Machine Learning

1.Ans- R-squared is a commonly used measure of goodness of fit in regression analysis. It is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. R-squared ranges from 0 to 1, with a higher value indicating a better fit.
Residual Sum of Squares (RSS) is a measure of the difference between the observed data and the predicted values by the model. It is used to determine the variance of residuals. Lower RSS indicates a better fit.
In general, R-squared is a more widely used measure of goodness of fit in regression analysis because it is easily interpretable and it indicates the proportion of variation in the dependent variable that can be explained by the independent variables. However, it is important to note that R-squared does not indicate whether a model is a good fit, it only indicates the proportion of variation that can be explained by the model.

2.Ans- TSS (Total Sum of Squares) is a measure of the total variance in the dependent variable, regardless of whether it is explained by the independent variables in the regression model. It is calculated as the sum of the squared differences between the mean of the dependent variable and each individual data point.
ESS (Explained Sum of Squares) is a measure of the variance in the dependent variable that is explained by the independent variables in the regression model. It is calculated as the sum of the squared differences between the predicted values of the dependent variable and the mean of the dependent variable.
RSS (Residual Sum of Squares) is a measure of the variance in the dependent variable that is not explained by the independent variables in the regression model. It is calculated as the sum of the squared differences between the observed values of the dependent variable and the predicted values of the dependent variable.
The relationship between TSS, ESS, and RSS can be represented by the following equation:
TSS = ESS + RSS
This equation states that the total variance in the dependent variable (TSS) is equal to the explained variance (ESS) plus the residual variance (RSS). In other words, TSS measures the total amount of variation in the dependent variable, while ESS and RSS measure the explained and residual variation, respectively.

3.Ans- Regularization helps in the following stages of machine learning: i) Providing a good accuracy in the model. ii) Helps in the problem of overfitting- A high-dimensional dataset having too many features can sometimes lead to overfitting. iii) Helps in underfitting- It prevents the loss of important data due to underfitting.

4.Ans- Gini Index is one of the most popular algorithm which is used by Decision Tree for selecting the best split. Gini Impurity measures the impurity of the nodes in a Decision Tree, hence lower the Gini impurity we can safely infer the purity will be more and a higher chance of the homogeneity of the nodes. Gini Impurity is always between 0 to 1.Gini Impurity=1-Gini

5.Ans- Yes, if decision trees are left unregularized will be prone overfitting. This is because it will grow which means each leaf node will represent one data point. In order to overcome this issue of overfitting, we should prune the tree. We can prune decision Tree by setting Max-depth of the tree or by setting minimum data points in each node

6.Ans- Ensemble methods is a machine learning technique that combines several base models in order to produce one optimal predictive model.For example: Bagging or Bootstrap Aggregation and Random Forest Model.

7.Ans- Bagging attempts to tackle the over-fitting issue. Each model in bagging is trained parallelly and independently where in a final prediction is created from the prediction of every models. Boosting tries to reduce bias. Boosting is an interative process which trains all the models together and gets a certain prediction, a second model is then built which tries to correct the errors present in the first model. This procedure is continued and models are added until either the complete training data set is predicted correctly or the maximum number of models is added.

8.Ans- The out-of-bag (OOB) error is the average error for each calculated using predictions from the trees that do not contain in their respective bootstrap sample. This allows the RandomForestClassifier to be fit and validated whilst being trained

9.Ans- Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation.

10. Ans- Hyperparameter tuning consists of finding a set of optimal hyperparameter values for a learning algorithm while applying this optimized algorithm to any data set.In machine learning, a hyperparameter is a parameter whose value is used to control the learning process That combination of hyperparameters maximizes the model's performance, minimizing a predefined loss function to produce better results with fewer errors. Hyperparameter is required as it can give the optimized values for hyperparameters, which maximizes the model's predictive accuracy.

11.Ans- If the learning rate is very large then it will skip the optimal solution. A higher rate could result in a model that might not be able to predict anything accurately

12.Ans- Logistic regression is neither linear nor is it a classifier. The idea of a "decision boundary" haslittle to do with logistic regression, which is instead a direct probability estimation method that separates predictions from decision

13.Ans- AdaBoost is the first designed boosting algorithm with a particular loss function. On the other hand, Gradient Boosting is a generic algorithm that assists in searching the approximate solutions to the additive modelling problem. This makes Gradient Boosting more flexible than AdaBoost.

14.Ans- Bias Variance Trade off is a tradeoff between a model’s ability to minimize bias and variance which is referred to as the best solution for selecting a value of regularization constant. Proper understanding of these errors would help to avoid the overfitting and underfitting of a data set while training the algorithm.

15.Ans- Linear kernel: The linear kernel is the simplest kernel function. It simply computes the dot product of the two input vectors. It is used when the data is linearly separable.

RBF (Radial basis function) kernel: The RBF kernel is a popular kernel function used in SVMs. It is defined as the exponential of the negative Euclidean distance between the input vectors. It is often used in non-linear classification problems.

Polynomial kernel: The polynomial kernel is a non-linear kernel function that computes the dot product of the input vectors raised to a power. The degree of the polynomial is a parameter that can be tuned. It is often used when the data is not linearly separable.

STATISTICS WORKSHEET

1.Ans: D

2.Ans: C

3.Ans: C

4.Ans: B

5.Ans: C

6.Ans: C

7.Ans: A

8.Ans: A

9.Ans: B

10.Ans: A