Task 2

MACHINE LEARNING ASSIGNMENT – 5

A1.

R-squared is a better measure because it provides a normalized value (between 0 and 1) that explains the proportion of variance in the dependent variable that is predictable from the independent variables. RSS alone cannot provide this normalized information.

A2.

- **TSS**: Measures the total variance in the observed data.
- **ESS**: Measures the variance explained by the model.
- **RSS**: Measures the variance that the model fails to explain (the error).
- **Equation**: TSS = ESS + RSS

A3.

Regularization is needed to prevent 'overfitting' by penalizing overly complex models. It reduces model variance by adding constraints, like L1 (Lasso) or L2 (Ridge), to the loss function.

A4.

It is a measure of how often a randomly chosen element from the set would be incorrectly classified. It is used in decision trees to decide splits, with lower Gini values indicating purer nodes.

A5.

Yes, because unregularized decision trees can grow too complex by fitting noise in the data, especially if no depth or leaf constraints are imposed, leading to poor generalization.

A6.

Ensemble techniques combine multiple models (e.g., decision trees) to produce better results than a single model. Examples include Bagging (Random Forests) and Boosting (AdaBoost).

A7.

- **Bagging**: Uses random sampling with replacement to create multiple models and averages their predictions.
- **Boosting**: Sequentially builds models, where each model tries to correct the errors of the previous one.

A8.

Out-of-bag (OOB) error is the error estimate for random forests, computed by using each tree's unused data (not part of the bootstrap sample) to test its performance, providing an unbiased estimate.

A9.

K-fold cross-validation splits the dataset into 'K' subsets or folds. The model is trained on 'K-1' folds and validated on the remaining one. This is repeated K times to ensure robust evaluation.

A10.

Hyperparameter tuning is the process of optimizing hyperparameters (parameters not learned from the data) to improve model performance. It is done to find the best configuration for better model accuracy and generalization.

A11.

A large learning rate can cause the model to 'overshoot' the minimum of the loss function, leading to divergence or oscillation rather than convergence to the optimal solution.

A12.

Logistic Regression assumes a 'linear relationship' between input variables and the log-odds of the outcome, so it cannot directly handle non-linear data. For non-linear data, other models (e.g., kernel methods, neural networks) are better suited.

A13.

- **AdaBoost**: Emphasizes misclassified instances in each iteration and adjusts weights accordingly.
- **Gradient Boosting**: Sequentially adds models to minimize the residual errors from the previous models, focusing on correcting mistakes through gradient descent.

A14.

The bias-variance trade-off refers to the balance between:

- **Bias**: Error due to overly simple models that underfit the data.
- **Variance**: Error due to overly complex models that overfit the data. Achieving a balance is key to good generalization.

A15.

- **Linear Kernel**: Assumes a linear relationship between the input features, used when data is linearly separable.

- **RBF (Radial Basis Function) Kernel**: Maps data into higher dimensions, making it effective for non-linear classification.

- **Polynomial Kernel**: Captures interactions between features by using polynomial combinations of the inputs, effective for non-linear problems.

STATISTICS WORKSHEET-5

1 (d)- Expected

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