ASSIGNMENT 4

MACHINE LEARNING

1. **C]** Between -1 and 1
2. **B]** PCA
3. **A]** Linear
4. **A]** Logistic Regression
5. **D]** Cannot be determined
6. **B]** Increase
7. **A]** Random Forests reduce overfitting
8. **D]** all of the above
9. **D]** Indentifying different segments of diseases based on BMI, blood pressure, cholesterol, blood sugar levels
10. **D]** min_ samples_ leaf
11. An outlier is a single data point that goes far outside the average value of a group of statistics. Outliers may be expectations that stand outside individual samples of population as well. To detect the outliers using this method, we define a new range, let's call it decision range, and any data point lying outside this range is considered as outlier and is accordingly **Quartile Range or IQR.**

    IOR = Q3 – Q1

12. Bagging is the simplest way of combining predictions that belong to the same type while Boosting is a way of combining predictions that belong to different types. **Bagging aims to decrease variance, not bias while Boosting aims to decrease bias, not variance.**

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. Adjusted R-squared value can be calculated based on value of r-squared, number of independent variables (predictors), total sample size.

    R2 adjusted = 1-(1 – R2) (N – 1)/N – p -1

14. In Normalization, the changes in values is that they are at a standard scale without distorting the difference in values. Whereas, Standardization assumes that the dataset is in Gaussian distribution and measures the variable at different scales, making all the variables equally contribute to the analysis. **Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1.**

15. **Cross – Validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segment :** one used to learn or train a model and the other used to validate the model.

ADVANTAGES: Cross-Validation is a statistical method used to estimate the performance (or accuracy) of machine learning model. It is used to protect against overfitting in a predictive model, particularly in a case where the amount of data may be limited.

DISADVANTAGES: The training algorithm has to be rerun from scratch **k time**, which means it takes k times as much computation to make an evaluation. A variant of this method is to randomly divide the data into a test and training set k different times.

STATISTICS

1. The CLT is a statistical theory that states that – if you take a sufficiently large sample size from a population with a finite level o variance, the mean of all sample from that population will be roughly equal to the population mean. 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 advantages of statistical techniques that assume a normal distribution, as we will see in the next section.

2. A Sample is a subset of individuals from a large population. **Sampling** means selecting the group that you will actually collect the data from in your research. For example, if you are researching the opinions of students in your university, you could survey a sample of 100 students. In statistics, Sampling allows you to test a hypothesis about the

characteristics of a population. There are two primary types of sampling methods that you can use in research:

- **PROBABILITY SAMPLING** involves random selection, allowing you to make strong statistical inferences about the whole group. It is mainly used in **quantitative research.**

1. **Simple Random Sampling :-** In a simple random sample, every member of the population has an equal chance of being selected. Your sampling frame should include the whole population.
2. **Systematic Sampling :-** Systematic sampling is similar to simple random sampling, but it is usually slightly easier to conduct. Every member of the population is listed with a number, but instead of randomly generating numbers, individuals are chosen at regular intervals.
3. **Stratified Sampling :-** Stratified Sampling involves dividing the population into subpopulations that may differ in important ways. It allows you draw more precise conclusion by ensuring that every subgroup is properly represented in the sample.
4. **Cluster Sampling :-** Cluster sampling also involves dividing the population into subgroup, but each subgroup should have similar characteristics to the whole sample. Instead of sampling individuals from each subgroup, you randomly select entire subgroup.

- **NON-PROBABILITY SAMPLING** involves non-random selection based on convenience or other criteria, allowing you to easily collect data. It is often used in **exploratory** and **qualitative research.**

1. **Snowball Sampling** :- If the population is hard to access, snowball sampling can be used to recruit participants via other participants. The number of people you have access to "snowball" as you get in contact with more people. The downside here is also representativeness, as you have no way of knowing how representative your sample is due to the reliance on participants recruiting others.

3.TYPE – 1 ERROR :-

- It is also known as a false – positive.
- It occurs if the researcher rejects a correct null hypothesis in the population i.e ., incorrect rejection of the null hypothesis.

- Measured by alpha (significance level)
- If the significance level is fixed at 5%, it means there are about five chance of type- 1 error out of 100.
- The significance level is decided before testing the hypothesis.
- Sample size is not considered.
- It can be reduced by decreasing the level of significance.

TYPE – 2 ERROR :-

- It is also known as a false- negative.
- It occurs if a researcher fails to reject a null hypothesis that is actually a false hypothesis.
- Measured by beta (the power of test)
- The probability of committing a type-2 error is calculated by 1-beta (the power of test)
- A statistical test is not powerful enough.
- It is caused by a smaller sample size, it may hide the significance level of the items being tested.
- It can be reduced by increasing the level of significance.
- It can be reduced by decreasing the level of significance.

4.A normal distribution is a type of continuous probability distribution in which most data points cluster toward the middle of the range, while the rest taper off symmetrically toward either extreme. The middle of the range is also known as the mean of the distribution. The normal distribution is also known as a Gaussian distribution or probability bell curve. It is symmetric about the mean and indicates the values near the mean occur more frequently than the value that are farther away from the mean.

