Ml viva questions

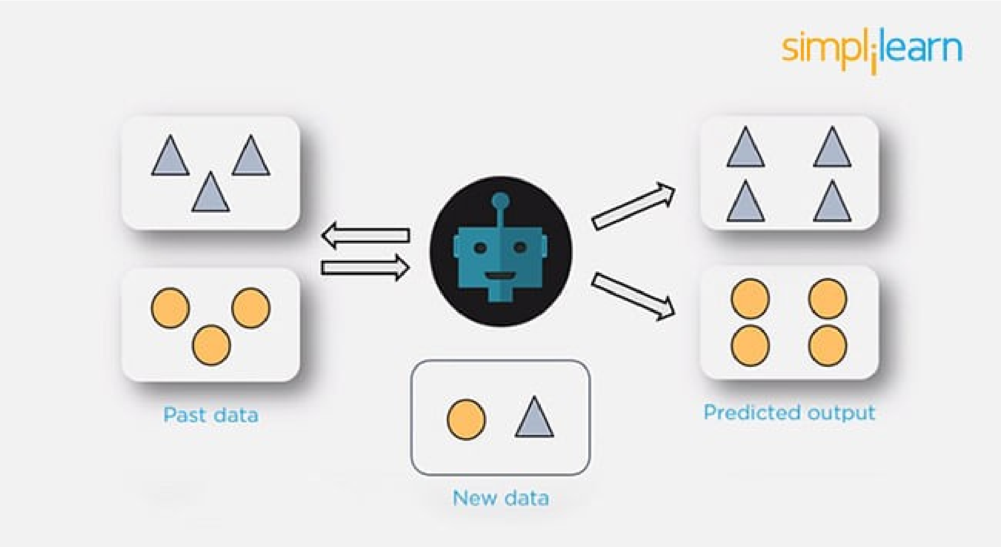
Machine Learning (University of Mumbai)

1. What Are the Different Types of Machine Learning?

There are three types of machine learning:

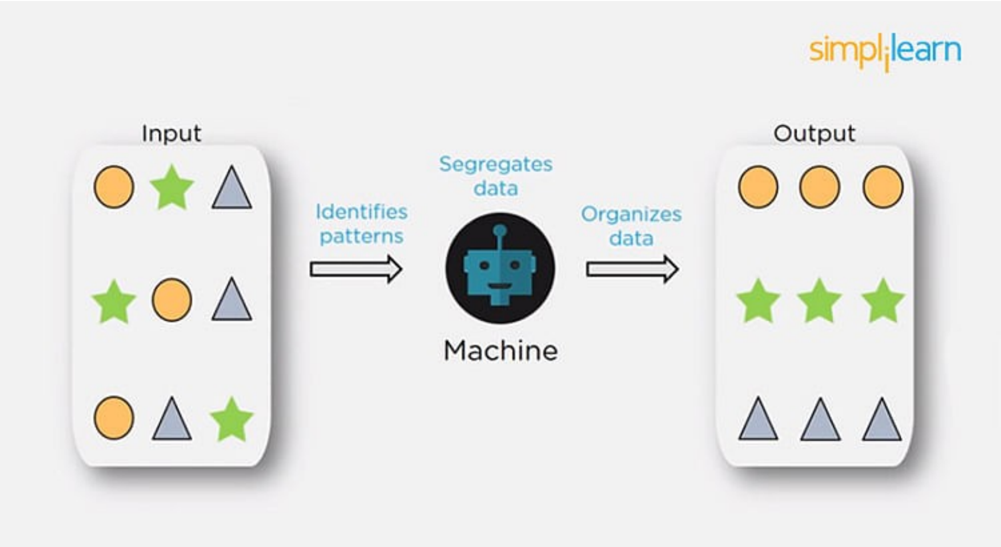
Supervised Learning

In supervised machine learning, a model makes predictions or decisions based on past or labeled data. Labeled data refers to sets of data that are given tags or labels, and thus made more meaningful.



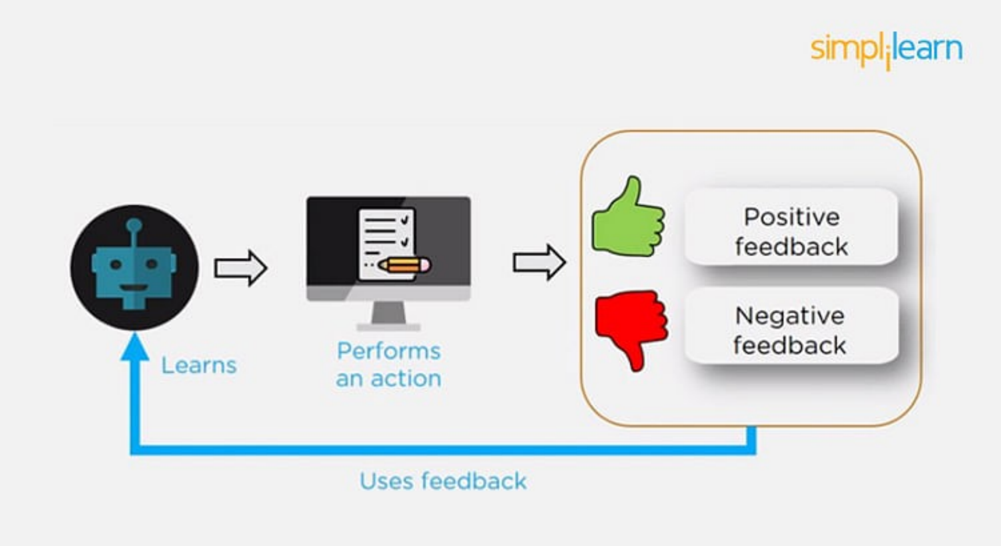
Unsupervised Learning

In unsupervised learning, we don't have labeled data. A model can identify patterns, anomalies, and relationships in the input data.



Reinforcement Learning

Using reinforcement learning, the model can learn based on the rewards it received for its previous action.



Consider an environment where an agent is working. The agent is given a target to achieve. Every time the agent takes some action toward the target, it is given positive feedback. And, if the action taken is going away from the goal, the agent is given negative feedback.

Also Read: Supervised and Unsupervised Learning in Machine Learning

2. What is Overfitting, and How Can You Avoid It?

The Overfitting is a situation that occurs when a model learns the training set too well, taking up random fluctuations in the training data as concepts. These impact the model's ability to generalize and don't apply to new data.

When a model is given the training data, it shows 100 percent accuracy—technically a slight loss. But, when we use the test data, there may be an error and low efficiency. This condition is known as overfitting.

There are multiple ways of avoiding overfitting, such as:

- Regularization. It involves a cost term for the features involved with the objective function
- Making a simple model. With lesser variables and parameters, the variance can be reduced
- Cross-validation methods like k-folds can also be used

- If some model parameters are likely to cause overfitting, techniques for regularization like LASSO can be used that penalize these parameters

Also Read: Overfitting and Underfitting in Machine Learning

3. What is ‘training Set’ and ‘test Set’ in a Machine Learning Model? How Much Data Will You Allocate for Your Training, Validation, and Test Sets?

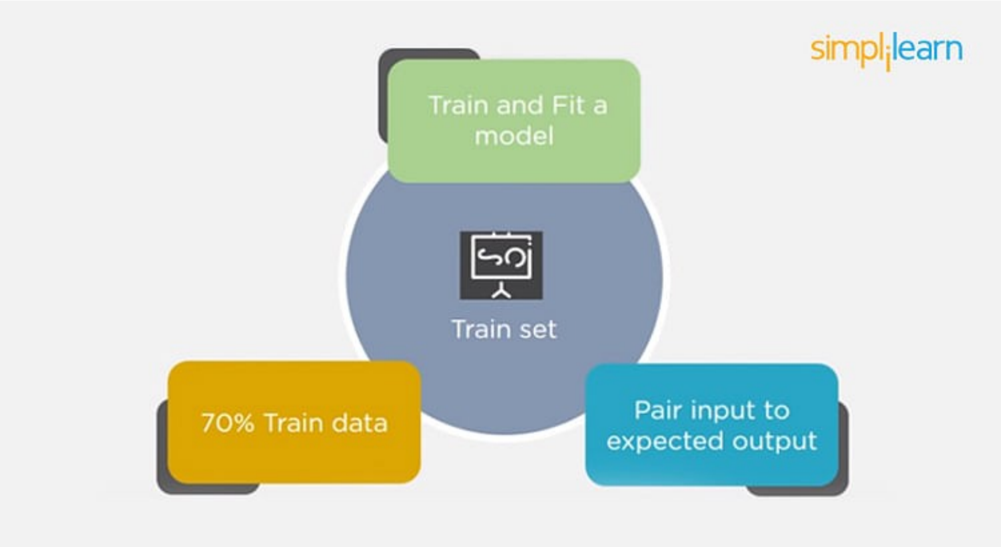
There is a three-step process followed to create a model:

1. Train the model
2. Test the model
3. Deploy the model

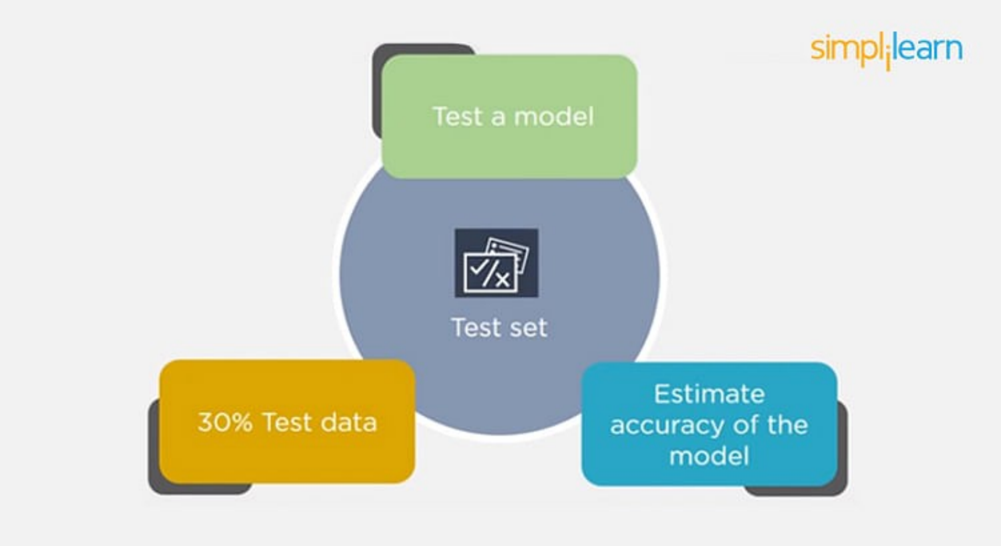
Training Set	Test Set
• The training set is examples given to the model to analyze and learn • 70% of the total data is typically taken as the training dataset • This is labeled data used to train the model	• The test set is used to test the accuracy of the hypothesis generated by the model • Remaining 30% is taken as testing dataset • We test without labeled data and then verify results with labels

Consider a case where you have labeled data for 1,000 records. One way to train the model is to expose all 1,000 records during the training process. Then you take a small set of the same data to test the model, which would give good results in this case.

But, this is not an accurate way of testing. So, we set aside a portion of that data called the ‘test set’ before starting the training process. The remaining data is called the ‘training set’ that we use for training the model. The training set passes through the model multiple times until the accuracy is high, and errors are minimized.



Now, we pass the test data to check if the model can accurately predict the values and determine if training is effective. If you get errors, you either need to change your model or retrain it with more data.



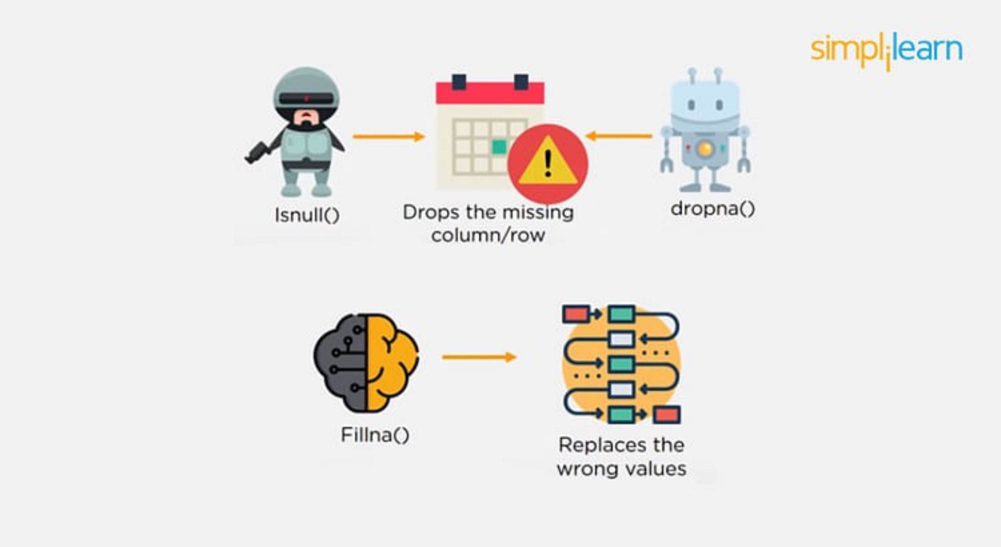
Regarding the question of how to split the data into a training set and test set, there is no fixed rule, and the ratio can vary based on individual preferences.

4. How Do You Handle Missing or Corrupted Data in a Dataset?

One of the easiest ways to handle missing or corrupted data is to drop those rows or 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



5. How Can You Choose a Classifier Based on a Training Set Data Size?

When the training set is small, a model that has a right bias and low variance seems to work better because they are less likely to overfit.

