Machine Learning

1. A. Least Square Error
2. A. Linear regression is sensitive to outliers
3. B. Negative
4. A. Regression
5. C. Low bias and high variance
6. A. Descriptive model
7. D. Regularization
8. D. SMOTE
9. C. Sensitivity and Specificity
10. A.True
11. B. Apply PCA to project high dimensional data
12. A, B & C

Subjective answer type questions: Brief

13. **Regularization**:Regularization is a technique used to reduce the errors by fitting the function appropriately on the given training set and avoid overfitting.
    The commonly used regularization techniques are :

- L1 regularization
- L2 regularization
- Dropout regularization

14. Algorithms are used for regularization:

- Ridge Regression
- LASSO (Least Absolute Shrinkage and Selection Operator) Regression
- Elastic-Net Regression

15. An error term represents the margin of error within a statistical model; it refers to the sum of the deviations within the regression line, which provides an explanation for the difference between the theoretical value of the model and the actual observed results. The regression line is used as a point of analysis when attempting to determine the correlation between one independent variable and one dependent variable.

PYTHON

1.C %

2.B. 0

3. C.24

4.B.True

5.A.2

6. B

7.A

8.C

9. C

10.D

STATISTICS WORKSHEET-1

1.A. TRUE

2.A. Central Limit Method

3.B

4.D

5.C

6.B

7.A

8.A

9.C

10. Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve.

11. Understanding the nature of missing data is critical in determining what treatments can be applied to overcome the lack of data. Data can be missing in the following ways:

- **Missing Completely At Random (MCAR)**
- **Missing At Random (MAR)**
- **Not Missing At Random (NMAR)**

Common methods:
1. Mean or Median Imputation
2. Multivariate Imputation by Chained Equations (MICE)
3. Random Forest

12. A/B testing, also known as split testing, refers to a randomized experimentation process wherein two or more versions of a variable (web page, page element, etc.) . are shown to different segments of website visitors at the same time to determine which version leaves the maximum impact and drives business metrics.

13. The process of replacing null values in a data collection with the data's mean is known as mean imputation.

Mean imputation is typically considered terrible practice since it ignores feature correlation. Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does.

Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

14. Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a "least squares" method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).

15.Various branches of statistics:

- Descriptive Statistics
- Inferential Statistics