

MACHINE LEARNING INTERVIEW QUESTIONS :

1) What do you understand by Machine learning?

Machine learning is the form of Artificial Intelligence that deals with system programming and automates data analysis to enable computers to learn and act through experiences without being explicitly programmed.

For example, Robots are coded in such a way that they can perform the tasks based on data they collect from sensors. They automatically learn programs from data and improve with experiences.

2) Differentiate between inductive learning and deductive learning?

In inductive learning, the model learns by examples from a set of observed instances to draw a generalized conclusion. On the other side, in deductive learning, the model first applies the conclusion, and then the conclusion is drawn.

- Inductive learning is the method of using observations to draw conclusions.
- Deductive learning is the method of using conclusions to form observations.

For example, if we have to explain to a kid that playing with fire can cause burns. There are two ways we can explain this to a kid; we can show training examples of various fire accidents or images of burnt people and label them as "Hazardous". In this case, a kid will understand with the help of examples and not play with the fire. It is the form of Inductive machine learning. The other way to teach the same thing is to let the kid play with the fire and wait to see what happens. If the kid gets a burn, it will teach the kid not to play with fire and avoid going near it. It is the form of deductive learning.

3) What is the difference between Data Mining and Machine Learning?

Data mining can be described as the process in which the structured data tries to abstract knowledge or interesting unknown patterns. During this process, machine learning algorithms are used.

Machine learning represents the study, design, and development of the algorithms which provide the ability to the processors to learn without being explicitly programmed.

4) What is the meaning of Overfitting in Machine learning?

Overfitting can be seen in machine learning when a statistical model describes random error or noise instead of the underlying relationship. Overfitting is usually observed when a model is excessively complex. It happens because of having too many parameters concerning the number of training data types. The model displays poor performance, which has been overfitted.

5) Why overfitting occurs?

The possibility of overfitting occurs when the criteria used for training the model is not as per the criteria used to judge the efficiency of a model.

6) What is the method to avoid overfitting?

Overfitting occurs when we have a small dataset, and a model is trying to learn from it. By using a large amount of data, overfitting can be avoided. But if we have a small database and are forced to build a model based on that, then we can use a technique known as **cross-validation**. In this method, a model is usually given a dataset of a known data on which training data set is run and dataset of unknown data against which the model is tested. The primary aim

of cross-validation is to define a dataset to "test" the model in the training phase. If there is sufficient data, **'Isotonic Regression'** is used to prevent overfitting.

7) Differentiate supervised and unsupervised machine learning.

- In supervised machine learning, the machine is trained using labeled data. Then a new dataset is given into the learning model so that the algorithm provides a positive outcome by analyzing the labeled data. For example, we first require to label the data which is necessary to train the model while performing classification.
- In the unsupervised machine learning, the machine is not trained using labeled data and let the algorithms make the decisions without any corresponding output variables.

8) How does Machine Learning differ from Deep Learning?

- Machine learning is all about algorithms which are used to parse data, learn from that data, and then apply whatever they have learned to make informed decisions.
- Deep learning is a part of machine learning, which is inspired by the structure of the human brain and is particularly useful in feature detection.

9) How is KNN different from k-means?

KNN or K nearest neighbors is a supervised algorithm which is used for classification purpose. In KNN, a test sample is given as the class of the majority of its nearest neighbors. On the other side, K-means is an unsupervised algorithm which is mainly used for clustering. In k-means clustering, it needs a set of unlabeled points and a threshold only. The algorithm further takes unlabeled data and learns how to cluster it into groups by computing the mean of the distance between different unlabeled points.

10) What are the different types of Algorithm methods in Machine Learning?

The different types of algorithm methods in machine learning are:

- Supervised Learning
- Semi-supervised Learning
- Unsupervised Learning
- Transduction
- Reinforcement Learning

11) What do you understand by Reinforcement Learning technique?

Reinforcement learning is an algorithm technique used in Machine Learning. It involves an agent that interacts with its environment by producing actions & discovering errors or rewards. Reinforcement learning is employed by different software and machines to search for the best suitable behavior or path it should follow in a specific situation. It usually learns on the basis of reward or penalty given for every action it performs.

12) What is the trade-off between bias and variance?

Both bias and variance are errors. Bias is an error due to erroneous or overly simplistic assumptions in the learning algorithm. It can lead to the model under-fitting the data, making

it hard to have high predictive accuracy and generalize the knowledge from the training set to the test set.

Variance is an error due to too much complexity in the learning algorithm. It leads to the algorithm being highly sensitive to high degrees of variation in the training data, which can lead the model to overfit the data.

To optimally reduce the number of errors, we will need to tradeoff bias and variance.

13) How do classification and regression differ?

Classification	Regression
Classification is the task to predict a discrete class label.	Regression is the task to predict a continuous quantity.
In a classification problem, data is labeled into one of two or more classes.	A regression problem needs the prediction of a quantity.
A classification having problem with two classes is called binary classification, and more than two classes is called multi-class classification	A regression problem containing multiple input variables is called a multivariate regression problem.
Classifying an email as spam or non-spam is an example of a classification problem.	Predicting the price of a stock over a period of time is a regression problem.

14) What are the five popular algorithms we use in Machine Learning?

Five popular algorithms are:

- Decision Trees
- Probabilistic Networks
- Neural Networks
- Support Vector Machines
- Nearest Neighbor

15) What do you mean by ensemble learning?

Numerous models, such as classifiers are strategically made and combined to solve a specific computational program which is known as ensemble learning. The ensemble methods are also known as committee-based learning or learning multiple classifier systems. It trains various hypotheses to fix the same issue. One of the most suitable examples of ensemble modeling is the random forest trees where several decision trees are used to predict outcomes. It is used to improve the classification, function approximation, prediction, etc. of a model.

16) What is a model selection in Machine Learning?

The process of choosing models among diverse mathematical models, which are used to define the same data is known as **Model Selection**. Model learning is applied to the fields of **statistics, data mining, and machine learning**.

17) What are the three stages of building the hypotheses or model in machine learning?

There are three stages to build hypotheses or model in machine learning:

- **Model building**

It chooses a suitable algorithm for the model and trains it according to the requirement of the problem.

- **Applying the model**

It is responsible for checking the accuracy of the model through the test data.

- **Model testing**

It performs the required changes after testing and apply the final model.

18) What according to you, is the standard approach to supervised learning?

In supervised learning, the standard approach is to split the set of example into the training set and the test.

19) Describe 'Training set' and 'training Test'.

In various areas of information of machine learning, a set of data is used to discover the potentially predictive relationship, which is known as 'Training Set'. The training set is an example that is given to the learner. Besides, the 'Test set' is used to test the accuracy of the hypotheses generated by the learner. It is the set of instances held back from the learner. Thus, the training set is distinct from the test set.

20) What are the common ways to handle missing data in a dataset?

Missing data is one of the standard factors while working with data and handling. It is considered as one of the greatest challenges faced by the data analysts. There are many ways one can impute the missing values. Some of the common methods to handle missing data in datasets can be defined as **deleting the rows, replacing with mean/median/mode, predicting the missing values, assigning a unique category, using algorithms that support missing values**, etc.

21) What do you understand by ILP?

ILP stands for **Inductive Logic Programming**. It is a part of machine learning which uses logic programming. It aims at searching patterns in data which can be used to build predictive models. In this process, the logic programs are assumed as a hypothesis.

22) What are the necessary steps involved in Machine Learning Project?

There are several essential steps we must follow to achieve a good working model while doing a Machine Learning Project. Those steps may include **parameter tuning, data preparation, data collection, training the model, model evaluation, and prediction**, etc.