For example, Naive 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.

6. Explain the Confusion Matrix with Respect to Machine Learning Algorithms.

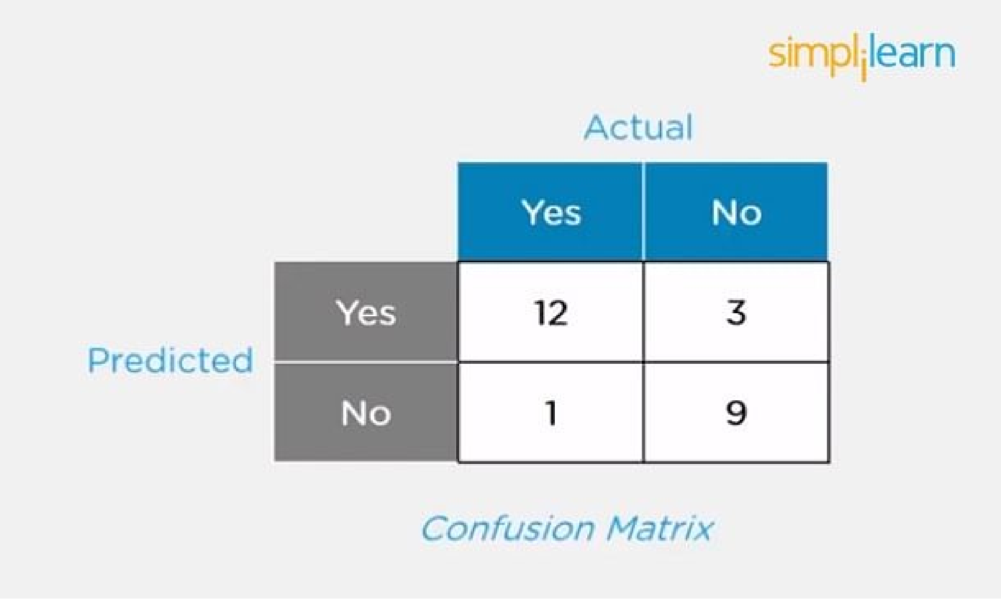
A confusion matrix (or error matrix) is a specific table that is used to measure the performance of an algorithm. It is mostly used in supervised learning; in unsupervised learning, it’s called the matching matrix.

The confusion matrix has two parameters:

- Actual
- Predicted

It also has identical sets of features in both of these dimensions.

Consider a confusion matrix (binary matrix) shown below:



		Actual	
		Yes	No
Predicted	Yes	12	3
	No	1	9

Confusion Matrix

Here,

For actual values:

Total Yes = 12+1 = 13

Total No = 3+9 = 12

Similarly, for predicted values:

Total Yes = 12+3 = 15

Total No = 1+9 = 10

For a model to be accurate, the values across the diagonals should be high. The total sum of all the values in the matrix equals the total observations in the test data set.

For the above matrix, total observations = 12+3+1+9 = 25

Now, accuracy = sum of the values across the diagonal/total dataset

= (12+9) / 25

= 21 / 25

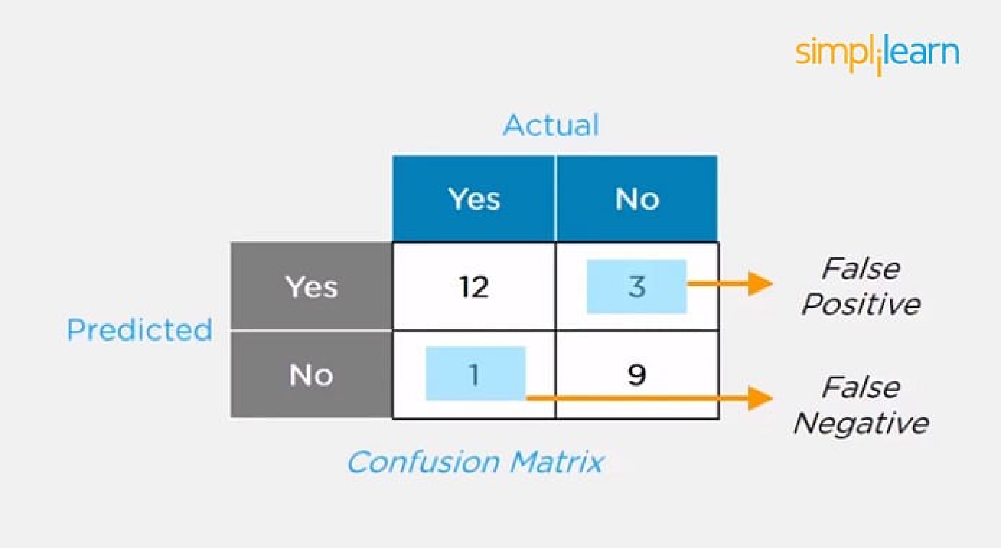
= 84%

7. What Is a False Positive and False Negative and How Are They Significant?

False positives are those cases that wrongly get classified as True but are False.

False negatives are those cases that wrongly get classified as False but are True.

In the term ‘False Positive,’ the word ‘Positive’ refers to the ‘Yes’ row of the predicted value in the confusion matrix. The complete term indicates that the system has predicted it as a positive, but the actual value is negative.



		Actual	
		Yes	No
Predicted	Yes	12	3
	No	1	9

Confusion Matrix

So, looking at the confusion matrix, we get:

False-positive = 3

True positive = 12

Similarly, in the term ‘False Negative,’ the word ‘Negative’ refers to the ‘No’ row of the predicted value in the confusion matrix. And the complete term indicates that the system has predicted it as negative, but the actual value is positive.

So, looking at the confusion matrix, we get:

False Negative = 1

True Negative = 9

8. What Are the Three Stages of Building a Model in Machine Learning?

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

Here, it's important to remember that once in a while, the model needs to be checked to make sure it's working correctly. It should be modified to make sure that it is up-to-date.

9. What is Deep Learning?

The Deep learning is a subset of machine learning that involves systems 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).

This is a commonly asked question asked in both Machine Learning Interviews as well as Deep Learning Interview Questions

10. What Are the Differences Between Machine Learning and Deep Learning?

Learn more: Difference Between AI,ML and Deep Learning

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

11. What Are the Applications of Supervised Machine Learning in Modern Businesses?

Applications of supervised machine learning include:

- Email Spam Detection

  Here we train the model using historical data that consists of emails categorized as spam or not spam. This labeled information is fed as input to the model.
- Healthcare Diagnosis

  By providing images regarding a disease, a model can be trained to detect if a person is suffering from the disease or not.
- 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

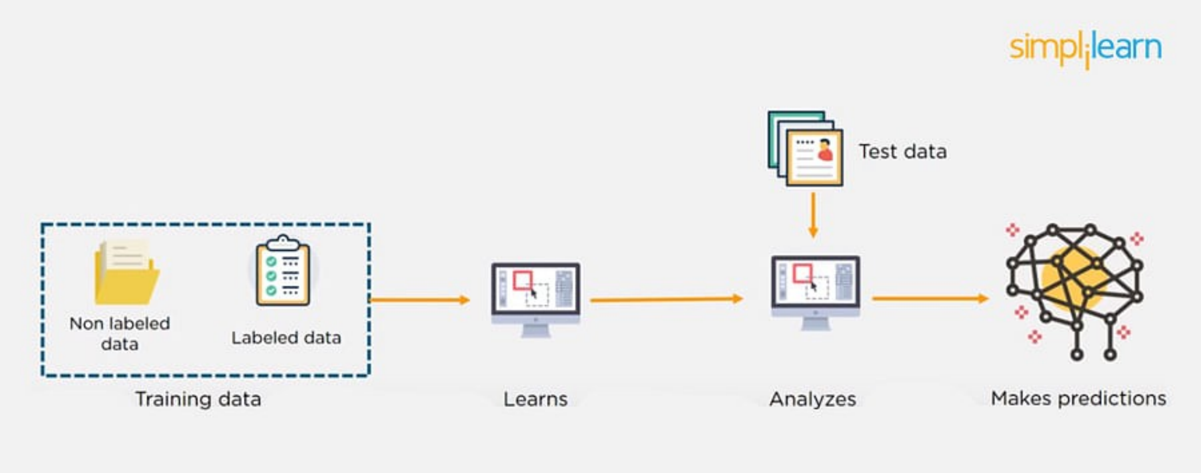
  By training the model to identify suspicious patterns, we can detect instances of possible fraud.

Related Interview Questions and Answers

12. What is Semi-supervised Machine Learning?

Supervised learning uses data that is completely labeled, whereas unsupervised learning uses no training data.

In the case of semi-supervised learning, the training data contains a small amount of labeled data and a large amount of unlabeled data.

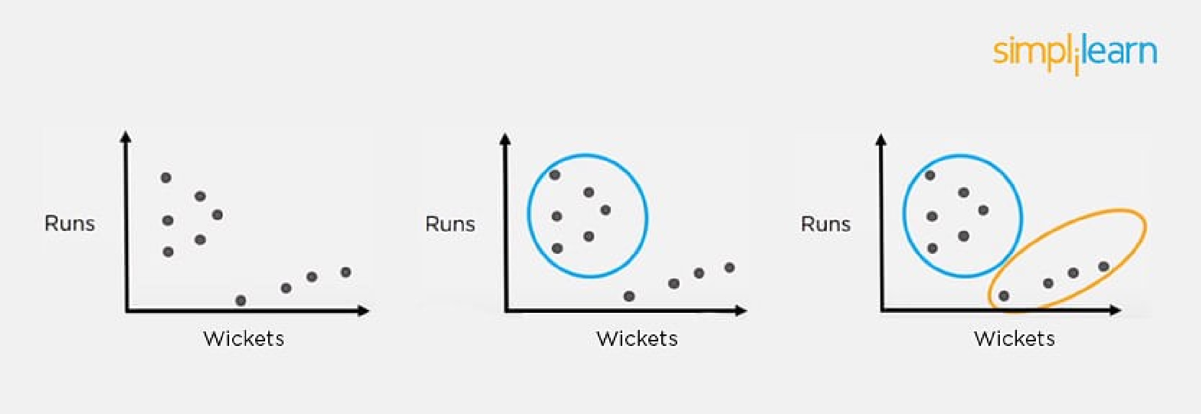


13. What Are Unsupervised Machine Learning Techniques?

There are two techniques used in unsupervised learning: clustering and association.

Clustering

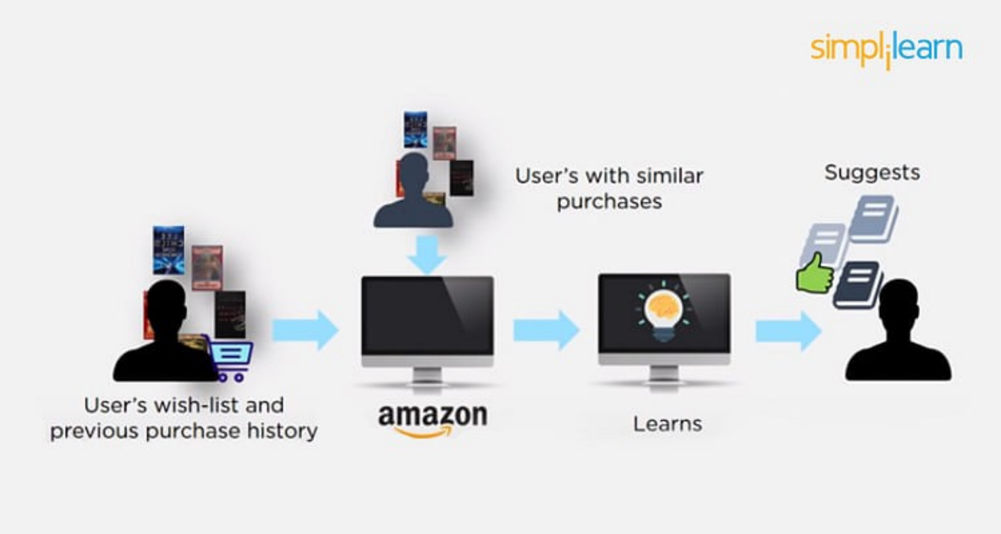
Clustering problems involve data to be divided into subsets. These subsets, also called clusters, contain data that are similar to each other. Different clusters reveal different details about the objects, unlike classification or regression.



Association

In an association problem, we identify patterns of associations between different variables or items.

For example, an e-commerce website can suggest other items for you to buy, based on the prior purchases that you have made, spending habits, items in your wishlist, other customers' purchase habits, and so on.


simplilearn
User's with similar purchases
Suggests
User's wish-list and previous purchase history
amazon
Learns

14. What is the Difference Between Supervised and Unsupervised Machine Learning?

- Supervised learning - This model learns from the labeled data and makes a future prediction as output
- Unsupervised learning - This model uses unlabeled input data and allows the algorithm to act on that information without guidance.

15. What is the Difference Between Inductive Machine Learning and Deductive Machine Learning?

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

16. Compare K-means and KNN Algorithms.

K-means	KNN
- K-Means is unsupervised - K-Means is a clustering algorithm - The points in each cluster are similar to each other, and each cluster is different from its neighboring clusters	- KNN is supervised in nature - KNN is a classification algorithm - It classifies an unlabeled observation based on its K (can be any number) surrounding neighbors

17. What Is ‘naive’ in the Naive Bayes Classifier?

The classifier is called ‘naive’ because it makes assumptions that may or may not turn out to be correct.

The algorithm assumes that the presence of one feature of a class is not related to the presence of any other feature (absolute independence of features), 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).

18. Explain How a System Can Play a Game of Chess Using Reinforcement Learning.

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 it towards the goal, it is rewarded. And, every time it takes a step that goes against that goal or in the reverse direction, it is penalized.

