**Assignment 4: Machine Learning Assignment (Questions 1 to 15)**

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

***C) Logarithmic Loss***

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

1. ***Linear regression is sensitive to outliers.***

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

1. ***Negative***

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

***B) Correlation***5, Which of the following is the reason for over fitting condition?

***C, Low bias and high variance***

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

***B) Predictive Model***

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

**D) Regularization**

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

**D) SMOTE**

9, The AUC Receiver Operator Characteristic (AUCROC) curve is an evaluation metric for binary classification problems. It uses \_\_\_\_\_ to make graph?

1. **TPR and FPR**

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

1. **FALSE**

11, Pick the feature extraction from below:  
**B) Apply PCA to project high dimensional data**

12, Which of the following is true about Normal Equation used to compute the coefficient of the Linear Regression?  
  
***A) We don’t have to choose the learning rate***

13, Explain the term regularization?

***It is a method used in machine learning to prevent overfitting. It does this by shrinking the model coefficients towards zero. This helps to reduce the complexity of the model and to improve its generalization.***

14, Which particular algorithms are used for regularization?  
  
***The algorithms used for regularization include Lasso Ridge and Elastic Net***

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

***This is the difference between the observed value of the dependent variable and the predicted value by the regression model. It represents the random variation or noise that is not explained by the model. The error term can be calculated by measuring the distance between each data point and the regression line, squaring them, and taking the mean.***

**Statistics Assignment**

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

1. **TRUE**

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?

1. ***Central Limit Theorem***

3, Which of the following is incorrect with respect to use of Poisson distribution?

1. **Modelling event/time data**

4, Point out the correct statement.  
  
**a) The exponent of a normally distributed random variables follows what is called the log- normal distribution.**

5, random variables are used to model rates.  
 **d) All of the mentioned**

6, Usually replacing the standard error by its estimated value does change the CLT   
 **b) False**

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

1. **Hypothesis**

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

**a) 0**

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

**d)None of the mentioned**

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

**It is a type of probability distribution that shows how data are symmetrically distributed around the mean. It has a bell-shaped curve that is determined by two parameters: the mean and the standard deviation.**

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

**When dealing with missing data two primary methods can be used to solve the error: imputation or the removal of data. Instead of deletion, imputation methods can be used to impute the value of missing data. Some examples of these techniques are:**

* **Using the mean, median and mode**
* **K Nearest neighbours**
* **Next or Previous value**
* **Maximum or Minimum value**
* **Missing value prediction (using ML models)**
* **Fixed Value**

12. What is A/B testing?

**A/B testing which is also called split testing or bucket testing compares the performance of two versions of content to see which one appeals more to visitors or viewers. It tests a control (A) version against a variant (B) version to measure which one is most successful based on your key metrics.**

13. Is mean imputation of missing data acceptable practice?

**Mean imputation is a simple and easy method to deal with missing data, but it has some disadvantages. Some of the pros and cons of mean imputation are:**

**Pros:**

**It preserves the mean of the variable.**

**It does not reduce the sample size.**

**It is computationally simple and easy to understand.**

**It can be helpful for missing data in time series of asset movements.**

**Cons:**

**It distorts the distribution of the variable by reducing its variance and standard deviation.**

**It may introduce bias if the missing data is not random.**

**It may roll forward an anomaly-type value (outlier) if unlucky.**

**It does not account for the uncertainty of the missing data.**

**Mean imputation is generally not recommended for serious statistical analysis, unless the missing data is very small and completely random. There are other methods that can handle missing data more effectively, such as multiple imputation or regression imputation.**

14. What is linear regression in statistics?

**It is an algorithm that is used to determine the line that represents the general trend of a data set. It models the relationship between a dependent variable (the outcome) and one more independent variables (predictors/feature) using a straight line (line of best fit).**

**The formula for linear regression is:**

**y = a + bx 1**

**where y is the dependent variable, x is the independent variable, b is the slope of the line, and a is the y-intercept. The slope of the line represents the change in y for every unit change in x. The y-intercept represents the value of y when x is equal to zero.**

15. What are the various branches of statistics?

**There are two main branches of statistics. These are:**

* **Descriptive statistics: Deals with the collection of data. There are two types of descriptive statistics. These are central tendency methods and variability measures.**
* **Central tendency method includes the mean, the mode, and the median.**
* **Variability measures include quartiles, ranges, variances, and standard deviation.**
* **Inferential statistics: Techniques that enable statisticians to use the information collected from the sample to conclude, bring decisions, or predict a defined population. The different types of inferential statistics include Regression analysis, ANOVA (Analysis of variance), Analysis of covariance (ANCOVA), Statistical significance(t-test), Correlation analysis**.

**Python Worksheet (Questions 1-10)**

**1, C %**

**2, B 0**

**3, C 24**

**4, A 2**

**5, D 6**

**6, C The finally block will be executed no matter if the try block raises an error or not.**

**7, A It is used to raise an exception**

**8, C in defining a generator**

**9, A and C**

**\_abc abc2**

**10, A and B**

**yield and raise**