Q.What is the difference between WHERE & HAVING Clause?

- The WHERE clause is applied to individual rows before they are grouped, while the HAVING clause is applied to groups of rows after they are grouped.
- The WHERE clause can be used to filter rows based on any condition, while the HAVING clause can only be used to filter groups based on the results of aggregate functions (e.g., SUM, AVG, MAX, etc.).
- The WHERE clause can be used with any SELECT, UPDATE, or DELETE statement, while the HAVING clause can only be used with a SELECT statement.

Q.What is the concept of bagging and explain any bagging algorithm?

Bagging, also known as bootstrap aggregation, is the ensemble learning method that is commonly used to reduce variance within a noisy dataset. In bagging, a random sample of data in a training set is selected with replacement—meaning that the individual data points can be chosen more than once.

One well-known bagging algorithm is the Random Forest algorithm. In a Random Forest, multiple decision trees are trained on different random subsets of the data. The final prediction is made by averaging the predictions of all the individual decision trees. This approach helps to reduce the variance of the model and improve its generalization performance.

To create a Random Forest, you first need to decide on the number of decision trees you want to include in the ensemble. Then, for each decision tree, you need to:

- Select a random subset of the training data
- Train a decision tree on the selected subset of the data
- Make predictions using the trained decision tree
- Once all the decision trees have been trained and have made their predictions, the final prediction is made by averaging the predictions of all the individual decision trees. This can be done by taking the mean of the predictions for regression tasks or by voting for the most common prediction for classification tasks.

Q.What is Ensemble of Decision Trees?

An ensemble of decision trees is a method for creating a more powerful machine learning model by training multiple decision trees and combining their predictions. Decision trees are a popular method for both classification and regression tasks, but they can be prone to overfitting, particularly when the trees are deep and have many nodes. By training multiple decision trees and combining their predictions, the ensemble can often make more accurate predictions than any of the individual trees.

There are several different methods for creating an ensemble of decision trees, including bagging, boosting, and random forests.

Q. What is Random Forest algorithm?

Random forests are a type of ensemble machine learning method that can be used for classification and regression tasks. They are an extension of decision trees, which are a popular method for both classification and regression.

In a random forest, a large number of decision trees are trained on random subsets of the data, and the final prediction is made by averaging the predictions of all the trees. The use of multiple decision trees helps to reduce the risk of overfitting, which can be a problem with individual decision trees.

3.What is boosting bagging and stacking?

Very roughly, we can say that bagging will mainly focus at getting an ensemble model with less variance than its components whereas boosting and stacking will mainly try to produce strong models less biased than their components (even if variance can also be reduced).

4.What are bagging and boosting?

Bagging is a technique for reducing prediction variance by producing additional data for training from a dataset by combining repetitions with combinations to create multi-sets of the original data. Boosting is an iterative strategy for adjusting an observation's weight based on the previous classification.

5.What is naive in naive bayes theorem?

In machine learning, a naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. The independence assumptions are that the presence or absence of a particular feature of a class is unrelated to the presence or absence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, a naive Bayes classifier would consider all of these properties to independently contribute to the probability that this fruit is an apple.

The "naive" in naive Bayes comes from the assumption that the features in a dataset are all independent of each other, which is generally a strong assumption and not necessarily true in real-world data. However, despite this assumption, the algorithm can still perform very well in practice, especially when the assumption of independence holds.

6.What is cross-validation and its types ?

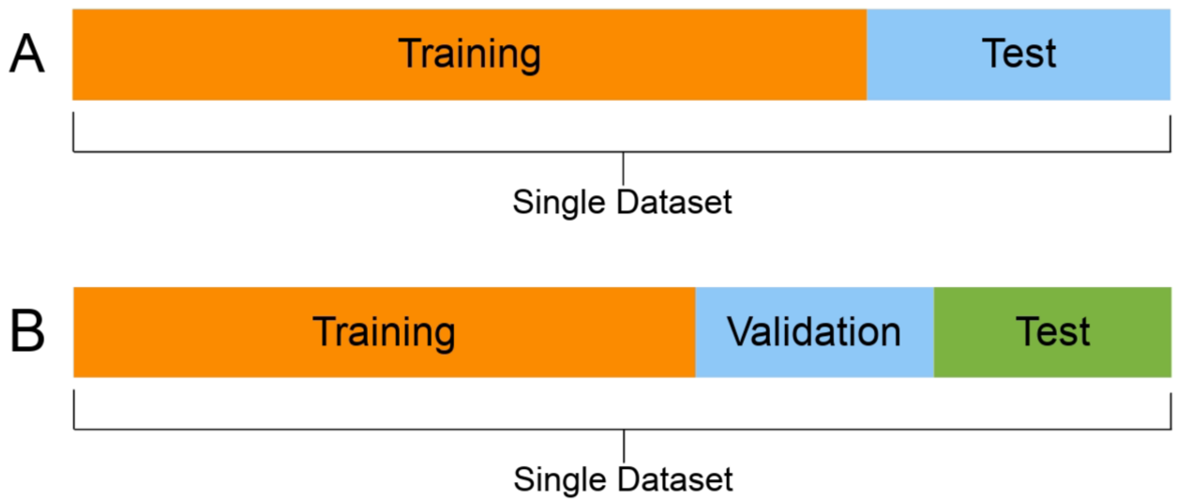
Definition. Cross-Validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to learn or train a model and the other used to validate the model.
7 most common types - Holdout, K-fold, Stratified k-fold, Rolling, Monte Carlo, Leave-p-out, and Leave-one-out method

7.What is train validation and test set?

Training Dataset: The sample of data used to fit the model

Validation Dataset: The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters.

Test Dataset: The sample of data used to provide an unbiased evaluation of a final model fit on the training dataset.



8.How much data will you allocate for your training validation and test sets?

The amount of data that you allocate for your training, validation, and test sets will depend on the size of your dataset and the specific needs of your model. Here are some general guidelines for how you might split your data:

Training set: This is the largest set of data and is used to train your model. A common split is to use 80% of the data for training and the remaining 20% for validation and test sets.

Validation set: This set of data is used to fine-tune your model by adjusting the hyperparameters. It is common to use 10-20% of the data for the validation set.

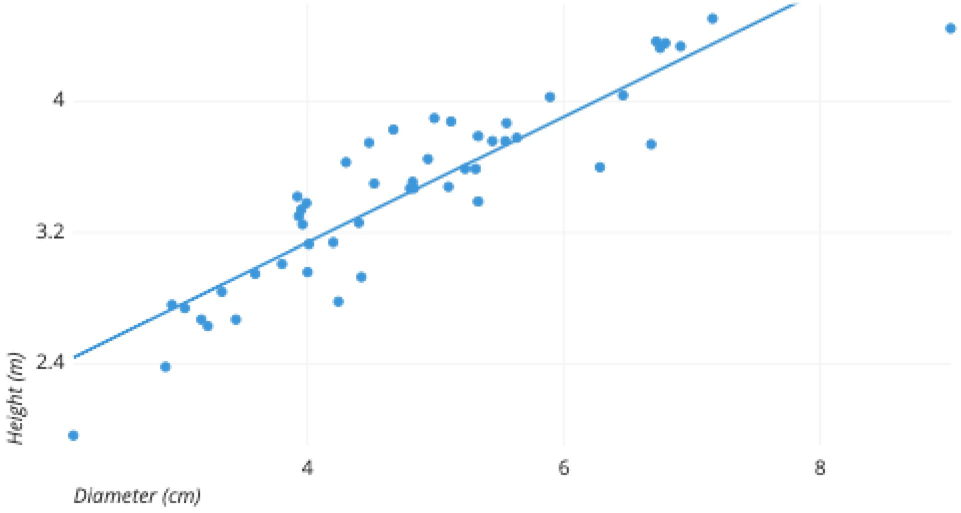
Test set: This is the final set of data that is used to evaluate the performance of the trained model. It is common to use 10-20% of the data for the test set.

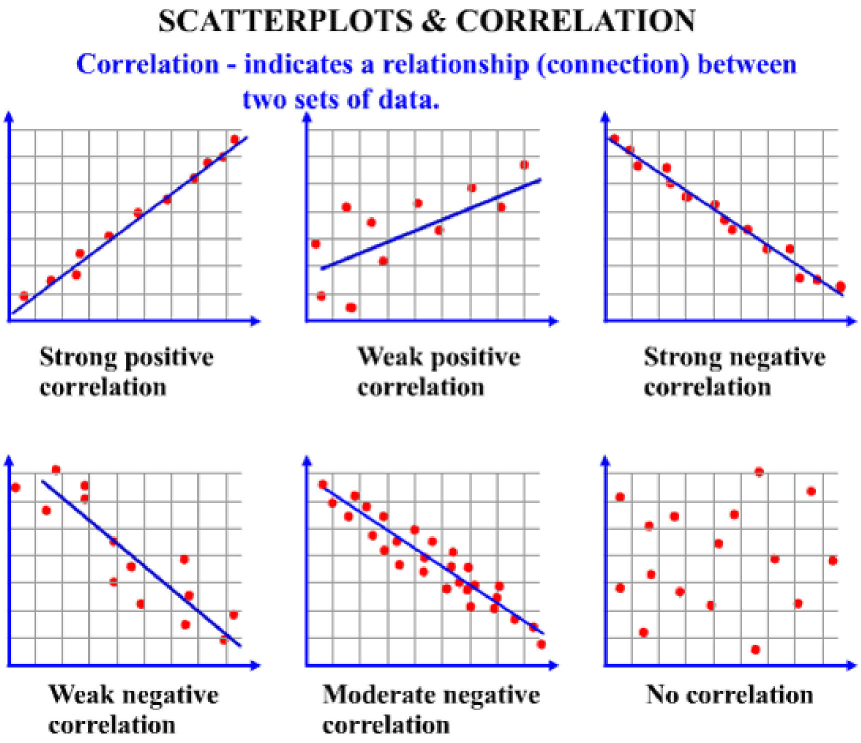
It is important to note that these splits are just general guidelines and the actual percentages may vary depending on the specific needs of your model and the size of your dataset.

In general, putting 80% of the data in the training set, 10% in the validation set, and 10% in the test set is a good split to start with. The optimum split of the test, validation, and train set depends upon factors such as the use case, the structure of the model, dimension of the data, etc

9.What is the purpose of the scatter plot?

The scatter plot explains the correlation between two attributes or variables. It represents how closely the two variables are connected and it is also used for detecting outliers.





SCATTERPLOTS & CORRELATION

Correlation - indicates a relationship (connection) between two sets of data.

Strong positive correlation

Weak positive correlation

Strong negative correlation

Weak negative correlation

Moderate negative correlation

No correlation

10.How do you create a drop down list in excel?

Create a drop-down list in excel:

A. Select the cells that you want to contain the lists.
B. On the ribbon, click DATA > Data Validation.
C. In the dialog, set Allow to List.
D. Click in Source, type the text or numbers (separated by commas, for a comma-delimited list) that you want in your drop-down list, and click OK.

11. What is RDBMS? How is it different from DBMS?

DBMS stands for Database Management System, and RDBMS is the acronym for the Relational Database Management system. In DBMS, the data is stored as a file, whereas in RDBMS, data is stored in the form of tables.
https://www.geeksforgeeks.org/difference-between-rdbms-and-dbms/

12. Write an SQL query to fetch the Emp_ID that are present in both tables

Emp_Details and Emp_Salary.
>>select Emp_ID.t1, Emp_ID.t2 from Emp_Details as t1, Emp_Salary as t2

13. Write an SQL query to fetch the Emp_ID and FullName of all the employees working under manager with ID = ‘986’.

>>select Emp_ID, FullName from Employee where Mng_ID = ‘986’

14. Fetch all the details of emp who is manager in position

15. What is the difference between tuple and set?

Tuple(round bracket) is immutable. In Sets {curly braces}, we cannot have repeated values. That means, we have unique values in Sets.

List, Set, and Dictionary are mutable.

The tuples refer to the collections of various objects of Python separated by commas between them. The sets are an unordered collection of data types.

16. Explain SMOTE in brief.

Synthetic Minority Oversampling Technique (SMOTE) is a statistical technique for increasing the number of cases in your dataset in a balanced way. The component works by generating new instances from existing minority cases that you supply as input using knn algorithm.

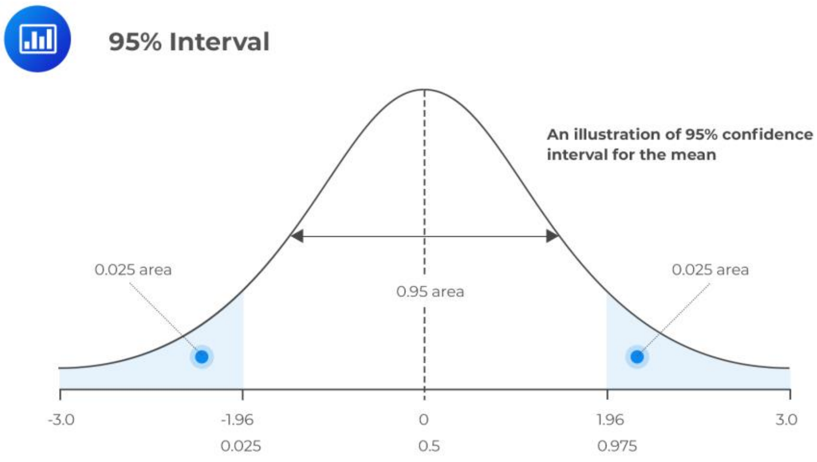
Q. What is Confidence Interval?

A confidence interval is a range of values that is calculated from a sample of data and is used to estimate a population parameter. It provides a way to express the uncertainty associated with estimating the population parameter based on the sample data.

For example, if you wanted to estimate the mean age of a population of people, you could take a sample of people and calculate the mean age of the sample. However, the mean age of the sample is likely to be different from the mean age of the population, due to sampling error. A confidence interval gives you a range of values that is likely to include the population mean, based on the mean and standard deviation of the sample and a specified level of confidence.

