-: ML ANSWER SHEET OF WORK SHEET SET-5 :-

1. Residual sum of squares- (RSS) is the absolute amount of explained variation, whereas R-squared is the absolute amount of variation as a proportion of total variation. Mainly both are commonly used measures to assess the goodness of fit of a regression model, but they capture different aspects of model performance, and the choice between them depends on the context and what we want to evaluate.

2. Total Sum of Squares (TSS):

- Represents the total variance in the dependent variable.
- Calculated by squaring the deviations of each data point from the mean of the dependent variable.
- Essentially, it shows the total amount of variability we're trying to explain with our model.

Explained Sum of Squares (ESS):

- Represents the portion of the total variance explained by the independent variables in the model.
- Calculated by squaring the deviations of the predicted values from the mean of the dependent variable.

Residual Sum of Squares (RSS):

- Represents the portion of the total variance that the model cannot explain.
- Calculated by squaring the deviations of the actual data points from the predicted values.
- Shows how much error remains after using the model.

The following equation shows the relationship between TSS, ESS and RSS:

TSS= ESS + RSS

3. Regularization: Regularization is a technique used in machine learning and deep learning to prevent overfitting and improve the generalization performance of a model.

The following scenarios are where to use regularization:-

- When it is important to consider all the independent variables in the model.
- When many interactions are present.
- When collinearity or codependency is present amongst the variables.

4. Gini-impurity index: This index is also known as the Gini coefficient, is a measure of impurity used in decision tree learning. It quantifies the probability of a randomly chosen element from a set being misclassified if it were labeled according to the class distribution of that set. It is a non-negative value, ranging from 0 to 1.

Here's how the Gini impurity index is calculated:

-Calculate the proportion of each class in the set.

-Square each proportion and subtract it from 1.

-Sum the squares for all classes.

5. Yes unregularized decision-trees prone to overfitting because the inherent structure and learning process of unregularized decision trees make them susceptible to overfitting. Regularization techniques are crucial to introduce bias, reduce variance, and improve the generalizability of decision tree models.

6. Ensemble techniques: This technique in machine learning involve combining multiple base models to create a single and more powerful model. This powerful model leverages the strengths of individual models while mitigating their weaknesses, leading to improved performance and accuracy.

7. Bagging:

- Bagging aims to reduce the variance of the model.
- Bagging is a learning approach that aids in enhancing the performance, execution and precision of machine learning algorithms.

  Boosting:
- The boosting method tries aims to reduce the bias to avoid underfitting the data.
- Boosting is an approach that iteratively modifies the weight of observation based on the last classification.

8. Out-of-bag error in random forests: OOB error is a valuable tool for random forests, providing a convenient and informative way to estimate the model's generalization performance without requiring a separate validation set. However, it's important to be aware of its limitations and consider it alongside other metrics for comprehensive model evaluation.

9. K-fold cross-validation: This is a powerful technique in machine learning used to evaluate the performance of a model on unseen data. It works by dividing the training data into k smaller, roughly equal-sized sets called folds. Then, it iteratively trains the model on k-1 folds and tests it on the remaining fold. This process I s repeated k times, ensuring that each data point is used for both training and testing at least once.

10. Hyper parameter tuning: This is a crucial step in the machine learning model development process. It involves finding the optimal set of values for the hyper parameters of your model. Importance of hyper parameter tuning:

- It improves model performance.
- Also prevents overfitting and underfitting.
- It optimizes for specific goals

11. Issues using of Gradient Descent: A large learning rate in Gradient Descent can lead to several issues, impacting both the convergence and stability of your model training. Here are some key problems to be aware of:

- Divergence- Parameters can be too big, causing the algorithm to take large jumps instead of smoothly approaching the minimum.
- Missing the minimum point by overshoot the cost function surface.
- Instability: Large updates can lead to erratic behavior during training. This can make it harder to diagnose and solve other issues arising during training.
- Increased sensitivity to noise: This can lead to increased variance in the model's performance and reduce its overall robustness and generalizability.
- Computational inefficiency: This can significantly increase the training time and computational resources required.

12. Logistic regression is a linear classifier that produces a linear decision surface. It's fast at classifying unknown records and performs well when the dataset is linearly separable. However, it can't solve non-linear problems because it assumes a linear relationship between the input features and the output.

13. Difference between AdaBoost and Gradient Boosting:

Features:	AdaBoost:	Gradient Boosting:
Base learners	Typically decision trees (stumps).	Can use various learners (trees, linear, models, etc.)
Loss Function	Exponential loss (focuses on misclassified points)	Any differentiable loss function.

Model Building	Sequentially adjusts data weights	Sequentially fits residuals from previous predictions
Noise Sensitivity	More sensitive to outliers	Less sensitive to outliers
Flexibility	Less flexible, tied to exponential loss	More flexible, adaptable to various loss functions
Generalizability	Can over fit more easily	Often more generalizable

14. Bias-variance tradeoff: In machine learning, the bias-variance tradeoff is a fundamental concept that describes the relationship between two sources of error in model predictions (bias and variance). Finding the right balance between two errors is crucial for creating an accurate and generalizable model.

15. Linear Kernel in SVM:

- Simple, good for linearly separable data.
- Decision boundary: straight line.

RBF Kernel in SVM:

- Versatile, good for non-linear data.
- Decision boundary: Smooth curve.
- More complex, needs hyper parameters tuning.

Polynomial Kernel in SVM:

- Powerful for complex relationships but can overfit.
- Decision boundary: curvy depending on degree.
- Less interpretable than Linear and RBF.