STATISTICS WORKSHEET – 5 ANSWERS

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

MACHINE LEARNING ASSIGNMENT – 5 ANSWERS

1) The percentage of the dependent variable's variation that can be predicted from the independent variables is measured by R-squared. RSS represents the sum of the squared differences between the actual and predicted values. R-squared is a better measure of goodness of fit model in regression because it provides a single number that summarizes the proportion of variance in the dependent variable that is explained by the model, which makes it easier to compare different models.
2) TSS: TSS (Total Sum of Squares) is the sum of the squared differences between each observed dependent variable value and the mean of all the dependent variable values.
   ESS: ESS (Explained Sum of Squares) is the sum of the squared differences between the predicted values of the dependent variable and the mean of all the dependent variable values.
   RSS: RSS (Residual Sum of Squares) represents the sum of the squared differences between the actual and predicted values.
   The equation relating these three metrics with each other:
   TSS=ESS+RSS
3) Regularization in machine learning prevents overfitting. When a model fits the training data too closely and is overly complicated, it is said to be overfitting and performs poorly on untrained data.
   Regularization also helps in providing a balance between bias (error due to underfitting) and variance (error due to overfitting), which leads to models with better performance.
4) Gini Impurity is a measurement is used in decision tree algorithms to quantify a dataset's impurity level or disorder. It ranges from 0 to 0.5, where 0 indicates a perfectly pure node (all instances belong to the same class), and 0.5 signifies maximum impurity (an equal distribution of classes). In decision trees, it helps in selecting the optimal split by identifying features that result in more homogeneous subsets of data, ultimately contributing to the creation of accurate and reliable predictive models.
5) When a model fits the training data too closely and is overly complicated, it is said to be overfitting and performs poorly on untrained data. Because they are too closely matched to the training set, unregularized decision trees may not be able to generalise well to new data.

They could overlook more general trends in the data in favour of identifying quirks unique to the training sets. In the absence of constraints, decision trees can continue splitting until each leaf node contains only a few instances, including noisy or irrelevant ones. As a result, the tree may memorize the training data instead of learning the underlying relationships, leading to poor generalization performance on unseen data.

6) Ensemble learning is a machine learning technique that enhances accuracy and resilience in forecasting by merging predictions from multiple models. It aims to mitigate errors or biases that may exist in individual models by leveraging the collective intelligence of the ensemble. Ensemble models in machine learning combine the decisions from multiple models to improve the overall performance.
7) Bagging is the simplest way to combine predictions that belong to the same type while Boosting is a way of combining predictions that belong to different types.
   Bagging aims to decrease variance whereas boosting aims to decrease bias.
   In bagging, each model receives equal weight whereas in boosting models are weighted according to their performance.
   Examples of Bagging: The Random Forest model uses bagging.
   Examples of Boosting: The AdaBoost uses boosting techniques.
8) Random Forests are often used for classification and regression tasks. Out-Of-Bag errors are an estimate of the performance of a random forest classifier or regressor on unseen data. The OOB error is computed using the samples that were not included in the training of the individual trees.
9) K-fold cross-validation is a technique for evaluating predictive models. The dataset is divided into k subsets or folds. The model is trained and evaluated k times, using a different fold as the validation set each time. Performance metrics from each fold are averaged to estimate the model's generalization performance.
10) When you're training machine learning models, each dataset and model needs a different set of hyperparameters, which are a kind of variable. The only way to determine these is through multiple experiments, where you pick a set of hyperparameters and run them through your model. This is called hyperparameter tuning. If we don't correctly tune our hyperparameters, our estimated model parameters produce suboptimal results, as they don't minimize the loss function. This means our model makes more errors.
11) The choice of learning rate can significantly impact the performance of gradient descent. If the learning rate is too high, the algorithm may overshoot the minimum, and if it is too low, the algorithm may take too long to converge.
12) Logistic regression cannot be used for classification of Non-linear data. Logistic regression is considered a generalized linear model because the outcome always depends on the sum of the inputs and parameters.
13) In AdaBoost, the weights of the samples are adjusted at each iteration. No reweighting of the samples take place in Gradient Boosting.
    The final model in AdaBoost is formed by combining the predictions from individual trees through a weighted sum whereas the final model in Gradient Boosting is an equal weighted sum of all the individual trees.
    In AdaBoost training process starts with a decision stump whereas Gradient Boosting uses gradient descent to iteratively fit new weak learners to the residuals of the previous ones minimizing a loss function.
14) The bias is known as the difference between the prediction of the values by the Machine Learning model and the correct value. The variability of model prediction for a given data point which tells us the spread of our data is called the variance of the model. If the

algorithm is too simple (hypothesis with linear equation) then it may be on high bias and low variance condition and thus is error prone. If algorithms fit too complex (hypothesis with high degree equation) then it may be on high variance and low bias. Finding the right balance of bias and variance is key to creating an effective and accurate model. This is called the bias-variance trade off.

15) Linear Kernel: Linear Kernel is used when the data is Linearly separable, that is, it can be separated using a single Line. It is one of the most common kernels to be used. It is mostly used when there are many Features in a particular Data Set.
RBF Kernel: Radial Basis Kernel is a kernel function that is used in machine learning to find a non-linear classifier or regression line. Radial Basis kernel is a very powerful kernel, which can give a curve fitting any complex dataset.
Polynomial Kernel: In machine learning, the polynomial kernel is a kernel function commonly used with support vector machines (SVMs) and other kernelized models, that represents the similarity of vectors (training samples) in a feature space over polynomials of the original variables, allowing learning of non-linear models.