The level of confidence is typically expressed as a percentage, such as 95% or 99%. A higher level of confidence corresponds to a wider confidence interval, which means that you can be more certain that the population parameter is contained within the interval. However, a wider interval also means that there is more uncertainty about the exact value of the population parameter.

Confidence intervals are commonly used in statistical analysis to understand the precision of estimates and to make inferences about a population based on a sample. They are an important tool for understanding the limitations of statistical analysis and for communicating the uncertainty of statistical conclusions.



17. What does the 95% CI (Confidence Interval) mean?

The 95% confidence interval is a range of values that you can be 95% certain contains the true value of the population parameter you are estimating. For example, if you are estimating the mean of a population, the 95% confidence interval would be a range of values that you can be 95% certain contains the true mean of the population. This is a common way to represent the precision of an estimate.

To construct a 95% confidence interval, you first need to calculate the point estimate, which is the estimate of the population parameter based on your sample data. Then, you need to calculate the margin of error, which is a measure of the precision of your estimate. The margin of error is determined by the sample size, the level of confidence, and the standard deviation of the population.

To construct the confidence interval, you add the margin of error to the point estimate to get the upper bound of the interval, and subtract the margin of error from the point estimate to get the lower bound. The resulting range is the 95% confidence interval.
For example, if you are estimating the mean of a population and your point estimate is 100, with a margin of error of 10, the 95% confidence interval would be 90 to 110. This means that you can be 95% certain that the true mean of the population is within this range.

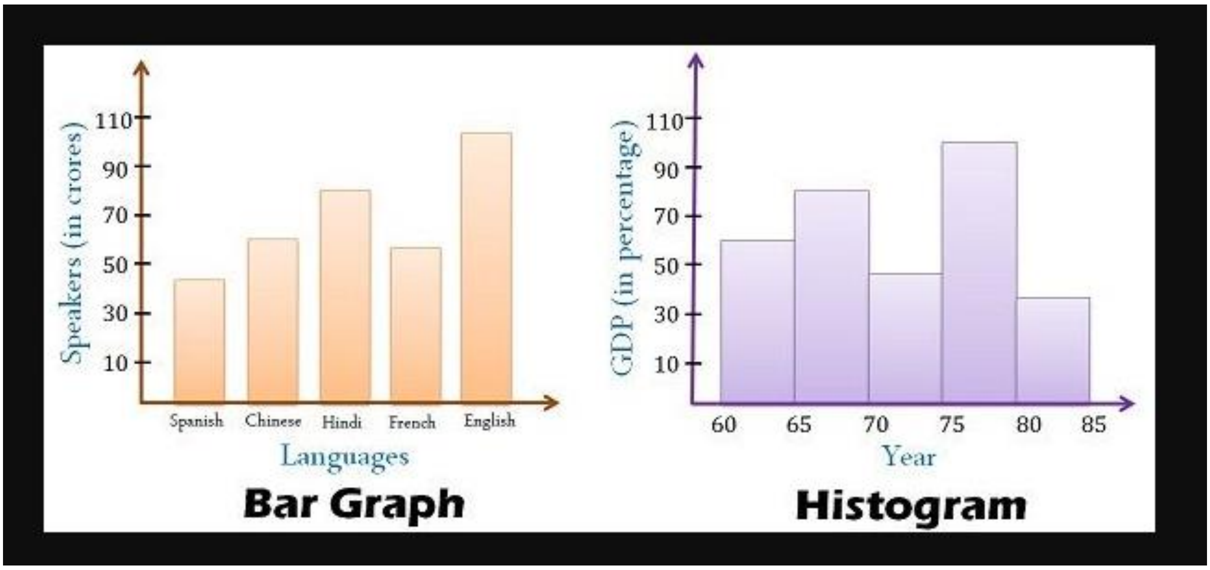
18. What is Polymorphism in Python?

In Python, polymorphism refers to the ability of a function or method to behave differently based on the number and/or type of its arguments. Polymorphism allows you to define a single function or method that can be used with multiple different types of objects, and the function or method will behave differently depending on the type of object it is called on.There are two main types of polymorphism in Python:

Ad-hoc polymorphism: This is achieved through function overloading, which allows you to define multiple functions with the same name but different signatures (number and/or type of arguments). Python does not support function overloading directly, but you can achieve a similar effect by using default arguments or by defining multiple functions with the same name and checking the number and/or type of the arguments in the function body.

Inheritance-based polymorphism: This is achieved through inheritance, which allows you to define a base class with one or more methods, and then define one or more derived classes that inherit from the base class. The derived classes can override the methods of the base class, which means that the same method name can behave differently depending on the type of object it is called on.

19. When will you use a Histogram and when will you use a bar chart. Explain with example.
Histograms visualize quantitative data or numerical data, whereas bar charts display categorical variables.



20. What are Pandas?

Pandas is an open-source library that is made mainly for working with relational or labeled data both easily and intuitively.
It provides various data structures and operations for manipulating numerical data and time series. This library is built on top of the NumPy library. Pandas is fast and it has high performance & productivity for users.

21. Define tokens in PostGRESQL

A token can be a key word, an identifier, a quoted identifier, a literal (or constant), or a special character symbol which is recognised by a parser.

Additionally, comments can occur in SQL input. They are not tokens, they are effectively equivalent to whitespace.

https://www.postgresql.org/docs/7.3/sql-syntax.html#:~:text=A%20token%20can%20be%20a,or%20a%20special%20character%20symbol.

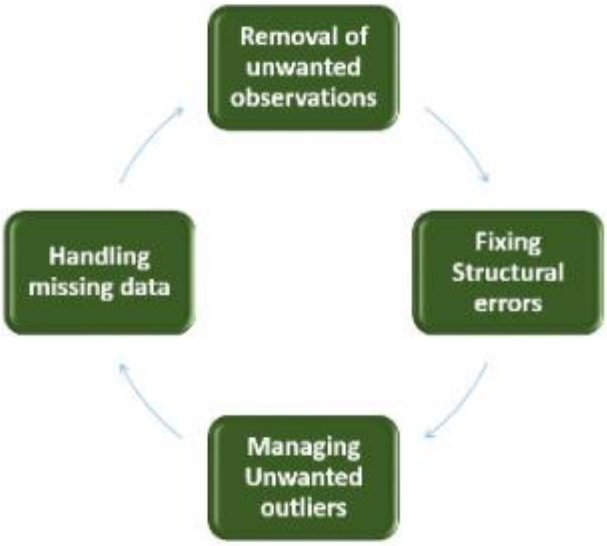
Tokens in PostgreSQL are the building blocks of any source code. They are known to comprise many of the special character symbols.

Tokens are normally separated by whitespace (space, tab, newline), but need not be if there is no ambiguity (which is generally only the case if a special character is adjacent to some other token type).
select * ; from table name;

22. Explain data cleaning in brief

Data Cleaning is the removal of unwanted observations.
https://www.geeksforgeeks.org/data-cleansing-introduction/
Steps involved in Data Cleaning:

(i) Removal of unwanted observations
(ii) Fixing Structural errors
(iii) Managing Unwanted outliers
(iv) Handling missing data



23. What is Descriptive and Inferential Statistics?

Descriptive statistics focus on describing the visible characteristics of a dataset (a population or sample).

Inferential statistics focus on making predictions or generalizations about a larger dataset, based on a sample of those data.

Descriptive: Describing
Inferential: Predictions or Generalizations

Descriptive statistics summarize the characteristics of a data set. Inferential statistics allow you to test a hypothesis or assess whether your data is generalizable to the broader population.

24. what is critical region?

The critical region is determined by the chosen level of significance, which is the probability of rejecting the null hypothesis when it is true. The size of the critical region determines the power of the test, which is the probability of correctly rejecting the null hypothesis when the alternative hypothesis is true.

25. What is significance level?

In statistical hypothesis testing, the significance level is the probability of rejecting the null hypothesis when it is true. It is denoted by the Greek letter alpha (α) and is usually set at a small value such as 0.01 or 0.05.

For example, if the significance level is set at 0.05, it means that there is a 5% chance of rejecting the null hypothesis when it is true. This means that if you run the test many times and the null hypothesis is true each time, about 5% of the tests will result in a rejection of the null hypothesis.

26. What does p-value means?

P-value is the level of marginal significance within a statistical hypothesis test, representing the probability of the occurrence of a given event, it considers the null hypothesis true.

27. What is Neural Network?

Neural networks, also known as artificial neural networks (ANNs) or simulated neural networks (SNNs), are a subset of machine learning and are at the heart of deep learning algorithms. Their name and structure are inspired by the human brain, mimicking the way that biological neurons signal to one another.

Q.What is forward propagation?

Forward propagation refers to the process of using an input to make a prediction through a model. It is a key part of the training process for many types of models, including neural networks.

In a neural network, forward propagation involves feeding the input through the network, starting at the input layer and working through each successive layer until the output layer is reached. At each layer, the input is transformed by a set of weights and biases, and it is passed through a nonlinear activation function before being passed to the next layer.

Q. What is Back Propagation?

Backpropagation is a method used to train artificial neural networks, which are a type of machine learning model. It is used to calculate the gradient of the error with respect to the model's weights, which can then be used to update the weights in order to minimize the error.

Backpropagation works by using the chain rule of calculus to calculate the gradient of the error with respect to each weight in the network. It begins at the output layer and works backwards through the layers of the network, propagating the error gradient backwards through the network.

Q. What is VLookUP and HLookup in Excel?

VLOOKUP and HLOOKUP are functions in Microsoft Excel that are used to search for specific information in a table or range of cells.

VLOOKUP (vertical lookup) searches for a value in the leftmost column of a table and returns a value from a specified column in the same row. For example, if you have a table of employee data and you want to find the salary of a specific employee, you could use the VLOOKUP function to search for the employee's name in the leftmost column of the table and return the salary from a column to the right.

HLOOKUP (horizontal lookup) is similar to VLOOKUP, but it searches for a value in the top row of a table and returns a value from a specified row in the same column. For example, if you have a table of sales data and you want to find the total sales for a specific product, you could use the

HLOOKUP function to search for the product name in the top row of the table and return the total sales from a row below.

Both VLOOKUP and HLOOKUP are useful for quickly finding specific information in large tables of data, but they have some limitations. They can only search in one direction (vertically or horizontally) and they require the lookup value to be in the leftmost column or top row of the table, respectively. Additionally, they can only return values from the same row or column as the lookup value, so they are not well-suited for more complex lookup tasks.

28. What is fillna()?

The fillna() method replaces the NULL values with a specified value. The fillna() method returns a new DataFrame object unless the inplace parameter is set to True , in that case the fillna() method does the replacing in the original DataFrame instead.
dataframe.fillna(value, method, axis, inplace, limit, downcast)
https://www.w3schools.com/python/pandas/ref_df_fillna.asp#:~:text=The%20fillna()%20method%20replaces,in%20the%20original%20DataFrame%20instead.

29. What is model explainability?

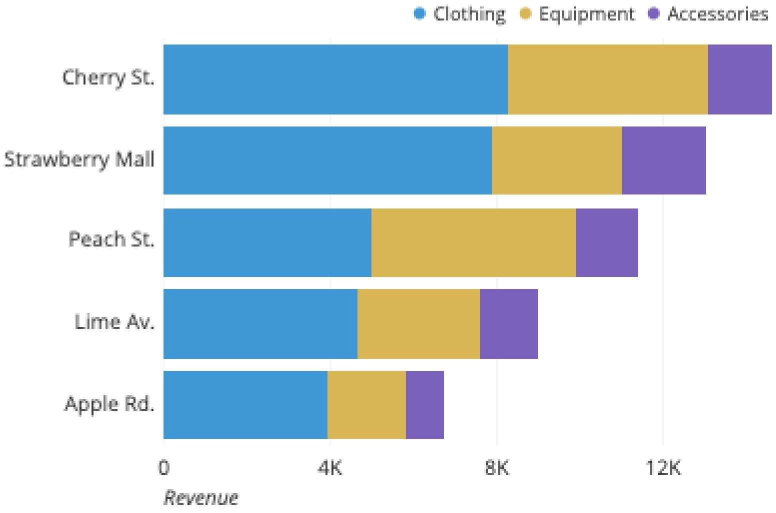
Model explainability refers to the ability to understand and explain how a machine learning model is making predictions or decisions. It is an important aspect of machine learning, particularly when the model is used in critical applications such as healthcare or finance, where the consequences of incorrect predictions or decisions can be significant
e.g. If a healthcare model is predicting whether a patient is suffering from a particular disease or not.

30. Why is Model Explainability required?

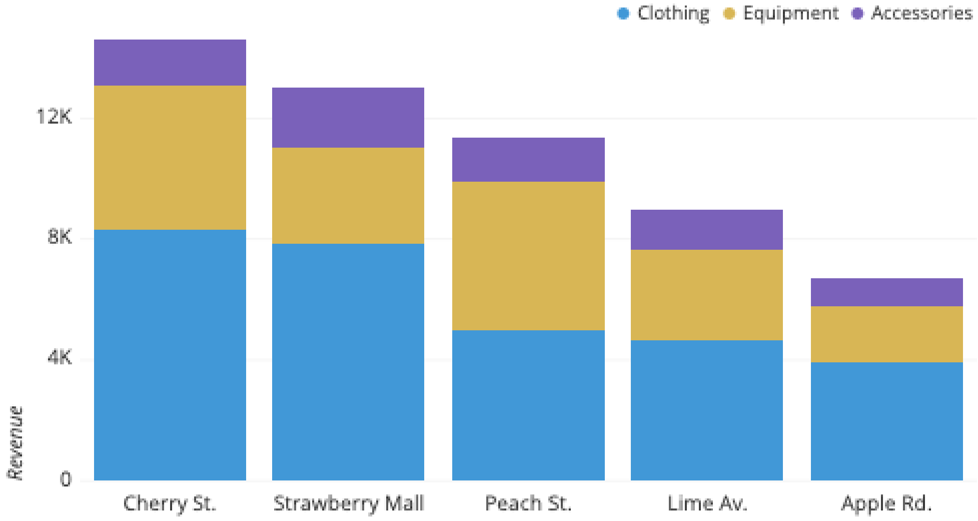
(i) Being able to interpret a model increases trust in a machine learning model.
(ii) Once we understand a model, we can detect if there is any bias present in the model.
(iii) Model Explainability becomes important while debugging a model during the development phase.
(iv) Model Explainability is critical for getting models to vet by regulatory authorities like FDA
https://www.analyticsvidhya.com/blog/2021/11/model-explainability/

31. What is line chart, stacked column chart, and stacked bar chart?

The ***stacked bar chart*** (aka stacked bar graph) extends the standard bar chart from looking at numeric values across one categorical variable to two. Each bar in a standard bar chart is divided into a number of sub-bars stacked end to end, each one corresponding to a level of the second categorical variable.



