Name-Vaishnavi Balaji
Email id-vaishnavi1716@gmail.com
Contact-6309876450

Statistics Worksheet 1

1. a
2. a
3. b
4. d
5. c
6. b
7. b
8. a
9. c
10. Normal Distribution which is also called as Gaussian Distribution is symmetric about the mean. It depicts the data near the mean are more frequent than the data far from the mean.
11. Missing data is the data that is not captured for the variable for the observation in question. Missing data can reduce the statistical strength of the analysis which can change the validity of the results. A Data Scientist doesn't intend to produce biased estimates which lead to invalid results.

    Imputation Techniques: The imputation technique replaces the missing value with a substituted value. There are many ways of imputing the missing value based on the nature of the data and the problem.

    (i) Imputation using Constant Value

    (ii) Imputation using Statistics

    (iii) Advanced Imputation Technique

    There is no exclusive method to handle missing values. Prior to application of any imputation method, it is necessary to understand the type of missing value, check the data type, skewness of the missing column and then determine which method is suitable for the particular problem.

12. A/B testing also known as split testing, refers to the random experimentation process where two or more versions of

Name-Vaishnavi Balaji
Email id-vaishnavi1716@gmail.com
Contact-6309876450

13. Mean imputation of missing values is not considered as an acceptable approach since it ignores feature correlation. By employing mean imputation variance of the data decreases while increasing the bias. Due to the reduced variance the model is less accurate and the confidence interval is narrower.
14. Linear Regression refers to the linear approach in developing a model for the relationship between dependent and independent variables.
15. There are 4 divisions into which Statistics is divided:

    (i) Mathematical Statistics-it helps in obtaining experimental and statistical distribution

    (ii) Statistical Methods-concerned with collection, tabulation and interpretation of the data, analyses the data and returns insights from the data

    (iii) Descriptive Statistics-summarizes and organizes data bases on set of characteristics, also helps in representation of data in classification and diagrammatic manner

    (iv) Inferential Statistics-utilised in finding conclusion regarding the population after the analysis on the sample drawn from it.

Name-Vaishnavi Balaji
Email id-vaishnavi1716@gmail.com
Contact-6309876450

Python Worksheet 1

1. C
2. B
3. C
4. A
5. D
6. C
7. A
8. C
9. A, C
10. A, B

Name-Vaishnavi Balaji
Email id-vaishnavi1716@gmail.com
Contact-6309876450

Machine Learning Worksheet

1. A
2. A
3. B
4. B
5. C
6. B
7. D
8. D
9. A
10. B
11. A
12. A, B
13. Overfitting occurs when the machine learning model blocks the test data and fails to perform well on unseen data. Regularization method is used to reduce errors by fitting the function appropriately on the given training data and avoid over-fitting.
14. Algorithms used for Regularization:

    (i) Ridge Regularization-modifies the over-fitted or under-fitted models by adding the penalty equivalent to the sum of the squares of the magnitude of coefficients. In statistics it is known as L2-norm.

    - Ridge method can be used when there are multiple independent variables which have multi-collinearity problem.
    - This method also helps to solve the problem when there are more parameters than samples.
    - This technique is not good for feature selection since it reduces the complexity of the model but does not reduce the number of independent variables and the coefficient is only minimized and does not lead to zero.

    (ii) LASSO Regression-which stands for Least Absolute and Selection Operator, it is similar to Ridge regression except that the penalty term includes absolute weights instead of square of weights. In statistics it is known as L1-norm.

    - In this method, the penalty has the effect of causing some of the coefficient estimates to be exactly equal to zero, which results in complete removal of some features. Hence, LASSO regression performs feature selection.
    - If the number of features is greater than the number of samples, LASSO will select n-features even if all features are relevant.
    - If there are two or more collinear variables then LASSO method selects one of them randomly which is not suitable for model interpretation.

15. In Statistics, error term is the sum of deviations of each data point from the model regression line.
    - Regression analysis is used to determine the degree of correlation between independent and dependent variable.
    - Error term in a linear regression equation corresponds to the unexplained difference between actually observed values of independent variable and the result predicted by the model.
    - Thereby, error term is a measure of how well the model reflects the relationship between the independent and dependent variable