5.Covariance is a statistical term that refer to a systematics relationships between two random variables in which a change in the other reflects a change in one variable. The covariance value can range from -0 and +0, with a negative value indicating a negative relationship and a positive value indicating a positive relationship. The greater this number, the more reliant the relationship. Positive covariance denotes a direct relationship and is represented by a positive number. A negative number, on the other hand denotes negative covariance, which indicates an inverse relationship between

the two variables. In statistics, correlation is a measure that determined the degree to which two or more random variable move in sequence. When an equivalent movement of another variable reciprocates the movement of one variable in some way or another during the study of two variables, the variables are said to be correlated. Positive correlation occurs when two variables move in same direction. When variables move in the opposite direction, they are said to be negatively correlated.

6.Univariate Analysis :- Univariate analysis is the most basic form of statistical data analysis technique. When the data contains only one variable and doesn't deal with a causes or effect relationships then a Univariate analysis technique is used. Univariate Analysis is conducted through several ways such as **Histograms, Pie charts, Bar charts, Frequency Distribution Tables, Frequency Polygons,** etc.

Bivariate Analysis :- Bivariate analysis is slightly more analytical than univariate analysis. When the data set contains two variables and researchers aim to undertake comparisons between the two data set then Bivariate analysis is the right type of analysis technique. Bivariate Analysis is conducted using **Correlation coefficients & Regression analysis.**

Multivariate Analysis :- Multivariate analysis is more complex from of statistical analysis technique and used when there are more than two variables in the data set. Commonly used Multivariate Analysis are **Cluster analysis, Variance analysis, Discriminant analysis, Multidimensional scaling,** etc.

7. Sensitivity Analysis is an analysis technique that work on the basis of what-if analysis like how independent factors can affect the dependent factor and is used to predict the outcome when analysis is performed under certain conditions. It is commonly used by investors who takes into consideration the conditions that affect their potential investment to test, predict and evaluate result.

$$Z = X2 + Y2$$

8. Hypothesis Testing is a type of **Statistical Analysis** in which you put assumption about a population parameter to the test. It is used to estimate the relationship between **two** statistical variable. In hypothesis testing there are two mutually exclusive hypotheses; **The Null Hypothesis (H0)** and **The**

Alternative Hypothesis **(H1).** In two tails, the test sample is checked to be greater or less than a range of values in **Two – Tailed Testing**, implying that the critical distribution area is two sided. If the sample falls within this range, the alternate hypothesis will be accepted, and the null hypothesis will be rejected.

9. Quantitative data are measures of values or counts and are expressed as numbers. Quantitative data are data about numeric variable (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 variable (e.g what type).

10. To calculate the range, you need to find the largest observed value of a variable (the maximum) and the subtract the smallest observed value (the minimum). The range only takes into account these two values and ignore the data points between the two extremities of the distribution. It's used as a supplement to other measures, but it is rarely used as the sole measure of dispersion because it's sensitive to extreme values. The interquartile range and semi-interquartile range give a better idea of the dispersion of data. To calculate these two measures, you need to know the values of the lower and upper quartiles. The lower quartile , or first quartile (Q1), is the value under which 25% of data points are found when they are arranged in increasing order. The upper quartile, or third quartile (Q3), is the value under which 75% of data points are found when arranged in increasing order. The median is considered the second quartile (Q2). The interquartile range is the difference between upper and lower quartile.

11. A bell is a type of graph that is used to visualize the distribution of a set of chosen values across a specified group that tend to have a central, normal values, as peak with low and high extremes tapering off relatively symmetrically on either side. Bell curves are visual representation of normal distribution, also called **Gaussian Distribution.** Bell curves are useful for quickly visualizing a data set's **Mean, Mode and Median** because when distribution is normal, the mean, median and mode are all the same.

12. Data Visualization Method :- You can use software to Visualize your data with a box plot, or a box-and-whisker-plot, so you can see the data distribution at a glance. This type of chart highlight minimum and maximum values (the range), the median, and the interquartile range for your data.

13. In the P-Value approach to hypothesis testing, a calculated probability is used to decide if there is evidence to reject the null hypothesis, also known as a conjecture. The conjecture is the initial claim about a data population, while the alternative hypothesis ascertains if the observed population parameter Effectively, the significance level is declared in advance how small the P- Value need to be such that the null hypothesis is rejected. The level of significance vary from one researcher to another, so it can get difficult for reader to compare results from two different tests. That is when P-value makes things easier. Reader could interpret the statistical significance by referring to the reported P-value of the hypothesis test. This is known as the P-value approach to hypothesis test.

14. Binomial Probability Formula is used to find the probability of getting a certain number of success, like successful basketball shots, out of a fixed number of trials. we use the binomial distribution to find discrete probabilities.

15. Analysis of variance, or ANOVA, is a statistical method that separates observed variance data into different component to use for additional tests. A one way ANOVA is used for three or more groups of data, to gain information about the relationship between the dependent and independent variable. If no true variance exists between the groups, the ANOVA's F – ratio should equal close to 1.

SQL

2.
```
SELECT date(order _placed_ date)
        COUNT(id) AS num _ orders
        SUM(order _total) AS daily_ total
FROM [Table]
GROUP BY date(order _placed_date)
```

3.
```
SELECT COUNT (Customer _num)
        AVG(Price)
FROM Products
```

6.
```
SELECT *
```

```
FROM table
WHERE(sal IN ( SELECT TOP  sal
                    FROM table as table1
                    GROUP BY sal
                    ORDER BY sal DESC))
```

11.

```
SELECT (ord_ no, ord_ date, purch_ amt)
   FROM orders
   WHERE quant _ ord < 10
```