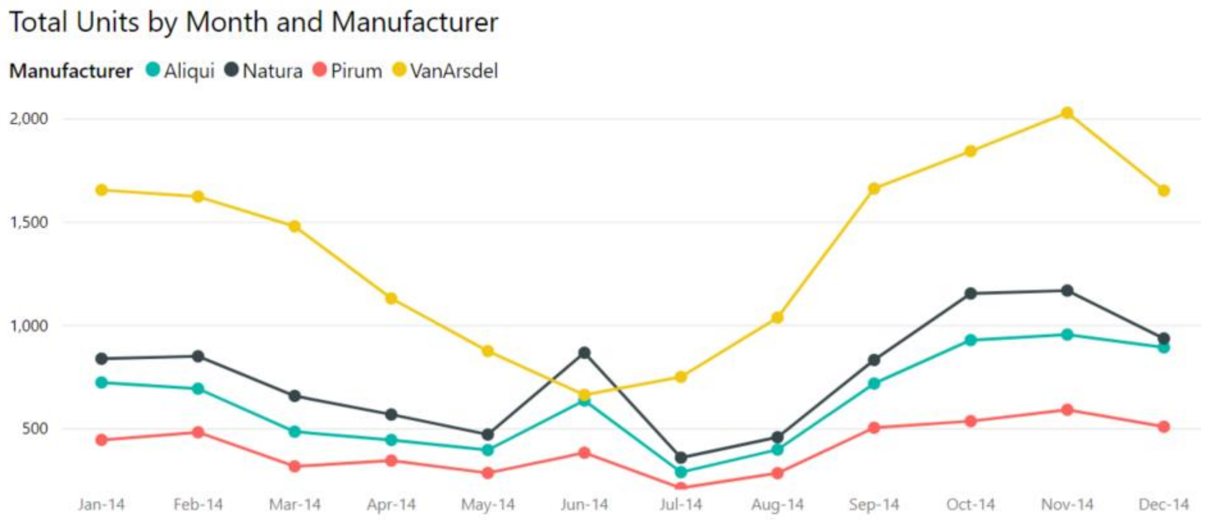
A ***stacked column chart*** is a chart type that is used to display data that is stacked in columns. It is used to show how much each value contributes to a total across a series of categories. In a stacked column chart, the values are plotted one on top of the other, with the values for each category stacked on top of each other to form columns. This allows you to compare the contribution of each value to the total for each category, and to see how the proportions of the different values change as you move from one category to the next.



A stacked column chart is similar to a stacked bar chart, with the main difference being the orientation of the bars.

- In a stacked column chart, the bars are vertical and are plotted one on top of the other, with the values for each category stacked on top of each other to form columns.
- In a stacked bar chart, the bars are horizontal and are plotted side by side, with the values for each category stacked on top of each other to form bars.

A ***line chart*** is a type of chart that provides a visual representation of data in the form of points that are connected in a straight line. The line can either be straight or curved depending on the data being researched.



https://www.cuemath.com/data/line-chart/

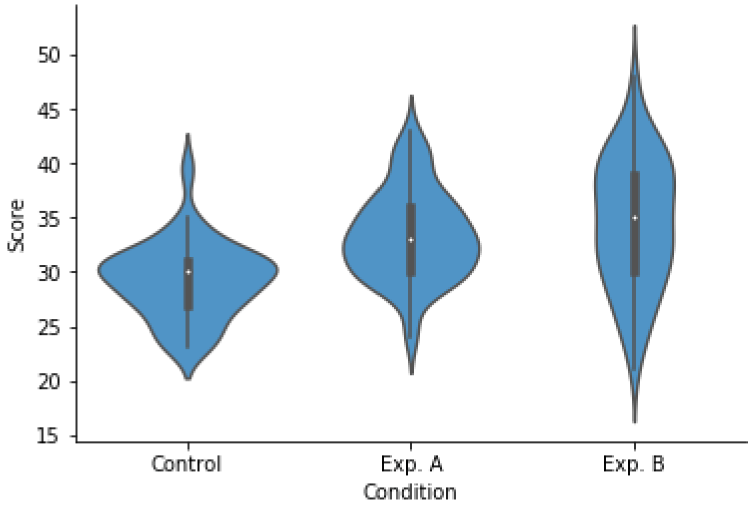
Q. What is Violin chart?

A violin plot is a hybrid of a box plot and a kernel density plot, which shows peaks in the data. It is used to visualize the distribution of numerical data. Unlike a box plot that can only show summary statistics, violin plots depict summary statistics and the density of each variable.

In a violin chart, the width of the chart at each level of the categorical variable is proportional to the number of observations at that level, and the thickness of the chart shows the distribution of the data at that level. The thicker parts of the chart represent areas where there is a higher density of data, while the thinner parts represent areas where there is a lower density of data.

Violin charts are useful for comparing the distribution of data across levels of a categorical variable and for visualizing the underlying distribution of the data. They are often used in combination with other types of plots, such as box plots or bar plots, to provide a more complete understanding of the data.

In Python, you can create a violin chart using the violinplot function in the seaborn library. This function takes in a DataFrame and the names of the variables to be plotted, and it generates a violin chart based on the values of those variables. You can customize the appearance of the chart by setting various options, such as the color and width of the violins and the visibility of the individual data points.



32. List out objects created by CREATE statement in MySQL.

The CREATE statement is used to create a variety of objects in the database
Following objects are created using CREATE statement:

- DATABASE
- EVENT
- FUNCTION
- INDEX
- PROCEDURE
- TABLE
- TRIGGER
- USER
- VIEW

33. What is TRIGGER used for in SQL?

A trigger is a special type of stored procedure that automatically runs when an event occurs in the database server.

34. What do you understand by IG (Information Gain)?

Information Gain = Entropy before splitting - Entropy after splitting

Information gain is used for determining the best features/attributes that render maximum information about a class. Information Gain, like Gini Impurity, is a metric used to train Decision Trees. Specifically, these metrics measure the quality of a split.

35. What do you mean by entropy?

In the context of machine learning, entropy is often used to measure the impurity of a node in a decision tree. A node is considered pure if all of the examples in the node belong to the same class, and impure if the examples belong to multiple classes. The entropy of a node is highest when the examples are evenly split between the classes, and lowest when the examples are completely pure.

Q. What is Gini in Gini index?

The Gini index is a measure of the impurity of a node in the tree. It is used to determine how to split the data at each node in the tree in order to create the most pure subnodes, where a pure subnode is one where all of the data points belong to the same class.

The Gini index is calculated based on the probability of a randomly chosen element belonging to a particular class. It is defined as:

Gini = 1 - ∑p(i)^2, where p(i) is the proportion of elements in the ith class.

36. What do you mean by Bag of Words?

The bag-of-words model is a way of representing text data when modeling text with machine learning algorithms.
The bag-of-words model is simple to understand and implement and has seen great success in problems such as language modeling and document classification.

https://machinelearningmastery.com/gentle-introduction-bag-words-model/

A bag-of-words model, or BoW for short, is a way of extracting features from text for use in modeling, such as with machine learning algorithms.

The approach is very simple and flexible, and can be used in a myriad of ways for extracting features from documents.

A bag-of-words is a representation of text that describes the occurrence of words within a document. It involves two things:

1. A vocabulary of known words.
2. A measure of the presence of known words.

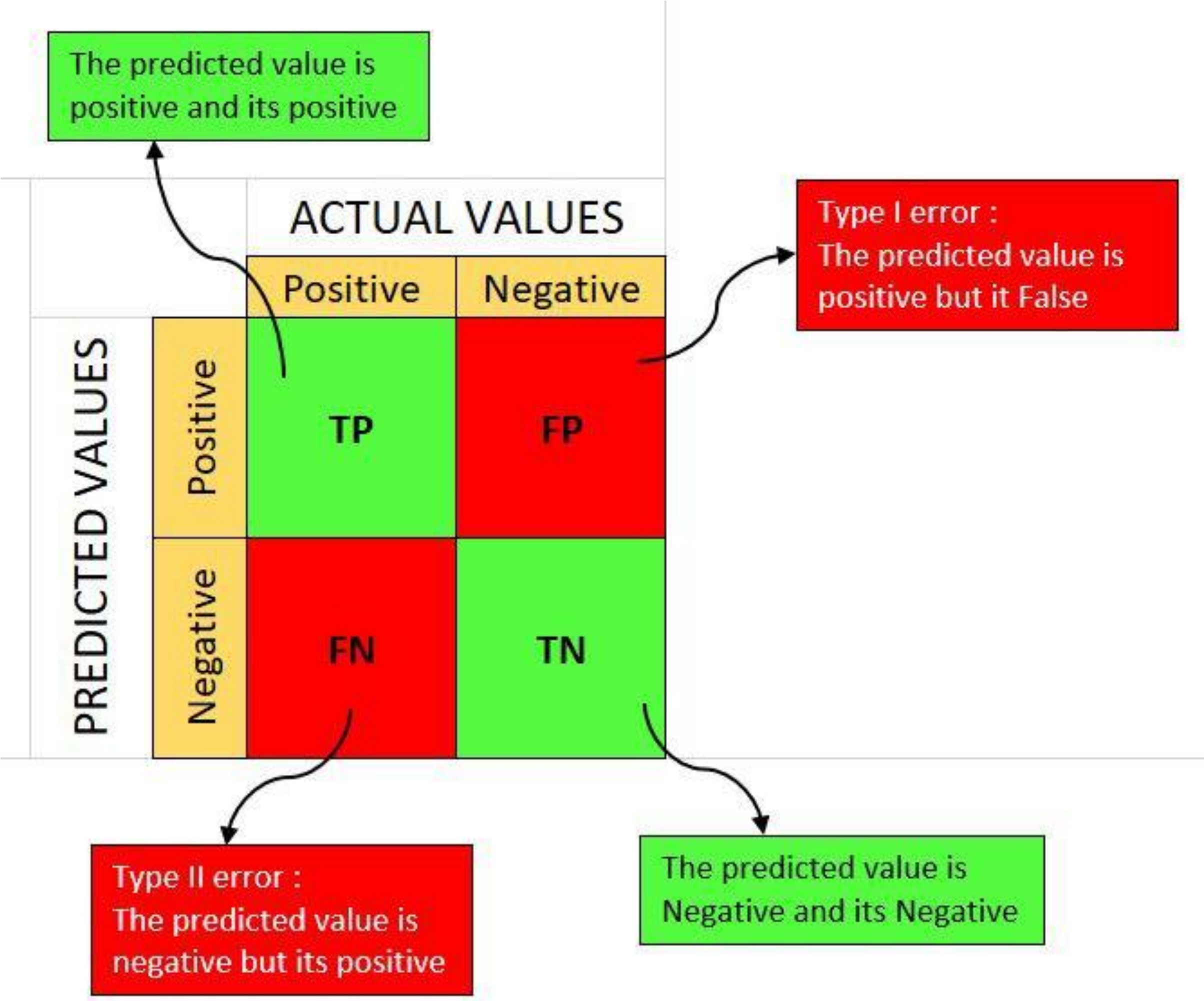
It is called a “bag” of words, because any information about the order or structure of words in the document is discarded. The model is only concerned with whether known words occur in the document, not where in the document.

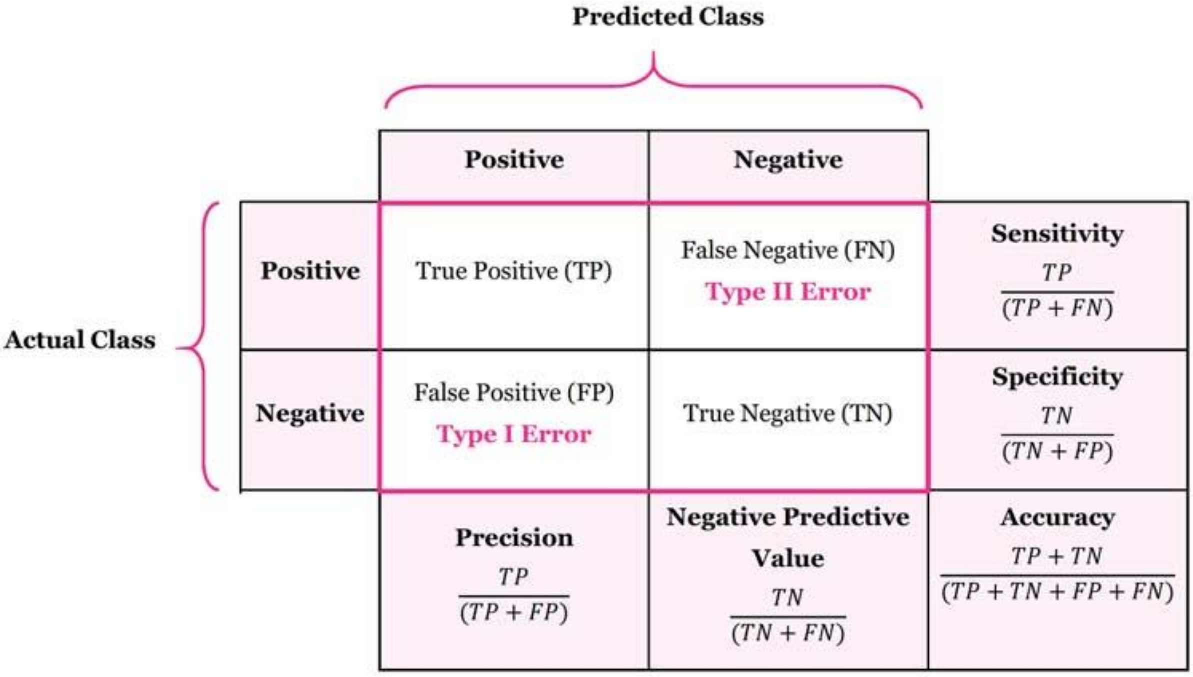
The most simple and known method is the Bag-Of-Words representation. It’s an algorithm that transforms the text into fixed-length vectors. This is possible by counting the number of times the word is present in a document. The word occurrences allow to compare different documents and evaluate their similarities for applications, such as search, document classification, and topic modeling.

