1. **What is the difference between WHERE & HAVING Clause?**

WHERE Clause is used to filter the records from the table based on the specified condition. HAVING Clause is used to filter records from the groups based on the specified condition.

2. **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.

3. **What is the concept of bagging?**

   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.

4. **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).

   Stacking ***mainly differs from bagging and boosting on two points. First stacking often considers heterogeneous weak learners***

5. **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.

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.

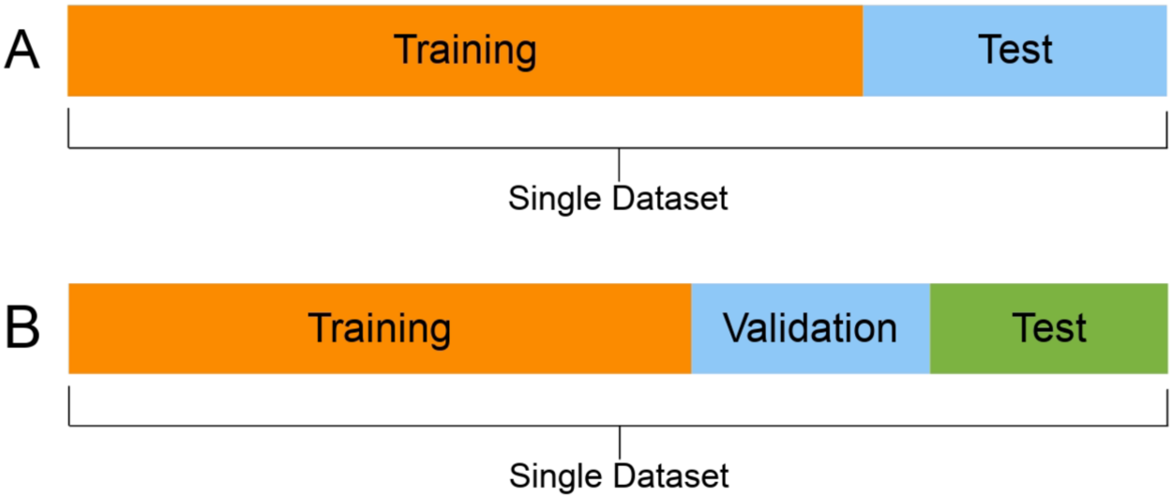
7 most common types - Holdout, K-fold, Stratified k-fold, Rolling, Monte Carlo, Leave-p-out, and Leave-one-out method

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. 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.

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

Create a drop-down list

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.

10. 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/

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

Emp_Details and Emp_Salary.

NOTE: Please check again

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

12. 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’

13. What is the difference between tuple and set?

Tuple is immutable. In Sets, 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.

14. 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.

15. 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 mean of the population*. As the sample size increases, the range of interval values will narrow, meaning that you know that mean with much more accuracy compared with a smaller sample.**

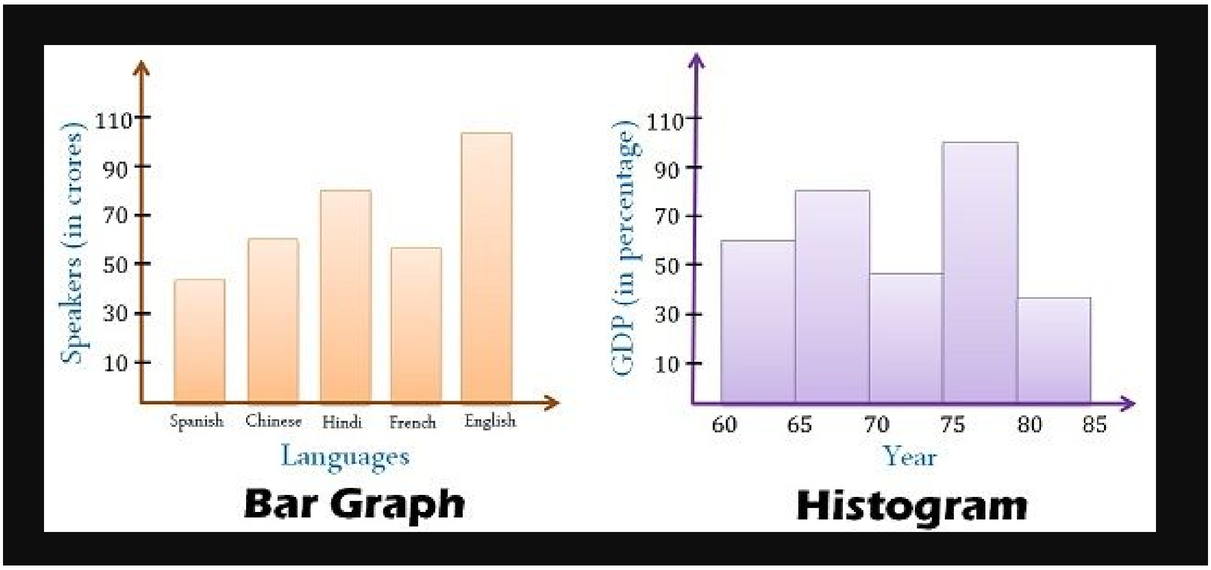
16. What is Polymorphism in Python?

