Data Science

1. Clustering

A way of grouping the data points into different clusters, consisting of similar data points or which groups the unlabelled dataset.

2. K means Clustering

K-means clustering algorithm tries to group similar items in the form of clusters.

3. Sensitivity

Sensitivity is a measure of the proportion of actual positive cases which got predicted as positive

4. Specificity

Specificity is defined as the proportion of actual negatives which got predicted as the negative (or true negative).

5. When to use specified and when to use unspecified

Unspecified codes are used when there isn't much information available about the patient's condition to specifically code it at a particular point in time. Specified are codes for which there is no exact code description for the condition described in the documentation.

6. Labeled and unlabelled data

Labelled data is data that comes with a tag, like a name, a type, or a number. Unlabelled data is data that comes with no tag.

7. ROC curve

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: True Positive Rate. False Positive Rate

8. Supervised and Unsupervised learning

supervised learning is when we teach or train the machine using data that is well labelled. Which means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples(data) so that the supervised learning algorithm analyses the training data (set of training examples) and produces a correct outcome from labelled data.

Unsupervised learning is the training of a machine using information that is neither classified nor labelled and allowing the algorithm to act on that information without guidance so, the task of the machine is to group unsorted information according to similarities, patterns, and differences without any prior training of data.

9. Sensitivity and specificity which is the best

Sensitivity

10. Independent variable

The independent variable is the variable the experimenter manipulates or changes, and is assumed to have a direct effect on the dependent variable.

11. Difference between logistic and linear 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 a discrete output.

12. When to use decision tree

Decision trees are extremely useful for data analytics and machine learning because they break down complex data into more manageable parts. They're often used in these fields for prediction analysis, data classification, and regression.

13. KNN

The abbreviation KNN stands for **"**K-Nearest Neighbour**"**. It is a supervised machine learning algorithm. The algorithm can be used to solve both

classification and regression problem statements. The number of nearest neighbours to a new unknown variable that has to be predicted or classified is denoted by the symbol 'K'.

14. Central limit theorem

The central limit theorem (CLT) states that the distribution of sample means approximates a normal distribution as the sample size gets larger, regardless of the population's distribution. Sample sizes equal to or greater than 30 are often considered sufficient for the CLT to hold.

15. Cross-validation

Cross-validation is a resampling method that uses different portions of the data to test and train a model on different iterations. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice.

16. Anova

Analysis of variance (ANOVA) is an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts: systematic factors and random factors. The systematic factors have a statistical influence on the given data set, while the random factors do not.

17. Example of anova -

Suppose an independent variable is social media use, and you assign groups to low, medium, and high levels of social media use to find out if there is a difference in hours of sleep per night.

18. Hypothesis testing

Hypothesis testing is a part of statistical analysis, where we test the assumptions made regarding a population parameter.

It is generally used when we were to compare:

- a single group with an external standard
- two or more groups with each other

19. What is overfitting

occurs when a statistical model fits exactly against its training data. When this happens, the algorithm unfortunately cannot perform accurately against unseen data, defeating its purpose.

20. What is logistic regression

Logistic regression is a statistical analysis method used to predict a data value based on prior observations of a data set. A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables.

21. What is linear regression

It is a regression model that estimates the relationship between one independent variable and one dependent variable using a straight line. Both variables should be quantitative.

22. What is the different cross validation method?

Holdout, K-fold, Stratified k-fold, Rolling, Monte Carlo, Leave-p-out, and Leave-one-out method.

23. What are the different types of data visualization?

scatter plots, line graphs, pie charts, bar charts, heat maps, area charts, choropleth maps and histograms.

24. Where to use scatter plot, histogram?

Scatter plot –

When you have paired numerical data.

When your dependent variable may have multiple values for each value of your independent variable.

When trying to determine whether the two variables are related, such as: When trying to identify potential root causes of problems.

Histogram-

- The data are numerical.

- You want to see the shape of the data's distribution, especially when determining whether the output of a process is distributed approximately normally.
- Analysing whether a process can meet the customer's requirements.

25. Precision -

The ability of a classification model to identify only the relevant data points. Mathematically, precision the number of true positives divided by the number of true positives plus the number of false positives.

26. Classifier Performance -

Precision and recall are objective measures of a classifier’s performance. The higher those numbers are, the better the classifier is doing. Unfortunately, precision and recall are often working against each other.

27. Data Visualisation

Data visualization (often abbreviated data viz is an interdisciplinary field that deals with the graphic representation of data. It is a particularly efficient way of communicating when the data is numerous as for example a time series.