The reason for its name, “Bag-Of-Words”, is due to the fact that it represents the sentence as a bag of terms. It doesn’t take into account the order and the structure of the words, but it only checks if the words appear in the document.

37. What is the Confusion Matrix?

A confusion matrix is a technique for summarizing the performance of a classification algorithm. Calculating a confusion matrix can give you a better idea of what your classification model is getting right and what types of errors it is making.
https://machinelearningmastery.com/confusion-matrix-machine-learning/


The predicted value is positive and its positive
ACTUAL VALUES
Positive
Negative
PREDICTED VALUES
Positive
Negative
TP
FP
FN
TN
Type I error : The predicted value is positive but it False
Type II error : The predicted value is negative but its positive
The predicted value is Negative and its Negative



		Predicted Class		
		Positive	**Negative**	
Actual Class	**Positive**	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

38. What is the ACCURACY Score?

Accuracy is one metric for evaluating classification models. Informally, accuracy is the fraction of predictions our model got right. Formally, accuracy has the following definition: Accuracy = Number of correct predictions Total number of predictions.
Accuracy = TP+TN/TP+FP+FN+TN

39. What is the PRECISION score?

The precision is the ratio TP / (TP + FP) where tp is the number of true positives and fp the number of false positives
OR
Out of all predicted positive values, how many are actually TRUE.

Depends on use cases like Mail spam classification : “when (type 1 FP) error is Dangerous”

40 .What is the RECALL score?

The recall is calculated as the ratio between the TP/(TP+FN) where tp is the number of true positives and fn the number of false negatives
OR
Out of all actual positive values, how many Predicted as TRUE.

Depends on use cases like Cancer detection: “when (type 2 FN) error is Dangerous”

41. What is F1 score?

harmonic mean between precision and recall

Definition: F1 score is defined as the harmonic mean between precision and recall. It is used as a statistical measure to rate performance.
F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

Q. What are Outliers?

Outliers are data points that are distant from other similar points due to variability in the measurement. Outliers should be excluded from the data set but detecting of those outliers is very difficult which is not always possible. The below blog clearly explains your effects of outliers in data and how to identify outliers in data.

Outliers are observations in a dataset that are significantly different from the majority of the other observations. They can be either higher or lower than the other values and can have a significant impact on statistical analyses and conclusions.

Outliers can occur for a variety of reasons, such as measurement errors, data entry errors, or truly unusual observations that are not representative of the rest of the data. It is important to identify and handle outliers appropriately because they can have a disproportionate effect on statistical measures, such as the mean and standard deviation.

There are several ways to identify outliers in a dataset. One common method is to use the interquartile range (IQR), which is the difference between the 75th and 25th percentiles of the data. Observations that are more than 1.5 times the IQR below the 25th percentile or above the 75th percentile are often considered outliers. Another method is to use standard deviations, where observations that are more than 3 standard deviations from the mean are considered outliers.

Once outliers have been identified, it is important to decide how to handle them. In some cases, it may be appropriate to remove the outliers from the dataset, if they are believed to be caused by errors or are not representative of the underlying population. In other cases, it may be more appropriate to keep the outliers and take them into account when analyzing the data. The appropriate approach will depend on the specific context and goals of the analysis.

42. What is precision, recall importance and scenarios?

Precision and recall are two evaluation metrics that are often used in the field of information retrieval and machine learning.

Precision is a measure of the accuracy of a classifier when it predicts the positive class. It is defined as the number of true positive predictions made by the classifier divided by the total number of positive predictions made by the classifier.

Recall is a measure of the ability of a classifier to find all the positive instances in a dataset. It is defined as the number of true positive predictions made by the classifier divided by the total number of positive instances in the dataset.

The importance of precision and recall depends on the specific application. In some cases, it is more important to have a classifier that has high precision, even if it means that some positive instances are missed. In other cases, it is more important to have a classifier that has high recall, even if it means that there are more false positive predictions.

One common scenario where precision and recall are important is in the field of information retrieval, where a search engine is trying to retrieve relevant documents from a large dataset. In this case, it is important to have a high precision, because it means that the search engine is returning relevant documents, but it is also important to have a high recall, because it means that the search engine is able to find all the relevant documents in the dataset.

43. What is the difference between MSE and MAE?

Mean Square Error (MSE): This measures the squared average distance between the real data and the predicted data.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$$

Advantages: 1. This method is differentiable at point zero so it can be used as loss function

Disadvantages: 1. Not Robust to outliers.
2. It does not give Same output as MAE like our predictor variable, meanwhile it gives a square of it.

Mean Absolute Error (MAE): This measures the absolute average distance between the real data and the predicted data, but it fails to punish large errors in prediction.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} | x_i - x |$$

Advantages: 1.It gives the same output like our predictor variable.
2. It is Robust to outliers. means it can handle outliers
.
Disadvantage: In this method we use modulus in its function, which is not differentiable at zero. and it is biggest drawback of MAE

44. What is the significance of the apply method?

The apply() method is a built-in Python function that allows you to apply a function to a set of data. It is part of the functools module, which is a standard library in Python. The apply() method is used to apply a function to a set of data, such as a list or a tuple. It is useful when you have a function that needs to be applied to multiple pieces of data, as it allows you to write the function once and then reuse it on different data without having to rewrite the function each time. The apply() method takes two arguments: the function to be applied, and the data to which the function should be applied. It returns the result of the function applied to the data

45. What is a Self-Join?

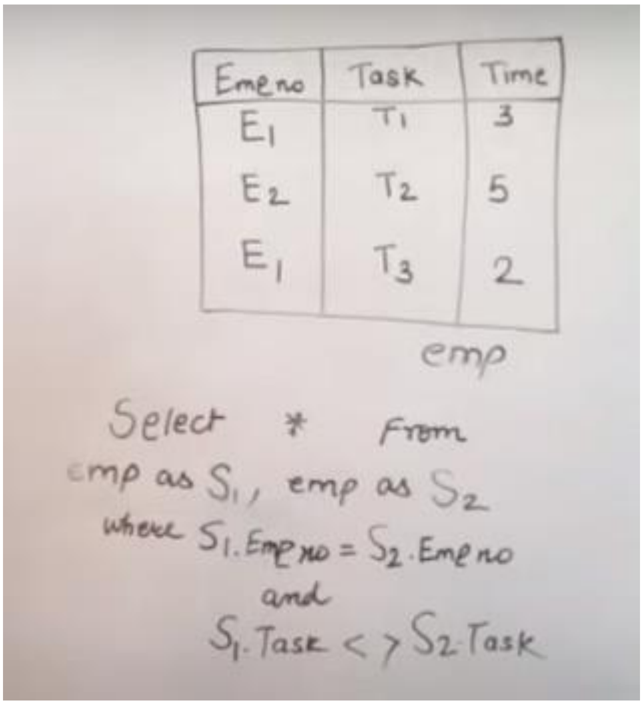
A self-join is a type of SQL join that allows you to join a table to itself. This is useful when you want to compare rows within a table or when you want to create a result set that combines data from a table with itself.

To perform a self-join, you need to specify the table that you want to join twice in the FROM clause of your SELECT statement. You also need to give each instance of the table a different alias. For example:

```
SELECT a.column1, a.column2, b.column1, b.column2
FROM table1 a
JOIN table1 b ON a.common_column = b.common_column
```

Another Example:

```
Select * from
emp as S1, emp as S2
where S1.emp no = S2.emp no
and S1.task <> S2.task
```



Emeno	Task	Time
E_1	T_1	3
E_2	T_2	5
E_1	T_3	2

emp

```
Select * From
mp as S1, emp as S2
where S1.Empno = S2.Empno
and
S1.Task <> S2.Task
```

In this example, table1 is being joined to itself using the JOIN clause and the ON clause specifies the condition that must be met in order for two rows to be considered a match.

Self-joins can be inner joins, left outer joins, or right outer joins, just like any other type of join. They can also be combined with other types of joins, such as cross joins or natural joins.

47. What is the Central Limit Theorem?

In probability theory, the central limit theorem establishes that, in many situations, when independent random variables are summed up, their properly normalized sum tends toward a normal distribution even if the original variables themselves are not normally distributed.

The Central Limit Theorem is a statistical theorem that states that, given a sufficiently large sample size from a population with a finite level of variance, the mean of the sample will be approximately equal to the mean of the population. Additionally, the distribution of the sample means will be approximately normal, even if the distribution of the population is not normal.

The Central Limit Theorem has important implications for statistical inference, as it allows us to use the normal distribution to approximate the distribution of sample means and to perform statistical tests, even if the underlying population is not normally distributed

.The Central Limit Theorem is a very important result in statistics and has many practical applications. It is used in a wide range of fields, including economics, finance, psychology, and engineering.

48. How do you explain standard deviation?

A standard deviation (or σ) is a measure of how dispersed the data is in relation to the mean.

49. R2 vs adjusted r2?

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

The most vital difference between adjusted R-squared and R-squared is simply that adjusted R-squared considers and tests different independent variables against the model and R-squared does not.

In a regression model, R-squared (R2) is a measure of how well the model fits the data. It is the proportion of the variance in the dependent variable that is explained by the model. R2 ranges from 0 to 1, with a higher value indicating a better fit.

Adjusted R-squared (adj. R2) is a modified version of R2 that adjusts for the number of predictors in the model.
It is a penalized version of R2 that increases only when the new term improves the model more than would be expected by chance.
The adjusted R2 increases only if the new term is significant and improves the model more than would be expected by chance.

In general, you should aim for the highest R2 value possible, but be aware that adding more predictors to the model will always increase R2, even if those predictors are not actually important.

The adjusted R2 can help you decide whether or not a predictor is worth including in the model, because it adjusts for the number of predictors.

If the adjusted R2 is not much higher than the R2, then the additional predictors are not adding much value to the model.

50. r2 vs mae?

The lower value of MAE, MSE, and RMSE implies higher accuracy of a regression model. However, a higher value of R square is considered desirable. R Squared & Adjusted R Squared are used for explaining how well the independent variables in the linear regression model explains the variability in the dependent variable.

51. r2 vs mse?

MSE represents the residual error which is nothing but sum of squared difference between actual values and the predicted / estimated values divided by total number of records.

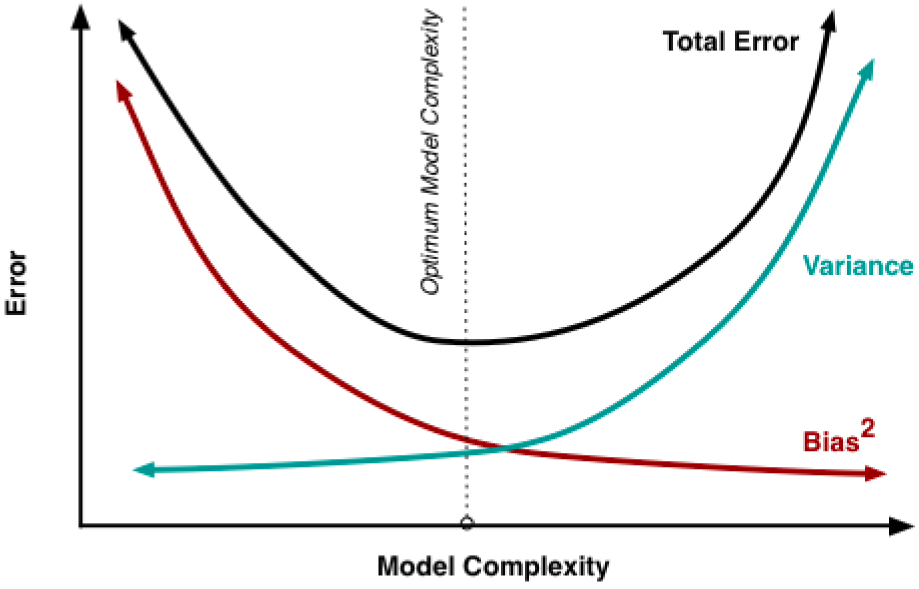
R-Squared represents the fraction of variance captured by the regression model.

52. What is overfitting in machine learning?

“Overfitting occurs when our machine learning model tries to cover all the data points or more than the required data points present in the given dataset.” Note: low bias and high variance

54. What is the bias-variance trade off?

In statistics and machine learning, 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.



55. What is bias and variance?

Simply stated, variance is the variability in the model prediction—how much the ML function can adjust depending on the given data set. Variance comes from highly complex models with a large number of features.

Bias: bias is the amount that a model's prediction differs from the target value, compared to the training data.

Models with high bias will have low variance. Models with high variance will have a low bias.

56. What's the similarities and differences between Bagging, Boosting, Stacking?

All three are so-called "meta-algorithms": approaches to combine several machine learning techniques into one predictive model in order to decrease the variance (bagging), bias (boosting) or improving the predictive force (stacking alias ensemble).

Bagging (stands for Bootstrap Aggregating) is a way to decrease the variance of your prediction by generating additional data for training from your original dataset using combinations with repetitions to produce multisets of the same cardinality/size as your original data.
By increasing the size of your training set you can't improve the model predictive force, but just decrease the variance, narrowly tuning the prediction to expected outcome.

Boosting is a two-step approach, where one first uses subsets of the original data to produce a series of averagely performing models and then "boosts" their performance by combining them together using a particular cost function (=majority vote).
Unlike bagging, in the classical boosting the subset creation is not random and depends upon the performance of the previous models: every new subsets contains the elements that were (likely to be) misclassified by previous models.

