Important Questions for Data Science

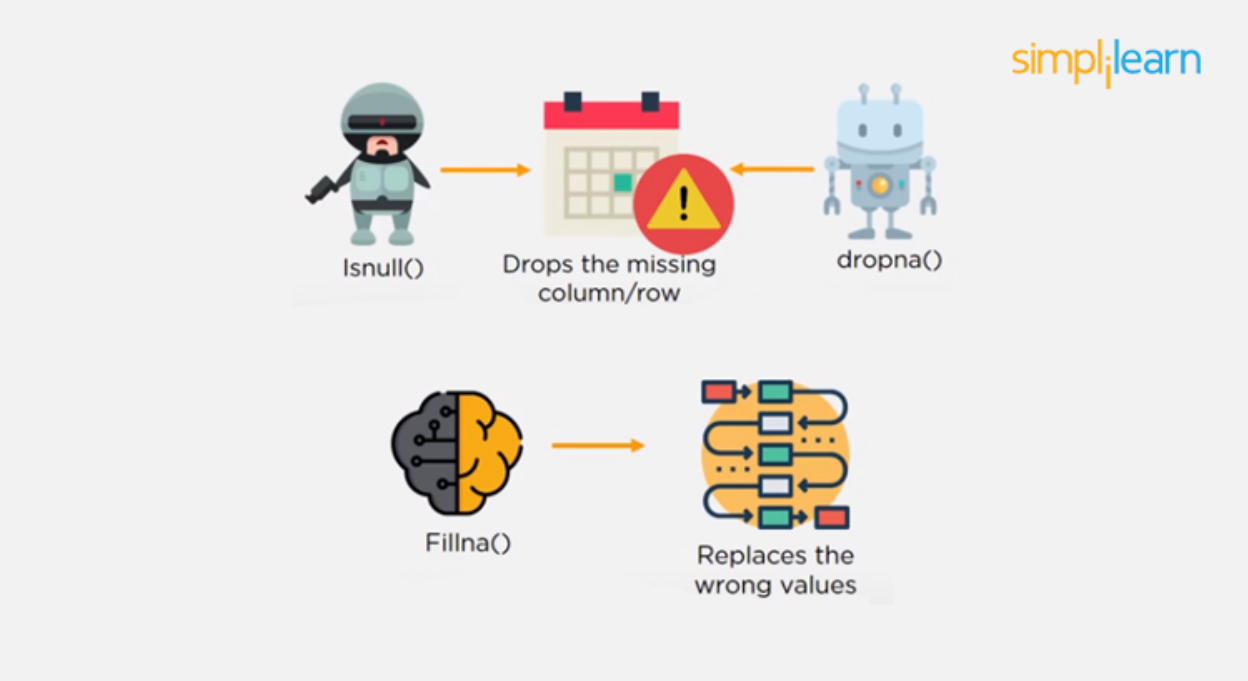
This document includes questions gathered from various study materials, Textbooks and Websites while preparing for the purpose of placements. – **Shobhit Goel**

Question 1: How Do You Handle Missing or Corrupted Data in a Dataset?

Answer 1: One of the easiest ways to handle missing or corrupted data is to drop those rows and columns or replace them entirely with some other value.

There are two useful methods in Pandas:

- IsNull() and dropna() will help to find the columns/rows with missing data and drop them.
- Fillna() will replace the wrong values with a placeholder value.



Question 2: How Can You Choose a Classifier Based on Training Set Data Size?

Answer 2: When the training data set is small, a model with a right bias and low variance seems to work better because they are less likely to overfit.

For example, Naïve Bayes works best when the training set is large. Models with low bias and high variance tend to perform better as they work fine with complex relationships.

Question 3: What Are the Three Stages of Building a Model in Machine Learning?

Answer 3: The three stages of building a machine learning model are:

- *Model Building*

Choose a suitable algorithm for the model and train it according to the requirement

- *Model Testing*

  Check the accuracy of the model through the test data

- *Applying the Model*

  Make the required changes after testing and use the final model for real-time projects.

Question 4: What is Deep Learning?

Answer 4: Deep Learning is a subset of Machine Learning that involves system that think and learn like humans using artificial neural networks. The term “deep” comes from the fact that you can have several layers of neural networks.