23) Describe Precision and Recall?

Precision and Recall both are the measures which are used in the information retrieval domain to measure how good an information retrieval system reclaims the related data as requested by the user.

Precision can be said as a positive predictive value. It is the fraction of relevant instances among the received instances.

On the other side, **recall** is the fraction of relevant instances that have been retrieved over the total amount or relevant instances. The recall is also known as **sensitivity**.

24) What do you understand by Decision Tree in Machine Learning?

Decision Trees can be defined as the Supervised Machine Learning, where the data is continuously split according to a certain parameter. It builds classification or regression models as similar as a tree structure, with datasets broken up into ever smaller subsets while developing the decision tree. The tree can be defined by two entities, namely **decision nodes**, and **leaves**. The leaves are the decisions or the outcomes, and the decision nodes are where the data is split. Decision trees can manage both categorical and numerical data.

25) What are the functions of Supervised Learning?

- Classification
- Speech Recognition
- Regression
- Predict Time Series
- Annotate Strings

26) What are the functions of Unsupervised Learning?

- Finding clusters of the data
- Finding low-dimensional representations of the data
- Finding interesting directions in data
- Finding novel observations/ database cleaning
- Finding interesting coordinates and correlations

27) What do you understand by algorithm independent machine learning?

Algorithm independent machine learning can be defined as machine learning, where mathematical foundations are independent of any particular classifier or learning algorithm.

28) Describe the classifier in machine learning.

A classifier is a case of a hypothesis or discrete-valued function which is used to assign class labels to particular data points. It is a system that inputs a vector of discrete or continuous feature values and outputs a single discrete value, the class.

29) What do you mean by Genetic Programming?

Genetic Programming (GP) is almost similar to an **Evolutionary Algorithm**, a subset of machine learning. Genetic programming software systems implement an algorithm that uses random mutation, a fitness function, crossover, and multiple generations of evolution to resolve a user-defined task. The genetic programming model is based on testing and choosing the best option among a set of results.

30) What is SVM in machine learning? What are the classification methods that SVM can handle?

SVM stands for **Support Vector Machine**. SVM are supervised learning models with an associated learning algorithm which analyze the data used for classification and regression analysis.

The classification methods that SVM can handle are:

- Combining binary classifiers
- Modifying binary to incorporate multiclass learning

31) How will you explain a linked list and an array?

An array is a datatype which is widely implemented as a default type, in almost all the modern programming languages. It is used to store data of a similar type.

But there are many use-cases where we don't know the quantity of data to be stored. For such cases, advanced data structures are required, and one such data structure is **linked list**.

There are some points which explain how the linked list is different from an array:

ARRAY	LINKED LIST
An array is a group of elements of a similar data type.	Linked List is an ordered group of elements of the same type, which are connected using pointers.
Elements are stored consecutively in the memory.	New elements can be stored anywhere in memory.
An Array supports Random Access . It means that the elements can be accessed directly using their index value, like arr[0] for 1st element, arr[5] for 6th element, etc. As a result, accessing elements in an array is fast with constant time complexity of O(1).	Linked List supports Sequential Access . It means that we have to traverse the complete linked list, up to that element sequentially which element/node we want to access in. To access the nth element of a linked list, the time

	complexity is $O(n)$.	
Memory is allocated at compile time as soon as the array is declared. It is known as Static Memory Allocation .	Memory is allocated at runtime , whenever a new node is added. It is known as Dynamic Memory Allocation .	
Insertion and Deletion operation takes more time in the array, as the memory locations are consecutive and fixed.	In case of a linked list, a new element is stored at the first free available. Thus, Insertion and Deletion operations are fast in the linked list.	n
Size of the array must be declared at the time of array declaration.	Size of a Linked list is variable. It grows at runtime whenever nodes are added to it.	

32) What do you understand by the Confusion Matrix?

A confusion matrix is a table which is used for summarizing the performance of a classification algorithm. It is also known as the **error matrix**.

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

Where,

TN= True Negative

TP= True Positive

FN= False Negative

FP= False Positive

33) Explain True Positive, True Negative, False Positive, and False Negative in Confusion Matrix with an example.

- **True Positive**

When a model correctly predicts the positive class, it is said to be a true positive.

For example, Umpire gives a Batsman NOT OUT when he is NOT OUT.

- **True Negative**

When a model correctly predicts the negative class, it is said to be a true negative.

For example, Umpire gives a Batsman OUT when he is OUT.

- **False Positive**

When a model incorrectly predicts the positive class, it is said to be a false positive. It is also known as '**Type I**' error.

For example, Umpire gives a Batsman NOT OUT when he is OUT.

- **False Negative**

When a model incorrectly predicts the negative class, it is said to be a false negative. It is also known as '**Type II**' error.

For example, Umpire gives a Batsman OUT when he is NOT OUT.

34) What according to you, is more important between model accuracy and model performance?

Model accuracy is a subset of model performance. The accuracy of the model is directly proportional to the performance of the model. Thus, better the performance of the model, more accurate are the predictions.

35) What is Bagging and Boosting?

- Bagging is a process in ensemble learning which is used for improving unstable estimation or classification schemes.
- Boosting methods are used sequentially to reduce the bias of the combined model.

36) What are the similarities and differences between bagging and boosting in Machine Learning?

Similarities of Bagging and Boosting

- Both are the ensemble methods to get N learns from 1 learner.
- Both generate several training data sets with random sampling.
- Both generate the final result by taking the average of N learners.
- Both reduce variance and provide higher scalability.

Differences between Bagging and Boosting

- Although they are built independently, but for Bagging, Boosting tries to add new models which perform well where previous models fail.

- Only Boosting determines the weight for the data to tip the scales in favor of the most challenging cases.
- Only Boosting tries to reduce bias. Instead, Bagging may solve the problem of over-fitting while boosting can increase it.

37) What do you understand by Cluster Sampling?

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

For example, if we are clustering the total number of managers in a set of companies, in that case, managers (sample) will represent elements and companies will represent clusters.

38) What do you know about Bayesian Networks?

Bayesian Networks also referred to as '**belief networks**' or '**casual networks**', are used to represent the graphical model for probability relationship among a set of variables.

For example, a Bayesian network can be used to represent the probabilistic relationships between diseases and symptoms. As per the symptoms, the network can also compute the probabilities of the presence of various diseases.

Efficient algorithms can perform inference or learning in Bayesian networks. Bayesian networks which relate the variables (e.g., speech signals or protein sequences) are called dynamic Bayesian networks.

39) Which are the two components of Bayesian logic program?

A Bayesian logic program consists of two components:

- **Logical**
It contains a set of Bayesian Clauses, which capture the qualitative structure of the domain.
- **Quantitative**
It is used to encode quantitative information about the domain.

40) Describe dimension reduction in machine learning.

Dimension reduction is the process which is used to reduce the number of random variables under considerations.

Dimension reduction can be divided into feature selection and extraction.

41) Why instance-based learning algorithm sometimes referred to as Lazy learning algorithm?

In machine learning, **lazy learning** can be described as a method where induction and generalization processes are delayed until classification is performed. Because of the same property, an instance-based learning algorithm is sometimes called lazy learning algorithm.

42) What do you understand by the F1 score?

The F1 score represents the measurement of a model's performance. It is referred to as a weighted average of the precision and recall of a model. The results tending to **1** are considered as the best, and those tending to **0** are the worst. It could be used in classification tests, where true negatives don't matter much.

43) How is a decision tree pruned?

