STATISTICS WORKSHEET 4

1. d) All of the mentioned
2. a) Discrete
3. a) pdf
4. c) mean
5. a) variance
6. a) variance
7. c) 0 and 1
8. b) bootstrap
9. b) summarized
10. Histograms are a special kind of bar graph that shows a bar for a range of data values instead of a single value. A box plot is a data display that draws a box over a number line to show the interquartile range of the data.

    Histograms indicate the whole frequency distribution of a variable, whereas the boxplot summarises its most prominent features. These features include median and spread as well as the extent and nature of departures from symmetry, and the possible presence of observations having extreme values (outliers).

11. Choice of metrics influences how the performance of machine learning algorithms is measured and compared. They influence how we weight the importance of different characteristics in the results.

    <u>Classification Metrics</u>
    - Accuracy.
    - Logarithmic Loss.
    - ROC, AUC.
    - Confusion Matrix.
    - Classification Report.

    <u>Regression Metrics</u>
    - Mean Absolute Error.
    - Mean Squared Error.

STATISTICS WORKSHEET 4

- Root Mean Squared Error.
- Root Mean Squared Logarithmic Error.
- R Square.
- Adjusted R Square.

In classification problems , we use two types of algorithms (dependent on the kind of output it creates):

- Class output : Algorithms like SVM and KNN create a class output. For instance, in a binary classification problem, the outputs will be either 0 or 1. SKLearn's/Other algorithms can convert these class outputs to probability.
- Probability output : Algorithms like Logistic Regression, Random Forest, Gradient Boosting, Adaboost etc. give probability outputs. Probability outputs can be converted to class output by creating a threshold probability.

In regression problems the output is always continuous in nature and requires no further treatment.

12. Hypothesis testing is used to find out the statistical significance of the insight. To elaborate, the null hypothesis and the alternate hypothesis are stated, and the p-value is calculated.
    After calculating the p-value, the null hypothesis is assumed true, and the values are determined. To fine-tune the result, the alpha value, which denotes the significance, is tweaked. If the p-value turns out to be less than the alpha, then the null hypothesis is rejected. This ensures that the result obtained is statistically significant.

13. **Exponential distributions** do not have a log-normal distribution or a Gaussian distribution.

14. Income is the classic example of when to use the median instead of the mean because its distribution tends to be skewed.

STATISTICS WORKSHEET 4

15. The likelihood is the probability that a particular outcome is observed when the true value of the parameter is , equivalent to the probability mass on ; it is not a probability density over the parameter . The likelihood, , should not be confused with , which is the posterior probability of given the data .