1 – DYNAMIC PROGRAMMING – it is Automatically select the data type.

2- FUNCTIONAL PROGRAMMING – it is nothing but it directly prints the value.

3- FUNDAMENTAL DATA TYPE- it can store only 1 value in one object.

4 ADVANCED DATA TYPE – it can able to store in multiple value in 1 object.

5- TYPE CASTING – it converts from one data type to another data type.

6- Can string converted into Integer.

Ans:- No, when it consist only text or with in string it can't be converted , only numeric value is converted.

7-INDEXING :- is nothing it store the single value of each characters.

- Positive indexing starts from 0 to length (-1)
- Negative indexing starts from -1 to -length.

8-Que:- What Is Difference Between Immutable And Mutable Object.

Ans:- Mutuable object- modifying the value within the same object is call mutable object.

Not able to modify the value with in the same object is called Immutable object.

9-What Is Difference B/W A Function And Method.

Ans:- Function -work irrespective of data type,

Method – a function which work with a particular data type.

10-difference b/w inner join and outer join merge in data frame.

Ans: - inner join by default it considered common value. While outer join considered all the value.

11- difference b/w append dataFrame and concat dataframe

Ans:- in append data frame only one data frame add and concat multiple date frame add a time.

Pandas and numpy are mutable

Numpy works on the element wise.

12- what is flattening?

Ans:- flattening is converting n-dimension array to 1-dimension array by using raval function .

13- difference b/w the stratified and cluster sampling

Ans:- in stratified ,population are divided into ratio but in cluster sampling ratio not work(ex-covid zone)

14:- how to identify the outliers.

Ans:- by using box plot we can identify the outliers.

15- how the treat the outliers.

Ans:- outlier treated by 3R technique-3R= remove , replace(by the max and min value of wishks) and retrain

15- how the treat missing value.

Ans:- remove the row and replace nan value or research value.

16- why required the heat map

Ans :- heat map are used to find out the correlation b/w the data.

17- what we different type of encoding techniques

Ans:-nominal data- nominal encoding- all category give equal weightage

Ordinal data- ordinal encoding & different weightage for each different category.

18- what do you mean dummy variable trap

Ans: - converting n columns to n-1 columns.

19- what is difference b/w label encoder and ordinal encoder

Ans: - labels encoding will take alphabetically orders of each categories

Ordinal encoder- will take in terms of order we have given.

20- what is decision tree

Ans-decision tree are supervised machine learning technique where the data is split according to certain parameter or condition.

we can solve both classification and regression problem statements.

Decision Tree is also a base tree that is used in Bagging and Boosting techniques such as Random Forest and Xgboost Classification And Regression Algorithms.

21- what is advantage of decision tree

Ans= -1. clear visualization

2.Simple and easy to under stand

3.Decision Tree can be used for both classification and regression problems.

4.Decision Tree can handle both continuous and categorical variables.

5.No feature scaling required.

22- disadvantage of decision tree

Ans = 1. overfitting problem occurs because train accuracy always 100%

2. unstable

3. not suitable for large datasets

23- what is hyperparameter tuning

Ans – there are different parameters with in the model , we r going to change different different parameters and identify which is the best parameter.

24- what is tuning

Ans- identifying the best parameter called tunning

Hyperparameter means – model parameter

25- what is optimization

Ans- identifying which is the best model and according to that your optimizing or changing which is reduce the overfitting problems.

26-what information gain in decision tree

Ans- Information gain is the difference between the entropy of a data segment before the split and after the split. The high difference represents high information gain.

- Information should be high and entropy should be low

27- Bessel's correction is the use of $n - 1$ instead of n in the formula for the sample variance and sample standard deviation,[1] where n is the number of observations in a sample. This method corrects the bias in the estimation of the population variance.

Bessel’s correction is a factor that is used to estimate a populations’ standard deviation from its sample. It causes the standard deviation to be less biased, thereby, providing more accurate results.

28- Standard deviation is a statistic that measures the dispersion of a dataset relative to its mean and is calculated as the square root of the variance.

29- Variance is a measure of how data points differ from the mean or a measure of how data points differ from the mean.

