DATA SCIENCE INTERVIEW QUESTIONS AND ANSWERS

1.What does the term Data Science mean?

Data Science is an interdisciplinary field that uses scientific methods, algorithms and systems to extract knowledge and insights from structured and unstructured data. It combines the principles and practices from a variety of fields such as mathematics, statistics, computer engineering and more.

The data science life cycle looks something like this:

- First, the problem is defined and the data needed for the problem is outlined.
- After that, the necessary data is collected through various sources.
- Then, the raw data collected is cleaned for inconsistencies and missing values.
- After that, the data is explored and a summary of the insights is collected.
- The cleansed data is then run through different algorithms such as text mining, recognition patterns, predictive analytics, etc.
- Finally, reports, charts, graphs, and other visualization techniques are used to present the results to the business stakeholders

2.Is there any difference between data science and data analytics?

Data science uses various tools and techniques including data analytics to gather meaningful insights and present them to business stakeholders. On the other hand, data analytics is one of the techniques that analyzes raw data to determine trends and patterns. These trends and patterns can help guide businesses in making effective and efficient decisions. Data analytics uses historical and present data to understand current trends. Whereas, data science uses predictive analytics to determine future problems and drive innovations. Answering this data science interview question can distinguish you from the rookies.

3.Mention some techniques used for sampling and their main advantages.

Sampling is at the core of data science and hence, this data science interview question gives you the opportunity to display your core knowledge. When the data set is very large in size, it is not feasible to conduct an analysis on the entire data set. In such cases, it is critical to select a sample from the given population and conduct data analytics on the selected dataset. This requires caution as a representative sample that represents the true characteristics of the entire population must be selected. The two main sampling techniques used as per statistical needs are:

- Probability samplings such as cluster sampling, random sampling, and stratified sampling

- Non-probability samplings such as quota sampling, convenience sampling, and snowball sampling

4.

Outline the differences between supervised and unsupervised learning.

This is an important data science statistics interview question. Let's outline the differences:

Supervised	Unsupervised
Known and labeled datasets are used as input	Unlabeled datasets are used as input
Has a feedback mechanism	Doesn't have a feedback mechanism
Common algorithms include decision trees, support vector machines, and logistic regression	Common algorithms include apriori algorithm, k-means clustering, and hierarchical clustering
Solves classification and regression problems	Solves clustering, association, and dimensionality reduction problems

5.

Mention the conditions for underfitting and overfitting.

Underfitting: Underfitting means that the statistical model does not fit the existing data set. Underfitting occurs when less training data is provided. The statistical model in underfitting is extremely weak in identifying the relationship in the data and thus, unable to identify any underlying trends. Underfitting can ruin the accuracy of the machine learning model. It can be avoided if more data is used and the number of features is reduced by using feature selection.

Overfitting: A statistical model is overfitted when a lot of data is used to train it. When too much data is used the model learns from the noise and inaccurate data as well, resulting in the inability of the model to categorize the data accurately. Overfitting occurs when non-parametric and non-linear methods are used. Solutions include using a linear algorithm and using parameters such as maximal depth.

Sometimes simple data science interview questions like the above can catch you off-guard, make sure you are prepared with such questions.

6.

What is imbalanced data?

When there is an unequal distribution of data across categories, the data is said to be imbalanced. Imbalanced data produces inaccurate results and model performance errors. Additionally, when training a model using an imbalanced dataset, the model pays more attention to the highly populated classes and poorly identifies the less populated classes.

7.

Which language is more popular for data science?

Python is the most popular language for data science, followed by R. This is so because Python provides great functionality for statistics, mathematics and scientific functions. Further, it offers rich libraries for data science applications.

8.

What are the three types of big data?

Structured, semi-structured, and unstructured data are the three types of data in big data.

9.

What is supervised learning?

Supervised learning is a type of machine learning where the algorithm is trained on a labeled dataset, either to classify data or predict outcomes.

10.

Name five V's of big data?

Volume, Velocity, Variety, Veracity, and Value are the five V's of big data.

11.

Can we process raw data more than once?

Raw data can be processed more than once. This is often done to clean or transform the data.

12.

What type of database is MongoDB?

MongoDB is a form of a NoSQL database.

