Data Science Interview Questions

1. What is logistic regression? Or State an example when you have used logistic regression recently?

- Ans :- Logistic Regression often referred as logit model is a technique to predict the binary outcome from a linear combination of predictor variables. For example, if you want to predict whether a particular political leader will win the election or not. In this case, the outcome of prediction is binary i.e. 0 or 1 (Win/Lose). The predictor variables here would be the amount of money spent for election campaigning of a particular candidate, the amount of time spent in campaigning, etc.

2. What are Recommender Systems?

- Ans :- A subclass of information filtering systems that are meant to predict the preferences or ratings that a user would give to a product. Recommender systems are widely used in movies, news, research articles, products, social tags, music, etc.

3. Why data cleaning plays a vital role in analysis?

- Ans :- Cleaning data from multiple sources to transform it into a format that data analysts or data scientists can work with is a cumbersome process because - as the number of data sources increases, the time take to clean the data increases exponentially due to the number of sources and the volume of data generated in these sources. It might take up to 80% of the time for just cleaning data making it a critical part of analysis task.

4. Differentiate between univariate, bivariate and multivariate analysis.

- Ans :- These are descriptive statistical analysis techniques which can be differentiated based on the number of variables involved at a given point of time. For example, the pie charts of sales based on territory involve only one variable and can be referred to as univariate analysis. If the analysis attempts to understand the difference between 2 variables at time as in a scatterplot, then it is referred to as bivariate analysis. For example, analysing the volume of sale and a spending can be considered as an example of bivariate analysis. Analysis that deals with the study of more than two variables to understand the effect of variables on the responses is referred to as multivariate analysis.

5. What is Machine Learning?

- Ans :- Machine learning is the science of getting computers to act without being explicitly programmed. Machine learning has given us self-driving cars, practical speech recognition, effective web search, and a vastly improved understanding of the human genome. It is so widespread that unknowingly we use it many a times in our daily life.

6. What is Supervised Machine Learning?

- Ans :- Supervised Machine Learning will be employed for the problem statements where in output variable (Nominal) of interest can be either classified or predicted. Examples: KNN, Naive Bayes, SVM, Decision Tree, Random Forest, Neural Network

7. What is Unsupervised Machine Learning?

- Ans :- In this category of Machine Learning, there won't be any output variable to be either predicted or classified. Instead the algorithm understands the patterns in the data. Examples: Segmentation, PCA, SVD, Market Basket Analysis, Recommender Systems.

8. What is Classification Modeling?

- Ans :- Classification Models are employed when the observations have to be classified in categories and not predicted. Examples being Cancerous and Non-cancerous tumor (2 categories), Bus, Rail, Car, Carpool (>2 categories)

9. Examples of Unsupervised Machine Learning

- Ans:-Segmentation, PCA, SVD, Market Basket Analysis, Recommender Systems

10. Examples of Supervised Machine Learning

- Ans:-KNN, Naive Bayes, SVM, Decision Tree, Random Forest, Neural Network

11. Why is hierarchial clustering called as Agglomerative clustering?

- Ans:-It is because of bottom up approach, where initially each observation is considered to be a single cluster and gradually based on the distance measure inidividual clusters will be paired and finally merged as one

12. When can you say that resultant clusters are good?

- Ans:-When the clusters are as much heterogenous as possible and when the observations within each cluster are as much homogeenous as possible.

13. In which domains can we employ clustering?

- Ans:-None of your data science topics are domain specific. They can be employed in any domain, provided data is available.

14. Example of clustering?

- Ans:-Using variables like income, education, profession, age, number of children, etc you come with different clusters and each cluster has people with similar socio-economic criteria

15. Is normalization of data required before applying clustering?

- Ans:-It would be better if we employ clustering on normalized data as you will get different results for with and without normalization

16. What is the range of Z transformed variable?

- Ans:-Theoretically it will be between - infinity to + inifinity but normally you have values between -3 to +3

17. What is the range of variable when ((x - min(X))/(max(X) - min(X)) normalization technique is employed?

- Ans:-0 to 1 is the range for this normalizaion technique

18. What does summary() command gives?

- Ans:-summary() command gives the distribution for numerical variables and proportion of observations for factor variables

19. What is str() command why is it required to run it?

- Ans:-str() command gives dimensions for your data drame. In addition to this it gives, class of the dataset & class of every variable

20. Packages to read excel files in R?

- Ans:-readxl or xlsx packages can be used to read excel files in R

21. What are linkages in hierarchical clustering?

- Ans:-Linkage is the criteria based on which distances between two clusters is computed. Single, Complete, Average are few of the examples for linkages Single - The distance between two clusters is defined as the shortest distance between two points in each cluster. Complete - The distance between two clusters is defined as the longest distance between two points in each cluster. Average - the distance between two clusters is defined as the average distance between each point in one cluster to every point in the other cluster.

22. How do we decide upon the number of clusters in hierarchial clustering?

- Ans:-In Hierarchial Clustering number of clusters will be decided only after looking at the dendrogram.

23. How to interpret clusterig output?

- Ans:-After computing optimal clusters, aggregate measure like mean has to be computed on all variables and then resultant values for all the variables have to be interpreted among the clusters

24. What is the use of set.seed() function ?

- Ans:-set.seed() function is to reproduce same results if the code is re-run again. Any number can be given within the paranthesis

25. Why is KNN called as non-parametric algorithm?

- Ans:-KNN makes no assumptions about the underlying data (unlike other algorithms, eg. Linear Regression)

26. Why is KNN called as Lazy Algorithm?

- Ans:-There is no or minimal training phase because of which training phase is pretty fast. Here the training data is used during the testing phase.

27. How do we choose the value of K in KNN algorithm?

- Ans:-K value can be selected using sqrt(no. of obs/2), kselection package, scree plot, k fold cross validation

28. Function in R to employ KNN?

- Ans:-knn() can be used from the class package

29. What is the r function to know the number of observations for the levels of a variable?

- Ans:-table() is the r function. It can be also be employed on any variable but it makes sense to employ on a factor variable.

30. What is the r function to know the persentae of observations for the levels of a variable?`

- Ans:-prop.table() employed on top of table() function i.e., prop.table(table()) is the r function. It can be also be employed on any variable but it makes sense to employ on a factor variable.

31. Difference between lapply & sapply function?

- Ans:-lapply returns the ouput as a list whereas sapply returns the ouput as a vector, matrix or array.