Earlier, chess programs had to determine the best moves after much research on numerous factors. Building a machine designed to play such games would require many rules to be specified.

With reinforced 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.

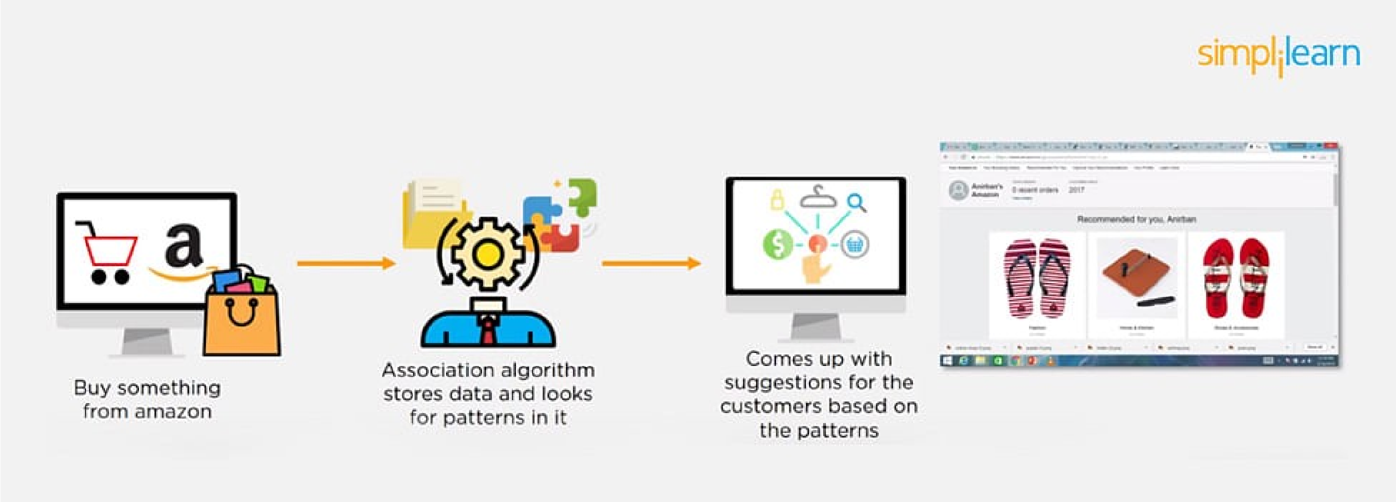
19. How Will You Know Which Machine Learning Algorithm to Choose for Your Classification Problem?

While there is no fixed rule to choose an algorithm for a classification problem, you can follow these guidelines:

- If accuracy is a concern, test different algorithms and cross-validate them
- If the training dataset is small, use models that have low variance and high bias
- If the training dataset is large, use models that have high variance and little bias

20. How is Amazon Able to Recommend Other Things to Buy? How Does the Recommendation Engine Work?

Once a user buys something from Amazon, Amazon stores that purchase data for future reference and finds products that are most likely also to be bought, it is possible because of the Association algorithm, which can identify patterns in a given dataset.


simplilearn
Buy something from amazon
Association algorithm stores data and looks for patterns in it
Comes up with suggestions for the customers based on the patterns
Recommended for you, Anirban

21. When Will You Use Classification over Regression?

Classification is used when your target is categorical, while regression is used when your target variable is continuous. Both classification and regression belong 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

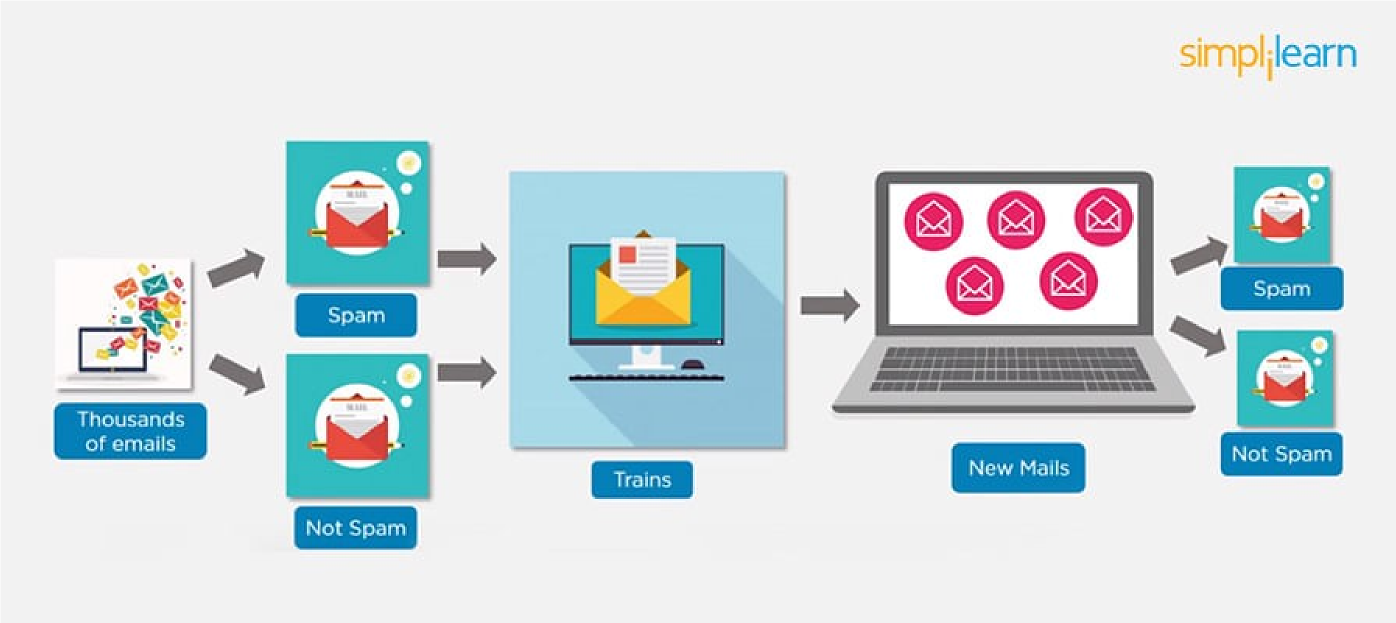
Examples of regression problems include:

- Estimating sales and price of a product
- Predicting the score of a team
- Predicting the amount of rainfall

22. How Do You Design an Email Spam Filter?

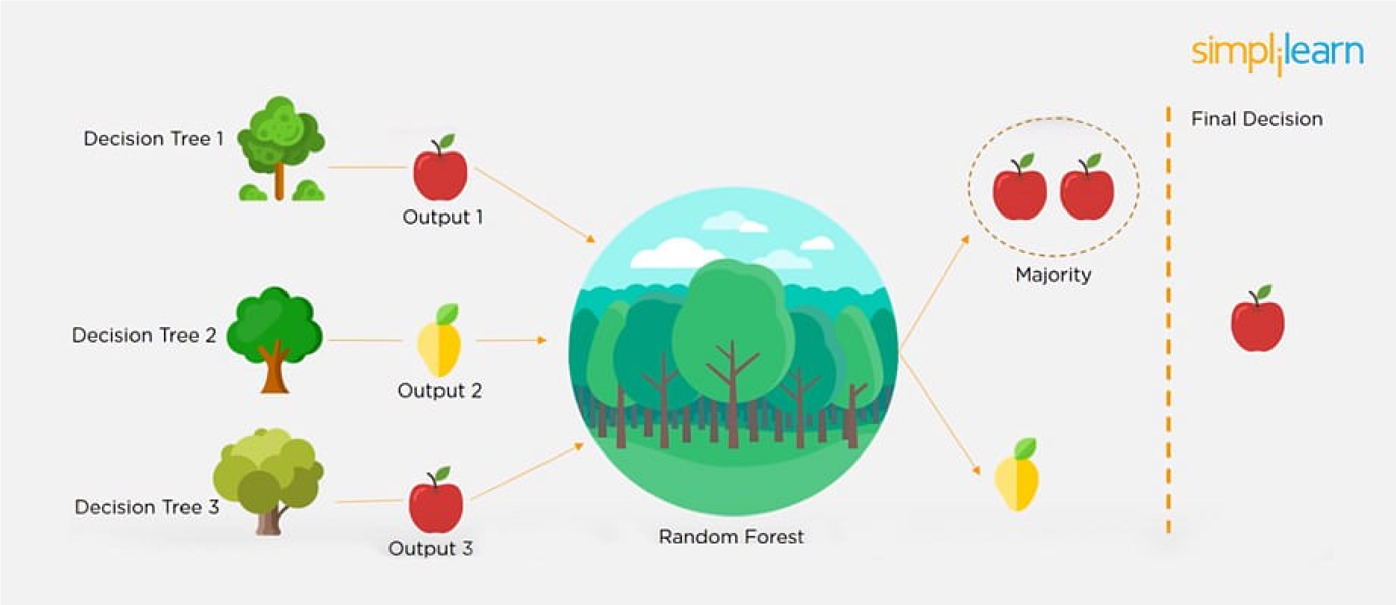
Building a spam filter involves the following process:

- 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



23. What is a Random Forest?

A ‘random forest’ is a supervised machine learning algorithm that is generally used for 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 decision.

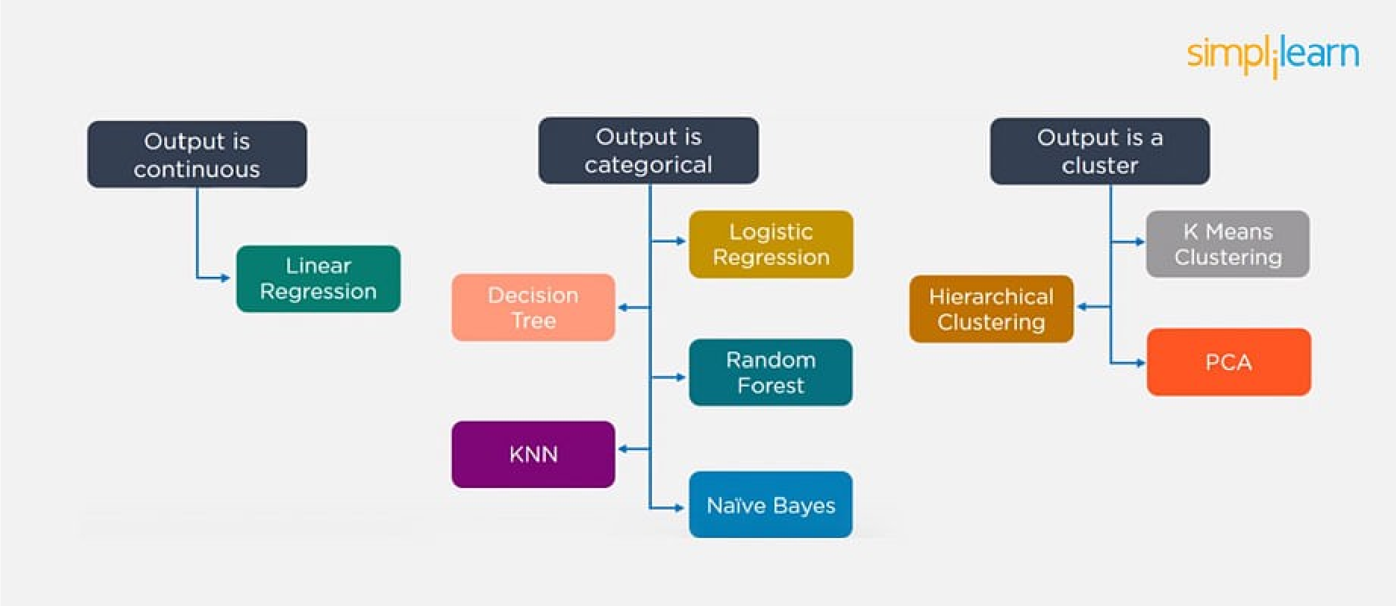


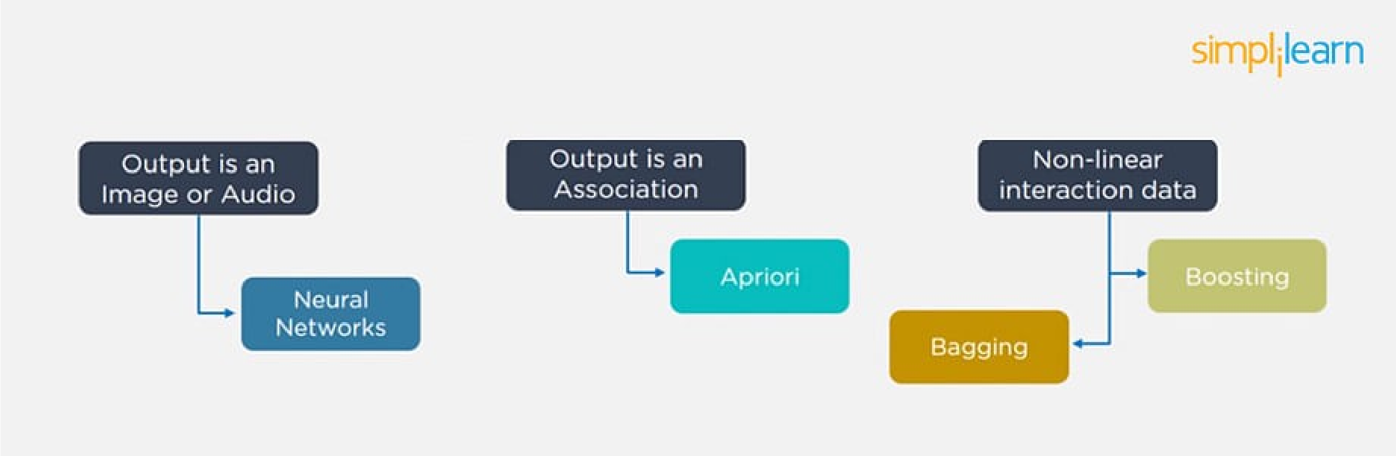
24. Considering a Long List of Machine Learning Algorithms, given a Data Set, How Do You Decide Which One to Use?