13.

Define Enumeration.

Enumeration is a process of assigning a numerical value to each member of a set or group. This can be used to count things or to identify members of a group.

14.

Is MICE a data imputation package?

MICE is a data imputation package, which can be used to fill in missing values in data.

15.

What is an Outlier?

Outliers are values that deviate significantly from the rest of the data and are sometimes caused by errors.

16.

Which language does relational database use?

Relational databases use a language called SQL (Structured Query Language) that is useful in manipulating data in the database.

17.

Which one is better for text analytics: R or Python?

Python would be best suited for text analytics because of rich libraries like Pandas.

18.

What does the P value greater than 0.5 indicate?

A P-value greater than 0.5 indicates that the null hypothesis is more likely true than the alternative hypothesis.

19.

Is Tuple an immutable data structure?

Yes, a tuple is an immutable data structure, which means that once it is created, it cannot be modified.

20.

How many expressions does a lambda function have?

A lambda function has only one expression.

21.

What is NLP?

NLP stands for Natural Language Processing, which is a process of extracting information from text data.

22.

Define disaggregation of data.

Disaggregation of data is the process of breaking down data into smaller, more manageable pieces.

23.

How to normalize variables?

To normalize variables, you need to standardize the data so that each variable has a mean of 0 and a standard deviation of 1.

24.

What is deep learning?

Deep learning is a subset of machine learning that that enables machines to learn from experience and understand the world in terms of a hierarchy of concepts. Deep learning can be used to build intelligent systems that can make decisions and predictions based on data.

25.

What is vertical representation of data called?

The vertical representation of data is known as column, while the horizontal representation of data is known as rows.

26.

What is the meaning of K in K-mean algorithm?

The "K" in K-means algorithm stands for the number of clusters that the algorithm will form. K-means is an unsupervised learning algorithm that clusters data into K distinct clusters.

INTERMEDIATE DATA SCIENCE INTERVIEW QUESTIONS & ANSWERS

1.

How do you explain variance in data science?

Variance in data science is a measure of the spread of a dataset. It is calculated by taking the average of the squared differences between each data point and the mean of the dataset.

2.

What is the primary key in SQL?

A primary key is a column in a table that we can use to identify all rows uniquely.

3.

Define Random forest algorithm.

An ensemble learning algorithm which is based on decision trees. Random forest is a machine learning algorithm for classification and regression.

4.

Are correlation and covariance interrelated?

Correlation and covariance are two measures of how two variables are related. Correlation is a measure of how two variables vary together, while covariance is a measure of how two variables vary in relation to each other.

5.

How to compare the distance b/w two binary strings?

The Hamming distance and the Levenshtein distance are two methods for comparing the distance between two binary strings. The number of bits that differ between two strings is defined as the Hamming distance. The number of edit operations (insert, delete, or replace) required to transform one string into another is represented by the Levenshtein distance.

6.

How many different data types are supported by Tableau?

Tableau supports a variety of data types, including numeric, string, date, and geographic data.

7.

What is R2 metrics?

R2 metrics is a statistical measure that represents the proportion of the variance in a data set that is explained by a linear regression model.

8.

What are some ways to measure the accuracy of a model?

There are several ways to measure the accuracy of a model, including the mean squared error, the mean absolute error, and the R-squared value.

9.

What is data mining?

Data mining is the process of obtaining useful information from large data sets.

10.

What is the difference between a classification, regression, and clustering model?

Classification, regression, and clustering are all types of machine learning models. Classification models are used to predict categorical values, regression models are used to predict numerical values, and clustering models are used to group data points into clusters.

11.

When is re-sampling needed?

When the data accuracy is questionable or there is uncertainty about the parameters of the given population, resampling is done. It is a method to improve the accuracy of the sample data and the quality of the model by training it on different datasets to handle variations.

12.

Can we use the KNN algorithm for both regression and classification problem statements?

Yes, this algorithm can be used for both classification and regression.

13.

What are descriptive statistics?

Descriptive statistics are numerical methods used to summarize and describe a given data set. They are used to quantify the data in order to better understand its characteristics.

14.

Name the types of sampling bias

Some of the popular sampling bias are - selection bias, under-coverage bias, non-response bias, survivorship bias, availability bias, among others.

