MACHINE LEARNING

1. R-squared or coefficient of determination is a better measure of how well the regression model fits the test data compared to the Residual Sum of Squares (RSS) because it measures how well the model can use the independent variables to predict the variance in the actual Y values, and it can be used to compare the goodness of different models, as opposed to the RSS that only provides the degree of error in the model's predictions. The R-squared is defined as the difference between one and the fraction of the RSS and the sum of squared differences between the actual values and their mean.

2. The total sum of squares (TSS) measures the summed deviation (squared difference) of each actual Y value from the mean Y value. The explained sum of squares (ESS) measures the summed deviation of the predicted Y values from the mean Y value. The residual sum of squares (RSS) measures the deviation of each predicted Y values from the actual Y value. The equation that relates the three measures is TSS=ESS+RSS.

3. Machine learning needs regularization to minimise the adjusted loss function which prevents overfitting and underfitting.

4. The Gini-impurity index is a measure used in the context of the decision tree classifier together with the entropy index to indicate the level of impurity or disorder in the data, i.e. how much a given dataset is heterogenous. It is useful as it assigns the different variables to higher or lower ranks in the hierarchy of branches in the decision tree, a higher Gini value indicates the root level, whereas a lower Gini value indicates the sub-branches levels.

5. Unregularized decision trees are more prone to overfitting because as the name suggests they are not regularised therefore they do not have constraints or penalties that prevent it from capturing the noise associated with the training data, therefore they are more prone to be complex and intricated due to overfitting.

6. The ensemble technique in machine learning is a machine learning paradigm where multiple models or weak learners are trained to solve the same problem and combined to obtain more accurate results.

7. The difference between Bagging and Boosting techniques lies in the manner in which multiple models are combined to learn big datasets; the Bagging technique combines models in a parallel way, whereas the Boosting technique combines them in a sequential order.

8. The out-of-bag error (OOB) is the error that is calculated when comparing the predictions made by each tree on the remaining dataset, called out-of-bag data, and the actual values of the out-of-bag data.

9. K-fold cross-validation is a resampling technique used to enable all the data to be used as both training and testing data. Data is split into a k number of groups, whereby each group is used once as the test data and the rest of the groups are used for training. K-fold cross-validation loops through these k groups in order to evaluate a less biased performance of the model.

10. Hyper parameter tuning is a technique used to select the best parameters for each classifier in order to obtain the most accurate results.

11. A large learning rate in Gradient Descent is a fast rate to adjust the weights or parameters in a neural network or regression model in order to achieve the optimal hyperparameter/s. If the

learning rate is too high, there is a risk that the Gradient Descent algorithm will skip the optimal solution due to large steps.

12. Since linear regression creates a linear boundary that it uses to make predictions, the presence of data that has a non-linear relationship between its predictor variables entails the incapacity for linear regression to classify non-linear data.

13. The Adaboost is different from the gradient boosting in such the former iteratively calculates the weights of the data by minimising the exponential loss function, whereas the gradient boosting uses an additive model, a loss function, and a weak learner to further optimise the loss function using different differentiable loss functions. Therefore, gradient boosting is more resistant to outliers compared to the Adaboost.

14. The variance-bias trade-off consists in the inability of a learning model to highly account for variance without capturing noise as well, which is high bias.

15. The linear kernel is used for data that has a linear relationship between its predictor variables that can be separated by a straight line.

The radial basis function or gaussian kernel transfers data from a bidimensional space to a high-dimensional space to then measure the similarity between data points.

The polynomial kernel transfers data points to a multi-dimensional space and uses a polynomial function to capture the similarity and relationship between the data points.

STATISTICS

1. D

2. C

3. C

4. B

5. C

6. B

7. A

8. A

9. B

10. A