32. Can we represent the output of a classifer having more than two levels using a confusion matrix?

- Ans:-We cannot use confusion matix when we have more than two levels in the output variable. Instead, we can use crosstable() function from gmodels package

33. What is Probability?

- Ans:-Probability is given by Number of interested events/Total number of events

34. What is Joint probability?

- Ans:-It is the probability of two events occuring at the same time. Classical example is probability of an email being spam wih the word lottery in it.Here the events are email being spam and email having the word lottery

35. What is the function to perform simple random sampling?

- Ans:-sample() is the function in R to employ Simple Random Sampling

36. Functions to find row & column count in R?

- Ans:-dim() function or nrow() & col() can be used to find row & column count

37. What is the function to compute accuracy of a classifier?

- Ans:-mean() function can be used to compute the accuracy. Within paranthesis actual labels have to compared with predicted labels

38. What is Bayes' Theorem?

- Ans:-Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Mathematically it is given as P(A|B) = [P(B|A)P(A)]/P(B) where A & B are events. P(A|B) called as Posterior Probability, is the probability of event A(response) given that B(independent) has already occured. P(B|A) is the likelihood of the training data i.e., probability of event B(indpendent) given that A(response) has already occured. P(A) is the probability of the response variable and P(B) is the probability of the training data or evidence.

39. What is the assumption of Naive Bayes Classifier?

- Ans:-The fundamental assumption is that each indepedent variable independently and equally contributes to the outcome.

40. What is SVM?

- Ans:-Here we plot each data point in n-dimensional space with the value of each dimension being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the classes very well

41. What are the tuning parameters of SVM?

- Ans:-Kernel, Regularization, Gamma and Margin are the tuning paramers of SVM

42. Explain Kernel in SVM?

- Ans:-Kernel tricks are nothing but the transformations applied on input variables which separate non-separable data to separable data. There are 9 different kernel tricks. Examples are Linear, RBF, Polynomial, etc.

43. Is there a need to convert categorical variables into numeric in SVM? If yes, explain.

- Ans:-All the categorical variables have to be converted to numeric by creating dummy variables, as all the data points have to be plotted on n dimensional space, in addition to this we have tuning paramteters like Kernel, Regularization, Gamma & Margin which are mathematical computations that require numeric variables. This is an assumpton of SVM.

44. What is Regularization in SVM?

- Ans:-The value of Regularization parameter tells the training model as to how much it can avoid misclassifying each training observation.

45. What is Gamma parameter in SVM?

- Ans:-Gamma is the kernel coefficient in the kernel tricks RBF, Polynomial, & Sigmoid. Higher values of Gamma will make the model more complex and overfits the model.

46. What do you mean by Margin in SVM?

- Ans:-Margin is the separation line to the closest class datapoints. Larger the margin width, better is the classification done. But before even achieving maximum margin, objective of te algorithm is to correctly classify data points.

47. What is the SVM package used for SVM in R?

- Ans:-kernlab is the package used in R for implementing SVM in R

48. What is the function name to implement SVM in R?

- Ans:-ksvm is the function in R to implement SVM in R

49. What is a decision tree?

- Ans:-Decision Tree is a superised machine learning algorithm used for classification and regression analysis. It is a tree-like structure in which internal node represents test on an attribute, each branch represents outcome of test and each leafe node represents class label.

50. What are rules in decision tree?

- Ans:-A path from root node to leaf node represents classification rules

51. Explain different types of nodes in nodes in decision tree and how are they selected.

- Ans:-We have Root Node, Internal Node, Leaf Node in a decision tree. Decision Tree starts at the Root Node, this is the first node of the decision tree. Data set is split based on Root Node, again nodes are selected to further split the already splitted data. This process of splitting the data goes on till we get leaf nodes, which are nothing but the classification labels. The process of selecting Root Nodes and Internal Nodes is done using the statistical measure called as Gain

52. What do you mean by impurity in Decision Tree?

- Ans:-We say a data set is pure or homoegenous if all of it's class labels is the same and impure or hetergenous if the class labels are different. Entropy or Gini Index or Classification Error can be used to measure impurity of the data set.

53. What is Pruning in Decision Tree?

- Ans:-The process of removal of sub nodes which contribute less power to the decision tree model is called as Pruning.

54. What is the advantage of Pruning?

- Ans:-Pruning reduces the complexity of the model which in turn reduces overfitting problem of Decision Tree. There are two strategies in Pruning. Propruning - discard unreliable parts from the fully grown tree, Prepruning - stop growing a branch when the information becomes unreliable. Postpruning is the preferred one.

55. What is the difference between Entropy and Information Gain?

- Ans:-Entropy is a probabilistic measure of uncertainity or impurity whereas Information Gain is the reduction of this uncertainity measure.

56. Explain the expression of Gain (of any column)?

- Ans:-Gain for any column is calculated by differencing Information Gain of a dataset with respect to a variable from the Information Gain of the entire dataset i.e., Gain(Age) = Info(D) - Info(D wrt Age)

57. What is the package required to implement Decision Tree in R?

- Ans:-C50 and tree packages can be used to implement a decision tree algorithm in R.

58. What is Random Forest?

- Ans:-Random Forest is an Ensemble Classifer. As opposed to building a single decision tree, random forest builds many decision trees and combines the output of all the decision trees to give a stable output.

59. How does Random Forest adds randomness and build a better model?

- Ans:-Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model. Additional randomness can be added by using random thresholds for each feature rather than searching for the best possible thresholds (like a normal decision tree does).

60. What is the R package to employ Random Forest in R?

- Ans:-randomForest is the package to employ Random Forest in R

61. What are the pros of using Random Forest?

- Ans:- Random Forest won't overfit the model, it is unexcelled in reliable accuracy, works very well on large data sets, can handle thousands of input variables without deletion, outputs significance of input variables, handles outliers and missing values very well

62. What is the limitation of Random Forest?

- Ans:- The main limitation of Random Forest is that a large number of trees can make the algorithm to slow and ineffective for real-time predictions. In most real-world applications the random forest algorithm is fast enough, but there can certainly be situations where run-time performance is important and other approaches would be preferred.

63. What is a Neural Network?

- Ans:- Neural Network is a supervised machine learning algorithm which is inspired by human nervous system and it replicates the similar to how human brain is trained. It consists of Input Layers, Hidden Layers, & Output Layers.

64. What are the various types of Neural Networks?

- Ans:- Artificial Neural Network, Recurrent Neural Networks, Convolutional Neural Networks, Boltzmann Machine Networks, Hopfield Networks are examples of the Neural Networks. There are a few other types as well.

65. What is the use of activation functions in neural network?

- Ans:- Activation function is used to convert a input signal of a node in a A-NN to an output signal. That output signal now is used as a input in the next layer in the stack.

