1. How Can You Choose a Classifier Based on a Training Set Data Size?

If the training set is small, high bias / low variance models (e.g. Naive Bayes) tend to perform better because they are less likely to overfit.

If the training set is large, low bias / high variance models (e.g. Logistic Regression) tend to perform better because they can reflect more complex relationships.

2. What Are Unsupervised Machine Learning Techniques?

There are two techniques used in unsupervised learning: clustering and association.

I) Clustering

Clustering problems involve data to be divided into subsets. These subsets, also called clusters, contain data that are similar to each other. Different clusters reveal different details about the objects, unlike classification or regression.

II) Association

In an association problem, we identify patterns of associations between different variables or items.

For example, an e-commerce website can suggest other items for you to buy, based on the prior purchases that you have made, spending habits, items in your wishlist, other customers' purchase habits, and so on.

3. Explain How a System Can Play a Game of Chess Using Reinforcement Learning.

Reinforcement learning has an environment and an agent. The agent performs some actions to achieve a specific goal. Every time the agent performs a task that is taking it towards the goal, it is rewarded. And, every time it takes a step that goes against that goal or in the reverse direction, it is penalized.

With reinforcement learning, the learning agent learns by playing the game. It will make a move (decision), check if it's the right move (feedback), and keep the outcomes in memory for the next step it takes (learning). There is a reward for every correct decision the system takes and punishment for the wrong one.

4. How is Amazon Able to Recommend Other Things to Buy?

Once a user buys something from Amazon, Amazon stores that purchase data for future reference and finds products that are most likely also to be bought, it is possible because of the Association algorithm, which can identify patterns in a given dataset.

5.What are techniques used for sampling?

There are various methods for drawing samples from data.

The two main Sampling techniques are

-Probability sampling

-Non-probability sampling

Probability sampling:

Probability sampling means that each individual of the population has a possibility of being included in the sample. Probability sampling methods include –

Simple random sampling:

In simple random sampling, each individual of the population has an equivalent chance of being selected or included.

Systematic sampling:

Systematic sampling is very much similar to random sampling. The difference is just that instead of randomly generating numbers, in systematic sampling every individual of the population is assigned a number and are chosen at regular intervals.

Stratified sampling:

In stratified sampling, the population is split into sub-populations. It allows you to conclude more precise results by ensuring that every sub-population is represented in the sample.

Cluster sampling:

Cluster sampling also involves dividing the population into sub-populations, but each subpopulation should have analogous characteristics to that of the whole sample. Rather than sampling individuals from each subpopulation, you randomly select the entire subpopulation.

-Non-probability sampling :

In non-probability sampling, individuals are selected using non-random ways and not every individual has a possibility of being included in the sample.

Convenience sampling:

Convenience sampling is a method where data is collected from an easily accessible group.

Voluntary Response sampling:

Voluntary Response sampling is similar to convenience sampling, but here instead of researchers choosing individuals and then contacting them, people or individuals volunteer themselves.

Purposive sampling:

Purposive sampling also known as judgmental sampling is where the researchers use their expertise to select a sample that is useful or relevant to the purpose of the research.

Snowball sampling:

Snowball sampling is used where the population is difficult to access. It can be used to recruit individuals via other individuals.

6.How to get the minimum, 25th percentile, median, 75th, and max of a numeric series?

```
"randomness= np.random.RandomState(100)

s = pd.Series(randomness.normal(100, 55, 5))

np.percentile(ser, q=[0, 25, 50, 75, 100])"
```

7. What is init in Python?

“init” is a reserved method in python classes. It is known as a constructor in object-oriented concepts. This method is called when an object is created from the class and it allows the class to initialise the attributes of the class.

8. Explain the Law of Large Numbers.

The ‘Law of Large Numbers’ states that if an experiment is repeated independently a large number of times, the average of the individual results is close to the expected value. It also states that the sample variance and standard deviation also converge towards the expected value.

9. What is denormalization?

Denormalization is the opposite of normalization; redundant data is added to speed up complex queries that have multiple tables that need to be joined. Optimization of the read performance of a database is attempted by adding or grouping redundant copies of data.

10. What are the applications of SQL?

The major applications of SQL include:

- Writing data integration scripts

- Setting and running analytical queries

- Retrieving subsets of information within a database for analytics applications and transaction processing

- Adding, updating, and deleting rows and columns of data in a database.

11. What is a DEFAULT constraint?

A default constraint is used to define a default value for a column so that it is added to all new records if no other value is specified. For example, if we assign a default constraint for the E_salary column in the following table and set the default value to 85000, then all the entries of this column will have the default value of 85000, unless no other value has been assigned during the insertion.

12. What is an index in SQL?

Indexes help speed up searching in a database. If there is no index on a column in the WHERE clause, then the SQL Server has to skim through the entire table and check each and every row to find matches, which may result in slow operations in large data

13. What is the ACID property in a database?

The full form of ACID is atomicity, consistency, isolation, and durability.

- Atomicity refers that if any aspect of a transaction fails, the whole transaction fails and the database state remains unchanged.

- Consistency means that the data meets all validity guidelines.

- Concurrency management is the primary objective of isolation.

- Durability ensures that once a transaction is committed, it will occur regardless of what happens in between such as a power outage, fire, or some other kind of disturbance.

14. What is the meaning of KPI in statistics?

KPI is an acronym for a key performance indicator. It can be defined as a quantifiable measure to understand whether the goal is being achieved or not. KPI is a reliable metric to measure the performance level of an organization or individual with respect to the objectives. An example of KPI in an organization is the expense ratio.

15. Explain One-hot encoding and Label Encoding. How do they affect the dimensionality of the given dataset?

One-hot encoding is the representation of categorical variables as binary vectors. Label Encoding is converting labels/words into numeric form. Using one-hot encoding increases the dimensionality of the data set. Label encoding doesn't affect the dimensionality of the data set. One-hot encoding creates a new variable for each level in the variable whereas, in Label encoding, the levels of a variable get encoded as 1 and 0.

16. How many report formats are available in Excel?

There are three report formats available in Excel; they are:

1. Compact Form

2. Outline Form

3. Tabular Form

17. What are sets in Tableau?

Sets are custom fields that define a subset of data based on some conditions. A set can be based on a computed condition, for example, a set may contain customers with sales over a certain threshold. Computed sets update as your data changes. Alternatively, a set can be based on specific data point in your view.

18. What is the difference between DROP and TRUNCATE commands?

DROP command removes a table and it cannot be rolled back from the database whereas TRUNCATE command removes all the rows from the table.

19. What is slicing in Python?

Slicing is used to access parts of sequences like lists, tuples, and strings. The syntax of slicing is- [start:end:step]. The step can be omitted as well. When we write [start:end] this returns all the elements of the sequence from the start (inclusive) till the end-1 element. If the start or end element is negative i, it means the ith element from the end.

20. What do you think the distribution of time spent per day on Facebook looks like? What metric would you use to describe the distribution.

In terms of the distribution of time spent per day on Facebook (FB), one can imagine there may be two groups of people on Facebook:

1. People who scroll quickly through their feed and don't spend too much time on FB.

2. People who spend a large amount of their social media time on FB.

Based on this, we make claim about the distribution of time spent on FB. The metrics to describe our distribution can be

1) Centre (mean, median, mode)

2) Spread (standard deviation, inter quartile range

3) Shape (skewness, kurtosis, uni or bimodal)

4) Outliers (Do they exist?)

We can give you a sample answer for your interview: –

If we assume that a person is visiting Facebook page, there is a probability(p) that after one unit of time(t) has passed that she will leave the page.

With a probability of p her visit will be limited to 1 unit of time. With a probability of (1−p)p her visit will be limited to 2 units of time. With a probability of (1−p)2p her visit will be limited to 3 units of time and so on. The probability mass function of this distribution is therefore (1−p)tp, and hence we can say this a geometric distribution.

21.How can we relate standard deviation and variance?

Standard deviation refers to the spread of your data from the mean. Variance is the average degree to which each point differs from the mean i.e. the average of all data points. We can relate Standard deviation and Variance because it is the square root of Variance.

22.Explain how a ROC curve works.

The ROC curve is a graphical representation of the contrast between true positive rates and the false positive rate at various thresholds. It's often used as a proxy for the trade-off between the sensitivity of the model (true positives) vs the fall-out or the probability it will trigger a false alarm (false positives).

23.Explain split() and join() functions in Python?

You can use split() function to split a string based on a delimiter to a list of strings.You can use join() function to join a list of strings based on a delimiter to give a single string.

24.What is pickling and unpickling?

Pickle module accepts any Python object and converts it into a string representation and dumps it into a file by using dump function, this process is called pickling. While the process of retrieving original Python objects from the stored string representation is called unpickling.

25. When does regularization come into play in Machine Learning?

At times when the model begins to underfit or overfit, regularization becomes necessary. It is a regression that diverts or regularizes the coefficient estimates towards zero. It reduces flexibility and discourages learning in a model to avoid the risk of overfitting. The model complexity is reduced and it becomes better at predicting.

26. What is the exploding gradient problem while using the back propagation technique?

When large error gradients accumulate and result in large changes in the neural network weights during training, it is called the exploding gradient problem. The values of weights can become so large as to overflow and result in NaN values. This makes the model unstable and the learning of the model to stall just like the vanishing gradient problem.

27. State some ways to improve the performance of Tableau?

Use an Extract to make workbooks run faster

Reduce the scope of data to decrease the volume of data

Reduce the number of marks on the view to avoid information overload

Try to use integers or Booleans in calculations as they are much faster than strings

Hide unused fields

Use Context filters

Reduce filter usage and use some alternative way to achieve the same result

Use indexing in tables and use the same fields for filtering

Remove unnecessary calculations and sheets.

28. What is Power Pivot & Power Query?

Power Pivot is an add-on provided by Microsoft for Excel since 2010. Power Pivot was designed to extend the analytical capabilities and services of Microsoft Excel.

Power Query is a business intelligence tool designed by Microsoft for Excel. Power Query allows you to import data from various data sources and will enable you to clean, transform and reshape your data as per the requirements. Power Query allows you to write your query once and then run it with a simple refresh.

29. What is macro in excel?

Macro refers to an algorithm or a set of actions that help automate a task in Excel by recording and playing back the steps taken to complete that task. Once the steps are stored, you create a Macro, and it can be edited and played back as many times as the user wants.

Macro is great for repetitive tasks and also eliminates errors. For example, suppose an account manager has to share reports regarding the company employees for non-payment of dues. In that case, it can be automated using a Macro and doing minor changes every month, as needed.

30.What are sets and groups in Tableau?

Sets and groups are used group data based on some specific conditions. The main difference between these two is that a group can divide the dataset into multiple groups whereas a set can have only two options which is either in or out. A user should choose to apply group or sets based on the requirements.

31. What are Support Vectors in SVM?

A Support Vector Machine (SVM) is an algorithm that tries to fit a line (or plane or hyperplane) between the different classes that maximizes the distance from the line to the points of the classes.

In this way, it tries to find a robust separation between the classes. The Support Vectors are the points of the edge of the dividing hyperplane.

32. What is Bias in Machine Learning?

Bias in data tells us there is inconsistency in data. The inconsistency may occur for several reasons which are not mutually exclusive.

For example, a tech giant like Amazon to speed the hiring process they build one engine where they are going to give 100 resumes, it will spit out the top five, and hire those.

When the company realized the software was not producing gender-neutral results it was tweaked to remove this bias.

33. Explain Correlation and Covariance?

Covariance signifies the direction of the linear relationship between two variables, whereas correlation indicates both the direction and strength of the linear relationship between variables.

34. What is SQL Injection?

SQL injection is a sort of flaw in website and web app code that allows attackers to take control of back-end processes and access, retrieve, and delete sensitive data stored in databases. In this approach, malicious SQL statements are entered into a database entry field, and the database becomes exposed to an attacker once they are executed. By utilising data-driven apps, this strategy is widely utilised to get access to sensitive data and execute administrative tasks on databases.

35. What is Perceptron? And how does it Work?

If we focus on the structure of a biological neuron, it has dendrites which are used to receive inputs. These inputs are summed in the cell body and using the Axon it is passed on to the next biological neuron. Similarly, a perceptron receives multiple inputs, applies various transformations and functions and provides an output. A Perceptron is a linear model used for binary classification. It models a neuron which has a set of inputs, each of which is given a specific weight. The neuron computes some function on these weighted inputs and gives the output.

36. What are the activation functions?

Activation function translates the inputs into outputs. Activation function decides whether a neuron should be activated or not by calculating the weighted sum and further adding bias with it. The purpose of the activation function is to introduce non-linearity into the output of a neuron.

There can be many Activation functions like:

Linear or Identity

Unit or Binary Step

Sigmoid or Logistic

Tanh

ReLU

Softmax.

37).What do you mean by Associative Rule Mining (ARM)?

Associative Rule Mining is one of the techniques to discover patterns in data like features (dimensions) which occur together and features (dimensions) which are correlated. It is mostly used in Market-based Analysis to find how frequently an itemset occurs in a transaction. Association rules have to satisfy minimum support and minimum confidence at the very same time. Association rule generation generally comprised of two different steps:

“A min support threshold is given to obtain all frequent item-sets in a database.”

“A min confidence constraint is given to these frequent item-sets in order to form the association rules.”

Support is a measure of how often the “item set” appears in the data set and Confidence is a measure of how often a particular rule has been found to be true.

38. What is x-velocity in Power Pivot?

X-Velocity is the in-memory analytics engine behind Power Pivot that loads and handles huge data in Power BI. It stores data in columnar storage that results in faster processing.

39. What is Gantt chart in Tableau ?

A Tableau Gantt chart illustrates the duration of events as well as the progression of value across the period. Along with the time axis, it has bars. The Gantt chart is primarily used as a project management tool, with each bar representing a project job.

40. What in Excel is a macro?

An Excel macro is an algorithm or a group of steps that helps automate an operation by capturing and replaying the steps needed to finish it. Once the steps have been saved, you may construct a Macro that the user can alter and replay as often as they like.

41. Power BI can connect to which data sources?

The data source is the point from which the data has been retrieved. It can be anything like files in various formats (.xlsx, .csv, .pbix, .xml, .txt etc), databases (SQL database, SQL Data Warehouse, Spark on Azure HDInsight), or form content packets like Google Analytics or Twilio.

42. What is DBSCAN Clustering?

DBSCAN groups ‘densely grouped’ data points into a single cluster. It can identify clusters in large spatial datasets by looking at the local density of the data points. The most exciting feature of DBSCAN clustering is that it is robust to outliers. It also does not require the number of clusters to be told beforehand, unlike K-Means, where we have to specify the number of centroids.

43. What are the different forms of joins in a table?

SQL has many kinds of different joins including INNER JOIN, SELF JOIN, CROSS JOIN, and OUTER JOIN. In fact, each join type defines the way two tables are related in a query. OUTER JOINS can further be divided into LEFT OUTER JOINS, RIGHT OUTER JOINS, and FULL OUTER JOINS.

44. How is the grid search parameter different from the random search?

Model Hyperparameter tuning is very useful to enhance the performance of a machine learning model. The only difference between both the approaches is in grid search we define the combinations and do training of the model whereas in RandomizedSearchCV the model selects the combinations randomly. Both are very effective ways of tuning the parameters that increase the model generalizability.

Random search is a technique where random combinations of the hyperparameters are used to find the best solution for the built model. The drawback of random search is that it yields high variance during computing. Since the selection of parameters is completely random; and since no intelligence is used to sample these combinations, luck plays its part.

45. Explain character-manipulation functions? Explains its different types in SQL.

Change, extract, and edit the character string using character manipulation routines. The function will do its action on the input strings and return the result when one or more characters and words are supplied into it.

The character manipulation functions in SQL are as follows:

A) CONCAT (joining two or more values): This function is used to join two or more values together. The second string is always appended to the end of the first string.

B) SUBSTR: This function returns a segment of a string from a given start point to a given endpoint.

C) LENGTH: This function returns the length of the string in numerical form, including blank spaces.

D) INSTR: This function calculates the precise numeric location of a character or word in a string.

E) LPAD: For right-justified values, it returns the padding of the left-side character value.

F) RPAD: For a left-justified value, it returns the padding of the right-side character value.

G) TRIM: This function removes all defined characters from the beginning, end, or both ends of a string. It also reduced the amount of wasted space.

H) REPLACE: This function replaces all instances of a word or a section of a string (substring) with the other string value specified.

46. How Do You Calculate the Daily Profit Measures Using LOD?

LOD expressions allow us to easily create bins on aggregated data such as profit per day.

Scenario: We want to measure our success by the total profit per business day.

Create a calculated field named LOD - Profit per day and enter the formula:

FIXED [Order Date] : SUM ([Profit])

Create another calculated field named LOD - Daily Profit KPI and enter the formula:

IF [LOD - Profit per day] > 2000 then “Highly Profitable.”

ELSEIF [LOD - Profit per day] <= 0 then “Unprofitable”

ELSE “Profitable”

END

To calculate daily profit measure using LOD, follow these steps to draw the visualization:

Bring YEAR(Order Date) and MONTH(Order Date) to the Columns shelf

Drag Order Id field to Rows shelf. Right-click on it, select Measure and click on Count(Distinct)

Drag LOD - Daily Profit KPI to the Rows shelf

Bring LOD - Daily Profit KPI to marks card and change mark type from automatic to area.

47. What are Super key and candidate key?

A super key may be a single or a combination of keys that help to identify a record in a table. Know that Super keys can have one or more attributes, even though all the attributes are not necessary to identify the records.

A candidate key is the subset of Super key, which can have one or more than one attributes to identify records in a table. Unlike Super key, all the attributes of the candidate key must be helpful to identify the records.

Note that all the candidate keys can be Super keys, but all the super keys cannot be candidate keys.

48.What is Database Cardinality?

Database Cardinality denotes the uniqueness of values in the tables. It supports optimizing query plans and hence improves query performance. There are three types of database cardinalities in SQL, as given below:

Higher Cardinality

Normal Cardinality

Lower Cardinality

49. Explain different character-manipulation functions in sql.

The character manipulation functions in SQL are as follows:

A) CONCAT (joining two or more values): This function is used to join two or more values together. The second string is always appended to the end of the first string.

B) SUBSTR: This function returns a segment of a string from a given start point to a given endpoint.

C) LENGTH: This function returns the length of the string in numerical form, including blank spaces.

D) INSTR: This function calculates the precise numeric location of a character or word in a string.

50. What is a UNIQUE constraint?

The UNIQUE Constraint prevents identical values in a column from appearing in two records. The UNIQUE constraint guarantees that every value in a column is unique.

51. What is BLOB and TEXT in MySQL?

BLOB stands for Binary Huge Objects and can be used to store binary data, whereas TEXT may be used to store a large number of strings. BLOB may be used to store binary data, which includes images, movies, audio, and applications.

TEXT values behave similarly to a character string or a non-binary string.

52. What do you mean by Denormalization?

Denormalization refers to a technique which is used to access data from higher to lower forms of a database. It helps the database managers to increase the performance of the entire infrastructure as it introduces redundancy into a table. It adds the redundant data into a table by incorporating database queries that combine data from various tables into a single table.

54. How can you select K for K-means Clustering?

There are two kinds of methods that include direct methods and statistical testing methods:

- Direct methods: It contains elbow and silhouette

- Statistical testing methods: It has gap statistics.

The silhouette is the most frequently used while determining the optimal value of k

56. What is Semi-supervised Machine Learning?

Supervised learning uses data that is completely labeled, whereas unsupervised learning uses no training data.

In the case of semi-supervised learning, the training data contains a small amount of labeled data and a large amount of unlabeled data

57.What is a stored procedure?

Stored Procedure is a function consists of many SQL statements to access the database system. Several SQL statements are consolidated into a stored procedure and execute them whenever and wherever required.

58.What is Power Query?

Power Query is a business intelligence tool designed by Microsoft for Excel. Power Query allows you to import data from various data sources and will enable you to clean, transform and reshape your data as per the requirements. Power Query allows you to write your query once and then run it with a simple refresh.

59.What is the Use of Dual-axis in Tableau?

Dual Axis allows you to compare measures, and this is useful when you want to compare two measures that have different scales.

60.What is an Alias in SQL?

An alias is a feature of SQL that is supported by most, if not all, RDBMSs. It is a temporary name assigned to the table or table column for the purpose of a particular SQL query. In addition, aliasing can be employed as an confusion technique to secure the real names of database fields. A table alias is also called a correlation name.

An alias is represented explicitly by the AS keyword but in some cases, the same can be performed without it as well.

61.Gantt chart in Tableau ?

A Tableau Gantt chart illustrates the duration of events as well as the progression of value across the period. Along with the time axis, it has bars. The Gantt chart is primarily used as a project management tool, with each bar representing a project job.

62. What is a kernel function in SVM?

In the SVM algorithm, a kernel function is a special mathematical function. In simple terms, a kernel function takes data as input and converts it into a required form. This transformation of the data is based on something called a kernel trick, which is what gives the kernel function its name. Using the kernel function, we can transform the data that is not linearly separable (cannot be separated using a straight line) into one that is linearly separable.

63. What do you understand by the F1 score?

The F1 score represents the measurement of a model's performance. It is referred to as a weighted average of the precision and recall of a model. The results tending to 1 are considered as the best, and those tending to 0 are the worst. It could be used in classification tests, where true negatives don't matter much.

64. What is the lambda function?

Lambda functions are an anonymous or nameless function.

These functions are called anonymous because they are not declared in the standard manner by using the def keyword. It doesn't require the return keyword as well. These are implicit in the function.

The function can have any number of parameters but can have just one statement and return just one value in the form of an expression. They cannot contain commands or multiple expressions.

An anonymous function cannot be a direct call to print because lambda requires an expression.

Lambda functions have their own local namespace and cannot access variables other than those in their parameter list and those in the global namespace.

Example:

```
x = lambda i,j: i+j
```

```
print(x(7,8))
```

```
Output: 15
```

65. What is a True positive rate and a false positive rate?

True positive rate or Recall: It gives us the percentage of the true positives captured by the model out of all the Actual Positive class.

TPR = TP/ (TP+FN)

False Positive rate: It gives us the percentage of all the false positives by my model prediction from the all Actual Negative class.

FPR = FP/(FP+TN)

66. Where is the data stored in Power BI?

Primarily, PowerBI uses two repositories to store its data: Azure Blob Storage and Azure SQL Database. Azure Blob Storage typically stores the data that is uploaded by the users. Azure SQL Database stores all the metadata and artifacts for the system itself.

67. What do you understand by query optimization?

The phase that identifies a plan for evaluation query which has the least estimated cost is known as query optimization.

The advantages of query optimization are as follows:

The output is provided fasterA larger number of queries can be executed in less timeReduces time and space complexity

68. Which questions should you ask the user/client before you create a dashboard?

Though this depends on the user's requirements, still some of the common questions that I would ask the client before creating a dashboard are :

What is the purpose of the dashboard?Should the dashboard be retrospective or real-time?How detailed the dashboard should be?How tech and data-savvy is the end-user?Does the data need to be segmented?Should I explain the dashboard design to you?

69. What are the common problems that data analysts encounter during analysis?

The common problems steps involved in any analytics project are:

Handling duplicate data

Collecting the meaningful right data at the right time

Handling data purging and storage problems

Making data secure and dealing with compliance issues

70. What is Clustering?

Clustering is the process of grouping a set of objects into a number of groups. Objects should be similar to one another within the same cluster and dissimilar to those in other clusters.

A few types of clustering are:

Hierarchical clustering

K means clustering

Density-based clustering

71. What is Reinforcement Learning?

Reinforcement learning is different from the other types of learning like supervised and unsupervised. In reinforcement learning, we are given neither data nor labels. Our learning is based on the rewards given to the agent by the environment.

72. Difference Between Sigmoid and Softmax functions?

Sigmoid is used for binary classification methods where we only have 2 classes, while SoftMax applies to multiclass problems. Sigmoid receives just one input and only outputs a single number that represents the probability of belonging to class1 (remember that we only have 2 classes so the probability of belonging to class2 = 1 - P(class1)). While on the other hand SoftMax is vectorized, meaning that takes a vector with the same number of entries as classes we have and outputs another vector where each component represents the probability of belonging to that class.

73. What is P-value?

P-values are used to make a decision about a hypothesis test. P-value is the minimum significant level at which you can reject the null hypothesis. The lower the p-value, the more likely you reject the null hypothesis.

74. What is the data analysis process?

Data analysis generally refers to the process of assembling, cleaning, interpreting, transforming, and modeling data to gain insights or conclusions and generate reports to help businesses become more profitable.

• Collect Data: The data is collected from a variety of sources and is then stored to be cleaned and prepared. This step involves removing all missing values and outliers.

• Analyse Data: As soon as the data is prepared, the next step is to analyze it. Improvements are made by running a model repeatedly. Following that, the model is validated to ensure that it is meeting the requirements.

• Create Reports: In the end, the model is implemented, and reports are generated as well as distributed to stakeholders.

75. Explain two different ways to detect outliers.

• Box Plot Method: According to this method, the value is considered an outlier if it exceeds or falls below 1.5*IQR (interquartile range), that is, if it lies above the top quartile (Q3) or below the bottom quartile (Q1).

• Standard Deviation Method: According to this method, an outlier is defined as a value that is greater or lower than the mean ± (3*standard deviation)

76. What is a Pivot table? Write its usage.

One of the basic tools for data analysis is the Pivot Table. With this feature, you can quickly summarize large datasets in Microsoft Excel. Using it, we can turn columns into rows and rows into columns. Furthermore, it permits grouping by any field (column) and applying advanced calculations to them. It is an extremely easy-to-use program since you just drag and drop rows/columns headers to build a report.

77. What is a DEFAULT constraint?

Constraints in SQL are used to specify some sort of rules for processing data and limiting the type of data that can go into a table.

A default constraint is used to define a default value for a column so that it is added to all new records if no other value is specified.

78. What do you mean by data integrity?

Data integrity is the assurance of accuracy and consistency of data over its whole life cycle. It is a critical aspect of the design, implementation, and usage of systems that store, process, or retrieve data.

Data integrity also defines integrity constraints for enforcing business rules on data when it is entered into a database or application.

79. What is AUTO_INCREMENT?

AUTO_INCREMENT is used in SQL to automatically generate a unique number whenever a new record is inserted into a table.Since the primary key is unique for each record, this primary field is added as the AUTO_INCREMENT field so that it is incremented when a new record is inserted.

80. What is the difference between DROP and TRUNCATE commands?

If a table is dropped, all things associated with that table are dropped as well. This includes the relationships defined on the table with other tables, access privileges, and grants that the table has, as well as the integrity checks and constraints.

However, if a table is truncated, there are no such problems as mentioned above. The table retains its original structure and the data is dropped.

81. How can we deal with problems that arise when the data flows in from a variety of sources?

There are many ways to go about dealing with multi-source problems. However, these are done primarily to solve the problems of:

Identifying the presence of similar/same records and merging them into a single recordRe-structuring the schema to ensure there is good schema integration

82. Where is Time Series Analysis used?

Since time series analysis (TSA) has a wide scope of usage, it can be used in multiple domains. Here are some of the places where TSA plays an important role:

Statistics

Signal processing

Econometrics

Weather forecasting

Earthquake prediction

Astronomy

Applied science

83. What are the ideal situations in which t-test or z-test can be used?

It is a standard practice that a t-test is used when there is a sample size less than 30 and the z-test is considered when the sample size exceeds 30 in most cases.

84. What is the usage of the NVL() function?

The NVL() function is used to convert the NULL value to the other value. The function returns the value of the second parameter if the first parameter is NULL. If the first parameter is anything other than NULL, it is left unchanged. This function is used in Oracle, not in SQL and MySQL. Instead of NVL() function, MySQL have IFNULL() and SQL Server have ISNULL() function.

85. How is Data modeling different from Database design?

Data Modeling: It can be considered as the first step towards the design of a database. Data modeling creates a conceptual model based on the relationship between various data models. The process involves moving from the conceptual stage to the logical model to the physical schema. It involves the systematic method of applying data modeling techniques.

Database Design: This is the process of designing the database. The database design creates an output which is a detailed data model of the database. Strictly speaking, database design includes the detailed logical model of a database but it can also include physical design choices and storage parameters.

86. What is the benefit of dimensionality reduction?

Dimensionality reduction reduces the dimensions and size of the entire dataset. It drops unnecessary features while retaining the overall information in the data intact. Reduction in dimensions leads to faster processing of the data.

The reason why data with high dimensions is considered so difficult to deal with is that it leads to high time consumption while processing the data and training a model on it. Reducing dimensions speeds up this process, removes noise, and also leads to better model accuracy.

87. Explain stacking in Data Science.

Just like bagging and boosting, stacking is also an ensemble learning method. In bagging and boosting, we could only combine weak models that used the same learning algorithms, e.g., logistic regression. These models are called homogeneous learners.

However, in stacking, we can combine weak models that use different learning algorithms as well. These learners are called heterogeneous learners. Stacking works by training multiple (and different) weak models or learners and then using them together by training another model, called a meta-model, to make predictions based on the multiple outputs of predictions returned by these multiple weak models.

88. What is the case when in SQL Server?

The CASE statement is used to construct logic in which one column's value is determined by the values of other columns. At least one set of WHEN and THEN commands makes up the SQL Server CASE Statement. The condition to be tested is specified by the WHEN statement. If the WHEN condition returns TRUE, the THEN sentence explains what to do.

When none of the WHEN conditions return true, the ELSE statement is executed. The END keyword brings the CASE statement to a close.

89. What is a relationship in SQL and what are they?

Database Relationship is defined as the connection between the tables in a database. There are various data base relationships, and they are as follows:.

One to One Relationship.

One to Many Relationship.

Many to One Relationship.

Self-Referencing Relationship.

90. What is the use of cycle fields in tableau?

Cycle fields help in switching and trying different colour combinations or views in a cyclic order. It will work only if we have a chart that allows more than one measure such as stacked bar chart and we are unable to finalize the visualizations then we can use cycle fields. To use cycle field, click on analysis menu in the toolbar then select cycle fields to take a quick look at an alternative visualization.

91. What is the difference between a function and a formula in Excel?

A formula is a user-defined expression that calculates a value. A function is pre-defined built-in operation that can take the specified number of arguments. A user can create formulas that can be complex and can have multiple functions in it. For example, =A1+A2 is a formula and =SUM(A1:A10) is a function.

92. Explain the difference between L1 and L2 regularization.

L2 regularization tends to spread error among all the terms, while L1 is more binary/sparse, with many variables either being assigned a 1 or 0 in weighting. L1 corresponds to setting a Laplacean prior on the terms, while L2 corresponds to a Gaussian prior.

93. What cross-validation technique would you use on a time series dataset?

Instead of using standard k-folds cross-validation, you have to pay attention to the fact that a time series is not randomly distributed data—it is inherently ordered by chronological order. If a pattern emerges in later time periods, for example, your model may still pick up on it even if that effect doesn't hold in earlier years!

You'll want to do something like forward chaining where you'll be able to model on past data then look at forward-facing data.

- Fold 1 : training [1], test [2]

- Fold 2 : training [1 2], test [3]

- Fold 3 : training [1 2 3], test [4]

- Fold 4 : training [1 2 3 4], test [5]

- Fold 5 : training [1 2 3 4 5], test [6]

94. What's the "kernel trick" and how is it useful?

The Kernel trick involves kernel functions that can enable in higher-dimension spaces without explicitly calculating the coordinates of points within that dimension: instead, kernel functions compute the inner products between the images of all pairs of data in a feature space. This allows them the very useful attribute of calculating the coordinates of higher dimensions while being computationally cheaper than the explicit calculation of said coordinates.

95. What are Query and Query language?

A query is nothing but a request sent to a database to retrieve data or information. The required data can be retrieved from a table or many tables in the database.

Query languages use various types of queries to retrieve data from databases. SQL, Datalog, and AQL are a few examples of query languages; however, SQL is known to be the widely used query language.

96. What do you mean by buffer pool and mention its benefits?

A buffer pool in SQL is also known as a buffer cache. All the resources can store their cached data pages in a buffer pool. The size of the buffer pool can be defined during the configuration of an instance of SQL Server.

The following are the benefits of a buffer pool:

Increase in I/O performance

Reduction in I/O latency

Increase in transaction throughput

Increase in reading performance

97. What is the difference between Zero and NULL values in SQL?

When a field in a column doesn't have any value, it is said to be having a NULL value. Simply put, NULL is the blank field in a table. It can be considered as an unassigned, unknown, or unavailable value. On the contrary, zero is a number, and it is an available, assigned, and known value.

98. What are Loss Function and Cost Functions? Explain the key Difference Between them?

When calculating loss we consider only a single data point, then we use the term loss function.

Whereas, when calculating the sum of error for multiple data then we use the cost function. There is no major difference.

In other words, the loss function is to capture the difference between the actual and predicted values for a single record whereas cost functions aggregate the difference for the entire training dataset.

The Most commonly used loss functions are Mean-squared error and Hinge loss.

Mean-Squared Error(MSE): In simple words, we can say how our model predicted values against the actual values.

MSE = summation(predicted value - actual value)**2/n(no of data points)

Hinge loss: It is used to train the machine learning classifier, which is

L(y) = max(0,1- y_true*y_pred)

Where y = -1 or 1 indicating two classes and y represents the output form of the classifier. The most common cost function represents the total cost as the sum of the fixed costs and the variable costs in the equation y = mx + b

101. Difference between a shallow and a deep copy?

It's faster to do shallow repetitions. It does, however, handle pointers and references in a "lazy" manner. It just copies over the pointer price rather than producing a current copy of the specific knowledge the pointer links to. As a result, each of the initial and subsequent copies can have pointers that relate to the same underlying knowledge. Deep repetition clones the underlying data completely. It is not shared by the first and, as a result, by the copy

102. What is the difference between primary key and unique key in SQL?

Both primary and unique keys carry unique values but a primary key cannot have a null value, while a unique key can. In a table, there cannot be more than one primary key, but there can be multiple unique keys.

103. What is a Stacked Column Chart in Tableau?

Stacked Column Chart, composed of multiple bars stacked vertically, one on another. The length of the bar depends on the value in the data point. A stacked column chart is the best one to know the changes in all variables. This type of chart should be checked when the number of series is higher than two.

104. Explain split() and join() functions in Python?

You can use split() function to split a string based on a delimiter to a list of strings.You can use join() function to join a list of strings based on a delimiter to give a single string.

105. Where is the data stored in Power BI?

Primarily, PowerBI uses two repositories to store its data: Azure Blob Storage and Azure SQL Database. Azure Blob Storage typically stores the data that is uploaded by the users. Azure SQL Database stores all the metadata and artifacts for the system itself.

106. What are the differences between OLTP and OLAP?

OLTP stands for online transaction processing, whereas OLAP stands for online analytical processing. OLTP is an online database modification system, whereas OLAP is an online database query response system.

107. What do you understand by query optimization?

The phase that identifies a plan for evaluation query which has the least estimated cost is known as query optimization.

The advantages of query optimization are as follows:

i) The output is provided faster.

ii) A larger number of queries can be executed in less time.

iii) Reduces time and space complexity

108. What is the difference between the RANK() and DENSE_RANK() functions?

The RANK() function in the result set defines the rank of each row within your ordered partition. If both rows have the same rank, the next number in the ranking will be the previous rank plus a number of duplicates. If we have three records at rank 4, for example, the next level indicated is 7.

The DENSE_RANK() function assigns a distinct rank to each row within a partition. This function will assign the same rank to the two rows if they have the same rank, with the next rank being the next consecutive number. If we have three records at rank 4, for example, the next level indicated is 5.

109. What is Normalization and what are the advantages of it?

Normalization in SQL is the process of organizing data to avoid duplication and redundancy. Some of the advantages are:

Better Database organization

More Tables with smaller rows

Efficient data access

Greater Flexibility for Queries

Quickly find the information

Easier to implement Security

Allows easy modification

Reduction of redundant and duplicate data

110. What is SRS and what are its key elements?

A System Requirements Specification (SRS) or a Software Requirements Specification is a document or set of documents that describe the features of a system or software application. The key elements of an SRS are:

Scope of Work

Functional Requirements

Non-Functional Requirements

Dependencies

Data Model

Assumptions

Constraints

Acceptance Criteria

111. What do you know about Kanban?

Kanban is a tool which helps the agile team to visually guide and manage the work as it progresses through the process. Besides, it works as a scheduling system in Agile just-in-time production. The Kanban board is used to describe the current development status.

112. What is clustered index in SQL?

A clustered index is actually a table where the data for the rows are stored. It determines the order of the table data based on the key values that can sort in only one direction. Each table can have only one clustered index. It is the only index, which has been automatically created when the primary key is generated. If many data modifications needed to be done in the table, then clustered indexes are preferred.

113. What are sets in Tableau?

Sets are custom fields that define a subset of data based on some conditions. A set can be based on a computed condition, for example, a set may contain customers with sales over a certain threshold. Computed sets update as your data changes. Alternatively, a set can be based on specific data point in your view.

114. Which of the following machine learning algorithms can be used for inputting missing values of both categorical and continuous variables?

K-means clustering

Linear regression

K-NN (k-nearest neighbor)

Decision trees

The K nearest neighbor algorithm can be used because it can compute the nearest neighbor and if it doesn't have a value, it just computes the nearest neighbor based on all the other features.

When you're dealing with K-means clustering or linear regression, you need to do that in your pre-processing, otherwise, they'll crash. Decision trees also have the same problem, although there is some variance

115. What is an RNN (recurrent neural network)?

RNN is an algorithm that uses sequential data. RNN is used in language translation, voice recognition, image capturing etc. There are different types of RNN networks such as one-to-one, one-to-many, many-to-one and many-to-many. RNN is used in Google's Voice search and Apple's Siri.

116. What is the goal of A/B Testing?

This is statistical hypothesis testing for randomized experiments with two variables, A and B. The objective of A/B testing is to detect any changes to a web page to maximize or increase the outcome of a strategy.

117. What is a star schema?

Star schema is the fundamental schema among the data mart schema and it is simplest. It is said to be star as its physical model resembles to the star shape having a fact table at its center and the dimension tables at its peripheral representing the star's points.

118. Explain data cleansing.

Data cleaning, also known as data cleansing or data scrubbing or wrangling, is basically a process of identifying and then modifying, replacing, or deleting the incorrect, incomplete, inaccurate, irrelevant, or missing portions of the data as the need arises. This fundamental element of data science ensures data is correct, consistent, and usable.

119. What is an Affinity Diagram?

An Affinity Diagram is an analytical tool used to cluster or organize data into subgroups based on their relationships. These data or ideas are mostly generated from discussions or brainstorming sessions and are used in analyzing complex issues.