Pruning is said to occur in decision trees when the branches which may consist of weak predictive power are removed to reduce the complexity of the model and increase the predictive accuracy of a decision tree model. Pruning can occur bottom-up and top-down, with approaches such as **reduced error pruning** and **cost complexity pruning**.

Reduced error pruning is the simplest version, and it replaces each node. If it is unable to decrease predictive accuracy, one should keep it pruned. But, it usually comes pretty close to an approach that would optimize for maximum accuracy.

44) What are the Recommended Systems?

Recommended System is a sub-directory of information filtering systems. It predicts the preferences or rankings offered by a user to a product. According to the preferences, it provides similar recommendations to a user. Recommendation systems are widely used in **movies, news, research articles, products, social tips, music**, etc.

45) What do you understand by Underfitting?

Underfitting is an issue when we have a low error in both the training set and the testing set. Few algorithms work better for interpretations but fail for better predictions.

46) When does regularization become necessary in Machine Learning?

Regularization is necessary whenever the model begins to overfit/ underfit. It is a cost term for bringing in more features with the objective function. Hence, it tries to push the coefficients for many variables to zero and reduce cost term. It helps to reduce model complexity so that the model can become better at predicting (generalizing).

47) What is Regularization? What kind of problems does regularization solve?

A regularization is a form of regression, which constrains/ regularizes or shrinks the coefficient estimates towards zero. In other words, it discourages learning a more complex or flexible model to avoid the risk of overfitting. It reduces the variance of the model, without a substantial increase in its bias.

Regularization is used to address overfitting problems as it penalizes the loss function by adding a multiple of an L1 (LASSO) or an L2 (Ridge) norm of weights vector w .

48) Why do we need to convert categorical variables into factor? Which functions are used to perform the conversion?

Most Machine learning algorithms require number as input. That is why we convert categorical values into factors to get numerical values. We also don't have to deal with dummy variables.

The functions **factor()** and **as.factor()** are used to convert variables into factors.

49) Do you think that treating a categorical variable as a continuous variable would result in a better predictive model?

For a better predictive model, the categorical variable can be considered as a continuous variable only when the variable is ordinal in nature.

50) How is machine learning used in day-to-day life?

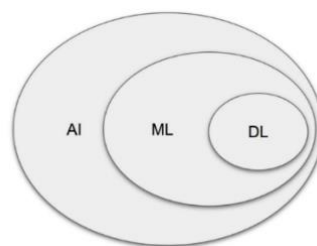
Most of the people are already using machine learning in their everyday life. Assume that you are engaging with the internet, you are actually expressing your preferences, likes, dislikes through your searches. All these things are picked up by cookies coming on your computer,

from this, the behavior of a user is evaluated. It helps to increase the progress of a user through the internet and provide similar suggestions.

The navigation system can also be considered as one of the examples where we are using machine learning to calculate a distance between two places using optimization techniques. Surely, people are going to more engage with machine learning in the near future.

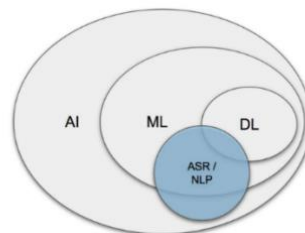
1. Explain the terms Artificial Intelligence (AI), Machine Learning (ML) and Deep Learning?

Artificial Intelligence (AI) is the domain of producing intelligent machines. ML refers to systems that can assimilate from experience (training data) and Deep Learning (DL) states to systems that learn from experience on large data sets. ML can be considered as a subset of AI. **Deep Learning (DL)** is ML but useful to large data sets. The figure below roughly encapsulates the relation between AI, ML, and DL:



In summary, DL is a subset of ML & both were the subsets of AI.

Additional Information: ASR (**Automatic Speech Recognition**) & **NLP** (Natural Language Processing) fall under AI and overlay with ML & DL as ML is often utilized for NLP and ASR tasks.



2. What are the different types of Learning/ Training models in ML?

ML algorithms can be primarily classified depending on the presence/absence of target variables.

A. Supervised learning: [Target is present]

The machine learns using labelled data. The model is trained on an existing data set before it starts making decisions with the new data.

The target variable is continuous: Linear Regression, polynomial Regression, and quadratic Regression.

The target variable is categorical: Logistic regression, Naive Bayes, KNN, SVM, Decision Tree, **Gradient Boosting**, ADA boosting, Bagging, **Random forest** etc.

B. Unsupervised learning: [Target is absent]

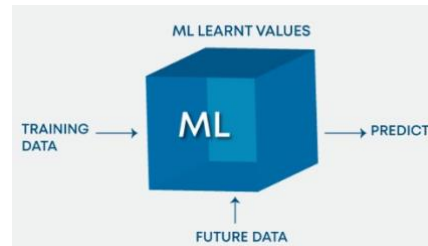
The machine is trained on unlabelled data and without any proper guidance. It automatically infers patterns and relationships in the data by creating clusters. The model

learns through observations and deduced structures in the data.
Principal component Analysis, Factor analysis, Singular Value Decomposition etc.

C. Reinforcement Learning:

The model learns through a trial and error method. This kind of learning involves an agent that will interact with the environment to create actions and then discover errors or rewards of that action.

3. What is the difference between deep learning and machine learning?



Machine Learning involves algorithms that learn from patterns of data and then apply it to decision making. Deep Learning, on the other hand, is able to learn through processing data on its own and is quite similar to the human brain where it identifies something, analyse it, and makes a decision.

The key differences are as follows:

- The manner in which data is presented to the system.
- Machine learning algorithms always require structured data and deep learning networks rely on layers of [artificial neural networks](#).

4. What is the main key difference between supervised and unsupervised machine learning?

Supervised learning	Unsupervised learning
The supervised learning technique needs labelled data to train the model. For example, to solve a classification problem (a supervised learning task), you need to have label data to train the model and to classify the data into your labelled groups.	Unsupervised learning does not need any labelled dataset. This is the main key difference between supervised learning and unsupervised learning.

5. How do you select important variables while working on a data set?

There are various means to select important variables from a data set that include the following:

- Identify and discard correlated variables before finalizing on important variables
- The variables could be selected based on 'p' values from Linear Regression
- Forward, Backward, and Stepwise selection
- [Lasso Regression](#)
- Random Forest and plot variable chart

- Top features can be selected based on information gain for the available set of features.

6. There are many machine learning algorithms till now. If given a data set, how can one determine which algorithm to be used for that?

Machine Learning algorithm to be used purely depends on the type of data in a given dataset. If data is linear then, we use linear regression. If data shows non-linearity then, the bagging algorithm would do better. If the data is to be analyzed/interpreted for some business purposes then we can use [decision trees](#) or SVM. If the dataset consists of images, videos, audios then, neural networks would be helpful to get the solution accurately.

So, there is no certain metric to decide which algorithm to be used for a given situation or a data set. We need to explore the data using EDA ([Exploratory Data Analysis](#)) and understand the purpose of using the dataset to come up with the best fit algorithm. So, it is important to study all the algorithms in detail.

8. 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.

9. We look at machine learning software almost all the time. How do we apply

Machine Learning to Hardware?

We have to build ML algorithms in System Verilog which is a Hardware development Language and then program it onto an FPGA to apply Machine Learning to hardware.

