**STATISTICS WORKSHEET-1, Python worksheet 1 and MachineLearning\_assignment**

**STATISTICS WORKSHEET-1**

1. Bernoulli random variables take (only) the values 1 and 0.

**Yes**

2. Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?

**Central Limit Theorem (CLT)**

3.Which of the following is incorrect with respect to use of Poisson distribution?  **Modeling bounded count data**

4. Point out the correct statement.

**The square of a standard normal random variable follows what is called chi-squared distribution**

5. \_\_\_\_\_\_ random variables are used to model rates.

**Poisson**

6. 10. Usually replacing the standard error by its estimated value does change the CLT.

**False**

7. 1. Which of the following testing is concerned with making decisions using data?

**Hypothesis**

8. 4. Normalized data are centered at\_\_\_\_\_\_and have units equal to standard deviations of the original data.

**0**

9. Which of the following statement is incorrect with respect to outliers?

**Outliers cannot conform to the regression relationship**

10. What do you understand by the term Normal Distribution?

**Ans:**

Normal distribution, also known as Gaussian distribution, is a symmetric probability distribution that forms a bell-shaped curve. It's characterized by a mean and standard deviation, and many natural phenomena tend to follow this pattern due to the central limit theorem.

11. How do you handle missing data? What imputation techniques do you recommend?

**Ans:**

Handling missing data involves using various techniques to fill in or estimate values for missing entries in a dataset. The recommended imputation technique depends on factors like data type, amount of missing data, and analysis goals. Common techniques include mean/median imputation, mode imputation, forward/backward fill, interpolation, K-nearest neighbors, regression, and domain-specific methods.

12. What is A/B testing?

**Ans:**

A/B testing is a method that compares two versions (A and B) of something, such as a webpage or marketing campaign, to determine which one performs better based on real-world data and user responses.

13. Is mean imputation of missing data acceptable practice?

**Ans:**

Yes, mean imputation of missing data is a common and simple practice, but it has potential drawbacks like distorting data distribution and not accounting for relationships among variables.

14. What is linear regression in statistics?

**Ans:**

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data.

15. What are the various branches of statistics?

**Ans:**

Various branches of statistics include descriptive statistics, inferential statistics, biostatistics, econometrics, social statistics, and mathematical statistics.

**Python worksheet 1**

1. Which of the following operators is used to calculate remainder in a division?

**Ans: %**

2. In python 2//3 is equal to?

**Ans: 0**

3. In python, 6<<2 is equal to?

**Ans: 24**

4. In python, 6&2 will give which of the following as output?

**Ans: 2**

5. In python, 6|2 will give which of the following as output?

**Ans: 4**

6. What does the finally keyword denotes in python?

**Ans: The finally block will be executed no matter if the try block raises an error or not.**

7. What does raise keyword is used for in python?

**Ans: It is used to raise an exception.**

8. Which of the following is a common use case of yield keyword in python?

**Ans: In defining a generator**

9. Which of the following are the valid variable names?

**Ans: \_abc, abc2**

10. Which of the following are the keywords in python?

**Ans: Yield, raise**

**programming questions (11-15)**

****

**MACHINE LEARNING**

1. Which of the following methods do we use to find the best fit line for data in Linear Regression?

**Ans: Both A and B**

2. Which of the following statement is true about outliers in linear regression?

**Ans: Linear regression is sensitive to outliers**

3. A line falls from left to right if a slope is \_\_\_\_\_\_?

**Ans: Negative**

4. Which of the following will have symmetric relation between dependent variable and independent

**Ans: Correlation**

5. Which of the following is the reason for over fitting condition?

**Ans: Low bias and high variance**

6. If output involves label then that model is called as:

**Ans: Predictive modal**

7. Lasso and Ridge regression techniques belong to \_\_\_\_\_\_\_\_\_?

**Ans: Regularization**

8. To overcome with imbalance dataset which technique can be used?

**Ans: SMOTE**

9. The AUC Receiver Operator Characteristic (AUCROC) curve is an evaluation metric for binary

**Ans: TPR and FPR**

10. In AUC Receiver Operator Characteristic (AUCROC) curve for the better model area under the

curve should be less.

**Ans: False**

11. Pick the feature extraction from below:

**Ans: Construction bag of words from a email**

12. Which of the following is true about Normal Equation used to compute the coefficient of the Linear

Regression?

**Ans: It becomes slow when number of features is very large.**

13. Explain the term regularization?

**Ans:**

**Regularization is a technique used in machine learning to prevent overfitting by adding a penalty term to the model's loss function, encouraging simpler or more generalized solutions.**

14. Which particular algorithms are used for regularization?

**Ans:**

**Popular algorithms that use regularization include Ridge Regression (L2 regularization), Lasso Regression (L1 regularization), and Elastic Net Regression (combination of L1 and L2 regularization).**

15. Explain the term error present in linear regression equation?

**Ans**

**The error in a linear regression equation represents the difference between the predicted values by the model and the actual observed values in the dataset.**