120. Which questions should you ask the user/client before you create a dashboard?

Though this depends on the user's requirements, still some of the common questions that I would rocask the client before creating a dashboard are :

What is the purpose of the dashboard?Should the dashboard be retrospective or real-time?How detailed the dashboard should be?How tech and data-savvy is the end-user?Does the data need to be segmented?Should I explain the dashboard design to you?

121. What's the difference between a generative and discriminative model?

A generative model will learn categories of data while a discriminative model will simply learn the distinction between different categories of data. Discriminative models will generally outperform generative models on classification tasks.

122. When should you use classification over regression?

Classification produces discrete values and dataset to strict categories, while regression gives you continuous results that allow you to better distinguish differences between individual points. You would use classification over regression if you wanted your results to reflect the belongingness of data points in your dataset to certain explicit categories (ex: If you wanted to know whether a name was male or female rather than just how correlated they were with male and female names.)

123. How do you think Google is training data for self-driving cars?

Google is currently using recaptcha to source labeled data on storefronts and traffic signs. They are also building on training data collected by Sebastian Thrun at GoogleX—some of which was obtained by his grad students driving buggies on desert dunes!

124.In Microsoft Excel, how do you create a drop-down list?

Start by selecting the Data tab from the ribbon.

Select Data Validation from the Data Tools group.

Go to Settings > Allow > List next.

Choose the source you want to offer in the form of a list array.

125. What is batch normalization?

Batch normalization is a technique through which attempts could be made to improve the performance and stability of the neural network. This can be done by normalizing the inputs in each layer so that the mean output activation remains 0 with the standard deviation at 1.

126. What is GAN?

The Generative Adversarial Network takes inputs from the noise vector and sends them forward to the Generator, and then to Discriminator, to identify and differentiate unique and fake inputs.

127. What is the case when in SQL Server?

The CASE statement is used to construct logic in which one column's value is determined by the values of other columns.At least one set of WHEN and THEN commands makes up the SQL Server CASE Statement. The condition to be tested is specified by the WHEN statement. If the WHEN condition returns TRUE, the THEN sentence explains what to do.

When none of the WHEN conditions return true, the ELSE statement is executed. The END keyword brings the CASE statement to a close.

129. How is KNN different from k-means?

KNN or K nearest neighbors is a supervised algorithm which is used for classification and regression purpose. In KNN, a test sample is given as the class of the majority of its nearest neighbors. On the other side, K-means is an unsupervised algorithm which is mainly used for clustering. In k-means clustering, it needs a set of unlabeled points and a threshold only. The algorithm further takes unlabeled data and learns how to cluster it into groups by computing the mean of the distance between different unlabeled points.

130. What do you understand by ILP?

ILP stands for Inductive Logic Programming. It is a part of machine learning which uses logic programming. It aims at searching patterns in data which can be used to build predictive models. In this process, the logic programs are assumed as a hypothesis.

131. What do you understand by Cluster Sampling?

Cluster Sampling is a process of randomly selecting intact groups within a defined population, sharing similar characteristics. Cluster sample is a probability where each sampling unit is a collection or cluster of elements.

For example, if we are clustering the total number of managers in a set of companies, in that case, managers (sample) will represent elements and companies will represent clusters.

132. What is Regularization? What kind of problems does regularization solve?

A regularization is a form of regression, which constrains/ regularizes or shrinks the coefficient estimates towards zero. In other words, it discourages learning a more complex or flexible model to avoid the risk of overfitting. It reduces the variance of the model, without a substantial increase in its bias.

Regularization is used to address overfitting problems as it penalizes the loss function by adding a multiple of an L1 (LASSO) or an L2 (Ridge) norm of weights vector w

133. What are the common problems that data analysts encounter during analysis?

The common problems steps involved in any analytics project are:

Handling duplicate data

Collecting the meaningful right data at the right time

Handling data purging and storage problems

Making data secure and dealing with compliance issues

134. Explain the Type I and Type II errors in Statistics?

In Hypothesis testing, a Type I error occurs when the null hypothesis is rejected even if it is true. It is also known as a false positive.

A Type II error occurs when the null hypothesis is not rejected, even if it is false. It is also known as a false negative.

135. How do you subset or filter data in SQL?

To subset or filter data in SQL, we use WHERE and HAVING clauses which give us an option of including only the data matching certain conditions.

136. What do you understand by a random forest model?

It combines multiple models together to get the final output or, to be more precise, it combines multiple decision trees together to get the final output. So, decision trees are the building blocks of the random forest model.

137. How are Data Science and Machine Learning related to each other?

Data Science and Machine Learning are two terms that are closely related but are often misunderstood. Both of them deal with data. Data Science is a broad field that deals with large volumes of data and allows us to draw insights out of this voluminous data. Machine Learning, on the other hand, can be thought of as a sub-field of Data Science. It also deals with data, but here, we are solely focused on learning how to convert the processed data into a functional model, which can be used to map inputs to outputs, e.g., a model that can expect an image as an input and tell us if that image contains a flower as an output.

138. What is a kernel function in SVM?

In the SVM algorithm, a kernel function is a special mathematical function. In simple terms, a kernel function takes data as input and converts it into a required form. This transformation of the data is based on something called a kernel trick, which is what gives the kernel function its name. Using the kernel function, we can transform the data that is not linearly separable (cannot be separated using a straight line) into one that is linearly separable.

139. Explain TF/IDF vectorization.

The expression ‘TF/IDF’ stands for Term Frequency–Inverse Document Frequency. It is a numerical measure that allows us to determine how important a word is to a document in a collection of documents called a corpus. TF/IDF is used often in text mining and information retrieval.

140. When does regularization become necessary in Machine Learning?

Regularization is necessary whenever the model begins to overfit/ underfit. It is a cost term for bringing in more features with the objective function. Hence, it tries to push the coefficients for many variables to zero and reduce cost term. It helps to reduce model complexity so that the model can become better at predicting (generalizing).

142. What according to you, is more important between model accuracy and model performance?

Model accuracy is only a subset of model performance, and sometimes a misleading one. Let’s say you wanted to detect fraud in a massive dataset with a sample of millions, a more accurate model would most likely predict no fraud at all if only a minority of cases were fraud. However, this would be useless for a predictive model—a model designed to find fraud that asserted there was no fraud at all! Questions like this help you demonstrate that model accuracy isn’t always great way to assess performance of the model. So Model performance is more important.

143. What are the necessary steps involved in Machine Learning Project?

There are several essential steps we must follow to achieve a good working model while doing a Machine Learning Project. Those steps may include parameter tuning, data preparation, data collection, training the model, model evaluation, and prediction, etc.

144. What are different types of Collation Sensitivity?

Following are the different types of Collation Sensitivity:

- Case sensitive: A and a, B and b

- Kana sensitive: Japanese Kana characters

- Width sensitive: single byte characters and double-byte characters.

- Accent Sensitive.

145. What is OLTP ?

OLTP or Online Transaction Processing is a type of data processing that consists of executing a number of transactions occurring concurrently—online banking, shopping, order entry, or sending text messages, for example. These transactions traditionally are referred to as economic or financial transactions, recorded and secured so that an enterprise can access the information anytime for accounting or reporting purposes.

146. What is OLAP?

OLAP stands for On-Line Analytical Processing. OLAP is a classification of software technology which authorizes analysts, managers, and executives to gain insight into information through fast, consistent, interactive access in a wide variety of possible views of data that has been transformed from raw information to reflect the real dimensionality of the enterprise as understood by the clients.

OLAP implement the multidimensional analysis of business information and support the capability for complex estimations, trend analysis, and sophisticated data modeling.

147. How OLAP Works?

Fundamentally, OLAP has a very simple concept. It pre-calculates most of the queries that are typically very hard to execute over tabular databases, namely aggregation, joining, and grouping. These queries are calculated during a process that is usually called 'building' or 'processing' of the OLAP cube. This process happens overnight, and by the time end users get to work - data will have been updated.

148. What are exploding gradients?

Exploding Gradients is the problematic scenario where large error gradients accumulate to result in very large updates to the weights of neural network models in the training stage. In an extreme case, the value of weights can overflow and result in NaN values. Hence the model becomes unstable and is unable to learn from the training data.

149. What is systematic sampling and cluster sampling ?

Systematic sampling is a type of probability sampling method. The sample members are selected from a larger population with a random starting point but a fixed periodic interval. This interval is known as the sampling interval. The sampling interval is calculated by dividing the population size by the desired sample size.

Cluster sampling involves dividing the sample population into separate groups, called clusters. Then, a simple random sample of clusters is selected from the population. Analysis is conducted on data from the sampled clusters.

150. What does KPI stand for in statistics?

A KPI is a quantifiable measure to evaluate whether the objectives are being met or not.

It is a reliable metric to measure the performance level of an organisation or individual.

An example of a KPI in an organisation such as the expense ratio.

In terms of performance, KPIs are an effective way of measuring whether an organisation or individual is meeting expectations.

151. What is Perceptron? And how does it Work?

If we focus on the structure of a biological neuron, it has dendrites which are used to receive inputs. These inputs are summed in the cell body and using the Axon it is passed on to the next biological neuron. Similarly, a perceptron receives multiple inputs, applies various transformations and functions and provides an output. A Perceptron is a linear model used for binary classification. It models a neuron which has a set of inputs, each of which is given a specific weight. The neuron computes some function on these weighted inputs and gives the output.

153. What's the difference between a feed-forward and a backpropagation neural network?

A Feed-Forward Neural Network is a type of Neural Network architecture where the connections are "fed forward", i.e. do not form cycles. The term "Feed-Forward" is also used when you input something at the input layer and it travels from input to hidden and from hidden to the output layer.

Backpropagation is a training algorithm consisting of 2 steps:

Feed-Forward the values.

Calculate the error and propagate it back to the earlier layers.

154. What is Dropout?

Dropout is a regularization technique to avoid overfitting thus increasing the generalizing power. Generally, we should use a small dropout value of 20%-50% of neurons with 20% providing a good starting point.

155. Is indentation required in python?

Indentation is necessary for Python. It specifies a block of code. All code within loops, classes, functions, etc is specified within an indented block. It is usually done using four space characters. If your code is not indented necessarily, it will not execute accurately and will throw errors as well.

156. What are Entities and Relationships?

Entity: An entity can be a real-world object that can be easily identifiable. For example, in a college database, students, professors, workers, departments, and projects can be referred to as entities.

Relationships: Relations or links between entities that have something to do with each other. For example – The employee's table in a company's database can be associated with the salary table in the same database.

157. What are Aggregate and Scalar functions?

An aggregate function performs operations on a collection of values to return a single scalar value. Aggregate functions are often used with the GROUP BY and HAVING clauses of the SELECT statement. A scalar function returns a single value based on the input value.

158. What are Custom Visuals in Power BI?

Custom Visuals are like any other visualizations, generated using Power BI. The only difference is that it develops the custom visuals using a custom SDK. The languages like JQuery and JavaScript are used to create custom visuals in Power BI

159. What is Auto Increment?

Auto increment keyword allows the user to create a unique number to be generated when a new record is inserted into the table. AUTO INCREMENT keyword can be used in Oracle and IDENTITY keyword can be used in SQL SERVER.

Mostly this keyword can be used whenever PRIMARY KEY is used.

160. Which operator is used in query for pattern matching?

LIKE operator is used for pattern matching, and it can be used as -.

1. % – Matches zero or more characters.

2. _(Underscore) – Matching exactly one character.

161. How to use Power BI in excel?

To use Power BI in Excel, there is an Analyse in Excel option for every report in the Power BI service. To use it, you will need to enable editing and enable content for the report for the first time. So, what this option provides us is that it gives us the underlying data set of our Power BI report. It comes as a data connection in excel. And we get to play with the data in excel. It is up to us how we analyze the same data, either through pivot tables, charts, etc. By default, when the data is extracted in excel for any report- it gives a Pivot table by default.

162.What is the difference between Deep Learning and Machine Learning?

Deep Learning allows machines to make various business-related decisions using artificial neural networks that simulate the human brain, which is one of the reasons why it needs a vast amount of data for training. Machine Learning gives machines the ability to make business decisions without any external help, using the knowledge gained from past data. Machine Learning systems require relatively small amounts of data to train themselves, and most of the features need to be manually coded and understood in advance.

163.What is Cross-validation in Machine Learning?

Cross-validation allows a system to increase the performance of the given Machine Learning algorithm. This sampling process is done to break the dataset into smaller parts that have the same number of rows, out of which a random part is selected as a test set and the rest of the parts are kept as train sets. Cross-validation consists of the following techniques:

•Holdout method

•K-fold cross-validation

•Stratified k-fold cross-validation

•Leave p-out cross-validation

164.What is Epoch in Machine Learning?

Epoch in Machine Learning is used to indicate the count of passes in a given training dataset where the Machine Learning algorithm has done its job. Generally, when there is a large chunk of data, it is grouped into several batches. All these batches go through the given model, and this process is referred to as iteration. Now, if the batch size comprises the complete training dataset, then the count of iterations is the same as that of epochs.

165. What is Dimensionality Reduction?

In the real world, Machine Learning models are built on top of features and parameters. These features can be multidimensional and large in number. Sometimes, the features may be irrelevant and it becomes a difficult task to visualize them. This is where dimensionality reduction is used to cut down irrelevant and redundant features with the help of principal variables. These principal variables conserve the features, and are a subgroup, of the parent variables.

166. What are the roles & responsibilities of a Power BI developer?

The specific responsibilities that a Power BI Developer performs vary widely based on the industry and the organization for which they work. A Power BI developer should expect to encounter some or all of the roles and responsibilities listed below:

Build Analysis Services reporting models.

Using Power BI desktop, build visual reports, dashboards, and KPI scorecards.

Data analysis skills

In Power BI, use row-level data security and learn about application security layer models.

On the Power BI desktop, run DAX queries.

Use the data set to do advanced level calculations.

Develop custom visuals for Power BI.

Integrate Power BI reports into other applications

Should be well-versed with secondary tools like Microsoft Azure, SQL data warehouse, PolyBase, Visual Studio, and others.

167. What is Power BI Desktop?

Power BI Desktop is an open-source application designed and developed by Microsoft. Power BI Desktop will allow users to connect to, transform, and visualize your data with ease. Power BI Desktop lets users build visuals and collections of visuals that can be shared as reports with your colleagues or your clients in your organization.

168. What is Power Pivot?

Power Pivot is an add-on provided by Microsoft for Excel since 2010. Power Pivot was designed to extend the analytical capabilities and services of Microsoft Excel.

169. What is Power Query?

Power Query is a business intelligence tool designed by Microsoft for Excel. Power Query allows you to import data from various data sources and will enable you to clean, transform and reshape your data as per the requirements. Power Query allows you to write your query once and then run it with a simple refresh.

170. Explain dashboard lifecycle?

Dashboard lifecycle in Tableau:

Functional Knowledge: Business Analysts give a current functional knowledge of the organization.

Requirement Analysis: Requirements that are kept in consideration are:

The requirement of the dashboard.

How is data flowing in the current system?

Blueprint or layout of the system.

Dashboard scope.

The value that is added to the business

required tools for the development of the project and its costs.

Planning Phase: It includes:

Timeline and needed resources.

Work and leave plan.

Dependencies and future challenges.

Methodologies to follow: Scrum, Agile, Waterfall, etc.

Technical Specs: It includes:

Technical details.

SQL, relations, and Joins.

Credentials for database access.

Business logic.

Development: It includes:

Query generation.

Connecting databases and creating dimension model

Publish it to the server.

Unit testing.

Q&A Testing: It includes:

Functionality and UI testing.

SQL testing and data validation

Security testing

Testing of applied customization.

Performance testing: Report opening time, with or without any webpage.

User Acceptance Testing (UAT): User validates data and functionality.

Production and Support: System is produced, and support is given once it goes live.

172. What is VIZQL in Tableau?

VIZQL is Visual Inquiry Language. It is a combination of VIZ and SQL. It is similar to SQL language. But instead of SQL commands, the VIZQL language converts data queries into visual images.

173. What is the difference between .twb and .twbx extensions?

.twb: .twb means Tableau workbook. .twb is an XML sheet, it stores the data about your documents, stories, and dashboards. This file is the reference to the source file such as Excel or tde. This file will be linked to your source file when you save the TWB file. If you want to share your workbook you need to send both the workbook and data source file.

.twbx: It is a compressed file, where you have all files. It includes data source files, twb, and other files to produce the workbook. TWBX is obsolete for sharing because it will share the copy of the file instead of an original source file. .twbx is used for reports and we can view using the tableau viewer.

174. How to use excel formulas?

Formulas in Excel are expressions that can make the calculation based on the information in your spreadsheet. The formula can be used in Excel by first selecting an empty cell where you want to enter the formula. The way to add formula syntax is by typing the "=" equal sign and then followed by operators to be used in the calculation.

The relevant operators used in the formula should be kept between opening and closing parenthesis. Press Enter to get the results.

175. How to maintain accounts in excel sheet format?

You can maintain accounts in an Excel sheet by using its rows and columns for record-keeping and keep track of account-related information such as account numbers, invoicing, and receipt of payments.

The cells can be used to input client or account holder information. The expenses can be calculated using the formula =SUM (F3:F6). You can also use tools such as expense tracking tools and contract tools for auto-invoicing, performing tracking, and billing of expenses.

176. How to create a serial number generator in excel?

You can create serial numbers in Excel using the Autofill method. Here are the steps –

Insert value (let's say 1) in a row and select it as the active cell.

You will notice that a small box or filler icon at the bottom right of the active cell will appear.

Double click on that box icon to drag to the desired cell.

Now go back to the second row and insert another value (let's say 2).

Now select both the cells at once, and the box or filler icon will appear again.Drag down the icon to the desired row, and you will notice that this time it shows serial numbers starting with the first value we inserted (1).

177. What is p-value in hypothesis testing?

If the p-value is more than then critical value, then we fail to reject the H0

If p-value = 0.015 (critical value = 0.05) – strong evidence

If p-value = 0.055 (critical value = 0.05) – weak evidence

If the p-value is less than the critical value, then we reject the H0

If p-value = 0.055 (critical value = 0.05) – weak evidence

If p-value = 0.005 (critical value = 0.05) – strong evidence

178. What is the difference between one tail and two tail hypothesis testing?

2-tail test: Critical region is on both sides of the distribution

H0: x = µ

H1: x <> µ

1-tail test: Critical region is on one side of the distribution

H1: x <= µ

H1: x > µ

179. What is the Six sigma in statistic?

In quality control, an error-free data set is generated using six sigma statistics. σ is known as standard deviation. The lower the standard deviation, the less likely that a process performs accurately and commits errors. If a process delivers 99.99966% error-free results, it is said to be six sigma. A six sigma model is one that outperforms 1σ, 2σ, 3σ, 4σ, and 5σ processes and is sufficiently reliable to deliver defect-free work.

181. Why Is Re-sampling Done?

Resampling is done to:

- Estimate the accuracy of sample statistics with the subsets of accessible data at hand.

- Substitute data point labels while performing significance tests.

- Validate models by using random subsets.

182. What are Autoencoders?

An autoencoder is a kind of artificial neural network. It is used to learn efficient data codings in an unsupervised manner. It is utilised for learning a representation (encoding) for a set of data, mostly for dimensionality reduction, by training the network to ignore signal “noise”. Autoencoder also tries to generate a representation as close as possible to its original input from the reduced encoding.

183. What are the steps to build a Random Forest Model?

A Random Forest is essentially a build up of a number of decision trees. The steps to build a random forest model include:

Step1: Select ‘k’ features from a total of ‘m’ features, randomly. Here k << m

Step2: Calculate node D using the best split point — along the ‘k’ features

Step3: Split the node into daughter nodes using best split

Step 4: Repeat Steps 2 and 3 until the leaf nodes are finalised

Step5: Build a Random forest by repeating steps 1-4 for ‘n’ times to create ‘n’ number of trees..

184. What is the difference between SQL and MySQL?

SQL is a standard language for retrieving and manipulating structured databases. On the contrary, MySQL is a relational database management system, like SQL Server, Oracle or IBM DB2, that is used to manage SQL databases.

185. What is Data Integrity?

Data Integrity is the assurance of accuracy and consistency of data over its entire life-cycle and is a critical aspect of the design, implementation, and usage of any system which stores, processes, or retrieves data.

186. What is the Use of Dual-axis in Tableau?

Dual Axis allows you to compare measures, and this is useful when you want to compare two measures that have different scales.

187. How Can You Embed a Webpage in a Dashboard?

Follow these simple steps to embed a webpage in a dashboard:

1. Go to dashboard

2. Double click the ‘Webpage’ option available under ‘Objects.’

3. Enter the URL (here https://www.cloudyml.com) of the webpage in the dialog box that appears

You can see the webpage appears on the dashboard.

189.Explain the Law of Large Numbers.

The ‘Law of Large Numbers’ states that if an experiment is repeated independently a large number of times, the average of the individual results is close to the expected value.

There are two forms of the law of large numbers, but the differences are primarily theoretical.The weak law of large numbers states that as n increases, the sample statistic of the sequence converges in probability to the population value.

The strong law of large numbers describes how a sample statistic converges on the population value as the sample size or the number of trials increases. For example, the sample mean will converge on the population mean as the sample size increases.

190. What is the importance of A/B testing.

The goal of A/B testing is to pick the best variant among two hypotheses, the use cases of this kind of testing could be a web page or application responsiveness, landing page redesign, banner testing, marketing campaign performance etc.

The first step is to confirm a conversion goal, and then statistical analysis is used to understand which alternative performs better for the given conversion goal.

191. Explain Eigenvectors and Eigenvalues.

Eigenvectors depict the direction in which a linear transformation moves and acts by compressing, flipping, or stretching. They are used to understand linear transformations and are generally calculated for a correlation or covariance matrix.

The eigenvalue is the strength of the transformation in the direction of the eigenvector.

An eigenvector's direction remains unchanged when a linear transformation is applied to it.

192. What is x-velocity in Power Pivot?

X-Velocity is the in-memory analytics engine behind Power Pivot that loads and handles huge data in Power BI. It stores data in columnar storage that results in faster processing.

193. Describe Joins in SQL.

SQL Join statement is used to combine data or rows from two or more tables based on a common field between them. Different types of Joins are as follows:

- INNER JOIN- returns rows when there is a match in both tables.

- LEFT JOIN- returns all rows from the left table and matching rows in the right table.

- RIGHT JOIN- returns all rows from the right table and matching rows in the left table.

- OUTER JOIN- returns all records from both tables that satisfy the join condition

194. What is a Calculated Field in Tableau?

A calculated field is used to create new (modified) fields from existing data in the data source. It can be used to create more robust visualizations and doesn't affect the original dataset.

195. What is a Parameter in Tableau? Give an Example.

A parameter is a dynamic value that a customer could select, and you can use it to replace constant values in calculations, filters, and reference lines. For example, when creating a filter to show the top 10 products based on total profit instead of the fixed value, you can update the filter to show the top 10, 20, or 30 products using a parameter.

196. How do you create a hyperlink in Excel?

Hyperlinks are used to navigate between worksheets and files/websites. To create a hyperlink, the shortcut used is Ctrl+K.

The 'Insert Hyperlink' box appears. Enter the address and the text to display.

197. What are the types of SQL Queries?

We have four types of SQL Queries:

DDL (Data Definition Language): the creation of objects

DML (Data Manipulation Language): manipulation of data

DCL (Data Control Language): assignment and removal of permissions

TCL (Transaction Control Language): saving and restoring changes to a database

198. What is Supervised Learning?

Supervised learning is a machine learning algorithm of inferring a function from labeled training data. The training data consists of a set of training examples.

Example: 01

Knowing the height and weight identifying the gender of the person. Below are the popular supervised learning algorithms.

Support Vector Machines

Regression

Naive Bayes

Decision Trees

K-nearest Neighbour Algorithm and Neural Networks.

Example: 02

If you build a T-shirt classifier, the labels will be “this is an S, this is an M and this is L”, based on showing the classifier examples of S, M, and L.

199. What is Unsupervised Learning?

Unsupervised learning is also a type of machine learning algorithm used to find patterns on the set of data given. In this, we don’t have any dependent variable or label to predict. Unsupervised Learning Algorithms:

Clustering,

Anomaly Detection,

Neural Networks and Latent Variable Models.

Example:

In the same example, a T-shirt clustering will categorize as “collar style and V neck style”, “crew neck style” and “sleeve types”.

201. What is Ensemble learning?

Ensemble learning is a method that combines multiple machine learning models to create more powerful models.

There are many reasons for a model to be different. Few reasons are:

Different Population

Different Hypothesis

Different modeling techniques

When working with the model’s training and testing data, we will experience an error. This error might be bias, variance, and irreducible error.

Now the model should always have a balance between bias and variance, which we call a bias-variance trade-off.

This ensemble learning is a way to perform this trade-off.

There are many ensemble techniques available but when aggregating multiple models there are two general methods:

Bagging, a native method: take the training set and generate new training sets off of it.

Boosting, a more elegant method: similar to bagging, boosting is used to optimize the best weighting scheme for a training set.

202. Can you name the wildcards in Excel?

There are 3 wildcards in Excel that can ve used in formulas.

Asterisk () – 0 or more characters. For example, Ex could mean Excel, Extra, Expertise, etc.

Question mark (?) – Represents any 1 character. For example, R?ain may mean Rain or Ruin.

Tilde () – Used to identify a wildcard character (, , ?). For example, If you need to find the exact phrase India in a list. If you use India* as the search string, you may get any word with India at the beginning followed by different characters (such as Indian, Indiana). If you have to look for India” exclusively, use ~.

Hence, the search string will be india~*. ~ is used to ensure that the spreadsheet reads the following character as is, and not as a wildcard.

203.What is cascading filter in tableau?

Cascading filters can also be understood as giving preference to a particular filter and then applying other filters on previously filtered data source. Right-click on the filter you want to use as a main filter and make sure it is set as all values in dashboard then select the subsequent filter and select only relevant values to cascade the filters. This will improve the performance of the dashboard as you have decreased the time wasted in running all the filters over complete data source.

204.What are the various Power BI versions?

Power BI Premium capacity-based license, for example, allows users with a free license to act on content in workspaces with Premium capacity. A user with a free license can only use the Power BI service to connect to data and produce reports and dashboards in My Workspace outside of Premium capacity. They are unable to exchange material or publish it in other workspaces. To process material, a Power BI license with a free or Pro per-user license only uses a shared and restricted capacity. Users with a Power BI Pro license can only work with other Power BI Pro users if the material is stored in that shared capacity. They may consume user-generated information, post material to app workspaces, share dashboards, and subscribe to dashboards and reports. Pro users can share material with users who don’t have a Power BI Pro subscription while workspaces are at Premium capacity.

205.What are the subsets in SQL ?

The following are the four significant subsets of the SQL:

Data definition language (DDL): It defines the data structure that consists of commands like CREATE, ALTER, DROP, etc.

Data manipulation language (DML): It is used to manipulate existing data in the database. The commands in this category are SELECT, UPDATE, INSERT, etc.

Data control language (DCL): It controls access to the data stored in the database. The commands in this category include GRANT and REVOKE.

Transaction Control Language (TCL): It is used to deal with the transaction operations in the database. The commands in this category are COMMIT, ROLLBACK, SET TRANSACTION, SAVEPOINT, etc.

206. What is the purpose of DCL Language?

Data control language allows users to control access and permission management to the database. It is the subset of a database, which decides that what part of the database should be accessed by which user at what point of time. It includes two commands, GRANT and REVOKE.

GRANT: It enables system administrators to assign privileges and roles to the specific user accounts to perform specific tasks on the database.

REVOKE: It enables system administrators to revoke privileges and roles from the user accounts so that they cannot use the previously assigned permission on the database.

207. What is a Database?

A database is an organized collection of data that is structured into tables, rows, columns, and indexes. It helps the user to find the relevant information frequently. It is an electronic system that makes data access, data manipulation, data retrieval, data storing, and data management very easy. Almost every organization uses the database for storing the data due to its easily accessible and high operational ease. The database provides perfect access to data and lets us perform required tasks.

The following are the common features of a database:

Manages large amounts of data

Accurate

Easy to update

Security

Data integrity

Easy to research data

208. What is meant by DBMS?

DBMS stands for Database Management System. It is a software program that primarily functions as an interface between the database and the end-user. It provides us the power such as managing the data, the database engine, and the database schema to facilitate the organization and manipulation of data using a simple query in almost no time. It is like a File Manager that manages data in a database rather than saving it in file systems. Without the database management system, it would be far more difficult for the user to access the database's data.

The following are the components of a DBMS:

Software

Data

Procedures

Database Languages

Query Processor

Database Manager

Database Engine

Reporting

209.How will you handle missing values in data?

There are several ways to handle missing values in the given data-

1.Dropping the values

2.Deleting the observation (not always recommended).

3.Replacing value with the mean, median and mode of the observation.

4.Predicting value with regression

5.Finding appropriate value with clustering

210. What is SVM? Can you name some kernels used in SVM?

SVM stands for support vector machine. They are used for classification and prediction tasks. SVM consists of a separating plane that discriminates between the two classes of variables. This separating plane is known as hyperplane. Some of the kernels used in SVM are –

Polynomial Kernel

Gaussian Kernel

Laplace RBF Kernel

Sigmoid Kernel

Hyperbolic Kernel

211.What is market basket analysis?

Market Basket Analysis is a modeling technique based upon the theory that if you buy a certain group of items, you are more (or less) likely to buy another group of items.

212.What is the benefit of batch normalization?

The model is less sensitive to hyperparameter tuning.

High learning rates become acceptable, which results in faster training of the model.

Weight initialization becomes an easy task.

Using different non-linear activation functions becomes feasible.

Deep neural networks are simplified because of batch normalization.

It introduces mild regularisation in the network.

213. Explain the data preprocessing steps in data analysis.

Data preprocessing transforms the data into a format that is more easily and effectively processed in data mining, machine learning and other data science tasks.

1. Data profiling.

2. Data cleansing.

3. Data reduction.

4. Data transformation.

5. Data enrichment.

6. Data validation.

214. What Are the Three Stages of Building a Model in Machine Learning?

The three stages of building a machine learning model are:

Model Building: Choosing a suitable algorithm for the model and train it according to the requirement

Model Testing: Checking the accuracy of the model through the test data

Applying the Model: Making the required changes after testing and use the final model for real-time projects

216. What is a Parameter in Tableau? Give an Example.

A parameter is a dynamic value that a customer could select, and you can use it to replace constant values in calculations, filters, and reference lines.

For example, when creating a filter to show the top 10 products based on total profit instead of the fixed value, you can update the filter to show the top 10, 20, or 30 products using a parameter.

217. Explain how the filter function works in python?

The filter() method filters a series using a function that checks if each element in the sequence is true or not. The filter() function takes two arguments: function - a function and iterable - an iterable like sets, lists, tuples etc.

218. How to remove duplicate elements from a list?

First we have a List that contains duplicates. Create a dictionary, using the List items as keys. This will automatically remove any duplicates because dictionaries cannot have duplicate keys. Then, convert the dictionary back into a list.

220. What is TF/IDF vectorization?

The TF-IDF statistic, which stands for term frequency–inverse document frequency, is a numerical measure of how essential a word is to a document in a collection or corpus. It's frequently used in information retrieval, text mining, and user modelling searches as a weighting factor. The tf–idf value rises in proportion to the number of times a word appears in a document and is offset by the number of documents in the corpus that contain the term, which helps to compensate for the fact that some words appear more frequently than others.

221.How to add drop down in excel?

Here is how to create a drop-down list in Excel –

Select the cells you want to contain in the drop-down list.

Go to ribbon and click “Data,” then “Data Validation.”

In the dialog box, select “Allow to List.”

Click on Source and type the numbers or text that you want in your drop-down lists.

Click OK to continue.

Note that commas should separate the numbers and texts.

222.What are the various refresh options available in Power BI?

There are essentially four major types of refresh options available in Power BI

Package refresh

Model/data refresh

Tile Refresh

Visual container refresh

223.What is bin in tableau?

Bins in tableau are containers of equal size used to store data values fitting in bin size. In other words, bins group the data into groups of equal size or data which can be used in systematic viewing of data. All the discrete fields in tableau can also be considered as set of bins.

224. Give some statistical methods that are useful for data analysts.

Two main statistical methods are used in data analysis: descriptive statistics, which summarizes data using indexes such as mean and median and another is inferential statistics, which draw conclusions from data using statistical tests such as student's t-test. The tests include ANOVA, Kruskal-Wallis H test, Friedman Test, etc.

225. Difference B/w Drop, Truncate and Delete?

Delete is used to delete one or more tuples of a table. With the help of the “DROP” command we can drop (delete) the whole structure in one go i.e. it removes the named elements of the schema. Truncate is used to delete all the rows of a table.

226. What is the importance of a dashboard in tableau?

Building dashboards with Tableau allows even non-technical users to create interactive, real-time visualizations in minutes. In just a few clicks, they can combine data sources, add filters, and drill down into specific information.

227. Explain the transformation phase of the data pipeline.

Ans. Transformation refers to operations that change data, which may include data standardization, sorting, deduplication, validation, and verification. The ultimate goal is to make it possible to analyze the data.

228. How would you define Power BI as an effective solution?

Power BI is a strong business analytical tool that creates useful insights and reports by collating data from unrelated sources. This data can be extracted from any source like Microsoft Excel or hybrid data warehouses. Power BI drives an extreme level of utility and purpose using interactive graphical interface and visualizations. You can create reports using the Excel BI toolkit and share them on-cloud with your colleagues.

229. Power BI can connect to which data sources?

The data source is the point from which the data has been retrieved. It can be anything like files in various formats (.xlsx, .csv, .pbix, .xml, .txt etc), databases (SQL database, SQL Data Warehouse, Spark on Azure HDInsight), or form content packets like Google Analytics or Twilio.

230. How to maintain stock in excel sheet format?

Navigate to the Search bar and type Inventory list.

Press Enter and a list of templates for inventory management will appear.

Select the inventory template that fit your requirement. It can be used to maintain stock in excel sheet format.

231. How to separate text in excel?

Select the cells or the range of cells with text.

Navigate to Data, Click Data Tools> Text to Columns.

In the Text to Column, select Delimited.

Click Next and select the Delimiters for data.

Click Finish.

232. What is a True positive rate and a false positive rate?

True positive rate or Recall: It gives us the percentage of the true positives captured by the model out of all the Actual Positive class.

TPR = TP/ (TP+FN)

False Positive rate: It gives us the percentage of all the false positives by my model prediction from the all Actual Negative class.

FPR = FP/(FP+TN)

233. Which classification metric works on actual probability values rather than the class labels?

Log Loss is the only metric which works on the actual probability values rather than class labels. ROC curve also takes the probability values in consideration(it uses AUC score, where AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0; one whose predictions are 100% correct has an AUC of 1.0.)

234. What is the geometric interpretation of Decision Tree?

Geometrically a decision tree divides the space by axis parallel hyperplanes which inturn divides a space into hypercubes and hypercuboids.

235. How is Surprise and Probability related?

Surprise and probability are inversely proportional. If my probability is 1 then my surprise will be 0 and when my probability is 0 then my surprise will be maximum.

236. How is the first principal component axis selected in PCA?

In Principal Component Analysis (PCA) we look to summarize a large set of correlated variables (basically a high dimensional data) into a smaller number of representative variables, called the principal components, that explains most of the variability in the original set.

The first principal component axis is selected in a way such that it explains most of the variation in the data and is closest to all n observations.

237. What are Hard-Margin and Soft-Margin SVMs?

Hard-Margin SVMs have linearly separable training data. No data points are allowed in the margin areas. This type of linear classification is known as Hard margin classification.

Soft-Margin SVMs have training data that are not linearly separable. Margin violation means choosing a hyperplane, which can allow some data points to stay either in between the margin area or on the incorrect side of the hyperplane.

238. What is the empirical rule?

In statistics, the empirical rule states that every piece of data in a normal distribution lies within three standard deviations of the mean. It is also known as the 68–95–99.7 rule. According to the empirical rule, the percentage of values that lie in a normal distribution follow the 68%, 95%, and 99.7% rule. In other words, 68% of values will fall within one standard deviation of the mean, 95% will fall within two standard deviations, and 99.75 will fall within three standard deviations of the mean.

239. What is the left-skewed distribution and the right-skewed distribution?

In the left-skewed distribution, the left tail is longer than the right side.

Mean < median < mode

In the right-skewed distribution, the right tail is longer. It is also known as positive-skew distribution.

Mode < median < mean

240.What is the Difference Between Joining and Blending?

Combining the data from two or more different sources is data blending, such as Oracle, Excel, and SQL Server. In data blending, each data source contains its own set of dimensions and measures.

Combining the data between two or more tables or sheets within the same data source is data joining. All the combined tables or sheets contain a common set of dimensions and measures.

241. What is the difference between count, counta, and countblank?

The count function is very often used in Excel. Here, let's look at the difference between count, and it's variants - counta and countblank.

1.COUNT

It counts the number of cells that contain numeric values only. Cells that have string values, special characters, and blank cells will not be counted.

2. COUNTA

It counts the number of cells that contain any form of content. Cells that have string values, special characters, and numeric values will be counted. However, a blank cell will not be counted.

3. COUNTBLANK

As the name suggests, it counts the number of blank cells only. Cells that have content will not be taken into consideration.

242.What is a Self-Join?

A self-join is a type of join that can be used to connect two tables. As a result, it is a unary relationship. Each row of the table is attached to itself and all other rows of the same table in a self-join. As a result, a self-join is mostly used to combine and compare rows from the same database table.

243. What relationships exist between a logistic regression's coefficient and the Odds Ratio?

The coefficients and the odds ratios then represent the effect of each independent variable controlling for all of the other independent variables in the model and each coefficient can be tested for significance.

244. What's the relationship between Principal Component Analysis (PCA) and Linear & Quadratic Discriminant Analysis (LDA & QDA)

LDA focuses on finding a feature subspace that maximizes the separability between the groups. While Principal component analysis is an unsupervised Dimensionality reduction technique, it ignores the class label. PCA focuses on capturing the direction of maximum variation in the data set.The PC1 the first principal component formed by PCA will account for maximum variation in the data.PC2 does the second-best job in capturing maximum variation and so on.

The LD1 the first new axes created by Linear Discriminant Analysis will account for capturing most variation between the groups or categories and then comes LD2 and so on.

245. What's the difference between logistic and linear regression? How do you avoid local minima?

Linear Regression is used to handle regression problems whereas Logistic regression is used to handle the classification problems.

Linear regression provides a continuous output but Logistic regression provides discreet output.