There is no master algorithm for all situations. Choosing an algorithm depends on the following questions:

- How much data do you have, and is it continuous or categorical?
- Is the problem related to classification, association, clustering, or regression?
- Predefined variables (labeled), unlabeled, or mix?
- What is the goal?

Based on the above questions, the following algorithms can be used:





25. What is Bias and Variance in a Machine Learning Model?

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.

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

Variance

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

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

26. What is the Trade-off Between Bias and Variance?

The bias-variance decomposition essentially decomposes the learning error from any algorithm by adding the bias, variance, and a bit of irreducible error due to noise in the underlying dataset.

Necessarily, if you make the model more complex and add more variables, you'll lose bias but gain variance. To get the optimally-reduced amount of error, you'll have to trade off bias and variance. Neither high bias nor high variance is desired.

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.

27. Define Precision and Recall.

Precision

Precision is the ratio of several events you can correctly recall to the total number of events you recall (mix of correct and wrong recalls).

Precision = (True Positive) / (True Positive + False Positive)

Recall

A recall is the ratio of the number of events you can recall the number of total events.

Recall = (True Positive) / (True Positive + False Negative)

28. What is a Decision Tree Classification?

A decision tree builds classification (or regression) models as a tree structure, with datasets broken up into ever-smaller subsets while developing the decision tree, literally in a tree-like way with branches and nodes. Decision trees can handle both categorical and numerical data.

29. What is Pruning in Decision Trees, and How Is It Done?

Pruning is a technique in machine learning that reduces the size of decision trees. It reduces the complexity of the final classifier, and hence improves predictive accuracy by the reduction of overfitting.

Pruning can occur in:

- Top-down fashion. It will traverse nodes and trim subtrees starting at the root
- Bottom-up fashion. It will begin at the leaf nodes

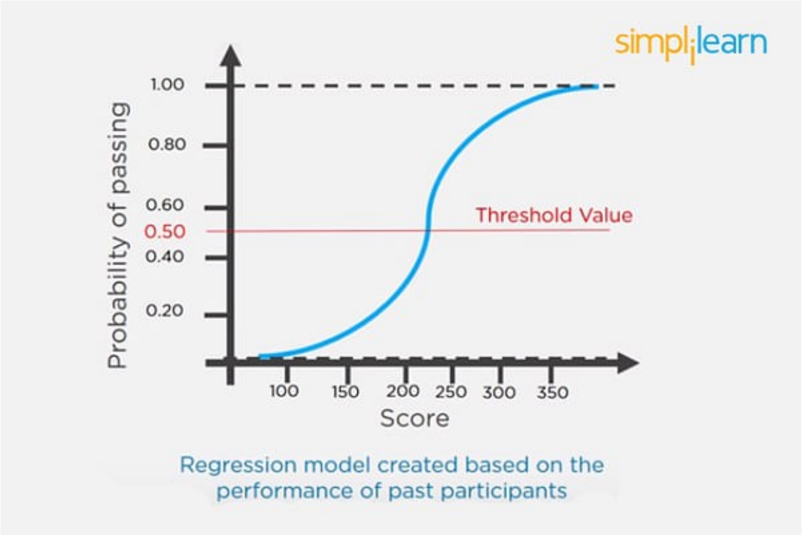
There is a popular pruning algorithm called reduced error pruning, in which:

- Starting at the leaves, each node is replaced with its most popular class
- If the prediction accuracy is not affected, the change is kept
- There is an advantage of simplicity and speed

30. Briefly Explain Logistic Regression.

Logistic regression is a classification algorithm used to predict a binary outcome for a given set of independent variables.

The output of logistic regression is either a 0 or 1 with a threshold value of generally 0.5. Any value above 0.5 is considered as 1, and any point below 0.5 is considered as 0.



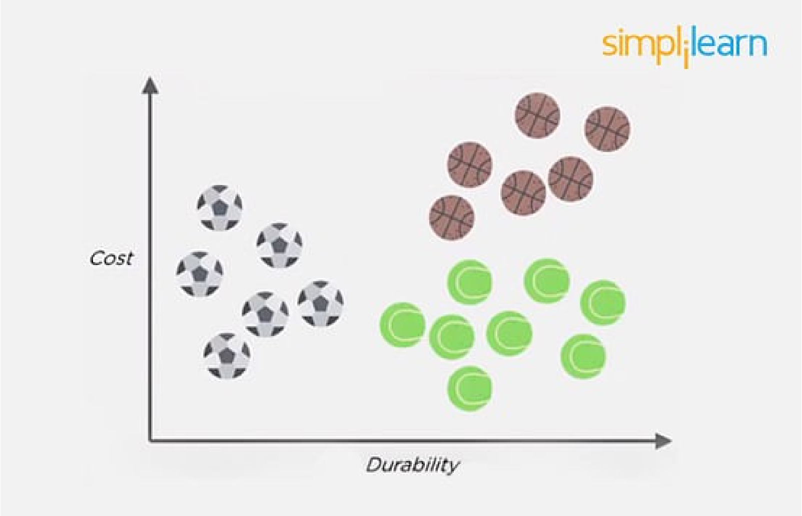
31. Explain the K Nearest Neighbor Algorithm.

K nearest neighbor algorithm is a classification algorithm that works in a way that a new data point is assigned to a neighboring group to which it is most similar.

In K nearest neighbors, K can be an integer greater than 1. So, for every new data point, we want to classify, we compute to which neighboring group it is closest.

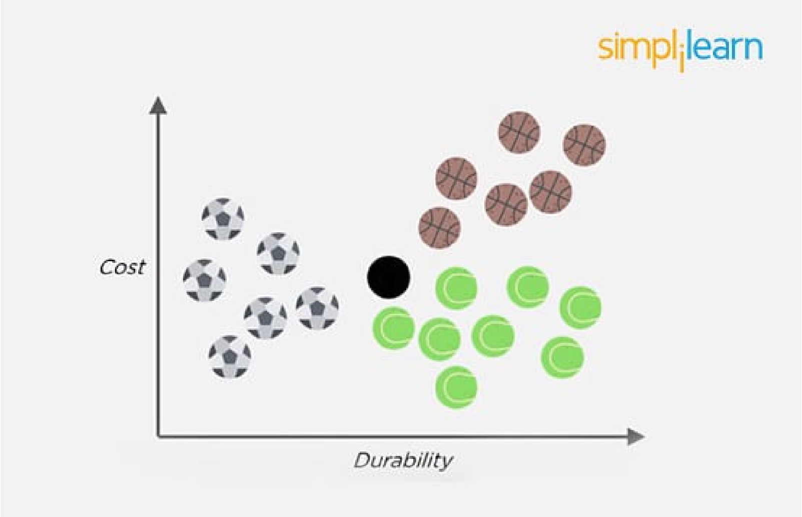
Let us classify an object using the following example. Consider there are three clusters:

- Football
- Basketball
- Tennis ball



Let the new data point to be classified is a black ball. We use KNN to classify it. Assume K = 5 (initially).

Next, we find the K (five) nearest data points, as shown.



Observe that all five selected points do not belong to the same cluster. There are three tennis balls and one each of basketball and football.

When multiple classes are involved, we prefer the majority. Here the majority is with the tennis ball, so the new data point is assigned to this cluster.

32. What is a Recommendation System?

Anyone who has used Spotify or shopped at Amazon will recognize a recommendation system: It’s an information filtering system that predicts what a user might want to hear or see based on choice patterns provided by the user.

33. What is Kernel SVM?

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.

34. What Are Some Methods of Reducing Dimensionality?

You can reduce dimensionality by combining features with feature engineering, removing collinear features, or using algorithmic dimensionality reduction.

Now that you have gone through these machine learning interview questions, you must have got an idea of your strengths and weaknesses in this domain.

35. What is Principal Component Analysis?

Principal Component Analysis or PCA is a multivariate statistical technique that is used for analyzing quantitative data. The objective of PCA is to reduce higher dimensional data to lower dimensions, remove noise, and extract crucial information such as features and attributes from large amounts of data.

36. What do you understand by the F1 score?

The F1 score is a metric that combines both Precision and Recall. It is also the weighted average of precision and recall.

The F1 score can be calculated using the below formula:

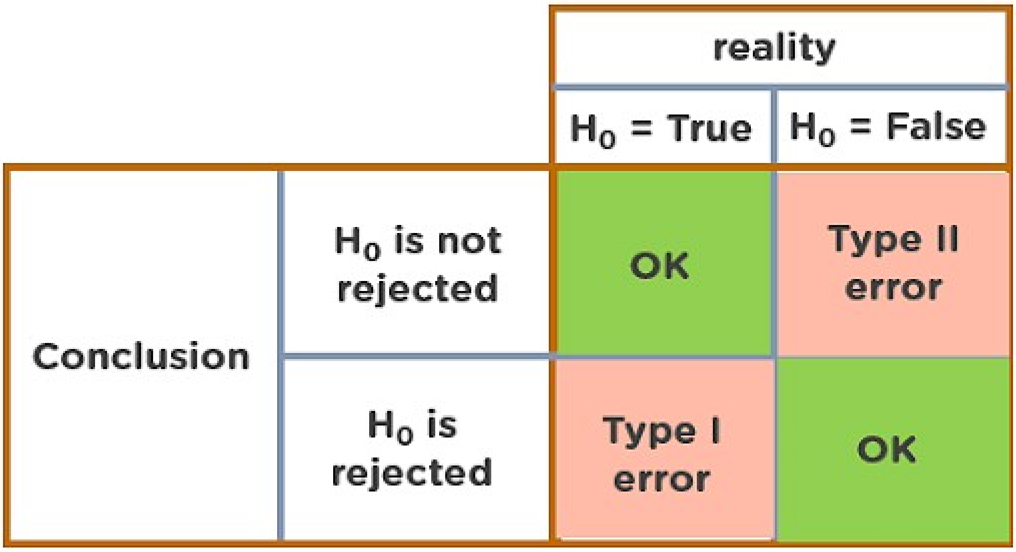
F1 = 2 * (P * R) / (P + R)

The F1 score is one when both Precision and Recall scores are one.

37. What do you understand by Type I vs Type II error?

Type I Error: Type I error occurs when the null hypothesis is true and we reject it.

Type II Error: Type II error occurs when the null hypothesis is false and we accept it.



		reality	
		H_0 = True	H_0 = False
Conclusion	H_0 is not rejected	OK	Type II error
	H_0 is rejected	Type I error	OK

38. Explain Correlation and Covariance?

Correlation: Correlation tells us how strongly two random variables are related to each other. It takes values between -1 to +1.

Formula to calculate Correlation:

$$\text{Correlation} = \frac{\text{Cov}(x, y)}{6x * 6y}$$

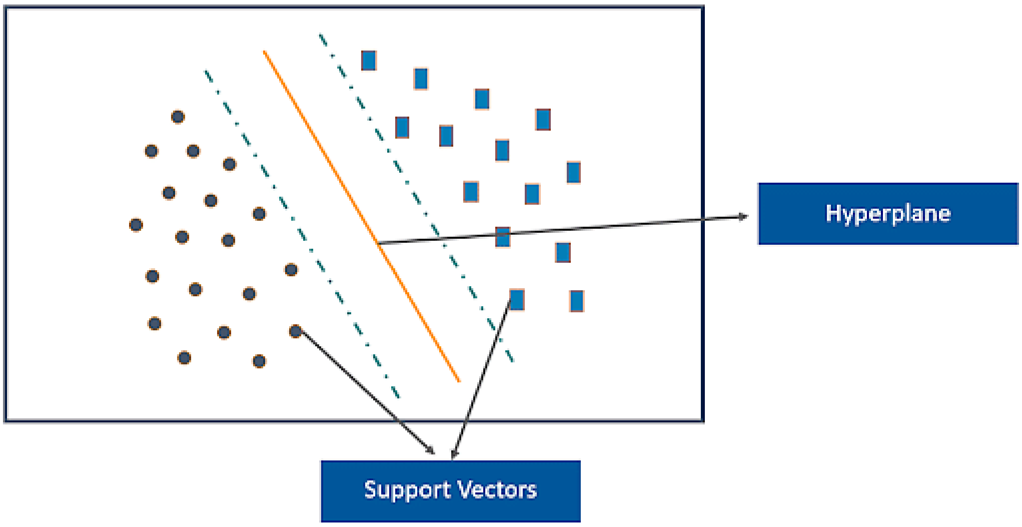
Covariance: Covariance tells us the direction of the linear relationship between two random variables. It can take any value between - ∞ and + ∞.

