Stats

1) Central tendency –

Central tendency is measuring the centre of the dataset which is used to know the location of datapoints and spread of datasets

2) Sampling and its types

- **It is selection of less than 100% datasets from population, it is called sampling,**

Types of sampling

a) **Stratified sampling**
b) **Random sampling**
c) **Cluster sampling**
d) **Systematic sampling**

3) Normal distribution

- **Normal distribution is the gaussion distribution, normal distribution is a form presenting data by arranging the probability distribution of each value in data. It is bell curved distribution**

4) Method to find outliers

- **Inter quantile range in which we calculate the higher fence and lower fence to find out outliers in dataset**
- **Lower fence formula Q1 – 1.5*IQR**
- **Higher fence formula Q3 + 1.5*IQR**

5) Type 1 error and type 2 error

- **Type 1 error - type 1 error are those error when null hypothesis of experiment is true but you still reject it called type 1 error**

- **Type 2 error - type 2 error are those error when null hypothesis is false but fail to reject it called type 2 error**

6) Hypothesis test. Ho and h1

- **Hypothesis test is the assumption, when we test the assumption after collecting the enough evidence to form conclusion, there can be two possible outcome**
- **1) null hypothesis – null hypothesis means assumption made before collecting evidence and test is right called null hypothesis**
- **2) alternative hypothesis - when you have reason to believe on basis of evidence previous assumption is not right and choose alternative assumption**

7) Covariance and correlation

- **Correlation is the measurement of relationship between two variables which could result between from -1 to 1.**
- **There are two type of relationship**
- **1) negative correlation**
- **2) positive correlation**

- **Covariance is the indicator which shows the dependency of one variable on other variable is called covariance which is usually used for measuring the relationship of dependent and independent variables**

8) P value hypothesis

- **P value hypothesis is probability value which is used to find whether null hypothesis is rejected or accepted. If p value is more than alpha then null hypothesis is rejected and if p value Is less than alpha then null hypothesis is rejected. P value is 0.05**

9) Range and interquartile range

- **Range is spread of dataset from lowest point to highest point called range**
- **Interquartile range is difference between Q3 and Q1**

10) Bell curve distribution

- **Bell curve distribution is the type of graph used to visualize the distribution of data on the basis of central tendency.**

11) Anova test

- **Anova test is tool which measure the significant difference between two categorical groups by testing difference of means using variance**

12) –

- **Univariate analysis is the analysis of single variable to determine the normal distribution**
- **Bivariate analysis is the examination of two variables**
- **Multivariate analysis is the examination of more than two variables simultaneously to determine the relationship**

Machine learning

1) **C) between -1 to 1**
2) **D) ridge regularisation**
3) **B) radial basis function**
4) **B) naïve bayes classifier**
5) **B) same as old coefficient of x**
6) **B) increase**
7) **C) random forest is easy to interpret**
8) **B) principle component are calculated using unsupervised learning technique**
9) **A) identifying the developed, under developed and under developing on basis of factors**
   e) Identifying the different segments of disease based on bmi , blood pressure and cholesterol
10) **A) max_depth**
    B) n_estimators
    d) min_sample leafs

11) Outliers and interquartile range

- Outliers are the datapoints which are significantly different from usual dataset called outliers

- Interquartile range is the difference between Q3 and Q1 which contains the 50% of data within these points

12) Bagging is ensemble technique used to improve the stability and accuracy of regression and classification model. It helps to decrease the variability and avoid the problem of overfitting. It decrease the variance not bias

- **Boosting is the attempt to build powerful and strong model from weaker models called boosting . it decrease the bias but not variance**

13) -

Adjusted r squared is the modified r2 score by considering predictors are not necessary and adding additional features

Adjusted R squared is calculated by dividing the residual mean square error by the total mean square error, the result is subtracted from 1

14) Normalisation is used to transform the feature to be on similar scale. This scale range between 0 to 1. It involves change in numeric columns in datasets to use a common scales. It cannot deal with outliers. It is useful when we don't know about distribution

- **Standardization**

- **Mean and standard deviation is used for scaling. It is bounded to certain range**
- **It is less affected by outliers**
- **It is also called standardscaler**
- **It is useful when database is normally distributed**
- **It is also called z score normalisation**

15) Cross validation

- **It is technique of resampling the dataset in order to evaluate the machine learning model. Parameters k donates number of sample**

Advantages –

1) **It prevents the model from overfitting from training datasets, chance of overfitting is less if dataset is large, cross validation is not required if dataset is enough or small**
2) **Cross validation helps in determining the accurate estimate of model prediction**

Disadvantage –

1) **It is required more time to implement the model while training**
2) **Cross validation is expensive to implement**

SQL

3) result = cursor.execute("SELECT productname, MSRP FROM products")

4) result = cursor.execute('SELECT productname FROM products WHERE quantityinstock = (SELECT MAX(quantityinstock) FROM products)')

5) RESULT=cursor.execute("SELECT products.productname, MAX(orderdetails.quantityordered) AS MAXORDER FROM products JOIN orderdetails ON products.productcode = orderdetails.productcode WHERE quantityordered = 'MAXORDER' ")

6) result = cursor.execute("SELECT customers.customername, MAX(amount) as pay FROM payments WHERE customers.customernumber = payments.customernumber ORDER BY PAY DESC")

7) result = cursor.execute("SELECT customername, customername, city FROM customers WHERE city = 'Melbourne city%'")

8) RESULT = cursor.execute("SELECT *FROM customers WHERE customername LIKE 'N%'")

9) RESULT = cursor.execute("SELECT *FROM customers WHERE phone LIKE '7%' AND city = 'Las Vegas'")

10) result= cursor.execute("""SELECT customername FROM customers WHERE creditlimit < 1000 AND city = 'Las Vegas' OR 'Nantes' OR 'Stavern'""")

11) result = cursor.execute("SELECT ordernumber FROM orderdetails WHERE quantityordered > 10")

12) result=cursor.execute("SELECT ordernumber FROM orders LEFT JOIN customers ON customers.customernumber = orders.customernumber WHERE customers.customername = 'N%'")

13) result= cursor.execute("SELECT customername FROM customers WHERE customernumber=(SELECT customernumber FROM orders WHERE status = 'Disputed')")

14) RESULT = cursor.execute("""SELECT customername, checknumber FROM customers JOIN payments ON customers.customernumber = payments.customernumber WHERE checknumber = 'h%' AND paymentdate = 2004-10-19""")

15) result=cursor.execute("SELECT checknumber FROM payments WHERE amount > 1000")