What is Polymorphism: The word polymorphism means having many forms. In programming, polymorphism means the same function name (but different signatures) being used for different types. The key difference is the data types and number of arguments used in function.

https://www.geeksforgeeks.org/polymorphism-in-python/

17. 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.



18. 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.

19. 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.

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.

20. Explain data cleaning in brief

Data Cleaning is the removal of unwanted observations.

https://www.geeksforgeeks.org/data-cleansing-introduction/

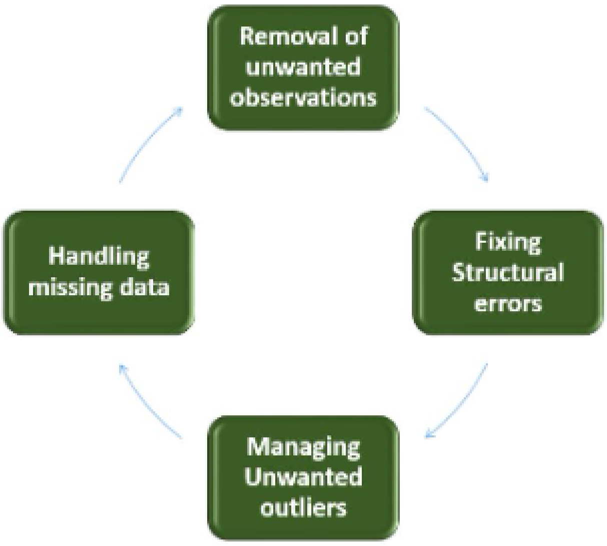
Steps involved in Data Cleaning:

(i) Removal of unwanted observations

(ii) Fixing Structural errors

(iii) Managing Unwanted outliers

(iv) Handling missing data



21. 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.

22. 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.

23. 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.

24. What is model explainability?

Model explainability refers to the concept of being able to understand the machine learning model.

e.g. If a healthcare model is predicting whether a patient is suffering from a particular disease or not.

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.

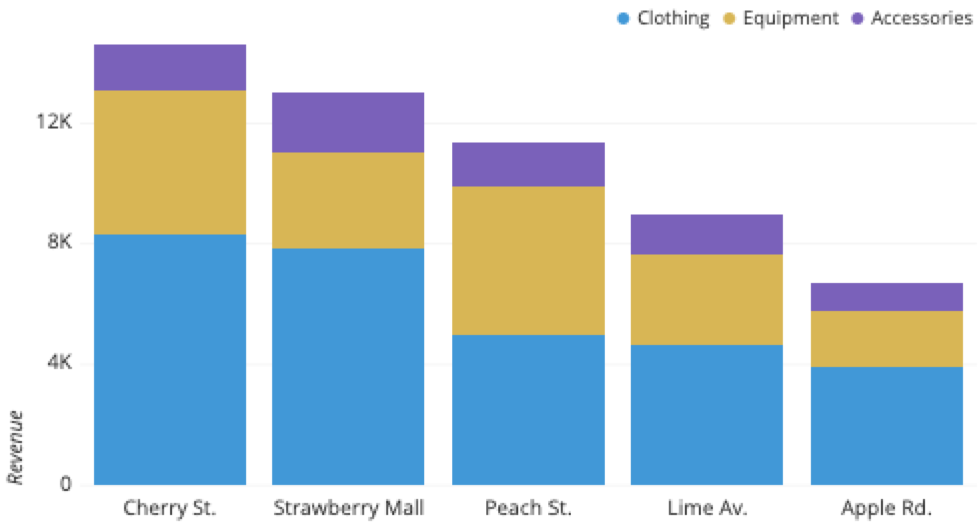
(iii) Model Explainability becomes important while debugging a model during the development phase.