One of the Primary differences between Machine Learning and Deep Learning is that feature engineering is done manually in Machine Learning. In the Case of Deep Learning, the model consisting of neural networks will automatically determine which features to use and which not to use.

Question 5: What Are the Differences Between Machine Learning and Deep Learning?

Answer 5:

Machine Learning	Deep Learning
• Enables machines to take decisions on their own, based on past data • It needs only a small amount of data for training • Works well on the low-end system, so you don't need large machines • Most features need to be identified in advance and manually coded • The problem is divided into two parts and solved individually and then combined	• Enables machines to take decisions with the help of artificial neural networks • It needs a large amount of training data • Needs high-end machines because it requires a lot of computing power • The machine learns the features from the data it is provided • The problem is solved in an end-to-end manner

Question 6: What Are the Applications of Supervised Machine Learning in Modern Businesses?

Answer 6: Applications of supervised machine learning include:

- Email Spam Detection

  Here we train the model on the historical data that consists of emails categorized as spam or not spam. This Labeled information is fed a input to the model.

- Sentiment Analysis

  This refers to the process of using algorithms to mine documents and determine whether they're positive, neutral or negative in sentiment.

- Fraud Detection

  Training the model to identify suspicious patterns, we can detect instances of possible fraud.

  Question 7: What is the Pruning in Decision Trees?

  Answer 7: Pruning is a technique in Machine Learning that reduces the length of the Decision Trees. It reduces the complexity of the final Classifier and hence, improves the predictive Accuracy by the Reduction of overfitting.

  Pruning can occur in:

  1. Top-Down Fashion: It will traverse nodes and trim sub trees starting at the root.
  2. Bottom- Up Fashion: It will begin at the leaf nodes.

  Question 8: Briefly Explain the Logistic Regression?

  Answer 8: Logistic Regression is a classification algorithm used to predict a binary outcome for a given set of independent variables.

  The Output of the Logistic Regression is either 0 or 1 with a threshold value of generally 0.5. Any value above 0.5 and vice versa.

  Question 9: What is Kernel SVM?

  Answer 9: Kernel SVM is the abbreviated version of the Kernel Support Vector Machine. Kernel

methods are a class of algorithms for pattern analysis and the most common one is the kernel SVM.

Question 10: When will you use Classification over Regression?

Answer 10: Classification is used when your target variable is categorical, while regression is used when your target variable is continuous. Both Classification and regression belonged to the category of Supervised Machine Learning Algorithms.

Examples of Classification Problems Include:

- Predicting yes or No.
- Estimating Gender.
- Breed of an Animal.
- Type of Color.

Examples of Regression Problems Include: 1. Estimating Sales and Price of a Product.
2. Predicting the score of team. 3. Predicting the amount of rainfall.

Question 11: How will you know which Machine Learning Algorithm to choose for your Classification Problem?

Answer 11: While there is no fixed rule to choose algorithm for a classification Problem but you can follow the below guidelines:

- If the accuracy is a concern, test different algorithms and cross validate them.
- If the training dataset is small, use models that have high bias and low variance to avoid the over-fitting problem.
- If the training dataset is large, use models that have low bias and high variance to avoid the under-fitting problem.

Question 12: Compare K-means Clustering and K-nearest(KNN) Algorithm?

Answer 12:

K-Means Clustering	KNN Algorithm
• K-means is unsupervised.	• KNN is supervised in nature.
• K-means is a clustering Algorithm.	• KNN is a classification algorithm.
• The Points in each cluster are similar to each other, and each	• It classifies an unlabeled observation based on its K-

cluster is different from its neighboring clusters.	nearest neighbors.

Question 13: What are the different types of Machine Learning?

Answer 13: There are three types of Machine Learning:

1. Supervised Learning: In Supervised Machine Learning, a model makes predictions and decisions based on the past or labeled data. Labeled data refers to a set of Data that are given tags or labels and thus made more meaningful.
2. Unsupervised Learning: In Unsupervised Learning, we don't have the labeled data. A model can identify patterns, anomalies and relationships in the input data.
3. Reinforcement Learning: Using Reinforcement learning, the model can learn based on the rewards it received for its previous actions.

Question 14: What is the Semi-Supervised Machine Learning?