Formula to calculate Covariance:

$$\text{Cov}(x, y) = \frac{\sum (x_i - x') * (y_i - y')}{N}$$

39. What are Support Vectors in SVM?

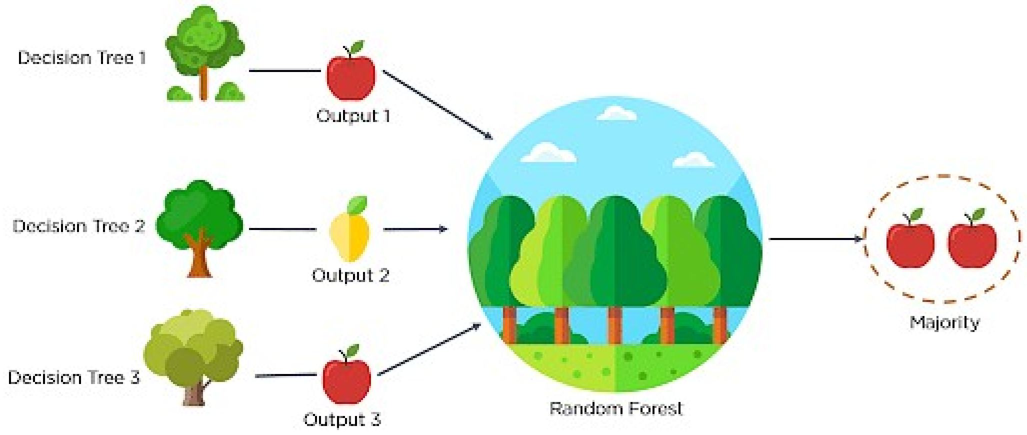
Support Vectors are data points that are nearest to the hyperplane. It influences the position and orientation of the hyperplane. Removing the support vectors will alter the position of the hyperplane. The support vectors help us build our support vector machine model.



40. What is Ensemble learning?

Ensemble learning is a combination of the results obtained from multiple machine learning models to increase the accuracy for improved decision-making.

Example: A Random Forest with 100 trees can provide much better results than using just one decision tree.



41. What is Cross-Validation?

Cross-Validation in Machine Learning is a statistical resampling technique that uses different parts of the dataset to train and test a machine learning algorithm on different iterations. The aim of cross-validation is to test the model's ability to predict a new set of data that was not used to train the model. Cross-validation avoids the overfitting of data.

K-Fold Cross Validation is the most popular resampling technique that divides the whole dataset into K sets of equal sizes.

42. What are the different methods to split a tree in a decision tree algorithm?

Variance: Splitting the nodes of a decision tree using the variance is done when the target variable is continuous.

$$\text{Variance} = \frac{\sum (X-\bar{X})^2}{N}$$

Information Gain: Splitting the nodes of a decision tree using Information Gain is preferred when the target variable is categorical.

$$IG = 1 - \text{Entropy}$$

$$\text{Entropy} = - \sum p_i \log_2 p_i$$

Gini Impurity: Splitting the nodes of a decision tree using Gini Impurity is followed when the target variable is categorical.

$$I_G(n) = 1 - \sum_{i=1}^{n} (p_i)^2$$

43. How does the Support Vector Machine algorithm handle self-learning?

The SVM algorithm has a learning rate and expansion rate which takes care of self-learning. The learning rate compensates or penalizes the hyperplanes for making all the incorrect moves while the expansion rate handles finding the maximum separation area between different classes.

44. What are the assumptions you need to take before starting with linear regression?

There are primarily 5 assumptions for a Linear Regression model:

- Multivariate normality
- No auto-correlation
- Homoscedasticity
- Linear relationship
- No or little multicollinearity

45. What is the difference between Lasso and Ridge regression?

Lasso(also known as L1) and Ridge(also known as L2) regression are two popular regularization techniques that are used to avoid overfitting of data. These methods are used to penalize the coefficients to find the optimum solution and reduce complexity. The Lasso regression works by penalizing the sum of the absolute values of the coefficients. In Ridge or L2 regression, the penalty function is determined by the sum of the squares of the coefficients.

Become Part of the Machine Learning Talent Pool

With technology ramping up, jobs in the field of data science and AI will continue to be in demand. Candidates who upgrade their skills and become well-versed in these emerging technologies can find many job opportunities with impressive salaries. Looking forward to becoming a Machine Learning Engineer? Enroll in Simplilearn's Caltech Post Graduate Program in AI & ML and get certified today. Based on your experience level, you may be asked to demonstrate your skills in machine learning, additionally, but this depends mostly on the role you're pursuing. These machine learning interview questions and answers will prepare you to clear your interview on the first attempt!

Apart from the above mentioned interview questions, it is also important to have a fair understanding of frequently asked Data Science interview questions.

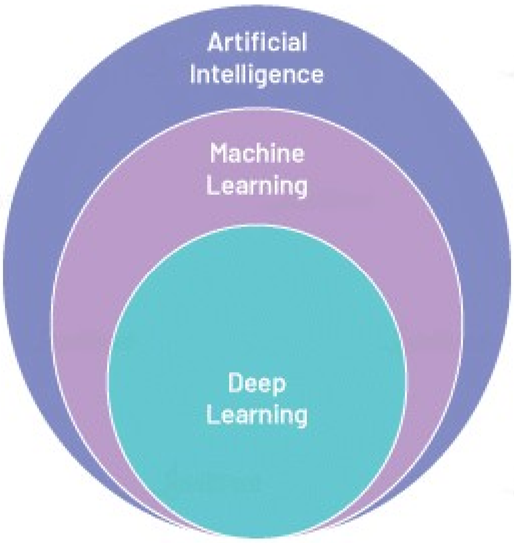
Considering this trend, Simplilearn offers Caltech Post Graduate Program in AI & ML certification course to help you gain a firm hold of machine learning concepts. This course is well-suited for those at the intermediate level, including:

- Analytics managers
- Business analysts
- Information architects
- Developers looking to become data scientists
- Graduates seeking a career in data science and machine learning

Facing the machine learning interview questions would become much easier after you complete this course.\

1. Explain Machine Learning, Artificial Intelligence, and Deep Learning

It is common to get confused between the three in-demand technologies, Machine Learning, Artificial Intelligence, and Deep Learning. These three technologies, though a little different from one another, are interrelated. While Deep Learning is a subset of Machine Learning, Machine Learning is a subset of Artificial Intelligence. Since some terms and techniques may overlap in these technologies, it is easy to get confused among them.



So, let us learn about these technologies in detail:

- Machine Learning: Machine Learning involves various statistical and Deep Learning techniques that allow machines to use their past experiences and get better at performing specific tasks without having to be monitored.
- Artificial Intelligence: Artificial Intelligence uses numerous Machine Learning and Deep Learning techniques that enable computer systems to perform tasks using human-like intelligence with logic and rules. Artificial intelligence is used in every sector hence it is necessary to pursue Artificial Intelligence Course to make your career in AI.
- Deep Learning: Deep Learning comprises several algorithms that enable software to learn from themselves and perform various business tasks including

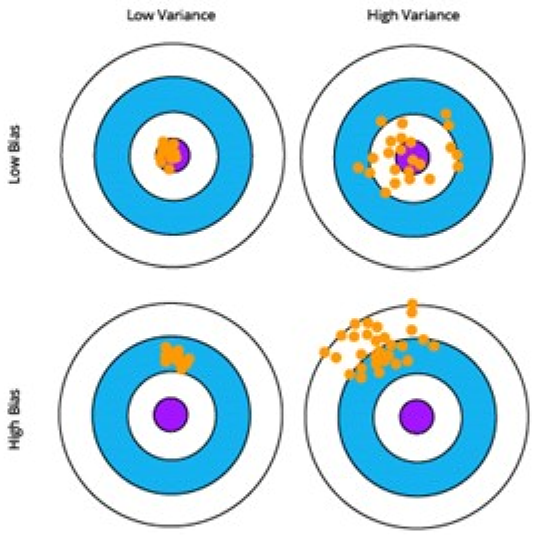
image and speech recognition. Deep Learning is possible when systems expose their multilayered neural networks to large volumes of data for learning.

Willing to master AI & ML skills? Check our [AI and Machine Learning Courses](#) in collaboration with top universities Now!

2. What is Bias and Variance in Machine Learning?

- Bias is the difference between the average prediction of a model and the correct value of the model. If the bias value is high, then the prediction of the model is not accurate. Hence, the bias value should be as low as possible to make the desired predictions.
- Variance is the number that gives the difference of prediction over a training set and the anticipated value of other training sets. High variance may lead to large fluctuation in the output. Therefore, a model's output should have low variance.

The following diagram shows the bias-variance trade-off:



Here, the desired result is the blue circle at the center. If we get off from the blue section, then the prediction goes wrong.

Interested in learning Machine Learning? Enroll in our [Machine Learning Training](#) now!

3. What is Clustering in Machine Learning?

Clustering is a technique used in unsupervised learning that involves grouping data points. The clustering algorithm can be used with a set of data points. This technique will allow you to classify all data points into their particular groups. The data points that are thrown into the same category have similar features and properties, while the data points that belong to different groups have distinct features and properties. Statistical data analysis can be performed by this method. Let us take a look at three of the most popular and useful clustering algorithms.

- K-means clustering: This algorithm is commonly used when there is data with no specific group or category. K-means clustering allows you to find the hidden patterns in the data, which can be used to classify the data into various groups. The variable *k* is used to represent the number of groups the data is divided into, and the data points are clustered using the similarity of features. Here, the centroids of the clusters are used for labeling new data.
- Mean-shift clustering: The main aim of this algorithm is to update the center-point candidates to be mean and find the center points of all groups. In mean-shift clustering, unlike k-means clustering, the possible number of clusters need not be selected as it can automatically be discovered by the mean shift.
- Density-based spatial clustering of applications with noise (DBSCAN): This clustering algorithm is based on density and has similarities with mean-shift clustering. There is no need to preset the number of clusters, but unlike mean-shift clustering, DBSCAN identifies outliers and treats them like noise. Moreover, it can identify arbitrarily-sized and -shaped clusters without much effort.

4. What is Linear Regression in Machine Learning?

Linear Regression is a supervised Machine Learning algorithm. It is used to find the linear relationship between the dependent and independent variables for predictive analysis.

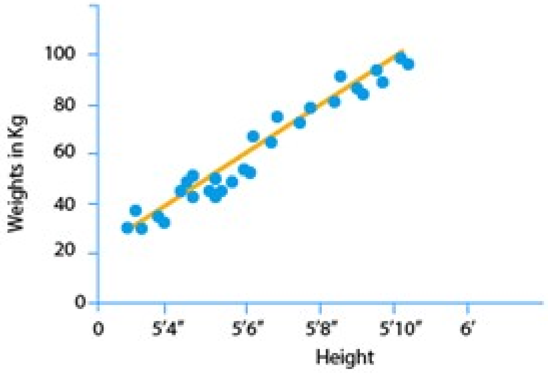
The equation for Linear Regression:

$$Y = A + B.X$$

where:

- X is the input or independent variable
- Y is the output or dependent variable
- a is the intercept, and b is the coefficient of X

Below is the best-fit line that shows the data of weight, Y or the dependent variable, and the ata of height, X or the independent variable, of 21-year-old candidates scattered over the plot. The straight line shows the best linear relationship that would help in predicting the weight of candidates according to their height.

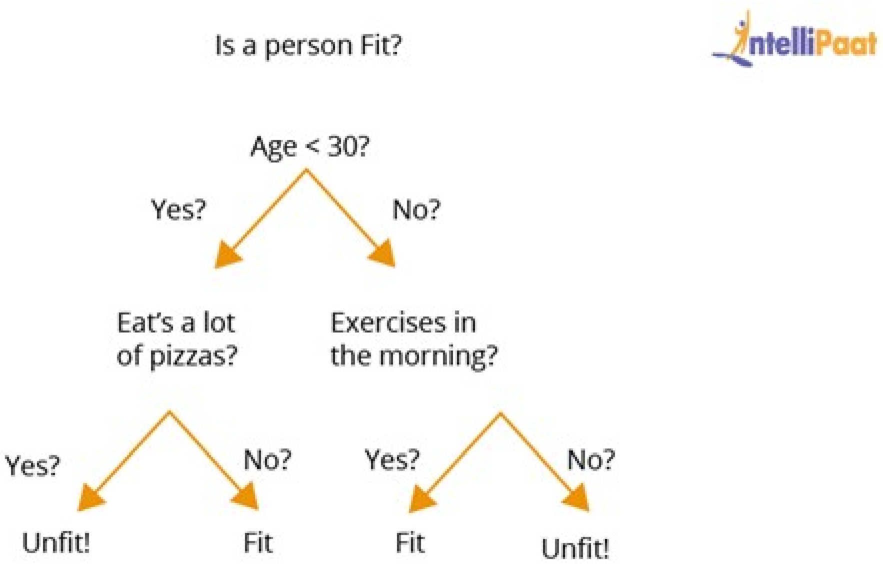


To get this best-fit line, the best values of a and b should be found. By adjusting the values of a and b, the errors in the prediction of Y can be reduced.

This is how linear regression helps in finding the linear relationship and predicting the output.

5. What is a Decision Tree in Machine Learning?

A decision tree is used to explain the sequence of actions that must be performed to get the desired output. It is a hierarchical diagram that shows the actions.



An algorithm can be created for a decision tree on the basis of the set hierarchy of actions.

In the above decision-tree diagram, a sequence of actions has been made for driving a vehicle with or without a license.

6. What is Overfitting in Machine Learning and how can it be avoided?

Overfitting happens when a machine has an inadequate dataset and tries to learn from it. So, overfitting is inversely proportional to the amount of data.

For small databases, overfitting can be bypassed by the cross-validation method. In this approach, a dataset is divided into two sections. These two sections will comprise the testing and training dataset. To train a model, the training dataset is used, and for testing the model for new inputs, the testing dataset is used. This is how to avoid overfitting.

7. What is Hypothesis in Machine Learning?

Machine Learning allows the use of available dataset to understand a specific function that maps input to output in the best possible way. This problem is known as function approximation. Here, approximation needs to be used for the unknown target function that maps all plausible observations based on the given problem in the best manner. Hypothesis in Machine learning is a model that helps in approximating the target function and performing the necessary input-to-output mappings. The choice and configuration of algorithms allow defining the space of plausible hypotheses that may be represented by a model.