10. Explain One-hot encoding and Label Encoding. How do they affect the dimensionality of the given dataset?

One-hot encoding is the representation of categorical variables as binary vectors. Label Encoding is converting labels/words into numeric form. Using one-hot encoding increases the dimensionality of the data set. Label encoding doesn't affect the dimensionality of the data set. One-hot encoding creates a new variable for each level in the variable whereas, in Label encoding, the levels of a variable get encoded as 1 and 0.

greatlearning
Learning for Life

LABEL ENCODING

Food Name	Categorical#	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

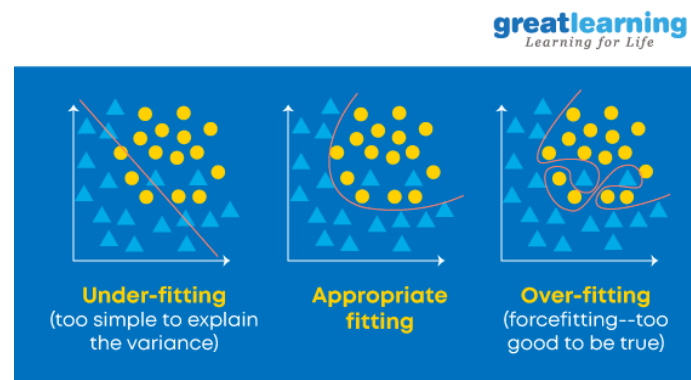
▼

ONE HOT ENCODING

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

11. When does regularization come into play in Machine Learning?

At times when the model begins to underfit or overfit, regularization becomes necessary. It is a regression that diverts or regularizes the coefficient estimates towards zero. It reduces flexibility and discourages learning in a model to avoid the risk of overfitting. The model complexity is reduced and it becomes better at predicting.



12. What is Bias, Variance and what do you mean by Bias-Variance Tradeoff?

Both are errors in [Machine Learning Algorithms](#). When the algorithm has limited flexibility to deduce the correct observation from the dataset, it results in bias. On the other hand, variance occurs when the model is extremely sensitive to small fluctuations.

If one adds more features while building a model, it will add more complexity and we will lose bias but gain some variance. In order to maintain the optimal amount of error, we perform a tradeoff between bias and variance based on the needs of a business.

Source: Understanding the Bias-Variance Tradeoff: Scott Fortmann – Roe

Bias stands for the error because of the erroneous or overly simplistic assumptions in the learning algorithm. This assumption can lead to the model underfitting the data, making

it hard for it to have high predictive accuracy and for you to generalize your knowledge from the training set to the test set.

Variance is also an error because of too much complexity in the learning algorithm. This can be the reason for the algorithm being highly sensitive to high degrees of variation in training data, which can lead your model to overfit the data. Carrying too much noise from the training data for your model to be very useful for your test data.

The bias-variance decomposition essentially decomposes the learning error from any algorithm by adding the bias, the variance and a bit of irreducible error due to noise in the underlying dataset. Essentially, if you make the model more complex and add more variables, you'll lose bias but gain some variance — in order to get the optimally reduced amount of error, you'll have to trade off bias and variance. You don't want either high bias or high variance in your model.

13. How can we relate standard deviation and variance?

Standard deviation refers to the spread of your data from the mean. Variance is the average degree to which each point differs from the mean i.e. the average of all data points. We can relate Standard deviation and Variance because it is the square root of Variance.

14. A data set is given to you and it has missing values which spread along 1

standard deviation from the mean. How much of the data would remain untouched?

It is given that the data is spread across mean that is the data is spread across an average. So, we can presume that it is a normal distribution. In a normal distribution, about 68% of data lies in 1 standard deviation from averages like mean, mode or median. That means about 32% of the data remains uninfluenced by missing values.

15. Is a high variance in data good or bad?

Higher variance directly means that the data spread is big and the feature has a variety of data. Usually, high variance in a feature is seen as not so good quality.

16. If your dataset is suffering from high variance, how would you handle it?

For datasets with high variance, we could use the bagging algorithm to handle it. Bagging algorithm splits the data into subgroups with sampling replicated from random data. After the data is split, random data is used to create rules using a training algorithm. Then we use polling technique to combine all the predicted outcomes of the model.

17. A data set is given to you about utilities fraud detection. You have built

aclassifier model and achieved a performance score of 98.5%. Is this a good model?

If yes, justify. If not, what can you do about it?

Data set about utilities fraud detection is not balanced enough i.e. imbalanced. In such a data set, accuracy score cannot be the measure of performance as it may only be predict the majority class label correctly but in this case our point of interest is to predict the minority label. But often minorities are treated as noise and ignored. So, there is a high probability of misclassification of the minority label as compared to the majority label. For evaluating the model performance in case of imbalanced data sets, we should use Sensitivity (True Positive rate) or Specificity (True Negative rate) to determine class label

wise performance of the classification model. If the minority class label's performance is not so good, we could do the following:

- We can use under sampling or over sampling to balance the data.
- We can change the prediction threshold value.
- We can assign weights to labels such that the minority class labels get larger weights.
- We could detect anomalies.

18. Explain the handling of missing or corrupted values in the given dataset.

An easy way to handle missing values or corrupted values is to drop the corresponding rows or columns. If there are too many rows or columns to drop then we consider replacing the missing or corrupted values with some new value.

Identifying missing values and dropping the rows or columns can be done by using `IsNull()` and `dropna()` functions in Pandas. Also, the `Fillna()` function in Pandas replaces the incorrect values with the placeholder value.

19. What is Time series?

A **Time series** is a sequence of numerical data points in successive order. It tracks the movement of the chosen data points, over a specified period of time and records the data points at regular intervals. Time series doesn't require any minimum or maximum time input. Analysts often use Time series to examine data according to their specific requirement.

20. What is a Box-Cox transformation?

Box-Cox transformation is a power transform which transforms non-normal dependent variables into normal variables as normality is the most common assumption made while using many statistical techniques. It has a lambda parameter which when set to 0 implies that this transform is equivalent to log-transform. It is used for variance stabilization and also to normalize the distribution.

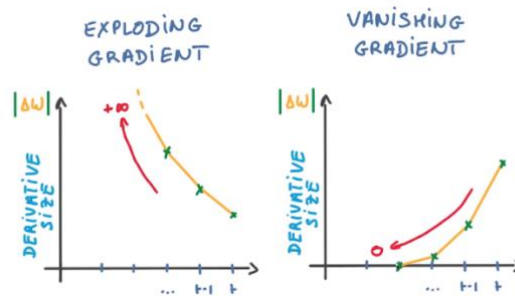
21. What is the difference between stochastic gradient descent (SGD) and gradient descent (GD)?

Gradient Descent and Stochastic Gradient Descent are the algorithms that find the set of parameters that will minimize a loss function.

The difference is that in Gradient Descent, all training samples are evaluated for each set of parameters. While in **Stochastic Gradient Descent** only one training sample is evaluated for the set of parameters identified.

22. What is the exploding gradient problem while using the back propagation technique?

When large error gradients accumulate and result in large changes in the neural network weights during training, it is called the exploding gradient problem. The values of weights can become so large as to overflow and result in NaN values. This makes the model unstable and the learning of the model to stall just like the **vanishing gradient problem**. This is one of the most commonly asked interview questions on machine learning.



23. Can you mention some advantages and disadvantages of decision trees?

The advantages of decision trees are that they are easier to interpret, are nonparametric and hence robust to outliers, and have relatively few parameters to tune. On the other hand, the disadvantage is that they are prone to overfitting.