Answer 14; Supervised Learning uses data that is completely labeled and Unsupervised Learning Algorithm uses data that is completely unlabeled.
In the Case of Semi-Supervised Learning, the training data contains a small amount of labeled data and large amount of unlabeled data.

Question 15: What is the difference between Inductive Machine Learning and Deductive Machine Learning?

Answer 15:

Inductive Learning	Deductive Learning
• It observes instances based on defined principles to draw a conclusion • Example: Explaining to a child to keep away from the fire by showing a video where fire causes damage	• It concludes experiences • Example: Allow the child to play with fire. If he or she gets burned, they will learn that it is dangerous and will refrain from making the same mistake again.

Question 16: What is the naïve in the Naïve Bayes Classifier?

Answer 16: The Classifier is called naïve because it makes assumptions that may or may not turn out to be correct.

The Algorithm assumes that the presence of one feature of class is not related to the presence of any other feature, given the class variable.

For instance, a fruit may be considered to be a cherry if it is red in color and round in shape, regardless of other features. This assumption may or may not be right (as an apple also matches the description).

Question 17: Explain how a system can play a game of Chess Using Reinforcement learning?

Answer 17: Reinforcement Learning has an environment and an agent. The agent performs some actions to achieve a Specific goal. Every time the agent performs a task that is taking towards the goal, it is rewarded. And every time, it takes a step which goes against the goal or in a reverse direction, it is penalized.

Earlier Chess Programs have to determine the best moves after much research on numerous factors.

With Enforced learning, we don't have to deal with this problem as the learning agent learns by playing the game. It will make a move (decision), check if it's the right move (feedback), and keep the outcomes in memory for the next step it takes (learning). There is a reward for every correct decision the system takes and punishment for the wrong one.

Question 18: How do you design an Email Spam Filter?

Answer 18: Building a Spam filter involves the following process:

- The email spam filter will be fed with thousands of emails
- Each of these emails already has a label: 'spam' or 'not spam.'
- The supervised machine learning algorithm will then determine which type of emails are being marked as spam based on spam words like the lottery, free offer, no money, full refund, etc.
- The next time an email is about to hit your inbox, the spam filter will use statistical analysis and algorithms like Decision Trees and SVM to determine how likely the email is spam
- If the likelihood is high, it will label it as spam, and the email won't hit your inbox

- Based on the accuracy of each model, we will use the algorithm with the highest accuracy after testing all the models.

Question 19: What is the Random Forest?

Answer 19: A Random Forest is a Supervised Machine Learning Algorithm that is generally used for the classification Problems. It operates by constructing multiple decision trees during the training phase. The Random forest chooses the decision of the majority of the trees as the final outcome.

Question 20: What is the Bias and Variance in the Machine Learning Model?

Answer 20: Bias: Bias in a machine learning model occurs when the predicted values are further from the actual values. Low bias indicates a model where the prediction values are very close to the actual ones.

Under-fitting: High bias can cause an algorithm to miss the relevant relations between features and target outputs.

Variance: Variance refers to the amount of target variable change when trained with different data. For a good model, the variance should be minimized.

Over-fitting: High variance can cause an algorithm to model the random noise in the training data rather than the intended outputs.

Important Tip:

High bias and low variance algorithms train models that are consistent, but inaccurate on average.

High variance and low bias algorithms train models that are accurate but inconsistent.

Important Questions as MCQ

1. Which of the following is a widely used and effective machine learning algorithm based on the idea of bagging? - Random Forest.

2. To find the minimum or the maximum of a function, we set the gradient to zero because: The value of the gradient at extreme of a function is always zero.

3. The Most widely used metrics and tools to assess the classification model are: Confusion Matrix, Cost-Sensitive Accuracy and Area under the ROC curve.

4. What is the disadvantage of the Decision Trees? - Decision trees are prone to over-fit.

5. How do you handle the missing or corrupted data in a dataset? – Drop missing rows and columns, Replace missing values with mean/median/mode and Assign a unique category to missing values.

6. What is the purpose of performing the Cross-Validation? – To Assess the predictive performance of the model, to judge how the trained model perform outside the sample on test data.

7. When Performing regression or classification, which of the following is the correct way to preprocess the data? – Normalize the data- PCA- Training.