The purpose of Linear Regression is to find the best-fitted line while Logistic regression is one step ahead and fitting the line values to the sigmoid curve.

The method for calculating loss function in linear regression is the mean squared error whereas for logistic regression it is maximum likelihood estimation.

We can try to prevent our loss function from getting stuck in a local minima by providing a momentum value. So, it provides a basic impulse to the loss function in a specific direction and helps the function avoid narrow or small local minima by using stochastic gradient descent.

246. Explain the difference between type 1 and type 2 errors.

Type 1 error is a false positive error that 'claims' that an incident has occurred when, in fact, nothing has occurred. The best example of a false positive error is a false fire alarm – the alarm starts ringing when there's no fire. Contrary to this, a Type 2 error is a false negative error that 'claims' nothing has occurred when something has definitely happened. It would be a Type 2 error to tell a pregnant lady that she isn't carrying a baby.

250. How do you analyse the performance of the predictions generated by regression models versus classification models?

In the regression model, the most commonly known evaluation metrics include:

- R-squared (R2)
- Root Mean Squared Error (RMSE)
- Residual Standard Error (RSE)
- Mean Absolute Error (MAE)

Classification is the problem of identifying to which of a set of categories/classes a new observation belongs, based on the training set of data containing records whose class label is known. Following are the performance metrics used for evaluating a classification model:

- Accuracy
- Precision and Recall
- Specificity
- F1-score
- AUC-ROC

252. What is cross-validation and why would you use it?

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, particularly in a case where the amount of data may be limited. In cross-validation, you make a fixed number of folds (or partitions) of the data, run the analysis on each fold, and then average the overall error estimate.

254.What is quick filter in tableau?

Whenever using a filter in Tableau, it comes with some options to change the functionality of filter very easily, such as using it as a single value drop down or single value list or multiple value list or multiple value drop down and various other options. After we set a filter to a sheet just right click on the sheet and there you can see all the quick filter options. Changes made to these options will also change the aesthetics of filter shown on the sheet.

255.How to calculate percentage in tableau?

To calculate the percentage of data on your worksheet. Go to Analysis pane and select Percentages of, there you will see a lot percentage options such as percentage of table, column, row, pane, row in pane, column in pane and cell. Select any of the above options then define the total value o which percentage is to be calculated. The option you choose will be uniform to all the rows and columns and there is no way to specify different options to rows and columns.

256.What's a Fourier transform?

A Fourier transform is a generic method to decompose generic functions into a superposition of symmetric functions. Or as this more intuitive tutorial puts it, given a smoothie, it's how we find the recipe. The Fourier transform finds the set of cycle speeds, amplitudes, and phases to match any time signal. A Fourier transform converts a signal from time to frequency domain—it's a very common way to extract features from audio signals or other time series such as sensor data.

257.How is a decision tree pruned?

Pruning is what happens in decision trees when branches that have weak predictive power are removed in order to reduce the complexity of the model and increase the predictive accuracy of a decision tree model. Pruning can happen bottom-up and top-down, with approaches such as reduced error pruning and cost complexity pruning. Reduced error pruning is perhaps the simplest version: replace each node. If it doesn't decrease predictive accuracy, keep it pruned. While simple, this heuristic actually comes pretty close to an approach that would optimize for maximum accuracy.

258. What's the F1 score? How would you use it?

The F1 score is a measure of a model's performance. It is a weighted average of the precision and recall of a model, with results tending to 1 being the best, and those tending to 0 being the worst.

259.Name an example where ensemble techniques might be useful?

Ensemble techniques use a combination of learning algorithms to optimize better predictive performance. They typically reduce overfitting in models and make the model more robust (unlikely to be influenced by small changes in the training data). You could list some examples of ensemble methods (bagging, boosting, the "bucket of models" method) and demonstrate how they could increase predictive power.

260. What is Marginalisation? Explain the process.

Marginalisation is summing the probability of a random variable X given joint probability distribution of X with other variables. It is an application of the law of total probability.

P(X=x) = ∑YP(X=x,Y)

Given the joint probability P(X=x,Y), we can use marginalization to find P(X=x). So, it is to find distribution of one random variable by exhausting cases on other random variables.

261.How do we check the normality of a data set or a feature?

Visually, we can check it using plots. There is a list of Normality checks, they are as follow:

Shapiro-Wilk W Test

Anderson-Darling Test

Martinez-Iglewicz Test

Kolmogorov-Smirnov Test

D'Agostino Skewness Test

262.What is OOB error and how does it occur?

For each bootstrap sample, there is one-third of data that was not used in the creation of the tree, i.e., it was out of the sample. This data is referred to as out of bag data. In order to get an unbiased measure of the accuracy of the model over test data, out of bag error is used. The out of bag data is passed for each tree is passed through that tree and the outputs are aggregated to give out of bag error. This percentage error is quite effective in estimating the error in the testing set and does not require further cross-validation.

263. What are the assumptions required for linear regression? What if some of these assumptions are violated?

The assumptions are as follows:

The sample data used to fit the model is representative of the population

The relationship between X and the mean of Y is linear

The variance of the residual is the same for any value of X (homoscedasticity)

Observations are independent of each other

For any value of X, Y is normally distributed.

Extreme violations of these assumptions will make the results redundant. Small violations of these assumptions will result in a greater bias or variance of the estimate.

264.What is multicollinearity and how to remove it?

Multicollinearity exists when an independent variable is highly correlated with another independent variable in a multiple regression equation. This can be problematic because it undermines the statistical significance of an independent variable.

You could use the Variance Inflation Factors (VIF) to determine if there is any multicollinearity between independent variables — a standard benchmark is that if the VIF is greater than 5 then multicollinearity exists.

265. What is overfitting and how to prevent it?

Overfitting is an error where the model 'fits' the data too well, resulting in a model with high variance and low bias. As a consequence, an overfit model will inaccurately predict new data points even though it has a high accuracy on the training data.

Few approaches to prevent overfitting are:

- Cross-Validation:Cross-validation is a powerful preventative measure against overfitting. Here we use our initial training data to generate multiple mini train-test splits. Now we use these splits to tune our model.

- Train with more data: It won't work every time, but training with more data can help algorithms detect the signal better or it can help my model to understand general trends in particular.

- We can remove irrelevant information or the noise from our dataset.

- Early Stopping: When you're training a learning algorithm iteratively, you can measure how well each iteration of the model performs.

Up until a certain number of iterations, new iterations improve the model. After that point, however, the model's ability to generalize can weaken as it begins to overfit the training data.

Early stopping refers stopping the training process before the learner passes that point.

- Regularization: It refers to a broad range of techniques for artificially forcing your model to be simpler. There are mainly 3 types of Regularization techniques:L1, L2,&,Elastic- net.

- Ensembling : Here we take number of learners and using these we get strong model. They are of two types : Bagging and Boosting.

266.How to calculate age in excel?

The easiest way to calculate age in MS Excel is using the formula –

"=DATEDIF(birth_date,as_of_date,"y")"

The result will be the number of complete years.

267.What is a legend in Power BI?

Legends are part of Power BI visuals. They represent categories in a visual and are usually color-coded. In some visuals, we can add a category dimension as legend explicitly. One example can be in stacked bar/column charts, where the stacks represent different categories, and these categories are color-coded. These categories are represented in the legend. And every visual where the legend is applicable has a separate formatting section where we can specify its font and font size, whether we need it to be visible or not, or even its position(top, bottom, left, right)

268.What is ETL in SQL?

ETL stands for Extract, Transform and Load. It is a three step process, where we would have to start off by extracting the data from sources. Once we collate the data from different sources, what we have is raw data. This raw data has to be transformed into tidy format, which will come in the second phase. Finally, we would have to load this tidy data into tools which would help us to find insights.

269.What is the main advantage of Naive Bayes?

A Naive Bayes classifier converges very quickly as compared to other models like logistic regression. As a result, we need less training data in the case of a naive Bayes classifier.

270.What should you do when your model is suffering from low bias and high variance?

When the model is suffering from low bias and high variance, it is essentially overfitting. In such a situation, techniques such as Regularization can be used or the model can be simplified by reducing the number of features in the dataset.

271.State the differences between causality and correlation?

Causality applies to situations where one action, say X, causes an outcome, say Y, whereas Correlation is just relating one action (X) to another action(Y) but X does not necessarily cause Y.

272.A data set is given to you and it has missing values which spread along 1 standard deviation from the mean. How much of the data would remain untouched?

It is given that the data is spread across mean that is the data is spread across an average. So, we can presume that it is a normal distribution. In a normal distribution, about 68% of data lies in 1 standard deviation from averages like mean, mode or median. That means about 32% of the data remains uninfluenced by missing values.

273.How will you define the number of clusters in a clustering algorithm?

By determining the Silhouette score and elbow method, we determine the number of clusters in the algorithm.

274.What is Ensemble Learning? Define types.

Ensemble learning is clubbing of multiple weak learners (ml classifiers) and then using aggregation for result prediction. It is observed that even if the classifiers perform poorly individually, they do better when their results are aggregated. An example of ensemble learning is random forest classifier.

276.What is pruning in Decision Tree?

Pruning is the process of reducing the size of a decision tree. The reason for pruning is that the trees prepared by the base algorithm can be prone to overfitting as they become incredibly large and complex

277. What is subquery in SQL?

A subquery is a query inside another query where a query is defined to retrieve data or information back from the database. In subquery, the outer query is called as the main query where the inner query is called subquery. Subqueries are always executed first and then result of subquery is passed to main query. It can be nested inside a SELECT, UPDATE or any other query.

278. What is Tower of Hanoi?

Tower of Hanoi is a mathematical puzzle that shows how recursion might be utilised as a device in building up an algorithm to take care of a specific problem. Using decision tree and a breadth-first search(BFS) algorithm in AI, we can solve the Tower of Hanoi.

279. What are the two main methods two detect outliers?

Box plot method: if the value is higher or lesser than 1.5*IQR (interquartile range) above the upper quartile (Q3) or below the lower quartile (Q1) respectively, then it is considered an outlier.

Standard deviation method: if value higher or lower than mean ± (3*standard deviation), then it is considered an outlier.

280. What is the role of a data model for any organization?

With the help of a data model, you can always keep your client informed in advance for a time period. However, when you enter a new market then you are facing new challenges almost every day. A data model helps you in understanding these challenges in the best way and deriving the accurate outputs from the same..

281.What are the various steps involved in an data analytics project?

The steps involved in a data analytics project are:

Data collection

Data cleansing

Data pre-processing

EDA

Creation of train test and validation sets

Model creation

Hyperparameter tuning

Model deployment

282. What is root cause analysis?

Root cause analysis is the process of tracing back of occurrence of an event and the factors which lead to it. It's generally done when a software malfunctions. In data science, root cause analysis helps businesses understand the semantics behind certain outcomes.

283. Define Confounding Variables.

A confounding variable is an external influence in an experiment. In simple words, these variables change the effect of a dependent and independent variable. A variable should satisfy below conditions to be a confounding variable :

Variables should be correlated to the independent variable.

Variables should be informally related to the dependent variable.

For example, if you are studying whether a lack of exercise has an effect on weight gain, then the lack of exercise is an independent variable and weight gain is a dependent variable. A confounder variable can be any other factor that has an effect on weight gain. Amount of food consumed, weather conditions etc. can be a confounding variable.

284. What is RUP methodology ?

Rational Unified Process is a product application improvement method with numerous devices to help with the coding the last product and assignments identified with this objective. RUP is an object oriented approach that guarantees successful project management and top-notch software production.

285. How does CATWOE help in business analysis and decision making?

Customers, Actors, Transformation process, Worldwide, Owner and Environment constraint (CATWOE) helps in making decisions ahead of time. It includes analysing how those decisions will affect customers (C);who are involved as actors(A);what difference transformation (T) processes are which might affect the system, global picture and Worldwide(W) issues; who is responsible/has ownership (O) for the business; what the environment (E) impacts will be of the projects/business.

286. What is RAD methodology ?

The Rapid Application Development model is a kind of incremental model. The phases of a project are produced in parallel as individuals projects. The development in the project are timeboxed, delivered and afterward assembled into a working model.

287. What is INVEST ?

INVEST is an abbreviated of Independent, Negotiable, Valuable, Estimable, Sized appropriately and Testable. This term is used by business analyst and project managers to deliver quality services and products.

288. What is correlation and covariance in statistics?

Correlation is defined as the measure of the relationship between two variables. If two variables are directly proportional to each other, then its positive correlation. If the variables are indirectly proportional to each other, it is known as a negative correlation. Covariance is the measure of how much two random variables vary together.

289.What is 'Naive' in a Naive Bayes?

A naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature, given the class variable. Basically, it's "naive" because it makes assumptions that may or may not turn out to be correct.

290. How can you select k for k-means?

The two methods to calculate the optimal value of k in k-means are:

Elbow method

Silhouette score method

Silhouette score is the most prevalent while determining the optimal value of k.

291.How to get the minimum, 25th percentile, median, 75th, and max of a numeric series?

```
"randomness= np.random.RandomState(100)

s = pd.Series(randomness.normal(100, 55, 5))

np.percentile(ser, q=[0, 25, 50, 75, 100])"
```

292. How would you deal with missing random values from a data set?

There are two forms of randomly missing values:

MCAR or Missing completely at random. Such errors happen when the missing values are randomly distributed across all observations.

We can confirm this error by partitioning the data into two parts –

One set with the missing values

Another set with the non-missing values.

After we have partitioned the data, we conduct a t-test of mean difference to check if there is any difference in the sample between the two data sets.

In case the data are MCAR, we may choose a pair-wise or a list-wise deletion of missing value cases.

MAR or Missing at random. It is a common occurrence. Here, the missing values are not randomly distributed across observations but are distributed within one or more sub-samples. We cannot predict the probability from the variables in the model. Data imputation is mainly performed to replace them.

293. Explain the difference between lists and tuples.

Both lists and tuples are made up of elements, which are values of any Python data type. However, these data types have a number of differences:

Lists are mutable, while tuples are immutable.

Lists are created in brackets (for example, my_list = [a, b, c]), while tuples are in parentheses (for example, my_tuple = (a, b, c)).

294.What is 'fsck'?

'fsck ' abbreviation for ' file system check.' It is a type of command that searches for possible errors in the file. fsck generates a summary report, which lists the file system's overall health and sends it to the Hadoop distributed file system.

295.What is method to convert date-strings to timeseries in a series.

Input:

s = pd.Series(['22 Feb 1984', '22-02-2013', '20170105', '2012/02/08', '2016-11-04', '2015-03-02T11:15'])

We will use the to_datetime() function

pd.to_datetime(s)

296.What is the Cartesian product of the table?

The output of Cross Join is called a Cartesian product. It returns rows combining each row from the first table with each row of the second table. For Example, if we join two tables having 15 and 20 columns the Cartesian product of two tables will be 15×20=300 rows.

297. Difference between NVL and NVL2 functions?

Both the NVL(exp1, exp2) and NVL2(exp1, exp2, exp3) functions check the value exp1 to see if it is null. With the NVL(exp1, exp2) function, if exp1 is not null, then the value of exp1 is returned; otherwise, the value of exp2 is returned, but case to the same data type as that of exp1. With the

NVL2(exp1, exp2, exp3) function, if exp1 is not null, then exp2 is returned; otherwise, the value of exp3 is returned.

300.What is market basket analysis?

Market Basket Analysis is a modeling technique based upon the theory that if you buy a certain group of items, you are more (or less) likely to buy another group of items.

301.What is the benefit of batch normalization?

The model is less sensitive to hyperparameter tuning.

High learning rates become acceptable, which results in faster training of the model.

Weight initialization becomes an easy task.

Using different non-linear activation functions becomes feasible.

Deep neural networks are simplified because of batch normalization.

It introduces mild regularisation in the network.

302. How do you assess logistic regression versus simple linear regression models?

• Linear Regression is used to handle regression problems whereas Logistic regression is used to handle the classification problems.

• Linear regression provides a continuous output but Logistic regression provides discreet output(actually it gives continuous probability output which is then converted into discreet values I.e. classes). • The purpose of Linear Regression is to find the best-fitted line while Logistic regression is one step ahead and fitting the line values to the sigmoid curve.

• The method for calculating loss function in linear regression is the mean squared error whereas for logistic regression it is maximum likelihood estimation.

303. What is cross-validation and why would you use it?

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, particularly in a case where the amount of data may be limited. In cross-validation, you make a fixed number of folds (or partitions) of the data, run the analysis on each fold, and then average the overall error estimate.

304. What's the name of the matrix used to evaluate predictive models?

Confusion Matrix: A confusion matrix is a tabular summary of the number of correct and incorrect predictions made by a classifier. It can be used to evaluate the performance of a classification model through the calculation of performance metrics like accuracy, precision, recall, and F1-score

305.How to create filters in Power BI?

Filters are an integral part of Power BI reports. They are used to slice and dice the data as per the dimensions we want. Filters are created in a couple of ways.

Using Slicers: A slicer is a visual under Visualization Pane. This can be added to the design view to filter our reports. When a slicer is added to the design view, it requires a field to be added to it. For example- Slicer can be added for Country fields. Then the data can be filtered based on countries.

Using Filter Pane: The Power BI team has added a filter pane to the reports, which is a single space where we can add different fields as filters. And these fields can be added depending on whether you want to filter only one visual(Visual level filter), or all the visuals in the report page(Page level filters), or applicable to all the pages of the report(report level filters)

306. What is MDX in Power BI?

MDX is an Analysis Services Tabular capability that is basically used to help improve the user experience in Excel and various other applications that use MDX query language against the tabular form of the models and Power BI datasets. To have a hands-on the MDX, try it by deploying your datasets in Power BI and then connecting with Excel by using analyze in Excel, the new Power BI Datasets within Excel feature, or you can connect to this dataset on Power BI Premium as well through the XMLA endpoint.

307.How to convert numbers to words in excel in rupees?

Select a cell adjacent to the cell you want the currency (Rupees) to be converted into words.

Enter the formula – NumberstoWords(A1) and press Enter. A1 is the cell containing the currency in Rupees.

Press Enter to confirm the formula.

308.How to round off in excel?

You can use the ROUNDUP function to round off in Excel. Here is how –

Select a blank cell.

Use the ROUND function by entering the syntax – "=ROUNDUP(number,num_digits).

Where number denotes the number to round up and num_digits refers to a number of digits to which it should be rounded up.

Press Return.

Using the button, you can round off in Excel by following these steps –

Select the cells you want to round off.

Go to the Home tab, click Increase Decimal or Decrease Decimal.

Press Enter.

310.What are the support vectors in SVM?

Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, we maximise the margin of the classifier. Deleting the support vectors will change the position of the hyperplane. These are the points that help us build our SVM.

311.What is pruning in Decision Tree?

Pruning is the process of reducing the size of a decision tree. The reason for pruning is that the trees prepared by the base algorithm can be prone to overfitting as they become incredibly large and complex.

312. Name the operator which is used in the query for pattern matching?

LIKE operator is used for pattern matching, and it can be used as - % _ It matches zero or more characters.

For eg: select * from student where studentname like 'a%'

_(underscore)- it matches exactly one character.

Select* from student where studentname like 'abc_'

313. In what terms DBSCAN is better than K- Means Clustering?

Outliers and noisy datasets are easily handled by DBScan clustering, whereas outliers and noisy datasets are not adequately handled by K-means clustering. The number of clusters does not need to be stated in DBScan clustering. The number of clusters specified in K-means clustering is important.

314.How to compare two columns in excel?

You can compare two columns in Excel using the simple IF formula. The syntax to be used is – "=IF(A2=B2,"Match," Mismatch")

If the values to be compared are case-sensitive, use these following formula –

=IF(EXACT(A2,B2),"Match","Mismatch")

315.How to add logo in tableau dashboard?

In the objects pane in dashboard there is an option to import image. Make sure that you have selected the floating type instead of by default tiled option. Drag and drop image object to dashboard and then select the logo from your system saved in .jpg, .png or .jpeg format. To use it as a background user can use the small drop down option on the right side of the image and select send to back and then increase the size of image so as to make it as a watermark logo.

316.Which two cross filter directions are available in Power BI table relationships?

When a relationship is created between two different tables in Power BI, then the relationship asks us the cross-filter direction. There are two options available for cross-filtering.

Single – When the cross filter direction is single, then the filtering between tables happens from left table to right table. It is the default setting. The first table can be used to filter the data in the second table.Both – When the cross filter direction is both, then the filtering between the tables will work in both ways. Either table can be used to filter the other table.

317.How should you maintain a deployed model?

A deployed model needs to be retrained after a while so as to improve the performance of the model. Since deployment, a track should be kept of the predictions made by the model and the truth values. Later this can be used to retrain the model with the new data. Also, root cause analysis for wrong predictions should be done.

318.What is the difference between append() and extend() methods?

append() is used to add items to list. extend() uses an iterator to iterate over its argument and adds each element in the argument to the list and extends it.

319.What is skewed Distribution & uniform distribution?

The skewed distribution is a distribution in which the majority of the data points lie to the right or left of the centre. A uniform distribution is a probability distribution in which all outcomes are equally likely.

320.What are lambda functions?

A lambda function is a small anonymous function. A lambda function can take any number of arguments, but can only have one expression.

322. Given two fair dices, what is the probability of getting scores that sum to 4 and 8?

There are 4 combinations of rolling a 4 (1+3, 3+1, 2+2):

P(rolling a 4) = 3/36 = 1/12

There are 5 combinations of rolling an 8 (2+6, 6+2, 3+5, 5+3, 4+4):

P(rolling an 8) = 5/36

323. Differentiate between univariate, bivariate, and multivariate analysis.

Univariate data, as the name suggests, contains only one variable. The univariate analysis describes the data and finds patterns that exist within it.

Bivariate data contains two different variables. The bivariate analysis deals with causes, relationships and analysis between those two variables.

Multivariate analysis involves analyzing multiple variables (more than two) to identify any possible association among them

324. What is dimensionality reduction? What are its benefits?

Dimensionality reduction is defined as the process of converting a data set with vast dimensions into data with lesser dimensions — in order to convey similar information concisely.

This method is mainly beneficial in compressing data and reducing storage space. It is also useful in reducing computation time due to fewer dimensions. Finally, it helps remove redundant features — for instance, storing a value in two different units (meters and inches) is avoided.

In short, dimensionality reduction is the process of reducing the number of random variables under consideration, by obtaining a set of principal variables. It can be divided into feature selection and feature extraction.

325. What is a ROC Curve? Explain how a ROC Curve works?

AUC – ROC curve is a performance measurement for the classification problem at various thresholds settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s.

326. What are the various steps involved in an analytics project?

The steps involved in a text analytics project are:

-Data collection

-Data cleansing

-Data pre-processing

-Creation of train test and validation sets

-Model creation

-Hyperparameter tuning

-Model deployment

327.What is the extension of the saved file in ms excel?

Microsoft Excel saved file has a .xls extension. The XLS-based extension has been used as the default format for Microsoft Excel 1997-2003. Files saved in the .xlsx extension are associated with Microsoft Excel 2007-10.

328.How to calculate EMI in excel?

EMI can be calculated in Excel using the simple formula –

“PMT (rate, nper, pv, [fv], [type])”

Where PMT – Payment;

Rate – the interest rate on the EMI;

Nper – total number of EMIs paid;

Pv – Present value or outstanding amount;

Fv (optional) – Future value or Cash balance;

Type (optional) – when payment is due. The default value is 0.

329. What is DAX?

Data Analysis Expressions (DAX) is a scripting language that is used to create calculated columns, measurements, and custom tables in Microsoft Power BI. It is a set of functions, operators, and constants that may be used to compute and return one or more values using a formula, or expression. You may utilise DAX to handle a variety of computations and data analysis challenges, allowing you to generate new data from existing data in your model.

330. How to connect JIRA to Power BI?

We connect to JIRA using JIRA content packs and JIRA Rest API. There are a few steps involved.

In order to connect JIRA to Power BI, we need an API token that can be generated from the Atlassian link.

Next, we download and install the Jira content pack. Then the API token is added to the Jira content pack.Once the connection is set, it automatically creates a workspace with a dashboard and a set of reports to visually analyze the data.

331.How to share Power BI dashboards?

Power BI reports/dashboards can be shared in multiple ways. If you and your end-users have a Power BI Pro license, then-

-We are using the share option in reports and dashboards :This option enables access to a report or dashboard to individual users.

-Using content packs :Publish your report/dashboard along with the data set as a content pack and then share it with either a group or individual or open it for the entire organization.

-Publish your dashboards and reports into App Workspace and share the App link to a group or individual or open it for the entire organization. When publishing as App- we can select the reports and datasets that we want to be included in the App.

-The report can be embedded into Web or SharePoint, for which we need the embed code. This embed code is added to the website code or the SharePoint code.

Another way to share the reports is by printing or exporting the report as PPT or PDF. With these options, we can share the report, but they are not interactive.

332. What is Lock Aspect in Power BI?

Lock Aspect is basically used to lock the proportion of a visual when you try to drag through the corners to resize it. So, for example, if you create a visualization chart & turn the Lock Aspect on – and then try to resize it with the help of your mouse, it won't change, and it keeps proportions the same as earlier.

You can find the lock aspect in the Format Tab in the Visualization pane.

333. How to delete duplicate rows in excel?

Select the range of cells you want to remove duplicates from.

Go to the Data tab and select the Data Tools group.

Click Remove Duplicates.

Under columns, select one or more columns that contain duplicates.

Click ok to continue.

334. How to protect an excel sheet?

Select the worksheet you want to protect.

Navigate to the Review tab and click Protect Sheet.

A dialog box will open with the "Allow all users of this worksheet to" list.

Check "Protect worksheet and contents of locked cells."

Check the options you want, such as "Select locked cells."

Click OK to continue.

335.Explain Normal Distribution.

Normal Distribution is also called the Gaussian Distribution. It is a type of probability distribution such that most of the values lie near the mean. It has the following characteristics:

-The mean, median, and mode of the distribution coincide

-The distribution has a bell-shaped curve

-The total area under the curve is 1

-Exactly half of the values are to the right of the centre, and the other half to the left of the centre

336. Mention some drawbacks of the Linear Model.

Here a few drawbacks of the linear model:

-The assumption regarding the linearity of the errors

-It is not usable for binary outcomes or count outcome

-It can't solve certain overfitting problem

-It also assumes that there is no multicollinearity in the data.

337. What is Power of Test?

Power of Test: The Power of the test is defined as the probability of rejecting the null hypothesis when the null hypothesis is false. Since β is the probability of a Type II error, the power of the test is defined as 1- β. In advanced statistics, we compare various types of tests based on their size and power, where the size denotes the actual proportion of rejections when the null is true and the power denotes the actual proportion of rejections when the null is false.

338. Which are the important steps of Data Cleaning?

Different types of data require different types of cleaning, the most important steps of Data Cleaning are:

Data Quality

Removing Duplicate Data (also irrelevant data)

Structural errors

Outliers

Treatment for Missing Data

Data Cleaning is an important step before analysing data, it helps to increase the accuracy of the model. This helps organisations to make an informed decision.

Data Scientists usually spends 80% of their time cleaning data.

340.What is the Confusion Matrix?

A confusion matrix is a tabular summary of the number of correct and incorrect predictions made by a classifier. It is used to measure the performance of a classification model. It can be used to evaluate the performance of a classification model through the calculation of performance metrics like accuracy, precision, recall, and F1-score.

For example, in a case of binary classifier, it predicts all data instances of a test dataset as either positive or negative. This produces four outcomes-

True positive(TP) — Correct positive prediction

False-positive(FP) — Incorrect positive prediction

True negative(TN) — Correct negative prediction

False-negative(FN) — Incorrect negative prediction

It helps in calculating various measures including error rate (FP+FN)/(P+N), specificity(TN/N), accuracy(TP+TN)/(P+N), sensitivity (TP/P), and precision(TP/(TP+FP)).

A confusion matrix is essentially used to evaluate the performance of a machine learning model when the truth values of the experiments are already known and the target class has two or more than two categories of data. It helps in visualisation and evaluation of the results of the statistical process.

341.What are the different types of functions generally used in Tableau?

In Tableau, we have a lot of processing and analytical freedom with the virtue of functions available.

With the help of different types of functions, we can perform a lot of analytical operations on the data.

The main categories of Tableau function are:

String function: These functions like ASCII, CHAR, FIND, ISDATE, LOWER, etc, are known as string functions because they work on the string values or characters to manipulate them.

Date function: We use date functions to apply logical as well as arithmetic operations on date values present at the data source. Using the date functions we can manipulate the date values by changing the old values, creating new ones or searching data on the basis of specific dates.

Some commonly used date functions in Tableau are DATEADD, MAKEDATE, ISDATE, MAKETIME, MONTH, MIN/MAX, TODAY, NOW, etc.

Logical function: We use logical functions to perform logical or relational operations on data in Tableau.

Some commonly used logical functions in Tableau are, CASE, IF, IFNULL, ISNULL, ZN, etc.

Aggregate function: We use aggregate functions to apply aggregation on data values in different ways.

Some important aggregation functions used in Tableau are; AVG, ATTR, MAX, MEDIAN, MIN, PERCENTILE, SUM, STDDEV, etc.

User function: We use functions to manage the users registered on Tableau Server or Tableau Online.

Commonly used user functions are, FULLNAME, ISFULLNAME, ISUSERNAME, USERDOMAIN, USERNAME, etc.

342.What do you understand by context filters?

Context filters are used to apply context on the data under analysis.

By applying a context we set a perspective according to which we can see the charts and graphs.

For example, we have sales data of an electronic store and we want to conduct our analysis only for the corporate sector or segment.

To do this, we have to apply a context filter on our Tableau sheet. Once we add the context for the Corporate segment from the Add to context option, all the charts present on the sheet will only show data relevant to the Corporate segment.

In this way, we can apply a context to our analysis in Tableau.

343. How do you add a Note to a cell?

To add a Note, select the cell and right-click on the same. then select the New Note option and type in any note that you wish to. In case you want to delete the Note, follow the same procedure and select the Delete Note option. Notes are indicated by a red triangle at the top-right corner of the cell.

344. How does a Decision Tree handle continuous(numerical) features?

Decision Trees handle continuous features by converting these continuous features to a threshold-based boolean feature.

To decide The threshold value, we use the concept of Information Gain, choosing that threshold that maximizes the information gain.

345. What are Loss Function and Cost Functions?

the loss function is to capture the difference between the actual and predicted values for a single record whereas cost functions aggregate the difference for the entire training dataset.

The Most commonly used loss functions are Mean-squared error and Hinge loss.

346. What is the difference between Python Arrays and lists?

Arrays in python can only contain elements of same data types i.e., data type of array should be homogeneous. It is a thin wrapper around C language arrays and consumes far less memory than lists.

Lists in python can contain elements of different data types i.e., data type of lists can be heterogeneous. It has the disadvantage of consuming large memory.

347. Write a Program to add two integers >0 without using the plus operator.

```
def add_nums(num1, num2):
while num2 != 0:
data = num1 & num2
num1 = num1 ^ num2
num2 = data << 1
return num1
print(add_nums(2, 10))
```

348. Write a query to fetch all employees who also hold the managerial position.

```
SELECT E.EmpFname, E.EmpLname, P.EmpPosition FROM EmployeeInfo E INNER JOIN
EmployeePosition P ON E.EmpID = P.EmpID AND P.EmpPosition IN ('Manager');
```

349. What is root cause analysis? How to identify a cause vs. a correlation? Give examples.

Ans. Root cause analysis: a method of problem-solving used for identifying the root cause(s) of a problem [5]

Correlation measures the relationship between two variables, range from -1 to 1. Causation is when a first event appears to have caused a second event. Causation essentially looks at direct relationships while correlation can look at both direct and indirect relationships.

Example: a higher crime rate is associated with higher sales in ice cream in Canada, aka they are positively correlated. However, this doesn't mean that one causes another. Instead, it's because both occur more when it's warmer outside. You can test for causation using hypothesis testing or A/B testing.

350.How would you track down the last line and segment in VBA?

To track down the last column, utilize the underneath lines code in the VBA module:

```
Sub FindingLastRow()

Faint lastRow As Long

lastRow = ActiveSheet.Cells.SpecialCells(xlLastCell).Row

MsgBox (lastRow)

End Sub
```

To track down the last section, utilize the beneath lines code in the VBA module:

```
Sub FindingLastColumn()

Faint lastRow As Long

lastColumn = ActiveSheet.Cells.SpecialCells(xlLastCell).Column

MsgBox (lastColumn)

End Sub
```

351.How to increase size of pie in tableau?

Creating a pie chart requires atleast one measure and one dimension in row and shelf column. Then you can select a pie chart from either the show me option on the right side of the screen or from the marks card change automatic to pie. Then give some detailing to the pie chart by using a dimension in colour and measure in angle. Option to increase the size also comes under marks card. Click on the size option and then move the slider towards right to increase its size.

352.How to calculate the average in Power BI?

Average can be calculated in two ways-

one is when we add a measure to a visual; by default, it summarizes any measure. When we click on the drop-down for the measure- we can change from Sum to Average. This gives us an average.

The second one is creating a calculated measure for average using the AVERAGE() DAX function.

353.Where is the data stored in Power BI?

The nation or area for the business identity is chosen by the first user in your organization who signs up for Power BI or Microsoft 365. The cloud's shared identity and access management service, Azure Active Directory (AAD), establishes a tenant in the data center region nearest to the chosen nation or area. AAD is a multi-tenant service, with each enterprise represented in the data center as a separate tenant. The data is saved in the area you choose at sign-up. This region will be the same for all users in your organization, regardless of their location. The chosen region should, ideally, be in the same geographical area as the majority of your consumers.

354.What is covariance?

Covariance signifies the direction of the linear relationship between the two variables. By direction we mean if the variables are directly proportional or inversely proportional to each other. (Increasing the value of one variable might have a positive or a negative impact on the value of the other variable).

The values of covariance can be between -infinity to +infinity.

355.How can you calculate the p-value using MS Excel?

The formula used in MS Excel to calculate p-value is –

=tdist(x,deg_freedom,tails)

The p-value is expressed in decimals in Excel. Here are the steps to calculate it –

-Find the Data tab

On the Analysis tab, click on the data analysis icon

-Select Descriptive Statistics and then click OK

-Select the relevant column

Input the confidence level and other variables

356. What is formula of exponential smoothing?

The simplest form of an exponential smoothing formula is given by:

$st = \alpha xt+(1 - \alpha)st-1= st-1+ \alpha(xt - st-1)$

Here,

st = smoothed statistic, it is the simple weighted average of current observation xt

$st-1$ = previous smoothed statistic

α = smoothing factor of data; $0 < \alpha < 1$

t = time period

If the value of the smoothing factor is larger, then the level of smoothing will reduce. Value of α close to 1 has less of a smoothing effect and give greater weight to recent changes in the data, while the value of α closer to zero has a greater smoothing effect and are less responsive to recent changes.

357. Consider a small unit of a factory where there are 5 employees : a supervisor and four labourers. The workers earn a salary of Rs. 5,000 per month each while the supervisor gets Rs. 15,000 per month. Calculate the mean, median and mode of the salaries.

Mean = (5000 + 5000 + 5000 + 5000 + 15000)/5 = 35000/5 = 7000

So, the mean salary is Rs. 7000 per month

To obtain the median, let us arrange the salaries in ascending order:

5000, 5000, 5000, 5000, 15000

Median = (n+1)/2 = (5+1)/2 = 6/2 = 3rd observation

Median = Rs. 5000/-

Mode = Number of times an observation is repeated = Rs.5000/

358.What is a heatmap? Give an example.

A heatmap is a type of visualization used to demonstrate a set of data through varying shades of colours where the darkest shade of a specific colour denotes an extreme value (high intensity/density). It is typically used to compare two or more measures.

A quick example of a heatmap would be to understand the anatomy of the human body and observe the level of warmth depending upon the temperature of specific organs. If the red-yellow combination of colours is used, the areas that show red will denote the maximum temperature.

359. What is DRIVE Program Methodology?

It is a product of iterative sessions previously used and tested by enterprise deployments. It is based on best practises and allows a user to follow a specific set of actions to avoid errors and expedite reporting or visualization process.

360. When does regularization come into play in Machine Learning?

At times when the model begins to underfit or overfit, regularization becomes necessary. It is a regression that diverts or regularizes the coefficient estimates towards zero. It reduces flexibility and discourages learning in a model to avoid the risk of overfitting. The model complexity is reduced and it becomes better at predicting.

361. What is the order of operations in Excel?

Excel follows PEMDAS: parentheticals, exponents, multiplication, division, addition, and then subtraction. If you type in “=1+2/4” the answer will be 3/2 rather than ¾.

362.How to remove space in excel?

To remove extra spaces in Excel –

Press Ctrl+Space to select cells in a column.

Open the “Find & Replace” dialog box using the shortcut Ctrl+H.

In the “Find What” field, press the spacebar.

Leave the “Replace with” field empty.

Click on “Replace all” and press OK. All spaces will be removed.

363.What is Power BI Cloud?

Power BI Cloud is a Cloud-Based Data Visualization set-up that gives access to its users for editing, viewing, creating, and sharing the dashboards’ reports & datasets across organizations and a group of people by protecting the Admin Rights & Credentials. Access is required to view & edit these reports & dashboards.

364. How to combine two columns in Power BI?

To combine two columns in Power BI, you need to select a new calculated column using the DAX expression.

For example, if you need to add the values of two columns & make the third column, then you can do it by the Following DAX Expression:

Total=Sheet Name(Col1) +sheet Name(Col2). In the image below, two text columns City & State, are added & concatenated to one new column. In the same way, you can add Numeric columns using DAX.

365.How to sort data in Power BI?

Sorting is available in multiple formats. In the data view, a common sorting option of alphabetical order is there. Apart from that, we have the option of Sort by column, where one can sort a column based on another column. The sorting option is available in visuals as well. Sort by ascending and descending option by the fields and measure present in the visual is also available.

366.How to convert pdf to excel?

Open the PDF document you want to convert in XLSX format in Acrobat DC.

Go to the right pane and click on the “Export PDF” option.

Choose spreadsheet as the Export format.

Select “Microsoft Excel Workbook.”

Now click “Export.”

Download the converted file or share it.

367. How to enable macros in excel?

Click the file tab and then click “Options.”

A dialog box will appear. In the “Excel Options” dialog box, click on the “Trust Center” and then “Trust Center Settings.”

Go to the “Macro Settings” and select “enable all macros.”

Click OK to apply the macro settings.

368. For the given points, how will you calculate the Euclidean distance in Python?

plot1 = [1,3]

plot2 = [2,5]

The Euclidean distance can be calculated as follows:

euclidean_distance = sqrt((plot1[0]-plot2[0])*2 + (plot1[1]-plot2[1])*2)

369.Which of the following machine learning algorithms can be used for inputting missing values of both categorical and continuous variables?

