ASSIGNMENT 2

STATISTICS ASSIGNMENT

1. Using a goodness of fit, we can assess whether a set of obtained frequencies differ from a set of frequencies.

Answer: Expected

2. Chi-square is used to analyze

Answer: Frequencies

3. What is the mean of a Chi-Square distribution with 6 degrees of freedom?

Answer: 6

4. Which of these distributions is used for goodness of fit testing?

Answer: Chi-squared distribution

5. Which of the following distributions is Continuous?

Answer: F Distribution

6. A statement made about a population for testing purpose is called?

Answer: Hypothesis

7. If the assumed hypothesis is tested for rejection considering it to be true is called?

Answer: Null Hypothesis

8. If the Critical region is evenly distributed, then the test is referred to as?

Answer: Two tailed

9. Alternative Hypothesis is also called as?

Answer: Research Hypothesis

10. In a Binomial Distribution, if ‘n’ is the number of trials and ‘p’ is the probability of success, then the mean value is given by

Answer: np

MACHINE LEARNING ASSIGNMENT

1. R-squared and Residual Sum of Squares (RSS)

R-squared shows how well the model fits the data, it is more informative

RSS measures the total error, it is less informative

2. TSS, ESS, RSS in regression

TSS is the total variation in the data

ESS is explained variation by the model

RSS is the remaining error

TSS = ESS + RSS

3. Need for regularization

It prevents the model from becoming too complex

It helps the model perform better on new data

4. Gini impurity index

It is a measure of how mixed classes are in decision trees

Lower gini means better splits

5. Are unregularized decision-trees prone to overfitting? If yes, why?

Yes, they can overfit because they can become too complex and memorize the data

6. Ensemble technique

It combines several models to improve predictions and reduce errors

7. Bagging and Boosting

Bagging trains models independently in parallel

Boosting trains models sequentially, each improving the previous one

8. Out of bag error

It is calculated using the data not used during tree training in random forests

It helps estimate model accuracy

9. K-fold cross validation

Data is split into K parts, with the model trained on K-1 parts and tested on the remaining part, rotating through all parts

10. Hyperparameter tuning

It is the process of finding the best settings for a machine learning model

It is used to adjust model settings to improve performance

11. Large learning rate issues

If we have large learning rate in gradient descent then the model might skip the optimal solution and fail to converge properly

12. Can we use Logistic Regression for classification of Non-Linear Data? If not, why?

No, because logistic regression assumes a straight line decision boundary

13. Adaboost and Gradient Boosting

Adaboost focuses on misclassified data

Gradient boosting minimizes error step by step

14. Bias variance trade off

It is used to find the right balance between simplicity and complexity

Bias – It happens when a model is too simple and misses important patterns

Variance – It happens when a model is too complex and tries to fit every detail in training data

Trade off – It is used to balance both and capture the main patterns without overfitting

15. SVM kernels

Linear – It is used for straight-line separations

RBF – It is used for complex, curved boundaries

Polynomial – It is used for non-linear separations in higher dimensions