8. Which of the following is an example of Feature Extraction? – Constructing Bag of words vector from an email, Applying PCA Projects to a large high dimensional data and removal of Stop Words.

9. What is pca components in sklearn? – set of all eigen vectors for the projection space.

10. How can you prevent a clustering algorithm from getting stuck in bad local optima? – Use Multiple Random Initializations.

11. Which of the following techniques can be used for normalization for the text mining? – Stemming and Lemmatization.

12. Data with Outliers, Data Points with different densities and Data Points with non-convex shapes gives bad results while performing the K-means Clustering.

13. Which of the following is the reasonable way to select the number of principal components "k"? – Choose K to be the smallest value so that at least 99% of the variance is retained.

14. Explain Ensemble Learning? - In Ensemble learning, many base models like classifiers and regressors are generated and combined together so that they give better results. It is used when we build component classifiers that are accurate and independent. There are sequential as well as parallel ensemble methods.

15. What are the parametric models? - Parametric models are those with a finite number of Parameters. To predict the new data, you only know the parameters of the model. Examples include linear regression, logistic regression and linear SVMs.
Non – parametric models are those with an unbounded number of parameters, allowing for more flexibility. To predict the new data, you need to know the parameters of the model and the state of the data that has been observed. Examples include decision trees, k-nearest neighbors.

16. What are PCA, KPCA and ICA used for? – PCA (Principal component Analysis), KPCA (Kernel- based Principal component Analysis) and ICA (Independent component Analysis) are important feature extraction techniques for the dimensionality reduction.

17. What are the Support Vector Machines? – SVMs are supervised learning Algorithms used for classification and regression Analysis.

18. When is ridge regression is favorable over Lasso regression? – You can quote ISLR's authors Hastie, in the presence of the few variables with medium and large sized effect, use lasso regression. In presence of many variables with small/medium sized effects use ridge regression. We can say that lasso regression (L1) does both variable selection and parameter shrinkage, whereas ridge regression only does parameter shrinkage and end up including all coefficients in the model. In the presence of correlated variables, ridge regression might be the preferred choice. Also, ridge regression works best in situations where the least square estimates have higher variance.

19. What's the F1 score? How would you use it? – The F1 score is the measure of a model's performance. It is a weighted average of the precision and recall of a model, with results to 1 being the best, and those tending to 0 being the worst.

20. Explain bagging? – Bagging and Bootstrap Aggregating, is an ensemble method in which the dataset is divided into multiple subsets through re-sampling.
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.

21. Mention some of the EDA Techniques? – EDA (Exploratory Data Analysis) helps analysts to understand the data better and forms the foundation of better models.

Visualizations:

1. Uni-variate visualizations.
2. Bi-variate visualizations.
3. Multivariate visualizations.

- Missing Value Treatment – Replace missing values with either mean/median.
- Outlier Detection – Use Box-plot to identify the distribution of Outliers, then apply IQR to set the boundary for IQR.

22. Do you think 50 small decision trees are better than a large one? Why? – Another way of asking this question is "Is a random forest a better model than a decision tree?" And the answer is yes because a random forest is an ensemble method that takes many decision trees to make a strong learner. Random forests are more accurate, more robust, and less prone to over-fitting.

23. State the differences between causality and correlation? – Causality applies to situations where one action, say X, causes an outcome, say Y, whereas Correlation is just relating one action(X) to another action(Y) but X does not necessarily cause Y.

24. How can you handle outliers of the data? – Outlier is an observation in the dataset which is far away from the other observations in the dataset. We can discover outliers using tools and

functions like box-plot, scatter-plot, Z- score, IQR score etc and handle them based on the visualizations we have got. To handle outliers, we can cap at some threshold, use transformation to reduce the Skew-ness of the data and remove outliers if they are anomalies of data.

25. What is the convex function? – This question is very often asked in Machine Learning Interviews. A convex function is a continuous function, and the value of the mid-point at every interval in its given domain is less than the numerical mean of the values at the two ends of the interval.

26. How can learning curves help create a better model? – Learning curves give the indication of the presence of over-fitting and under-fitting.
In a learning curve, the training error and the cross-validating error are plotted against the number of training data points.

27. Name and define techniques used to find similarities in the recommendation system? – Pearson correlation and Cosine correlation.