K-means clustering

Linear regression

K-NN (k-nearest neighbor)

Decision trees

The K nearest neighbor algorithm can be used because it can compute the nearest neighbor and if it doesn't have a value, it just computes the nearest neighbor based on all the other features.

When you're dealing with K-means clustering or linear regression, you need to do that in your pre-processing, otherwise, they'll crash. Decision trees also have the same problem, although there is some variance.

370.How are confidence tests and hypothesis tests similar? How are they different?

Confidence intervals and hypothesis testing are both tools used for to make statistical inferences.

The confidence interval suggests a range of values for an unknown parameter and is then associated with a confidence level that the true parameter is within the suggested range of. Confidence intervals are often very important in medical research to provide researchers with a stronger basis for their estimations. A confidence interval can be shown as “10 +/- 0.5” or [9.5, 10.5] to give an example.

Hypothesis testing is the basis of any research question and often comes down to trying to prove something did not happen by chance. For example, you could try to prove when rolling a dye, one number was more likely to come up than the rest.

371. What is the difference between observational and experimental data?

Observational data comes from observational studies which are when you observe certain variables and try to determine if there is any correlation.

Experimental data comes from experimental studies which are when you control certain variables and hold them constant to determine if there is any causality.

An example of experimental design is the following: split a group up into two. The control group lives their lives normally. The test group is told to drink a glass of wine every night for 30 days. Then research can be conducted to see how wine affects sleep.

372.What is a legend in Power BI?

Legends are part of Power BI visuals. They represent categories in a visual and are usually color-coded. In some visuals, we can add a category dimension as legend explicitly. One example can be in stacked bar/column charts, where the stacks represent different categories, and these categories are color-coded. These categories are represented in the legend. And every visual where the legend is applicable has a separate formatting section where we can specify its font and font size, whether we need it to be visible or not, or even its position(top, bottom, left, right)

373.What is drill down in Power BI?

Drill down, as the name suggests, is going downwards. And in Power BI, it is going down a hierarchical category or dimension. Basically, when we have data attributes like a date where we have a default hierarchy of year->quarter->month->day, then when using the same field in a visual, in Power BI, it by default takes the hierarchy form of the field. This gives the option of a drill-down feature denoted by up and down arrows. Also, we can right-click on the element in visual and get the option of drill-down.

374.What is load testing in Tableau? How do you perform it?

You need to load testing in Tableau to understand the server's capacity concerning its environment, workload, data, and use. You should conduct load testing at least 3 to 4 times a year. That is because workload, data, usage upgrade, or content authoring change whenever a new user joins in.

You must know that Tabjolt was created by Tableau to conduct point and run load and performance testing, particularly in Tableau servers. The features of Tabjolt are:

Eliminates dependency on script maintenance or script development

Automates user-specified loads' processes

Scales linearly by adding more nodes to the cluster

375.What is the maximum number of tables you can join in Tableau?

It is not possible to join more than 32 tables together in Tableau. It means you can join a maximum of 32 tables in Tableau.

376.What is the use of cycle fields in tableau?

Cycle fields help in switching and trying different colour combinations or views in a cyclic order. It will work only if we have a chart that allows more than one measure such as stacked bar chart and we are unable to finalize the visualizations then we can use cycle fields. To use cycle field, click on analysis menu in the toolbar then select cycle fields to take a quick look at an alternative visualization.

377.How to integrate tableau with website?

To integrate Tableau with a website, developer needs to have complete understanding of Javascript API for tableau. It contains all the types of functions required to view and control a tableau worksheet and dashboard directly from the website instead of interacting directly with the worksheet. Tableau has provided a complete list of functions used in Javascript API and implementation of some of the important functions is explained in the tutorial section also which is free for all users. All the tableau dashboards or worksheets when published on tableau public or online or server are integrated with the Javascrip API by default developer only needs to call that API in the HTML code and start interacting with it.

378.What is OData feed in Power BI?

OData stands for open data protocol. It is usually used to pull data from websites, SharePoint. It helps in getting data from URLs. It basically helps to extract data from a URL without getting into the details of URL-specific parameters like request, response, HTTP methods, etc. It takes care of these things in the back end, leaving us to focus on pulling the data and performing transformations, and creating useful analysis.

379.Explain Pivot tables along with their features?

Pivot Tables are statistical tables that condense data of those tables that have extensive information. The summary can be based on any field such as sales, averages, sums, etc that the pivot table represents in a simple and intelligent manner.

Features:

Some of the features of Excel Pivot Tables are as follows:

Allow the display of exact data you have to analyze

Provide various angles to view the data

Allow you to focus on important details

Comparison of data is very handy

Pivot tables can detect different patterns, relationships, data trends, etc

They can create instant data

Accurate reports

Serve the base for Pivot charts

380.What is a Horizontal Lookup in Microsoft Excel?

Horizontal Lookup or HLOOKUP looks for a value from the topmost row of the table horizontally and then moves in a downward direction.

Syntax of HLOOKUP is, = HLOOKUP (Val, giventable, row_no, [rnge_look]),

Where,

Val is the value to be searched in the first row of the table.

giventable is the row/rows that are sorted in ascending order.

row_no is the row from which the lookup value is to be recovered.

[rnge_look] is not a mandatory argument where TRUE (default) means inexact match and FALSE means exact match.

381.What is the use of cycle fields in tableau?

Cycle fields help in switching and trying different colour combinations or views in a cyclic order. It will work only if we have a chart that allows more than one measure such as stacked bar chart and we are unable to finalize the visualizations then we can use cycle fields. To use cycle field, click on analysis menu in the toolbar then select cycle fields to take a quick look at an alternative visualization.

382.What is the calculated column in Power BI?

Calculated columns are built to extend the data attributes. They are those columns that are created when the available columns in the data do not serve our purpose or we are not able to generate any useful insight from the same. That is when calculated columns come into the picture. And these are created using different DAX functions as per need. We can have a simple example of a date where we do not want to work with the complete date but individual day, month, or year. For this, in the data part of the Power BI desktop, we have the option of a ‘New Column.’ When we click that, we get a formula bar on top (like excel), where it asks us to type the new column name with its calculation. So, if we consider the data example, it will be something like this: yearcol = YEAR([Date column]) where YEAR is a DAX function. And ‘yearcol’ is your calculated column.

383.How to add custom colors in tableau?

Tableau offers various colour palettes which a user can use to define a legend or to be used in formatting. But sometimes due to client’s requirements we need to use some specific colours in our view. So instead of defining the colour each and every time you can create a custom palette which can be used whenever you open a workbook. To create a custom colour palette, go to “My tableau Repository” in the documents folder of your system. Open “Preference.tps” file in a text editor to create the custom palette. By default, the file will be empty with just opening and closing line of workbook. You need to define the colour palette in between these two lines and there are three types of colour categories that can be defined. One of them is categorical which is defined using type = “Regular” and second one is sequential and it comes under the type = “ordered-sequential” and last comes diverging colour which is of the type = “ordered-diverging”. Also, kindly note that the colour should be defined in the HTML #RRGGB order.

384.What is a legend in Power BI?

Legends are part of Power BI visuals. They represent categories in a visual and are usually color-coded. In some visuals, we can add a category dimension as legend explicitly. One example can be in stacked bar/column charts, where the stacks represent different categories, and these categories are color-coded. These categories are represented in the legend. And every visual where the legend is applicable has a separate formatting section where we can specify its font and font size, whether we need it to be visible or not, or even its position(top, bottom, left, right)

385.What is Power BI-Embedded?

Power BI Embedded is an analytics solution, provided as a Microsoft Azure Service, a platform-as-a-service (PaaS), wherein the developers & ISVs (Individual Software Vendors) can easily embed their dashboards, Visuals & reports into an application for their customers.

Fully interactive reports & visuals can be embedded into the applications.

The great thing about Power BI Embedded is that your customers are not required to have knowledge about Power BI. Power BI gives you the capability to create an embedded application using two different methods.

The first method is using a Power BI Pro account & the second method is using the service principal.

The Power BI Pro account acts as the master account of your application. This account will help you generate the required embed tokens for providing access to your customers to view the shared Power BI dashboards and reports.

The service principal embeds Power BI Dashboards/Reports/Visuals into an application using an app-only token. Now this will allow you to generate the embed tokens for providing access to your shared application's Power BI dashboards and reports.

386.What is the Ribbon and what does it contain?

The Ribbon refers to the row of buttons and icons at the top of your worksheet. These include common tabs like Home, Insert, Page Layout, and Data.

You can customize the Ribbon and collapse or expand it using CTRL+F1. Some tabs only appear when you select a relevant item, such as a chart or table.

387.What is Excel BI Toolkit.

Excel BI toolkit is the other half of the Microsoft self-service BI (one half being Power BI). This toolkit is composed of Excel and a few more add-ins such as Power Query, Power Pivot, Power View, Power Map. All the tools in the Excel BI toolkit serve a special purpose like importing, modeling, preparing and visualizing data. Generally, the tools are used to create reports by consolidating data from multiple data sources and to model the datasets. Excel and other add-ins can be used independently or along with each other to optimize BI capabilities of the toolkit.

388.Where is the Power BI data stored?

All the data that we import in Power BI from different data sources get stored in either of the two tables in a data warehouse; Fact tables and Dimension tables. The fact tables are the central/main table of a start schema which contains all the measure (quantitative data) values. It has primary keys and all the dimension tables are linked to the fact table. The fact table is not usually normalized.

While a dimension table is a table in a database which contains all the attribute values (information about data) for the data stored in fact table. Every dimension table in a star schema is linked to a fact table.

389. What is analysis in tableau?

Tableau comes with inbuilt features to analyze the data plotted on a chart. We have various tools such as adding an average line to the chart which tableau calculates itself after we drop the tool on the chart. Some other features include clustering, percentages, forming bands of a particular range and various other tools to explore and inspect data. All these tools are available in analyze tab on each sheet used to create any chart. The features become visible only when they are applicable to the worksheet.

390.How to create sets in tableau?

Sets are custom fields used to compare and ask questions about a subset of data. For creating a set on dimension, right-click on a dimension in data pane and select create -> set. In general tab select the fields that will be considered for computing the set. Specify the conditions to create set in conditions tab and you also have the option to select top N members in dataset based on any field in the top tab. When a set is created it divides the measure into two parts namely in and out of the set based on the conditions applied by the user.

391.Why and how would you use a custom visual file?

A custom visual file is used when none of the pre existing visuals fit the business needs. Custom visual files are generally created by Developers which can be used in the same way as prepackaged files.

392. What are the various type of users who can use Power BI?

PowerBI can be used by anyone for their requirements but there is a particular group of users who are more likely to use it:

Report Consumers: They consume the reports based on a specific information they need

Report Analyst: Report Analysts need detailed data for their analysis from the reports

Self Service Data Analyst: They are more experienced business data users. They have an in-depth understanding of the data to work with.

Basic Data Analyst: They can build their own datasets and are experienced in PowerBI Service

Advanced Data Analyst: They know how to write SQL Queries and have hands-on experience on PowerBI. They have experience in Advanced PowerBI with DAX training and data modelling.

393.How do you apply a single format to all the sheets present in a workbook?

To apply the same format to all the sheets of a workbook, follow the given steps:

-Right-click on any sheet present in that workbook

-Then, click on the Select All Sheets option

-Format any of the sheets and you will see that the format has been applied to all the other sheets as well

394.Explain SUM and SUMIF functions.

SUM: The SUM function is used to calculate the sum of all the values that are specified as a parameter to it. The syntax of this function is as follows:

SUM(number1, number2, …)

SUMIF: This function is used to calculate the sum of values that comply with a given condition.

SYNTAX:

SUMIF(range, criteria, [sum_range])

where,

range specifies the range of cells to be evaluated

criteria provides the condition to be met

sum_range is optional and provides the actual cells to be summed up

395. What are the different data types of Tableau.

There are 7 data types in Tableau

-Boolean (True/False)

-Date (Individual Value)

-Date and Time

-Geography

-Text or String

-Decimal Number

-Whole Number

396.What is a dual-axis?

It is a function in Tableau that showcases two scales of two measures in a single graph. This is very similar to the function found on Microsoft Office products where a single graph has line and bar elements. In most cases, it has either two X or two Y axes.

A dual-axis is typically used to show trend lines and historical data. An example would be total revenue vs profit across 12 months.

397. What is the shortcut to add a filter to a table?

The filter mechanism is used when you want to display only specific data from the entire dataset. By doing so, there is no change being made to the data. The shortcut to add a filter to a table is Ctrl+Shift+L.

398. What are the wildcards available in Excel?

Wildcards only work with text data. Excel has three wildcards.

1. * (Asterisk)

This refers to any number of characters.

2.? (Question mark)

It represents one single character.

3.~ (Tilde)

It is used to identify a wildcard character (~, *, ?) in the text

399. What is live and extract in tableau?

Extract is a snapshot of the data optimized for aggregation. Extracts are loaded into the system and hence improve the performance of tableau. Whereas extracts won't help in situations where data is updated continuously because then we manually need to refresh the data for all the updates but using a live connection might slow the processing but will definitely update the data source itself. So, live connection should be used only when data is continuously updating otherwise extract file is preferred.

400.Differentiate between K-Means and KNN algorithms?

KNN algorithms is Supervised Learning where-as K-Means is Unsupervised Learning. With KNN, we predict the label of the unidentified element based on its nearest neighbour and further extend this approach for solving classification/regression-based problems.

K-Means is Unsupervised Learning, where we don't have any Labels present, in other words, no Target Variables and thus we try to cluster the data based upon their coordinates and try to establish the nature of the cluster based on the elements filtered for that cluster.

401.What is absolute cell reference in excel?

Absolute cell reference is a locked reference and ensures that the rows and columns will not change on copying the cell. We add a '$' symbol in front of the row and column number. When the cell is A2, both rows and columns remain unchanged on copying. But if it is $A2, it means that the column remains unchanged, but the row changes on copying. Similarly, for A$2, the column changes, but the row remains unchanged on copying.

402. List the most popular distribution curves along with scenarios where you will use them in an algorithm.

The most popular distribution curves are as follows- Bernoulli Distribution, Uniform Distribution, Binomial Distribution, Normal Distribution, Poisson Distribution, and Exponential Distribution.

Each of these distribution curves is used in various scenarios.

Bernoulli Distribution can be used to check if a team will win a championship or not, a newborn child is either male or female, you either pass an exam or not, etc.

Uniform distribution is a probability distribution that has a constant probability. Rolling a single dice is one example because it has a fixed number of outcomes.

Binomial distribution is a probability with only two possible outcomes, the prefix 'bi' means two or twice. An example of this would be a coin toss. The outcome will either be heads or tails.

Normal distribution describes how the values of a variable are distributed. It is typically a symmetric distribution where most of the observations cluster around the central peak. The values further away from the mean taper off equally in both directions. An example would be the height of students in a classroom.

Poisson distribution helps predict the probability of certain events happening when you know how often that event has occurred. It can be used by businessmen to make forecasts about the number of customers on certain days and allows them to adjust supply according to the demand.

Exponential distribution is concerned with the amount of time until a specific event occurs. For example, how long a car battery would last, in months.

403.Why is logistic regression a type of classification technique and not a regression? Name the function it is derived from?

Since the target column is categorical, it uses linear regression to create an odd function that is wrapped with a log function to use regression as a classifier. Hence, it is a type of classification technique and not a regression. It is derived from cost function.

404. What is OOB error and how does it occur?

For each bootstrap sample, there is one-third of data that was not used in the creation of the tree, i.e., it was out of the sample. This data is referred to as out of bag data. In order to get an unbiased measure of the accuracy of the model over test data, out of bag error is used. The out of bag data is passed for each tree is passed through that tree and the outputs are aggregated to give out of bag error. This percentage error is quite effective in estimating the error in the testing set and does not require further cross-validation.

405. State the limitations of Fixed Basis Function.

Linear separability in feature space doesn't imply linear separability in input space. So, Inputs are non-linearly transformed using vectors of basic functions with increased dimensionality. Limitations of Fixed basis functions are:

Non-Linear transformations cannot remove overlap between two classes but they can increase overlap.

Often it is not clear which basis functions are the best fit for a given task. So, learning the basic functions can be useful over using fixed basis functions.

If we want to use only fixed ones, we can use a lot of them and let the model figure out the best fit but that would lead to overfitting the model thereby making it unstable.

406. You are given a data set. The data set has missing values that spread along 1 standard deviation from the median. What percentage of data would remain unaffected? Why?

This question has enough hints for you to start thinking! Since the data is spread across the median, let's assume it's a normal distribution. We know, in a normal distribution, ~68% of the data lies in 1 standard deviation from mean (or mode, median), which leaves ~32% of the data unaffected. Therefore, ~32% of the data would remain unaffected by missing values.

407. What do you mean by convex hull?

Convex hull is represents to outer boundaries of two-level group of data point. Once is convex hull has to been created data-set value, we get maximum data-set value level of margin hyperplane (MMH), which attempts to create data set value greatest departure between two groups data set value, as a vertical bisector between two convex hulls data set value.

A convex hull of a set S is the intersection of all convex set of which S is a subset. We denote it by [S] the convex hull of S.

Example:

S= {x: |x|=1} implies [S]= {x: |x|< or =1}

S= {x: |x| > or= 1} implies [S]=En

408. You've built a random forest model with 10000 trees. You got delighted after getting training error as 0.00. But, the validation error is 34.23. What is going on? Haven't you trained your model perfectly?

The model has overfitted. Training error 0.00 means the classifier has mimicked the training data patterns to an extent, that they are not available in the unseen data. Hence, when this classifier was run on an unseen sample, it couldn't find those patterns and returned predictions with higher error.

In a random forest, it happens when we use a larger number of trees than necessary. Hence, to avoid this situation, we should tune the number of trees using cross-validation.

409. State one real life applications of convex hulls?

One applications convex hulls is to computation/construction of convex relaxations. Can say this is a way to find 'closest' convex problem to a non-convex problem one is attempting to solve.

410. How Are Weights Initialized in a Neural network?

There are two methods here: we can either initialize the weights to zero or assign them randomly.

Initializing all weights to 0: This makes your model similar to a linear model. All the neurons and every layer perform the same operation, giving the same output and making the deep net useless.

Initializing all weights randomly: Here, the weights are assigned randomly by initializing them very close to 0. It gives better accuracy to the model since every neuron performs different computations. This is the most commonly used method.

411. What are the variants of Gradient descent?

Stochastic Gradient Descent: We use only a single training example for calculation of gradient and

update parameters.

Batch Gradient Descent: We calculate the gradient for the whole dataset and perform the update at

each iteration.

Mini-batch Gradient Descent: It’s one of the most popular optimization algorithms. It’s a variant of

Stochastic Gradient Descent and here instead of single training example, mini-batch of samples is

used.

412. What are the feature selection methods used to select the right variables?

There are two main methods for feature selection:

Filter Methods

This involves:

- Linear discrimination analysis
- ANOVA
- Chi-Square

The best analogy for selecting features is "bad data in, bad answer out." When we're limiting or selecting the features, it's all about selecting the useful feature.

Wrapper Methods

This involves:

- Forward Selection: We test one feature at a time and keep adding them until we get a good fit

• Backward Selection: We test all the features and start removing them to see what works better

• Recursive Feature Elimination: Recursively looks through all the different features and how they pair together. Wrapper methods are very labor-intensive, and high-end computers are needed if a lot of data analysis is performed with the wrapper method.

413. What is joint sampling and separate sampling?

· Joint sampling is done when there are equal number of events and non-events. Not appropriate for imbalanced data

· Separate sampling is done for imbalanced data. For rare event, all observations are kept when target = 1 and only few observations are kept when target = 0.

414. Say your model is suffering from low bias and high variance !! Which algorithm you think will help you to tackle it? Why?

Low bias occurs when the model's predicted values are near to actual values. In other words, model becomes flexible enough to mimic the training data distribution. Here not to forget a flexible model has no generalization capabilities. It means, when this model is tested on an unseen data, it gives disappointing results.

In such situations, we can use Bagging Algorithm (like random forest) to tackle high variance problem. Bagging algorithms divides a data set into subsets made with repeated randomized sampling. Then these samples are used to generate a set of models using a single learning algorithm. Later model predictions are combined using voting (classification) or averaging (regression).

Also to overcome high variance one can:

1. Use Regularization technique, where higher model coefficients get penalized, hence lowering model complexity.

2. Use top n features from variable importance chart. Maybe, with all the variable in the data set, algorithm is having difficulty in finding meaningful signal

415. You are given a train data set having 2000 columns and 2 million rows. The data set is based on a classification problem. Your manager has asked you to reduce the dimension of this data so that model computation time can be reduced. Your machine has memory constraints? How you will deal with this?

1.Problem is because of Lower RAM - I will close all not required applications running on my machine using task manager/activity tracker - aiming to use most of memory.

2. I will try random sampling on data set. By this I mean, we can create a smaller data set, let's say having 2000 variables and 300000 rows and do computations.

3. I will try to reduce dimensionality, I will separate numerical and categorical variables and remove correlated variables.

4. For numerical variables, I will use correlation. For categorical variables, I will use chi-square test.

5. I will also try to use PCA and pick components which can explain maximum variance in given data set

6. I guess building a Linear Model using Stochastic Gradient Descent will be helpful

7. If possible I will try to apply business understanding to estimate which all predictors can impact response variable (domain knowledge will be required). Cons can be - as this is an intuitive approach failing to identify useful predictors might result in significant loss of information

416. Why are Validation and Test Datasets Needed?

Data is split into three different categories while creating a model:

- Training dataset: Training dataset is used for building a model and adjusting its variables. The correctness of the model built on the training dataset cannot be relied on as the model might give incorrect outputs after being fed new inputs.

- Validation dataset: Validation dataset is used to look into a model's response. After this, the hyperparameters on the basis of the response of the model(on the validation dataset data) are tuned.When a model's response is evaluated by using the validation dataset, the model is indirectly trained with the validation set. This may lead to the overfitting of the model to specific data. So, this model will not be strong enough to give the desired response to real-world data.

- Test dataset: Test dataset is the subset of the actual dataset, which is not yet used to train the model. The model is unaware of this dataset. So, by using the test dataset, the response of the created model can be computed on hidden data. The model's performance is tested on the basis of the test dataset.Note: The model is always exposed to the test dataset after tuning the hyperparameters on top of the validation dataset.

As we know, the evaluation of the model on the basis of the validation dataset would not be enough. Thus, the test dataset is used for computing the efficiency of the model.

417. What are Gaussian Mixture Model? Where they are used?

Probabilistic model for representing normally distributed subpopulations within an overall population. Mixture models in general don't require knowing which subpopulation a data point belongs to, allowing the model to learn the subpopulations automatically. Since subpopulation assignment is not known, this constitutes a form of unsupervised learning. These models have been used for feature extraction from speech data, and have also been used extensively in object tracking of multiple objects, where number of mixture components and their means predict object locations at each frame in a video sequence. change inplace of 1st.

418. What are Hard-Margin and Soft-Margin SVMs?

Hard-Margin SVMs have linearly separable training data. No data points are allowed in the margin areas. This type of linear classification is known as Hard margin classification.

Soft-Margin SVMs have training data that are not linearly separable. Margin violation means choosing a hyperplane, which can allow some data points to stay either in between the margin area or on the incorrect side of the hyperplane.

419. What is the empirical rule?

In statistics, the empirical rule states that every piece of data in a normal distribution lies within three standard deviations of the mean. It is also known as the 68–95–99.7 rule. According to the empirical rule, the percentage of values that lie in a normal distribution follow the 68%, 95%, and 99.7% rule. In other words, 68% of values will fall within one standard deviation of the mean, 95%

will fall within two standard deviations, and 99.75 will fall within three standard deviations of the mean.

420. What is the left-skewed distribution and the right-skewed distribution?

In the left-skewed distribution, the left tail is longer than the right side.

Mean < median < mode

In the right-skewed distribution, the right tail is longer. It is also known as positive-skew distribution.

Mode < median < mean

421. What are Different Kernels in SVM?

There are six types of kernels in SVM:

Linear kernel - used when data is linearly separable.

Polynomial kernel - When you have discrete data that has no natural notion of smoothness.

Radial basis kernel - Create a decision boundary able to do a much better job of separating two classes than the linear kernel.

Sigmoid kernel - used as an activation function for neural networks.

422. According to you what are some advantages in using a CNN over a DNN (dense neural network) for an image classification task?

Ans: Both models can capture relationship between close pixels but CNNs have some useful properties like: translation invariant — exact location of pixel is irrelevant for filter less likely to overfit — typical number of parameters in a CNN is much smaller than that of a DNN gives better understanding of model — one can look at filters, weights and visualize what network learned hierarchical nature — learns patterns in by describing complex patterns using simpler ones.

423. What are your intuitions about pruning in decision tree?

Ans: Pruning a decision tree helps to prevent overfitting the training data so that our model generalizes well to unseen data. Pruning a decision tree means to remove a subtree that is redundant and not a useful split and replace it with a leaf node. Decision tree pruning can be divided into two types: pre-pruning and post-pruning.

Pre-pruning, also known as Early Stopping Rule, is the method where the subtree construction is halted at a particular node after evaluation of some measure. These measures can be the Gini Impurity or the Information Gain. In pre-pruning, we evaluate the pruning condition based on the above measures at each node.

post-pruning means to prune after the tree is built. You grow the tree entirely using your decision tree algorithm and then you prune the subtrees in the tree in a bottom-up fashion. You start from the bottom decision node and, based on measures such as Gini Impurity or Information Gain, you decide whether to keep this decision node or replace it with a leaf node.

424. Give some intuition about Epoch, Batch, and Iteration in deep learning how they are different?

1. Epoch - Represents one iteration over entire dataset (everything put into training model)

2. Batch - Refers to when we cannot pass entire dataset into neural network at once, so we divide dataset into several batches

3. Iteration - if we have 100,000 images as data and a batch size of 400. then an epoch should run 250 iterations (100,000 divided by 400)

425. Explain ways to train hyperparameters in a neural network?

Hyperparameters in a neural network can be trained using four components:

1. Batch size: Indicates size of input data

2. Epochs: Denotes the number of times the training data is visible to neural network to train

3. Momentum: Used to get an idea of next steps that occur with data being executed

4. Learning rate: Represents time required for network to update parameters and learn

426. What happens if the learning rate is set too high or too low?

If the learning rate is too low, your model will train very slowly as minimal updates are made to the weights through each iteration. Thus, it would take many updates before reaching the minimum point.

If the learning rate is set too high, this causes undesirable divergent behavior to the loss function due to drastic updates in weights, and it may fail to converge.

427. Explain Crude working of LSTM? ==> How Does an LSTM Network Work?

Long-Short-Term Memory (LSTM) is a special kind of recurrent neural network RNN capable of learning long-term dependencies, remembering information for long periods as its default behaviour. Crude steps in an LSTM network are as:

1: Network decides what to forget and what to remember

2: It selectively updates cell state values

3: Network decides what part of current state makes it to output

428. Explain what is Swish Function?

Swish is an activation function proposed by Google which is an alternative to ReLU activation function.

It is represented as: f(x) = x * sigmoid(x)

Swish function works better than ReLU for a variety of deeper models.

The derivative of Swist can be written as: y' = y + sigmoid(x) * (1 - y)

429. What is Information gain?

Information gain is the difference of entropy of the parent node and the weighted entropies of the child nodes.

Information Gain = Entropy before splitting - Entropy after splitting

Suppose we have multi features in a dataset, so we will choose that feature as a root node which will have maximum Information Gain.

Information gain is used for determining the best features/attributes that render maximum information about a class. It follows the concept of entropy while aiming at decreasing the level of entropy, beginning from the root node to the leaf nodes.

430. When accuracy should not be used as a parameter to measure the performance of Classification model?

There are 2 major situations when you will not use accuracy as a parameter for classification performance measures-

A) When you have severely imbalanced dataset. If suppose you have a dataset with 90% as positive class and 10% as negative class, then even a dumb model can classify all the datapoints as positive class and attain the accuracy as 90%, which is not sensible.

B) Since accuracy does not considers probability values so if I have a threshold of 0.5 and my model M1 gives the probability value for a "x" as 0.9 and my second model gives the probability value of 0.55, both of them will be labelled as positive class but we know the performance of M1 is much better than M2 when you see the actual probabilities.

431. Why do we use the summary function?

Summary functions are used to give the summary of all the numeric values in a dataframe. Eg. The describe() function can be used to provide the summary of all the data values given to it.

column_name.describe() will give the following values of all the numeric data in the column-

Count

Mean

Std-Standard deviation

Min-Minimum

25%

50%

75%

max-Maximum

432. How are Entropy and Gini index related?

Both Entropy and Gini index are parabolic in nature when it's plotted by the probability. However the maximum value of my Gini Impurity is 0.5 while that of entropy is 1 which only occurs when my classes are Equi-probable. Both of them will tend to zero values when one of the classes are dominant.

433. Why is KNN algorithm called Lazy Learner?

When it gets the training data, it does not learn and make a model, it just stores the data. It does not derive any discriminative function from the training data. It uses the training data when it actually needs to do some prediction. So, KNN does not immediately learn a model, but delays the learning, that is why it is called lazy learner.

434. What is stratify in Train_test_split?

Stratification means that the train_test_split method returns training and test subsets that have the same proportions of class labels as the input dataset. So if my input data has 60% 0's and 40% 1's as my class label, then my train and test dataset will also have similar proportions.

435. Explain the difference between lists and tuples.

Both lists and tuples are made up of elements, which are values of any Python data type.

Lists are mutable, while tuples are immutable.

Lists are created using Square brackets (for example, my_list = [a, b, c]), while tuples are in parentheses (for example, my_tuple = (a, b, c)).

436. What is the difference between classification and regression?

Classification is about predicting a label and regression is about predicting a quantity. In classification we predict a discrete class label . Examples of Classification algorithm are Decision Tree, KNN, Naive Bayes, Logistic regression etc.

In regression we predict a continuous quantity output. Examples of regression models are Linear regression, Random Forest regressor etc.

437. Should Outliers be treated for Logistic regression?

Logistics regression uses Sigmoid function to sqashing the Best fit line which restricts the value of yi or the target variable between 0 and 1. Hence it's not much impacted by outliers.

However when there is huge number of outliers so it will affect my logistics regression as well as then my Sigmoid function will fit on dataset in such a way that it would try to compensate for the maximum outliers thus affecting the best fit line.

438. How would you choose the best feature for the root node as per the entropy criterion?

We will always choose that feature which gives us the maximum information gain once it splits. Information gain is the difference of entropy of the feature as root node and the summation of the entropy of the child nodes

439. Why do we need to have regularisation in Logistics Regression?

Since the optimization function, which has weight vectors , features and target variables, can only get minimum value of zero when weight vectors will tend to infinity. This will be a perfect case of overfitting since my optimization function goes to the minimum possible value which is zero. Hence we need to use regularisation to avoid overfitting.

440. What are independent events?

In probability, we say two events are independent if knowing one event occurred doesn't change the probability of the other event. For example, the probability that a fair coin shows "heads" after being flipped is 1 / 2.

441. What is the interquartile range (IQR)?

The IQR is the difference between the third quartile (Q3) and the first quartile (Q1) of a set of data. Specifically, IQR = Q3–Q1.

442.What is the difference between a univariate analysis and a bivariate analysis?

The univariate analysis involves analyzing the distribution of a single variable. Bivariate analysis, on the other hand, considers the relationship between two distinct variables.For example, calculating the average amount of coffee consumed by a certain population would be a univariate analysis. In contrast, understanding the relationship between coffee consumption and age, gender, or time of the year would be examples of bivariate analysis.

443.What is the difference between the central limit theorem and the law of large numbers?

The central limit theorem states that the sampling distribution of the sample means approaches a normal distribution as the sample size increases, regardless of the shape of the original distribution.

The law of large numbers states that as the number of trials increases, the average of the results will become closer to the expected value. For example, the proportion of heads in 1,000 fair coin tosses is closer to 0.5 than it is in 10 tosses.

444.Does feature scaling always result in improvement in model accuracy?

Feature scaling is not necessary when using tree-based algorithms e.g. decision trees, random forest, and gradient boosting as they are invariant to the scale of the features. Decision trees split each node using one feature at a time and as a result, they are unaffected by the other features in the dataset.For gradient descent and distance-based algorithms, however, feature scaling can result in improvement in model accuracy.

445. How is random forest better than decision tree?

Random forest is an ensemble method that takes many weak decision trees to make a strong learner. Random forests are more accurate, more robust, and less prone to overfitting.

446. Explain what the bootstrap sampling method is and give an example of when it's used.

Bootstrap sampling method is a resampling method that uses random sampling with replacement.

It's an essential part of the random forest algorithm, as well as other ensemble learning algorithms.

447. What happens if the learning rate is set too high or too low?

If the learning rate is too low, your model will train very slowly as minimal updates are made to the weights through each iteration. Thus, it would take many updates before reaching the minimum point.

If the learning rate is set too high, this causes undesirable divergent behavior to the loss function due to drastic updates in weights, and it may fail to converge.

448. What is the difference between supervised learning and unsupervised learning? Give concrete examples.

Supervised learning involves learning a function that maps an input to an output.

For example, if I had a dataset with two variables, age (input) and height (output), I could implement a supervised learning model to predict the height of a person based on their age.

Unlike supervised learning, unsupervised learning is used to draw inferences and find patterns from input data without references to labeled outcomes. A common use of unsupervised learning is grouping customers by purchasing behavior to find target markets.

449.How do you assess the statistical significance of an insight?

You would perform hypothesis testing to determine statistical significance. First, you would state the null hypothesis and alternative hypothesis.

Second, you would calculate the p-value, the probability of obtaining the observed results of a test assuming that the null hypothesis is true. Last, you would set the level of the significance (alpha) and if the p-value is less than the alpha, you would reject the null — in other words, the result is statistically significant.

450. What is the Law of Large Numbers?

The Law of Large Numbers is a theory that states that as the number of trials increases, the average of the result will become closer to the expected value.

Eg. flipping heads from fair coin 100,000 times should be closer to 0.5 than 100 times.

451.If a Company says that they want to double the number of ads in Newsfeed, how would you figure out if this is a good idea or not?

You can perform an A/B test by splitting the users into two groups: a control group with the normal number of ads and a test group with double the number of ads. Then you would choose the metric to define what a “good idea” is. For example, we can say that the null hypothesis is that doubling the number of ads will reduce the time spent on Facebook and the alternative hypothesis is that doubling the number of ads won’t have any impact on the time spent on Facebook. However, you can choose a different metric like the number of active users or the churn rate. Then you would conduct the test and determine the statistical significance of the test to reject or not reject the null.

452.What is Selection Bias and what are various types?

Selection bias takes place when data is chosen in a way that is not reflective of real-world data distribution. This happens because proper randomization is not achieved when collecting data.

Types of selection bias -

• Sampling bias: occurs when randomization is not properly achieved during data collection. To give an example, imagine that there are 10 people in a room and you ask if they prefer grapes or bananas. If you only surveyed the three females and concluded that the majority of people like grapes, you’d have demonstrated sampling bias.

•Convergence bias: occurs when data is not selected in a representative manner. e.g. when you collect data by only surveying customers who purchased your product and not another half, your dataset does not represent the group of people who did not purchase your product.

•Participation bias: occurs when the data is unrepresentative due to participations gaps in the data collection process.

453.Can you compare the validation test with the test set?

Validation Dataset: The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters. The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration.

Test Dataset: The sample of data used to provide an unbiased evaluation of a final model fit on the training dataset.

454.What is the aim of conducting A/B Testing?

Creating a website or email marketing campaign is just the first step in marketing. Once you have a website, you'll want to know if it helps or hinders sales.

A/B testing lets you know what words, phrases, images, videos, testimonials, and other elements work best. Even the simplest changes can impact conversion rates.

To understand how A/B testing works, let's take a look at an example.

Imagine you have two different designs for a landing page—and you want to know which one will perform better.

After you create your designs, you give one landing to one group and you send the other version to the second group. Then you see how each landing page performs in metrics such as traffic, clicks, or conversions. To understand how A/B testing works, let's take a look at an example.

Imagine you have two different designs for a landing page—and you want to know which one will perform better.

After you create your designs, you give one landing to one group and you send the other version to the second group. Then you see how each landing page performs in metrics such as traffic, clicks, or conversions.

455.What is Hierarchical Clustering and what are it's 2 types?

Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters.

Hierarchical Clustering is of two types:

•Agglomerative: This is a "bottom-up" approach where each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

• Divisive: This is a "top-down" approach where all observations start in one big cluster, and splits are performed recursively as one moves down the hierarchy.

456. What is a logistic function? What is the range of values of a logistic function?

f(z) = 1/(1+e -z)

The values of a logistic function will range from 0 to 1. The values of Z will vary from -infinity to +infinity.

457. What is the difference between R square and adjusted R square?

R square and adjusted R square values are used for model validation in case of linear regression. R square indicates the variation of all the independent variables on the dependent variable. i.e. it considers all the independent variable to explain the variation. In the case of Adjusted R squared, it considers only significant variables(P values less than 0.05) to indicate the percentage of variation in the model.

Thus Adjusted R2 is always lesser then R2.

458. What is stratify in Train_test_split?

Stratification means that the train_test_split method returns training and test subsets that have the same proportions of class labels as the input dataset. So if my input data has 60% 0's and 40% 1's as my class label, then my train and test dataset will also have the similar proportions.

459. What is Backpropagation in Artificial Neuron Network?

Backpropagation is the method of fine-tuning the weights of a neural network based on the error rate obtained in the previous epoch (i.e., iteration). Proper tuning of the weights allows you to reduce error rates and make the model reliable by increasing its generalization.

460. Why is dimension reduction important?

This is important mainly in the case when you want to reduce variance in your model (overfitting).

There are four advantages of dimensionality reduction :

*It reduces the time and storage space required

*Removal of multi-collinearity improves the interpretation of the parameters of the machine learning model

*It becomes easier to visualize the data when reduced to very low dimensions such as 2D or 3D

*It avoids the curse of dimensionality

461. What are the drawbacks of a linear model?

There are a couple of drawbacks of a linear model:

A linear model holds some strong assumptions that may not be true in application. It assumes a linear relationship, multivariate normality, no or little multicollinearity, no auto-correlation, and homoscedasticity

A linear model can’t be used for discrete or binary outcomes.

You can’t vary the model flexibility of a linear model.

462. Is mean imputation of missing data acceptable practice? Why or why not

Mean imputation is the practice of replacing null values in a data set with the mean of the data.