66. What are the different types of activation functions in neural network?

- Ans:- Sigmoid or Logistic, Tanh or Hyperbolic tangent, ReLu or Rectified Linear units are examples of activation functions in neural network

67. What is the package name to implement neural network in R?

- Ans:- neuralnet package can be used to implement neural network in R

68. What is a probability distribution?

- Ans:- A probability distribution is a function that provides the probabilities of occurrence of different possible outcomes in an experiment. In a probability distribution, random variable is plotted on X axis and associated probabilities are plotted on Y axis

69. What are the classifications of probability distributions?

- Ans:- Probability distributions are categorized in two, Discrete and Continuous probability distributions. In discrete probability distribution underlying random variable is discrete whereas in conitnusous probability distribution underlying random variable is continuous

70. What do you mean by discrete random variable?

- Ans:- A discrete random variable is a random variable that has countable values, such as a list of non-negative integers.

71. Examples of discrete random variables.

- Ans:- Number of students present, number of red marbles in a jar, number of heads when flipping three coins, students' grade level are few of the examples of discrete random variables

72. What do you mean by continuous random variable?

- Ans:- A continuous random variable is a random variable with a set of possible values (known as the range) that is infinite and uncountable.

73. Examples of continuous random variables

- Ans:- Height of students in class, weight of students in class, time it takes to get to school, distance traveled between classes are few of the examples of continuous random variables

74. What do you mean by Expected Value?

- Ans:- Expected value (EV), also known as mean value, is the expected outcome of a given experiment, calculated as the weighted average of all possible values of a random variable based on their probabilities. EV(n) = x1P1 + X2P2+X3P3+...+XnPn

75. What do you mean by Data Type?

- Ans:- A data type, in programming, is a classification that specifies which type of value a variable has and what type of mathematical, relational or logical operations can be applied to it without causing an error.

76. What are the classifications of data types in Statistics?

- Ans:- Qualitative and Quantitative are the broader classifications in R, however these are further classified into Nominal, Ordinal, Interval, & Ratio data types.

77. What is the difference between a nominal and an ordinal data type?

- Ans:- A nominal data type merely is a name or a label. Languages spoken by a person, jersey numbers of football players are examples of Nominal data type. Whereas, on top of being a name or a label, Ordinal data type has some natural ordering associated with it. Shirt sizes, Likert scale rating, Ranks of a competition, Educational background of a person are examples of Ordinal data type

78. How is Interval data type different from Ratio?

- Ans:- Interval scales are numeric scales in which we know not only the order, but also the exact differences between the values, but the problem with the problem with interval scales is that they don’t have a “true zero". Temperature and Dates are examples of Interval data type. Whereas Ratio data type tell us about the order, exact value between units, and they also have an absolute zero. Heights & Weights of people, length of a object

79. Why Data Types are important?

- Ans:- Any statistical method, be it descriptive, predictive or prescriptive can be employed only based on the data type of the variable. Incorrect identification of data types leads to incorrect modeling which in turn leads to incorrect solution.

80. What do you mean by an Absolute Zero?

- Ans:- Absolute zero means true absence of a value. We do not have any absolute zero in Interval data type. One such example is 0 Celsius temperature which does not mean that the temperature is absent.

81. Which data object in R is used to store and process categorical data?

- Ans:- The Factor data objects in R are used to store and process categorical data in R.

82. What is a Factor variable?

- Ans:- Factor variable is a variable which can take only limited set of values. In other words, the levels of a factor variable will be limited.

83. What is the difference between "%%" and "%/%"?

- Ans:- "%%" gives remainder of the division of first vector with second while "%/%" gives the quotient of the division of first vector with second.

84. What is lazy function evaluation in R?

- Ans:- The lazy evaluation of a function means, the argument is evaluated only if it is used inside the body of the function. If there is no reference to the argument in the body of the function then it is simply ignored.

85. What is the output for the below expression all(NA==NA)?

- Ans:- NA

86. What is the difference between subset() function and sample() function in R?

- Ans:- The subset() functions is used to select variables and observations. The sample() function is used to choose a random sample of size n from a dataset.

87. Is an array a matrix or a matrix an array?

- Ans:- Every matrix can be called an array but not the reverse. Matrix is always two dimensional but array can be of any dimension.

88. How do you convert the data in a JSON file to a data frame?

- Ans:- Using the function as.data.frame()

89. What is R Base package?

- Ans:- This is the package which is loaded by default when R environment is set. It provides the basic functionalities like input/output, arithmetic calculations etc. in the R environment.

90. What is recycling of elements in a vector? Give an example.

- Ans:- When two vectors of different length are involved in a operation then the elements of the shorter vector are reused to complete the operation. This is called element recycling. Example - v1 <- c(4,1,0,6) and V2 <- c(2,4) then v1*v2 gives (8,4,0,24). The elements 2 and 4 are repeated.

validation

91. What is reshaping of data in R?

- Ans:- In R the data objects can be converted from one form to another. For example we can create a data frame by merging many lists. This involves a series of R commands to bring the data into the new format. This is called data reshaping.

92. What is the output of runif(4)?

- Ans:- It generates 4 random numbers between 0 and 1.

93. What are the different types of Discrete Probability Distributions?

- Ans:- Binomial, Poisson, Negative Binomial, Geometric, Hypergeometric are the examples of Discrete Probability Dsitributions

94. What are the different types of Continuous Probability Distribution?

- Ans:- Normal, Exponential, t, f, Chi-square, Unifrom, Weibull are few of the examples of Continuous Probability Distributions

95. What do you mean by Binomial Distribution?

- Ans:- Binomial Distribution can be simply thought of as the probability of Success of Failure outcome in an experiment that is conducted multiple times. Exmaples: Head and Tail outcomes after tossing a coin, Pass or Fail after appearing for an examination.

96. What is a Poisson distribution?

- Ans:- Poisson Distribution gives the probability of a number of events happening in a fixed interval or space. Number of customers visiting a restaurant every day

97. What do you mean by Negative Binomial Distribution?

- Ans:- It is the distribution of number of successes occuring in a sequence of independently and identically distributed Bernoulli trials before a specied number of failures occurs Example: From a lot of tires containing 5% defectives, if 4 tires have to be chosen at random then what is the probability of finding 2 defective tires before 4 good ones.

98. Explain Normal Distribution?

- Ans:- The normal distribution or a bell curve is a probability function that describes how the values of a variable are distributed by its mean and standard deviation. Distribution of heights, weights, salaries of people are examples of Normal distribution

100. What do you mean by Uniform Distribution?

- Ans:- Uniform Distribution is the simplest of all the statistical distributions. It is sometimes also known as a rectangular distribution, is a distribution that has constant probability. This distribution is defined by two parameters, a and b, a being minimum value and b the maximum value. Examples: Probability of a flight landing between 25 to 30 minutes when it is anounced that the flight will be landing in 30 minutes. Continuous Uniform Distribution (resembles rectangle) and Discrete Uniform Distribution (rectangle in the form of a dots) are the two types of Uniform Distribution. Examples:

101. What is T Distribution?

- Ans:- The T distribution also known as, Student’s T-distribution is a probability distribution that is used to estimate population parameters when the sample size is small and/or when the population variance is unknown.