15.

What is the Nunique function?

Nunique function is an aggregation function in PostgreSQL. It used to calculate the number of unique values in a data set.

16.

What is the purpose of bagging ?

Bagging, a machine learning technique, improves the accuracy and stability of models by combining the predictions from multiple models.

17.

Is SVM a classification Algorithm?

Yes, Support Vector Machine or SVM is a classification algorithm.

18.

What is a decision tree?

A decision tree is a tree-like structure where each node represents a decision. The leaves of the tree represent the output of the decision tree.

19.

What is data wrangling?

Data wrangling is the procedure of cleaning and preparing data for analysis.

20.

What is the difference between univariate and bivariate analysis?

Univariate analysis has one variable, whereas bivariate analysis has two variables. Univariate analysis is used to describe data and find patterns within it. On the other hand, bivariate data focuses on finding how two variables are related to each other.

21.

What is data visualization?

The method of creating visual representations of the data is referred to as data visualization.

22.

Define Pandas Index.

A Pandas Index is a mutable, ordered set that can be used to index data in a Pandas DataFrame.

23.

What is exploratory data analysis?

Exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. This involves visualizing the data, computing summary statistics, etc.

24.

Are data dredging and data snooping the same thing?

Yes. Both have the same meaning.

25.

In exponential smoothing, what is the sum of weights?

In exponential smoothing, the sum of weights is equal to 1.

26.

What is the full form of ANOVA?

The full form of ANOVA is ‘Analysis of Variance’.

27.

What is the Central Limit theorem?

The Central Limit Theorem states that the distribution of the sample will be normal if the sample size is large enough.

28.

What is the purpose of the head and tail function?

The head and tail function is used to return the first or last n elements of a vector, respectively. This can be useful to see the elements at the beginning or end of a dataset.

29.

What is data augmentation? Give examples.

In machine learning, data augmentation is the process of artificially increasing the size of your training dataset by adding new, modified, or synthetic data samples. This can be done by adding more samples of the existing data, or by synthesizing new samples from the existing data.

30.

What is the append method?

The append() method is used to add data to a Panda DataFrame.

31.

List some benefits of using TensorFlow

There are many advantages to using TensorFlow. Some of the most notable ones include the ability to scale to large datasets, the ability to use GPUs for speed, and the ability to automatically differentiate between various data types.

32.

Why does skewed distribution occur?

A skewed distribution occurs when the data in a data set are not evenly distributed.

33.

What is the range of recall ratio?

The range of recall ratio is typically 0 to 1, with higher values indicating a better recall rate.

34.

Name a few clustering algorithms

There are a number of different clustering algorithms, including k-means clustering, hierarchical clustering, fuzzy C-means clustering, density-based clustering, etc.

35.

Is Matplotlib an open-source library?

Yes, Matplotlib is an open-source library.

36.

What is the difference between the long and wide formats of data?

Long Format	Wide Format
One row shows one response for one subject. For different responses, different rows are used.	Different responses about a subject part of different columns
For data recognition, rows are used as groups.	For data recognition, columns are used as groups.
Mostly used in R analyses and for writing into log files after trials.	Mostly used in statistical packages for repeated measures ANOVA.

37.

Mention feature selection methods for selecting the right variables

Through this data science interview question, the interviewer wants to understand whether you have experience handling critical situations. The two main methods of feature selection are wrapper and filter methods.

Wrapper method includes:

- Forward selection: One feature is tested at a time and added till a good fit is achieved
- Backward selection: All features are tested and those not fit are removed to find which fits best.
- Recursive feature elimination: Different features are checked and their pairs are tested to see how they work together recursively.
- Wrapper methods need high-end computers and a lot of labor if the data sets for analysis are huge.

Filter method includes:

- Chi-square
- ANOVA
- Linear discrimination analysis
- Filter methods involve cleaning up the data. In order to select the most suitable features, various statistical methods are used by these filter methods.