29- Variance is the average squared deviations from the mean, while standard deviation is the square root of this number. Both measures reflect variability in a distribution, but their units differ.

Standard deviation measures how far apart numbers are in a data set. Variance, on the other hand, gives an actual value to how much the numbers in a data set vary from the mean.

30- What is Linear Regression

Ans - Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It’s used to predict values within a continuous range, (e.g. sales, price) rather than trying to classify them into categories.

31- How does a Non-Linear regression analysis differ from Linear regression analysis.

Ans- Non-linear functions have variables with powers greater than 1.

- *Linear* functions have variables with only powers of 1. They form a straight line if it is graphed.
- Non-linear regression analysis tries to model a non-linear relationship between the independent and dependent variables.

- *Linear* regression analysis tries to model a linear relationship between the independent and dependent variables.

32- How is the Error calculated in a Linear Regression model?

Ans- The smaller the Mean Squared Error, the closer you are to finding the *line of best fit*

33-How would you detect Overfitting in Linear Models?

Ans- The common pattern for overfitting can be seen on learning curve plots, where model performance on the training dataset continues to improve. So, an overfit model will have extremely low training error but a high testing error.

34- Name a disadvantage of R-squared and explain how would you address it?

Ans- R-squared (R^2) is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.

drawback of R-squared is that its values can increase if we add predictors to the regression model, leading to a possible *overfitting*.

35- What are the Assumptions of Linear Regression?

Ans- There are four assumptions associated with a linear regression model:

Linearity: The relationship between X and the mean of Y is linear.

Homoscedasticity: The variance of residual is the same for any value of X.

Independence: Observations are independent of each other.

Normality: For any fixed value of X, Y is normally distributed.

36- Why would you use Normalisation vs Standardisation for Linear Regression?

Ans- Normalization transforms your data into a range between 0 and 1

Standardization transforms your data such that the resulting distribution has a mean of 0 and a standard deviation of 1

37- What is multi-collinearity?

Ans- Co-linearity is the relationship between two variables. Multi-collinearity is the relationship between more than two variables.

38- What are the key matrices used to check the performance of logistic regression?

Ans- **Accuracy — (True positive + True negative) / Total cases**

Error Rate — (False positive + False negative) / Total cases

Sensitivity — True positive / Total actual positive

Specificity — True negative / Total actual negative

Positive pred value — True positive / Total predicted positive

Negative pred value — True negative / Total predicted negative

KS — it measures the distance between cumulative good and cumulative bad. The maximum distance is KS.

AUCROC — measures the performance of the model across all cut-offs. Sensitivity is on the y-axis and 1-specificty is on the x-axis

Gain chart — positive prediction rate is on y-axis and percentage of cases allocated to event is on x-axis.

39- What is heteroscedasticity?

Ans- Heteroscedasticity is exactly the opposite of homoscedasticity, which means that the error terms are not equally distributed. To correct this phenomenon, usually, a log function is used.

40- What is feature engineering? How do you apply it in the process of modelling?

Ans: Feature engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models

41- What is the use of regularization?

Ans: Regularization is a technique that is used to reduce the overfitting problem and reduce the model complexity.

42- How to choose the value of the parameter learning rate (α)?

Ans: If the value is too small, the gradient descent algorithm takes ages to converge to the optimal solution. On the other hand, if the value of the learning rate is high, the gradient descent will overshoot the optimal solution and most likely never converge to the optimal solution.

43- What is VIF? How do you calculate it?

Ans: Variance Inflation Factor (VIF) is used to check the presence of multicollinearity in a dataset.

44- Explain gradient descent with respect to linear regression.

Ans:- Gradient descent is an optimization algorithm. In linear regression, it is used to optimize the cost function and find the values of the βs (estimators) corresponding to the optimized value of the cost function.

45- Ridge regression:

Ans: - it is a type of regularization in which it reducing the overfitting problem by adding the penalty term to alpha* slope^2

46: what is machine learning

Ans - “In classic terms, machine learning is a type of artificial intelligence that enables self-learning from data and then applies that learning without the need for human intervention. OR identifying the relation between the input vs output variable without writing any code.

