perfectly fit the training data.

6. Ensemble technique is a method that combines the predictions of multiple individual models to create a more robust and accurate predictive model
   The main idea is to leverage the diversity of multiple models to improve overall performance, reduce overfitting, and enhance generalization.

   2 types of Ensemble technique -**Bagging (Bootstrap Aggregating)and Boosting**

7. **Bagging (Bootstrap Aggregating):**
   In bagging, multiple instances of the same learning algorithm are trained on different subsets of the training data, which are created by random sampling with replacement (bootstrap samples). Each model is trained independently, and their predictions are combined through averaging (for regression) or voting (for classification).
   Example algorithms: Random Forest.

   Boosting:
   Boosting involves training multiple weak learners sequentially, with each model focusing on the mistakes of its predecessor. It assigns weights to data points, emphasizing the misclassified points in subsequent models. The final prediction is typically a weighted sum of the individual model predictions.
   Example algorithms: AdaBoost, Gradient Boosting (e.g., XGBoost, LightGBM, CatBoost).

8. The out-of-bag (OOB) error is a concept specific to Random Forests, a popular ensemble learning algorithm. In Random Forests, each tree in the forest is constructed using a bootstrap sample of the original training data. Since the bootstrap sample is a random sample with replacement, some observations are likely to be omitted from each sample, and some may be repeated.

9. K-fold cross-validation is a resampling technique commonly used in machine learning to assess the performance of a predictive model and to reduce the risk of overfitting. The dataset is divided into K subsets (or folds), and the learning algorithm is trained K times, each time using K-1 folds for training and the remaining fold for validation/testing. This process is repeated K times, with each fold used exactly once as a validation while the K-1 remaining folds form the training set.

10. Hyperparameter tuning, also known as hyperparameter optimization, is the process of finding the best set of hyperparameters for a machine learning model. Hyperparameters are external configuration settings that are not learned from the data but are set prior to the training process. Examples include the learning rate in gradient boosting, the regularization parameter in linear models, or the depth of a decision tree.
11.
12. Logistic Regression is a linear classification algorithm, and by its nature, it assumes a linear relationship between the features and the log-odds of the target variable. Consequently, Logistic Regression is generally not suitable for handling highly non-linear relationships in the data. If the decision boundary between classes is highly non-linear, Logistic Regression may struggle to capture and model that complexity effectively.
13. Adaboost is computed with a specific loss function and becomes more rigid when comes to few iterations. But in gradient boosting, it assists in finding the proper solution to additional iteration modeling problem as it is built with some generic features
14. The bias-variance trade-off is a fundamental concept in machine learning that addresses the trade-off between a model's ability to fit the training data well and its ability to generalize to new, unseen data. It is crucial to strike a balance between bias and variance to create a model that performs well on both the training set and new data
15. Support Vector Machines (SVMs) are a class of supervised learning algorithms that can be used for both classification and regression tasks. SVMs use kernels to transform the input data into a higher-dimensional space, allowing for the creation of non-linear decision boundaries. Here are short descriptions of commonly used kernels in SVM:

    Linear Kernel:
    Description: The linear kernel is the simplest kernel, and it computes the inner product of the feature vectors in the original space. It is suitable for linearly separable data or when the decision boundary is expected to be close to a hyperplane.
    Radial Basis Function (RBF) Kernel:
    Description: The RBF kernel, also known as the Gaussian kernel, is a widely used kernel that allows SVMs to model complex, non-linear decision boundaries. It maps the input data into an infinite-dimensional space. It is effective when there is no prior knowledge about the structure of the data.
    Polynomial Kernel:
    Description: The polynomial kernel is used to handle non-linear data by mapping it into a higher-dimensional space. It is suitable for data with polynomial structures. The degree of the polynomial is a hyperparameter that determines the flexibility of the decision boundary.