**MACHINE LEARNING**

Answer No. 1: C

Answer No. 2: D

Answer No. 3: C

Answer No. 4: A

Answer No. 5: C

Answer No. 6: B

Answer No. 7: C

Answer No. 8: B & C

Answer No. 9: A & D

Answer No. 10: A, B&D

Answer No. 11: An *outlier* is an observation that lies an abnormal value from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst to decide what will be considered abnormal. To channelize abnormal observation, one must have to characterize normal observations.

The **interquartile range** is a number that indicates the spread of the middle half or the middle 50% of the data. It is the difference between the third quartile (Q3) and the first quartile (Q1).

IQR = Q3 – Q1

The IQR can help determine outliers.

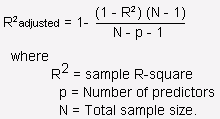
A data is a potential outlier if and only if the data is {smaller than Q1 - 1.5 \* IQR

Or larger than Q3 + 1.5 \* IQR)

Answer No. 12: Bagging is a method of merging the same type of predictions while boosting is a method of merging different types of predictions.

Bagging decrease variance, not bias, and solves over-fitting issues in a model while in boosting decreases bias, not variance.

In bagging, each model receives an equal weight. In boosting, models are weighed based on their performance, models are built independently in bagging while in boosting new models are affected by a previously build model’s performance.

Answer No. 13: It measures the proportion of variation explained by only those independent variables that really help in explaining the dependent variable. It penalizes you for adding independent variable that do not help in predicting the dependent variable. [](https://4.bp.blogspot.com/-qEGt3DaQIF0/V2meLITZj3I/AAAAAAAAEp4/WKCs0FrI1JsovDMwaw1r1iUboULfRI7MwCLcB/s1600/stb1.png)

Answer No. 14: Cross-validation is a statistical method used to estimate the performance (or accuracy) of machine learning models. It is used to protect against overfitting in a predictive model.

Advantage: - **Reduces Overfitting:** In Cross Validation, we split the dataset into multiple folds and train the algorithm on different folds. This prevents our model from overfitting the training dataset. So, in this way, the model attains the generalization capabilities which is a good sign of a robust algorithm.

Disadvantage: - **Needs Expensive Computation:** Cross Validation is computationally very expensive in terms of processing power required.

**STATISTICS WORKSHEET-4**

Answer No. 1: The Central Limit Theorem tells us that as sample sizes get larger, the sampling distribution of the mean will become normally distributed*,*even if the data within each sample are not normally distributed*.*

The Central Limit Theorem is important for statistics because it allows us to safely assume that the sampling distribution of the mean will be normal in most cases. This means that we can take advantage of statistical techniques that assume a normal distribution

Answer No. 2: Sampling is a technique of selecting individual members or a subset of the population to make statistical inferences from them and estimate characteristics of the whole population.

**Probability sampling:**  is a sampling technique where a researcher sets a selection of a few criteria and chooses members of a population randomly.

**Non-probability sampling:** the researcher chooses members for research at random. This sampling method is not a fixed or predefined selection process.

Answer No. 3: Type 1 error, in statistical hypothesis testing, is the error caused by rejecting a null hypothesis when it is true. Type II error is the error that occurs when the null hypothesis is accepted when it is not true

Answer No. 4: Normal 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. Normal distribution will appear as a bell curve in graph formation.

Answer No. 5:

**Correlation –**

1. It shows whether and how strongly pairs of variables are related to each other.
2. Correlation takes values between -1 to +1, wherein values close to +1 represents strong positive correlation and values close to -1 represents strong negative correlation.
3. In this variable are indirectly related to each other.
4. It gives the direction and strength of relationship between variables.

**Covariance –**

1. It is the relationship between a pair of random variables where change in one variable causes change in another variable.
2. It can take any value between -infinity to +infinity, where the negative value represents the negative relationship whereas a positive value represents the positive relationship.
3. It is used for the linear relationship between variables.
4. It gives the direction of relationship between variables.

Answer No. 6: Univariate contains only one variable. The purpose of the univariate analysis is to describe the data and find patterns that exist within it.

Bivariate involves two different variables. The analysis of this type of data deals with causes and relationships and the analysis is done to determine the relationship between the two variables.

Multivariate compare more than two variables.

Answer No. 7: A sensitivity analysis determines how different values of an independent variable affect a particular dependent variable under a given set of assumptions.

Sensitivity analysis works on the simple principle**: Change the model and observe the behavior. It is calculated as: TN/TN+FP.**

Answer No. 8: A statistical hypothesis is an assertion or conjecture concerning one or more populations. To prove that a hypothesis is true, or false, with absolute certainty, we would need absolute knowledge. That is, we would have to examine the entire population. Instead, hypothesis testing concerns on how to use a random sample to judge if it is evidence that supports or not the hypothesis

Ho: the null hypothesis

H1: the alternate hypothesis.