47- How does supervised machine learning work?

Ans- to train the algorithm with both labeled inputs and desired outputs. Supervised learning algorithms are good for the following tasks:

Binary classification: Dividing data into two categories.

Multi-class classification: Choosing between more than two types of answers.

Regression modeling: Predicting continuous values.

Ensembling: Combining the predictions of multiple machine learning models to produce an accurate prediction.

48- How does unsupervised machine learning work?

Ans- Unsupervised machine learning algorithms do not require data to be labeled. They sift through unlabeled data to look for patterns that can be used to group data points into subsets.

Unsupervised learning algorithms are good for the following tasks:

Clustering: Splitting the dataset into groups based on similarity.

Anomaly detection: Identifying unusual data points in a data set.

Association mining: Identifying sets of items in a data set that frequently occur together.

Dimensionality reduction: Reducing the number of variables in a data set.

49 - Who's using machine learning and what's it used for?

Ans- Today, machine learning is used in a wide range of applications. Perhaps one of the most well-known examples of machine learning in action is the recommendation engine that powers Facebook's news feed.

-Customer relationship management.

-Business intelligence

-Human resource information systems.

-Self-driving cars.

-Virtual assistants.

50- What do you understand by Natural Language Processing?

Ans - Natural Language Processing is a field of computer science that deals with communication between computer systems and humans.

51- List any two real-life applications of Natural Language Processing.

Ans -google assistant & chat boat

52- What are stop words?

Ans- Stop words are said to be useless data for a search engine. Words such as articles, prepositions, etc. are considered as stop words. There are stop words such as was, were, is, am, the, a, an, how, why, and many more.

53- What is NLTK?

Ans- NLTK is a Python library, which stands for Natural Language Toolkit. We use NLTK to process data in human spoken languages. NLTK allows us to apply techniques such as parsing, tokenization, lemmatization, stemming, and more to understand natural languages.

54- What is Syntactic Analysis?

Ans - Syntactic analysis is a technique of analyzing sentences to extract meaning from it. Using syntactic analysis, a machine can analyze and understand the order of words arranged in a sentence.

55- What is Semantic Analysis?

Ans- Semantic analysis helps make a machine understand the meaning of a text. It uses various algorithms for the interpretation of words in sentences. It also helps understand the structure of a sentence.

56 - What is TF-IDF?

Ans- TFIDF or Term Frequency-Inverse Document Frequency indicates the importance of a word in a set.When TF*IDF is high, the frequency of the term is less and vice versa.

57- What are unigrams, bigrams, trigrams, and n-grams in NLP?

Ans - When we parse a sentence one word at a time, then it is called a unigram. The sentence parsed two words at a time is a bigram. When the sentence is parsed three words at a time, then it is a trigram. Similarly, n-gram refers to the parsing of *n* words at a time.

58-What is precision and recall?

Ans -**Precision** is the ratio of true positive instances and the total number of positively predicted instances.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$= \frac{True\ Positive}{Total\ Predicted\ Positive}$$

Recall is the ratio of true positive instances and the total actual positive instances.

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$= \frac{True\ Positive}{Total\ Actual\ Positive}$$

59- what do you mean by tokenization.

Ans- tokenization means break into the words , Tokenization is a process used in NLP to split a sentence into tokens .sentence tokenization and word tokenization

60 – what is stemming ?

Ans - bringing the words to the base word , it is a text classification , some times having base word meaning may not be meaning full. It take less time and used whare human interaction not there.

61 – what do u mean by lemmatization.

Ans- when there is human interaction lemmatization are used, its always give base word meaningful.

Example – chat boat , Alexa, google assistant

62- what is vectorization?

Ans - it is used to convert the unstructured text data into numeric data, its similar to the encoding.

63- Bag of word or count vectorization or word frequency?

Ans – it count the repetition of word in a given sentences.

Disadvantage – it give same weightage for all words

64 – what is N- grams

Ans – combination of adjacent words of length.

65 - What is Parts-of-speech Tagging?

Ans - POS tagging is one of the most essential tools in Natural Language Processing. It helps in making the machine understand the meaning of a sentence.