In the hypothesis, lowercase h (h) is used for a specific hypothesis, while uppercase h (H) is used for the hypothesis space that is being searched. Let us briefly understand these notations:

- Hypothesis (h): A hypothesis is a specific model that helps in mapping input to output; the mapping can further be used for evaluation and prediction.
- Hypothesis set (H): Hypothesis set consists of a space of hypotheses that can be used to map inputs to outputs, which can be searched. The general constraints include the choice of problem framing, the model, and the model configuration.

8. What are the differences between Deep Learning and Machine Learning?

- Deep Learning: Deep Learning allows machines to make various business-related decisions using artificial neural networks, which is one of the reasons why it needs a vast amount of data for training. Since there is a lot of computing power required, Deep Learning requires high-end systems as well. The systems acquire various properties and features with the help of the given data, and the problem is solved using an end-to-end method.

- Machine Learning: Machine Learning gives machines the ability to make business decisions without any external help, using the knowledge gained from past data. Machine Learning systems require relatively small amounts of data to train themselves, and most of the features need to be manually coded and understood in advance. In Machine Learning, a given business problem is dissected into two and then solved individually. Once the solutions of both have been acquired, they are then combined.

9. What are the differences between Supervised and Unsupervised Machine Learning?

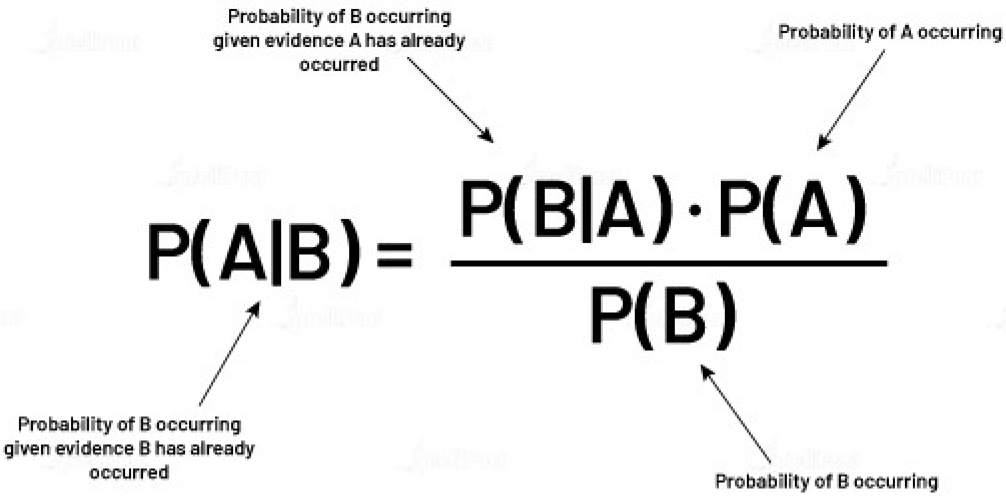
- Supervised learning: The algorithms of supervised learning use labeled data to get trained. The models take direct feedback to confirm whether the output that is being predicted is, indeed, correct. Moreover, both the input data and the output data are provided to the model, and the main aim here is to train the model to predict the output upon receiving new data. Supervised learning offers accurate results and can largely be divided into two parts, classification and regression.
- Unsupervised learning: The algorithms of unsupervised learning use unlabeled data for training purposes. In unsupervised learning, the models identify hidden data trends and do not take any feedback. The unsupervised learning model is only provided with input data. Unsupervised learning's main aim is to identify hidden patterns to extract information from unknown sets of data. It can also be classified into two parts, clustering, and associations. Unfortunately, unsupervised learning offers results that are comparatively less accurate.

Learn more about Machine Learning from this Machine Learning tutorial!

10. What is Bayes's Theorem in Machine Learning?

Bayes’s theorem offers the probability of any given event to occur using prior knowledge. In mathematical terms, it can be defined as the true positive rate of the given sample condition divided by the sum of the true positive rate of the said condition and the false positive rate of the entire population.

Two of the most significant applications of Bayes’s theorem in Machine Learning are Bayesian optimization and Bayesian belief networks. This theorem is also the foundation behind the Machine Learning brand that involves the Naive Bayes classifier.


Probability of B occurring given evidence A has already occurred
Probability of A occurring
P(A|B) = P(B|A) · P(A) / P(B)
Probability of B occurring given evidence B has already occurred
Probability of B occurring

11. What is PCA in Machine Learning?

Multidimensional data is at play in the real world. Data visualization and computation become more challenging with the increase in dimensions. In such a scenario, the dimensions of data might have to be reduced to analyze and visualize it easily. This is done by:

- Removing irrelevant dimensions
- Keeping only the most relevant dimensions

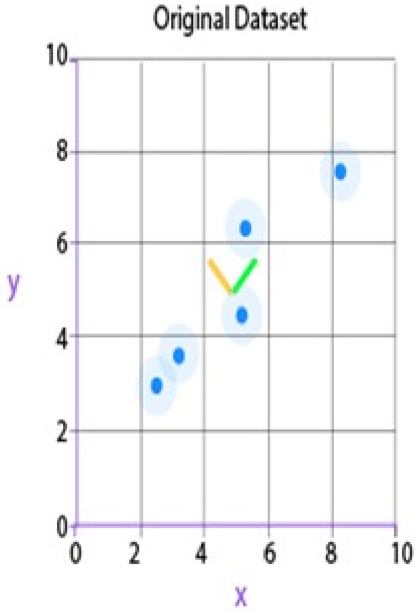
This is where Principal Component Analysis (PCA) is used.

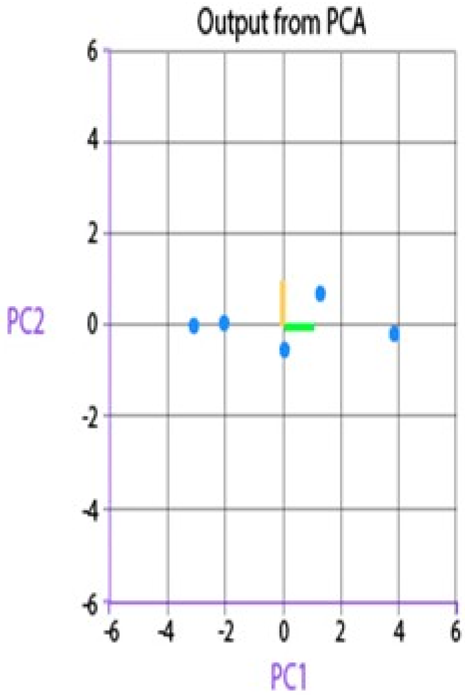
The goal of PCA is to find a fresh collection of uncorrelated dimensions (orthogonal) and rank them on the basis of variance.

Mechanism of PCA:

- Compute the covariance matrix for data objects
- Compute eigenvectors and eigenvalues in descending order
- Select the initial *N* eigenvectors to get new dimensions
- Finally, change the initial n-dimensional data objects into N-dimensions

Example: Below are two graphs showing data points or objects and two directions, one is green and the other is yellow. Graph 2 is arrived at by rotating Graph 1 so that the x-axis and y-axis represent the green and yellow direction respectively.





After the rotation of data points, it can be inferred that the green direction, the x-axis, gives the line that best fits the data points.

Here, two-dimensional data is being represented; but in real life, the data would be multidimensional and complex. So, after recognizing the importance of each direction, the area of dimensional analysis can be reduced by cutting off the less-significant directions.

Now, we will go through another important Machine Learning interview question on PCA.

Career Transition

-
-

12. What is Support Vector Machine (SVM) in Machine Learning?

SVM is a Machine Learning algorithm that is majorly used for classification. It is used on top of the high dimensionality of the characteristic vector.

The following is the code for SVM classifier:

```
# Introducing required libraries
from sklearn import datasets
from sklearn.metrics import confusion_matrix
from sklearn.model_selection import train_test_split
# Stacking the Iris dataset
iris = datasets.load_iris()
# A -> features and B -> label
A = iris.data
B = iris.target
# Breaking A and B into train and test data
A_train, A_test, B_train, B_test = train_test_split(A, B,
random_state = 0)
# Training a linear SVM classifier
from sklearn.svm import SVC
svm_model_linear = SVC(kernel = 'linear', C = 1).fit(A_train,
B_train)
svm_predictions = svm_model_linear.predict(A_test)
# Model accuracy for A_test
accuracy = svm_model_linear.score(A_test, B_test)
# Creating a confusion matrix
cm = confusion_matrix(B_test, svm_predictions)
```

13. What is Cross-validation in Machine Learning?

Cross-validation allows a system to increase the performance of the given Machine Learning algorithm, which is fed a number of sample data from the dataset. This sampling process is done to break the dataset into smaller parts that have the same number of rows, out of which a random part is selected as a test set and the rest of the parts are kept as train sets. Cross-validation consists of the following techniques:

- Holdout method
- K-fold cross-validation
- Stratified k-fold cross-validation
- Leave p-out cross-validation

14. What is Entropy in Machine Learning?

Entropy in Machine Learning measures the randomness in the data that needs to be processed. The more entropy in the given data, the more difficult it becomes to draw any useful conclusion from the data. For example, let us take the flipping of a coin. The result of this act is random as it does not favor heads or tails. Here, the result for any number of tosses cannot be predicted easily as there is no definite relationship between the action of flipping and the possible outcomes.

15. What is Epoch in Machine Learning?

Epoch in Machine Learning is used to indicate the count of passes in a given training dataset where the Machine Learning algorithm has done its job. Generally, when there is a large chunk of data, it is grouped into several batches. All these batches go through the given model, and this process is referred to as iteration. Now, if the batch size comprises the complete training dataset, then the count of iterations is the same as that of epochs.
In case there is more than one batch, d*e=i*b is the formula used, wherein d is the dataset, e is the number of epochs, i is the number of iterations, and b is the batch size.

Intermediate Machine Learning Interview Questions and Answers

16. What are the types of Machine Learning?

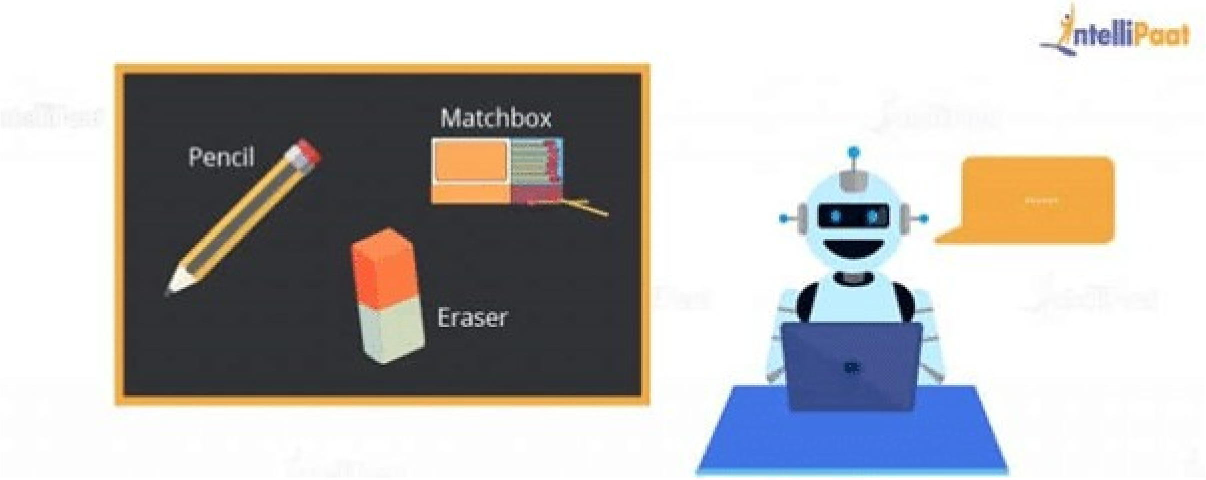
This is one of the most basic interview questions that everyone must know.

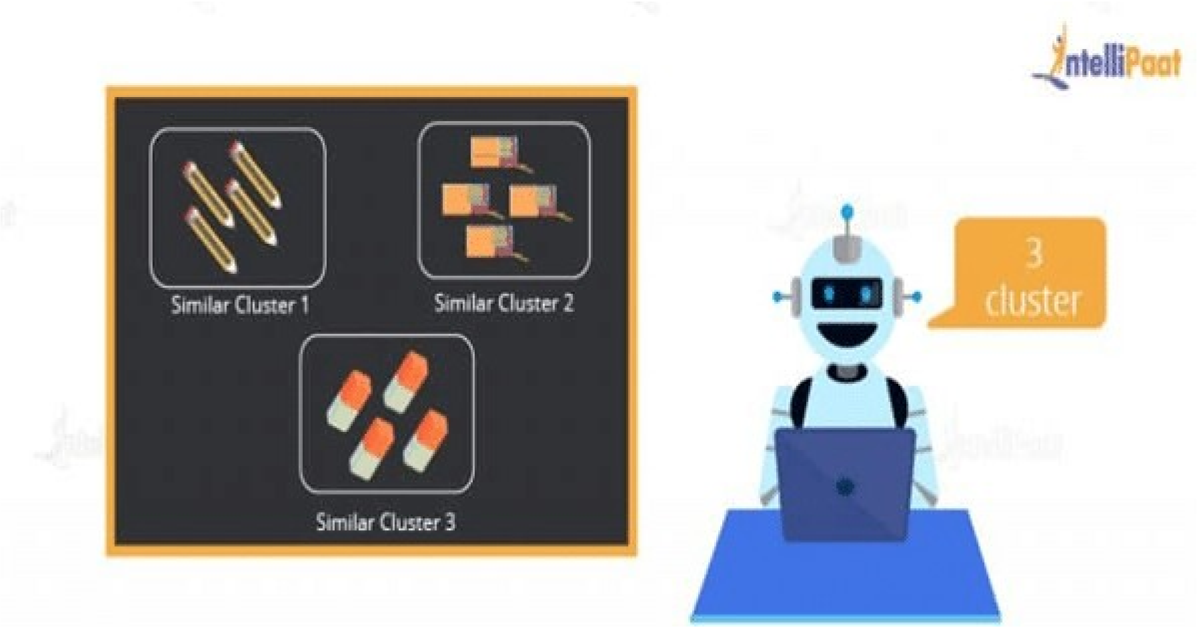
So, basically, there are three types of Machine Learning. They are described as follows:

Supervised learning: In this type of Machine Learning, machines learn under the supervision of labeled data. There is a training dataset on which a machine is trained, and it gives the output according to its training.