24. Explain the differences between Random Forest and Gradient Boosting machines.

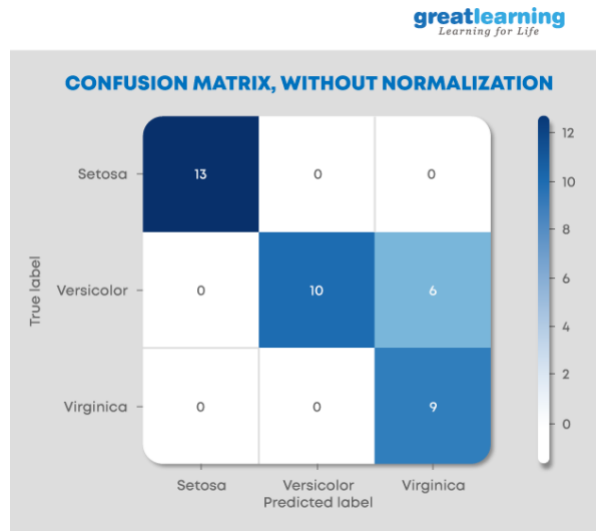
Random Forests	Gradient Boosting
Random forests are a significant number of decision trees pooled using averages or majority rules at the end.	Gradient boosting machines also combine decision trees but at the beginning of the process, unlike Random forests.
The random forest creates each tree independent of the others while gradient boosting develops one tree at a time.	Gradient boosting yields better outcomes than random forests if parameters are carefully tuned but it's not a good option if the data set contains a lot of outliers/anomalies/noise as it can result in overfitting of the model.
Random forests perform well for multiclass object detection.	Gradient Boosting performs well when there is data which is not balanced such as in real-time risk assessment.

25. What is a confusion matrix and why do you need it?

Confusion matrix (also called the error matrix) is a table that is frequently used to illustrate the performance of a classification model i.e. classifier on a set of test data for which the true values are well-known.

It allows us to visualize the performance of an algorithm/model. It allows us to easily identify the confusion between different classes. It is used as a performance measure of a model/algorithm.

A confusion matrix is known as a summary of predictions on a classification model. The number of right and wrong predictions were summarized with count values and broken down by each class label. It gives us information about the errors made through the classifier and also the types of errors made by a classifier.



26. What's a Fourier transform?

Fourier Transform is a mathematical technique that transforms any function of time to a function of frequency. Fourier transform is closely related to Fourier series. It takes any time-based pattern for input and calculates the overall cycle offset, rotation speed and strength for all possible cycles. Fourier transform is best applied to waveforms since it has functions of time and space. Once a Fourier transform applied on a waveform, it gets decomposed into a sinusoid.

27. What do you mean by Associative Rule Mining (ARM)?

Associative Rule Mining is one of the techniques to discover patterns in data like features (dimensions) which occur together and features (dimensions) which are correlated. It is mostly used in Market-based Analysis to find how frequently an itemset occurs in a transaction. Association rules have to satisfy minimum support and minimum confidence at the very same time. Association rule generation generally comprised of two different steps:

- “A min support threshold is given to obtain all frequent item-sets in a database.”
- “A min confidence constraint is given to these frequent item-sets in order to form the association rules.”

Support is a measure of how often the “item set” appears in the data set and Confidence is a measure of how often a particular rule has been found to be true.

28. What is Marginalisation? Explain the process.

Marginalisation is summing the probability of a random variable X given joint probability distribution of X with other variables. It is an application of the law of total probability.

$$P(X=x) = \sum_Y P(X=x, Y)$$

Given the joint probability $P(X=x, Y)$, we can use marginalization to find $P(X=x)$. So, it is to find distribution of one random variable by exhausting cases on other random variables.

29. Explain the phrase “Curse of Dimensionality”.

The [Curse of Dimensionality](#) refers to the situation when your data has too many features.

The phrase is used to express the difficulty of using brute force or grid search to optimize a function with too many inputs.

It can also refer to several other issues like:

- If we have more features than observations, we have a risk of overfitting the model.
- When we have too many features, observations become harder to cluster. Too many dimensions cause every observation in the dataset to appear equidistant from all others and no meaningful clusters can be formed.

Dimensionality reduction techniques like PCA come to the rescue in such cases.

30. What is the Principle Component Analysis?

The idea here is to reduce the dimensionality of the data set by reducing the number of variables that are correlated with each other. Although the variation needs to be retained to the maximum extent.

The variables are transformed into a new set of variables that are known as Principal Components'. These PCs are the eigenvectors of a covariance matrix and therefore are orthogonal.

31. Why is rotation of components so important in Principle Component Analysis (PCA)?

Rotation in PCA is very important as it maximizes the separation within the variance obtained by all the components because of which interpretation of components would become easier. If the components are not rotated, then we need extended components to describe variance of the components.

32. What are outliers? Mention three methods to deal with outliers.

A data point that is considerably distant from the other similar data points is known as an outlier. They may occur due to experimental errors or variability in measurement. They are problematic and can mislead a training process, which eventually results in longer training time, inaccurate models, and poor results.

The three methods to deal with outliers are:

Univariate method – looks for data points having extreme values on a single variable

Multivariate method – looks for unusual combinations on all the variables

Minkowski error – reduces the contribution of potential outliers in the training process

Also Read - [Advantages of pursuing a career in Machine Learning](#)

33. What is the difference between regularization and normalisation?

Normalisation	Regularisation
Normalisation adjusts the data; . If your data is on very different scales (especially low to high), you would want to normalise the data. Alter each column to have compatible basic statistics. This can be helpful to make sure there is no loss of	Regularisation adjusts the prediction function. Regularization imposes some control on this by providing simpler fitting functions over complex ones.

accuracy. One of the goals of model training is to identify the signal and ignore the noise if the model is given free rein to minimize error, there is a possibility of suffering from overfitting.

34. Explain the difference between Normalization and Standardization.

Normalization and Standardization are the two very popular methods used for feature scaling.

Normalisation	Standardization
Normalization refers to re-scaling the values to fit into a range of [0,1]. Normalization is useful when all parameters need to have an identical positive scale however the outliers from the data set are lost.	Standardization refers to re-scaling data to have a mean of 0 and a standard deviation of 1 (Unit variance)

35. List the most popular distribution curves along with scenarios where you will use them in an algorithm.

The most popular distribution curves are as follows- Bernoulli Distribution, Uniform Distribution, Binomial Distribution, Normal Distribution, [Poisson Distribution](#), and Exponential Distribution.

Each of these distribution curves is used in various scenarios.

Bernoulli Distribution can be used to check if a team will win a championship or not, a newborn child is either male or female, you either pass an exam or not, etc.

Uniform distribution is a probability distribution that has a constant probability. Rolling a single dice is one example because it has a fixed number of outcomes.

Binomial distribution is a probability with only two possible outcomes, the prefix 'bi' means two or twice. An example of this would be a coin toss. The outcome will either be heads or tails.

Normal distribution describes how the values of a variable are distributed. It is typically a symmetric distribution where most of the observations cluster around the central peak. The values further away from the mean taper off equally in both directions. An example would be the height of students in a classroom.

Poisson distribution helps predict the probability of certain events happening when you know how often that event has occurred. It can be used by businessmen to make forecasts about the number of customers on certain days and allows them to adjust supply according to the demand.

Exponential distribution is concerned with the amount of time until a specific event occurs. For example, how long a car battery would last, in months.

36. How do we check the normality of a data set or a feature?

Visually, we can check it using plots. There is a list of Normality checks, they are as follow:

- Shapiro-Wilk W Test
- Anderson-Darling Test
- Martinez-Iglewicz Test
- Kolmogorov-Smirnov Test
- D'Agostino Skewness Test

37. What is Linear Regression?

Linear Function can be defined as a Mathematical function on a 2D plane as, $Y = Mx + C$, where Y is a dependent variable and X is Independent Variable, C is Intercept and M is slope and same can be expressed as Y is a Function of X or $Y = F(x)$.