(iv) Model Explainability is critical for getting models to vet by regulatory authorities like FDA

https://www.analyticsvidhya.com/blog/2021/11/model-explainability/

25. 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 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/

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

https://www.boopathirajan.com/what-are-the-objects-can-be-created-using-create-statement-in-mysql/

Following objects are created using CREATE statement:

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

27. 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.

28. 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.

29. 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.

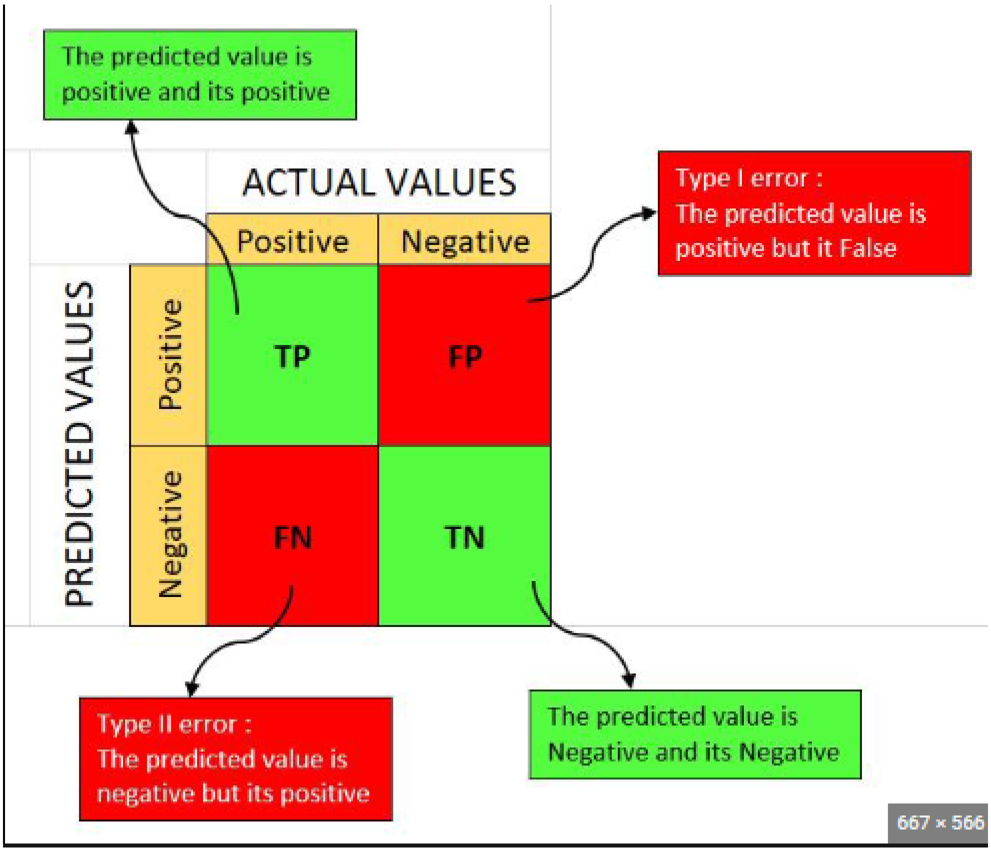
It is called a "*bag*" of words, because any information about the order or structure of words in the document is discarded. The model is only concerned with whether known words occur in the document, not where in the document.

30. 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

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

31. 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.

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.

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

32. What is the significance of the apply method?

The apply() method allows you to apply a function along one of the axis of the DataFrame, default 0, which is the index (row) axis.

33. What is a Self-Join?

SELF JOIN: As the name signifies, in SELF JOIN a table is joined to itself.

In other words we can say that it is a join between two copies of the same table.

34. 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.

35. 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.