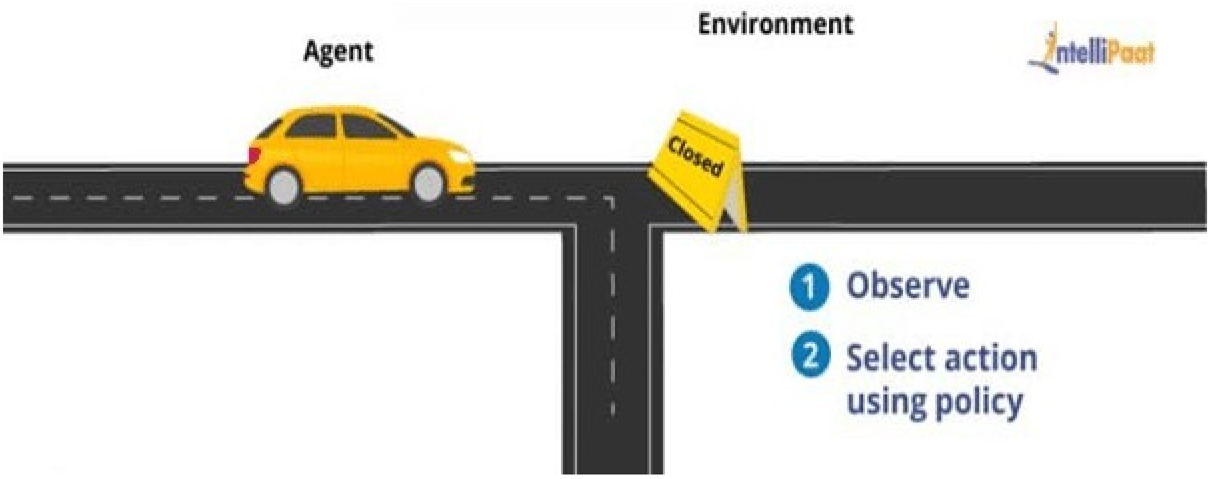
Patient Symptoms Positive or Negative Tag

Unsupervised learning: This type of Machine Learning has unlabeled data unlike supervised learning. Unsupervised learning works on data under absolutely no supervision. Unsupervised learning tries to identify patterns in data and makes clusters of similar entities. After that, when a new input data is fed into the model, it does not identify the entity; rather, it puts the entity in a cluster of similar objects.





Reinforcement learning: Reinforcement learning includes models that learn and traverse to find the best possible move. The algorithms for reinforcement learning are constructed in a way that they try to find the best possible suite of action on the basis of the reward and punishment theory.



-50 Points
Closed
3 Action
4 Reward or Penalty

For next time....
Closed
4 Update policy
5 Iterate the Process

17. Differentiate between Classification and Regression in Machine Learning

In Machine Learning, there are various types of prediction problems based on supervised and unsupervised learning. They are classification, regression, clustering, and association. Here, we will discuss classification and regression.

Classification: In classification, a Machine Learning model is created that assists in differentiating data into separate categories. The data is labeled and categorized based on the input parameters.

For example, predictions have to be made on the churning out customers for a particular product based on some recorded data. Either the customers will churn out or they will not. So, the labels for this would be “Yes” and “No.”

Regression: It is the process of creating a model for distinguishing data into continuous real values, instead of using classes or discrete values. It can also identify the distribution movement depending on historical data. It is used for predicting the occurrence of an event depending on the degree of association of variables.

For example, the prediction of weather conditions depends on factors such as temperature, air pressure, solar radiation, elevation, and distance from the sea. The relation among these factors assists in predicting the weather condition.

18. How is the suitability of a Machine Learning Algorithm determined for a particular problem?

To identify a Machine Learning Algorithm for a particular problem, the following steps should be followed:

Step 1: Problem classification: Classification of the problem depends on the classification of input and output:

- Classifying the input: Classification of the input depends on whether there is data labeled (supervised learning) or unlabeled (unsupervised learning), or whether a model has to be created that interacts with the environment and improves itself (reinforcement learning.)
- Classifying the output: If the output of a model is required as a class, then some classification techniques need to be used.

If the output is a number, then regression techniques must be used; if the output is a different cluster of inputs, then clustering techniques should be used.

Step 2: Checking the algorithms in hand: After classifying the problem, the available algorithms that can be deployed for solving the classified problem should be considered.

Step 3: Implementing the algorithms: If there are multiple algorithms available, then all of them are to be implemented. Finally, the algorithm that gives the best performance is selected.

19. What is the Variance Inflation Factor?

Variance inflation factor (VIF) is the estimate of the volume of multicollinearity in a collection of many regression variables.

VIF = Variance of the model / Variance of the model with a single independent variable

This ratio has to be calculated for every independent variable. If VIF is high, then it shows the high collinearity of the independent variables.

20. What is a Confusion Matrix?

Confusion matrix is used to explain a model's performance and gives a summary of predictions of the classification problems. It assists in identifying the uncertainty between classes.

Confusion matrix gives the count of correct and incorrect values and error types. Accuracy of the model:

For example, consider the following confusion matrix. It consists of values as true positive, true negative, false positive, and false negative for a classification model. Now, the accuracy of the model can be calculated as follows:

So, in the example:

```
Accuracy = (200 + 50) / (200 + 50 + 10 + 60) = 0.78
```

This means that the model’s accuracy is 0.78, corresponding to its True Positive, True Negative, False Positive, and False Negative values.

21. What are Type I and Type II Errors?

Type I Error: Type I Error, false positive, is an error where the outcome of a test shows the nonacceptance of a true condition.

For example, suppose a person gets diagnosed with depression even when they are not suffering from the same, it is a case of false positive.

Type II Error: Type II Error, false negative, is an error where the outcome of a test shows the acceptance of a false condition.

For example, the CT scan of a person shows that they do not have a disease but in fact they do have the disease. Here, the test accepts the false condition that the person does not have the disease. This is a case of false negative.

22. When should Classification be used over Regression?

Both classification and regression are associated with prediction. Classification involves the identification of values or entities that lie in a specific group. Regression entails predicting a response value from consecutive sets of outcomes.

Classification is chosen over regression when the output of the model needs to yield the belongingness of data points in a dataset to a particular category.

For example, If you want to predict the price of a house, you should use regression since it is a numerical variable. However, if you are trying to predict whether a house situated in a particular area is going to be high-, medium-, or low-priced, then a classification model should be used.

23. Explain Logistic Regression

Logistic regression is the proper regression analysis used when the dependent variable is categorical or binary. Like all regression analyses, logistic regression is a technique for predictive analysis. Logistic regression is used to explain data and the relationship between one dependent binary variable and one or more independent variables. Logistic regression is also employed to predict the probability of categorical dependent variables.

Logistic regression can be used in the following scenarios:

- To predict whether a citizen is a Senior Citizen (1) or not (0)
- To check whether a person has a disease (Yes) or not (No)

There are three types of logistic regression:

- Binary logistic regression: In this type of logistic regression, there are only two outcomes possible.

Example: To predict whether it will rain (1) or not (0)

- Multinomial logistic regression: In this type of logistic regression, the output consists of three or more unordered categories.

Example: Predicting whether the prize of the house is high, medium, or low.

- Ordinal logistic regression: In this type of logistic regression, the output consists of three or more ordered categories.

Example: Rating an Android application from one to five stars.

Interested in learning Machine Learning? Enroll in this Machine Learning Training in Bangalore!

24. How to handle Missing or Corrupted Data in a Dataset?

In Python pandas, there are two methods to locate lost or corrupted data and discard those values:

- isNull(): It can be used for detecting the missing values.
- dropna(): It can be used for removing columns or rows with null values.

fillna() can be used to fill the void values with placeholder values.

25. Why is rotation required in PCA? What will happen if the components are not rotated?

Rotation is a significant step in principal component analysis (PCA.) Rotation maximizes the separation within the variance obtained by the components. This makes the interpretation of the components easier.

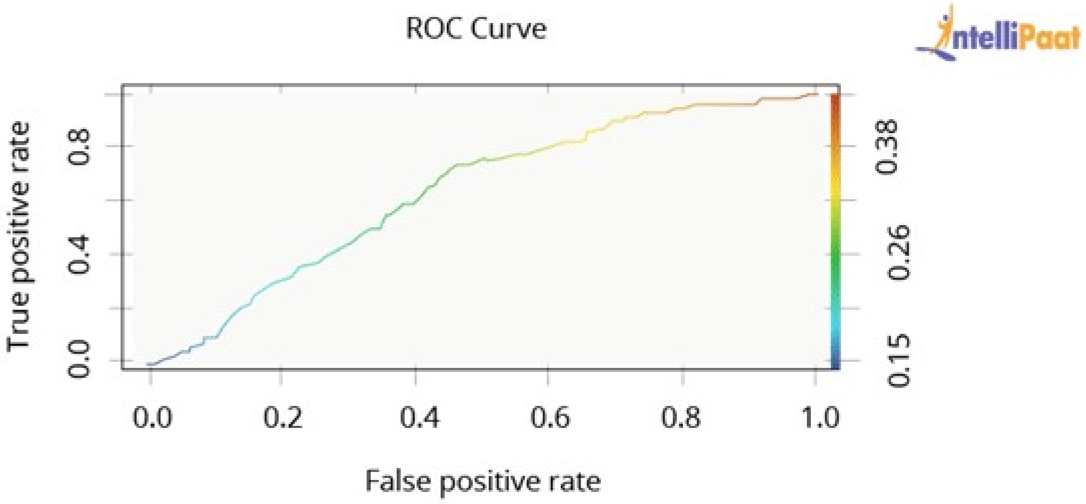
The motive behind conducting PCA is to choose fewer components that can explain the greatest variance in a dataset. When rotation is performed, the original coordinates of the points get changed. However, there is no change in the relative position of the components.

If the components are not rotated, then there needs to be more extended components to describe the variance.

26. What is ROC Curve and what does it represent?

ROC stands for receiver operating characteristic. ROC Curve is used to graphically represent the trade-off between true and false-positive rates.

In ROC, the area under the curve (AUC) gives an idea about the accuracy of the model.



The above graph shows a ROC curve. The greater the AUC, the better the performance of the model.

Next, we will be taking a look at Machine Learning interview questions on rescaling, binarizing, and standardizing.

27. Why are Validation and Test Datasets Needed?

Data is split into three different categories while creating a model:

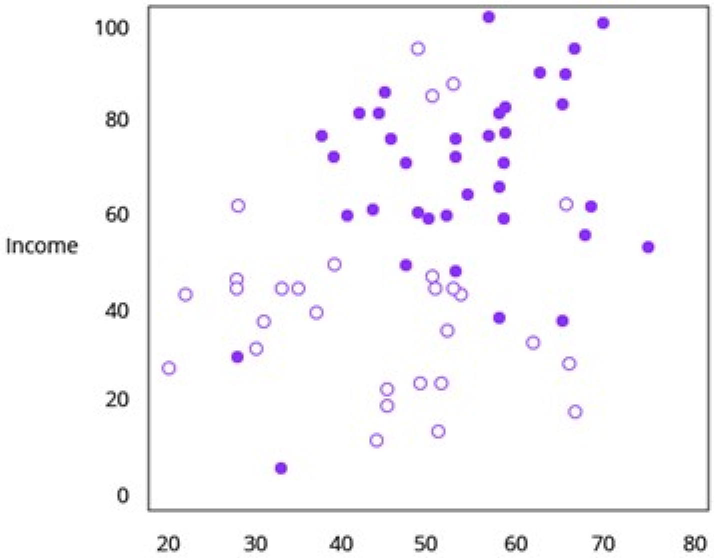
- **Training dataset:** Training dataset is used for building a model and adjusting its variables. The correctness of the model built on the training dataset cannot be relied on as the model might give incorrect outputs after being fed new inputs.
- **Validation dataset:** Validation dataset is used to look into a model's response. After this, the hyperparameters on the basis of the estimated benchmark of the validation dataset data are tuned.When a model's response is evaluated by using the validation dataset, the model is indirectly trained with the validation set. This may lead to the overfitting of the model to specific data. So, this model will not be strong enough to give the desired response to real-world data.
- **Test dataset:** Test dataset is the subset of the actual dataset, which is not yet used to train the model. The model is unaware of this dataset. So, by using the test dataset, the response of the created model can be computed on hidden data. The model's performance is tested on the basis of the test dataset.Note:

The model is always exposed to the test dataset after tuning the hyperparameters on top of the validation dataset.