ADVANCED DATA SCIENCE INTERVIEW QUESTIONS AND ANSWERS

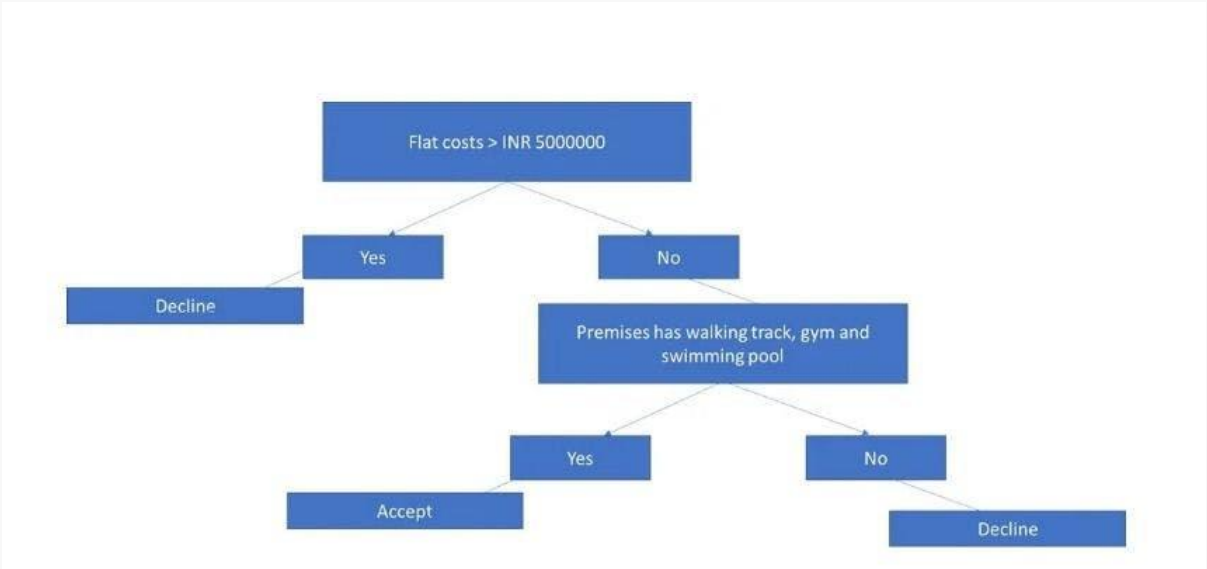
1.

Outline the steps for building a decision tree.

This data science interview question establishes your own decision-making prowess.
Below are the steps for building a decision tree:
For input, take the whole data set

- Calculate the entropy of the class variables and predictor attributes
- Calculate entropy after splitting the attributes
- Calculate information gain of all attribute splits
- Select as root node the attribute with the highest information gain
- Repeat the process for all branches until you finalize the decision node of each branch

For example, if you want to make a decision tree for deciding whether you should buy a certain flat or not, this is how the decision tree may look like:



We can see from the decision tree that the flat will be bought if:
The cost of the flat is less than INR 5000000
The premises has walking track, gym, and swimming pool
2.

Does overfitting occur only when you have a large amount of data for training?

No, overfitting may occur even if the size of data is not large.
3.

Do hybrid Bayesian networks take only continuous variables?

No, hybrid bayesian networks take both continuous and discrete variables as numerical inputs.
4.

Explain neural networks.

Replicated from the neuron of the human brain, neural networks is a technique in AI to teach computers how to process data. They are made up of many interconnected processing nodes, or neurons, that can learn to recognize patterns in input data.
5.

Why do we use cross-validation in machine learning?

Cross-validation is a method of evaluating a machine learning model by training it on a portion of the data and then testing it on another portion of the data. This allows you to assess the accuracy of the model and avoid overfitting.
6.

Give some drawbacks of linear regression model

Some drawbacks of linear regression model are -

- It can't capture non-linear relationships between the independent and dependent variables.
- It is sensitive to outlier data sets.
- It cannot be used to predict categorical outcomes.

7.

How do we measure bias in a model?

Bias in a model can be measured by looking at the difference between the predicted values of the model and the actual values of the data. Bias can be caused by factors, such as selection bias and data leakage.
8.

The process of reducing the size of a decision tree is called?

This process is called pruning.
9.

How can overfitting be avoided?

Overfitting is a problem that can occur when a machine learning model is too complex and does not generalize well to new data. Overfitting can be avoided by using regularization methods such as early stopping and cross-validation.
10.