102. Explain F Distribution?

- Ans:- Probability distribution for the statistical measure 'F-statistic' is called as F Distribution. It will be a skewed dsitribution used for ANOVA testing. Minimum value will be 0 and there is no standard maximum value. Here F statistic is nothing but the value that you get in the output of ANOVA or Regression analysis. F test will tell you if a group of variables are staisitically significant.

103. Explain Weibull Distribution?

- Ans:- The Weibull distribution is particularly useful in reliability work since it is a general distribution which, by adjustment of the distribution parameters, can be made to model a wide range of life distribution characteristics of different classes of engineered items. Weibull distribution is widely used in assess product reliability, analyze life data, and model failure times i.e, it is widely used in Reliability and Survival Analysis Based on the Beta parameter, Weibull distriution can take different distributions. If Beta<1 then Gamma, Beta=1 then Exponential, Beta=2 then Lognormal, Beta=3.5 then Normal.

104. When is it appropriate to employ a Bar plot?

- Ans:- A Bar plot of a numerical variable will be cluttered and makes it difficult for interpretation, whereas it makes sense to employ bar plot on categorical variable as we can interpret it in an efficient way.

105. Why Standard Deviation when we have Variance measure?

- Ans:- Variance is calculated to find how the individual data points are away from the mean, nothing but dispersion in the data. It is calculated as the averge of the square of the difference of mean from each data point. So from this calculation, we know for a fact that units are getting squared. There is way to get rid of squared units without having the necessity of standard deviation is by taking an absolute instead of square in the variance calculation. But the problem with taking absolute is it will lead to misleading results, for example, two variable X1(4,4,-4,-4) & X2(7,1,-6,-2) you get same variance as 4 if absolute is used and different variances as 4 & 4.74 when squared is used. For this reason we resort to squaring the difference of each value from it's mean. At this stage, if we interpret dispersion of data based on Variance, it shall confusion as the values & units are squared. Hence, we resort to Standard Deviation.

106. Why the probability associated with a single value of a continuous random variable is considered to be zero?

- Ans:- A continuous random variable takes an infinite number of possible values. As the number of values assumed by the random variable is infinite, the probability of observing a single value is zero.

107. List out the different Sampling Techniques.

- Ans:- Probability Sampling and Non-Probability Sampling are the broader classifications of Sampling techniques. The difference lies between the above two is whether the sample selection is based on randomization or not. With randomization, every element gets equal chance to be picked up and to be part of sample for study.

108. What do you mean by Sampling Error?

- Ans:- An error occured during the sampling process if referred to as a Sampling Error. It can include both Systematic Sampling Error and Random Sampling Error. Systematic sampling error is the fault of the investigation, but random sampling error is not.

109. Explain Probability Sampling and it's types

- Ans:- This Sampling technique uses randomization to make sure that every element of the population gets an equal chance to be part of the selected sample. It’s alternatively known as random sampling. Simple Random Sampling, Stratified Sampling, Systematic Sampling, Cluster Sampling, Multi stage Sampling are the types of Probability Sampling

110. Explain Simple Random Sampling.

- Ans:- Every element has an equal chance of getting selected to be the part sample. It is used when we don’t have any kind of prior information about the target population. For example: Random selection of 20 students from class of 50 student. Each student has equal chance of getting selected. Here probability of selection is 1/50

111. Explain Stratified Sampling.

- Ans:- This technique divides the elements of the population into small subgroups (strata) based on the similarity in such a way that the elements within the group are homogeneous and heterogeneous among the other subgroups formed. And then the elements are randomly selected from each of these strata. We need to have prior information about the population to create subgroups.

112. Explain Cluster Sampling.

- Ans:- Our entire population is divided into clusters or sections and then the clusters are randomly selected. All the elements of the cluster are used for sampling. Clusters are identified using details such as age, sex, location etc. Single Stage Cluster Sampling or Two Stage Cluster Sampling can be used to perform Cluster Sampling

113. Explain Multi-Stage Sampling

- Ans:- It is the combination of one or more probability sampling techniques. Population is divided into multiple clusters and then these clusters are further divided and grouped into various sub groups (strata) based on similarity. One or more clusters can be randomly selected from each stratum. This process continues until the cluster can’t be divided anymore. For example country can be divided into states, cities, urban and rural and all the areas with similar characteristics can be merged together to form a strata.

114. Explain Non-Probability Sampling and it's types

- Ans:- It does not rely on randomization. This technique is more reliant on the research era€™s ability to select elements for a sample. Outcome of sampling might be biased and makes difficult for all the elements of population to be part of the sample equally. This type of sampling is also known as non-random sampling. Convenience Sampling, Purposive Sampling, Quota Sampling, Referral/Snowball Sampling are the types of Non-Probability Sampling

115. Explain Convenience Sampling

- Ans:- Here the samples are selected based on the availability. This method is used when the availability of sample is rare and also costly. So based on the convenience samples are selected. For example: Researchers prefer this during the initial stages of survey research, as it’s quick and easy to deliver results.

116. Explain Purposive Sampling

- Ans:- This is based on the intention or the purpose of study. Only those elements will be selected from the population which suits the best for the purpose of our study. For Example: If we want to understand the thought process of the people who are interested in pursuing master’s degree then the selection criteria would be “Are you interested for Masters in..?”. All the people who respond with a “No” will be excluded from our sample.

117. Explain Quota Sampling

- Ans:- This type of sampling depends of some pre-set standard. It selects the representative sample from the population. Proportion of characteristics/ trait in sample should be same as population. Elements are selected until exact proportions of certain types of data are obtained or sufficient data in different categories is collected. For example: If our population has 45% females and 55% males then our sample should reflect the same percentage of males and females.

118. Explain Referral /Snowball Sampling

- Ans:- This technique is used in the situations where the population is completely unknown and rare. Therefore we will take the help from the first element which we select for the population and ask him to recommend other elements who will fit the description of the sample needed. So this referral technique goes on, increasing the size of population like a snowball. For example: It’s used in situations of highly sensitive topics like HIV Aids. Not all the victims will respond to the questions asked so researchers can contact people they know or volunteers to get in touch with the victims and collect information

119. Explain Systematic Sampling Error.

- Ans:- When errors are systematic, they bias the sample in one direction. Under these circumstances, the sample does not truly represent the population of interest. Systematic error occur s when the sample is not drawn properly. It can also occur if names are dropped from the sample list because some individuals were difficult to locate or uncooperative. Individuals dropped from the sample could be different from those retained. Those remaining could quite possibly produce a biased sample. Political polls often have special problems that make prediction difficult.

