MACHINE LEARNING ASSIGNMENT 5

1.R-squared vs RSS for goodness of Fit in Regression:

R-squared is generally a better measure in RSS because it is normalized and indicates the proportion of variance in the dependent variable is predictable from the independent variables. RSS on the other hand, is an absolute measure that depends on the scale of data, making it less suitable for comparing models across different datasets

2.TSS, ESS, and RSS in Regression:

TSS (Total SUM Of Squares): The total variance in the response variable

ESS (Explained Sum of Squares): The variance explianed by the model

RSS (Residual Sum of Squares): The variance that not explained by the model

3.Need for regularization in Machine Learning:

Regularization is used to prevent overlifting by penalizing the large coefficients in model ensuring the models complexity is balanced with performance on unseen data

4.Gini impurity text:

Gini impurity text measures the degree of probability of specific variable being wrongly classified when it randomly chosen. It's a key metric in decision trees

5.Unreguralized Decision Trees and Overlifting:

Yes, unreguralized decision trees are prone to overlifting because they can create complex trees that fit in and noise in the training data rather than capturing the true pattern

6.Ensemble Technique in Machine Learning:

An ensemble technique combines multiple models to improve the overall performance, robustness, and accuracy of prediction compared to single modekl

7.Difference between Bagging and Boosting:

Bagging: involves in training multiple models in parallel, each on random subset of data and averaging their predictions

Boosting: Sequantially trains models, where each model attempts to correct the errors made by the previous ones

8.Out-Of-Bag Error in Random Forests:

Out-of-bag error is the error estimate of random forest model on the training data using only tress that didn't have a particular data in their bootstrap sample

9.K-Fold Cross Validation:

It involves dividing the dataset into K subsets and repeatedly training the model K times each time using the subset as the test set and others are the training set to validate the test models

10.Hyperparameter Tuning in Machine Learning:

Hyperparameter tuning involves systematically searching for the optimal paramaeters of a model to improve its performance.it provides the highest accuracy or performance metric.

11.Large Learning Rate in Gradient Descent Issues:

A large rate can cause the algorithm to converge too quickly to suboptimal solution or even diverge, missing the global minimum

12.Logistic Regression and Non-Linear Data:

Logistics Regression typically used for the linear classifications. For non-linear data, it can struggle unless feature engineering is used to capture nonlinearity

13.Differentiating Adaboost and Gradient Boosting:

Adaboost: Focused on increasing the weighs of misclassified data points after each iteration

Gradient Booting: Optimizes a loss function directly, adjusting for residuel in iteration

14.Bias-Variance Trade –Off in Machine Learning:

It's the trade-off between the model's ability to generalized well to new data (low variance) and its accuracy on training data (low bias) .Caregiving a balance is key to good predictive performance

15.Linear, RBF, Polynomial Kernels in SVM:

Linear Kernel: Good for linear separable data uses the original feature space.

RBF (Radial Bias Function) Kernel: Effective for non-linear data can map into infinite dimensions

Polynomial Kernel: Suitable for non-linear data maps data into higher dimensions using polynomial kernel.

STATISTICS WORKSHEET –G

1.C) Predicted

2.c) Frequencies

3.c)6

4.b) Chi squared Distribution

5.c) F Distribution

6.b) Hypothesis

7.a) Null hypothesis

8.a) Two tailed

9.b) Research hypothesis

10.a) np