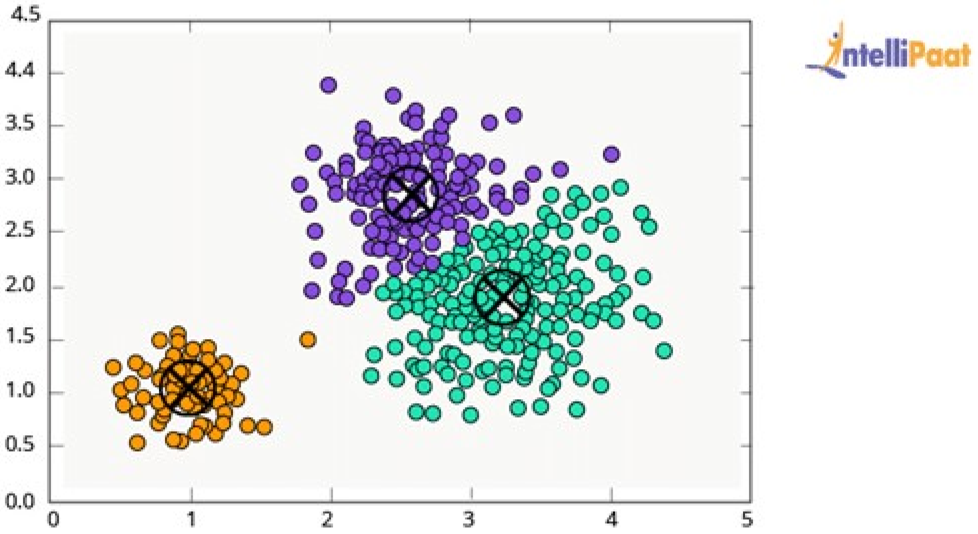
As we know, the evaluation of the model on the basis of the validation dataset would not be enough. Thus, the test dataset is used for computing the efficiency of the model.

28. Explain the difference between KNN and K-means Clustering

K-nearest neighbors (KNN): It is a supervised Machine Learning algorithm. In KNN, identified or labeled data is given to the model. The model then matches the points based on the distance from the closest points.



K-means clustering: It is an unsupervised Machine Learning algorithm. In K-means clustering, unidentified or unlabeled data is given to the model. The algorithm then creates batches of points based on the average of the distances between distinct points.



29. What is Dimensionality Reduction?

In the real world, Machine Learning models are built on top of features and parameters. These features can be multidimensional and large in number. Sometimes, the features may be irrelevant and it becomes a difficult task to visualize them.

This is where dimensionality reduction is used to cut down irrelevant and redundant features with the help of principal variables. These principal variables conserve the features, and are a subgroup, of the parent variables.

30. Both being Tree-based Algorithms, how is Random Forest different from Gradient Boosting Machine (GBM)?

The main difference between a random forest and GBM is the use of techniques. Random forest advances predictions using a technique called bagging. On the other hand, GBM advances predictions with the help of a technique called boosting.

- **Bagging:** In bagging, we apply arbitrary sampling and we divide the dataset into *N.* After that, we build a model by employing a single training algorithm.

Following that, we combine the final predictions by polling. Bagging helps to increase the efficiency of a model by decreasing the variance to eschew overfitting.

- **Boosting:** In boosting, the algorithm tries to review and correct the inadmissible predictions at the initial iteration. After that, the algorithm's sequence of iterations for correction continues until we get the desired prediction. Boosting assists in reducing bias and variance for strengthening the weak learners.

31. What is meant by Parametric and Non-parametric Models?

Parametric models refer to the models having a limited number of parameters. In case of parametric models, only the parameter of a model is needed to be known to make predictions regarding the new data.

Non-parametric models do not have any restrictions on the number of parameters, which makes new data predictions more flexible. In case of non-parametric models, the knowledge of model parameters and the state of the data needs to be known to make predictions.

32. Differentiate between Sigmoid and Softmax Functions

Sigmoid and Softmax functions differ based on their usage in Machine Learning task classification. Sigmoid function is used in the case of binary classification, while Softmax function is used in case of multi-classification.

33. In Machine Learning, for how many classes can Logistic Regression be used?

Logistic regression cannot be used for more than two classes. Logistic regression is, by default, a binary classifier. However, in cases where multi-class classification problems need to be solved, the default number of classes can be extended, i.e., multinomial logistic regression.

34. What do you understand about the P-value?

P-value is used in decision-making while testing a hypothesis. The null hypothesis is rejected at the minimum significance level of the P-value. A lower P-value indicates that the null hypothesis is to be rejected.

35. What is meant by Correlation and Covariance?

Correlation is a mathematical concept used in statistics and probability theory to measure, estimate, and compare data samples taken from different populations. In simpler terms, correlation helps in establishing a quantitative relationship between two variables.

Covariance is also a mathematical concept; it is a simpler way to arrive at a correlation between two variables. Covariance basically helps in determining what change or affect does one variable has on another.

36. What are the Various Tests for Checking the Normality of a Dataset?

In Machine Learning, checking the normality of a dataset is very important. Hence, certain tests are performed on a dataset to check its normality. Some of them are:

- D’Agostino Skewness Test
- Shapiro-Wilk Test
- Anderson-Darling Test
- Jarque-Bera Test
- Kolmogorov-Smirnov Test

37. What are the Two Main Types of Filtering in Machine Learning? Explain.

The two types of filtering are:

- Collaborative filtering
- Content-based filtering

Collaborative filtering refers to a recommender system where the interests of the individual user are matched with preferences of multiple users to predict new content.

Content-based filtering is a recommender system where the focus is only on the preferences of the individual user and not on multiple users.

38. Outlier Values can be Discovered from which Tools?

The various tools that can be used to discover outlier values are scatterplots, boxplots, Z-score, etc.

39. What is meant by Ensemble Learning?

Ensemble learning refers to the combination of multiple Machine Learning models to create more powerful models. The primary techniques involved in ensemble learning are bagging and boosting.

Watch this complete course video on Machine Learning Interview Questions

40. What are the Various Kernels that are present in SVM?

The various kernels that are present in SVM are:

- Linear
- Polynomial
- Radial Basis
- Sigmoid

Advanced Machine Learning interview questions for experienced

41. Suppose you found that your model is suffering from high variance. Which algorithm do you think could handle this situation and why?

Handling High Variance

- For handling issues of high variance, we should use the bagging algorithm.
- The bagging algorithm would split data into subgroups with a replicated sampling of random data.
- Once the algorithm splits the data, we can use random data to create rules using a particular training algorithm.
- After that, we can use polling for combining the predictions of the model.

42. What is Rescaling of Data and how is it done?

In real-world scenarios, the attributes present in data are in a varying pattern. So, rescaling the characteristics to a common scale is beneficial for algorithms to process data efficiently.

We can rescale data using Scikit-learn. The code for rescaling the data using MinMaxScaler is as follows:

```
#Rescaling data
import pandas
import scipy
import numpy
from sklearn.preprocessing import MinMaxScaler
```

```
names = ['Abhi', 'Piyush', 'Pranay', 'Sourav', 'Sid', 'Mike',
'pedi', 'Jack', 'Tim']
Dataframe = pandas.read_csv(url, names=names)
Array = dataframe.values
# Splitting the array into input and output
X = array[:,0:8]
Y = array[:,8]
Scaler = MinMaxScaler(feature_range=(0, 1))
rescaledX = scaler.fit_transform(X)
# Summarizing the modified data
numpy.set_printoptions(precision=3)
print(rescaledX[0:5,:])
```

Apart from the theoretical concepts, some interviewers also focus on the implementation of Machine Learning topics. The following Interview Questions are related to the implementation of theoretical concepts.

43. What is Binarizing of Data? How to Binarize?

Converting data into binary values on the basis of threshold values is known as binarizing of data. The values that are less than the threshold are set to 0 and the values that are greater than the threshold are set to 1. This process is useful when feature engineering has to be performed. This can also be used for adding unique features. Data can be binarized using Scikit-learn. The code for binarizing data using Binarizer is as follows:

```
from sklearn.preprocessing import Binarizer
import pandas
import numpy
names = ['Abhi', 'Piyush', 'Pranay', 'Sourav', 'Sid', 'Mike',
'pedi', 'Jack', 'Tim']
dataframe = pandas.read_csv(url, names=names)
array = dataframe.values
# Splitting the array into input and output
X = array[:,0:8]
Y = array[:,8]
binarizer = Binarizer(threshold=0.0).fit(X)
binaryX = binarizer.transform(X)
# Summarizing the modified data
```

```
numpy.set_printoptions(precision=3)
print(binaryX[0:5,:])
```

44. How to Standardize Data?

Standardization is the method that is used for rescaling data attributes. The attributes are likely to have a mean value of 0 and a value of the standard deviation of 1. The main objective of standardization is to prompt the mean and standard deviation for the attributes.

Data can be standardized using Scikit-learn. The code for standardizing the data using StandardScaler is as follows:

```
# Python code to Standardize data (0 mean, 1 stdev)
from sklearn.preprocessing import StandardScaler
import pandas
import numpy
names = ['Abhi', 'Piyush', 'Pranay', 'Sourav', 'Sid', 'Mike',
'pedi', 'Jack', 'Tim']
dataframe = pandas.read_csv(url, names=names)
array = dataframe.values
# Separate the array into input and output components
X = array[:,0:8]
Y = array[:,8]
scaler = StandardScaler().fit(X)
rescaledX = scaler.transform(X)
# Summarize the transformed data
numpy.set_printoptions(precision=3)
print(rescaledX[0:5,:])
```

45. We know that one-hot encoding increases the dimensionality of a dataset, but label encoding doesn’t. How?

When one-hot encoding is used, there is an increase in the dimensionality of a dataset. The reason for the increase in dimensionality is that every class in categorical variables, forms a different variable.

Example: Suppose there is a variable “Color.” It has three sublevels, “Yellow,” “Purple,” and “Orange.” So, one-hot encoding “Color” will create three different variables as Color.Yellow, Color.Purple, and Color.Orange.

In label encoding, the subclasses of a certain variable get the value 0 and 1. So, label encoding is only used for binary variables.

This is why one-hot encoding increases the dimensionality of data and label encoding does not.

Now, if you are interested in doing an end-to-end certification course in Machine Learning, you can check out Intellipaat’s Machine Learning Course with Python.

46. Executing a binary classification tree algorithm is a simple task. But how does tree splitting take place? How does the tree determine which variable to break at the root node and which at its child nodes?

Gini index and Node Entropy assist the binary classification tree to make decisions. Basically, the tree algorithm determines the feasible feature that is used to distribute data into the most genuine child nodes.

According to the Gini index, if we arbitrarily pick a pair of objects from a group, then they should be of identical class and the probability for this event should be 1.

The following are the steps to compute the Gini index:

1. Compute Gini for sub-nodes with the formula: The sum of the square of probability for success and failure (p^2 + q^2)
2. Compute Gini for split by weighted Gini rate of every node of the split

Now, Entropy is the degree of indecency that is given by the following:

Where *a* and *b* are the probabilities of success and failure of the node

When Entropy = 0, the node is homogenous

When Entropy is high, both groups are present at 50–50 percent in the node.

Finally, to determine the suitability of the node as a root node, the entropy should be very low.

47. Imagine you are given a dataset consisting of variables having more than 30% missing values. Let’s say, out of 50 variables, 16 variables have missing values, which is higher than 30%. How will you deal with them?

To deal with the missing values, we will do the following:

- We will specify a different class for the missing values.
- Now, we will check the distribution of values, and we will hold those missing values that are defining a pattern.
- Then, we will charge these values into yet another class while eliminating others.

48. Explain False Negative, False Positive, True Negative, and True Positive with a simple example.

True Positive (TP): When the Machine Learning model correctly predicts the condition, it is said to have a True Positive value.

True Negative (TN): When the Machine Learning model correctly predicts the negative condition or class, then it is said to have a True Negative value.

False Positive (FP): When the Machine Learning model incorrectly predicts a negative class or condition, then it is said to have a False Positive value.

False Negative (FN): When the Machine Learning model incorrectly predicts a positive class or condition, then it is said to have a False Negative value.

49. What is F1-score and How Is It Used?

F-score or F1-score is a measure of overall accuracy of a binary classification model. Before understanding F1-score, it is crucial to understand two more measures of accuracy, i.e., precision and recall.

Precision is defined as the percentage of True Positives to the total number of positive classifications predicted by the model. In other words,

Precision = (No. of True Positives / No. True Positives + No. of False Positives)

Recall is defined as the percentage of True Positives to the total number of actual positive labeled data passed to the model. In other words,

Precision = (No. of True Positives / No. True Positives + No. of False Negatives)

Both precision and recall are partial measures of accuracy of a model. F1-score combines precision and recall and provides an overall score to measure a model's accuracy.

F1-score = 2 × (Precision × Recall) / (Precision + Recall)

This is why, F1-score is the most popular measure of accuracy in any Machine-Learning-based binary classification model.

50. How to Implement the KNN Classification Algorithm?

Iris dataset is used for implementing the KNN classification algorithm.

```
# KNN classification algorithm
from sklearn.datasets import load_iris
from sklearn.neighbors import KNeighborsClassifier
import numpy as np
from sklearn.model_selection import train_test_split
iris_dataset=load_iris()
A_train, A_test, B_train, B_test =
ztrain_test_split(iris_dataset["data"],
iris_dataset["target"], random_state=0)
kn = KNeighborsClassifier(n_neighbors=1)
kn.fit(A_train, B_train)
A_new = np.array([[8, 2.5, 1, 1.2]])
```

```
prediction = kn.predict(A_new)
print("Predicted target value: {}\n".format(prediction))
print("Predicted feature name: {}\n".format
(iris_dataset["target_names"][prediction]))
print("Test score: {:.2f}".format(kn.score(A_test, B_test)))
Output:
Predicted Target Name: [0]
Predicted Feature Name: [‘ Setosa’]
Test Score: 0.92
```