120. Explain Random Sampling Error.

- Ans:- Random sampling error, as contrasted to systematic sampling error, is often referred to as chance error. Purely by chance, samples drawn from the same population will rarely provide identical estimates of the population parameter of interest. These estimates will vary from sample to sample. For example, if you were to flip 100 unbiased coins, you would not be surprised if you obtained 55 heads on one trial, 49 on another, 52 on a third, and so on.

121. Sigma in statistics represents Standard Deviation. Say for a normally distributed data, the Standard Deviation is 5 and the mean is 50, then +_1sigma will have the range of 50-1(5) and 50+1(5) i.e., [45,55], 31% of the given data will be in this range. On similar lines, if calculated, you get the range as [40,60], 69% for +_2sigma and [35,65], 93% for +_3sigma respectively.

- Ans:- Sigma in statistics represents Standard Deviation. Say the Standard Deviation is 5 and the mean is 50, then +_1sigma will have the range of 50-1(5) and 50+1(5) i.e., [45,55] . 31% of the given data will be in this range. On similar lines, if calculated, you get the range as [40,60] i.,e 69% for +_2sigma and [35,65] i.e., 93% for +_3sigma respectively

122. What do you mean by Empirical Rule?

- Ans:- In Statistics, 68–95–99.7 rule is also known as the empirical rule or three sigma rule. For a Gaussian distribution the the mean (arithmetic average), median (central value), and mode (most frequent value) coincide. Here, area under the curve between ± 1s (1 sigma) includes 68% of all values (of the population), while ± 2s (2 sigma) includes 95% and ± 3s (3 sigma) includes 99.7% of all values.

123. In order to come up with a Linear Regression output a minimum of how many observations are required.

- Ans:- a. 1, b. 2, c. 30, d. None . Correct Answer is b which is 2. Output of Linear Regression is in the form of equation of straight line which requires atleast 2 observations.

124. How can you say that Standard Normal Distribution is better than Normal Distribution?

- Ans:- It is inappropriate to say that Sam with 80 score in English Literature is better than Tom with 60 score in Psychology, as the variability of scores within the subjects may vary. In order to compare the scores of two different subjects, we need to standardize the deviations of the subjects and then compare the results.. This can be done using Z transformation, which gives 0 as mean and 1 as Standard Deviation for any normally distributed data. Assuming SD=77, Mean=3 for English Literature and SD=56 , Mean=2 and Psychology, we get 1,2 as z scores or SD away from Mean for English Literature and Psychology. Now you can say that English Literature Tom performed better than Sam.

125. What do you mean by a Quantile?

- Ans:- Often referred to as Percentiles, Quantiles are the point(values) in your data below which certain proportion of data falls. For example Median is also a quantile or 50th percentile below which 50% of the data falls.

126. How is a Normal QQ plot plotted , what is it's use and why is it called as a Normal QQ plot?

- Ans:- A Q-Q Plot or a Quantile-Quantile Plot is plotted by considering raw values of a variable on Y axis and it's standardized values on X axis. It is used to assess the distribution of the underlying data, the distribution could be any of the theoretical distributions like Normal, Exponential, etc. Mostly we will be interested to find if the distribution of underlying data (variable) follows normal distribution or not, Q-Q Plot is called as Normal Q-Q Plot

127. What is the use of the reference line in Normal Q-Q plot?

- Ans:- Reference Line indicates the normal distribution of the data. If most of the data points in a Normal Q-Q Plot are falling across the refence line then we say that the distribution of the underlying data (variable) follows Normal Distribution.

128. What are the R functions to plot QQ Plot and the reference line in a Q-Q Plot

- Ans:- qqnorm() is used to plot a Q-Q Plot whereas qqline() is used to plot the refence line

129. Differentiate between Sample Variance and Sampling Variation.

- Ans:- Sample Variance refers to the variation of observations in a single sample whereas Sampling Variance refers to the variation of a statistical measure (eg., Mean) among multiple samples.

130. How is Standard Deviation different from Standard Error?

- Ans:- Stadard Deviation and Standard Error are both measures of dispersion or spreadness. Standard Deviation uses Population data and Standard Error uses Sample data. Standard Error tells you have far a sample statistic (eg., sample mean) deviates from the actual Population mean. This deviation is referred as the Standard Error. Larger the sample size, less will be the deviation (SE) between the sample mean and the population mean.

131. Explain Central Limit Theorem

- Ans:- Central Limit Theorem explains about the distribution of the sample data. The distribution will be normal, if the population data is normally distributed or if the propulation data is not normal but the sample size is fairly large.

132. What is the necessity of a Confidence Interval?

- Ans:- We cannot trust a point esitmate (for example a sample mean) to infer about Population mean, reason being, if we draw another sample it is more likely that we will get a different sample mean all together. To overcome this problem, we come up with an Interval associated with some Confidence. This can be achieved by including Magin of Error with Point Estimate which gives us the Confidence Interval

133. What is the R function to calculate z value from probability value?

- Ans:- a. pnorm, b. qnorm, c. qt d. None Ans is b

134. What is the R function to calculate t value from probability value?

- Ans:- a. pnorm, b. qnorm, c. qt d. None Ans is c

135. Do we have standard z values for different probability values, explain?

- Ans:- Yes, we have standard z values for different probability values. For example, 1.64 for 90%, 1.96 for 95%, & 2.58 for 99% probability values

136. Do we have standard t values for different probability values, explain?

- Ans:- We will not have standard t values for different probability values, reason being the computation of t value includes degrees of freedom, which is dependent on the sample size. Hence for the same probability with different degrees of freedom we get different t values.

137. Why do we have to include Random Sample while interpreting Confidence Interval?

- Ans:- If we were asked to comment about Population Mean (a single value, which do not change) by using Sample Data (randomly selected from the Population), we do not calculate Population Mean(i.e., Point Estimate) instead we come up with a Confidence Interval. Now, if another Sample Data is randomly drawn and CI is computed then then it is quite obvious that we will get a different CI all together. Hence, you can say that CI is dependent on the drawn sample. Therefore, it is always mandatory to interpret CI by including random sample.

138. What do you mean by Degrees of Freedom?

- Ans:- Degrees of freedom are the number of independent values that a statistical analysis can estimate. You can also think of it as the number of values that are free to vary as you estimate parameters They appear in the output of Hypothesis Tests, Probability Distributions, Regression Analysis. Degrees of freedom is equal your sample size minus the number of parameters you need to calculate during an analysis. It is usually a positive whole number.