At any given value of X, one can compute the value of Y, using the equation of Line. This relation between Y and X, with a degree of the polynomial as 1 is called [Linear Regression](#).

In Predictive Modeling, LR is represented as $Y = B_0 + B_1x_1 + B_2x_2$

The value of B1 and B2 determines the strength of the correlation between features and the dependent variable.

Example: Stock Value in \$ = Intercept + (+/-B1)*(Opening value of Stock) + (+/-B2)*(Previous Day Highest value of Stock)

38. Differentiate between regression and classification.

Regression and classification are categorized under the same umbrella of supervised machine learning. The main difference between them is that the output variable in the regression is numerical (or continuous) while that for classification is categorical (or discrete).

Example: To predict the definite Temperature of a place is Regression problem whereas predicting whether the day will be Sunny cloudy or there will be rain is a case of classification.

39. What is target imbalance? How do we fix it? A scenario where you have performed target imbalance on data. Which metrics and algorithms do you find suitable to input this data onto?

If you have categorical variables as the target when you cluster them together or perform a frequency count on them if there are certain categories which are more in number as compared to others by a very significant number. This is known as the target imbalance.

Example: Target column – 0,0,0,1,0,2,0,0,1,1 [0s: 60%, 1: 30%, 2:10%] 0 are in majority. To fix this, we can perform up-sampling or down-sampling. Before fixing this problem let's assume that the performance metrics used was confusion metrics. After fixing this problem we can shift the metric system to AUC: ROC. Since we added/deleted data [up sampling or downsampling], we can go ahead with a stricter algorithm like SVM, Gradient boosting or ADA boosting.

40. List all assumptions for data to be met before starting with linear regression.

Before starting linear regression, the assumptions to be met are as follow:

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

41. When does the linear regression line stop rotating or finds an optimal spot where it is fitted on data?

A place where the highest RSquared value is found, is the place where the line comes to rest. RSquared represents the amount of variance captured by the virtual linear regression line with respect to the total variance captured by the dataset.

42. Why is logistic regression a type of classification technique and not a regression?

Name the function it is derived from?

Since the target column is categorical, it uses linear regression to create an odd function that is wrapped with a log function to use regression as a classifier. Hence, it is a type of classification technique and not a regression. It is derived from cost function.

43. What could be the issue when the beta value for a certain variable varies way too much in each subset when regression is run on different subsets of the given dataset?

Variations in the beta values in every subset implies that the dataset is heterogeneous. To overcome this problem, we can use a different model for each of the clustered subsets of the dataset or use a non-parametric model such as decision trees.

44. What does the term Variance Inflation Factor mean?

Variance Inflation Factor (VIF) is the ratio of the variance of the model to the variance of the model with only one independent variable. VIF gives the estimate of the volume of multicollinearity in a set of many regression variables.

$VIF = \text{Variance of model with one independent variable}$

45. Which machine learning algorithm is known as the lazy learner and why is it called so?

KNN is a Machine Learning algorithm known as a lazy learner. K-NN is a lazy learner because it doesn't learn any machine learnt values or variables from the training data but dynamically calculates distance every time it wants to classify, hence memorising the training dataset instead.

Machine Learning Interview Questions for Experienced

We know what the companies are looking for, and with that in mind, we have prepared the set of Machine Learning interview questions an experienced professional may be asked. So, prepare accordingly if you wish to ace the interview in one go.

46. Is it possible to use KNN for image processing?

Yes, it is possible to use KNN for image processing. It can be done by converting the 3-dimensional image into a single-dimensional vector and using the same as input to KNN.

47. Differentiate between K-Means and KNN algorithms?

KNN algorithms	K-Means
<p>KNN algorithms is Supervised Learning where-as K-Means is Unsupervised Learning. With KNN, we predict the label of the unidentified element based on its nearest neighbour and further extend this approach for solving classification/regression-based problems.</p>	<p>K-Means is Unsupervised Learning, where we don't have any Labels present, in other words, no Target Variables and thus we try to cluster the data based upon their coord</p>

48. How does the SVM algorithm deal with self-learning?

SVM has a learning rate and expansion rate which takes care of this. The [learning rate](#) compensates or penalises the hyperplanes for making all the wrong moves and expansion rate deals with finding the maximum separation area between classes.

49. What are Kernels in SVM? List popular kernels used in SVM along with a scenario of their applications.

The function of the kernel is to take data as input and transform it into the required form. A few popular Kernels used in SVM are as follows: RBF, Linear, Sigmoid, Polynomial, Hyperbolic, Laplace, etc.

50. What is Kernel Trick in an SVM Algorithm?

Kernel Trick is a mathematical function which when applied on data points, can find the region of classification between two different classes. Based on the choice of function, be it linear or radial, which purely depends upon the distribution of data, one can build a classifier.

51. What are ensemble models? Explain how ensemble techniques yield better learning as compared to traditional classification ML algorithms.

An ensemble is a group of models that are used together for prediction both in classification and regression classes. Ensemble learning helps improve ML results because it combines several models. By doing so, it allows for a better predictive

performance compared to a single model.

They are superior to individual models as they reduce variance, average out biases, and have lesser chances of overfitting.

52. What are overfitting and underfitting? Why does the decision tree algorithm suffer often with overfitting problems?

Overfitting is a statistical model or machine learning algorithm which captures the noise of the data. Underfitting is a model or machine learning algorithm which does not fit the data well enough and occurs if the model or algorithm shows low variance but high bias.

In decision trees, overfitting occurs when the tree is designed to perfectly fit all samples in the training data set. This results in branches with strict rules or sparse data and affects the accuracy when predicting samples that aren't part of the training set.

53. What is OOB error and how does it occur?

For each bootstrap sample, there is one-third of data that was not used in the creation of the tree, i.e., it was out of the sample. This data is referred to as out of bag data. In order to get an unbiased measure of the accuracy of the model over test data, out of bag error is used. The out of bag data is passed for each tree is passed through that tree and the outputs are aggregated to give out of bag error. This percentage error is quite effective in estimating the error in the testing set and does not require further [cross-validation](#).

54. Why boosting is a more stable algorithm as compared to other ensemble algorithms?

Boosting focuses on errors found in previous iterations until they become obsolete. Whereas in bagging there is no corrective loop. This is why boosting is a more stable algorithm compared to other ensemble algorithms.

55. How do you handle outliers in the data?

Outlier is an observation in the data set that is far away from other observations in the data set. We can discover outliers using tools and functions like box plot, scatter plot, Z-Score, IQR score etc. and then handle them based on the visualization we have got. To handle outliers, we can cap at some threshold, use transformations to reduce skewness of the data and remove outliers if they are anomalies or errors.

56. List popular cross validation techniques.

There are mainly six types of cross validation techniques. They are as follow:

- K fold
- Stratified k fold
- Leave one out
- Bootstrapping
- Random search cv
- Grid search cv

57. Is it possible to test for the probability of improving model accuracy without cross-validation techniques? If yes, please explain.

Yes, it is possible to test for the probability of improving model accuracy without cross-validation techniques. We can do so by running the ML model for say **n** number of iterations, recording the accuracy. Plot all the accuracies and remove the 5% of low probability values. Measure the left [low] cut off and right [high] cut off. With the remaining 95% confidence, we can say that the model can go as low or as high [as mentioned within cut off points].

58. Name a popular dimensionality reduction algorithm.

Popular dimensionality reduction algorithms are [Principal Component Analysis](#) and Factor Analysis.

Principal Component Analysis creates one or more index variables from a larger set of measured variables. Factor Analysis is a model of the measurement of a latent variable. This latent variable cannot be measured with a single variable and is seen through a relationship it causes in a set of **y** variables.