When should we use linear regression and when should we use logistic regression or another type of model?

Linear regression is a type of machine learning algorithm that is used to predict continuous values. Logistic regression is a type of machine learning algorithm that is useful in predicting binary values.
11.

How do you interpret the coefficients from a linear regression model with multiple predictors (e.g., age and income)?

Both types of regression can be used with multiple predictors, but the interpretation of the coefficients may be different.

12.

How do you overcome survivorship bias?

Survivorship bias is a type of cognitive bias that occurs when people only pay attention to information that confirms their preexisting beliefs. This can lead to distorted conclusions about what is true and what isn't.

One way to overcome survivorship bias is to be aware of it. Pay attention to information that goes against your beliefs and try to understand why that information exists. Be open to the possibility that you might be wrong about something and be willing to change your beliefs if new evidence suggests that you should.

13.

What is a confounding variable?

A confounding variable is an extraneous variable that interacts with the independent and dependent variables, making it difficult to determine the true effect of the independent variable on the dependent variable.

14.

How to calculate the precision rate?

Precision rate is calculated by dividing the number of true positives (TP) by the sum of the true positives and false positives (FP), like so:

Precision Rate = TP / (TP + FP)

15.

What does SMOTE stand for?

SMOTE - Synthetic Minority Oversampling Technique

16.

What is bivariate analysis?

Bivariate analysis is the study of two variables. This can involve things like looking at the relationship between them or predicting one variable based on the other.

17.

How does the K-means algorithm work?

The K-means algorithm works by partitioning a data set into a number of clusters and then assigning each data point to the cluster that is closest to it.

18.

Explain the difference between gradient and gradient descent?

The gradient is a measure of the steepness of a slope. Gradient descent is a method of minimizing a model's error by determining the best gradient.

19.

What is the basic principle of Pareto?

Also known as the 80/20 rule, the Pareto principle states that "80% of the effects result from 20% of the causes." In other sayings, a limited number of factors account for a large proportion of the results.

20.

Why does a neural network require an activation function?

An activation function is a mathematical function that determines the output of a node. The purpose of an activation function is to introduce non-linearity into the network so that it can learn complex relationships.

21.

What is a chi-square test?

A chi-square is a statistical test used to determine whether there is a significant difference between two groups/variables.

22.

Is logistic regression a supervised machine learning algorithm?

Yes, logistic regression is a supervised machine learning algorithm that is used to predict the probability of a binary outcome. The output is a value between 0 and 1 that represents the likelihood of the occurrence of the event.

23.

Define bias in a neural network?

Bias is a term used in machine learning to refer to the error introduced by the simplified assumptions made by the model. A biased model has been oversimplified and does not accurately represent the true relationship between the input and output variables

24.

What if we use a ReLU activation followed by a sigmoid as the final layer?

If we use a ReLU activation and then a sigmoid as the final layer, the output will be a value between 0 and 1. The sigmoid function is used to squash the output of the neurons so that it is interpretable as a probability.

25.

What is inner join in SQL?

The function of inner join is to combine two or more tables. It returns all rows from the tables that have matching values in the specified columns.

26.

Explain A/B testing

A/B testing is an evidence-based approach to making decisions. A/B testing is a way of comparing two groups of data to see which one is better.
27.

How to avoid selection bias?

One way to avoid selection bias is to use a randomized sampling technique. This method randomly selects an equal number of cases from each group and then combines the cases into one data set.
28.

Is PyTorch a deep learning framework?

PyTorch is a deep-learning framework that is used for building and training neural networks.
29.

Write output of this code?

```
calc = lambda m:[i*m for i in range(4)]
print(calc(4))
```

Output:

[0, 4, 8, 12]
30.

Karen has two children, one of whom is a girl. What is the likelihood that the second child will also be a girl?

1/3
31.

What is the formula of standard deviation in binomial distribution?

$\sigma 2=npq$
32.

Given x and y of shapes (10,) and (10,20) respectively, what would be a valid broadcasting statement?

X[:, np.newaxis] + Y
33.

Solve the below code and give its output

```
def fast():
  ls = []
   ls.append(2)
   return ls
print(ls())
```

Output:

231
34.