28. Do you suggest that treating a categorical variable as a categorical variable would result in a better predictive model? – For better prediction, the categorical variable is considered as a continuous variable only when the variable is ordinal in nature.

29. You've just finished training a decision tree for spam classification, and it is getting abnormally bad performance on both your training and test sets. You know that your implementation has no bugs, so what could be causing the problem? – Your trees are too shallow.
Shallow decision trees- trees that are too shallow might lead to overly simple models that can't fit the data.
A model that is under-fit will have the high training and testing error. Hence, bad performance on training and test sets indicates under-fitting which means the set of hypotheses are not complex enough to include the true but unknown predictions.
The shallower the tree the less variance we have in our predictions.

30. Expectation Maximization is a clustering algorithm, CART is a decision tree algorithm, Gaussian Naïve Bayes is a Bayesian algorithm, Apriori is a association rule learning algorithm.

31. You trained a binary classifier model which gives very high accuracy on the training data, but very much lower accuracy on validation data. Which of the following may be true? – Over-fitting, the training data was not well regularized, the training and testing examples are sampled from different distributions.

32. Averaging the output of the multiple decision trees help? – Decrease the Variance

33. How to combat Over-fitting and under-fitting? – To Combat Over-fitting: Add Noise,

Feature Selection, Increase training Set, Use Cross-Validation techniques, such as k-folds cross-validation, Boosting and bagging, Dropout technique, Perform early stopping.
To combat under-fitting: Add features and Increase time of training.

34. What is regularization? Why is it useful? –- Regularization is the process of adding tuning parameter (penalty term) to a model to induce smoothness to prevent over fitting. This is most often done by adding a constant multiple to an existing weight vector. This constant is often the L1 (Lasso – alpha) or L2 (Ridge – alpha squared). The model predictions should then minimize the loss function calculated on the regularized training set.

35. What are confounding variables? – In Statistics, a confounder is a variable that influences both the independent and dependent variable.

If you are researching whether a lack of exercise leads to weight gain:
Lack of exercise = independent variable
Weight gain = dependent variable
A confounding variable here would be any other variable that affects both of these variables, such as the age of the subject.

36. What are the types of Biases that can occur during sampling? – Selection bias, under coverage bias, Survivorship bias.

37. What is the Survivorship bias? – It is the logical error of focusing aspects that support surviving some process and casually overlooking that did not work because of their lack of prominence. This can lead to the wrong conclusions in numerous different means. For example, during a recession you look just at the survived businesses, noting that they are performing poorly. However, they perform better than the rest, which is failed, thus being removed from the time series.

38. What is the Selection bias and under coverage bias? - Selection bias occurs when the sample obtained is not representative of the population intended to be analyzed. For instance, you select only Asians to perform a study on the world population height.

Under coverage bias occurs when some members of the population are inadequately represented in the sample. A classic example of under coverage is the Literary Digest voter survey, which predicted that Alfred Landon would beat Franklin Roosevelt in the 1936 presidential election. The survey sample suffered from under coverage of low-income voters, who tended to be Democrats. How did this happen? The survey relied on a convenience sample, drawn from telephone directories and car registration lists. In 1936, people who owned cars and telephones tended to be more affluent. Under coverage is often a problem with convenience samples.

39. Explain how a ROC Curve works? – The ROC Curve is a graphical representation of the contrast between True positive rates and False Positive rates at various thresholds. It is often

used as a proxy for the trade-off between the sensitivity (true positive rate) and false positive rate.

40. What is TF-IDF Vectorization? - TF-IDF is short for term frequency-inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining.

The TF-IDF value increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general.

41. Python or R- which one would you prefer for text analytics? – We will prefer Python because of the following reasons:
1. Python would be the best option because it has the Pandas library that provides easy to use data structures and high-performance data analysis tools.
2. R would be more suitable for the machine learning.
3. Python performs faster for all types of text analytics.

41. How does a data cleaning play a vital role in the analysis? – Data cleaning can help in analysis because:
Cleaning data helps transform it into a format that data analysts or data scientists can work with.
• Data Cleaning helps to increase the accuracy of the model in machine learning.
• It is a cumbersome process because as the number of data sources increases, the time taken to clean the data increases exponentially due to the number of sources and the volume of data generated by these sources.
• It might take up to 80% of the time for just cleaning data making it a critical part of the analysis task.