Stacking is a similar to boosting: you also apply several models to your original data. The difference here is, however, that you don't have just an empirical formula for your weight function, rather you introduce a meta-level and use another model/approach to estimate the input together with outputs of every model to estimate the weights or, in other words, to determine what models perform well and what badly given these input data.

57. What are Weak Learners?

In machine learning, a weak learner is a model that is only slightly better than random guessing. Weak learners are used as building blocks in ensemble learning methods, where multiple weak learners are combined to create a strong learner that can make accurate predictions.

One common example of a weak learner is a decision tree with a shallow depth. These trees are only slightly better than random guessing, but they can be combined with other weak learners to create a more accurate model.

Ensemble learning methods that use weak learners include boosting algorithms, such as AdaBoost, and bagging algorithms, such as Random Forest. These methods train multiple weak learners on different subsets of the training data and combine their predictions to make a final prediction.

By combining the predictions of multiple weak learners, ensemble learning methods can often achieve higher accuracy than could be achieved by any of the individual weak learners alone.

In ensemble learning theory, we call weak learners (or base models) models that can be used as building blocks for designing more complex models by combining several of them. Most of the time, these basics models perform not so well by themselves either because they have a high bias (low degree of freedom models, for example) or because they have too much variance to be robust (high degree of freedom models, for example).

58. Given a list, write a Python program to swap first and last element of the list?

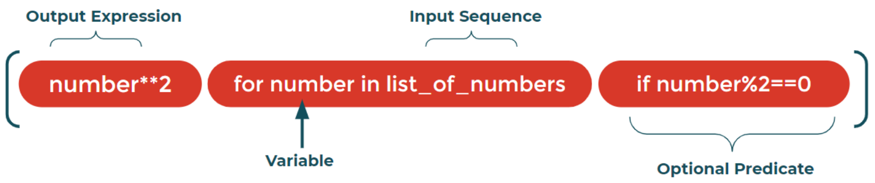
```
l1=[1,2,3,4,5,7,8,9,9,0]
print('list before swapping',l1)
l1[0], l1[-1] = l1[-1], l1[0] if len(l1) > 1 else (l1[0], l1[0])
print(f"list after swapping  {l1}")
```

59. Write a function to find the words in a string that greater than the given number

```
def string_k(k, str): string = [] text = str.split(" ") for x in text: if len(x) > k: string.append(x) return
string # Driver Program k = 3 str ="almabetter" print(string_k(k, str))
```

60. What are the different parts of syntax of list comprehension?

An Input Sequence. A Variable representing members of the input sequence. An Optional Predicate expression. An Output Expression producing elements of the output list from members of the Input Sequence that satisfy the predicate.



List comprehension offers a shorter syntax when you want to create a new list based on the values of an existing list.

eg:-

```
fruits = ["apple", "banana", "cherry", "kiwi", "mango"]
newlist = [x for x in fruits if "a" in x]
```

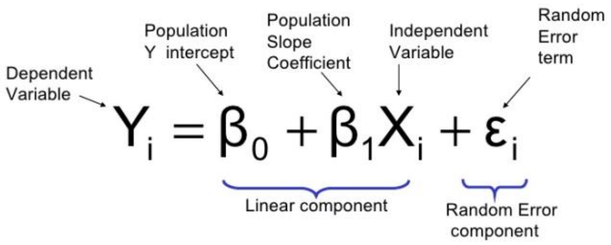
eg:-

```
odd_squared_numbers = [number**2 if number%2!=0 else number**3 for number in
list_of_numbers ]
print(odd_squared_numbers)
```

Q. What is linear regression?

Linear regression is one of the most basic types of regression in supervised machine learning. The linear regression model consists of a predictor variable and a dependent variable related linearly to each other.
We try to find the relationship between independent variable(input) and a corresponding dependent variable (output).

This can be expressed in the form of a straight line :



Linear regression, also known as ordinary least squares (OLS) and linear least squares, is the real workhorse of the regression world.

Q. What are the assumptions of linear regression?

LEHM - ***Linear relation - Error(Mean of Residual error) - Homoscedastity - Multicollinearity***

Assumption of regression line:

1. The relation between the dependent and independent variables
should be almost Linear.

2. Mean of residuals (error) should be zero or close to 0 as much as possible. It
is done to check whether our line is actually the line of “best fit”.

3. There should be homoscedasticity or equal variance in a regression
model. This assumption means that the variance around the
regression line is the same for all values of the predictor variable (X).

4. There should not be multicollinearity in regression model.
Multicollinearity generally occurs when there are high correlations
between two or more independent variables.

Q.What is logistic regression?

Logistic regression is a statistical method for modeling a binary outcome. In logistic regression, the dependent variable is a binary variable that takes on values of 0 or 1. The goal of logistic regression is to find the best model that can predict the probability that a given data point belongs to one of the two classes.

To do this, logistic regression uses an equation with a set of weights and an intercept term that is learned from the data. The inputs to the equation are a set of features or predictors, and the output is the predicted probability that a given data point belongs to one of the two classes. The weights and intercept are learned by optimizing an objective function that measures the difference between the predicted probabilities and the actual outcomes in the training data.

Logistic regression is often used in classification tasks, where the goal is to predict which of a set of predefined classes a data point belongs to. It is also used in other applications such as predicting the likelihood of a customer clicking on an ad or the likelihood of a patient having a certain medical condition.

formula : log(□1−□)=□0+□1□1,...,□□□□=□□□

Q.What is regularised liner regression?

Regularized linear regression is a method for modeling data that tries to minimize the complexity of the model by adding a penalty term to the objective function that is being optimized. This penalty term is called the regularization term and helps to prevent overfitting of the model to the training data.
There are several different types of regularization that can be used in linear regression, including L1 regularization, L2 regularization, and Elastic Net regularization.
Each of these methods uses a different approach to adding the regularization term to the objective function and can lead to different behavior in the resulting model.

Q. What is difference between Linear And Logistic Regression?

- 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.

Q. What is ridge and lasso regression?

Ridge regression and Lasso regression are both techniques for regularized linear regression, which is a method of modeling data that seeks to minimize the complexity of the model by adding a penalty term to the objective function being optimized. The main difference between Ridge and Lasso regression is in the form of the regularization term that is added to the objective function.

In Ridge regression, the regularization term is the sum of the squares of the coefficients of the independent variables in the model. This has the effect of shrinking the coefficients of the independent variables towards zero, which can help to reduce the complexity of the model and prevent overfitting.

On the other hand, Lasso regression uses the sum of the absolute values of the coefficients of the independent variables as the regularization term. This has the effect of setting some of the coefficients of the independent variables exactly to zero, which can be useful for feature selection as it can help to identify which features are most important for predicting the outcome.

Both Ridge and Lasso regression are useful for controlling overfitting and improving the generalization of the model to new data. The choice of which method to use will depend on the specific characteristics of the data and the goals of the modeling process.

Q. What is cross validation and rolling cross validation?

Cross-validation is a method for evaluating the performance of a machine learning model on unseen data. It involves dividing the data into a training set and a test set, training the model on the training set, and evaluating its performance on the test set. This process is repeated a number of times, with different combinations of training and test sets, in order to get a better estimate of the model's performance.

There are several different types of cross-validation, including k-fold cross-validation and stratified k-fold cross-validation. In k-fold cross-validation, the data is divided into k folds, and the model is trained and evaluated k times, with each fold being used as the test set once and the training set the other k-1 times. Stratified k-fold cross-validation is similar, but it ensures that the proportion of each class is approximately the same in each fold as it is in the whole dataset.

Rolling cross-validation is a type of cross-validation that can be used when the data is time series data, where the observations are ordered in time. In rolling cross-validation, the model is trained on a window of consecutive data points and evaluated on the next point in the sequence. This process is repeated, rolling the window along the time series, until the model has been trained and evaluated on all the data.

Cross-validation is a useful method for evaluating the performance of machine learning models and comparing the performance of different models. It can help to reduce the risk of overfitting, as the model is trained and evaluated on different subsets of the data, and it can provide a more robust estimate of the model's performance.

Q. What is k fold cross validation?

K-fold cross-validation is a method for evaluating the performance of a machine learning model. It involves dividing the data into a training set and a test set and training the model on the training set. The model is then evaluated on the test set, and this process is repeated a number of times, with different combinations of training and test sets.

In k-fold cross-validation, the data is divided into k folds, and the model is trained and evaluated k times, with each fold being used as the test set once and the training set the other k-1 times. For example, in 5-fold cross-validation, the data is divided into 5 folds, and the model is trained and evaluated 5 times, with each fold being used as the test set once.

The performance of the model is then evaluated by averaging the performance across all k iterations. This can provide a more robust estimate of the model's performance, as it is less sensitive to the specific split of the data into training and test sets.

K-fold cross-validation is a useful method for evaluating the performance of machine learning models and comparing the performance of different models. It can help to reduce the risk of overfitting, as the model is trained and evaluated on different subsets of the data, and it can provide a more robust estimate of the model's performance.

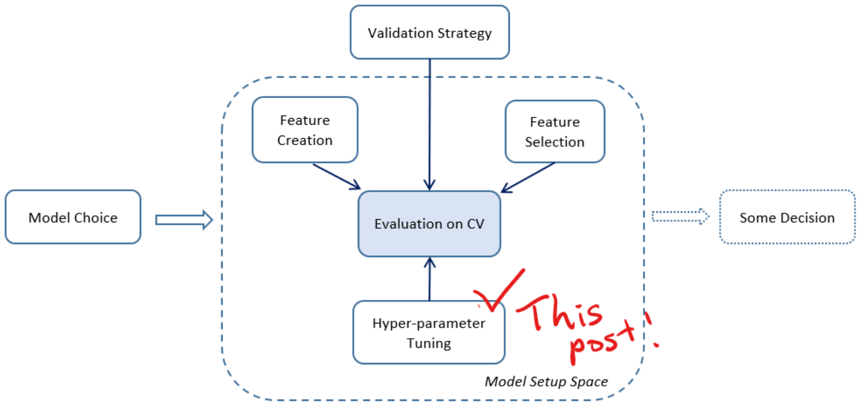
Q. What is hyper parameter tuning

Hyperparameter tuning is the process of optimizing the hyperparameters of a machine learning model. Hyperparameters are the parameters of a model that are set prior to training and that control the learning process. They are different from the model parameters, which are learned from the data during training.

Hyperparameter tuning is important because the performance of a machine learning model can depend heavily on the hyperparameters chosen. Different hyperparameter values can result in models with very different behaviors, and finding the optimal hyperparameters for a particular problem can require a significant amount of experimentation.

There are several methods for hyperparameter tuning, including manual tuning, grid search, and random search. In manual tuning, the hyperparameters are chosen manually by the practitioner based on their experience and knowledge of the problem. Grid search is a method for systematically searching through a predefined set of hyperparameter values, training a model for each combination of values, and evaluating the performance of each model. Random search is a method for sampling random combinations of hyperparameter values and evaluating the performance of the resulting models.

Hyperparameter tuning can be a time-consuming process, but it is important for getting the best performance out of a machine learning model. It can be especially important for complex models with many hyperparameters, as the performance of these models can be sensitive to even small changes in the hyperparameter values.



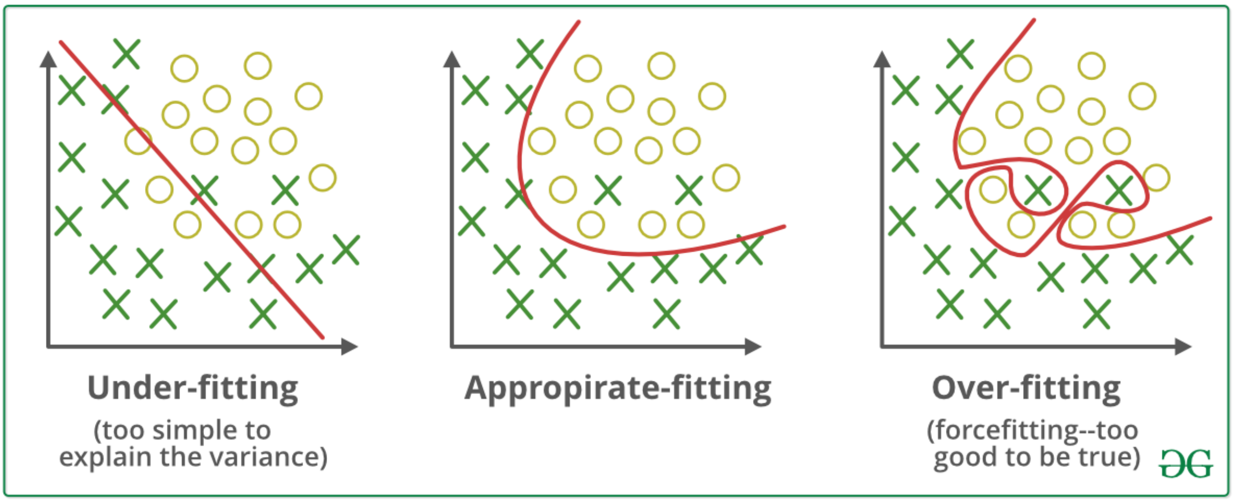
Q.What is underfitting and overfitting?

Underfitting and overfitting are common problems that can occur when training a machine learning model.

Underfitting occurs when a model is not able to capture the underlying trends in the data and has poor performance on both the training data and new data. This can happen when the model is too simple, or when there is not enough data to learn from. Underfitting can be identified by a poor performance on the training data, as well as a poor generalization to new data.

On the other hand, ***overfitting occurs*** when a model is too complex and has learned the noise in the training data rather than the underlying trends. An overfitted model will have excellent performance on the training data, but it will not generalize well to new data and will have poor performance on test or validation data.