Mean imputation is generally bad practice because it doesn’t take into account feature correlation. For example, imagine we have a table showing age and fitness score and imagine that an eighty-year-old has a missing fitness score. If we took the average fitness score from an age range of 15 to 80, then the eighty-year-old will appear to have a much higher fitness score that he actually should.

Second, mean imputation reduces the variance of the data and increases bias in our data. This leads to a less accurate model and a narrower confidence interval due to a smaller variance.

463. Give examples of data that does not have a Gaussian distribution, nor log-normal.

Any type of categorical data won't have a gaussian distribution or lognormal distribution.

Exponential distributions — eg. the amount of time that a car battery lasts or the amount of time until an earthquake occurs.

464. The covid spread rate in India fell last year to 99 from 115 the year before. Is this reported change really noteworthy?

Since this is a Poisson distribution question, mean = lambda = variance, which also means that standard deviation = square root of the mean

a 95% confidence interval implies a z score of 1.96

one standard deviation = sqrt(115) = 10.724

Therefore the confidence interval = 115+/- 21.45 = [93.55, 136.45]. Since 99 is within this confidence interval, we can assume that this change is not very noteworthy.

465.What are the conditions for Overfitting and Underfitting?

- In Overfitting the model performs well for the training data, but for any new data it fails to provide output. For Underfitting the model is very simple and not able to identify the correct relationship. Following are the bias and variance conditions.

- Overfitting – Low bias and High Variance results in the overfitted model. The decision tree is more prone to Overfitting.

- Underfitting – High bias and Low Variance. Such a model doesn't perform well on test data also. For example – Linear Regression is more prone to Underfitting.

466. Which models are more prone to Overfitting?

Complex models, like the Random Forest, Neural Networks, and XGBoost are more prone to overfitting. Simpler models, like linear regression, can overfit too – this typically happens when there are more features than the number of instances in the training data.

467. When does feature scaling should be done?

We need to perform Feature Scaling when we are dealing with Gradient Descent Based algorithms (Linear and Logistic Regression, Neural Network) and Distance-based algorithms (KNN, K-means, SVM) as these are very sensitive to the range of the data points.

468.Which model is not affected by outliers?

Tree-based models are generally not affected by outliers, while regression-based models are.

469. What does the word 'Naive' mean in Naive Bayes?

Naive Bayes is a Data Science algorithm. It has the word 'Bayes' in it because it is based on the Bayes theorem, which deals with the probability of an event occurring given that another event has already occurred.

It has 'naive' in it because it makes the assumption that each variable in the dataset is independent of the other. This kind of assumption is unrealistic for real-world data. However, even with this assumption, it is very useful for solving a range of complicated problems, e.g., spam email classification, etc.

470. You are given a data set consisting of variables with more than 30 percent missing values. How will you deal with them?

If the data set is large, we can just simply remove the rows with missing data values. It is the quickest way, we use the rest of the data to predict the values.

For smaller data sets, we can substitute missing values with values like mean, mode, forward or backward fill. There are different ways to do so, such as df.mean(), df.fillna(mean).

471.What do you understand about true positive rate and false-positive rate?

The True Positive Rate (TPR) defines the probability that an actual positive will turn out to be positive.

The True Positive Rate (TPR) is calculated by taking the ratio of the [True Positives (TP)] and [True Positive (TP) & False Negatives (FN)].

Formula: TPR=TP/TP+FN

The False Positive Rate (FPR) defines the probability that an actual negative result will be shown as a positive one i.e the probability that a model will generate a false alarm.

The False Positive Rate (FPR) is calculated by taking the ratio of the [False Positives (FP)] and [True Negatives (TN) & False Positives(FP)].

Formula: FPR=FP/TN+FP

472. What are the feature selection methods you use to select important variables?

1. Filter Methods

There are various filter methods such as the Chi-Square test, Fisher's Score method, Correlation Coefficient, Variance Threshold, Mean Absolute Difference (MAD) method, Dispersion Ratios, etc.

2. Wrapper Methods

There are three types of wrapper methods, they are:

Forward Selection: Here, one feature is tested at a time and new features are added until a good fit is obtained.

Backward Selection: Here, all the features are tested and the non-fitting ones are eliminated one by one to see while checking which works better.

Recursive Feature Elimination: The features are recursively checked and evaluated how well they perform.

3. Embedded Method

Examples of embedded methods: LASSO Regularization (L1), Random Forest Importance.

473. What is a lambda expression in Python?

With the help of lambda expression, you can create an anonymous function. Unlike conventional functions, lambda functions occupy a single line of code. The basic syntax of a lambda function is –

lambda arguments: expression

Eg

```
x = lambda a : a * 4

print(x(5))
```

Output of 20.

474. Write the code to find the length of a string without using the string functions.

```
string=raw_input("Enter string:")

 count=0

for i in string:

     count=count+1

     print("Length of the string is:")

     print(count)
```

475. How to find the positions of numbers that are multiples of 4 from a series in python?

For finding the multples of 4, we will use the argwhere() function. First, we will create a list of 10 numbers –

```
s1 = pd.Series([1, 2, 3, 4, 5, 6, 7, 8, 9, 10]) np.argwhere(s1 % 4==0)
```

Output > [3], [7]

476. What is the difference between recall and precision?

While calculating the Precision of a model, we should consider both Positive as well as Negative samples that are classified.

While calculating the Recall of a model, we only need all positive samples while all negative samples will be neglected.

Hence Precision quantifies the number of positive class predictions that actually belong to the positive class. Recall quantifies the number of positive class predictions made out of all positive examples in the dataset.

477. How will you get second highest salary of an employee emp from employee_table?

```
SELECT TOP 1 salary FROM( SELECT TOP 2 salary FROM employee_table ORDER BY salary DESC) AS
emp ORDER BY salary ASC;
```

478. What is the central limit theorem?

The central limit theorem states that the distribution of an average will tend to be Normal as the sample size increases, regardless of the distribution from which the average is taken except when the moments of the parent distribution do not exist.

479. Difference between Normalisation and Standardization?

Both Normalisation and Standardization are methods of Features Conversion. However, the methods are different in terms of the conversions. The data after Normalisation scales in the range of 0-1. While in case of Standardization the data is scaled such that it means comes out to be 0.

480. What relationships exist between a logistic regression's coefficient and the Odds Ratio?

The coefficients and the odds ratios then represent the effect of each independent variable controlling for all of the other independent variables in the model and each coefficient can be tested for significance.

481. What's the relationship between Principal Component Analysis (PCA) and Linear & Quadratic Discriminant Analysis (LDA & QDA)

LDA focuses on finding a feature subspace that maximizes the separability between the groups. While Principal component analysis is an unsupervised Dimensionality reduction technique, it ignores the class label. PCA focuses on capturing the direction of maximum variation in the data set.The PC1 the first principal component formed by PCA will account for maximum variation in the data.PC2 does the second-best job in capturing maximum variation and so on.

The LD1 the first new axes created by Linear Discriminant Analysis will account for capturing most variation between the groups or categories and then comes LD2 and so on.

482. How to ensure that your model is not overfitting?

To ensure that your model is not overfitting are:

• Keep the model simpler: remove some of the noise in the training data.

• Use cross-validation techniques such as k-folds cross-validation.

• Use regularization techniques such as LASSO.

484. Forward and Backward selection? It's working?

Forward Selection chooses a subset of the predictor variables for the final model. Unlike forward stepwise selection, it begins with the full least squares model containing all p predictors, and then iteratively removes the least useful predictor, one-at-a-time. forward selection starts with a null model (with no predictors) and proceeds to add variables one at a time, and so unlike backward selection, it DOES NOT have to consider the full model (which includes all the predictors).

485. Fix multicollinearity?

To remove multicollinearities, we can do two things.

1. We can create new features

2. remove those features from our data.

486. Difference between Gradient boosting and random forest?

The two main differences are: How trees are built: random forests builds each tree independently while gradient boosting builds one tree at a time. Combining results: random forests combine results at the end of the process (by averaging or "majority rules") while gradient boosting combines results along the way.

487. What are Support Vectors in SVM?

A Support Vector Machine (SVM) is an algorithm that tries to fit a line (or plane or hyperplane) between the different classes that maximizes the distance from the line to the points of the classes.

In this way, it tries to find a robust separation between the classes. The Support Vectors are the points of the edge of the dividing hyperplane.

488. What is Bias in Machine Learning?

Bias in data tells us there is inconsistency in data. The inconsistency may occur for several reasons which are not mutually exclusive.

For example, a tech giant like Amazon to speed the hiring process they build one engine where they are going to give 100 resumes, it will spit out the top five, and hire those.

When the company realized the software was not producing gender-neutral results it was tweaked to remove this bias.

489. Explain Correlation and Covariance?

Covariance signifies the direction of the linear relationship between two variables, whereas correlation indicates both the direction and strength of the linear relationship between variables.

490. What do you think the distribution of time spent per day on Facebook looks like? What metric would you use to describe the distribution.

In terms of the distribution of time spent per day on Facebook (FB), one can imagine there may be two groups of people on Facebook:

1. People who scroll quickly through their feed and don’t spend too much time on FB.

2. People who spend a large amount of their social media time on FB.

Based on this, we make claim about the distribution of time spent on FB. The metrics to describe our distribution can be

1) Centre (mean, median, mode)

2) Spread (standard deviation, inter quartile range

3) Shape (skewness, kurtosis, uni or bimodal)

4) Outliers (Do they exist?)

We can give you a sample answer for your interview: –

If we assume that a person is visiting Facebook page, there is a probability(p) that after one unit of time(t) has passed that she will leave the page.

With a probability of p her visit will be limited to 1 unit of time. With a probability of (1−p)p her visit will be limited to 2 units of time. With a probability of (1−p)2p her visit will be limited to 3 units of time and so on. The probability mass function of this distribution is therefore (1−p)tp, and hence we can say this a geometric distribution.

491.What is the difference between WHERE and HAVING in SQL?

WHERE Clause is used to filter the records from the table based on the specified condition. HAVING Clause is used to filter records from the groups based on the specified condition

WHERE Clause implements in row operations. HAVING Clause implements in column operation.

WHERE Clause can be used without GROUP BY clause.

HAVING Clause cannot be used without GROUP BY clause.

WHERE Clause is used before GROUP BY clause. HAVING Clause is used after GROUP BY clause.

WHERE Clause is used with single row functions like UPPER, LOWER etc. HAVING Clause is used with multiple row functions like SUM, COUNT etc.

492. What is a trigger in MySQL? How can we use it?

A trigger in MySQL is a set of SQL statements that reside in a system catalog. It is a special type of stored procedure that is invoked automatically in response to an event. Each trigger is associated with a table, which is activated on any DML statement such as INSERT, UPDATE, or DELETE.

A trigger is called a special procedure because it cannot be called directly like a stored procedure. The main difference between the trigger and procedure is that a trigger is called automatically when a data modification event is made against a table. In contrast, a stored procedure must be called explicitly.

Generally, triggers are of two types according to the SQL standard: row-level triggers and statement-level triggers.

Row-Level Trigger: It is a trigger, which is activated for each row by a triggering statement such as insert, update, or delete. For example, if a table has inserted, updated, or deleted multiple rows, the row trigger is fired automatically for each row affected by the insert, update or delete statement.

Statement-Level Trigger: It is a trigger, which is fired once for each event that occurs on a table regardless of how many rows are inserted, updated, or deleted.

For creating a new trigger, we need to use the CREATE TRIGGER statement. Its syntax is as follows:

```
CREATE TRIGGER trigger_name trigger_time trigger_event

ON table_name

FOR EACH ROW

BEGIN

...

END;
```

Trigger_name is the name of the trigger which must be put after the CREATE TRIGGER statement.

Trigger_time is the time of trigger activation and it can be BEFORE or AFTER. We must have to specify the activation time while defining a trigger.

Trigger_event can be INSERT, UPDATE, or DELETE. This event causes the trigger to be invoked. A trigger only can be invoked by one event.

Table_name is the name of the table. Actually, a trigger is always associated with a specific table.

BEGIN…END is the block in which we will define the logic for the trigger.

493.What is the difference between Skewness and kurtosis?

The characteristic of a frequency distribution that ascertains its symmetry about the mean is called skewness. On the other hand, Kurtosis means the relative pointedness of the standard bell curve, defined by the frequency distribution.

Skewness is characteristic of the deviation from the mean, to be greater on one side than the other, i.e. attribute of the distribution having one tail heavier than the other. Skewness is used to indicate the shape of the distribution of data. Conversely, kurtosis is a measure to indicate the flatness or peakedness of the frequency distribution curve and measures the tails or outliers of the distribution.

Skewness is an indicator of lack of symmetry, i.e., both left and right sides of the curve are unequal, with respect to the central point. As against this, kurtosis is a measure of data, that is either peaked or flat, with respect to the probability distribution.

Skewness shows how much and in which direction, the values deviate from the mean. In contrast, kurtosis explain how tall and sharp the central peak is.

In a skewed distribution, the curve is extended to either left or right side. So, when the plot is extended towards the right side more, it denotes positive skewness, wherein mode < median < mean. On the other hand, when the plot is stretched more towards the left direction, then it is called as negative skewness and so, mean < median < mode. Positive kurtosis represents that the distribution is more peaked than the normal distribution, whereas negative kurtosis shows that the distribution is less peaked than the normal distribution.

494.How would you build the algorithm for type-ahead search for Netflix?

We want a recommendation algorithm sounds like RNN which is a recurrent neural network which is not easy to setup at all, but we can try building recommendation with much simpler approach that is with a simple prefix matching algorithm and we can certainly go into expanding it on until we have something that will be on par with RNN.

We will use Lookup in this database table/ prefix table. This prefix table starts with an input string and that is your prefix and it will output the suggested string or suffixes. Example: what does “hello” prefix to and its suffixes to the model.

Scoping is very important by doing fuzzy matching, context matching like what if you were using a different language. So, if you are trying to input “big” that could output any number of suffixes example: big shot or big sky or the big year etc.

In existing search corpus of billions of searches what proportion of the time do people writing the big actually click on the big shot and what proportion of the time do they output the big sky, so you can just have a simple thing that has every possible search prefix that has ever been typed on Netflix and output that to the most common thing that they clicked on. Boom! That’s your prefix matching a recommendation algorithm for type ahead search.

Context matching is also important here, if u have string input and a user profile with various number of features into a string output. You can convert user profile into a K means clustering we can output this into either John Stamos’s fan and not John Stamos’s fan, so if you are John Stamos’s fan and if you type the big, every time Netflix is going to recommend the big shot or else other way round. Also user profile can be set to right dimensionality.

495.What is central tendency ?

The central tendency is a descriptive description of a dataset represented by a single value that represents the data distribution's center. It does not provide information on individual values in the dataset, but it does provide a thorough overview of the entire dataset. The following measurements can be used to describe the central tendency of a dataset in general: Mean: Represents the sum of all values in a dataset divided by total number of values in the dataset. The median is the midpoint value in an ascending-ordered dataset. The most often occurring value in a dataset is defined by the mode.

496. Which central tendency method is used If there exists any outliers ?

Because the value of the mean might be skewed by outliers, the median is frequently recommended in these instances. It will, however, depend on how powerful the outliers are. Using the mean as the measure of central tendency is usually preferred if they do not significantly alter the mean.

497. Explain Central limit theorem ?

The central limit theorem (CLT) asserts that as the sample size grows higher, the distribution of a sample variable approaches a normal distribution (i.e., a "bell curve"), assuming that all samples are identical in size and independent of the population's actual distribution shape.

498. What is Chi-Square test ?

A hypothesis testing method is the Chi-square test. Checking if observed frequencies in one or more categories match expected frequencies is one of two frequent Chi-square tests. The Chi-square test is used to determine whether your data is as expected. The test's core premise is to compare the observed values in your data to the expected values that you would see if the null hypothesis is true. The Chi-square goodness of fit test and the Chi-square test of independence are two regularly used Chi-square tests. Both tests use variables to categorize your data into groups.

500.Explain ROC curve.

A Receiver Operator Characteristic (ROC) curve is a graphical plot used to show the diagnostic ability of binary classifiers.A ROC curve is constructed by plotting the true positive rate (TPR) against the false positive rate (FPR). The true positive rate is the proportion of observations that were correctly predicted to be positive out of all positive observations (TP/(TP + FN)). Similarly, the false positive rate is the proportion of observations that are incorrectly predicted to be positive out of all negative observations (FP/(TN + FP)).

501. What is learning rate in gradient descent?

A. Gradient descent is the popular optimization algorithm used in machine learning to estimate the model parameters. Learning rate, generally represented by the symbol ‘α’, is a hyper-parameter used to control the rate at which an algorithm updates the parameter estimates or learns the values of the parameters.

502. What do you understand by Recall and Precision?

Precision is defined as the fraction of relevant instances among all retrieved instances. Recall, sometimes referred to as ‘sensitivity, is the fraction of retrieved instances among all relevant instances. A perfect classifier has precision and recall both equal to 1.

503. What is Lasso Regression and why do we use it for?

The lasso regression allows you to shrink or regularize these coefficients to avoid overfitting and make them work better on different datasets. This type of regression is used when the dataset shows high multicollinearity or when you want to automate variable elimination and feature selection.

504. How do you deal with overfitting in Logistic regression?

One of the ways to combat over-fitting is to increase the training data size. Another way to combat over-fitting is to perform early stopping. A way to combat over-fitting is through regularization. Regularization techniques can be viewed as imposing certain prior distribution on the model parameters.

505. What is Multicollinearity and how to deal with it?

Multicollinearity occurs when independent variables in a regression model are correlated. This correlation is a problem because independent variables should be independent. If the degree of correlation between variables is high enough, it can cause problems when you fit the model and interpret the results. Remove some of the highly correlated independent variables. Linearly combine the independent variables, such as adding them together. Perform an analysis designed for highly correlated variables, such as principal components analysis or partial least squares regression. LASSO and Ridge regression are advanced forms of regression analysis that can handle multicollinearity.

506. What is meant by 'curse of dimensionality'?

Curse of Dimensionality refers to a set of problems that arise when working with high-dimensional data. The dimension of a dataset corresponds to the number of attributes/features that exist in a dataset. A dataset with a large number of attributes, generally of the order of a hundred or more, is referred to as high dimensional data. Some of the difficulties that come with high dimensional data manifest during analyzing or visualizing the data to identify patterns, and some manifest while training machine learning models. The difficulties related to training machine learning models due to high dimensional data is referred to as 'Curse of Dimensionality'.

507. What is Gradient Descent?

Gradient descent is an optimization algorithm which is commonly-used to train machine learning models and neural networks. Training data helps these models learn over time, and the cost function within gradient descent specifically acts as a barometer, gauging its accuracy with each iteration of parameter updates. Until the function is close to or equal to zero, the model will continue to adjust its parameters to yield the smallest possible error.

508. How would you deal with missing random values from a data set?

Deletions. Pairwise Deletion. Listwise Deletion/ Dropping rows. Dropping complete columns. Basic Imputation Techniques. Imputation with a constant value. Imputation using the statistics (mean, median, mode) K-Nearest Neighbor Imputation.

509. What is a Gaussian distribution and it's use?

In a Gaussian distribution(normal distribution), data is symmetrically distributed with no skew. When plotted on a graph, the data follows a bell shape, with most values clustering around a central region and tapering off as they go further away from the center.

510. What is ‘cluster sampling’?

Cluster sampling is a probability sampling technique where researchers divide the population into multiple groups (clusters) for research. Researchers then select random groups with a simple random or systematic random sampling technique for data collection and data analysis. Cluster sampling is defined as a sampling method where the researcher creates multiple clusters of people from a population where they are indicative of homogeneous characteristics and have an equal chance of being a part of the sample.

511. What are the alternatives to PyTorch?

scikit-learn, Keysight Eggplant Platform, machine-learning in Python, V7, Personalizer, Kubeflow, Google Cloud TPU, python-recsys.

512. Why do you use feature selection?

Feature selection offers a simple yet effective way to overcome the challenge of high dimensional data analysis by eliminating redundant and irrelevant data. Removing the irrelevant data improves learning accuracy, reduces the computation time, and facilitates an enhanced understanding for the learning model or data. .

513. What’s the difference between Gaussian Mixture Model and K-Means?

The first visible difference between K-Means and Gaussian Mixtures is the shape the decision boundaries. GMs are somewhat more flexible and with a covariance matrix ∑ we can make the boundaries elliptical, as opposed to circular boundaries with K-means.

Another thing is that GMs is a probabilistic algorithm. If we compare both algorithms, the Gaussian mixtures seem to be more robust. However, GMs usually tend to be slower than K-Means because it takes more iterations of the EM algorithm to reach the convergence.

514. How do you pick k for K-Means?

Calculate the Within-Cluster-Sum of Squared Errors (WSS) for different values of k, and choose the k for which WSS becomes first starts to diminish. In the plot of WSS-versus-k, this is visible as an elbow. Also, the silhouette method is used. The silhouette value measures how similar a point is to its own cluster (cohesion) compared to other clusters (separation).

515. How do you know when Gaussian Mixture Model is applicable?

An approach is to find the clusters using soft clustering methods and then see if they are gaussian. If they are then you can apply a GMM model which represents the whole dataset.

516. How can you assess a good logistic model?

An approach to determining the goodness of fit is through the Homer-Lemeshow statistics, which is computed on data after the observations have been segmented into groups based on having similar predicted probabilities. It examines whether the observed proportions of events are similar to the predicted probabilities of occurrence in subgroups of the data set using a Pearson chi-square test. Small values with large p-values indicate a good fit to the data while large values with p-values below 0.05 indicate a poor fit. The null hypothesis holds that the model fits the data and in the below example we would reject H0.

517. What is bias, variance trade off ?

Bias is the difference between the average prediction of our model and the correct value which we are trying to predict. Variance is the variability of model prediction for a given data point or a value which tells us spread of our data. If our model is too simple and has very few parameters then it may have high bias and low variance. On the other hand if our model has large number of parameters then it's going to have high variance and low bias. So we need to find the right/good balance without overfitting and underfitting the data. This tradeoff in complexity is why there is a tradeoff between bias and variance. An algorithm can't be more complex and less complex at the same time.

518. Why is mean square error a bad measure of model performance?

A disadvantage of the mean-squared error is that it is not very interpretable because MSEs vary depending on the prediction task and thus cannot be compared across different tasks. Assume, for example, that one prediction task is concerned with estimating the weight of trucks and another is concerned with estimating the weight of apples. Then, in the first task, a good model may have an RMSE of 100 kg, while a good model for the second task may have an RMSE of 0.5 kg. Therefore, while RMSE is viable for model selection, it is rarely reported and R2 is used instead.

519. How can the outlier values be treated

Below are some of the methods of treating the outliers

Trimming/removing the outlier: In this technique, we remove the outliers from the dataset.

Quantile based flooring and capping : In this technique, the outlier is capped at a certain value above the 90th percentile value or floored at a factor below the 10th percentile value.

Mean/Median imputation : As the mean value is highly influenced by the outliers, it is advised to replace the outliers with the median value.

520. What is residual error?

The difference between a group of values observed and their arithmetical mean. The difference between what was expected and what was predicted is called the residual error.

521. Select an appropriate value of k in k-means?

A. The elbow approach is a popular way for determining the ideal value of K when using the K-Means Clustering Algorithm. The essential concept behind this method is that it plots various cost values as k changes. There will be fewer elements in the cluster as the value of K grows.

522. How do you deal missing values in datasets?

Deleting Rows with missing values, Impute missing values for continuous variable, Impute missing values for categorical variable, Using Algorithms that support missing values, Prediction of missing values, Imputation using Deep Learning Library — Datawig, Imputation using Machine Learning libraries like SimpleImputer, KNNImputer,etc.

523. What is a confusion matrix?

A confusion matrix is a method of summarising a classification algorithm's performance. Calculating a confusion matrix can help you understand what your classification model is getting right and where it is going wrong. It gives us: "true positive" for correctly predicted event values, "false positive" for incorrectly predicted event values, "true negative" for correctly predicted no-event values, "false negative" for incorrectly predicted no-event values.

524. Difference between correlation and Multicollinearity.?

Correlation refers to the linear relationship between 2 variables. Multicollinearity is a special case of collinearity where a strong linear relationship exists between 2 or more independent variables even if no pair of variables has a high correlation. Number of variables involved is 2 in correlation while in multicollinearity is 2 or more.

525. What is pruning in Decision Tree?

Pruning a decision tree helps to prevent overfitting the training data so that our model generalizes well to unseen data. Pruning a decision tree means to remove a subtree that is redundant and not a useful split and replace it with a leaf node. Decision tree pruning can be divided into two types: pre-pruning and post-pruning.

526. What is loss and cost function?

Loss function is a method of evaluating how well your algorithm models your data set. If your predictions are totally off, your loss function will output a higher number. If they're pretty good, it'll output a lower number. A loss function/error function is for a single training example/input. A cost function, on the other hand, is the average loss over the entire training dataset.

527. How does decision tree makes a decision?

Decision trees are upside down which means the root is at the top and then this root is split into various several nodes. Decision trees are nothing but a bunch of if-else statements in layman terms. It checks if the condition is true and if it is then it goes to the next node attached to that decision.

528. What is skewed Distribution & uniform distribution?

Uniform distribution refers to a condition when all the observations in a dataset are equally spread across the range of distribution. Skewed distribution refers to the condition when one side of the graph has more dataset in comparison to the other side.

529. What is bias and variance in data science?

Bias is the amount that a model's prediction differs from the target value, compared to the training data. Bias error results from simplifying the assumptions used in a model so the target functions are easier to approximate. Variance indicates how much the estimate of the target function will alter if different training data were used. In other words, variance describes how much a random variable differs from its expected value.

530. Difference between RNN and CNN?

The main difference between a CNN and an RNN is the ability to process temporal information — data that comes in sequences, such as a sentence. Recurrent neural networks are designed for this very purpose, while convolutional neural networks are incapable of effectively interpreting temporal information. There are differences in the structures of the neural networks themselves to fit those different use cases. CNNs employ filters within convolutional layers to transform data (more on that later), whereas RNNs are predictive, reusing activation functions from other data points in the sequence to generate the next output in a series.

531. What's the Central Limit Theorem and why is it important?

The central limit theorem tells us that no matter what the distribution of the population is, the shape of the sampling distribution will approach normality as the sample size (N) increases. This is useful, as the research never knows which mean in the sampling distribution is the same as the population mean, but by selecting many random samples from a population the sample means will cluster together, allowing the research to make a very good estimate of the population mean. Thus, as the sample size (N) increases the sampling error will decrease.

534. Explain Central limit theorem ?

The central limit theorem (CLT) asserts that as the sample size grows higher, the distribution of a sample variable approaches a normal distribution (i.e., a "bell curve"), assuming that all samples are identical in size and independent of the population's actual distribution shape.

535. What is Chi-Square test ?

A hypothesis testing method is the Chi-square test. Checking if observed frequencies in one or more categories match expected frequencies is one of two frequent Chi-square tests. The Chi-square test is used to determine whether your data is as expected. The test's core premise is to compare the observed values in your data to the expected values that you would see if the null hypothesis is true. The Chi-square goodness of fit test and the Chi-square test of independence are two regularly used Chi-square tests. Both tests use variables to categorize your data into groups.

536. How will you define the number of clusters in a clustering algorithm?

In k-means clustering, the number of clusters that you want to divide your data points into i.e., the value of K has to be pre-determined whereas in Hierarchical clustering data is automatically formed into a tree shape form (dendrogram).

537. Ways to avoid overfitting

Some steps that we can take to avoid it:

1. Data augmentation

2. L1/L2 Regularization

3. Remove layers / number of units per layer

4. Cross-validation

538. Image classification algorithms

Image Classification algorithms are the algorithms which are used to classify labels for images using their characteristics. Example: Convolutional Neural Networks.

539. args will return?

The special syntax *args in function is used to pass a variable number of arguments to a function. It is used to pass a non-key worded, variable-length argument list. The syntax is to use the symbol * to take in a variable number of arguments; by convention, it is often used with the word args.

540. Difference between having and where clause in SQL.

WHERE Clause is used to filter the records from the table based on the specified condition. HAVING Clause is used to filter record from the groups based on the specified condition.

541. How do you handle categorical data?

One-Hot Encoding is the most common, correct way to deal with non-ordinal categorical data. It consists of creating an additional feature for each group of the categorical feature and mark each observation belonging (Value=1) or not (Value=0) to that group.

542. What Is Interpolation And Extrapolation?

Interpolation is the process of calculating the unknown value from known given values whereas extrapolation is the process of calculating unknown values beyond the given data points.

543. SQL joins and Groups

The SQL Joins clause is used to combine records from two or more tables in a database. The GROUP BY statement groups rows that have the same values into summary rows, like "find the number of customers in each country".

544. How do you handle null values and which Imputation method is more favorable?

Ways to handle missing values in the dataset:

Deleting Rows with missing values.

Impute missing values for continuous variable.

Impute missing values for categorical variable.

Other Imputation Methods.

Using Algorithms that support missing values.

Prediction of missing values.

Multiple imputation is more advantageous than the single imputation because it uses several complete data sets and provides both the within-imputation and between-imputation variability.

545. Difference between Xgboost and Random Forest?

In Random Forest, the decision trees are built independently so that if there are five trees in an algorithm, all the trees are built at a time but with different features and data present in the algorithm. This makes developers look into the trees and model them in parallel. XGBoost builds one tree at a time so that each data pertaining to the decision tree is taken into account and the data is filled if there are any missing data. This helps developers to work with gradient algorithms along with the decision tree algorithm for better results.

546. What is Homoscedasticity?

Homoscedasticity, or homogeneity of variances, is an assumption of equal or similar variances in different groups being compared. This is an important assumption of parametric statistical tests because they are sensitive to any dissimilarities. Uneven variances in samples result in biased and skewed test results. In statistics, a sequence of random variables is homoscedastic if all its random variables have the same finite variance.

547. What is DBSCAN Clustering?

DBSCAN is a density-based clustering algorithm that works on the assumption that clusters are dense regions in space separated by regions of lower density. It groups ‘densely grouped’ data points into a single cluster. It can identify clusters in large spatial datasets by looking at the local density of the data points. The most exciting feature of DBSCAN clustering is that it is robust to outliers. It also does not require the number of clusters to be told beforehand, unlike K-Means, where we have to specify the number of centroids.

548. What is box cox transformation?

A Box Cox transformation is a transformation of non-normal dependent variables into a normal shape. Normality is an important assumption for many statistical techniques; if your data isn't normal, applying a Box-Cox means that you are able to run a broader number of tests.

549. Conditions for Overfitting and Underfitting.

If both the training accuracy and test accuracy are close then the model has not overfit. If the training result is very good and the test result is poor then the model has overfitted. If the training accuracy and test accuracy is low then the model has underfit.

550. What does it mean when the p-values are high and low?

The p-value is a number between 0 and 1 and interpreted in the following way: A small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, so you reject the null hypothesis.

A large p-value (> 0.05) indicates weak evidence against the null hypothesis, so you fail to reject the null hypothesis. p-values very close to the cutoff (0.05) are considered to be marginal (could go either way).

551. What are the differences between correlation and covariance?

Covariance is a measure to indicate the extent to which two random variables change in tandem.Correlation is a measure used to represent how strongly two random variables are related to each other.Covariance is nothing but a measure of correlation.Correlation refers to the scaled form of covariance.

552. How to create a sparse Matrix in Python?

Python's SciPy gives tools for creating sparse matrices using multiple data structures, as well as tools for converting a dense matrix to a sparse matrix. The function csr_matrix() is used to create a sparse matrix of compressed sparse row format whereas csc_matrix() is used to create a sparse matrix of compressed sparse column format.

553. How will you remove duplicates from dataframe based on particular column?

An important part of Data analysis is analyzing Duplicate Values and removing them. Pandas drop_duplicates() method helps in removing duplicates from the data frame. Subset parameter takes a column or list of column label. After passing columns, it will consider them only for duplicates.

555. What is learning rate in gradient descent?

Gradient descent is the popular optimization algorithm used in machine learning to estimate the model parameters. Learning rate, generally represented by the symbol 'α', is a hyper-parameter used to control the rate at which an algorithm updates the parameter estimates or learns the values of the parameters.

556. How do you clean raw data ?

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. While the techniques used for data cleaning may vary according to the types of data your company stores, you can follow these basic steps to map out a framework for your organization. Step 1: Remove duplicate or irrelevant observations, Step 2: Fix structural errors, Step 3: Filter unwanted outliers and Step 4: Handle missing data.

557. Explain any feature selection method?

The goal of feature selection in machine learning is to find the best set of features that allows one to build useful models of studied phenomena. Regularization consists of adding a penalty to the different parameters of the machine learning model to reduce the freedom of the model, i.e. to avoid over-fitting. In linear model regularization, the penalty is applied over the coefficients that multiply each of the predictors. From the different types of regularization, Lasso or L1 has the property that is able to shrink some of the coefficients to zero. Therefore, that feature can be removed from the model.

558. Why random forest better than multiple decided trees?

Overfitting, error due to variation, and error due to bias are all important difficulties with decision trees. A Random Forest is a group of decision trees that produce a single, aggregated outcome. Overfitting is less likely when numerous trees are used in the random forest. They're also difficult to comprehend. A decision tree is simple to read and comprehend, whereas a random forest is more difficult to comprehend. A single decision tree isn't very good at forecasting outcomes, but it's simple to set up. More trees result in a more robust model that avoids overfitting. Each tree in the forest must be generated, processed, and analysed. As a result, this is a lengthy process that might take hours or even days.

559. What is a bias-variance trade-off?

If the algorithm is too basic (hypothesis with linear eq.), it may be prone to errors due to strong bias and low variance. If the algorithms are too sophisticated (hypothesis with a high degree eq.), the variance and bias may be considerable. The new entries will not fare well in the latter scenario. Trade-off, also known as Bias Variance Trade-off, is something that exists between these two situations. There is a tradeoff between bias and variance because of this tradeoff in complexity. It's impossible for an algorithm to be both more complex and less complex at the same time.

560. Describe different regularization methods, such as L1 and L2 regularization?

L1 regularization, also known as L1 norm or Lasso (in regression problems), combats overfitting by shrinking the parameters towards 0. This makes some features obsolete. L2 regularization, or the L2 norm, or Ridge (in regression problems), combats overfitting by forcing weights to be small, but not making them exactly 0.

561. What is the the difference between normal distribution, standard normal distribution and uniform distribution?

A uniform distribution is one in which all values are equally likely within a range (and impossible beyond that range). A Normal distribution is one in which values cluster around the mean, or average, a outlying values are very unlikely. The standard normal distribution, also called the z-distribution, is a special normal distribution where the mean is 0 and the standard deviation is 1.

562. Which should be preferred among Gini impurity and Entropy?

When using sklearn, the Gini impurity is a good default because it is slightly faster to compute. When they work in opposite directions, Gini impurity isolates the most common class in its own branch of the Tree, whereas entropy produces slightly more balanced Trees.

563. Explain the time and space complexity of training and testing in the case of a Decision Tree.

In the training stage for features (dimensions) in the dataset, we sort the data which takes O(n log n) time following which we traverse the data points to find the right threshold which takes O(n) time. Subsequently, for d dimensions, the total time complexity would be: O(nlogn*d). Moreover, the testing time complexity is O(depth) as we have to traverse from the root to a leaf node of the decision tree i.e., testing space complexity is O(nodes) .

564. What do you mean by Bootstrap Sample in random forest?

Each tree in a random forest learns from a random selection of data points during training. Because the samples are drawn via replacement, or bootstrapping, certain samples will be utilised numerous times in a single tree. The theory is that by training each tree on various samples, even if each tree has a lot of variance with regard to a specific set of training data, the overall variance of the forest will be minimised, but not at the expense of bias.

565. What are the Limitations of Bagging Trees?

The loss of interpretability of a model is one of the drawbacks of bagging. When the right technique is ignored, the resulting model can have a lot of bias. Despite its great accuracy, bagging can be computationally expensive, which may deter its adoption in some situations.

566. What is skewed Distribution & uniform distribution?

The term "uniform distribution" refers to a situation in which all of the observations in a dataset are evenly distributed over the distribution range. When one side of a graph has more dataset than the other, the circumstance is known as skewed distribution. A uniform distribution, as opposed to a skewed distribution, contains relatively consistent data, with the frequency of each class being similar to the others.

567. Find the positions of numbers that are multiples of 4 from a series?

```
import pandas as pd

import numpy as np

num_series = pd.Series(np.random.randint(1, 10, 9))

result = np.argwhere(num_series % 4==0)

print(result)
```

568. Solve curse of dimensionality?

A. The curse of dimensionality states that as the number of characteristics grows, the error grows as well. It refers to the fact that high-dimensional algorithms are more difficult to build and often have a running duration that is proportional to the dimensions. Dimensionality reduction is a technique for transforming high-dimensional variables into lower-dimensional variables while preserving the variables' particular information. The curse of dimensionality is caused by unstable parameter estimates; thus, regularising these estimates will assist the parameters in making accurate estimations. One of the most common tools for dimension reduction is PCA (Principal Component Analysis).

569. Get third highest scored student in class table sql?

SELECT * FROM student_table ORDER BY marks DESC LIMIT 2,1

570. Difference between type I and type II errors?

In statistics, a Type I error is a false positive conclusion, while a Type II error is a false negative conclusion. Type I vs. Type II Error as an Example Based on your modest symptoms, you decide to get tested for COVID-19. There are two types of errors that could occur: Type I error (false positive): the test result indicates that you have coronavirus when you don't. Type II error (false negative): the test result indicates that you are free of coronavirus, while you are in fact infected.

571. When underfitting occurs in a static model?

When a mathematical model or a machine learning algorithm fails to capture the primary trend of the data, this is known as underfitting. When a model is too simplistic — informed by too few features or overly regularised — it becomes inflexible in learning from the dataset, resulting in underfitting.

572. While working on a data set,how can you select important variables?

1. Variance: You can remove characteristics with low variance because they don't provide any insight into the forecast and overfit the model.

2. Correlated Features: There may be features with significant correlations among them; we can keep a few and discard the rest. 3. From the Model: This approach of feature selection takes longer but provides us with the most important features supplied by the models themselves. In this case, we can employ 2-3 models and place the train set on top of them, instructing the model to output a certain amount of relevant functions while ignoring the rest. In the Sklearn RFE and RFECV feature selection modules, there are two functions that we can use to pick features.

573. What is Autoencoder?

Autoencoder is an unsupervised artificial neural network that learns how to compress and encode data effectively before reconstructing it back to a representation that is as similar to the original input as feasible. By definition, an autoencoder decreases data dimensionality by learning to ignore noise in the data. Encoder, Bottleneck, Decoder, and Reconstruction Loss: Autoencoders are made up of four primary parts: encoder, bottleneck, decoder, and reconstruction loss.