36. 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

37. r2 vs adjusted r2?

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.

38. 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.

39. 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.

40. 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

41. 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”

42 .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”

43. 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.

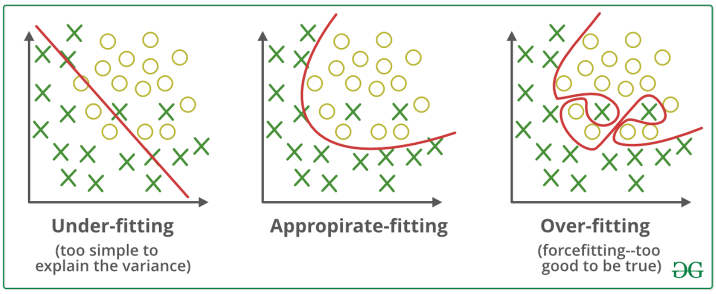
F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

44. 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**

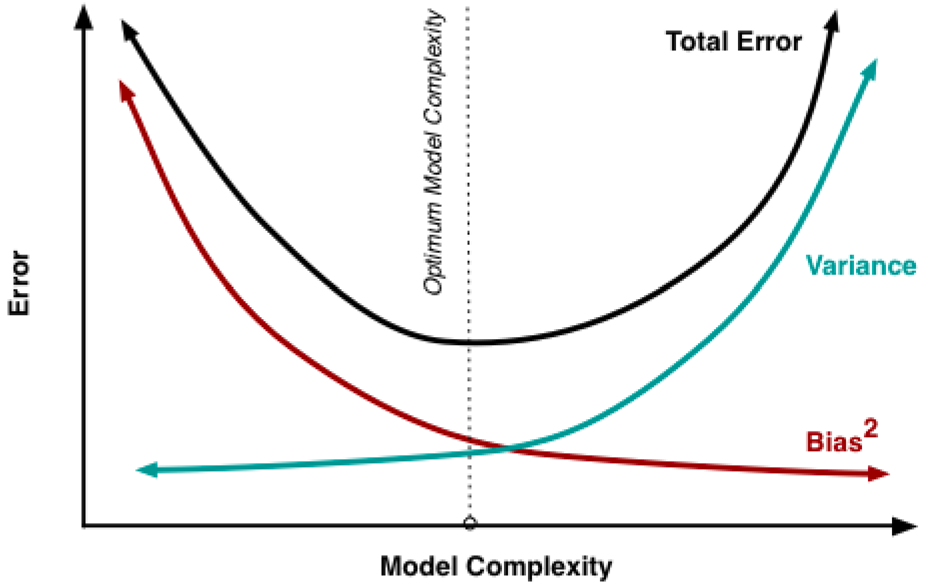
45. What is underfitting in machine learning?

Underfitting: *A statistical model or a machine learning algorithm is said to have underfitting when it cannot capture the underlying trend of the data*, Note: high bias and low variance


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

46. 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.



47. 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.

Q1. 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.

Q2. What are Weak Learners?

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).

Q3)Given a list, write a Python program to swap first and last element of the list?

```
# Swap function def swapList(newList): size = len(newList) # Swapping temp = newList[0]
newList[0] = newList[size - 1] newList[size - 1] = temp return newList # Driver code
newList = [12, 35, 9, 56, 24] print(swapList(newList))
```

Q4. 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))

Q5. 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.

Q1. What are the assumptions of linear regression?Question

MasterSolution Linear relationship. No or little multicollinearity. No auto-correlation. Homoscedasticity.

Q2. Explain Line chart, stacked bar chart and stacked column chart.Question

MasterSolution 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.

Q3. Given a list, write a Python program to swap first and last element of the list.Question MasterSolution # Swap function def swapList(newList): size = len(newList) # Swapping temp = newList[0] newList[0] = newList[size - 1] newList[size - 1] = temp return newList # Driver code newList = [12, 35, 9, 56, 24] print(swapList(newList))

Q4. Write a python program to find factorial of a number.Question MasterSolution def factorial(n): if n < 0: return 0 elif n == 0 or n == 1: return 1 else: fact = 1 while(n > 1): fact *= n n -= 1 return fact # Driver Code num = 5; print("Factorial of",num,"is", factorial(num))