139. When do we go for T distribution?

- Ans:- T Distribution or Student's T Distribution is employed when the Population Standard Deviation is unknonw and the sample size is less than 30. If the Sample Size is >= 30 then T Distribution appears similar to Normal Distribution

140. What do you mean by Hypothesis Testing?

- Ans:- It is the way of testing results of an experiment whether they are valid and meaningful and have not occured just by chance. If the results have happened just by chance then the experiment cannot be repeated and is not reusable.

141. What is Null Hypothesis in Hypothesis Testing?

- Ans:- Null Hypothesis is nothing but a statement which is usually true. On top of Null Hypothesis we conduct various Hypothesis Tests to see if Null Hypothesis holds true or not. Null Hypothesis is denoted by Ho.

142. Why is naive Bayes so ‘naive’ ?

- Ans:- naive Bayes is so ‘naive’ because it assumes that all of the features in a data set are equally important and independent. As we know, these assumption are rarely true in real world scenario.

143. What do you mean by Prior Probability?

- Ans:- Prior probability is the proportion of dependent variable in the data set. It is the closest guess you can make about a class, without any further information.

144. How is True Positive Rate and Recall related? Write the equation.

- Ans:- True Positive Rate = Recall. Yes, they are equal having the formula (TP/TP + FN).

145. How do you select important variables in EDA?

- Ans:- 1 Remove the correlated variables prior to selecting important variables 2 Use linear regression and select variables based on p values 3 Use Forward Selection, Backward Selection, Stepwise Selection 4 Use Random Forest, Xgboost and plot variable importance chart 5 Use Lasso Regression 6 Measure information gain for the available set of features and select top n features accordingly.

146. What is the difference between covariance and correlation?

- Ans:- Covariance is a measure to know how to variables change together.Covariances are difficult to compare. For example: if we calculate the covariances of salary ($) and age (years), we'll get different covariances which can't be compared because of having unequal scales. To combat such situation, we calculate correlation to get a value between -1 and 1, irrespective of their respective scale. Correlation is the standardized form of covariance.

147. Is there a way to capture the correlation between continuous and categorical variable?

- Ans:- ANCOVA (analysis of covariance) technique to capture association between continuous and categorical variables.

148. what is the difference between covariance and correlation?

- Ans:- A classification trees makes decision based on Gini Index and Node Entropy. Gini index says, if we select two items from a population at random then they must be of same class and probability for this is 1 if population is pure. Entropy is the measure of impurity in the dataset. Entropy is zero when a node is homogeneous. It is maximum when a both the classes are present in a node at 50% – 50%. Lower entropy is desirable.

149. What is a LogLoss function and where is it used?

- Ans:- In classification techniques, instead of predicting the actual classes, a measure called as LogLoss is used to predict the probabilities for an observation,.

150. What do you mean by Cross Entropy?

- Ans:- Cross Entropy essenitally is similar to log loss function used to measure the probabilities of an actual label. Generally, Log loss term is used in Binary classifications, whereas Cross Entropy is used for multiple classification.

151. Given Decision Tree & Random Forest, which one do you think might create an overfitting problem and which one solves the overfitting problem

- Ans:- Decision Tree has the tendency of overfitting because for the fact, it tries to build as much accurate model as possible by selecting the root node & the internal nodes based on the measure Gain. This Decision Tree will behave very well on the training data but might not generalize it's predictions on the test data. To overcome this problem, we have a reliable ensemble algorithm called as Random Forest which helps in tackling the overfitting problem by creating creating a lot of decision trees (built using a fewer input variables) and just not a single one. Finally, the results will be considered based on majority voting or an average of all the results.

152. Draw parallels between KMeans Clustering & KNN?

- Ans:- Both are purely trial and error methods, we try with different values of K to find the best value. Another similarity is the distance measure involved. Both the algorithms have distance measure calculations.

153. For a coefficient value of -0.65123 for an input variable cost.car what has to be the interpretation of Log(Carpool/Car) in a multinomial regression?

- Ans:- First of all, the sign (+ve,-ve) indicates the impact of the input variable on the output mode. In this case, if there is a unit increase in the input variable i.e., cost.car, the Log(Carpool/Car) decreases by 0.65123

154. For Logiistic Regression, is it a good practice to decide on the goodness of the model based on just accuracy, or is there anything else we can look at.

- Ans:- Output of the Logistic Regression is great, you have multiple measures using which you can comment about the accuracy and reliability of the model. Like, probabilitites of parameters, Null Deviance, Residual Deviance, stepAIC (to compare mutliple models), confusion matrix, overall accuracy, Sensitivity (Recall), Specificity, ROC Curve, Lift Chart are the measures you might want to look at based on the context of the business objective.

155. How does Multinomial Regression predicts the probabilities of class labels, given the fact that you have more than 2 class labels?

- Ans:- In a way, Multinomial Regression builds n-1 individual Logistic Regression models, here n is the number of class labels. Applying exponential on either sides of the the n-1 model outputs and then solving them gives us the individual probabilities for the n class labels. Once we get the probabilities we then classify observations as the class labels.

156. Why is SVM called as a black box technique?

- Ans:- SVM is termed as a black box technique, as internally the algorithm applies complex transformations on the input variables based on the Kernel trick applied. Although, the math of these tranformations is not hidden but slightly complex. Becasue of this complexity, SVM is known as a black box technique.

157. Why are Ensemble techniques more preferred than other classification models?

- Ans:- Firstly the ensemble techniques assure about the reliability of the accuracy. This however can also be achieved for non-ensemble techniques by employing various reliability techniques. One such popular technique is k-fold cross validation. Secondly, it's the way how intelligently the classifications are predicted in ensemble techniques.

158. Which pre-processing steps can be considered before building a recommendation system?

- Ans:- Imputing the missing values, normalization, SVD or PCA or Clustering, similarity measures can be considered as the pre-processing steps before Recommendation Systems.

159. What is the need of having Confidence and Lift Ratio, when you have the Support measure?

- Ans:- Support measure helps us in filtering out all the possible combination of rules which are expnential. Effect of Antecedent or Conseqent being a generalized product cannot be filtered out just by definig Support. Confidence helps in filtering out Antecedents being generalized products and Lift Ratio helps in filtering our Consequents being generalized ones.

160. Are you aware of the algorithm which employs Affinity Analysis ?

- Ans:- Apriori is the algorithm which employs Affinity Analysis.

161. What is User based collaborative filtering?

- Ans:- In User Based Collaborative Filtering, users act as rows and items as columns. Here we try to find the similarity among the users.