42. What is Cluster Sampling? –Cluster Sampling is a technique used when it becomes difficult to study the target population spread across a wide area and simple random sampling cannot be applied. Cluster Sample is a probability sample where each sampling unit is a collection or cluster of elements.
For example, a researcher wants to survey the academic performance of high school students in Japan. He can divide the entire population of Japan into different clusters (cities). Then the researcher selects a number of clusters depending on his research through simple or systematic random sampling.

43. Can you cite some examples where a false positive is important than a false negative? – Let us first understand what false positives and false negatives are:
1. False positive are the cases where you wrongly classified a non-event as an event ,that is, Type 1 error.
2. False negative rate are the cases when you wrongly classified a event as a non-event, that is,

Type 2 error.

Example 1: In the medical field, assume you have to give chemotherapy to patients. Assume a patient comes to that hospital and he is tested positive for cancer, based on the lab prediction but he actually doesn't have cancer. This is a case of false positive. Here it is of utmost danger to start chemotherapy on this patient when he actually does not have cancer. In the absence of cancerous cell, chemotherapy will do certain damage to his normal healthy cells and might lead to severe diseases, even cancer.

Example 2: Let's say an e-commerce company decided to give $1000 Gift voucher to the customers whom they assume to purchase at least $10,000 worth of items. They send free voucher mail directly to 100 customers without any minimum purchase condition because they assume to make at least 20% profit on sold items above $10,000. Now the issue is if we send the $1000 gift vouchers to customers who have not actually purchased anything but are marked as having made $10,000 worth of purchase.

44. Can you cite some examples where a false negative is important than a false postive? -
Example 1 FN: What if Jury or judge decides to make a criminal go free?
Example 2 FN: Fraud detection.

45. Can you cite some examples where a false negative is equally important to false postive? - In the Banking industry giving loans is the primary source of making money but at the same time if your repayment rate is not good you will not make any profit, rather you will risk huge losses. Banks don't want to lose good customers and at the same point in time, they don't want to acquire bad customers. In this scenario, both the false positives and false negatives become very important to measure.

46. Can you explain the difference between a validation set and a test set? -

A Training Set: • to fit the parameters i.e. weights
A Validation set: • part of the training set • for parameter selection • to avoid over-fitting.
A Test set: • for testing or evaluating the performance of a trained machine learning model, i.e. evaluating the predictive power and generalization.

47. What are the drawbacks of the linear model? - Some drawbacks of the linear model are:
• The assumption of linearity of the model.
• It can't be used for count outcomes or binary outcomes.
• There is over-fitting or under-fitting problems that it can't solve.

48. What is pruning in decision trees? - Pruning is a technique in machine learning and search algorithms that reduces the size of decision trees by removing sections of the tree that provide little power to classify instances. So, when we remove sub-nodes of a decision node, this process is called pruning or opposite process of splitting.

49. How the outliers can be treated? - Outlier values can be identified by using Univariate or any other graphical analysis method. If the number of outlier values is few then they can be assessed individually but for a large number of outliers, the values can be substituted with either the 99th or the 1st percentile values.

All extreme values are not outlier values. The most common ways to treat outlier values:
1. Change it with a mean or median
2. Standardize the feature, changing the distribution but smoothing the outliers
3. Log transform the feature (with many outliers)
4. Drop the value
5. First/third quartile value if more than 2 S.D.

50. What are the most popular cloud services used in Data Science? – Amazon leads 100 Billion dollars cloud market 33% (Microsoft azure on second at 18%).

51. What is Data Science? List the differences between Supervised and Unsupervised learning algorithms? – Data Science is a blend of various tools, algorithms, and machine learning principles with the goal to discover the hidden patterns from the raw data. How is this difference from what statisticians have been doing for years? The answer lies in the difference between explaining and prediction: Statisticians work a posteriori, explaining the results and designing a plan; a data scientist use historical data to make predictions.

Supervised	Unsupervised
Input data is labeled.	Input data is unlabeled.
Split in training/validation/test.	No split.
Used for prediction	Used for analysis.