The hypothesis we want to test is if H1 is “likely” true.

So, there are two possible outcomes:

• Reject H0 and accept H1 because of sufficient evidence in the sample in favor or H1;

• Do not reject H0 because of insufficient evidence to support H1.

Answer No. 9: **Quantitative**data are measures of values or counts and are expressed as numbers.

Quantitative data are data about numeric variables (e.g. how many; how much; or how often).  
**Qualitative**data are measures of 'types' and may be represented by a name, symbol, or a number code.  
Qualitative data are data about categorical variables (e.g. what type).

Answer No. 10: The **interquartile range** is a measure of where the ‘middle fifty’ is in a data set. Where a range is a measure of where the beginning and end are in a set.

The interquartile range formula is the first quartile subtracted from the third quartile: IQR = Q3 – Q1

Answer No. 11. The term "bell curve" is used to describe a graphical depiction of a normal probability distribution, whose underlying standard deviations from the mean create the curved bell shape.

Answer No. 12. Outlier values can be measured by using boxplot or by using any graphical method. If the outliers are few it can be treated individually but for the large number of outliers it can be treated with mean or the percentile values.

To treat the outlier values one can bring it in the range of the columns values or he can simply drop/remove them.

Answer No. 13. The P value, or calculated probability, is the probability of finding the observed, or more extreme, results when the **null hypothesis (H0)** of a study question is true – the definition of ‘extreme’ depends on how the hypothesis is being tested. P is also described in terms of rejecting **H0** when it is actually true, however, it is not a direct probability of this state

Answer No. 14. A **binomial distribution** can be thought of as simply the probability of a SUCCESS or FAILURE outcome in an experiment or survey that is repeated multiple times. The binomial is a type of distribution that has **two possible outcomes.**

 For example, a coin toss has only two possible outcomes: heads or tails and taking a test could have two possible outcomes: pass or fail.

- The first variable in the binomial formula, n, stands for the number of times the experiment runs.

- The second variable, p, represents the probability of one specific outcome.

Answer No. 15. ANOVA (Analysis of Variance) is used when we have more than two sample groups and determine whether there are any statistically significant differences between the means of two or more independent sample groups.

- *One Way ANOVA* **–** It is also known as one factor ANOVA. Here, we are using one criterion variable (or called as a factor) and analyze the difference between more than two sample groups. Suppose in glass industry, we want to compare the variation of three batches (glass) for their average weight (factor).

- *Two Way ANOVA* **–**Here, we are using two independent variables (factors) and analyze the difference between more than two sample groups. Similarly, we want to compare the variation of three batches of glass w.r.t weight and hardness (two factors).

**SQL WORKSHEET-4**

Answer No. 1: WITH x AS (SELECT `shippedDate`, COUNT(`orderNumber`) AS `total\_orders` FROM Orders) SELECT AVG(`total\_orders`) AS `AverageNumberOfOrdersShipped` FROM x;

Answer No. 2: WITH x AS (SELECT `orderDate`, COUNT(`orderNumber`) AS `total\_orders` FROM Orders) SELECT AVG(`total\_orders`) AS `AverageNumberOfOrdersPlaced` FROM x;

Answer No. 3:

SELECT `productName` FROM Products

ORDER BY MSRP

LIMIT 1;

Answer No. 4: SELECT `productName` FROM Products

ORDER BY `quantityInStock` DESC LIMIT 1

Answer No. 5:

Answer No. 6:

Answer No. 7:

SELECT `customerNumber`, `customerName` FROM Customers

WHERE `city`= "Melbourne";

Answer No. 8:

SELECT `customerName` FROM Customers

WHERE `customerName` REGEXP `^N\*`;

Answer No. 9:

SELECT `customerName` FROM Customers

WHERE `phones` REGEXP “^7.\*” AND `city` = “Las Vegas”;

Answer No. 10:

SELECT `customerName` FROM Customers

WHERE `creditLimit` < 1000 AND `city` IN ("Las Vegas", "Nantes", "Stavern");

Answer No. 11:

SELECT `orderNumber` FROMorderDetails

WHERE `quantityOrdered` < 10;

Answer No. 12:

SELECT `orderNumber`

FROM Customers AS a INNER JOIN orders AS b ON a.customerNumber = b.customerNumber WHERE `customerName` REGEXP "^B.\*";

Answer No. 13:

SELECT `customerName`

FROM Orders AS a INNER JOIN Customers AS b ON

a. customer Number=b

Answer No. 14:

SELECT `customerName`

FROM Payments INNER JOIN Customers USING (`customerNumber`)

Answer No. 15:

SELECT `checkNumber` FROM Payments

WHERE `amount` > 1000;