574. Difference between violin plot and box plot?

Although violin plots are similar to box plots, they provide additional information such as the sample data distribution (density trace). Box plots show data points outside 1.5 times the inter-quartile range as outliers above or below the whiskers by default, whereas violin plots show the entire range of the data.

575. What is statistical power of sensitivity and how do you calculate it?

The statistical power of an A/B test refers to the test's sensitivity to certain magnitudes of effect sizes. More precisely, it is the probability of observing a statistically significant result at level alpha (α) if a true effect of a certain magnitude (MEI) is in fact present.

576. Why do we use Cross validation?

Cross-validation is a technique used in applied machine learning to estimate a machine learning model's skill on unknown data. That is, to use a small sample to assess how the model will perform in general when used to generate predictions on data that was not utilised during the model's training.

577. How will you work in a machine learning project if there is a huge imbalance in the data?

Machine Learning algorithms tend to produce unsatisfactory classifiers when faced with imbalanced datasets. To overcome this, methods like randome oversampling and random undersampling can be used. Synthetic Minority Over-sampling Technique, Modified synthetic minority oversampling technique (MSMOTE), Algorithmic Ensemble Techniques can also be used.

578. Formula of sigmoid function?

A sigmoid function is a mathematical function having a characteristic "S"-shaped curve or sigmoid curve. Sigmoid functions most often show a return value (y axis) in the range 0 to 1. A common example of a sigmoid function is the logistic function shown in the first figure and defined by the formula

S(x) = 1/(1+exp(-x))

579. Can we use sigmoid function in case of multiple classification?

You must use Softmax if your task is a type of classification where the labels are mutually exclusive and each input only has one label. If you have numerous labels for each input in your classification work, your classes are not mutually exclusive, and you can utilise Sigmoid for each output. In the first scenario, the output should be the output entry with the highest value. In the latter scenario, you have an activation value for each class that belongs to the last sigmoid. You can assert that an entry exists in the input if each activation is greater than 0.5.

580. How will you measure the Euclidean distance between the two arrays in NumPy?

The magnitude or length of the line segment between two locations is known as the Euclidean distance in mathematics. We first create two numpy arrays in the first procedure. The Euclidean distance is then computed directly using numpy's linalg.norm() function.

581. Difference between Normalisation and Standardization?

Normalization is the process of rescaling values into a range of [0,1]. Typically, standardisation entails rescaling data to a mean of 0 and a standard deviation of 1. (unit variance).

582. What is Ensemble Learning?

Ensemble learning is the process of systematically generating and combining many models, such as classifiers or experts, to tackle a specific computational intelligence problem. Ensemble learning is largely used to improve a model's performance (classification, prediction, function approximation, etc.) or to lessen the risk of an unintentional poor model selection.

583. Difference between Batch and Stochastic Gradient Descent?

Batch Gradient Descent is highly slow on very big training sets since it entails calculations over the entire training set at each step. As a result, Batch GD becomes extremely computationally expensive. SGD is stochastic in nature, which means it chooses up a "random" instance of training data at each step and then computes the gradient, which is significantly faster than Batch GD because there is much less data to modify at once.

584. How can outlier values be treated?

A. An outlier is an observation in a dataset that differs significantly from the rest of the data. This signifies that an outlier is much larger or smaller than the rest of the data.

Given are some of the methods of treating the outliers: Trimming or removing the outlier, Quantile based flooring and capping, Mean/Median imputation.

585. What is root cause analysis?

A. A root cause is a component that contributed to a nonconformance and should be eradicated permanently through process improvement. The root cause is the most fundamental problem—the most fundamental reason—that puts in motion the entire cause-and-effect chain that leads to the problem (s). Root cause analysis (RCA) is a word that refers to a variety of approaches, tools, and procedures used to identify the root causes of problems. Some RCA approaches are more directed toward uncovering actual root causes than others, while others are more general problem-solving procedures, and yet others just provide support for the root cause analysis core activity.

586. What is bias and variance in Data Science?

The model's simplifying assumptions simplify the target function, making it easier to estimate. Bias is the difference between the Predicted Value and the Expected Value in its most basic form. Variance refers to how much the target function's estimate will fluctuate as a result of varied training data. In contrast to bias, variance occurs when the model takes into account the data's fluctuations, or noise.

587. What is a confusion matrix?

A confusion matrix is a method of summarising a classification algorithm's performance. Calculating a confusion matrix can help you understand what your classification model is getting right and where it is going wrong. This gives us the following: "True positive" for event values that were successfully predicted. "False positive" for event values that were mistakenly predicted. For successfully anticipated no-event values, "true negative" is used. "False negative" for no-event values that were mistakenly predicted.

588. Is skewness in data bad for the model? Why?

A. In a statistical distribution, skewed data is defined as a curve that seems deformed or skewed to the left or right. Many statistical models fail when there is too much skewness in the data. The tail portion of skewed data may act as an outlier for the statistical model, and we know that outliers have a negative impact on model performance, particularly regression-based models.

589. How to train a model robust to outliers?

A. You can employ an outlier-resistant model. Outliers have little effect on tree-based models, but they do alter regression-based models. If you're doing a statistical test, instead of using a parametric test, use a non-parametric one. A robust error metric can be used: The influence of outliers is reduced by switching from mean squared error to mean absolute difference. Set a limit on how much data you can collect. Try a log transformation if your data has a strong right tail.

590. Show me how lamda and map function works together in python

In Python, the map() function accepts two arguments: a function and a list. The function is called with a lambda function and a list, and it returns a new list with all of the lambda modified items returned by that function for each item.

591. Combat Overfitting?

When a model performs well on training data but not on new data, it is said to be overfitted. To avoid overfitting, enhance training data and simplify the model. During the training phase, you should end sooner rather than later (have an eye over the loss over the training period as soon as loss begins to increase stop training). Ridge Regularization and Lasso Regularization are two types of regularisation. To combat overfitting in neural networks, use dropout.

592. What are Entropy and Information gain in Decision tree algorithm?

Entropy is a measure of impurity or uncertainty in a set of data used in information theory. It determines how data is split by a decision tree. The quantity of information improved in the nodes before splitting them for making subsequent judgments can be characterized as the information obtained in the decision tree.

593. What Will Happen If the Learning Rate Is Set inaccurately (Too Low or Too High)?

A high learning rate in gradient descent will cause the learning to jump over global minima, whereas a low learning rate will cause the learning to take too long to converge or become stuck in an unwanted local minimum.

594. What is meant by 'curse of dimensionality'?

The problem produced by the exponential rise in volume associated with adding extra dimensions to Euclidean space is known as the "curse of dimensionality." The curse of dimensionality states that as the number of characteristics grows, the error grows as well. It refers to the fact that high-dimensional algorithms are more difficult to build and often have a running duration that is proportional to the dimensions. A higher number of dimensions theoretically allows for more information to be stored, but in practice, it rarely helps because real-world data contains more noise and redundancy.

595. Difference between remove, del and pop?

remove function removes the first matching value/object. It does not do anything with the indexing. del function removes the item at a specific index. And pop removes the item at a specific index and returns it.

596. What is the difference between squared error and absolute error?

Squared error is the squared difference between the predicted values and the actual value. Absolute Error is the difference between the measured value and true value. The squared error is differentiable everywhere, whereas the absolute error is not (its derivative is undefined at 0). This makes the squared error more susceptible to mathematical optimization strategies.

597. Under what aspects Naive Bayes is bad?

In Naive Bayes, all predictors (or traits) are assumed to be independent, which is rarely the case in real life. This limits the algorithm's usability in real-world scenarios. The 'zero-frequency problem' occurs when an algorithm assigns zero probability to a categorical variable whose category in the test data set was not present in the training dataset. To get over this problem, you should employ a smoothing approach. You shouldn't take its probability outputs seriously because its estimations can be off in some instances.

598. How will you tackle an exploding gradient problem?

The problem of exploding gradients can be solved by reducing the number of layers in the network. Using a smaller batch size when training the network may also be beneficial. Long Short-Term Memory (LSTM) memory units and maybe similar gated-type neuron structures can be used to reduce exploding gradients. When you have an exploding gradient problem, you can use gradient clipping. First, we choose a threshold value, and if the value produced by the gradient function is greater than this threshold, we change it to something else. It may also be beneficial to use suitable weight initialization techniques.

599. How will you prevent overfitting when creating a statistical model?

cross-validation is an effective tool for avoiding overfitting. The aim is to create many micro train-test splits using your initial training data. These divisions can be used to fine-tune your model. More data can help algorithms recognize the signal more accurately. Also, make a list of relevant variables and terms that you'll probably use in your own model.

600. How to add a border that is filled with 0s around an existing array?

For doing this the pad function of numpy can be used by passing a pad width of 1 and constant values as 0.

601. An array has shape (7,4,2) what is the index (x,y,z) of the 20th element?

The answer is (2, 2, 0). This can be found out using the unravel_index function of numpy. Use "print (np.unravel_index(20, (7,4,2)))"

602. For a model an accuracy of 100% is obtained but with the validation set, the accuracy score is 75%. What should be looked out for?

There could be an issue with overfitting. When analysing machine learning algorithms, there are a few key strategies to avoid overfitting: To estimate model accuracy, use a resampling strategy, hold back a validation dataset, use extra training data, or pick relevant features.

603. How is skewness different from kurtosis?

Skewness is a metric for symmetry, or more specifically, the lack of it. If a distribution, or data collection, looks the same to the left and right of the centre point, it is said to be symmetric. Kurtosis is a measure of how heavy-tailed or light-tailed the data are in comparison to a normal distribution. Data sets having a high kurtosis are more likely to contain heavy tails, or outliers. Light tails or a lack of outliers are common in data sets with low kurtosis. The main distinction between skewness and kurtosis is that the former refers to the degree of symmetry in the frequency distribution, whilst the latter refers to the degree of peakedness.

604. In experimental design, is it necessary to do randomization? If yes, why?

A. Yes, in experimental design, randomization is required. Randomization eliminates biases and ensures that the outcomes are balanced. The sample that is randomly picked is designed to be representative of the population, and it is fairly selected because it is not influenced by the researcher. You can achieve the best cause-effect linkages between the variables by randomizing the experiments.

605. What is the benefit of weight initialisation in neural networks?

A. The fundamental goal of weight initialization is to prevent exploding or vanishing gradients in layer activation outputs during forward propagation. If either of these issues arises, loss gradients will be either too great or too little, and the network will take longer to converge, if it can at all. If we appropriately set the weights, we will achieve our goal of optimising the loss function in the shortest time possible; otherwise, converging to a minimum via gradient descent will be impossible.

606. How will you evaluate the performance of a logistic regression model?

The confusion matrix can be used to evaluate a logistic regression model. The accuracy, sensitivity, and specificity of the model can be useful indicators of what you want to do with it - focusing on true positives or false negatives. We can also utilize precision and recall to evaluate your model, as well as the f1 score.

607. How can you make data normal using Box-Cox transformation?

A Box Cox transformation turns non-normal dependent variables into normal shapes. The Lambda value specifies the level of data that should be raised to. The Box-Cox power transformation does this by searching from Lambda = -5 to Lambda = +5 until the best value is discovered.

608. What is the loss function SVM tries to minimize?

Although there is no "loss function" for hard-margin SVMs, the loss does exist when solving soft-margin SVMs. The hinge loss is a loss function used in machine learning to train classifiers. For "maximum-margin" classification, the hinge loss is utilised, most notably for support vector machines (SVMs).

609. Detect heteroscedasticity in simple linear regression?

Heteroscedasticity refers to the situation where the spread of the residuals changes in a systematic way over the range of observed values. A fitted value vs. residual plot is the simplest technique to determine heteroscedasticity. The "cone" form is a clear marker of heteroscedasticity if the residuals become significantly more spread out as the fitted values get greater. The Breusch-Pagan test is a more formal, mathematical method of determining heteroscedasticity.

610. Explain ANOVA?

The analysis of variance (ANOVA) is a statistical technique for determining if the means of two or more groups differ significantly. One-way ANOVA, two-way ANOVA, and multivariate ANOVA are the three types. An ANOVA's null hypothesis is that there is no significant difference between the groups. The alternative hypothesis proposes that the groups have at least one substantial difference. The null hypothesis is rejected and the alternative hypothesis is validated if the p-value associated with the F is less than.05. If the null hypothesis is rejected, one concludes that the means of all the groups are not equal.

611. Determine no. of neighbors in KNN?

The number of neighbors(K) in KNN is a hyperparameter that must be chosen during model construction. According to research, there is no ideal number of neighbors for all types of data sets. A small number of neighbors is the most flexible fit, resulting in low bias but high variation, whereas a big number of neighbors results in a smoother decision boundary, resulting in reduced variance but higher bias. If the number of classes is even, data scientists usually choose an odd number. You can also test the model's performance by creating it with different values of k and comparing the results. Elbow technique is another option.

612. What do you mean by central trend?

The central trend is a description of a dataset represented by a single value that represents the data distribution's center. The following measurements can be used to describe the central tendency of a dataset. The sum of all values in a dataset divided by the total number of values is the mean. The middle value in an ascending-ordered dataset is called the median. The most often occurring value in a dataset is defined by the mode. Despite the fact that the measures listed above are the most generally employed to describe central tendency, there are others, such as geometric mean, harmonic mean, midrange, and geometric median.

613. Significance of Gamma and Regularization in SVM?

Gamma and C(behaves as a regularization parameter in svm) are the two parameters which is very important in SVM. The behavior of the model is very sensitive to the gamma parameter. If gamma is too large, the radius of the area of influence of the support vectors only includes the support vector itself and no amount of regularization with C will be able to prevent overfitting. When gamma is very small, the model is too constrained and cannot capture the complexity or “shape” of the data. The resulting model will behave similarly to a linear model with a set of hyperplanes that separate the centers of high density of any pair of two classes.

614. Is ARIMA model a good fit for every time series problem?

No we cannot use it for every problem although it's a popular method in time series. ARIMA models are applied in some cases where data show evidence of non-stationarity in the sense of mean (but not variance/autocovariance), where an initial differencing step (corresponding to the "integrated" part of the model) can be applied one or more times to eliminate the non-stationarity of the mean function.

615. Difference between stochastic gradient descent (SGD) and gradient descent (GD)?

In both gradient descent (GD) and stochastic gradient descent (SGD), you update a set of parameters in an iterative manner to minimize an error function.

While in GD, you have to run through ALL the samples in your training set to do a single update for a parameter in a particular iteration, in SGD, on the other hand, you use ONLY ONE or SUBSET of training sample from your training set to do the update for a parameter in a particular iteration. If you use SUBSET, it is called Minibatch Stochastic gradient Descent.

Thus, if the number of training samples are large, in fact very large, then using gradient descent may take too long because in every iteration when you are updating the values of the parameters, you are running through the complete training set. On the other hand, using SGD will be faster because you use only one training sample and it starts improving itself right away from the first sample.

SGD often converges much faster compared to GD but the error function is not as well minimized as in the case of GD. Often in most cases, the close approximation that you get in SGD for the parameter values are enough because they reach the optimal values and keep oscillating there.

616. Explain How a System Can Play a Game of Chess Using Reinforcement Learning?

The chess-playing reinforcement learning model starts with a clean slate, and is only given the basic rules of moving the pieces and the ultimate goal, which is to drive the opponent into check mate. At the beginning, the AI knows nothing about the tactics of the game and makes random moves.

But after playing against itself thousands and millions of times, it starts to develop a statistical model of what sequences of moves are likely to win each situation. With reinforced learning, It will make a move (decision), check if it's the right move (feedback), and keep the outcomes in memory for the next step it takes (learning). There is a reward for every correct decision the system takes and punishment for the wrong one.

617. Difference between Gini Impurity and Entropy in a Decision Tree?

The criteria typically used to decide which feature to split on are the Gini index and information entropy. Both of these measures are pretty similar numerically.

*The Gini index is used by the CART (classification and regression tree) algorithm, whereas information gain via entropy reduction is used by algorithms like C4.5.

*Entropy takes slightly more computation time than Gini Index because of the log calculation, maybe that's why Gini Index has become the default option for many ML algorithms.

*The range of Entropy lies in between 0 to 1 and the range of Gini Impurity lies in between 0 to 0.5.

*Gini measurement is the probability of a random sample being classified incorrectly if we randomly pick a label according to the distribution in a branch.Whereas entropy is a measure of information that indicates the disorder of the features with the target.

618. Suppose you found that your model is suffering from low bias and high variance. How would you tackle it?

Low bias and High variance is the classic term of overfitting which means your model is performing very good in the training dataset, but not good in the test data or unseen data

regularization, creating a random forest or any bagging technique work as well. Rationale of bagging is as simple as following - if X1 , ..., Xn are independently distributed with variance v , variance of their average becomes v/n and thereby reducing the variance.

619. What are crosstab in python?

Crosstab function builds a cross-tabulation table that can show the frequency with which certain groups of data appear.

620. Describe how Gradient Boosting works.

Gradient boosting is a type of machine learning boosting. It relies on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error. If a small change in the prediction for a case causes no change in error, then next target outcome of the case is zero. Gradient boosting produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.

621. Describe the decision tree model.

Decision Trees are a type of Supervised Machine Learning where the data is continuously split according to a certain parameter. The leaves are the decisions or the final outcomes. A decision tree is a machine learning algorithm that partitions the data into subsets.

622. What is a neural network?

Neural networks are a set of algorithms, modeled loosely after the human brain, that are designed to recognize patterns. They interpret sensory data through a kind of machine perception, labeling or clustering raw input. They, also known as Artificial Neural Networks, are the subset of Deep Learning.

623. Explain the Bias-Variance Tradeoff

The bias–variance tradeoff is the property of a model that the variance of the parameter estimated across samples can be reduced by increasing the bias in the estimated parameters.

625. Precision and Recall? How they are related to ROC curve?

The recall is the ratio of the relevant results returned by the search engine to the total number of the relevant results that could have been returned. The precision is the proportion of relevant results in the list of all returned search results. When dealing with highly skewed datasets, Precision-Recall (PR) curves in give a more informative picture of an algorithm's performance. We show that a deep connection exists between ROC space and PR space, such that a curve dominates in ROC space if and only if it dominates in PR space.

626. What is feature scaling and why is it necessary?

A. Feature scaling in machine learning is one of the most critical steps during the pre-processing of data before creating a machine learning model. Scaling can make a difference between a weak machine learning model and a better one.

The most common techniques of feature scaling are Normalization and Standardization.

627. Difference between likelihood and probability?

Probability basically corresponds to finding the chance of something given a sample distribution of the data, while on the other hand, likelihood refers to finding the best distribution of the data given a particular value of some feature or some situation in the data.

While calculating the probability, feature value can be varied, but the characteristics(mean, standard and deviation) of the data distribution cannot be altered.

The likelihood in very simple terms means to increase the chances of a particular situtation to happen by varrying the characteristics of the dataset distribution.

628. Difference between Sigmoid and Softmax function?

The Sigmoid Activation Function is a mathematical function with a recognizable “S” shaped curve. It is used for the logistic regression and basic neural network implementation. If we want to have a classifier to solve a problem with more than one right answer, the Sigmoid Function is the right choice. We should apply this function to each element of the raw output independently. The return value of Sigmoid Function is mostly in the range of values between 0 and 1 or -1 and 1.

Whereas the Softmax Activation Function, also know as SoftArgMax or Normalized Exponential Function is a fascinating activation function that takes vectors of real numbers as inputs, and normalizes them into a probability distribution proportional to the exponentials of the input numbers. Before applying, some input data could be negative or greater than 1. Also, they might not sum up to 1. After applying Softmax, each element will be in the range of 0 to 1, and the elements will add up to 1. This way, they can be interpreted as a probability distribution. For more clarification, the larger the input number, the larger the probabilities will be

629. Which machine learning algorithm is known as the lazy learner ?

KNN is a Machine Learning algorithm known as a lazy learner. K-NN is a lazy learner because it doesn't learn any machine learnt values or variables from the training data but dynamically calculates distance every time it wants to classify, hence memorises the training dataset instead.

630. Why does overfitting occur?

A. Overfitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data. This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model

631. What is ensemble learning?

A. Ensemble learning is the process by which multiple models, such as classifiers or experts, are strategically generated and combined to solve a particular computational intelligence problem. Ensemble learning is primarily used to improve the (classification, prediction, function approximation, etc.) performance of a model, or reduce the likelihood of an unfortunate selection of a poor one.

632. What is F1 score?

A. The F1 score is defined as the harmonic mean of precision and recall. As a short reminder, the harmonic mean is an alternative metric for the more common arithmetic mean. It is often useful when computing an average rate. In the F1 score, we compute the average of precision and recall.

633. What is pickling and unpickling?

."Pickling" is the process whereby a Python object hierarchy is converted into a byte stream, and "unpickling" is the inverse operation, whereby a byte stream (from a binary file or bytes-like object) is converted back into an object hierarchy.

634. What is lambda function?

Python Lambda Functions are anonymous function means that the function is without a name. As we already know that the def keyword is used to define a normal function in Python. Similarly, the lambda keyword is used to define an anonymous function in Python.

635. What is the trade of between bias and variance ?

Bias is the simplifying assumptions made by the model to make the target function easier to approximate. Variance is the amount that the estimate of the target function will change given different training data. Trade-off is tension between the error introduced by the bias and the variance.

636.Autoencoder methods

A. Autoencoder is a type of neural network where the output layer has the same dimensionality as the input layer. In simpler words, the number of output units in the output layer is equal to the number of input units in the input layer. Various techniques exist to prevent autoencoders from learning the identity function and to improve their ability to capture important ' information and learn richer representations. 1.Sparse autoencoder (SAE) 2. Denoising autoencoder (DAE) 3. Contractive autoencoder (CAE) 4. Principal component analysis.

637. L1 and L2 regularization?

A. L1 regularization gives output in binary weights from 0 to 1 for the model's features and is adopted for decreasing the number of features in a huge dimensional dataset. L2 regularization disperse the error terms in all the weights that leads to more accurate customized final models.

638. How to measure the Euclidean distance betweeen the two arrays in numpy?

A. Euclidean distance is defined in mathematics as the magnitude or length of the line segment between two points. There are multiple methods for measuring the euclidean methods.

Method 1. In this method, we first initialize two numpy arrays. Then, we use linalg.norm() of numpy basically to compute the euclidean distance directly.

Method 2. In this method, we first initialize two numpy arrays. Then, we take the difference of the two arrays, compute the dot product of the result, and transpose of the result. Then we take the square root of the answer. This is another way to implement Euclidean distance.

Method 3. In this method, we first initialize two numpy arrays. Then, we compute the difference of these arrays and take their square. We take the sum of the squared elements, and after that, we take the square root in the end. This is another way to implement Euclidean distance.

639.What are the support vectors in SVM?

Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, we maximize the margin of the classifier. Deleting the support vectors will change the position of the hyperplane. These are the points that help us build our SVM.

640. How do you handle categorical data?

One-Hot Encoding is the most common, correct way to deal with non-ordinal categorical data. It consists of creating an additional feature for each group of the categorical feature and mark each observation belonging (Value=1) or not (Value=0) to that group.

641. What is coerrelation?

Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate). It's a common tool for describing simple relationships without making a statement about cause and effects

642. What is covariance?

Covariance is nothing but a measure of correlation. Covariance is a measure of how much two random variables vary together. It’s similar to variance, but where variance tells you how a single variable varies, co variance tells you how two variables vary together

644. Kernals in svm, there difference

Kernel Function in SVM is a method used to take data as input and transform into the required form of processing data.

Gaussian Kernel Radial Basis Function (RBF) : It is used to perform transformation, when there is no prior knowledge about data and it uses radial basis method to improve the transformation.

Sigmoid Kernel: this function is equivalent to a two-layer, perceptron model of neural network, which is used as activation function for artificial neurons.

Polynomial Kernel: It represents the similarity of vectors in training set of data in a feature space over polynomials of the original variables used in kernel.

Linear Kernel: used when data is linearly separable.

645. Explain the difference between an array and a linked list.

A. An array is a collection of elements of a similar data type. A linked list is a collection of objects known as a node where node consists of two parts, i.e., data and address. Array elements store in a contiguous memory location. Linked list elements can be stored anywhere in the memory or randomly stored.

646. How do you ensure you are not overfitting a model?

A. Keep your model simple. Use regularization technique. Use cross-validation.

647. How do you fix high variance in a model?

A. You can reduce High variance, by reducing the number of features in the model. There are several methods available to check which features don't add much value to the model and which are of importance. Increasing the size of the training set can also help the model generalize.

648. What are hyperparameters? How do they differ from model parameters?

A. Model Parameters: These are the parameters in the model that must be determined using the training data set. These are the fitted parameters. Hyperparameters: These are adjustable parameters that must be tuned in order to obtain a model with optimal performance.

649. You are told that your regression model is suffering from multicollinearity. How do verify this is true and build a better model?

A simple method to detect multicollinearity in a model is by using something called the variance inflation factor or the VIF for each predicting variable. We can follow these steps in order to build a better model:

Remove some of the highly correlated independent variables.

Linearly combine the independent variables, such as adding them together.

Perform an analysis designed for highly correlated variables, such as principal components analysis or partial least squares regression.

650. You build a random forest model with 10,000 trees. Training error as at 0.00, but validation error is 34.23. Explain what went wrong ?

It means that the model has mimicked the training pattern perfectly that it will cause overfitting problem in test samples. To avoid this overfitting, use techniques like less complex model or cross validation etc.

651. What is the recall, specificity and precision of the confusion matrix?

The recall is the ratio of the relevant results returned by the search engine to the total number of the relevant results that could have been returned. Specificity (SP) is calculated as the number of correct negative predictions divided by the total number of negatives. The precision is the proportion of relevant results in the list of all returned search results.

652. Imputation methods?

A. They are:

List wise or case deletion

Pairwise deletion

Mean substitution

Regression imputation

Maximum likelihood.

654. Gridsearch vs Random search?

A. Random search differs from grid search in that we no longer provide an explicit set of possible values for each hyperparameter; rather, we provide a statistical distribution for each hyperparameter from which values are sampled. Essentially, we define a sampling distribution for each hyperparameter to carry out a randomized search.

655. Hyperparameters in SVM?

A. kernel: It maps the observations into some feature space. Ideally the observations are more easily (linearly) separable after this transformation. There are multiple standard kernels for this transformations, e.g. the linear kernel, the polynomial kernel and the radial kernel.

C: It is a hypermeter in SVM to control error. The C parameter tells the SVM optimization how much you want to avoid misclassifying each training example. For large values of C, the optimization will choose a smaller-margin hyperplane if that hyperplane does a better job of getting all the training points classified correctly.

gamma: Gamma is used when we use the Gaussian RBF kernel. if you use linear or polynomial kernel then you do not need gamma only you need C hypermeter. Somewhere it is also used as sigma. Gamma decides that how much curvature we want in a decision boundary.

656. Ridge vs lasso?

Ridge and Lasso regression are some of the simple techniques to reduce model complexity and prevent over-fitting which may result from simple linear regression . Ridge Regression : In ridge regression, the cost function is altered by adding a penalty equivalent to square of the magnitude of the coefficients. lasso (least absolute shrinkage and selection operator; also Lasso or LASSO) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model.

657.Inter quartile ranges?

The most effective way to find all of your outliers is by using the interquartile range (IQR). The IQR contains the middle bulk of your data, so outliers can be easily found once you know the IQR. Quartiles divide the entire set into four equal parts. So, there are three quartiles, first, second and third represented by Q 1 , Q 2 and Q 3 , respectively. Q 2 is nothing but the median.

662. What's the difference between L1 and L2 regularization?

The main intuitive difference between the L1 and L2 regularization is that L1 regularization tries to estimate the median of the data while the L2 regularization tries to estimate the mean of the data to avoid overfitting. That value will also be the median of the data distribution mathematically.

663. Deal with unbalanced binary classification?

A. Techniques to Handle unbalanced Data:

1. Use the right evaluation metrics

2. Use K-fold Cross-Validation in the right way

3. Ensemble different resampled datasets

4. Resample with different ratios

5. Design your own models

664. Activation function?

Activation functions are mathematical equations that determine the output of a neural network model. It is a non-linear transformation that we do over the input before sending it to the next layer of neurons or finalizing it as output.

666. Why is mean square error a bad measure of model performance?

Mean Squared Error (MSE) gives a relatively high weight to large errors — therefore, MSE tends to put too much emphasis on large deviations.

667. long-tailed distribution ?

A. A long tail distribution of numbers is a kind of distribution having many occurrences far from the "head" or central part of the distribution. Most of occurrences in this kind of distributions occurs at early frequencies/values of x-axis.

668. Outlier? Deal with it?

A. An outlier is an object that deviates significantly from the rest of the objects. They can be caused by measurement or execution error.

Removing outliers is legitimate only for specific reasons. Outliers can be very informative about the subject-area and data collection process. If the outlier does not change the results but does affect assumptions, you may drop the outlier. Or just trim the data set, but replace outliers with the nearest "good" data, as opposed to truncating them completely.

669. Example where the median is a better measure than the mean ?

If your data contains outliers, then you would typically rather use the median because otherwise the value of the mean would be dominated by the outliers rather than the typical values. In conclusion, if you are considering the mean, check your data for outliers, if any then better choose median

671. What is the difference between covariance and correlation?

A. "Covariance" indicates the direction of the linear relationship between variables. "Correlation" on the other hand measures both the strength and direction of the linear relationship between two variables.

672. What is negative indexing? Why is it needed? Can you give an example for the same in python

A. This means that the index value of -1 gives the last element, and -2 gives the second last element of an array. The negative indexing starts from where the array ends.

example: for list L = [0,2,35,3]; L[-1] will print 3 in Python.

673. What is the condition for using a t-test or a z-test?

z-test is used for it when sample size is large, generally n >30. Whereas t-test is used for hypothesis testing when sample size is small, usually n < 30 where n is used to quantify the sample size.

674. What is the main difference between overfitting and underfitting?

Overfitting is a modeling error which occurs when a function is too closely fit to a limited set of data points. Underfitting refers to a model that can neither model the training data nor generalize to new data.

675. What is the KNN imputation method?

The idea in kNN methods is to identify 'k' samples in the dataset that are similar or close in the space. Then we use these 'k' samples to estimate the value of the missing data points. Each sample's missing values are imputed using the mean value of the 'k'-neighbors found in the dataset.

676. Time Series (ARIMA)?

Ans. ARIMA, short for 'AutoRegressive Integrated Moving Average', is a forecasting algorithm based on the idea that the information in the past values of the time series can alone be used to predict the future values.

677. How to reduce overfitting ?

Ans. Techniques to reduce overfitting:

Increase training data.

*Reduce model complexity.

*Early stopping during the training phase.

*Ridge Regularization and Lasso Regularization.

678. What is precision/recall ratio?

When it comes to precision we're talking about the true positives over the true positives plus the false positives. As opposed to recall which is the number of true positives over the true positives and the false negatives.

679. Dimensionality reduction?

Ans. Dimensionality Reduction is used to reduce the feature space with consideration by a set of principal features.

680. Bias and variance?

Bias is one type of error which occurs due to wrong assumptions about data such as assuming data is linear when in reality, data follows a complex function. On the other hand, variance gets introduced with high sensitivity to variations in training data.

681. Difference between classification and clustering?

In classification data are grouped by analyzing data objects whose class label is known. Clustering analyzes data objects without knowing class label. There is some prior knowledge of attributes of each classification. There is no prior knowledge of attributes of data to form clusters.

683. Assumptions in Multiple linear regression

Ans. The regression has five key assumptions:

Linear relationship.

Multivariate normality.

No or little multicollinearity.

No auto-correlation.

Homoscedasity

684. Entropy

Ans. Entropy is a measure of disorder or uncertainty and the goal of machine learning models and Data Scientists in general is to reduce uncertainty. Entropy can be defined as a measure of the purity of the sub split. Entropy always lies between 0 to 1.

685. Random forest algorithm

Ans. Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. It is based on model ensemble learning technique.

686. XGBoost Algorithm

Ans. XGBoost is an algorithm that has recently been dominating applied machine learning and Kaggle competitions for structured or tabular data. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.

687. Central limit theorem

Ans. The central limit theorem states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement , then the distribution of the sample means will be approximately normally distributed.

688. VIF

Ans. Variance Inflation Factor (VIF) is used to detect the presence of multicollinearity. Variance inflation factors (VIF) measure how much the variance of the estimated regression coefficients are inflated as compared to when the predictor variables are not linearly related.

689. Difference Between Bagging and Boosting

Ans. Bagging is a way to decrease the variance in the prediction by generating additional data for training from dataset using combinations with repetitions to produce multi-sets of the original data. Boosting is an iterative technique which adjusts the weight of an observation based on the last classification.

690. P value and it's significance

Ans. The p-value is the probability that the null hypothesis is true. (1 – the p-value) is the probability that the alternative hypothesis is true. A low p-value shows that the results are replicable. A low p-value shows that the effect is large or that the result is of major theoretical, clinical or practical importance.

691. Type I and Type II error

A type I error (false-positive) occurs if an investigator rejects a null hypothesis that is actually true in the population; a type II error (false-negative) occurs if the investigator fails to reject a null hypothesis that is actually false in the population.

692. Difference between Correlation and Regression.

Ans. The main difference in correlation vs regression is that the measures of the degree of a relationship between two variables; let them be x and y. Here, correlation is for the measurement of degree, whereas regression is a parameter to determine how one variable affects another.

693. Why do we square the residuals instead of using modulus?

Ans. It is because of the extra penalty for higher errors and squaring the residuals for mean deviation were observed to be more efficient than mean absolute deviation.

694. Which evaluation metric should you prefer to use for a dataset having a lot of outliers in it?

Ans. Mean Absolute Error(MAE) is preferred when we have too many outliers present in the dataset because MAE is robust to outliers whereas MSE and RMSE are very susceptible to outliers and these start penalizing the outliers by squaring the error terms.

695. Heteroscedasticity? How to detect it?

Ans. Heteroskedasticity refers to situations where the variance of the residuals is unequal over a range of measured values. When running a regression analysis, heteroskedasticity results in an unequal scatter of the residuals (also known as the error term). To check for heteroscedasticity, you need to assess the residuals by fitted value plots specifically.

696. p-value?

Ans. A p-value is a measure of the probability that an observed difference could have occurred just by random chance. The lower the p-value, the greater the statistical significance of the observed difference. P-value can be used as an alternative to or in addition to pre-selected confidence levels for hypothesis testing.

697. Root Cause Analysis?

Ans. Root cause analysis (RCA) is defined as a collective term that describes a wide range of approaches used to uncover causes of problems. Some RCA approaches are geared more toward identifying true root causes than others.

698. Regularization?

Ans. Regularization is the process which regularizes or shrinks the coefficients towards zero. In simple words, regularization discourages learning a more complex or flexible model, to prevent overfitting.

699. DBSCAN Clustering?

Density-based spatial clustering of applications with noise (DBSCAN) is a well-known data clustering algorithm that is commonly used in data mining and machine learning. The main idea behind DBSCAN is that a point belongs to a cluster if it is close to many points from that cluster. It is able to find arbitrary shaped clusters and clusters with noise (i.e. outliers).

700. How can you assess a good logistic model?

Ans. Measuring the performance of Logistic Regression:

1. One can evaluate it by looking at the confusion matrix and count the misclassifications (when using some probability value as the cutoff)

2. One can evaluate it by looking at statistical tests such as the Deviance or individual Z-scores.

701. How Regularly Must an Algorithm be Updated?

Ans. It can vary time to time depending upon number of updates happened in the algorithm as per the requirement.

702. Why Is Resampling Done?

Ans. Resampling methods are used to ensure that the model is good enough and can handle variations in data. The model does that by training it on the variety of patterns found in the dataset.

703. Write the formula to calculate R-square?

Ans. R^2 = 1 - (RSS/TSS) where RSS = sum of squares of residual and TSS = Total sum of squares

704. Goal of A/B Testing?

Ans. A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

705. Do Gradient Descent Methods at All-Time Converge to a Similar Point?

Ans. No, they always don't. That's because in some cases it reaches a local minima or a local optima point.

706. Feature Vectors?

Ans. A feature vector is a vector containing multiple elements about an object. Putting feature vectors for objects together can make up a feature space.

707. If p value is more than 0.05 what does it mean?

A p-value higher than 0.05 (> 0.05) is not statistically significant and indicates strong evidence for the null hypothesis. This means we retain the null hypothesis and reject the alternative hypothesis.

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

Normal distribution, also known as the Gaussian 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. In graph form, normal distribution will appear as a bell curve.

710. How can you compute significance using p-value

Ans. The p-value is the probability that the null hypothesis is true. (1 – the p-value) is the probability that the alternative hypothesis is true. The smaller the p-value, the stronger the evidence that you should reject the null hypothesis.

711. Differentiate between a multi-label classification problem and a multi-class classification problem.

Ans. Multinomial classification is the problem of classifying instances into one of three or more classes (classifying instances into one of two classes is called binary classification). Multi-label classification involves predicting zero or more class labels.

712. If the training loss of your model is high and almost equal to the validation loss, what does it mean? What should you do?

Ans. If the training loss is high and validation loss are almost equal, it means that you're avoiding case of overfitting which is good. But, if your loss is high it means your model is going through underfitting. In this case, you can surely increase layers to increase your accuracy and decrease your training loss.

713. Why L1 regularizations cause parameter sparsity whereas L2 regularization does not?

This is due to the shape of Bias region formed by L1 Norm. When compared to L2 Norm, L1 doesn't concede any area around the axes.

714. What is the advantage of performing dimensionality reduction before fitting an SVM?

Support Vector Machine Learning Algorithm performs better in the reduced space. It is beneficial to perform dimensionality reduction before fitting an SVM if the number of features is large when compared to the number of observations.

715. Different activation function?

Binary Step Function, Linear Activation Function, Sigmoid/Logistic Activation Function, Tanh Function (Hyperbolic Tangent), ReLU Activation Function.

716. How do you handle imbalance data?

Follow these techniques:

Use the right evaluation metrics.

Use K-fold Cross-Validation in the right way.

Ensemble different resampled datasets.

Resample with different ratios.

Cluster the abundant class.

Design your own models.

717. Difference between sigmoid and softmax ?

The sigmoid function is used for the two-class logistic regression, whereas the softmax function is used for the multiclass logistic regression (a.k.a. MaxEnt, multinomial logistic regression, softmax Regression, Maximum Entropy Classifier).