52. What is the Bias-variance trade-off? – Bias: Bias is an error introduced in the model due to the oversimplification of the algorithm used (does not fit the data properly). It can lead to under-fitting.
Low-Bias machine learning algorithms- Decision trees, KNN and SVM.
High-Bias machine learning algorithms- Linear Regression, Logistic Regression.

Variance is an error introduced in the model due to a complex Algorithm, it performs very well in the training set but poorly in the test set. It can lead to high sensitivity and over-fitting.
Possible high variance - polynomial regression.

Normally, as you increase the complexity of your model, you will see a reduction in error due to lower bias in the model. However, this only happens until a particular point. As you continue to

make your model more complex, you end up over-fitting your model and hence your model will start suffering from high variance.

Bias-Variance trade-off: The goal of any supervised machine learning algorithm is to have low bias and low variance to achieve good prediction performance.

1. The k-nearest neighbor algorithm has low bias and high variance, but the trade-off can be changed by increasing the value of k which increases the number of neighbors that contribute to the prediction and in turn increases the bias of the model.
3. The decision tree has low bias and high Variance, you can decrease the depth of the tree or use fewer attributes.
4. The linear regression has low variance and high bias, you can increase the number of features or use another regression that better fits the data.

There is no escaping the relationship between bias and variance in machine learning. Increasing the bias will decrease the variance. Increasing the variance will decrease the bias.

53. What is the confusion matrix? – The confusion matrix is a 2*2 table that contains 4 output provided by the binary classifier.

54. What do you understand by the term Normal-distribution? - Data is usually distributed in different ways with a bias to the left or to the right or it can all be jumbled up. However, there are chances that data is distributed around a central value without any bias to the left or right and reaches normal distribution in the form of a bell-shaped curve.

The random variables are distributed in the form of a symmetrical, bell-shaped curve. Properties of Normal Distribution are as follows:

1. Unimodal (Only one mode)
2. Symmetrical (left and right halves are mirror images)
3. Bell-shaped (maximum height (mode) at the mean)
4. Mean, Mode, and Median are all located in the center
5. Asymptotic.

55. What is the goal of A/B Testing? – It is the hypothesis testing for a randomized experiment with two variables A and B.

The goal of A/B testing is to identify any changes to the web page to maximize or increase the outcome of interest. A/B testing is a fantastic method for figuring out the best online promotional and marketing strategies for your business.

56. What is P-value? - When you perform a hypothesis test in statistics, a p-value can help you determine the strength of your results. p-value is the minimum significance level at which you can reject the null hypothesis. The lower the p-value, the more likely you reject the null hypothesis.

57. What are the differences between Under-fitting and over-fitting? – In over-fitting, a statistical model describes random error or noise instead of the underlying relationship. Over-fitting occurs when the model is too complex, such as having too many parameters relative to the number of observations. A model that has been over-fitted, has poor-predictive performance, as it overreacts to minor fluctuations in the training data.

Under-fitting occurs when a statistical model or machine learning algorithm cannot capture the underlying trend of the data. Under-fitting would occur, for example, when fitting a linear model to a non-linear model. A model too have poor-predictive performance.

58. What is the law of large numbers? – It is a theorem that describes the result of performing the same experiment a large number of times. It says that the sample means, the sample variance and sample standard deviation converge to what are they trying to estimate. According to the law, the average of the results obtained from the large number of trials should be close to the expected value and tend to move closer as more number of trials are performed.

59. How is logistic regression done? – Logistic regression measures the relationship the dependent variable (our label of what we want to predict) and one or more independent variables (or features) by estimating probability using its underlying logistic function (sigmoid).

The logistic regression function models the probability of x. p(x) as a function that outputs either values between 0 and 1 for all values of x. This is given by:

$$p(X) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+}}$$

60. What are odds ratio in logistic regression? – Odds is another way of representing probabilities which can be derived from re-arranging the logistic regression. Unlike probabilities, odds takes values from 0 to infinity. The value of odds close to 0 indicates low probability while value of odds close to infinity represents high probability.

Odds are given by: $\frac{p(X)}{1 - p(X)} = e^{\beta_0+\beta}$

61. How is Bayes theorem related to the logistic regression? – Logistic regression is a probabilistic model that is based on the Bayes Theorem to determine the class with the largest probability for given observation.