STATISTICS WORKSHEET-5

Q1 to Q10 are mcQs with only one correct answer. Q1 to Q10 are MCQs Choose the correct option.

- 1)Ans Expected
- 2)Ans Rank
- 3)Ans 6
- 4)Ans Chisquared distributions
- 5)Ans f Distribution
- 6)Ans Hypothesis
- 7)Ans Null Hypothesis
- 8)Ans Two tailed
- 9)Ans Research Hypothesis
- 10)Ans NP

Q1 to Q15 are subjective answer type questions, Answer them briefly.

1)

Ans R-squared is generally a better measure of the goodness of fit for a regression model than the residual sum of squares because R-squared shows how well the data fit the regression model (the good of fit) it is a statistical measure that represents the proportion of the variance for the dependent variable that explained by the independent variables in the model.

2)

Ans   particular, the explained sum of squares measures how much variation there is in the modelled values and this is compared to the total sum of squares (TSS), which measures how much variation there is in the observed data, and to the residual sum of squares, which measures the variation in the error between observed data and modelled values.

3)

Ans   IT need to prevent overfitting and improve the generalization performance of models of regularization in machine learning.

4)

Ans   Gini-impurity index is a measure of how mixed or impure a dataset is. It ranges between 0 & 1 where 0 represents a pure dataset and 1 represents a completely impure dataset.

5)

Ans   yes unregularized decision-trees prone to overfitting because they learn too much from the training data and fail to generalize well to new data.it is a popular power full method for data mining ,as we can handle both numerical and categorical data.

6)

Ans   ensemble technique in machine learning combine multiple models to make predictions or classifications.

7)

Ans   Bagging is a learning approach that aids in enhancing the performance, execution and precision of machine learning algorithms & boosting is an approach that iteratively modifies the weight of observation based on the last classification.

8)

Ans   The out-of-bag error is a way to measure the prediction error of a random forest model.

9)

Ans   k -fold cross-validation is a technique for evaluating machine learning models by splitting a data set into subsets or folds and using each fold for training &testing.

10)

Ans

Ans   Hyper parameter tuning is the process of finding the best set hyper parameters are variables that control the model training process and set before the learning process begins.

11)

Ans   A large learning rate in Gradient Descent can cause a model to very-shoot the optimal solution & fail to converge which can lead to poor performance.   Because a large learning rate cause the algorithm to tale big steps in the direction of the negative gradient, we can cause it to skip the minimum point.

12

Ans No we can’t used logistic regression for classification of Non-linear data because it may not perform that well when the relations hip between the feature and out comes is non-linear.

13)

Ans Boosting and AdaBoost are both ensemble machine learning techniques that combine multiple weak learners into a strong predictor. The key differences are:

AdaBoost	Gradient Boosting
In AdaBoost “”shortcomings” are identified by high-weight data points.	In Gradient boost “shortcomings” are identified by gradients.
In AdaBoost, shift is done by up-weighting observations that were misclassified before.	Gradients boost identifies difficult observations by large residuals computed in the previous iterations.
Exponential loss of AdaBoost gives more weights for those sample fitted worse.	Gradient boost further dissect error components to bring in more explanation.

14)

Ans Skriking balance between accuracy and the ability to make predictions beyond the training data in an ml model is called the bias-variance trade off.

15)

Ans Short description of:

linear =linear kernel is used the data is linearly separated, that is, it can be separated using a single line.

RBF = Radial basis function (RBF) is a popular kernel function used in various kernelized learning algorithms. It is commonly used in support vector machine classification.

Polynomial kernel = The polynomial kernel is the input data into a higher- dimensional feature space using polynomial functions of the original features.