718. Explain about optimizers?

Ans. Optimizers are algorithms or methods used to change the attributes of the neural network such as weights and learning rate to reduce the losses. Optimizers are used to solve optimization problems by minimizing the function.

719. Precision-Recall Trade off ?

Ans. The Idea behind the precision-recall trade-off is that when a person changes the threshold for determining if a class is positive or negative it will tilt the scales. It means that it will cause precision to increase and recall to decrease, or vice versa.

720. Decision Tree Parameters?

Ans. These are the parameters used for building Decision Tree: min_samples_split, min_samples_leaf, max_features and criterion.

721. equation to calculate the precision and recall rate.?

Precision = True positives/ (True positives + False positives) = TP/ (TP + FP).

Recall = TruePositives / (TruePositives + FalseNegatives) = TP / (TP + FN).

722. Difference between Point Estimates and Confidence Interval?

Ans. The two are closely related. In fact, the point estimate is located exactly in the middle of the confidence interval. However, confidence intervals provide much more information and are preferred when making inferences.

723. TF/IDF vectorization?

Ans. tf-idf vectorization gives a numerical representation of words entirely dependent on the nature and number of documents being considered. The same words will have different vector representations in another corpus.

724. Entropy and Information gain in Decision tree algorithm?

Ans. Entropy can be defined as a measure of the purity of the sub split. Entropy always lies between 0 to 1. The information gain is based on the decrease in entropy after a dataset is split on an attribute. Constructing a decision tree is all about finding attribute that returns the highest information gain (i.e., the most homogeneous branches).

725. Define Degree of Freedom with example.

Ans. Degrees of Freedom Definition Degrees of freedom (df) refers to the number of independent values (variable) in a data sample used to find the missing piece of information (fixed) without violating any constraints imposed in a dynamic system. Exp: degree of freedom for given sequence: x = 2, 8, 3, 6, 4, 2, 9, 5.

Given n=8

Therefore,

DF = n-1

DF = 8-1

DF = 7.

726. Explain the difference between Variance and R squared error.

Ans. Considering this aspect in regression analysis, the variance is the mean squared error that measures the squared and thus, the summed difference between the actual values and the values predicted through the formed regression equation. R-squared error is completely different in concept as compared to variance.

727. What is an example of a data set with a non-Gaussian distribution?

Ans. Any distribution of money or value will be non--Gaussian. For example: distributions of income; distributions of house prices; distributions of bets placed on a sporting event. These distributions cannot have negative values and will usually have extended right hand tails.

728. Correlation and what is its range?

Ans. “Correlation” on the other hand measures both the strength and direction of the linear relationship between two variables. It is scaled between the range, -1 and +1.

729. You have two dices to play with. You win Rs10 every time you roll a 5. If you play till you win and then stop, what is the expected winning money?

Ans. 1/6 = Probability of getting 5 on dice. It will be (1/6 + 5/6*1/6 + 5/6*5/6*1/6 +....)*10 = 1/6*1/(1-(5/6))*10 = 1*10. Expected money to win is 10Rs. As eventually, we're not stopping till win, so prize money is 10Rs.

730. Choose algorithm if your data has noise?

Ans. Algorithms like Probabilistic Random Forest and DBSCAN can work well when you've noise in data.

731. Is it fine if we don't do scaling before training model? Reason?

Ans. If the values of the features are closer to each other there are chances for the algorithm to get trained well and faster instead of the data set where the data points or features values have high differences with each other will take more time to understand the data and the accuracy will be lower.

So if the data in any conditions has data points far from each other, scaling is a technique to make them closer to each other or in simpler words, we can say that the scaling is used for making data points generalized so that the distance between them will be lower.

732. Tricks you use to faster your model training?

Ans. Reduce Calculations by normalization or standardization. Use high computational memory while training(like GPU instead of local memory).

733. What is SelectK best? how does it works?

Ans. The SelectKBest method selects the features according to the k highest score. By changing the 'score_func' parameter we can apply the method for both classification and regression data. Selecting best features is important process when we prepare a large dataset for training. The SelectKBest class just scores the features using a function (in this case f_classif but could be others) and then "removes all but the k highest scoring features".

734. How will you handle heavy data ?

Ans. 1. Allocate more memory 2. Work with smaller sample 3. Use a relational database like SQL 4. Use a big data platform such as Hadoop.

735. GMM vs K means?

Ans. Gaussian mixture models (GMMs) are often used for data clustering. You can use GMMs to perform either hard clustering or soft clustering on query data. To perform hard clustering, the GMM assigns query data points to the multivariate normal components that maximize the component posterior probability, given the data.

K-means clustering aims to partition data into k clusters in a way that data points in the same cluster are similar and data points in the different clusters are farther apart. Similarity of two points is determined by the distance between them.

736. Regression imputation?

Ans. Regression imputation fits a statistical model on a variable with missing values. Predictions of this regression model are used to substitute the missing values in this variable.

737. Z score for outliers treatment?

Ans. Z score test is one of the most commonly used methods to detect outliers. It measures the number of standard deviations away the observation is from the mean value. A z score of 1.5 indicated that the observation is 1.5 standard deviations above the mean and -1.5 means that the observation is 1.5 standard deviations below or less than the mean.

738. Handle data with lot of noise?

Ans. Noisy data can be handled by following the given procedures:

1) Binning:

- Binning methods smooth a sorted data value by consulting the values around it.

- The sorted values are distributed into a number of “buckets,” or bins.

2) Regression:

- Here data can be smoothed by fitting the data to a function.

- Linear regression involves finding the “best” line to fit two attributes, so that one attribute can be used to predict the other.

- Multiple linear regressionis an extension of linear regression, where more than two attributes are involved and the data are fit to a multidimensional surface.

3) Clustering:

- Outliers may be detected by clustering, where similar values are organized into groups, or “clusters.”

739. Min sample leaf and max depth in random forest?

Ans. Min sample leaf specifies the minimum number of samples that should be present in the leaf node after splitting a node. The max_depth of a tree in Random Forest is defined as the longest path between the root node and the leaf node. Using the max_depth parameter, I can limit up to what depth I want every tree in my random forest to grow.

740. Dict and list comprehension?

Ans. The only difference between list and dictionary comprehension is that the dictionary has the keys and values in it. So, the keys and values will come as the expression or value.

741. What is the condition for using a t-test or a z-test?

Ans. z-test is used for it when sample size is large, generally n >30. Whereas t-test is used for hypothesis testing when sample size is small, usually n < 30 where n is used to quantify the sample size.

742. What is the KNN imputation method?

Ans. The idea in kNN methods is to identify 'k' samples in the dataset that are similar or close in the space. Then we use these 'k' samples to estimate the value of the missing data points. Each sample's missing values are imputed using the mean value of the 'k'-neighbors found in the dataset.

743. What is Yolo?

Ans. YOLO - You Only Look Once is an algorithm proposed by by Redmond et. al in a research article published at the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) as a conference paper, winning OpenCV People’s Choice Award.

Compared to the approach taken by object detection algorithms before YOLO, which repurpose classifiers to perform detection, YOLO proposes the use of an end-to-end neural network that makes predictions of bounding boxes and class probabilities all at once.

744. Select perfect k for k means

Ans. There is a popular method known as elbow method which is used to determine the optimal value of K to perform the K-Means Clustering Algorithm. The basic idea behind this method is that it plots the various values of cost with changing k. As the value of K increases, there will be fewer elements in the cluster.

745. Model metrics when you have outliers?

Ans. The bigger the MAE, the more critical the error is. It is robust to outliers. Therefore, by taking the absolute values, MAE can deal with the outliers.

746. Features for food delivery data to give discount to selected customers?

Ans. The features for same can be: Total number of orders, Frequency of ordering per week, Amount paid per order, Distance travelled by delivery man etc.

747. Filter and wrapper feature selection methods?

Ans. Wrapper methods measure the “usefulness” of features based on the classifier performance. In contrast, the filter methods pick up the intrinsic properties of the features (i.e., the “relevance” of the features) measured via univariate statistics instead of cross-validation performance. So, wrapper methods are essentially solving the “real” problem.

Filter methods: information gain, chi-square test, fisher score, correlation coefficient, variance threshold

Wrapper methods: recursive feature elimination, sequential feature selection algorithms, genetic algorithms

748. C and degree in SVM?

Ans. The C parameter tells the SVM optimization how much you want to avoid misclassifying each training example. For large values of C, the optimization will choose a smaller-margin hyperplane if that hyperplane does a better job of getting all the training points classified correctly. degree: It is the degree of the polynomial kernel function ('poly') and is ignored by all other kernels. The default value is 3.

749. Dbscan vs kmeans?

Ans. In Data Science and Machine Learning, KMeans and DBScan are two of the most popular clustering(unsupervised) algorithms. Density clustering algorithms use the concept of reachability i.e. how many neighbors has a point within a radius. DBScan is more lovely because it doesn’t need parameter, k, which is the number of clusters we are trying to find, which KMeans needs. When you don’t know the number of clusters hidden in the dataset and there’s no way to visualize your dataset, it’s a good decision to use DBScan. DBSCAN produces a varying number of clusters, based on the input data.

750. adjusted R2 and R2?

Ans. R2 is statistical measurement used to explain the dependent and independent variables. Adjusted R Squared is a measurement that predicts the regression variables. This model will take additional input variable that predicts to solve the problems. Adjusted R2 is the better model when you compare models that have a different amount of variables.

751. How to decide imputation method if there are null values in a dataset?

Ans. Mean/Median Imputation:-the mean or a median value of a variable is used in place of the missing data value for that same variable. Median over mean when the data column has any outliers.

Mode substitution:- the highest occuring value for categorical value is used in place of the missing data value of the same variable.

Deleting Column: If the number of null values are more than 70%, then prefer deleting the column as it doesn't contribute to the model.

752.   How to improve model performance?

Ans. Follow these techniques:

1. Use Validation methods

2. Add more data

3. Apply feature engineering techniques(Normalization, Imputation etc)

4. Compare Multiple algorithms

5. Hyperparameter Tuning

753. Standardization vs log transformation?

Ans. Standardization is the process of putting different variables on the same scale. This process allows you to compare scores between different types of variables. Typically, to standardize variables, you calculate the mean and standard deviation for a variable. Log transformation is a data transformation method in which it replaces each variable x with a log(x). Log-transform decreases skew in some distributions, especially with large outliers. But, it may not be useful as well if the original distributed is not skewed. Also, log transform may not be applied to some cases (negative values), but standardization is always applicable (except σ=0).

754. Object detection?

Ans. Object detection is a computer vision technique for locating instances of objects in images or videos. Object detection algorithms typically leverage machine learning or deep learning to produce meaningful results.

755. Data cleaning steps?

Ans. Step 1: Remove duplicate or irrelevant observations. Remove unwanted observations from your dataset, including duplicate observations or irrelevant observations.

*Step 2: Fix structural errors.

*Step 3: Filter unwanted outliers.

*Step 4: Handle missing data.

*Step 5: Validate your data if it's appropriate according to problem statement

756. Upsampling and downsampling methods

Ans. In a classification task, there is a high chance for the algorithm to be biased if the dataset is imbalanced. An imbalanced dataset is one in which the number of samples in one class is very higher or lesser than the number of samples in the other class.

To counter such imbalanced datasets, we use a technique called up-sampling and down-sampling.

In up-sampling, we randomly duplicate the observations from the minority class in order to reinforce its signal. The most common way is to resample with replacement.

In down-sampling, we randomly remove the observations from the majority class. Thus after up-sampling or down-sampling, the dataset becomes balanced with same number of observations in each class.

757. hypothesis testing?

Ans. Hypothesis testing is defined as the process of choosing hypotheses for a particular probability distribution, based on observed data. We use this to test whether a hypothesis can be accepted or not.

758. Inter quartile ranges?

Ans. The most effective way to find all of your outliers is by using the interquartile range (IQR). The IQR contains the middle bulk of your data, so outliers can be easily found once you know the IQR. Quartiles divide the entire set into four equal parts. So, there are three quartiles, first, second and third represented by Q 1 , Q 2 and Q 3 , respectively. Q 2 is nothing but the median.

759. Imputation methods?

Ans. They are:

*List wise or case deletion

*Pairwise deletion

*Mean substitution

*Regression imputation

*Maximum likelihood.

761. Gridsearch vs Random search?

Ans. Random search differs from grid search in that we no longer provide an explicit set of possible values for each hyperparameter; rather, we provide a statistical distribution for each hyperparameter from which values are sampled. Essentially, we define a sampling distribution for each hyperparameter to carry out a randomized search.

762. Hyperparameters in SVM?

Ans. kernel: It maps the observations into some feature space. Ideally the observations are more easily (linearly) separable after this transformation. There are multiple standard kernels for this transformations, e.g. the linear kernel, the polynomial kernel and the radial kernel.

*C: It is a hypermeter in SVM to control error. The C parameter tells the SVM optimization how much you want to avoid misclassifying each training example. For large values of C, the optimization will choose a smaller-margin hyperplane if that hyperplane does a better job of getting all the training points classified correctly.

*gamma: Gamma is used when we use the Gaussian RBF kernel. if you use linear or polynomial kernel then you do not need gamma only you need C hypermeter.

763. Ridge vs lasso?

Ans. Ridge and Lasso regression are some of the simple techniques to reduce model complexity and prevent over-fitting which may result from simple linear regression . Ridge Regression : In ridge regression, the cost function is altered by adding a penalty equivalent to square of the magnitude of the coefficients. lasso (least absolute shrinkage and selection operator; also Lasso or LASSO) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model.

764. What is a sensitivity analysis in the decision making process?

Ans: Sensitivity analysis is a method for predicting the outcome of a decision if a situation turns out to be different compared to the key predictions. It helps in assessing the riskiness of a strategy. Helps in identifying how dependent the output is on a particular input value.

765. How to decide imputation method if there are null values in a dataset?

Ans: Mean/Median Imputation:- In a mean or median substitution, the mean or a median value of a variable is used in place of the missing data value for that same variable. Median over mean when the data column has any outliers.

Mode substitution:- In mode substitution, the highest occurring value for categorical value is used in place of the missing data value of the same variable.

Deleting Column: If the number of null values are more than 70%, then prefer deleting the column as it doesn't contribute to the model. If it does, try other methods such as regression or k-means imputation.

766. Difference between LSTM and Simple RNN?

Ans. The main difference between RNN and LSTM is in terms of which one maintain information in the memory for the long period of time. Here LSTM has advantage over RNN as LSTM can handle the information in memory for the long period of time as compare to RNN.

767. How do you know if you have enough data for your model?

Ans. Right kind of data is much more important than amount of data for your model, and should be prioritize first. If your model has diverse features, then surely you'll require more data as possible in order to understand it better. Having as much data possible is always preferable though one can always check with plotting learning curve in order to determine enough data.

768. P-value and confidence interval?

Ans. The P-value is known as the level of marginal significance within the hypothesis testing that represents the probability of occurrence of the given event. A confidence interval is a kind of interval calculation, obtained from the observed data that holds the actual value of the unknown parameter. It displays the probability that a parameter will fall between a pair of values around the mean.

769. ANOVA vs Chi-Square test

Ans. Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples. A chi-square test is a statistical test used to compare observed results with expected results. The purpose of this test is to determine if a difference between observed data and expected data is due to chance, or if it is due to a relationship between the variables you are studying.

770. How model behaves when you have multicollinearity?

Ans. Multicollinearity causes the following two basic types of problems: The coefficient estimates can swing wildly based on which other independent variables are in the model. The coefficients become very sensitive to small changes in the model.

771. Name algorithms which does not need scaling?

Ans. CART, Random Forests, Gradient Boosted Decision Trees. Algorithms that do not require normalization/scaling are the ones that rely on rules. They would not be affected by any monotonic transformations of the variables. Scaling is a monotonic transformation

772. Detect outliers, what are quartiles?

Ans. The most effective way to find all of your outliers is by using the interquartile range (IQR). The IQR contains the middle bulk of your data, so outliers can be easily found once you know the IQR. Quartiles divide the entire set into four equal parts. So, there are three quartiles, first, second and third represented by Q 1 , Q 2 and Q 3 , respectively. Q 2 is nothing but the median.

773. Clustering, methods. Density based approach? Why?

Ans. Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them. It is basically a type of unsupervised technique.

Types of Clustering Methods:

1. Hierarchical based clustering

2. K-means clustering

3. Density based clustering: Density-based clustering methods provide a safety valve. Instead of assuming that every point is part of some cluster, we only look at points that are tightly packed and assume everything else is noise. This approach requires two parameters: a radius ϵ and a neighborhood density θ. The idea of density-based is that we need to compare the density around an object with the density around its local neighbors.

It is a better approach because it does not require a-priori specification of number of clusters and is able to identify noise data while clustering.

774. Do classification for 2 classes if Classes has 95: 5 ratio of records.

Ans. This is called the case of imbalanced classification dataset. You can follow these techniques:

Use your own right evaluation metrics according to problem statement.

*Use K-fold Cross-Validation in the right way.

*Ensemble different resampled datasets.

*Resample with different ratios.

*Cluster the abundant class.

*Design your own models.

775. Model Pipeline? Benefits

Ans. Automating the machine learning workflow by enabling data to be transformed and correlated into a model that can then be analyzed to achieve outputs is done with ML pipelines. This type of ML pipeline makes the process of inputting data into the ML model fully automated. The main objective of having a proper pipeline for any ML model is to exercise control over it. A well-organised pipeline makes the implementation more flexible.

776. Multicollinearity? Deal?

Ans. Multicollinearity occurs when two or more independent variables(also known as predictor) are highly correlated with one another in a regression model. This means that an independent variable can be predicted from another independent variable in a regression model.

To remove multicollinearities, we can do two things.

1. We can create new features

2. remove them from our data.

777. Time Series (ARIMA)?

Ans. ARIMA, short for 'AutoRegressive Integrated Moving Average', is a forecasting algorithm based on the idea that the information in the past values of the time series can alone be used to predict the future values.

778. What is precision/recall ratio?

Ans. When it comes to precision we're talking about the true positives over the true positives plus the false positives. As opposed to recall which is the number of true positives over the true positives and the false negatives.

779. Bias and variance?

Ans. Bias is one type of error which occurs due to wrong assumptions about data such as assuming data is linear when in reality, data follows a complex function. On the other hand, variance gets introduced with high sensitivity to variations in training data.

780. Conditional Probability

Ans. Conditional probability formula gives the measure of the probability of an event given that another event has occurred.

781. Hypothesis Testing. Null and Alternate hypothesis

Ans. Hypothesis testing is defined as the process of choosing hypotheses for a particular probability distribution, on the basis of observed data Hypothesis testing is simply a core and important topic in statistics.

A null hypothesis is a statistical hypothesis in which there is no significant difference exist between the set of variables. It is the original or default statement, with no effect, often represented by H0 (H-zero). It is always the hypothesis that is tested.

Alternative Hypothesis is a statistical hypothesis used in hypothesis testing, which states that there is a significant difference between the set of variables. It is often referred to as the hypothesis other than the null hypothesis, often denoted by H1 (H-one). The acceptance of alternative hypothesis depends on the rejection of the null hypothesis i.e. until and unless null hypothesis is rejected, an alternative hypothesis cannot be accepted.

782. Why use Decision Trees?

Ans. First, a decision tree is a visual representation of a decision situation (and hence aids communication). Second, the branches of a tree explicitly show all those factors within the analysis that are considered relevant to the decision (and implicitly those that are not).

783. PCA Advantages and Disadvantages?

Advantages: Removes correlated features, Improves algorithm features, Reduces overfitting, Improves Visualization.

Disadvantages: Independent variables become less interpretable, Data Standardization is must, Information loss.

785. Meaning when the p-values are high and low?

Ans. High P values: your data are likely with a true null. Low P values: your data are unlikely with a true null. A low P value suggests that your sample provides enough evidence that you can reject the null hypothesis for the entire population.

786. Missing values of more than 30%, how will you deal with such a dataset?

Ans. Simple approaches include taking the average of the column and use that value, or if there is a heavy skew the median or mode might be better. A better approach, you can perform regression or nearest neighbor imputation on the column to predict the missing values. Then continue on with your analysis/model.

787. random forest or multiple decision trees?

Ans. Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. It is based on model ensemble learning technique.

788. regularization method?

Ans. Regularization is a technique used for tuning the function by adding an additional penalty term in the error function. Regularization, significantly reduces the variance of the model, without substantial increase in its bias.

789. RMSE and MSE ?

Ans. MSE (Mean Squared Error) represents the difference between the original and predicted values which are extracted by squaring the average difference over the data set. It is a measure of how close a fitted line is to actual data points. The lesser the Mean Squared Error, the closer the fit is to the data set. The MSE has the units squared of whatever is plotted on the vertical axis. RMSE (Root

Mean Squared Error) is the error rate by the square root of MSE. RMSE is the most easily interpreted statistic, as it has the same units as the quantity plotted on the vertical axis or Y-axis. RMSE can be directly interpreted in terms of measurement units, and hence it is a better measure of fit than a correlation coefficient.

790. Can time-series data be declared as stationery? How?

Ans. It is stationary when the variance and mean of the series are constant with time.

791. Pruning?

Ans. Pruning is a data compression technique in machine learning and search algorithms that reduces the size of decision trees by removing sections of the tree that are non-critical and redundant to classify instances.

792. Assumptions required for carrying out linear regression?Q2. Pickling and Unpickling?

Ans. Pickling in python refers to the process of serializing objects into binary streams, while unpickling is the inverse of that. It's called that because of the pickle module in Python which implements the methods to do this.

793. Lambda function?

Ans. Python Lambda Functions are anonymous function means that the function is without a name. As we already know that the def keyword is used to define a normal function in Python. Similarly, the lambda keyword is used to define an anonymous function in Python. This function can have any number of arguments but only one expression, which is evaluated and returned.

794. print(len(list1 + list2)). What is the output?

Ans. list1 + list2 gives a new list which is combined both of list1 and list2. so the length will be combined both length, i.e, len(list1) + len(list2).

795. Shallow copy and deep copy?

Ans. In Shallow copy, a copy of the original object is stored and only the reference address is finally copied. In Deep copy, the copy of the original object and the repetitive copies both are stored. Shallow copy is faster than Deep copy. The changes made in the copied object also reflect the original object in shallow copy. There is no reflection on the original object when the changes are made in the copied object in deep copy.

796. Specificity?

Ans. Specificity (SP) is calculated as the number of correct negative predictions divided by the total number of negatives.

797. Combat overfitting and underfitting?

Ans. To avoid overfitting: Keep your model simple, Use regularization technique, Use cross-validation. To avoid underfitting: Decrease regularization, increase the training duration.

798. Activation function?

Ans. Activation functions are mathematical equations that determine the output of a neural network model. It is a non-linear transformation that we do over the input before sending it to the next layer of neurons or finalizing it as output.

799. Dimension reduction?

Ans. Dimensionality Reduction is used to reduce the feature space with consideration by a set of principal features.

800. Why is mean square error a bad measure of model performance?

Ans. Mean Squared Error (MSE) gives a relatively high weight to large errors — therefore, MSE tends to put too much emphasis on large deviations.

801. Example where the median is a better measure than the mean ?

Ans. If your data contains outliers, then you would typically rather use the median because otherwise the value of the mean would be dominated by the outliers rather than the typical values. In conclusion, if you are considering the mean, check your data for outliers, if any then better choose median.

802. Feature selection methods for selecting the right variables for building efficient predictive models?

Ans. Some of the Feature selection techniques are: Information Gain, Chi-square test, Correlation Coefficient, Mean Absolute Difference (MAD), Exhaustive selection, Forward selection, Regularization.

803. Treat missing values?

Ans. They are:

1. List wise or case deletion

2. Pairwise deletion

3. Mean substitution

4. Regression imputation

5. Maximum likelihood.

804. **assumptions used in linear regression? What would happen if they are violated?**

Ans. 1. Linear relationship.

2. Multivariate normality.

3. no or little multicollinearity.

4. no auto-correlation.

5. Homoscedasticity.

Data to be analyzed by linear regression were sampled violate one or more of the linear regression assumptions, the results of the analysis may be incorrect or misleading.

805. How is the grid search parameter different from the random search tuning strategy?

Ans. Random search differs from grid search in that we no longer provide an explicit set of possible values for each hyperparameter; rather, we provide a statistical distribution for each hyperparameter from which values are sampled. Essentially, we define a sampling distribution for each hyperparameter to carry out a randomized search.

806. Is it good to do dimensionality reduction before fitting a Support Vector Model?

Ans. Support Vector Machine Learning Algorithm performs better in the reduced space. It is beneficial to perform dimensionality reduction before fitting an SVM if the number of features is large when compared to the number of observations.

807. ROC Curve?

Ans ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds.

808. How to identify a cause vs. a correlation? Give examples.

Ans. While causation and correlation can exist at the same time, correlation does not imply causation. Causation explicitly applies to cases where action A causes outcome B. On the other hand, correlation is simply a relationship. Correlation between Ice cream sales and sunglasses sold. As the sales of ice creams is increasing so do the sales of sunglasses. Causation takes a step further than correlation.

809. precision, accuracy and recall?

Ans. The recall is the ratio of the relevant results returned by the search engine to the total number of the relevant results that could have been returned. The precision is the proportion of relevant results in the list of all returned search results. Accuracy is the measurement used to determine which model is best at identifying relationships and patterns between variables in a dataset based on the input, or training, data.

810. choose k in k-means?

Ans. There is a popular method known as elbow method which is used to determine the optimal value of K to perform the K-Means Clustering Algorithm. The basic idea behind this method is that it plots the various values of cost with changing k. As the value of K increases, there will be fewer elements in the cluster.

811. word2vec methods?

Ans. Word2vec is a technique for natural language processing published in 2013. The word2vec algorithm uses a neural network model to learn word associations from a large corpus of text. Once trained, such a model can detect synonymous words or suggest additional words for a partial sentence.

812. Pruning in case of decision trees?

Ans. Pruning is a data compression technique in machine learning and search algorithms that reduces the size of decision trees by removing sections of the tree that are non-critical and redundant to classify instances.

813. Central tendency

A. Central tendency is defined as “the statistical measure that identifies a single value as representative of an entire distribution.” It aims to provide an accurate description of the entire data. It is the single value that is most typical/representative of the collected data.

814. Chi-Square test

A. A chi-square test is a statistical test used to compare observed results with expected results. The purpose of this test is to determine if a difference between observed data and expected data is due to chance, or if it is due to a relationship between the variables you are studying.

815. A/B testing

A. A/B testing is an optimisation technique often used to understand how an altered variable affects audience or user engagement. It's a common method used in marketing, web design, product development, and user experience design to improve campaigns and goal conversion rates.

816. Outlier treatment method

A. Removing outliers is legitimate only for specific reasons. Outliers can be very informative about the subject-area and data collection process. If the outlier does not change the results but does affect assumptions, you may drop the outlier. Or just trim the data set, but replace outliers with the nearest "good" data, as opposed to truncating them completely.

817. ANOVA test

A. Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples.

818. Cross validation

A. Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation.

819. How will you work in a machine learning project if there is a huge imbalance in the data

A. Follow these techniques:

*Use the right evaluation metrics.

*Use K-fold Cross-Validation in the right way.

*Ensemble different resampled datasets.

*Resample with different ratios.

*Cluster the abundant class.

*Design your own models.

820. Formula of sigmoid function

A. It is a mathematical function having a characteristic that can take any real value and map it to between 0 to 1 shaped like the letter "S".

Y = 1 / 1+(e*-z)

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

A. The Gini Index and the Entropy and Information gain metrics are the metrics to use in the algorithm to create a decision tree

822. What is ensemble learning

A. Ensemble methods is a machine learning technique that combines several base models in order to produce one optimal predictive model.

823. Do gradient descent methods always converge to similar points?

Ans. Gradient Descent need not always converge at global minimum. It all depends on following conditions; If the line segment between any two points on the graph of the function lies above or on the graph then it is convex function.

824. For the given points, how will you calculate the euclidean distance in python?

Ans. Example:

```
import numpy as np

# initializing points in

# numpy arrays

point1 = np.array((1, 2, 3))

point2 = np.array((1, 1, 1))

 # finding sum of squares

sum_sq = np.sum(np.square(point1 - point2))

 # Doing squareroot and # printing Euclidean distance

print(np.sqrt(sum_sq))
```

825. What is a box-cox transformation?

Ans. A Box Cox transformation is a transformation of non-normal dependent variables into a normal shape.

826. What is collaborative filtering? And a content based?

Ans. Collaborative filtering filters information by using the interactions and data collected by the system from other users. It's based on the idea that people who agreed in their evaluation of certain items are likely to agree again in the future. Content-based filtering uses item features to recommend other items similar to what the user likes, based on their previous actions or explicit feedback.

827. What is pooling on cnn, and how does it work?

Ans. Pooling layers are used to reduce the dimensions of the feature maps. Thus, it reduces the number of parameters to learn and the amount of computation performed in the network. The pooling layer summarises the features present in a region of the feature map generated by a convolution layer.

828. What are recurrent neural networks (rnns)?

Ans. Recurrent Neural Networks (RNN) are a class of Artificial Neural Networks that can process a sequence of inputs in deep learning and retain its state while processing the next sequence of inputs. Traditional neural networks will process an input and move onto the next one disregarding its sequence.

829. What is dropout and batch normalization?

Ans. Dropout is a technique where randomly selected neurons are ignored during training. Batch normalization is a technique for training very deep neural networks that standardizes the inputs to a layer for each mini-batch. This has the effect of stabilizing the learning process and dramatically reducing the number of training epochs required to train deep networks.

830. Fix multi-collinearity in a regression model?

Ans. Follow these methods:

Remove some of the highly correlated independent variables.

Linearly combine the independent variables, such as adding them together.

Perform an analysis designed for highly correlated variables, such as principal components analysis or partial least squares regression.

831. Can you write a user-defined function in SQL? how?

Ans. SQL Server allows us to create our functions called as user defined functions in SQL Server. For example, if we want to perform some complex calculations, then we can place them in a separate function, and store it in the database. Whenever we need the calculation, we can call it.

832. Hash tables? Where do you use it?

Ans. A hash table is a data structure that is used to store keys/value pairs. It uses a hash function to compute an index into an array in which an element will be inserted or searched. They are widely used in many kinds of computer software, particularly for associative arrays, database indexing, caches and sets. The idea of a hash table is to provide a direct access to its items. So that is why the it calculates the "hash code" of the key and uses it to store the item, instead of the key itself.

833. in a scenario, what would you prioritize: bias or variance?

Ans. Bias is an error between the actual values and the model's predicted values. Variance is also an error but from the model's sensitivity to the training data. A prioritization of Bias over Variance will lead to a model that overfits the data. Prioritizing Variance will have a model underfit the data.

834. How would you remove bias?

Ans. Ways to minimize Bias in ML:

Choose the correct learning model. There are two types of learning models, and each has its own pros and cons.

Use the right training dataset.

Perform data processing mindfully.

Monitor real-world performance across the ML lifecycle.

Make sure that there are no infrastructural issues.

835. In what situation would you consider mean over median

Ans. If your data contains outliers, then you would typically rather use the median because otherwise the value of the mean would be dominated by the outliers rather than the typical values. In conclusion, if you are considering the mean, check your data for outliers, if any then better choose median.

836. LASSO & RIDGE?

Ans. Ridge and Lasso regression are some of the simple techniques to reduce model complexity and prevent over-fitting which may result from simple linear regression . Ridge Regression : In ridge regression, the cost function is altered by adding a penalty equivalent to square of the magnitude of the coefficients. lasso (least absolute shrinkage and selection operator; also Lasso or LESSONS) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model.

837. Optimization?

Ans. Optimization is the problem of finding a set of inputs to an objective function that results in a maximum or minimum function evaluation. It is the challenging problem that underlies many machine learning algorithms, from fitting logistic regression models to training artificial neural networks.

838. What is standard error of the mean?

Ans. The standard error of mean tells you how accurate the mean of any given sample from that population is likely to be compared to the true population mean. When the standard error increases, i.e. the means are more spread out, it becomes more likely that any given mean is an inaccurate representation of the true population mean.

839. What is bias?

Ans. The bias is known as the difference between the prediction of the values by the ML model and the correct value. Being high in biasing gives a large error in training as well as testing data. Its recommended that an algorithm should always be low biased to avoid the problem of underfitting.

840. Normal Distribution? Skewed distribution? Solution?

Ans. The normal distribution is a continuous probability distribution that is symmetrical on both sides of the mean, so the right side of the center is a mirror image of the left side. The area under the normal distribution curve represents probability and the total area under the curve sums to one. If one tail is longer than another, the distribution is skewed. These distributions are sometimes called asymmetric or asymmetrical distributions as they don't show any kind of symmetry. Symmetry means that one half of the distribution is a mirror image of the other half.

Dealing with skew data: log transformation: transform skewed distribution to a normal distribution:

1. Remove outliers.

2. Normalize (min-max)

3. Cube root: when values are too large.

4. Square root: applied only to positive values.

5. Reciprocal.

6. Square: apply on left skew.

841. capture the correlation between continuous and categorical variable? How?

Ans. There are three big-picture methods to understand if a continuous and categorical are significantly correlated - point biserial correlation, logistic regression, and Kruskal Wallis H Test. The point biserial correlation coefficient is a special case of Pearson's correlation coefficient.

842. variance? Is it good or bad in data?

Ans. Variance refers to the changes in the model when using different portions of the training data set. Simply stated, variance is the variability in the model prediction—how much the ML function can adjust depending on the given data set.

843. error and a residual error?

Ans. The error (or disturbance) of an observed value is the deviation of the observed value from the (unobservable) true value of a quantity of interest (for example, a population mean), and the residual of an observed value is the difference between the observed value and the estimated value of the quantity of interest.

844. select an appropriate value of k in k-means?

Ans. There is a popular method known as elbow method which is used to determine the optimal value of K to perform the K-Means Clustering Algorithm. The basic idea behind this method is that it plots the various values of cost with changing k. As the value of K increases, there will be fewer elements in the cluster.

845. Meaning when p values are high or low?

Ans. High p-values indicate that your evidence is not strong enough to suggest an effect exists in the population. An effect might exist but it's possible that the effect size is too small, the sample size is too small, or there is too much variability for the hypothesis test to detect it. The smaller the p-value, the stronger the evidence that you should reject the null hypothesis.

846. Difference between expected and mean value

Ans. While mean is the simple average of all the values, expected value of expectation is the average value of a random variable which is probability-weighted.

847. How time series problems different from regression problem

Ans. Regression is Intrapolation. Time-series refers to an ordered series of data. Time-series models usually forecast what comes next in the series - much like our childhood puzzles where we extrapolate and fill patterns.

848. RoC curve

Ans. An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: True Positive Rate. False Positive Rate.

849. Random forest or multiple decision trees. Which is better?

Ans. Random Forest is suitable for situations when we have a large dataset, and interpretability is not a major concern. Decision trees are much easier to interpret and understand. Since a random forest combines multiple decision trees, it becomes more difficult to interpret.

850. Example when false positive is important than false negative

Ans. A false positive is where you receive a positive result for a test, when you should have received a negative results. Some examples of false positives: A pregnancy test is positive, when in fact you aren't pregnant. A cancer screening test comes back positive, but you don't have the disease. Innocent party is found guilty in such cases

851. What is a sensitivity analysis in the decision making process?

Ans. Sensitivity analysis is a method for predicting the outcome of a decision if a situation turns out to be different compared to the key predictions. It helps in assessing the riskiness of a strategy. Helps in identifying how dependent the output is on a particular input value.

852. How do you interpret the data using statistical techniques

Ans. Most Important Methods For Statistical Data Analysis Mean.

1. Standard Deviation.

2. Regression.

3. Sample Size.

4. Determination.

5. Hypothesis Testing.

853. KNN imputation method?

Ans. The idea in kNN methods is to identify 'k' samples in the dataset that are similar or close in the space. Then we use these 'k' samples to estimate the value of the missing data points. Each sample's missing values are imputed using the mean value of the 'k'-neighbors found in the dataset.

854. Map Reduce?

Ans. MapReduce facilitates concurrent processing by splitting petabytes of data into smaller chunks, and processing them in parallel on Hadoop commodity servers. In the end, it aggregates all the data from multiple servers to return a consolidated output back to the application.

855. What is a Pivot Table?

Ans. A pivot table is a table of grouped values that aggregates the individual items of a more extensive table within one or more discrete categories.

856. difference between 1-Sample T-test, and 2-Sample T-test?

Ans. The 2-sample t-test takes your sample data from two groups and boils it down to the t-value. The process is very similar to the 1-sample t-test, and you can still use the analogy of the signal-to-noise ratio. Unlike the paired t-test, the 2-sample t-test requires independent groups for each sample.

857. variance and covariance difference?

Ans. Variance and covariance are mathematical terms frequently used in statistics and probability theory. Variance refers to the spread of a data set around its mean value, while a covariance refers to the measure of the directional relationship between two random variable.

858. What is R2? What are some other metrics that could be better than R2 and why?

Ans. R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. R-squared does not measure goodness of fit. R-squared does not measure predictive error. R-squared does not allow you to compare models using transformed responses. R-squared does not measure how one variable explains another. Some better metrics that could be better than R2 are:

Mean Squared Error (MSE).

Root Mean Squared Error (RMSE).

Mean Absolute Error (MAE)

859. What is the curse of dimensionality?

Ans. The curse of dimensionality basically means that the error increases with the increase in the number of features. It refers to the fact that algorithms are harder to design in high dimensions and often have a running time exponential in the dimensions.

860. What are advantages of plotting your data before performing analysis ?

Ans. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

861. How would you explain a confidence interval to an engineer with no statistics background? What does 95% confidence mean?

Ans. A 95% confidence interval, for example, implies that were the estimation process repeated again and again, then 95% of the calculated intervals would be expected to contain the true parameter value.

862. How do you deal with some of your predictors being missing?

Ans. Simple approaches include taking the average of the column and use that value, or if there is a heavy skew the median or mode might be better. A better approach, you can perform regression or nearest neighbor imputation on the column to predict the missing values. Then continue on with your analysis/mode

863. How can you assess a good logistic model?

A. Measuring the performance of Logistic Regression:

1. One can evaluate it by looking at the confusion matrix and count the misclassifications (when using some probability value as the cutoff)

2. One can evaluate it by looking at statistical tests such as the Deviance or individual Z-scores.

864. How Regularly Must an Algorithm be Updated?

A. It can vary time to time depending upon number of updates happened in the algorithm as per the requirement.

865. Why Is Resampling Done?

A. Resampling methods are used to ensure that the model is good enough and can handle variations in data. The model does that by training it on the variety of patterns found in the dataset.