28. Data Visualisation Techniques -

Univariate Analysis Techniques for Data Visualization

1. Distribution Plot
2. Box and Whisker Plot
3. Violin Plot

Bivariate Analysis Techniques for Data Visualization

1. Line Plot
2. Bar Plot
3. Scatter Plot

29. Univariate, Bivariate, and Multivariate -

When it comes to the level of analysis in statistics, there are three different analysis techniques that exist. These are –

- **Univariate analysis**
- **Bivariate analysis**
- **Multivariate analysis**

Univariate analysis

Univariate analysis is the most basic form of statistical data analysis technique. When the data contains only one variable and doesn't deal with a causes or effect relationships then a Univariate analysis technique is used.

- Frequency Distribution Tables
- Histograms
- Frequency Polygons
- Pie Charts
- Bar Charts

Bivariate analysis

Bivariate analysis is slightly more analytical than Univariate analysis. When the data set contains two variables and researchers aim to undertake comparisons between the two data sets then Bivariate analysis is the right type of analysis technique.

Bivariate analysis is conducted using –

- Correlation coefficients
- Regression analysis

Multivariate analysis

Multivariate analysis is a more complex form of statistical analysis technique and used when there are more than two variables in the data set.

Commonly used multivariate analysis technique include –

- Factor Analysis
- Cluster Analysis
- Variance Analysis
- Discriminant Analysis
- Multidimensional Scaling
- Principal Component Analysis

30. Random Forest -

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned.

31. Accuracy -

Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positives and false negatives are almost the same. Therefore, you have to look at other parameters to evaluate the performance of your model. For our model, we have got 0.803 which means our model is approx. 80% accurate.

Accuracy = TP+TN/TP+FP+FN+TN

32. Precision -

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all passengers that are labeled as survived, how many actually survived? High precision relates to the low false positive rate. We have got 0.788 precision which is pretty good.

Precision = TP/TP+FP

33. Histogram -

A histogram is a graphical representation that organizes a group of data points into user-specified ranges. Similar in appearance to a bar graph, the histogram condenses a data series into an easily interpreted visual by taking many data points and grouping them into logical ranges or bins.

34. Normal Distribution -

Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve.

35. Bagging and Boosting -

Bagging is a method of merging the same type of predictions. Boosting is a method of merging different types of predictions.

Bagging decreases variance, not bias, and solves over-fitting issues in a model. Boosting decreases bias, not variance.

In Bagging, each model receives an equal weight. In Boosting, models are weighed based on their performance.

36. Bagging

Bagging is an acronym for 'Bootstrap Aggregation' and is used to decrease the variance in the prediction model. Bagging is a parallel method that fits different, considered learners independently from each other, making it possible to train them simultaneously.

37. Boosting

Boosting is a sequential ensemble method that iteratively adjusts the weight of observation as per the last classification. If an observation is incorrectly classified, it increases the weight of that observation. The term 'Boosting' in a layman language, refers to algorithms that convert a weak learner to a stronger one. It decreases the bias error and builds strong predictive models.

38. Noise -

Noise is unwanted data items, features or records which don't help in explaining the feature itself, or the relationship between feature & target. Noise often causes the algorithms to miss out patterns in the data. Noisy data is meaningless data. The term has been used as a synonym for corrupt data.

39. Binomial Distribution -

A distribution where only two outcomes are possible, such as success or failure, gain or loss, win or lose and where the probability of success and failure is same for all the trials is called a Binomial Distribution. The outcomes need not be equally likely.

40. 68 97 99.7 Rule -

In statistics, the 68–95–99.7 rule, also known as the empirical rule, is a shorthand used to remember the percentage of values that lie within an interval estimate in a normal distribution: 68%, 95%, and 99.7% of the values lie within one, two, and three standard deviations of the mean, respectively.

41. Euclidean distance -

It is a distance measure that best can be explained as the length of a segment connecting two points.

42. Data Science Pipeline -

A data science pipeline is the set of processes that convert raw data into actionable answers to business questions. Data science pipelines automate the flow of data from source to destination, ultimately providing you insights for making business decisions.

43. Normal and Binomial Distribution -

Normal distribution describes continuous data which have a symmetric distribution, with a characteristic 'bell' shape.

Binomial distribution describes the distribution of binary data from a finite sample. Thus it gives the probability of getting r events out of n trials.

44. What is the difference between binomial and normal distribution?

The main difference between normal distribution and binomial distribution is that binomial distribution is discrete.

This means that in binomial distribution there are no data points between any two data points. This is very different from a normal distribution which has continuous data points.

45. Best Classifier -

Logistic Regression

Logistic regression is a calculation used to predict a binary outcome: either something happens, or does not. This can be exhibited as Yes/No, Pass/Fail, Alive/Dead, etc.