162. What is Item based collaborative filtering?

- Ans:- In Item Based Collaborative Filtering, items act as rows and users as columns. Here we try to find the similarity among the items.

163. How is Item based collaborative filtering different from User based collaborative filtering?

- Ans:- When compared to Users, count of Items will be more. And in Item based collaborative filtering, we try to find similarity among the items which inturn makes the process computationally expensive. In addition to this, in User based collaborative filtering, by trying to find the similarity among the users we try to connect to the usre's taste. Whereas Item based collaborative filtering is somewhat similar to Market Basket Analysis where in we generalize the results.

164. Can we normalize the data before employing Recommendation Systems?

- Ans:- It is appropriate to normalize the data when we have the values like ratings(1-5) as opposed to having values as purchased/not purchased or rated/not rated.

165. What is the first thing that you need to look at when you are given a dataset?

- Ans:- The first check should be made on NA values. Check if there are any NA values present in the data or not. If present, then impute the NA values rather than deleting the observations having NAs.

166. What happens when missing values are not treated?

- Ans:- Missing data can reduce the power/fit of a model or can lead to a biased model making incorrect predictions or classifications.

167. What could be the reasons for the presence of NA values in the data?

- Ans:- Data Extraction and Data Collection are considered to be the major reasons for missing values.

168. What are the reasons for the NAs while collecting the data?

- Ans:- Missing completely at random, Missing at random, Missing that depends on unobserved predictors, Missing that depends on the missing value itself are the reasons for NAs while collecting the data.

169. What are the various imputation techniques?

- Ans:- Listwise deletion, Pairwise deletion, Mean/Mode Substitution, Prediction model, KNN Imputation, Hot Deck Imputation, Maximum Likelihood, Multiple Imputation are the various imputation techniques

170. Explain Pairwise Deletion imputation technique?

- Ans:- For each record, correlation between each combination of variables is computed. If the correlation is a junk value for two subsequent correlations, then the common value will be dropped from any computation.

171. How can we employ prediction modeling in imputation?

- Ans:- We divide dataset into two halves. One with no missing values (train data) and the other one with the missing values (test data). Variable with missing values is treated as the target variable. Next, we create a model to predict the target variable based on other attributes of the training data set.

172. What are the drawbacks of the prediction model imputation?

- Ans:- There are two drawbacks for this approach. First is that the model estimated values are usually more well-behaved than the true values. Second is that, if there is no relationships with attributes in the data set and the attribute with missing values, then the model will not be precise for estimating missing values.

173. Explain KNN Imputation

- Ans:- In this method, the missing values of an attribute are imputed using the given number of attributes that are most similar to the attribute whose values are missing. The similarity to the attribute is determined using a discrete function.

174. What are the advantages of using KNN imputation?

- Ans:- KNN can predict both qualitative and quantitaive attributes Creation of predictive model for each attribute with missing data is not required Attributes with multiple missing values can be easily treated Correlation structure of the data is take into consideration

175. What are the disadvantages of using KNN imputation?

- Ans:- KNN imputation is a very time-consuming in analyzing large database. It searches throughout the dataset looking for most similar instances. Choice of K value is critical. Higher values of K would include attributes which are significantly different from what we need whereas lower value of K implies missing out of significant attributes.

176. Explain Hot Deck Imputation

- Ans:- Algorithm traverses from top to bottom in a column, if any NA is found, it makes note of the other values of that record and traverses down till the end and goes up and comes back to the same position of NA. During this traverse it looks for for the exact matches of the record values it noted down and replace the value of NA with the exact matched record. But mostly, we did not find an exact match. If no exact match found, then we need to resort to other techniques like Mean/Mode imputation.

177. Why are ensemble methods superior to individual models?

- Ans :-Ensemble methods average out biases, reduce variance, and are less likely to overfit. General practice in machine learning is: "ensemble and get 2%." This implies that you can build your models as usual and typically expect a small performance boost from ensembling.

178. Explain Bagging?

- Ans :-Bagging, or Bootstrap Aggregating, is an ensemble method in which the dataset is first divided into multiple subsets through resampling. Then, each subset is used to train a model, and the final predictions are made through voting or averaging the component models. Bagging is performed in parallel.

179. What do you understand by selection bias?

- Ans :-It is a statistical error that causes a bias in the sampling portion of an experiment. The error causes one sampling group to be selected more often than other groups included in the experiment. Selection bias may produce an inaccurate conclusion if the selection bias is not identified.

180. Explain how a ROC curve works.

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

181. Define precision and recall?

- Ans :- Recall is also known as the true positive rate: the amount of positives your model claims compared to the actual number of positives there are throughout the data. Precision is also known as the positive predictive value, and it is a measure of the amount of accurate positives your model claims compared to the number of positives it actually claims. It can be easier to think of recall and precision in the context of a case where you've predicted that there were 10 apples and 5 oranges in a case of 10 apples. You'd have perfect recall (there are actually 10 apples, and you predicted there would be 10) but 66.7% precision because out of the 15 events you predicted, only 10 (the apples) are correct.

182. What's the trade-off between bias and variance?

- Ans :-Bias is error due to erroneous or overly simplistic assumptions in the learning algorithm you're using. This can lead to the model underfitting your data, making it hard for it to have high predictive accuracy and for you to generalize your knowledge from the training set to the test set. Variance is error due to too much complexity in the learning algorithm you're using. This leads to the algorithm being highly sensitive to high degrees of variation in your training data, which can lead your model to overfit the data. You'll be carrying too much noise from your training data for your model to be very useful for your test data.
- The bias-variance decomposition essentially decomposes the learning error from any algorithm by adding the bias, the variance and a bit of irreducible error due to noise in the underlying dataset. Essentially, if you make the model more complex and add more variables, you'll lose bias but gain some variance — in order to get the optimally reduced amount of error, you'll have to tradeoff bias and variance. You don't want either high bias or high variance in your model.

183. Explain the difference between L1 and L2 regularization?

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

184. What's the difference between Type I and Type II error in hypothesis?

- Ans :-Type I error is a false positive, while Type II error is a false negative. Briefly stated, Type I error means claiming something has happened when it hasn't, while Type II error means that you claim nothing is happening when in fact something is.

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

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

186. How is a decision tree pruned?

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

187. What are the advantages and disadvantages of decision trees?

- Ans :-Advantages: Decision trees are easy to interpret, nonparametric (which means they are robust to outliers), and there are relatively few parameters to tune. Disadvantages: Decision trees are prone to be overfit. However, this can be addressed by ensemble methods like random forests or boosted trees.