866. Write the formula to calculate R-square?

A. R^2 = 1 - (RSS/TSS) where RSS = sum of squares of residual and TSS = Total sum of squares

867. Goal of A/B Testing?

A. A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

868. Do Gradient Descent Methods at All-Time Converge to a Similar Point?

A. No, they always don't. That's because in some cases it reaches a local minima or a local optima point.

869. Feature Vectors?

A. A feature vector is a vector containing multiple elements about an object. Putting feature vectors for objects together can make up a feature space.

870. If p value is more than 0.05 what does it mean?

A. A p-value higher than 0.05 (> 0.05) is not statistically significant and indicates strong evidence for the null hypothesis. This means we retain the null hypothesis and reject the alternative hypothesis.

871. Difference between R square and Adjusted R Square.

Ans. One main difference between R2 and the adjusted R2: R2 assumes that every single variable explains the variation in the dependent variable. The adjusted R2 tells you the percentage of variation explained by only the independent variables that actually affect the dependent variable.

872. Difference between Precision and Recall.

Ans. When it comes to precision we're talking about the true positives over the true positives plus the false positives. As opposed to recall which is the number of true positives over the true positives and the false negatives.

873. 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. The fourth one is normality.

874. Difference between Random Forest and Decision Tree.

Ans. A decision tree combines some decisions, whereas a random forest combines several decision trees. Thus, it is a long process, yet slow. Whereas, a decision tree is fast and operates easily on large data sets, especially the linear one. The random forest model needs rigorous training.

875. How does K-means work?

Ans. K-means clustering uses “centroids”, K different randomly-initiated points in the data, and assigns every data point to the nearest centroid. After every point has been assigned, the centroid is moved to the average of all of the points assigned to it.

876. How do you generally choose among different classification models to decide which one is performing the best?

Ans. Here are some important considerations while choosing an algorithm:

Size of the training data, Accuracy and/or Interpretability of the output, Speed or Training time, Linearity and number of features.

877. How do you perform feature selection?

Ans. Unsupervised: Do not use the target variable (e.g. remove redundant variables). Correlation.

Supervised: Use the target variable (e.g. remove irrelevant variables). Wrapper: Search for well-performing subsets of features. RFE.

878. What is an intercept in a Linear Regression? What is its significance?

Ans. The intercept (often labeled as constant) is the point where the function crosses the y-axis. In some analysis, the regression model only becomes significant when we remove the intercept, and the regression line reduces to Y = b*X + error. The intercept (often labeled the constant) is the expected mean value of Y when all X="0. Start with a regression equation with one predictor, X. If X sometimes equals 0, the intercept is simply the expected mean value of Y at that value. If X never equals 0, then the intercept has no intrinsic meaning

879. How to find the multi collinearity in the data set

Ans. A simple method to detect multicollinearity in a model is by using something called the variance inflation factor or the VIF for each predicting variable.

880. Explain the difference ways to treat multi collinearity!

Ans. 1. Remove some of the highly correlated independent variables.

2. Linearly combine the independent variables, such as adding them together.

3. Perform an analysis designed for highly correlated variables, such as principal components analysis or partial least squares regression.

881. How you decide which feature to keep and which feature to eliminate after performing multi collinearity test?

Ans. It is advisable to get rid of variables iteratively. We would begin with a variable with the highest VIF score since other variables are likely to capture its trend. As a result of removing this variable, other variables’ VIF values are likely to reduce.

882. Explain logistic regression

Ans. Logistic regression is a statistical analysis method used to predict a data value based on prior observations of a data set. A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables.

883. P value and its significance in statistical testing?

Ans. In statistics, the p-value is the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct. A smaller p-value means that there is stronger evidence in favor of the alternative hypothesis.

884. If through training all the features in the dataset, an accuracy of 100% is obtained but with the validation set, the accuracy score is 75%. What should be looked out for?

A. Training accuracy is much higher than validation accuracy, proving that it's the case of overfitting, so in this case, try regularization or making less complex model or any other method to avoid overfitting.

885. How is skewness different from kurtosis?

A. Skewness basically measures the asymmetry in data. Kurtosis on the other hand, measures the bulge / peak of a distribution curve.

886. How to calculate the accuracy of a binary classification algorithm using its confusion matrix?

A. Accuracy is calculated as the total number of two correct predictions (TP + TN) divided by the total number of a dataset (P + N).

887. How will you measure the Euclidean distance between the two arrays in numpy?

A. eucl_distance = np. linalg. norm(point_a - point_b) where np stands for numpy.

888. In a survey conducted, the average height was 164cm with a standard deviation of 15cm. If Alex had a z-score of 1.30, what will be his height?

A. Alex's height = 164 + 1.30*15 = 183.5 cm.

889. Kernels in SVM?

A. Kernel Function is a method used to take data as input and transform into the required form of processing data. “Kernel” is used due to set of mathematical functions used in Support Vector Machine provides the window to manipulate the data.

890. Overfitting and Underfitting? How do you handle them?

A. Your model is underfitting the training data when the model performs poorly on the training data. Your model is overfitting your training data when you see that the model performs well on the training data but does not perform well on the evaluation data.

To avoid overfitting: Keep your model simple, Use regularization technique, Use cross-validation. To avoid underfitting: Decrease regularization, increase the training duration.

891. P value? Why threshold is 0.05 or less?

A. A p-value is a measure of the probability that an observed difference could have occurred just by random chance. The lower the p-value, the greater the statistical significance of the observed difference. P-value can be used as an alternative to or in addition to pre-selected confidence levels for hypothesis testing. A p-value less than 0.05 is statistically significant. It indicates strong evidence against the null hypothesis, as there is less than a 5% probability the null is correct (and the results are random). Therefore, we reject the null hypothesis, and accept the alternative hypothesis.

892. Regularization and its use?

A. Regularization is a technique used for tuning the function by adding an additional penalty term in the error function. Regularization, significantly reduces the variance of the model, without substantial increase in its bias.

893. Type 1 and type 2 error in confusion matrix?

A. A type I error is the mistaken rejection of the null hypothesis (also known as a "false positive" finding or conclusion; example: "an innocent person is convicted"), while a type II error is the mistaken acceptance of the null hypothesis (also known as a "false negative" finding or conclusion; example: "a guilty person is not convicted").

894. How root node is selected in decision tree.

A. While building the decision tree, we would prefer choosing the attribute/feature with the least Gini index as the root node.

895. Std deviation vs variance?

A. 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 is expressed in the same units as the original values (e.g., minutes or meters).

Variance is expressed in much larger units (e.g., meters squared).

Variance helps to find the distribution of data in a population from a mean, and standard deviation also helps to know the distribution of data in population, but standard deviation gives more clarity about the deviation of data from a mean

896. Explain the difference between an array and a linked list.

A. An array is a collection of elements of a similar data type. A linked list is a collection of objects known as a node where node consists of two parts, i.e., data and address. Array elements store in a contiguous memory location. Linked list elements can be stored anywhere in the memory or randomly stored.

897. How do you ensure you are not overfitting a model?

A. Keep your model simple. Use regularization technique. Use cross-validation.

898. What are hyperparameters? How do they differ from model parameters?

A. Model Parameters: These are the parameters in the model that must be determined using the training data set. These are the fitted parameters. Hyperparameters: These are adjustable parameters that must be tuned in order to obtain a model with optimal performance.

899. You are told that your regression model is suffering from multicollinearity. How do verify this is true and build a better model?

A. A simple method to detect multicollinearity in a model is by using something called the variance inflation factor or the VIF for each predicting variable. We can follow these steps in order to build a better model:

Remove some of the highly correlated independent variables.

Linearly combine the independent variables, such as adding them together.

Perform an analysis designed for highly correlated variables, such as principal components analysis or partial least squares regression.

900. You build a random forest model with 10,000 trees. Training error as at 0.00, but validation error is 34.23. Explain what went wrong.

A. It means that the model has mimicked the training pattern perfectly that it will cause overfitting problem in test samples. To avoid this overfitting, use techniques like less complex model or cross validation etc.

901. What is the recall, specificity and precision of the confusion matrix?

A. The recall is the ratio of the relevant results returned by the search engine to the total number of the relevant results that could have been returned. Specificity (SP) is calculated as the number of correct negative predictions divided by the total number of negatives. The precision is the proportion of relevant results in the list of all returned search results.

902. Difference between WHERE and HAVING in SQL

A. The main difference between them is that the WHERE clause is used to specify a condition for filtering records before any groupings are made, while the HAVING clause is used to specify a condition for filtering values from a group.

903. Explain confusion matrix ?

A. 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.

904. Explain PCA

A. The principal components are eigenvectors of the data's covariance matrix. Thus, the principal components are often computed by eigen decomposition of the data covariance matrix or singular value decomposition of the data matrix. PCA is the simplest of the true eigenvector-based multivariate analyses and is closely related to factor analysis.

905. How do you cut a cake into 8 equal parts using only 3 straight cuts ?

A. Cut the cake from middle first, then pile up the one piece on another, and then again cut it straight from the middle which will leave you with 4 pieces. Finally, put all the 4 pieces on one another, and cut it for the third time. This is how with 3 straight cuts, you can cut cake into 8 equal pieces.

906. Explain kmeans clustering

A. K-means clustering aims to partition data into k clusters in a way that data points in the same cluster are similar and data points in the different clusters are farther apart. Similarity of two points is determined by the distance between them.

907. How is KNN different from k-means clustering?

A. K-means clustering represents an unsupervised algorithm, mainly used for clustering, while KNN is a supervised learning algorithm used for classification.

908. Stock market prediction: You would like to predict whether or not a certain company will declare bankruptcy within the next 7 days (by training on data of similar companies that had previously been at risk of bankruptcy). Would you treat this as a classification or a regression problem?

A. It is a classification problem.

909. Heteroscedasticity? How to detect it?

A. Heteroskedasticity refers to situations where the variance of the residuals is unequal over a range of measured values. When running a regression analysis, heteroskedasticity results in an unequal scatter of the residuals (also known as the error term). To check for heteroscedasticity, you need to assess the residuals by fitted value plots specifically. Typically, the telltale pattern for heteroscedasticity is that as the fitted values increases, the variance of the residuals also increases.

910. Root Cause Analysis?

A. Root cause analysis (RCA) is defined as a collective term that describes a wide range of approaches used to uncover causes of problems. Some RCA approaches are geared more toward identifying true root causes than others, some are more general problem-solving techniques, and others simply offer support for the core activity of root cause analysis.

911. Difference between array and list

A. The main difference between these two data types is the operation you can perform on them. Lists are containers for elements having differing data types but arrays are used as containers for elements of the same data type.

912. Which is faster dictionary or list for look up

A. Dictionary is faster because you used a better algorithm. The reason is because a dictionary is a lookup, while a list is an iteration. Dictionary uses a hash lookup, while your list requires walking through the list until it finds the result from beginning to the result each time.

913. How much time SVM takes to complete if 1 iteration takes 10sec for 1st class.

And there are 4 classes.

A. It would take 4*10 = 40 seconds to train one-vs-all method one to one.

914. How would you build a model to predict credit card fraud?

A. Use Kaggle's Credit card fraud dataset, start with EDA (Exploratory Data Analysis). Applying train, test split over the data and then finally choosing any model like logistic regression, XGBoost or Random Forest. After Hyperparameter tuning and fitting the model, the final step would be evaluating its performance.

915. How would you derive new features from features that already exist?

A. Feature engineering is applied first to generate additional features, and then feature selection is done to eliminate irrelevant, redundant, or highly correlated features. This includes techniques like Binning, Data manipulation etc.

916. If you're attempting to predict a customer's gender, and you only have 100 data points, what problems could arise?

A. Overfitting because we might learn too much into some particular patterns within this small sample set so we lose generalization abilities on other datasets.

917. Suppose you were given two years of transaction history. What features would you use to predict credit risk?

A. Following are the features that can be used in such case.

Transaction amount,

Transaction count,

Transaction frequency,

transaction category: bar, grocery, jwery etc.

transaction channels: credit card, debit card, international wire transfer etc.

distance between transaction address and mailing address,

fraud/ risk score

918. Why does SVM need to maximize the margin between support vectors?

A. Our goal is to maximize the margin because the hyperplane for which the margin is maximum is the optimal hyperplane. Thus SVM tries to make a decision boundary in such a way that the separation between the two classes is as wide as possible in the plane.

919. You have been asked to evaluate a regression model based on R^2, adjusted R^2 and tolerance. What will be your criteria?

Tolerance (1 / VIF) is used as an indicator of multicollinearity. It is an indicator of percent of variance in a predictor which cannot be accounted by other predictors. Large values of tolerance are desirable.

We will consider adjusted R^2 as opposed to R^2 to evaluate model fit because R^2 increases irrespective of improvement in prediction accuracy as we add more variables.

But, adjusted R^2 would only increase if an additional variable improves the accuracy of model, otherwise stays same. It is difficult to commit a general threshold value for adjusted R^2 because it varies between data sets.

For example: a gene mutation data set might result in lower adjusted R^2 and still provide fairly good predictions, as compared to a stock market data where lower adjusted R^2 implies that model is not good.

920. How many iPhones are currently being used in India?

Clarify with the interviewers whether the question is about only a single version of the iPhone or all versions put together. Here, we shall assume that all iPhones put together are being talked about.

The first step toward solving this query will be segmentation. There are many ways in which India's population can be segmented. Here, we shall first assume that only people who have attained a working age and are under the age of retirement own an iPhone. Children and old citizens do not own an iPhone. This removes 20% of the population as children and 20% as senior citizens.

The next assumption will be that only the upper stratum of India's income range can afford an iPhone. This metric assumes that only 5% of the eligible citizens from the previous filter can own an iPhone.

Now, it is not necessary that every member of this upper stratum will own an iPhone. Other options, such as OnePlus, Samsung, etc., are also available. However, a fair assumption would be that 50% of the eligible population from the previous filter owns an iPhone.

Calculating the proportion of the population that owns an iPhone –

0.6 x 0.05 x 0.5 = 0.015

Total iPhones in India = 0.015 x 130 crore = 1.95 crore

921. I know that a linear regression model is generally evaluated using Adjusted R^2 or F value. How would you evaluate a logistic regression model?

We can use the following methods:

- Since logistic regression is used to predict probabilities, we can use AUC-ROC curve along with confusion matrix to determine its performance.

- Also, the analogous metric of adjusted R^2 in logistic regression is AIC. AIC is the measure of fit which penalizes model for the number of model coefficients. Therefore, we always prefer model with minimum AIC value.

- Null Deviance indicates the response predicted by a model with nothing but an intercept. Lower the value, better the model. Residual deviance indicates the response predicted by a model on adding independent variables. Lower the value, better the model.

922.Two candidates Aman and Mohan appear for a Data Science Job interview. The probability of Aman cracking the interview is 1/8 and that of Mohan is 5/12. What is the probability that at least one of them will crack the interview?

The probability of Aman getting selected for the interview is 1/8 P(A) = 1/8 The probability of Mohan getting selected for the interview is 5/12 P(B)=5/12

Now, the probability of at least one of them getting selected can be denoted at the Union of A and B, which means

P(A U B) =P(A)+ P(B) – (P(A ∩ B)) ………………………(1)

Where P(A ∩ B) stands for the probability of both Aman and Mohan getting selected for the job. To calculate the final answer, we first have to find out the value of P(A ∩ B) So, P(A ∩ B) = P(A) * P(B)

1/8 * 5/12

5/96

Now, put the value of P(A ∩ B) into equation 1

P(A U B) =P(A)+ P(B) – (P(A ∩ B))

1/8 + 5/12 -5/96

So, the answer will be 47/96.

923. We know that one hot encoding increases the dimensionality of a dataset, but label encoding doesn't. How?

- When we use one-hot encoding, there is an increase in the dimensionality of a dataset.
- The reason for the increase in dimensionality is that, for every class in the categorical variables, it forms a different variable.
- Example: Suppose, there is a variable 'Color.' It has three sub-levels as Yellow, Purple, and Orange. So, one hot encoding 'Color' will create three different variables as Color.Yellow, Color.Purple, and Color.Orange.
- In label encoding, the sub-classes of a certain variable get the value as 0 and 1. So, we use label encoding only for binary variables.
- This is the reason that one hot encoding increases the dimensionality of data and label encoding does not.

924. If your model suffers from low bias and high variance, which algorithm would you use to tackle it? Why?

The error of a model can either be of bias and/or variance. Very low bias but high variance indicates overfitting, as well as complexity of the model. By averaging these out, we can reduce the variance and increase the bias.

a) A bagging algorithm can handle the high variance. The dataset is randomly subsampled mm times and the model trained using each subsample. Then the models are averaged by averaging out the predictions of each mode.

b) By using the k-nearest neighbour algorithm, the trade-off between bias and variance can be achieved. The value of k is increased to increase the number of neighbours that contribute to the prediction, and this in turn increases the bias of the model.

c) By using the support vector machine algorithm, the trade-off can be achieved by increasing the C parameter that influences the number of violations of the margin allowed in the training data, and this in turn increases the bias but decreases the variance.

925. Three zebras are sitting on each corner of an equilateral triangle. Each zebra randomly picks a direction and only runs along the outline of the triangle to either opposite edge of the triangle. What is the probability that none of the zebras collide?

- Let's imagine all of the zebras on an equilateral triangle. They each have two options of directions to go in if they are running along the outline to either edge. Given the case is random, let's compute the possibilities in which they fail to collide.
- There are only really two possibilities. The zebras will either all choose to run in a clockwise direction or a counter-clockwise direction.

- Let's calculate the probabilities of each. The probability that every zebra will choose to go clockwise will be the product of each zebra choosing the clockwise direction. Given there are two choices (counterclockwise or clockwise), that would be 1/2 * 1/2 * 1/2 = 1/8

- The probability of every zebra going counter-clockwise is the same at 1/8. Therefore, if we sum up the probabilities, we get the correct probability of 1/4 or 25%

926. What is backpropagation? How does it work? Why do we need it?

The Backpropagation algorithm looks for the minimum value of the error function in weight space using a technique called the delta rule or gradient descent. The weights that minimize the error function is then considered to be a solution to the learning problem.

We need backpropagation because,

Calculate the error – How far is your model output from the actual output.

Minimum Error – Check whether the error is minimized or not.

Update the parameters – If the error is huge then, update the parameters (weights and biases). After that again check the error.

Repeat the process until the error becomes minimum.

Model is ready to make a prediction – Once the error becomes minimum, you can feed some inputs to your model and it will produce the output.

927. How would you predict who will renew their subscription next month? What data would you need to solve this? What analysis would you do? Would you build predictive models? If so, which algorithms?

Let's assume that we're trying to predict renewal rate for Netflix subscription. So our problem statement is to predict which users will renew their subscription plan for the next month.

Next, we must understand the data that is needed to solve this problem. In this case, we need to check the number of hours the channel is active for each household, the number of adults in the household, number of kids, which channels are streamed the most, how much time is spent on each channel, how much has the watch rate varied from last month, etc. Such data is needed to predict whether or not a person will continue the subscription for the upcoming month.

After collecting this data, it is important that you find patterns and correlations. For example, we know that if a household has kids, then they are more likely to subscribe. Similarly, by studying the watch rate of the previous month, you can predict whether a person is still interested in a subscription. Such trends must be studied.

The next step is analysis. For this kind of problem statement, you must use a classification algorithm that classifies customers into 2 groups:

Customers who are likely to subscribe next month

Customers who are not likely to subscribe next month

Would you build predictive models? Yes, in order to achieve this you must build a predictive model that classifies the customers into 2 classes like mentioned above.

Which algorithms to choose? You can choose classification algorithms such as Logistic Regression, Random Forest, Support Vector Machine, etc.

Once you've opted the right algorithm, you must perform model evaluation to calculate the efficiency of the algorithm. This is followed by deployment.

928. What are some ways I can make my model more robust to outliers?

There are several ways to make a model more robust to outliers, from different points of view (data preparation or model building). An outlier in the question and answer is assumed being unwanted, unexpected, or a must-be-wrong value to the human's knowledge so far (e.g. no one is 200 years old) rather than a rare event which is possible but rare.

Outliers are usually defined in relation to the distribution. Thus outliers could be removed in the pre-processing step (before any learning step), by using standard deviations (Mean +/- 2*SD), it can be used for normality. Or interquartile ranges Q1 - Q3, Q1 - is the "middle" value in the first half of the rank-ordered data set, Q3 - is the "middle" value in the second half of the rank-ordered data set. It can be used for not normal/unknown as threshold levels.

Moreover, data transformation (e.g. log transformation) may help if data have a noticeable tail. When outliers related to the sensitivity of the collecting instrument which may not precisely record small values, Winsorization may be useful. This type of transformation has the same effect as clipping signals (i.e. replaces extreme data values with less extreme values). Another option to reduce the influence of outliers is using mean absolute difference rather mean squared error.

For model building, some models are resistant to outliers (e.g. tree-based approaches) or non-parametric tests. Similar to the median effect, tree models divide each node into two in each split. Thus, at each split, all data points in a bucket could be equally treated regardless of extreme values they may have.

929. Why do we need a validation set and a test set?

The data is split into three different categories while creating a model:

Training set: We use the training set for building the model and adjusting the model's variables. But we cannot rely on the correctness of the model build on top of the training set. The model might give incorrect outputs on feeding new inputs.

Validation set: We use a validation set to look into the model's response on top of the samples that don't exist in the training dataset. Then, we will tune hyperparameters on the basis of the estimated benchmark of the validation data.

When we are evaluating the model's response using the validation set, we are indirectly training the model with the validation set. This may lead to the overfitting of the model to specific data. So, this model won't be strong enough to give the desired response to the real-world data.

Test set: The test dataset is the subset of the actual dataset, which is not yet used to train the model. The model is unaware of this dataset. So, by using the test dataset, we can compute the response of the created model on hidden data. We evaluate the model's performance on the basis of the test dataset.

The model is always exposed to the test dataset after tuning the hyperparameters on top of the validation set.

As we know, the evaluation of the model on the basis of the validation set would not be enough. Thus, we use a test set for computing the efficiency of the model.

930. Difference between forecasting and prediction?

A forecast refers to a calculation or an estimation which uses data from previous events, combined with recent trends to come up a future event outcome.

On the other hand, a prediction is an actual act of indicating that something will happen in the future with or without prior information.

Accuracy: A Forecast is more accurate compared to a prediction. This is because forecasts are derived by analyzing a set of past data from the past and presents trends.

On the other hand, a prediction can be right or wrong. For example, if you predict the outcome of a football match, the result depends on how well the teams played no matter their recent performance or players.

Bias: Forecasting uses mathematical formulas and as a result, they are free from personal as well as intuition bias.

On the other hand, predictions are in most cases subjective and fatalistic in nature.

Quantification: When using a model to do a forecast, it's possible to come up with the exact quantity. For example, the World Bank uses economic trends, and the previous GDP values and other inputs to come up with a percentage value for a country economic growth.

However, when doing prediction, since there is no data for processing, one can only say the economy of a given country will grow or not.

Application: Forecasts are only applicable in the economic and meteorology field where there is a lot of information about the subject matter.

On the contrary, prediction can be applied anywhere as long as there is an expected future outcome.

931. What is Pruning in Decision Trees, and How Is It Done?

Pruning is a data compression technique in machine learning and search algorithms that reduces the size of decision trees by removing sections of the tree that are non-critical and redundant to classify instances. Pruning reduces the complexity of the final classifier, and hence improves predictive accuracy by the reduction of overfitting.

Pruning processes can be divided into two types (pre- and post-pruning).

Pre-pruning procedures prevent a complete induction of the training set by replacing a stop () criterion in the induction algorithm. Pre-pruning methods are considered to be more efficient because they do not induce an entire set, but rather trees remain small from the start

Post-pruning is the most common way of simplifying trees. Here, nodes and subtrees are replaced with leaves to reduce complexity.

The procedures are differentiated on the basis of their approach in the tree (top-down or bottom-up).

Top-down fashion: It will traverse nodes and trim subtrees starting at the root.

Bottom-up fashion: It will begin at the leaf nodes.

There is a popular pruning algorithm called reduced error pruning, in which starting at the leaves, each node is replaced with its most popular class. If the prediction accuracy is not affected, the change is kept.

932. How would you evaluate a logistic regression model?

Model Evaluation is a very important part in any analysis to answer the following questions,

How well does the model fit the data? Which predictors are most important? Are the predictions accurate?

So, the following are the criterion to access the model performance,

1. Akaike Information Criteria (AIC): In simple terms, AIC estimates the relative amount of information lost by a given model. So, the less information lost the higher the quality of the model. Therefore, we always prefer models with minimum AIC.

2. Receiver operating characteristics (ROC curve): ROC curve illustrates the diagnostic ability of a binary classifier. It is calculated/ created by plotting True Positive against False Positive at various threshold settings. The performance metric of ROC curve is AUC (area under curve). Higher the area under the curve, better the prediction power of the model.

3. Confusion Matrix: In order to find out how well the model does in predicting the target variable, we use a confusion matrix/ classification rate. It is nothing but a tabular representation of actual Vs predicted values which helps us to find the accuracy of the model.

933. If 70% of Facebook users on iOS use Instagram, but only 35% of Facebook users on Android use Instagram, how would you investigate the discrepancy?

There are a number of possible variables that can cause such a discrepancy that I would check to see:

The demographics of iOS and Android users might differ significantly. For example, according to Hootsuite, 43% of females use Instagram as opposed to 31% of men. If the proportion of female users for iOS is significantly larger than for Android then this can explain the discrepancy (or at least a part of it). This can also be said for age, race, ethnicity, location, etc.…

Behavioral factors can also have an impact on the discrepancy. If iOS users use their phones more heavily than Android users, it's more likely that they'll indulge in Instagram and other apps than someone who spent significantly less time on their phones.

Another possible factor to consider is how Google Play and the App Store differ. For example, if Android users have significantly more apps (and social media apps) to choose from, that may cause greater dilution of users.

Lastly, any differences in the user experience can deter Android users from using Instagram compared to iOS users. If the app is more buggy for Android users than iOS users, they'll be less likely to be active on the app.

934. What are Loss Function and Cost Functions? Explain the key Difference Between them?

When calculating loss we consider only a single data point, then we use the term loss function.

Whereas, when calculating the sum of error for multiple data then we use the cost function.

In other words, the loss function is to capture the difference between
the actual and predicted values for a single record whereas cost functions aggregate the difference for the entire training dataset.

The most commonly used loss functions are Mean-squared error and Hinge loss.

Mean-Squared Error(MSE): In simple words, we can say how our model predicted values against the actual values.

MSE = √(predicted value - actual value)2

Hinge loss: It is used to train the machine learning classifier, which is

L(y) = max(0,1- yy)

Where y = -1 or 1 indicating two classes and y represents the output form of the classifier. The most common cost function represents the total cost as the sum of the fixed costs and the variable costs in the equation y = mx + b.

935. What is Bayes' Theorem? How is it useful in a machine learning context?

Bayes' theorem, also known as Bayes' rule or Bayes' law, is a theorem in statistics that describes the probability of one event or condition as it relates to another known event or condition.

Mathematically, it's expressed as the true positive rate of a condition sample divided by the sum of the false positive rate of the population and the true positive rate of a condition.

Let's understand this theorem with an example:

For instance, say you had a 60% chance of actually having the flu after a flu test, but out of people who had the flu, the test will be false 50% of the time, and the overall population only has a 5% chance of having the flu. Would you actually have a 60% chance of having the flu after having a positive test?

Bayes' Theorem says no. It says that you have a (.6 * 0.05) (True Positive Rate of a Condition Sample) / (.6*0.05)(True Positive Rate of a Condition Sample) + (.5*0.95) (False Positive Rate of a Population) = 0.0594 or 5.94% chance of getting a flu.

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

Naive Bayes uses a similar method to predict the probability of different class based on various attributes. This algorithm is mostly used in text classification and with problems having multiple classes.

936. How will you estimate the number of weddings that take place in a year in India?

Facts

India's population in a year - 1.3 bill

Population breakup - Rural - 70% and Urban - 30%

Assumptions

Every year India's population would grow steadily, but the growth won't be very fast-paced.

Every man and women would be eventually married (homogeneously or heterogeneously). They won't prematurely die or prefer not to marry. People would be married only once.

In rural areas the age of marriage (in average) is between 15 - 35 year range. Similarly, in urban areas = 20 - 35 years. India is a young country, and 15 - 35 year range has around 50% of the total population.

Rural estimation

Rural population = 70% * 1.3 bill = 900 mill

Population within marriage age in a year = 50% * 900 mill = 450 mill

Number of marriages to happen = 450 / 2 = 225 mill marriages

These people will marry within a 20 year time period according to our assumptions.

Number of rural marriages in a year = 225 mill / 20 = 11.25 mill marriages

Urban estimation

Urban population = 30% * 1.3 bill = 400 mill

Population within marriage age in a year = 50% * 400 mill = 200 mill

Number of marriages to happen = 200 / 2 = 100 mill marriages

These people will marry within a 15 year time period according to our assumptions.

Number of urban marriages in a year = 100 mill / 15 = 6.6 mill marriages

Many people die in accidents prematurely, and won’t marry. In addition, most people don't marry as well as a consumer preference parameter. So, our market number is over-estimated. Even if we try to normalize it by introducing an error percentage of around 10%, the final number number will be lesser by around 10%-15%.

Answer = Approximately 14 million marriages occur in a year in India

937. What is the statistical power of sensitivity?

The statistical power of a study (sometimes called sensitivity) is how likely the study is to distinguish an actual effect from one of chance.

It’s the likelihood that the test is correctly rejecting the null hypothesis. For example, a study that has an 80% power means that the study has an 80% chance of the test having significant results.

A high statistical power means that the test results are likely valid. As the power increases, the probability of making a Type II error decreases.

A low statistical power means that the test results are questionable.

Statistical power helps you to determine if your sample size is large enough.

It is possible to perform a hypothesis test without calculating the statistical power. If your sample size is too small, your results may be inconclusive when they may have been conclusive if you had a large enough sample.

938. If the probability of seeing a shooting star in 15mins is 20%. What is the probability of seeing at least one shooting star in one hour?

Here it means, 20% probability = 20/100 = 1/5

Probability of Seeing a Star in 15 minutes = 1/5

Probability of not seeing a Star in 15 minutes = 1 - 1/5 = 4/5

Probability that you see at least one shooting star in the period of an hour

= 1 - Probability of not seeing any Star in 60 minutes

= 1 - Probability of not seeing any Star in 15 * 4 minutes

= 1 - $(4/5)^4$

= 1 - 0.4096

= 0.5904

So, the probability of seeing at least one shooting star in a period of an hour is 0.594

939. What is the difference between Rank and Dense Rank function?

An analytic function computes values over a group of rows and returns a single result for each row. This is different from an aggregate function, which returns a single result for a group of rows.

With analytic functions you can compute moving averages, rank items, calculate cumulative sums, and perform other analyses.

RANK gives you the ranking within your ordered position. Ties are assigned the same rank, with the next ranking(s) skipped.

DENSE_RANK again gives you the ranking within your ordered partition, but the ranks are consecutive. No ranks are skipped if there are ranks with multiple items.

Example below.. which is order partitioned on salary

Salary.	Rank.	Dense rank
1000	1	1
2000	2	2
2000	2	2
2000	2	2
3000	5	3

Due to next rankings skipped in the case of RANKS, generally DENSE_RANK is preferred as it gives proper ranking when we want to calculate nth highest salary

940. Difference Between Bagging and Boosting?

Bagging and Boosting are two types of Ensemble Learning, which helps to improve machine learning results by combining several models. This approach allows the production of better predictive performance compared to a single model.

So, let's understand the difference between Bagging and Boosting?

Bagging(Bootstrap aggregation): It is a homogeneous weak learners' model that learns from each other independently in parallel and combines them for determining the model average. Boosting: It is also a homogeneous weak learners' model. In this model, learners learn sequentially and adaptively to improve model predictions of a learning algorithm.

If the classifier is unstable (high variance), then we need to apply bagging. If the classifier is steady and straightforward (high bias), then we need to apply boosting.

In bagging, different training data subsets are randomly drawn with replacement from the entire training dataset. In boosting, every new subset contains the elements that were misclassified by previous models.

Bagging simplest way of combining predictions that belong to the same type. Boosting is the way of combining predictions that belong to the different types.

Each model is built independently for bagging. While in the case of boosting, new models are influenced by the performance of previously built models.

Bagging attempts to tackle the overfitting issue. Boosting tries to reduce bias.

Example: The Random Forest model uses Bagging. The AdaBoost uses Boosting techniques

941. How is oversampling different from under sampling?

Oversampling and under sampling are 2 important techniques used in machine learning – classification problems in order to reduce the class imbalance thereby increasing the accuracy of the model.

Classification is nothing but predicting the category of a data point to which it may probably belong by learning about past characteristics of similar instances. When the segregation of classes is not approximately equal then it can be termed as a “Class imbalance” problem. To solve this scenario in our data set, we use oversampling and under sampling.

Oversampling is used when the amount of data collected is insufficient. A popular over-sampling technique is SMOTE (Synthetic Minority Over-sampling Technique), which creates synthetic samples by randomly sampling the characteristics from occurrences in the minority class.

Conversely, if a class of data is the overrepresented majority class, under sampling may be used to balance it with the minority class. Under sampling is used when the amount of collected data is sufficient. Common methods of under sampling include cluster centroids and To meke links, both of which target potential overlapping characteristics within the collected data sets to reduce the amount of majority data.

Ex: Let's say in a bank majority of the customers are from a specific Race and very few customers are from other races , hence if the model is trained with this data , it is most likely that Model will reject the loan for Minority Race.

So, what should we do about it?

For, oversampling we increase the number of records belonging to the “minority race” category by duplicating its presence. So that the difference between the numbers of records belonging to both of the classes will narrow down.

Under sampling we reduce the number of records belonging to the “majority race”. The records for the deletion are selected strictly through a random process and are not influenced by any constraints or bias.

To conclude, over sampling is preferable as under sampling can result in the loss of important data. Under sampling is suggested when the amount of data collected is larger than ideal and can help data mining tools to stay within the limits of what they can effectively process.

942. What is multicollinearity and how to detect it in a dataset?

Let’s start with understanding correlation.

The correlation between two variables can be measured with a correlation coefficient which can range between -1 to 1. If the value is 0, the two variables are independent and there is no correlation. If the measure is extremely close to one of the extreme values, it indicates a linear relationship and highly correlated with each other. This means a change in one variable is associated with a significant change in other variables.

Multicollinearity is a condition when there is a significant dependency or association between the independent variables or the predictor variables. A significant correlation between the independent variables is often the first evidence of presence of multicollinearity.

How to test Multicollinearity?

Correlation matrix / Correlation plot: A correlation plot can be used to identify the correlation or bivariate relationship between two independent variables

Variation Inflation Factor (VIF): VIF is used to identify the correlation of one independent variable with a group of other variables.

Consider that we have 9 (assume V1 to V9) independent variables. To calculate the VIF of variable V1, we isolate the variable V1 and consider as the target variable and all the other variables(i.e V2 to V9) will be treated as the predictor variables.

We use all the other predictor variables and train a regression model and find out the corresponding R2 value.

Using this R2 value, we compute the VIF value gives as the image below.

It is always desirable to have VIF value as small as possible. A threshold is also set, which means that any independent variable greater than the threshold will have to be removed.

943. What is cross-validation and why would you use it?

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

In typical cross-validation, the training and validation sets must cross-over in successive rounds such that each data point has a chance of being validated against.

The basic form of cross-validation is k-fold cross-validation. Other forms of cross-validation are special cases of k-fold cross-validation or involve repeated rounds of k-fold cross-validation.

In k-fold cross-validation, the data is first partitioned into k equally (or nearly equally) sized segments or folds. Subsequently k iterations of training and validation are performed such that within each iteration a different fold of the data is held-out for validation while the remaining k − 1 folds are used for learning.

So, why do we use cross-validation?

It allows us to get more metrics and draw important conclusion about our algorithm and our data.

Helps to tune the hyper parameters of a given machine learning algorithm, to get good performance according to some suitable metric.

It mitigates overfitting while building a pipeline of models, such that second's models input will be real predictions on data that our first model never seen before.

K-fold cross validation also significantly reduces bias as we are using most of the data for fitting, and also significantly reduces variance as most of the data is also being used in validation set.

944. How to detect an anomaly in a distribution?

Anomaly detection is a technique used to identify unusual patterns that do not conform to expected behavior. It can be considered the thoughtful process of determining what is normal and what is not. Anomalies are also referred to as outliers, novelties, noise, exceptions and deviations. Simply, anomaly detection is the task of defining a boundary around normal data points so that they can be distinguishable from outliers.

Anomalies can be broadly categorized as:

Point anomalies: A single instance of data is anomalous if it's too far off from the rest. Business use case: Detecting credit card fraud based on "amount spent."

Contextual anomalies: The abnormality is context specific. This type of anomaly is common in time-series data. Business use case: Spending $100 on food every day during the holiday season is normal, but may be odd otherwise.

Collective anomalies: A set of data instances collectively helps in detecting anomalies. Business use case: Someone is trying to copy data from a remote machine to a local host unexpectedly, an anomaly that would be flagged as a potential cyber-attack.

The different types of methods for anomaly detection are as follows:

Simple Statistical Methods

The simplest approach to identifying irregularities in data is to flag the data points that deviate from common statistical properties of a distribution, including mean, median, mode, and quantiles. When an anomalous data point deviates by a certain standard deviation from the mean, then traversing mean over time-series data isn't exactly trivial, as it's not static. Thus, a rolling window to compute the average across the data points and it's intended to smooth short-term fluctuations and highlight long-term ones.

Machine Learning-Based Approaches for Anomaly Detection:

(a) Clustering-Based Anomaly Detection:

The approach focuses on unsupervised learning, similar data points tend to belong to similar groups or clusters, as determined by their distance from local centroids.

The k-means algorithm can be used which partition the dataset into a given number of clusters. Any data points that fall outside of these clusters are considered as anomalies.

(b) Density-based anomaly detection:

This approach is based on the K-nearest neighbors algorithm. It’s evident that normal data points always occur around a dense neighborhood and abnormalities deviate far away. To measure the nearest set of a data point, you can use Euclidean distance or similar measure according to the type of data you have.

(c) Support Vector Machine-Based Anomaly Detection:

A support vector machine is another effective technique for detecting anomalies. One-Class SVMs have been devised for cases in which one class only is known, and the problem is to identify anything outside this class.

This is known as novelty detection, and it refers to automatic identification of unforeseen or abnormal phenomena, i.e. outliers, embedded in a large amount of normal data.

Anomaly detections helps to monitor any data source, including user logs, devices, networks, and servers. This rapidly helps in identifying zero-day attacks as well as unknown security threats.