Independent variables are analyzed to determine the binary outcome with the results falling into one of two categories. The independent variables can be categorical or numeric, but the dependent variable is always categorical. Written like this:

46. Data Processing Steps -

Six stages of data processing

1. Data collection

Collecting data is the first step in data processing. Data is pulled from available sources, including data lakes and data warehouses. It is important that the data sources available are trustworthy and well-built so the data collected (and later used as information) is of the highest possible quality.

2. Data preparation

Once the data is collected, it then enters the data preparation stage. Data preparation, often referred to as "pre-processing" is the stage at which raw data is cleaned up and organized for the following stage of data processing. During preparation, raw data is diligently checked for any errors. The purpose of this step is to eliminate bad data (redundant, incomplete, or incorrect data) and begin to create high-quality data for the best business intelligence.

3. Data input

The clean data is then entered into its destination (perhaps a CRM like Salesforce or a data warehouse like Redshift), and translated into a language that it can understand. Data input is the first stage in which raw data begins to take the form of usable information.

4. Processing

During this stage, the data inputted to the computer in the previous stage is actually processed for interpretation. Processing is done using machine learning algorithms, though the process itself may vary slightly depending on the source of data being processed (data lakes, social networks,

connected devices etc.) and its intended use (examining advertising patterns, medical diagnosis from connected devices, determining customer needs, etc.).

5. Data output/interpretation

The output/interpretation stage is the stage at which data is finally usable to non-data scientists. It is translated, readable, and often in the form of graphs, videos, images, plain text, etc.). Members of the company or institution can now begin to self-serve the data for their own data analytics projects.

6. Data storage

The final stage of data processing is storage. After all of the data is processed, it is then stored for future use. While some information may be put to use immediately, much of it will serve a purpose later on. Plus, properly stored data is a necessity for compliance with data protection legislation like GDPR. When data is properly stored, it can be quickly and easily accessed by members of the organization when needed.

47. Cross Validation -

Cross-validation is a resampling method that uses different portions of the data to test and train a model on different iterations. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice.

48. ANOVA Significance -

ANOVA is helpful for testing three or more variables. It is similar to multiple two-sample t-tests. However, it results in fewer type I errors and is appropriate for a range of issues. ANOVA groups differences by comparing the means of each group and includes spreading out the variance into diverse sources.

49. Balanced and Imbalanced Dataset -

Balanced Dataset: — Let's take a simple example if in our data set we have positive values which are approximately the same as negative values. Then we can balance our dataset in balance.

Imbalanced Dataset: — If there is a very high difference between the positive values and negative values. Then we can say our dataset in Imbalance Dataset.

50. Normalization -

Normalization is a technique often applied as part of data preparation for machine learning. The goal of normalization is to change the values of numeric columns in the dataset to use a common scale, without distorting differences in the ranges of values or losing information. Normalization is also required for some algorithms to model the data correctly.

51. Classification vs Regression -

Classification:

Classification is a process of finding a function which helps in dividing the dataset into classes based on different parameters. In Classification, a computer program is trained on the training dataset and based on that training, it categorizes the data into different classes.

The task of the classification algorithm is to find the mapping function to map the input(x) to the discrete output(y).

Types of ML Classification Algorithms:

- Logistic Regression
- K-Nearest Neighbours
- Support Vector Machines

Regression:

Regression is a process of finding the correlations between dependent and independent variables. It helps in predicting the continuous variables such as prediction of Market Trends, prediction of House prices, etc.

The task of the Regression algorithm is to find the mapping function to map the input variable(x) to the continuous output variable(y).

Types of Regression Algorithm:

- Simple Linear Regression
- Multiple Linear Regression
- Polynomial Regression
- Support Vector Regression

52. Ensemble Learning -

Ensemble learning is a way of generating various base classifiers from which a new classifier is derived which performs better than any constituent classifier.

53. Hypothesis -

Hypothesis testing is a part of statistical analysis, where we test the assumptions made regarding a population parameter. It is generally used when we were to compare: a single group with an external standard. two or more groups with each other.

54. Supervised Learning

Supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to find a mapping function to map the input variable(x) with the output variable(y).

55. Unsupervised Learning

We do not need to supervise model, instead we allow model to work on its own

56. Line Fitting

Line fitting is the process of constructing a straight line that has the best fit to a series of data points.

A line of best fit can be roughly determined using an eyeball method by drawing a straight line on a scatter plot so that the number of points above the line and below the line is about equal (and the line passes through as many points as possible).

57. Line Fitting by Least Squares Regression

The Least Squares Regression Line is the line that minimizes the sum of the residuals squared.

58. Outliers in Linear Regression

Outliers in regression are observations that fall far from the "cloud" of points. These points are especially important because they can have a strong influence on the least squares line.