When we are working on the sematic analysis pos tagging are used.

66- What is the Confusion Matrix?

Ans- **A confusion matrix is a summary of prediction results on a classification problem**. The number of correct and incorrect predictions are summarized with count values and broken down by each class.

67-. Which metric is used to split a node in Decision Tree.

Ans -information gain 2. Gini impurity 3. Chi square 4. Reduction in variance

68 - **Logistic Regression vs Linear Regression with a real-life example**

Ans – **pricing of new house linear regression and cancer patients or not .**

69. What is a random forest?

Random Forest is a machine learning method that performs classification and regression tasks, and consists of a large number of decision trees that work together to achieve dimensionality reduction, outlier values, and treat missing values.

70. What is a confusion matrix?

The Confusion Matrix calculates the performance of a classification model, where the rows represent instances in the predicted class, and columns represent instances in the actual class. It is called so because it helps to check if the system is confusing the two classes.

71. What are recommender systems?

Recommender systems are a part of information filtering systems that predict whether the user would prefer an item or not.

72. **What is collaborative filtering?**

Collaborative Filtering is a data filtering technique commonly used by recommender systems to identify useful patterns and information using several data sources, agents, and other collaborative sources to simplify the automatic prediction process for users.

73. What is cluster sampling?

Cluster Sampling is a sampling technique that is used by dividing the target area into clusters and is applied when the population is spread across a large area or when Simple Random Sampling is not possible to apply. In Cluster Sampling, each sampling unit is a cluster of individual elements.

74.Please explain Gradient Descent.

Gradient descent is an optimization algorithm that is used to find the coefficients of a function that minimizes the cost function. By trying different coefficient values, we can evaluate them and easily find the lowest cost.

75- Difference between CHAR and VARCHAR in MySQL

Ans-CHAR is fixed length while VARCHAR is variable length.

76- How does DISTINCT work in MySQL? How MySQL Optimises DISTINCT?

Ans- MySQL DISTINCT clause is usually used to eliminate the duplicate records from the table and fetch only the unique records. It is only used with the SELECT statement in MySQL.

77- How do you differentiate between Oracle and MySQL?

Ans- Oracle is a multi-model database with a single, integrated back-end, while MySQL is an open-source relational database management system (RDBMS).

78- How to test for NULL values in a database?

Ans- A null value is usually utilised in databases to signify a missing or unknown value. MySQL interprets the NULL value differently from other data types. To compare the fields with NULL values, one must use the “IS NULL” or “IS NOT NULL” operator.

79- What is a foreign key? How to implement the same in MySQL?

Ans- A foreign key is used to link two tables together. MySQL supports foreign keys, that permit cross-referencing related data across tables, and foreign key constraints, which help keep the related data consistent.

80-**What is SQL?**

Ans-*Structured Query Language*. It is a programming language specifically designed for working with databases.

81-What is a Database? What is a DBMS?

Ans-Database is data stored on a computer and organized in a way that makes it easy to access and manipulate.

The software tool that allows the user to interact with the data stored in the database is called a database management system – DBMS.

82-What is the difference between DDL, DML, DCL, and TCL?

Ans-*DDL* stands for **Data Definition Language** and includes commands which allow you to CREATE, DROP, ALTER, and TRUNCATE data structures.

DML, instead, involves commands for *manipulating* information. It means “**Data Manipulation Language**”, and regards the possibility to SELECT, INSERT, UPDATE, and DELETE data.

DCL, **Data Control Language**, consists of commands that are typically used by database administrators. This category allows the programmer to GRANT and REVOKE rights delineating how much *control* you can have over the information in the database.

Similarly, TCL, which is the **Transaction Control Language**, also contains commands applied by database administrators. They ensure the transactions occurring within the database will happen in such a way that minimalizes the danger of suffering from data loss.

83- What is a GAN used for?

Generative adversarial networks (GANs) are algorithmic architectures that use two neural networks, pitting one against the other (thus the “adversarial”) in order to generate new, synthetic instances of data that can pass for real data. They are used widely in **image generation, video generation and voice generation**.

84-