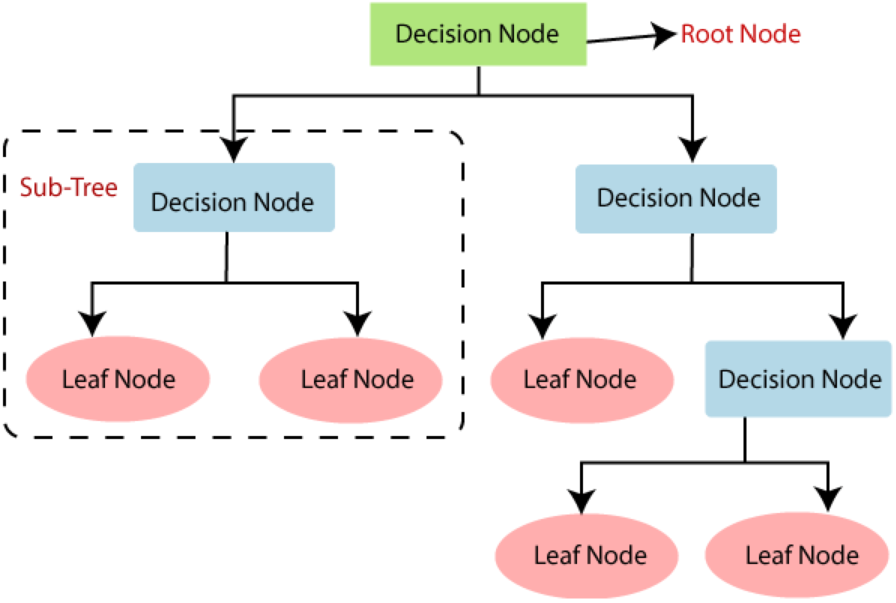
Both underfitting and overfitting can be addressed by using a more appropriate model, adding more data, or using regularization techniques to constrain the model and prevent it from becoming too complex.


Under-fitting
(too simple to explain the variance)
Appropirate-fitting
Over-fitting
(forcefitting--too good to be true)

Q.What is decision tree?
A decision tree is a type of machine learning model that is used for classification and regression tasks. It is called a decision tree because it is a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. Decision trees are a popular method for both classification and regression because they are easy to interpret, handle categorical data well, and can handle multiple classes.

In a decision tree, the model creates a tree-like structure, with an internal node representing an attribute or feature of the data, and branches representing the possible values that the attribute can take. The model splits the data into subsets based on the values of the attributes, and continues to do so recursively until a stopping criterion is reached. The final nodes of the tree, called leaf nodes, represent the final classification or prediction made by the model.

Decision trees are often used in combination with other models, such as random forests, which are an ensemble of decision trees that work together to make more accurate predictions. They can also be used to make decisions based on uncertain or incomplete information, by using probabilities and utility to evaluate the potential outcomes of different decisions.



46. Why do we require pruning in the Decision Tree? Explain.

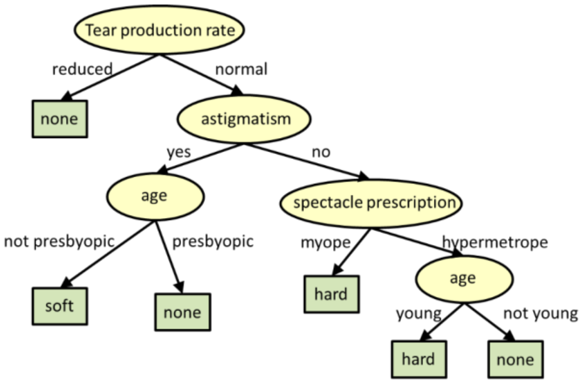
Pruning reduces the size of decision trees by removing parts of the tree that do not provide power to classify instances. Decision trees are the most susceptible out of all the machine learning algorithms to overfitting and effective pruning can reduce this likelihood.

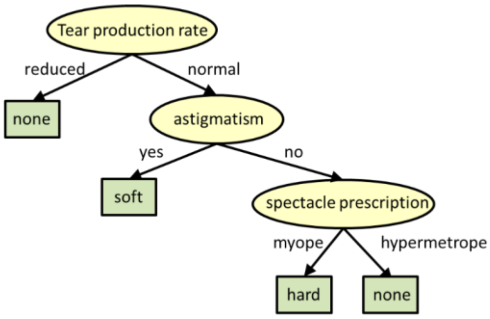
Pruning is the process of removing unnecessary nodes from a decision tree. It is used to reduce the complexity of the tree and to improve the generalization ability of the model.

There are two main reasons why pruning is important in decision trees:

Overfitting: Without pruning, decision trees are prone to overfitting, which means that they can perform well on the training data, but not generalize well to new data. Pruning helps to prevent overfitting by removing nodes that do not contribute significantly to the overall accuracy of the model.

Complexity: Decision trees can become very complex, with a large number of nodes and branches. This can make them difficult to understand and interpret, and can also lead to slower performance when making predictions. Pruning helps to reduce the complexity of the tree, making it easier to understand and faster to use. Overall, pruning is an important technique that can help to improve the performance and interpretability of decision tree models.





Q. What is KNN ?

K-nearest neighbors (KNN) is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The KNN algorithm stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). A new case is classified by a majority vote of its neighbors, with the case being assigned to the class most common among its K nearest neighbors measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbor.

Q. What is bernoulli distribution?

The Bernoulli distribution is a discrete probability distribution that models the outcome of a single binary event, such as a coin flip. It is defined by a single parameter, p, which is the probability of the event occurring.

The probability mass function (PMF) of the Bernoulli distribution is defined as:

f(x) = p^x (1-p)^(1-x), where x is either 0 (indicating that the event did not occur) or 1 (indicating that the event did occur).

Q. Explain Line chart, stacked bar chart and stacked column chart.Question

The line chart is a popular type of diagrammatic way for visualizing the data, it connects the individual data points to view the data. We can easily visualize the series of values, we can see trends over time or predict future values. The horizontal axis holds the category to which it belongs and the vertical axis holds the values.

Stacked Bar Chart, composed of multiple bars stacked horizontally, one below the other. The length of the bar depends on the value in the data point. A stacked bar chart makes the work easier, they will help us to know the changes in all variables presented, side by side. We can watch the changes in their total and forecast future values.

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.

63. Given a list, write a Python program to swap first and last element of the list.

```
x = input("Enter a list of elements: ")#modified the list by asking the input from user
num = x.split()                         #splitting the input
print('list before swapping',num)
num[0], num[-1] = num[-1], num[0]  #swapping the variable positions
print(f"list after swapping  {num}")
```

64. Write a python program to find factorial of a number.

```
num = 5
factorial = 1
for i in range(1, num+1):
  factorial *= i
print(factorial)  # prints 120
```

65. Top earners hackerrank solution mysql

Top Earners | HackerRank

```
select months*salary, count(*) from employee
group by months*salary
order by months*salary desc
limit 1;
```

output : **108064 7**

66. You randomly draw a coin from 100 coins - 1 unfair coin (head-head), 99 fair coins (head-tail) and roll it 10 times. If the result is 10 heads, whats the probability that the coin is unfair?

If you randomly draw a coin from a group of 100 coins, where 1 of the coins is unfair (two heads) and 99 of the coins are fair (one head and one tail), and you roll the coin 10 times and get 10 heads, the probability that the coin is unfair is very high.

To calculate the probability exactly, you can use the formula for the probability of an event occurring multiple times in a row:
P(unfair coin) = (1/100) * (1/100) * (1/100) * ... * (1/100) = (1/100)^10
Plugging in the values, you get:
P(unfair coin) = (1/100)^10 = 1/10000000000 = 1.0 * 10^(-10)
This probability is very close to 1, which means that it is highly likely that the coin you drew and rolled 10 times is the unfair coin with two heads.

67. What are limitations of L2 Regularization?

L2 regularization, also known as weight decay, is a method used to improve the generalization ability of a machine learning model. It does this by adding a penalty term to the objective function that the model is trying to optimize. The penalty term is the sum of the squares of the model weights, multiplied by a regularization coefficient.

One limitation of L2 regularization is that it can lead to a slower convergence of the optimization algorithm, especially if the regularization coefficient is set too high. This is because the penalty term slows down the optimization process by making the objective function "stiffer" and harder to optimize.

Another limitation of L2 regularization is that it tends to push the weights of the model towards smaller values, but it does not guarantee that the weights will be exactly zero. This means that L2 regularization can still allow some unimportant features to have non-zero weights, which can lead to suboptimal performance.

Overall, L2 regularization can be a useful tool for improving the generalization ability of a machine learning model, but it is important to carefully tune the regularization coefficient and consider other regularization methods as well, depending on the specific problem at hand.

Q.What is class imbalance?

Class imbalance refers to a situation where the number of data points belonging to one class is significantly different from the number of data points belonging to other classes. This can be a problem in machine learning because it can lead to a model that is biased towards the more common class.

For example, consider a binary classification problem where the goal is to predict whether a customer will make a purchase or not. If the dataset has a class imbalance, with many more negative examples (customers who did not make a purchase) than positive examples (customers who did make a purchase), then a classifier trained on this data may be more accurate at predicting the negative class than the positive class. This can be a problem if the goal is to identify customers who are likely to make a purchase, as the classifier may not be able to accurately identify these customers.

Q.What is over sampling and undersampling

Oversampling involves increasing the number of data points belonging to the minority class by generating synthetic data points or by duplicating existing data points. The goal of oversampling is to balance the class distribution by making the minority class more prevalent in the dataset.

Undersampling involves reducing the number of data points belonging to the majority class. This can be done by randomly selecting a subset of the majority class data points or by using a targeted sampling method to select specific data points to remove. The goal of undersampling is to balance the class distribution by making the majority class less prevalent in the dataset.

Q. What is black box?

A black box model is a model whose internal workings are not transparent and are not easily interpretable by humans. They are trained using a set of input data and a set of desired output labels, and they make predictions based on this training data. However, the process by which the model arrives at these predictions is not easily understood by humans, and the model's internal decision-making process is not transparent.

Q.What is multinomial distribution?

The multinomial distribution is a probability distribution that describes the outcome of a multi-class classification problem. It is a generalization of the binomial distribution, which describes the outcome of a binary classification problem.

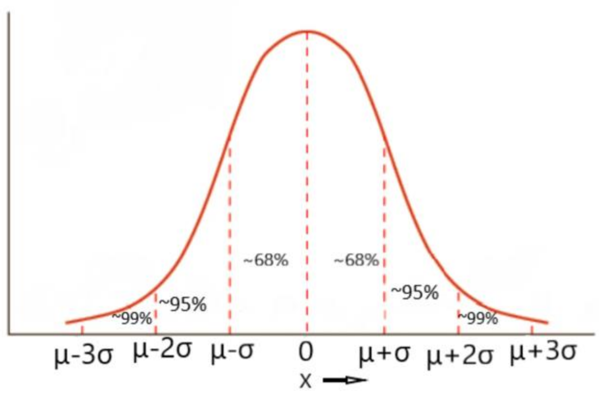
In a multi-class classification problem, there are K possible classes, and the goal is to predict which class a given data point belongs to. The multinomial distribution models the probability of each class given the data point, and it is defined by K parameters, p1, p2, ..., pK, which represent the probability of each class.

Q. What is Gaussian Distribution?

The Gaussian distribution, also known as the normal distribution, is a continuous probability distribution that is defined by a mean and a standard deviation. It is a bell-shaped curve that is symmetrical about the mean, and it is often used to model real-valued random variables.

The probability density function (PDF) of the Gaussian distribution is defined as:

f(x) = (1/(sigma * sqrt(2pi))) * exp(-((x-mu)^2)/(2sigma^2))

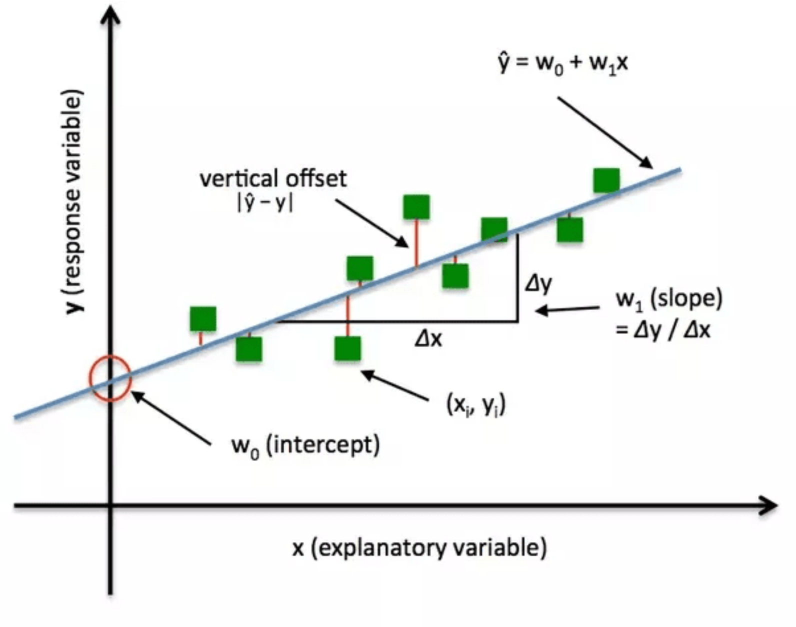


Q. What is Loss funtion?

A loss function is a measure of how good or bad a model's predictions are, compared to the true values. It is a key part of the training process for many types of machine learning models, and it is used to optimize the model's parameters in order to improve its performance.

The loss function is a mathematical function that takes in the predicted output of the model and the true output, and it calculates a numeric value that represents the error or loss between the two. The goal of the training process is to minimize this loss, typically by adjusting the model's parameters

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$



Q. What is activation function ?

An activation function is a function used in artificial neural networks to transform the input into an output. The goal of the activation function is to introduce nonlinearity into the network, which makes it possible to learn and model complex relationships in the data.

There are many different activation functions that can be used, each with its own characteristics and properties. Some common activation functions include sigmoid, tanh, ReLU, and softmax.