188. What is the difference between Gini Impurity and Entropy in a Decision Tree?

- Ans :-Gini Impurity and Entropy are the metrics used for deciding how to split a Decision Tree. Gini measurement is the probability of a random sample being classified correctly if you randomly pick a label according to the distribution in the branch.
- Entropy is a measurement to calculate the lack of information. You calculate the Information Gain (difference in entropies) by making a split. This measure helps to reduce the uncertainty about the output label.

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

- Ans :-The F1 score is a measure of a model's performance. It is a weighted average of the precision and recall of a model, with results tending to 1 being the best, and those tending to 0 being the worst. You would use it in classification tests where true negatives don't matter much.

190. How would you handle an imbalanced dataset?

- Ans :-An imbalanced dataset is when you have, for example, a classification test and 90% of the data is in one class. That leads to problems: an accuracy of 90% can be skewed if you have no predictive power on the other category of data! Few tactics to handle imbalanced sets 1- Collect more data to even the imbalances in the dataset. 2- Resample the dataset to correct for imbalances.

191. What are ensemble techniques?

- Ans :-Ensemble techniques use a combination of learning algorithms to optimize better predictive performance. They typically reduce overfitting in models and make the model more robust (unlikely to be influenced by small changes in the training data).

192. What is Overfitting? And how do you ensure you're not overfitting with a model?

- Ans :-Over-fitting occurs when a model studies the training data to such an extent that it negatively influences the performance of the model on new data. Few methods to avoid overfitting:
- Keep the model simpler: reduce variance by taking into account fewer variables and parameters, thereby removing some of the noise in the training data.
- Collect more data so that the model can be trained with varied samples. Use cross-validation techniques such as k-folds cross-validation.

- Use regularization techniques such as LASSO that penalize certain model parameters if they're likely to cause overfitting.
- Use ensembling methods, such as Random Forest. It is based on the idea of bagging, which is used to reduce the variation in the predictions by combining the result of multiple Decision trees on different samples of the data set.

193. What evaluation approaches would you work to gauge the effectiveness of a machine learning model?

- Ans :-You would first split the dataset into training and test sets, or perhaps use cross-validation techniques to further segment the dataset into composite sets of training and test sets within the data. You should then implement a choice selection of performance metrics such as the F1 score, the accuracy, confusion matrix alongside error measures in relevance to the problem being solved

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

- Ans :-The Kernel trick involves kernel functions that can enable in higher dimension. The kernel function maps the lower dimensional data to higher dimension and nonlinear data will become linear in higher dimension data and using the kernel trick enables us effectively run algorithms in a high-dimensional space with lower-dimensional data.

195. What are collinearity and multicollinearity?

- Ans :-Collinearity occurs when two predictor variables (e.g., x1 and x2) in a multiple regression have some correlation. Multicollinearity occurs when more than two predictor variables (e.g., x1, x2, and x3) are inter-correlated.

196. What are Eigenvectors and Eigenvalues?

- Ans :-Eigenvectors: Eigenvectors are those vectors whose direction remains unchanged even when a linear transformation is performed on them. Eigenvalues: Eigenvalue is the scalar that is used for the transformation of an Eigenvector.

197. What is A/B Testing?

- Ans :-A/B is Statistical hypothesis testing for randomized experiment with two variables A and B. It is used to compare two models that use different predictor variables in order to check which variable fits best for a given sample of data. Consider a scenario where you've created two models (using different predictor variables) that can be used to recommend products for an e-commerce platform. A/B Testing can be used to compare these two models to check which one best recommends products to a customer.

198. What is Cluster Sampling?

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

199. How are Python libraries NumPy and SciPy related?

- Ans :-NumPy defines arrays along with some basic numerical functions like indexing, sorting, reshaping, etc. SciPy implements computations such as numerical integration, optimization and machine learning using NumPy's functionality.

200. When these library would you prefer for plotting in Python language: Seaborn, Matplotlib, Bokeh?

- Ans :-Matplotlib: Used for basic plotting like bars, pies, lines, scatter plots, etc Seaborn: Is built on top of Matplotlib and Pandas to ease data plotting. It is used for statistical visualizations like creating heatmaps or showing the distribution of your data Bokeh: Used for interactive visualization. In case your data is too complex and you haven't found any "message" in the data, then use Bokeh to create interactive visualizations that will allow your viewers to explore the data themselves

201. How is kNN different from kmeans clustering?

- Ans :-kmeans algorithm partitions a data set into clusters such that a cluster formed is homogeneous and the points in each cluster are close to each other. The algorithm tries to maintain enough separability between these clusters. Due to unsupervised nature, the clusters have no labels. kNN algorithm is a classification algorithm which classify an unlabeled observation based on its k (can be any number) surrounding neighbors. It is also known as lazy learner because it involves minimal training of model. Hence, it doesn't use training data to make generalization on unseen data set.

202. Explain Principle Component Analysis (PCA).

- Ans :-PCA is a method for transforming features in a dataset by combining them into uncorrelated linear combinations. These new features, or principal components, sequentially maximize the variance represented (i.e. the first principal component has the most variance, the second principal component has the second most, and so on). As a result, PCA is useful for dimensionality reduction because you can set an arbitrary variance cutoff.

203. What are the advantages and disadvantages of neural networks?

- Ans :-Advantages: Neural networks (specifically deep NNs) have led to performance breakthroughs for unstructured datasets such as images, audio, and video. Their incredible flexibility allows them to learn patterns that no other ML algorithm can learn. Disadvantages: However, they require a large amount of training data to converge. It's also difficult to pick the right architecture, and the internal "hidden" layers are incomprehensible.

204. What is the Box-Cox transformation used for?

- Ans :-The Box-Cox transformation is a generalized "power transformation" that transforms data to make the distribution more normal. For example, when its lambda parameter is 0, it's equivalent to the log-transformation. It's used to stabilize the variance (eliminate heteroskedasticity) and normalize the distribution.

205. What is the difference between stochastic gradient descent (SGD) and gradient descent (GD)?

- Ans :-Both algorithms are methods for finding a set of parameters that minimize a loss function by evaluating parameters against data and then making adjustments. In standard gradient descent, you'll evaluate all training samples for each set of parameters. This is akin to taking big, slow steps toward the solution. In stochastic gradient descent, you'll evaluate only 1 training sample for the set of parameters before updating them. This is akin to taking small, quick steps toward the solution.

206. When would you use GD over SDG, and vice-versa?

- Ans :-GD theoretically minimizes the error function better than SGD. However, SGD converges much faster once the dataset becomes large. That means GD is preferable for small datasets while SGD is preferable for larger ones. In practice, however, SGD is used for most applications because it minimizes the error function well enough while being much faster and more memory efficient for large datasets.