59. How can we use a dataset without the target variable into supervised learning algorithms?

Input the data set into a [clustering algorithm](#), generate optimal clusters, label the cluster numbers as the new target variable. Now, the dataset has independent and target variables present. This ensures that the dataset is ready to be used in supervised learning algorithms.

60. List all types of popular recommendation systems? Name and explain two personalized recommendation systems along with their ease of implementation.

Popularity based recommendation, content-based recommendation, user-based collaborative filter, and item-based recommendation are the popular types of recommendation systems.

Personalised [Recommendation systems](#) are- Content-based recommendation, user-based collaborative filter, and item-based recommendation. User-based collaborative filter and item-based recommendations are more personalised. Ease to maintain: Similarity matrix can be maintained easily with Item-based recommendation.

61. How do we deal with sparsity issues in recommendation systems? How do we measure its effectiveness? Explain.

Singular value decomposition can be used to generate the prediction matrix. RMSE is the measure that helps us understand how close the prediction matrix is to the original matrix.

62. Name and define techniques used to find similarities in the recommendation system.

Pearson correlation and Cosine correlation are techniques used to find similarities in recommendation systems.

63. State the limitations of Fixed Basis Function.

Linear separability in feature space doesn't imply linear separability in input space. So, Inputs are non-linearly transformed using vectors of basic functions with increased dimensionality. Limitations of Fixed basis functions are:

- Non-Linear transformations cannot remove overlap between two classes but they can increase overlap.
- Often it is not clear which basis functions are the best fit for a given task. So, learning the basic functions can be useful over using fixed basis functions.
- If we want to use only fixed ones, we can use a lot of them and let the model figure out the best fit but that would lead to overfitting the model thereby making it unstable.

64. Define and explain the concept of Inductive Bias with some examples.

Inductive Bias is a set of assumptions that humans use to predict outputs given inputs that the learning algorithm has not encountered yet. When we are trying to learn Y from X and the hypothesis space for Y is infinite, we need to reduce the scope by our beliefs/assumptions about the hypothesis space which is also called inductive bias.

Through these assumptions, we constrain our hypothesis space and also get the capability to incrementally test and improve on the data using hyper-parameters. Examples:

1. We assume that Y varies linearly with X while applying Linear regression.
2. We assume that there exists a hyperplane separating negative and positive examples.

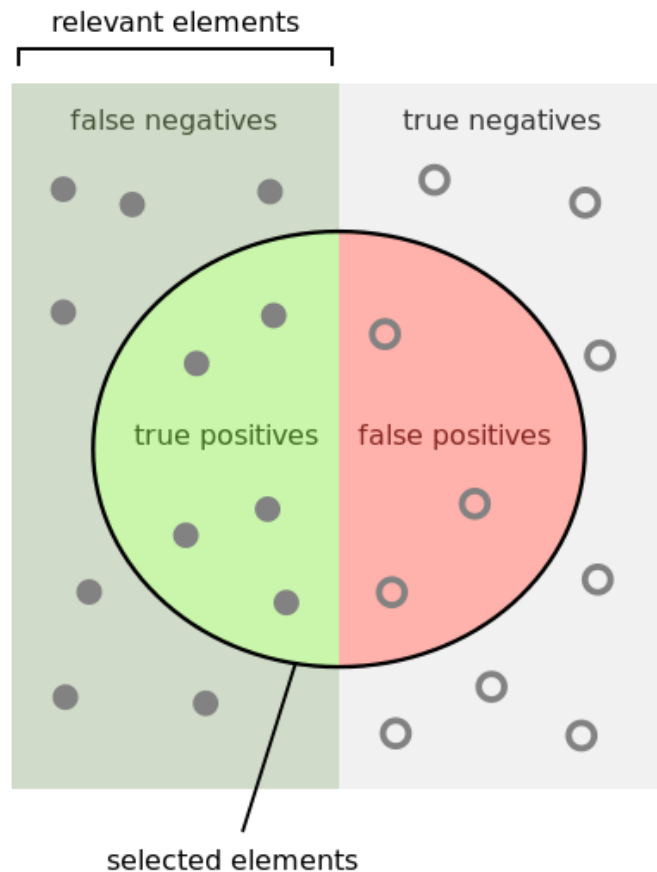
65. Explain the term instance-based learning.

Instance Based Learning is a set of procedures for regression and classification which produce a class label prediction based on resemblance to its nearest neighbors in the training data set. These algorithms just collect all the data and get an answer when required or queried. In simple words they are a set of procedures for solving new problems based on the solutions of already solved problems in the past which are similar to the current problem.

66. Keeping train and test split criteria in mind, is it good to perform scaling before the split or after the split?

Scaling should be done post-train and test split ideally. If the data is closely packed, then scaling post or pre-split should not make much difference.

67. Define precision, recall and F1 Score?



How many selected items are relevant?

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

The metric used to assess the performance of the classification model is Confusion Metric. Confusion Metric can be further interpreted with the following terms:-

True Positives (TP) – These are the correctly predicted positive values. It implies that the value of the actual class is yes and the value of the predicted class is also yes.

True Negatives (TN) – These are the correctly predicted negative values. It implies that the value of the actual class is no and the value of the predicted class is also no.

False positives and false negatives, these values occur when your actual class contradicts with the predicted class.

Now,

Recall, also known as Sensitivity is the ratio of true positive rate (TP), to all observations

in actual class – yes

$$\text{Recall} = \text{TP}/(\text{TP}+\text{FN})$$

Precision is the ratio of positive predictive value, which measures the amount of accurate positives model predicted viz a viz number of positives it claims.

$$\text{Precision} = \text{TP}/(\text{TP}+\text{FP})$$

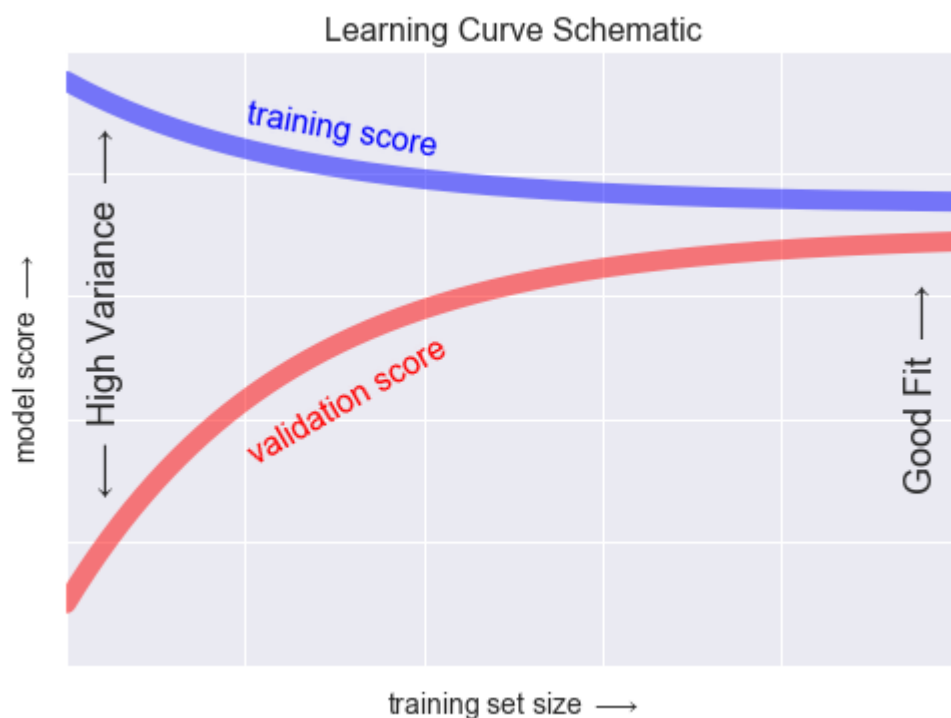
Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations.

$$\text{Accuracy} = (\text{TP}+\text{TN})/(\text{TP}+\text{FP}+\text{FN}+\text{TN})$$

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have a similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall.