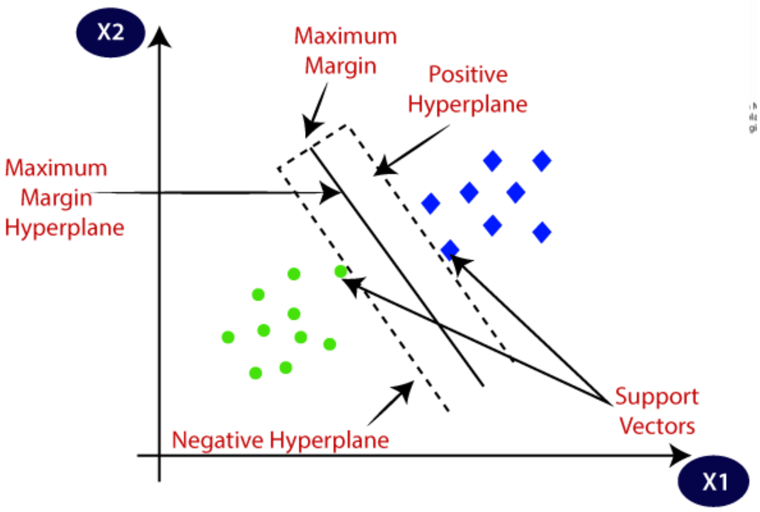
The sigmoid activation function is a smooth function that maps input values to output values between 0 and 1. It is often used in the output layer of a binary classification model.

Q. What is SVM?

Support vector machines (SVMs) are a type of supervised learning algorithm that can be used for classification or regression tasks. They are based on the idea of finding the hyperplane in a high-dimensional space that maximally separates the classes.

In the case of a classification task, the goal is to find the hyperplane that maximally separates the data points belonging to different classes. Once the hyperplane is found, new data points can be easily classified by checking on which side of the hyperplane they lie.

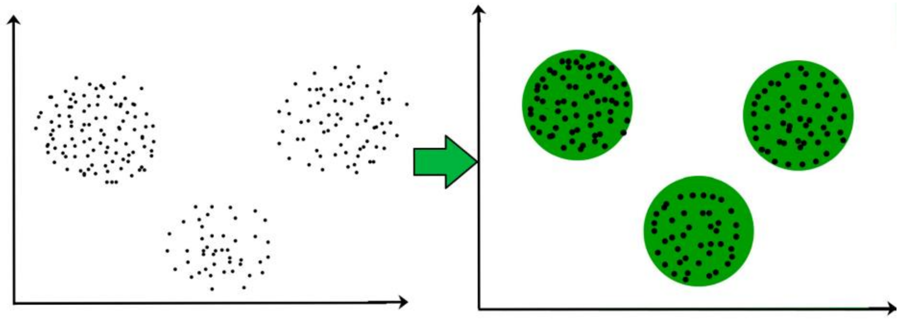
SVMs are particularly useful in cases where the data is not linearly separable, which means that it cannot be separated into different classes by a straight line. In these cases, SVMs can use the kernel trick to transform the data into a higher-dimensional space where it becomes linearly separable.



Q. What is clustering?

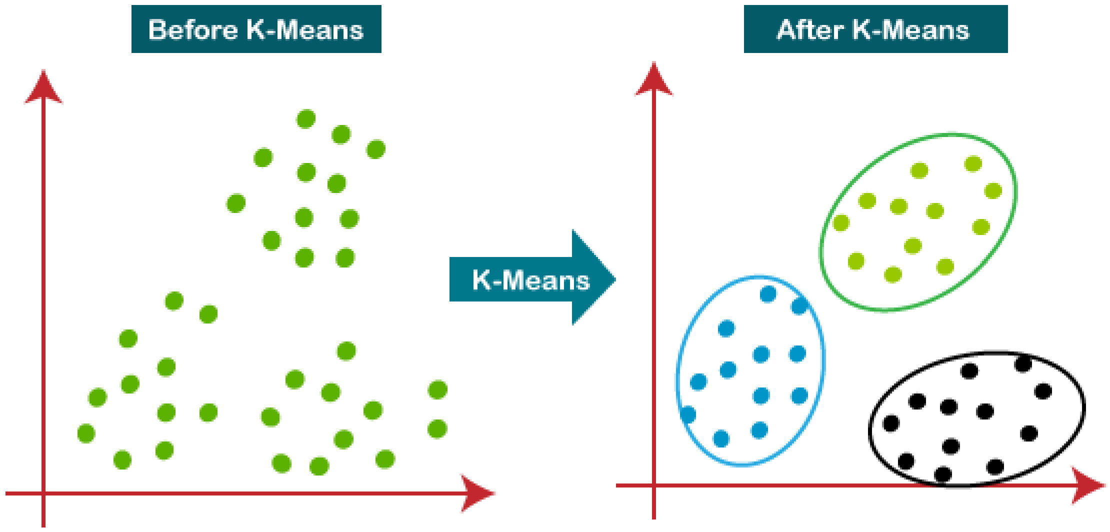
Clustering is a type of unsupervised learning that involves dividing a set of data points into groups (called clusters) based on their similarity. The goal of clustering is to group together data points that are similar to each other and to separate out data points that are dissimilar.

There are many different algorithms that can be used for clustering, and the choice of algorithm will depend on the specific characteristics of the data and the goals of the clustering process. Some common clustering algorithms include k-means, hierarchical clustering, and DBSCAN.



Q. What is K-mean clustering?

K-means clustering is an unsupervised machine learning algorithm that divides a dataset into a specified number of clusters (K) based on the patterns in the data. The algorithm works by first randomly selecting K points (centroids) in the data, and then assigning each data point to the closest centroid. It then recalculates the centroids based on the mean of the points assigned to them, and repeats the process of reassigning points and recalculating centroids until the centroids stop moving or the assignments stop changing. The result is a partitioning of the data into K clusters, with each cluster containing points that are similar to each other.

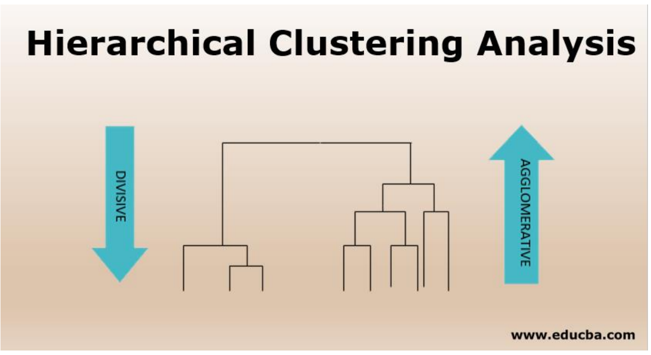


Q.What is hierarchical clustering?

A Hierarchical clustering method works via grouping data into a tree of clusters. Hierarchical clustering begins by treating every data point as a separate cluster. Then, it repeatedly executes the subsequent steps:

- Identify the 2 clusters which can be closest together, and
- Merge the 2 maximum comparable clusters. We need to continue these steps until all the clusters are merged together.

In Hierarchical Clustering, the aim is to produce a hierarchical series of nested clusters. A diagram called Dendrogram (A Dendrogram is a tree-like diagram that statistics the sequences of merges or splits) graphically represents this hierarchy and is an inverted tree that describes the order in which factors are merged (bottom-up view) or clusters are broken up (top-down view).



Q. What is topic modeling?

Topic modeling is a method of unsupervised machine learning that is used to discover the underlying topics that are present in a collection of documents. It does this by analyzing the words in the documents and grouping them into clusters of related words, called "topics." Each topic is represented by a distribution of words, and each document can be represented by a distribution of topics.

Topic modeling is useful for a variety of tasks, such as information retrieval, document classification, and text summarization. It can also be used to discover latent structure in the data, such as the main themes or topics that are present in a collection of documents.

There are several algorithms that can be used for topic modeling, such as Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA).

Q. What is NLP?
Natural Language Processing (NLP) is a field of artificial intelligence that deals with the interaction between computers and humans through the use of natural language. It involves the use of computational techniques to process and analyze large amounts of natural language data, with the goal of understanding and extracting useful information from it.

NLP is used in a wide range of applications, including language translation, text classification, sentiment analysis, and question answering. It has many practical applications, such as in chatbots, voice assistants, and customer service systems.

Q. What is tokenization?

Tokenization is the process of breaking a stream of text up into individual words, phrases, symbols, or other meaningful elements, known as tokens. Tokenization is an essential preprocessing step in many natural language processing (NLP) applications, as it allows the machine to analyze and work with the text in a more structured and efficient way.

There are many different approaches to tokenization, and the appropriate approach will depend on the specific application and the characteristics of the text being processed.

Q. What are stop words?
Stop words are words that are commonly used in natural language, but are not particularly meaningful in the context of a particular search query or document. Examples of stop words include "the," "a," "and," "is," "are," "was," etc.

Stop words are often removed from text data as a preprocessing step in natural language processing (NLP) and information retrieval tasks, as they do not contribute much to the meaning of a document and can be safely discarded without losing much information. Removing stop words can also reduce the size of the data and make the subsequent processing steps more efficient.

Q. What is stemming ?
Stemming is the process of reducing a word to its base or root form. The goal of stemming is to reduce a word to its most general form, so that it can be recognized as the same word in different variations of inflection or tense.

For example, the stem of the word "jumps" might be "jump," and the stem of the word "stemmer," might be "stem." The stem of a word is not always the same as the root of the word, which is the form from which it is derived, but it is usually a shortened form of the root.

Q. What is lemmatization?
Lemmatization is the process of reducing a word to its base form, known as the lemma. The lemma of a word is the dictionary form of the word, which is typically the infinitive form of a verb or the singular form of a noun.

For example, the lemma of the word "jumps" is "jump," and the lemma of the word "stemmer" is "stemmer."

Q. What is PCA?
Principal Component Analysis (PCA) is a statistical method that is used to reduce the dimensionality of a dataset. It does this by identifying the directions in which the data vary the most, and then projecting the data onto a new set of axes that are aligned with these directions.

The new axes, known as principal components, are ranked by the amount of variation they capture in the data. The first principal component captures the most variation, the second principal component captures the second most, and so on. By retaining only the first few principal components, it is possible to reduce the dimensionality of the data while retaining a large portion of the variation.

Q. What is dimensionality reduction?
Dimensionality reduction is the process of reducing the number of dimensions, or features, in a dataset. The goal of dimensionality reduction is to remove redundant or irrelevant features from the data, while preserving as much of the information in the data as possible.

There are many techniques that can be used for dimensionality reduction, including feature selection, feature extraction, and manifold learning. These techniques can be applied for a variety of purposes, such as data visualization, data compression, and feature engineering.

Q. What are advantages of dimensionality reduction?

- It reduces the time and storage space required.
- The removal of multicollinearity 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.
- Reduce space complexity.
- More interpretable because it removes noise thus provides a simpler explanation.
- To mitigate “curse of dimensionality”.

Q. What is feature selection and feature extraction?
Feature selection is the process of selecting a subset of relevant features for use in model construction. The goal of feature selection is to select a small number of features that are most relevant to the problem, while discarding the irrelevant ones. This can be useful for a number of reasons:

- It can reduce the complexity of the model, and make it easier to interpret.
- It can improve the generalization of the model, by reducing overfitting.
- It can speed up the training of the model, by reducing the number of features that the model needs to process.

Feature extraction, on the other hand, is the process of creating new features from existing ones. This can be done using a variety of techniques, such as principal component analysis (PCA), independent component analysis (ICA), and singular value decomposition (SVD). The goal of feature extraction is to extract the most important information from the data, and express it in a compact form that is easy to use. This can be useful for a number of reasons:

- It can reduce the dimensionality of the data, and make it easier to visualize.
- It can extract the most important features from the data, and discard the less important ones.
- It can identify patterns in the data that may not be immediately apparent.

Q. What is recommendation system?

A recommendation system, or a recommendation engine, is a subclass of information filtering system that seeks to predict the "rating" or "preference" a user would give to an item. Recommendation systems are utilized in a variety of areas, and are most commonly recognized

as playlist generators for video and music services, product recommenders for online stores, or content recommenders for social media platforms.
There are several different approaches to building recommendation systems, including:

- Collaborative filtering: This approach makes recommendations based on the preferences of other users with similar tastes.
- Content-based filtering: This approach makes recommendations based on the characteristics of the items themselves.
- Hybrid systems: These systems combine collaborative filtering and content-based filtering to make recommendations.

Q. What is content based filtering?

Content-based filtering is a technique used in recommendation systems to make recommendations based on the characteristics of the items themselves. It works by analyzing the attributes of the items and recommending items that are similar to ones that the user has liked in the past.

For example, suppose a recommendation system is being used to recommend movies to users. A content-based filtering system might recommend movies to a user based on the genres of movies that the user has previously watched and enjoyed. If the user has a history of watching and enjoying romantic comedies, the system might recommend other romantic comedies to the user.

Q. What is collaborative filtering?

Collaborative filtering is a technique used in recommendation systems to make recommendations based on the preferences of other users. It works by identifying users who have similar tastes, and recommending items that they have liked to the current user.
There are two main types of collaborative filtering:

- User-based collaborative filtering: This approach looks for other users who have similar tastes to the current user, and recommends items that they have liked.
- Item-based collaborative filtering: This approach looks at the items that the current user has liked, and recommends similar items to the user.

Q. What is time series analysis?

Time series analysis is a statistical technique that deals with time series data, or trend analysis. Time series data is a series of data points collected at regular intervals of time. These data points may be related to a variety of topics, such as economics, weather, stock prices, and so on. Time series analysis is used to understand the underlying patterns, trends, and correlations in the data, and to make forecasts based on these patterns. Some common techniques used in time series analysis include trend analysis, seasonality analysis, and autoregressive moving average (ARMA) models.

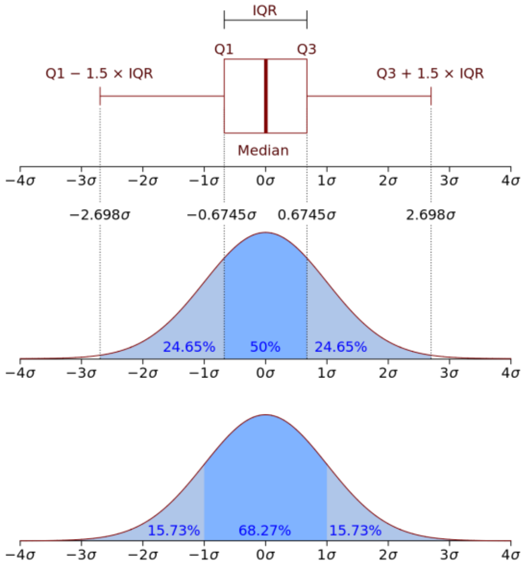
Q. What is aggregate function?