59. Simple Linear regression

A model that assumes a linear relationship between the input variables (x) and the single output variable (y)

60. Multiple Regression -

Multiple regression is a statistical technique that can be used to analyze the relationship between a single dependent variable and several independent

variables. The objective of multiple regression analysis is to use the independent variables whose values are known to predict the value of the single dependent value.

61. Logistic Regression

Is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model.

62. Nearest Neighbour Classification

is a machine learning method that aims at labeling previously unseen query objects while distinguishing two or more destination classes.

63. K-Nearest Neighbour

Is one of the simplest Machine Learning algorithms based on Supervised Learning technique. The algorithm can be used to solve both classification and regression problems.

64. K means clustering

an algorithm to cluster n objects on attributes into k partitions

65. Naive Bayes algorithm

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

66. Posterior probability = likelihood*prior probability/marginal likelihood

67. Evaluation of model performance

The three main metrics used to evaluate a classification model are accuracy, precision, and recall. Accuracy is defined as the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions.

68. F-measure

Is a measure of a test's accuracy. ... The highest possible value of an F-score is 1.0, indicating perfect precision and recall, and the lowest possible value is 0, if either the precision or the recall is zero.

69. A confusion matrix

Is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.

70. Data sciences

Is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from noisy, structured and unstructured data,

71. Recall (Sensitivity)

Recall is the ratio of correctly predicted positive observations to the all observations in actual class

72. Error

Fraction of negative examples that are incorrectly classified

73. Sensitivity

fraction of positive examples classified as positive.

74. Specificity

Fraction of negative examples classified correctly

75. False Alarm rate

Fraction of negative examples classified as positive.

76. Central Limit theorem

The distribution of sample means approximates a normal distribution as the sample size gets larger, regardless of the population's distribution.

77. Cross-validation

Is a resampling method that uses different portions of the data to test and train a model on different iterations.

78. Analysis of variance (ANOVA)

Is an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts: systematic factors and random factors.

79. Hypothesis testing

Is a part of statistical analysis, where we test the assumptions made regarding a population parameter.

80. Overfitting

Is a modeling error in statistics

81. What is a normal distribution in data science?

Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.

82. Which is an ensemble method?

Ensemble methods are techniques that create multiple models and then combine them to produce improved results. Ensemble methods usually produce more accurate solutions than a single model would.

83. What are the techniques of data visualization?

Data visualization is defined as a graphical representation that contains the information and the data. By using visual elements like charts, graphs, and maps, data visualization techniques provide an accessible way to see and understand trends, outliers, and patterns in data.

Common general types of data visualization:

- Charts
- Tables
- Graphs
- Maps
- Infographics
- Dashboards

Application - Healthcare Industries

A dashboard that visualises a patient's history might aid a current or new doctor in comprehending a patient's health. It might give faster care facilities based on illness in the event of an emergency. Instead than sifting through hundreds of pages of information, data visualisation may assist in finding trends.

84. What is meant by cross-validation in machine learning?

Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample.

The purpose of cross–validation is to test the ability of a machine learning model to predict new data

85. What are the different cross validation methods ?

Holdout, K-fold, Stratified k-fold, Rolling, Monte Carlo, Leave-p-out, and Leave-one-out method.

86. What are the different types of data visualization ?

scatter plots, line graphs, pie charts, bar charts, heat maps, area charts, choropleth maps and histograms.

87. Where to use scatter plot , histogram ?

Scatter plot –

1. When you have paired numerical data.
2. When your dependent variable may have multiple values for each value of your independent variable.
3. When trying to determine whether the two variables are related, such as: When trying to identify potential root causes of problems.

Histogram-

1. The data are numerical.
2. You want to see the shape of the data's distribution, especially when determining whether the output of a process is distributed approximately normally.
3. Analyzing whether a process can meet the customer's requirements.

88. Regression vs. Classification in Machine Learning

Regression and Classification algorithms are Supervised Learning algorithms.

The main difference between Regression and Classification algorithms that Regression algorithms are used to **predict the continuous** values such as

price, salary, age, etc. and Classification algorithms are used to **predict/Classify the discrete values** such as Male or Female, True or False, Spam or Not Spam, etc.

Types of ML Classification Algorithms:

Classification Algorithms can be further divided into the following types:

- Logistic Regression
- K-Nearest Neighbours
- Support Vector Machines
- Kernel SVM
- Naïve Bayes

Types of Regression Algorithm:

- Simple Linear Regression
- Multiple Linear Regression
- Polynomial Regression

89. How do you explain the 68 95 and 99.7 rule?

What is the 68 95 99.7 rule?

About 68% of values fall within one standard deviation of the mean.

About 95% of the values fall within two standard deviations from the mean.

Almost all of the values—about 99.7%—fall within three standard deviations from the mean.