Based on the below table, write a query to find out the name of the student whose age is 18

StudentData

Student_id	Name	Year	Age	Department
22	James	2022	21	Business
09	Zoe	2019	20	Bio Tech
18	John	2016	22	Computer science
23	Fransica	2021	21	Bio Tech

SELECT Name FROM StudentData WHERE Age = '18'
35.

Based on the above table, write a query to find out the names of those students who are from the Biotech department and are 21 years old

SELECT Name FROM StudentData WHERE Department = 'Bio Tech' and Age = '21'
36.

If you roll a dice three times, what is the probability to get two consecutive sixes?

The probability is 11/216
37.

Using the Euclidean distance formula calculate the distance between the following points P(4, 5) Q (3, 2).

For:

(X1, Y1) = (4, 5)
(X2, Y2) = (3, 2)

$$d = \sqrt{(3-4)^2 + (2-5)^2}$$

$$d = \sqrt{(-1)^2 + (-3)^2}$$

$$d = \sqrt{1+9}$$

$$d = \sqrt{10}$$

$$d = 3.162278$$

38.

Write a code to build ROC curve for model Build

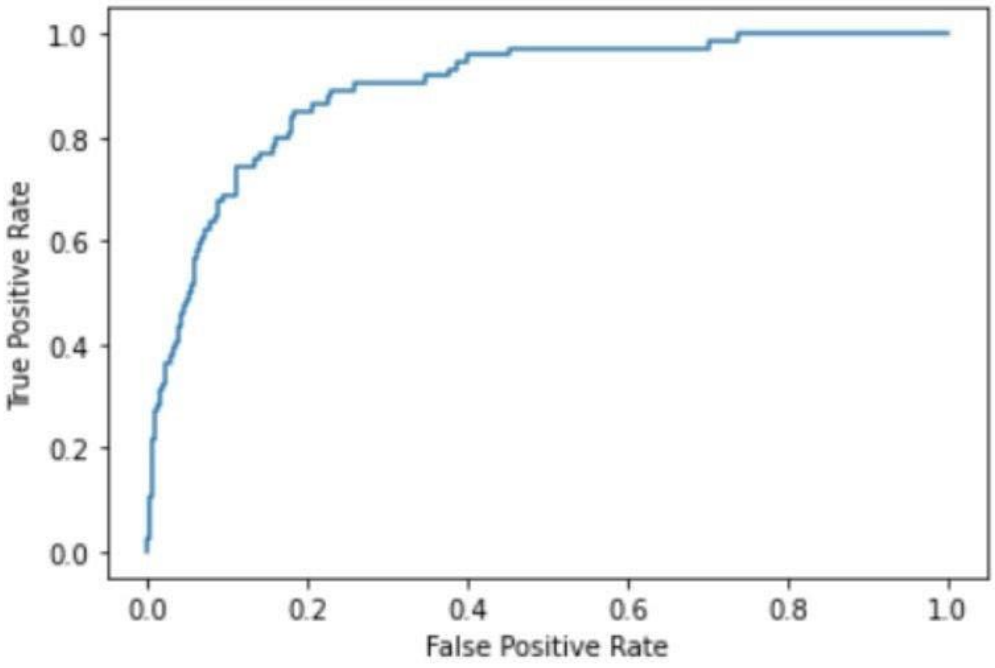
```
[1]  import pandas as pd
     import numpy as np
     from sklearn.model_selection import train_test_split
     from sklearn.linear_model import LogisticRegression
     from sklearn import metrics
     import matplotlib.pyplot as plt
```

```
[2]  url = "https://raw.githubusercontent.com/Statology/Python-Guides/main/default.csv"
     data = pd.read_csv(url)
     x = data[['student', 'balance', 'income']]
     y = data['default']
```

```
     x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=0.2,random_state=0)

     logistic_regression = LogisticRegression()
     logistic_regression.fit(x_train,y_train)
     y_pred_prob = logistic_regression.predict_proba(x_test)[::,1]
     FPR, TPR, _ = metrics.roc_curve(y_test, y_pred_prob)
```

```
[4]  plt.plot(FPR,TPR)
     plt.ylabel('True Positive Rate')
     plt.xlabel('False Positive Rate')
     plt.show()
```


True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0