An aggregate function is a function that performs a calculation on a set of values and returns a single value. Aggregate functions are often used in database management systems, spreadsheets, and other software to perform operations such as summing the values in a column or finding the average of a group of values. Some common examples of aggregate functions include SUM, AVG, MIN, and MAX. These functions can be used to perform calculations on groups of rows or columns of data, and the results of the calculations can be used to gain insights into the data or to create reports.

Q. What is interquartile range?

The interquartile range (IQR) is a measure of dispersion in a data set. It is defined as the difference between the 75th percentile and the 25th percentile of a data set. In other words, it is the range of values that encompasses the middle 50% of the data. The interquartile range is a useful measure of dispersion because it is not affected by outliers or extreme values in the data.

To calculate the interquartile range, you first need to find the 75th percentile and the 25th percentile of the data set. These values are also known as the upper quartile and the lower quartile, respectively. The interquartile range is then simply the difference between these two values.



Q. What is Standardization vs Normalization?

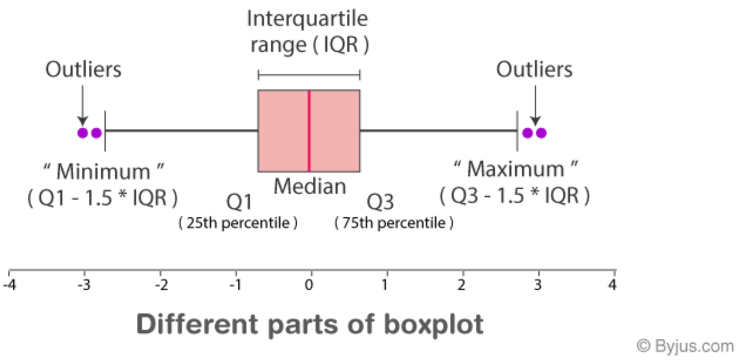
Standardization involves scaling the data so that it has a mean of 0 and a standard deviation of 1. This is done by subtracting the mean of the data from each value and dividing by the standard deviation. Standardization is useful when the data has a Gaussian (normal) distribution, or when the data does not have a clear set of minimum and maximum values.

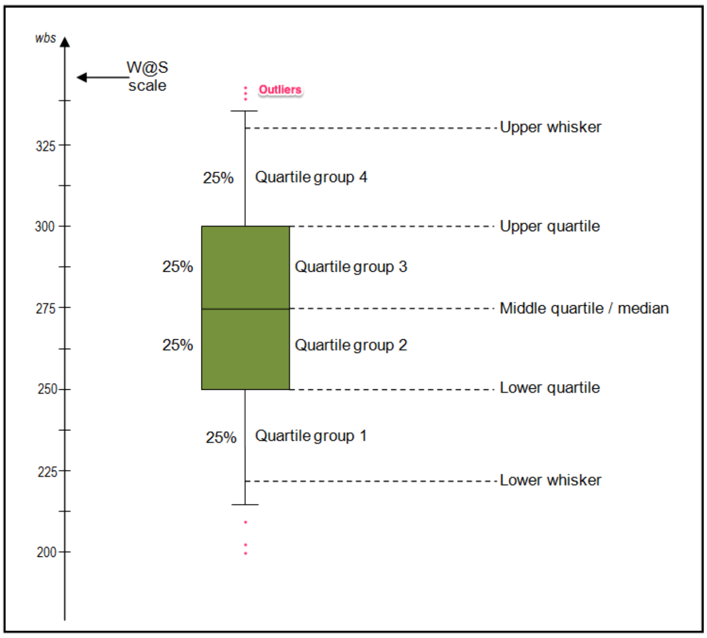
Normalization, on the other hand, scales the data so that it has a minimum value of 0 and a maximum value of 1. This is done by subtracting the minimum value of the data from each value and dividing by the range of the data (i.e., the difference between the minimum and maximum values). Normalization is useful when the data has a clear set of minimum and maximum values, and when the data does not have a Gaussian distribution.

Q. What is box plot?

A box plot is constructed by drawing a box from the first quartile to the third quartile, with a line (called the median) drawn inside the box to represent the second quartile (the median). The "whiskers" of the plot are lines that extend from the box to the minimum and maximum values of the data set. Outliers, which are values that fall outside of the range defined by the first and third quartiles, are represented by dots outside of the whiskers.

Box plots are useful for visualizing the distribution of a data set and identifying outliers. They are often used in statistical analysis and data mining, and are particularly useful for comparing the distribution of data between different groups or categories.





Q. What is view in SQL?

A view is a virtual table that is based on a SELECT statement. A view contains no data itself, but rather displays data that is stored in other tables. A view can be thought of as a "saved SELECT statement," since it is a SELECT statement that has been saved in the database with a name and can be used to retrieve data just like a regular table.

One of the main benefits of views is that they can be used to simplify complex queries by breaking them down into smaller, more manageable pieces. This can make it easier to understand and maintain the queries, as well as to reuse parts of the query in multiple places.

```
CREATE VIEW view_name AS
SELECT column1, column2, ...
FROM table_name
WHERE condition;
```

Q. What are Eigen Vectors and Eigen Values?

Eigenvectors and eigenvalues are useful for understanding the behavior of certain types of linear transformations, such as rotations, stretches, and compressions. They can also be used to solve systems of linear equations, to decompose matrices, and to diagonalize certain types of matrices.

To find the eigenvectors and eigenvalues of a matrix, you need to solve the matrix's characteristic equation. The characteristic equation is a polynomial equation whose roots are the eigenvalues of the matrix. Once the eigenvalues have been found, the corresponding eigenvectors can be found by solving a set of linear equations.

Q. What is Correlation and Covariance?

Correlation and covariance are measures of the relationship between two variables.

Correlation is a measure of the strength and direction of a linear relationship between two variables. It ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no correlation. Positive correlation means that the variables are positively related, meaning that as one variable increases, the other variable also increases. Negative correlation means that the variables are negatively related, meaning that as one variable increases, the other variable decreases.

Covariance is a measure of the degree to which two variables are linearly related. It is calculated as the product of the standard deviations of the two variables and the Pearson correlation coefficient between the variables. Like correlation, covariance can range from -1 to 1, with 0 indicating no relationship. However, unlike correlation, covariance is not standardized, so it is not comparable between variables with different scales

Q. What is iloc and loc in python?

iloc and loc are attributes in the Pandas library in Python that are used to slice and index data in a DataFrame.

iloc stands for integer-location based indexing, and is used to slice data based on the integer indices of the rows and columns. It is used when you have a DataFrame with integer-based row labels, and you want to slice the data based on these labels.
eg:

```
import pandas as pd
# Create a sample DataFrame
df = pd.DataFrame({'A': [1, 2, 3], 'B': [4, 5, 6], 'C': [7, 8, 9]})
# Print the first two rows of the DataFrame
print(df.iloc[:2])
# Print the second and third rows of the DataFrame
print(df.iloc[1:3])
# Print the first and third rows of the DataFrame
print(df.iloc[[0, 2]])
```

loc stands for label-location based indexing, and is used to slice data based on the row and column labels. It is used when you have a DataFrame with non-integer-based row labels, and you want to slice the data based on these labels.

eg:

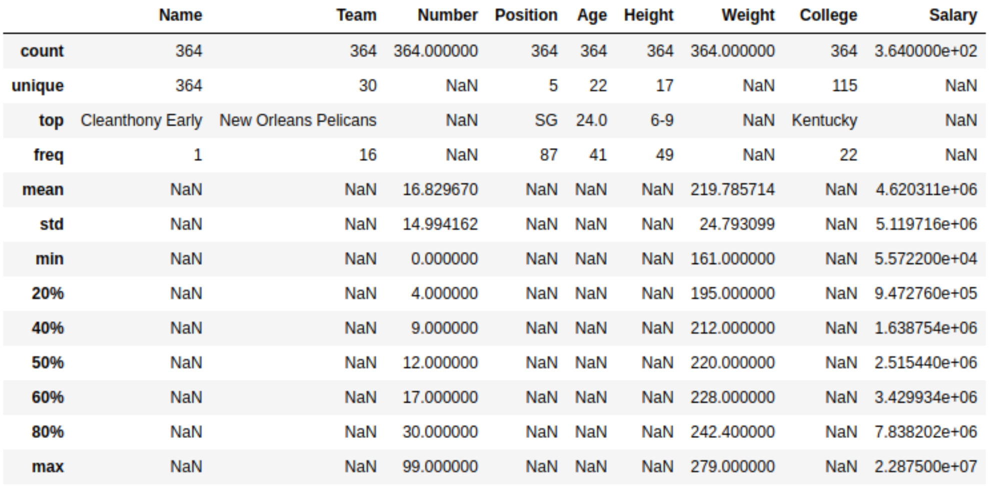
```
import pandas as pd
# Create a sample DataFrame
df = pd.DataFrame({'A': [1, 2, 3], 'B': [4, 5, 6], 'C': [7, 8, 9]}, index=['a', 'b', 'c'])
# Print the row with label 'a'
print(df.loc['a'])
# Print the rows with labels 'a' and 'c'
print(df.loc[['a', 'c']])
# Print the rows with labels 'a' and 'c' and the column with label 'B'
```

print(df.loc[['a', 'c'], 'B'])

68. What df.describe() and df.info show in panda?

In a Pandas DataFrame, the describe method returns a summary of statistical information for the numeric columns in the DataFrame. This includes the count, mean, standard deviation, minimum, maximum, and quartiles of the data.

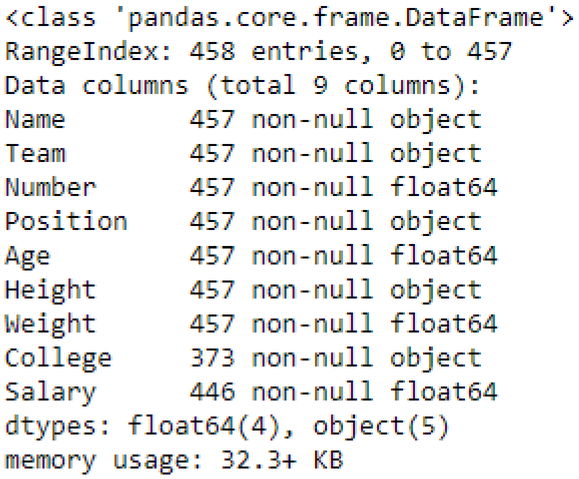
df.describe() - This will return a summary of statistical information for the numeric columns in the DataFrame.



	Name	Team	Number	Position	Age	Height	Weight	College	Salary
count	364	364	364.000000	364	364	364	364.000000	364	3.640000e+02
unique	364	30	NaN	5	22	17	NaN	115	NaN
top	Cleanthony Early	New Orleans Pelicans	NaN	SG	24.0	6-9	NaN	Kentucky	NaN
freq	1	16	NaN	87	41	49	NaN	22	NaN
mean	NaN	NaN	16.829670	NaN	NaN	NaN	219.785714	NaN	4.620311e+06
std	NaN	NaN	14.994162	NaN	NaN	NaN	24.793099	NaN	5.119716e+06
min	NaN	NaN	0.000000	NaN	NaN	NaN	161.000000	NaN	5.572200e+04
20%	NaN	NaN	4.000000	NaN	NaN	NaN	195.000000	NaN	9.472760e+05
40%	NaN	NaN	9.000000	NaN	NaN	NaN	212.000000	NaN	1.638754e+06
50%	NaN	NaN	12.000000	NaN	NaN	NaN	220.000000	NaN	2.515440e+06
60%	NaN	NaN	17.000000	NaN	NaN	NaN	228.000000	NaN	3.429934e+06
80%	NaN	NaN	30.000000	NaN	NaN	NaN	242.400000	NaN	7.838202e+06
max	NaN	NaN	99.000000	NaN	NaN	NaN	279.000000	NaN	2.287500e+07

On the other hand, the info method returns a summary of the DataFrame, including the data types of the columns, the number of non-null values, and the memory usage of the DataFrame.

df.info() - This will return a summary of the DataFrame, including the data types of the columns, the number of non-null values, and the memory usage of the DataFrame.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 458 entries, 0 to 457
Data columns (total 9 columns):
Name        457 non-null object
Team        457 non-null object
Number      457 non-null float64
Position    457 non-null object
Age         457 non-null float64
Height      457 non-null object
Weight      457 non-null float64
College     373 non-null object
Salary      446 non-null float64
dtypes: float64(4), object(5)
memory usage: 32.3+ KB
```

Both the describe and info methods can be useful for getting a quick overview of the data in a Pandas DataFrame.

69. Can we write every loop as list comprehension?

Every list comprehension can be rewritten in for loop, but every for loop can’t be rewritten in the form of list comprehension.

Not every loop can be written as a list comprehension. List comprehensions are a concise way to create a list based on a single line of code, and they are often used to apply a function to a sequence of elements or to filter a sequence of elements.
some loops cannot be written as list comprehensions because they include conditions or statements that cannot be expressed in a single line of code. For example, the following loop cannot be rewritten as a list comprehension:

```
# For loop
numbers = [1, 2, 3, 4, 5]
squares = []
for number in numbers:
    if number % 2 == 0:
        squares.append(number ** 2)

print(squares)
```

Q. What is a pivot table?

A pivot table is a tool in data processing that allows you to rearrange and summarize data in a table format. It can be used to sort, count, total, or average the data, as well as create new calculated fields from the data.
Pivot tables are especially useful when you have a large dataset and you want to quickly understand and analyze it in different ways. You can use pivot tables to create charts and graphs that can help you visualize the data and see trends or patterns more clearly.