68. Plot validation score and training score with data set size on the x-axis and another plot with model complexity on the x-axis.

For high bias in the models, the performance of the model on the validation data set is similar to the performance on the training data set. For high variance in the models, the performance of the model on the validation set is worse than the performance on the training set.



69. What is Bayes' Theorem? State at least 1 use case with respect to the machine learning context?

Bayes' Theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event. For example, if cancer is related to age, then, using Bayes' theorem, a person's age can be used to more accurately assess the probability that they have cancer than can be done without the knowledge of the person's age.

Chain rule for Bayesian probability can be used to predict the likelihood of the next word in the sentence.

70. What is Naive Bayes? Why is it Naive?

Naive Bayes classifiers are a series of classification algorithms that are based on the Bayes theorem. This family of algorithm shares a common principle which treats every pair of features independently while being classified.

Naive Bayes is considered Naive because the attributes in it (for the class) is independent of others in the same class. This lack of dependence between two attributes of the same class creates the quality of naiveness.

Read more about [Naive Bayes](#).

71. Explain how a Naive Bayes Classifier works.

Naive Bayes classifiers are a family of algorithms which are derived from the Bayes theorem of probability. It works on the fundamental assumption that every set of two features that is being classified is independent of each other and every feature makes an equal and independent contribution to the outcome.

$$P(y|x_1, \dots, x_n) = \frac{P(x_1|y)P(x_2|y)\dots P(x_n|y)P(y)}{P(x_1)P(x_2)\dots P(x_n)}$$

$$P(y|x_1, \dots, x_n) \propto P(y) \prod_{i=1}^n P(x_i|y)$$

$$y = \operatorname{argmax}_y P(y) \prod_{i=1}^n P(x_i|y)$$

72. What do the terms prior probability and marginal likelihood in context of Naive Bayes theorem mean?

Prior probability is the percentage of dependent binary variables in the data set. If you are given a dataset and dependent variable is either 1 or 0 and percentage of 1 is 65% and percentage of 0 is 35%. Then, the probability that any new input for that variable of being 1 would be 65%.

Marginal likelihood is the denominator of the Bayes equation and it makes sure that the posterior probability is valid by making its area 1.

$$\overbrace{p(\boldsymbol{\theta}|\mathbf{X})}^{\text{posterior}} = \frac{\overbrace{p(\mathbf{X}|\boldsymbol{\theta})}^{\text{likelihood}} \overbrace{p(\boldsymbol{\theta}|\alpha)}^{\text{prior}}}{\underbrace{\int_{\boldsymbol{\theta}} p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\alpha)}_{\text{marginal likelihood}}} = \frac{\overbrace{p(\mathbf{X}|\boldsymbol{\theta})}^{\text{likelihood}} \overbrace{p(\boldsymbol{\theta}|\alpha)}^{\text{prior}}}{\underbrace{p(\mathbf{X}|\alpha)}_{\text{marginal likelihood}}} = \frac{\overbrace{p(\mathbf{X},\boldsymbol{\theta})}^{\text{joint probability}}}{\underbrace{p(\mathbf{X}|\alpha)}_{\text{marginal likelihood}}} = Kp(\mathbf{X}, \boldsymbol{\theta}) \propto p(\mathbf{X}, \boldsymbol{\theta})$$

73. Explain the difference between Lasso and Ridge?

Lasso(L1) and Ridge(L2) are the regularization techniques where we penalize the coefficients to find the optimum solution. In ridge, the penalty function is defined by the sum of the squares of the coefficients and for the Lasso, we penalize the sum of the absolute values of the coefficients. Another type of regularization method is ElasticNet, it is a hybrid penalizing function of both lasso and ridge.

74. What's the difference between probability and likelihood?

Probability is the measure of the likelihood that an event will occur that is, what is the certainty that a specific event will occur? Where-as a likelihood function is a function of parameters within the parameter space that describes the probability of obtaining the observed data.

So the fundamental difference is, Probability attaches to possible results; likelihood attaches to hypotheses.

75. Why would you Prune your tree?

In the context of data science or AIML, pruning refers to the process of reducing redundant branches of a decision tree. Decision Trees are prone to overfitting, pruning the tree helps to reduce the size and minimizes the chances of overfitting. Pruning involves turning branches of a decision tree into leaf nodes and removing the leaf nodes from the original branch. It serves as a tool to perform the tradeoff.

76. Model accuracy or Model performance? Which one will you prefer and why?

This is a trick question, one should first get a clear idea, what is Model Performance? If Performance means speed, then it depends upon the nature of the application, any application related to the real-time scenario will need high speed as an important feature. Example: The best of Search Results will lose its virtue if the Query results do not appear fast.

If Performance is hinted at Why Accuracy is not the most important virtue – For any imbalanced data set, more than Accuracy, it will be an F1 score than will explain the business case and in case data is imbalanced, then Precision and Recall will be more important than rest.

77. List the advantages and limitations of the Temporal Difference Learning

Method.

Temporal Difference Learning Method is a mix of Monte Carlo method and Dynamic programming method. Some of the advantages of this method include:

- It can learn in every step online or offline.
- It can learn from a sequence which is not complete as well.
- It can work in continuous environments.

- It has lower variance compared to MC method and is more efficient than MC method.

Limitations of TD method are:

- It is a biased estimation.
- It is more sensitive to initialization.

78. How would you handle an imbalanced dataset?

Sampling Techniques can help with an imbalanced dataset. There are two ways to perform sampling, Under Sample or Over Sampling.

In Under Sampling, we reduce the size of the majority class to match minority class thus help by improving performance w.r.t storage and run-time execution, but it potentially discards useful information.

For Over Sampling, we upsample the Minority class and thus solve the problem of information loss, however, we get into the trouble of having Overfitting.

There are other techniques as well –

Cluster-Based Over Sampling – In this case, the K-means clustering algorithm is independently applied to minority and majority class instances. This is to identify clusters in the dataset. Subsequently, each cluster is oversampled such that all clusters of the same class have an equal number of instances and all classes have the same size

Synthetic Minority Over-sampling Technique (SMOTE) – A subset of data is taken from the minority class as an example and then new synthetic similar instances are created which are then added to the original dataset. This technique is good for Numerical data points.

79. Mention some of the EDA Techniques?

Exploratory Data Analysis (EDA) helps analysts to understand the data better and forms the foundation of better models.

Visualization

- Univariate visualization
- Bivariate visualization
- Multivariate visualization

Missing Value Treatment – Replace missing values with Either Mean/Median

Outlier Detection – Use Boxplot to identify the distribution of Outliers, then Apply IQR to set the boundary for IQR

Transformation – Based on the distribution, apply a transformation on the features

Scaling the Dataset – Apply MinMax, Standard Scaler or Z Score Scaling mechanism to scale the data.

Feature Engineering – Need of the domain, and SME knowledge helps Analyst find derivative fields which can fetch more information about the nature of the data

Dimensionality reduction — Helps in reducing the volume of data without losing much information

80. Mention why feature engineering is important in model building and list out some of the techniques used for feature engineering.

Algorithms necessitate features with some specific characteristics to work appropriately. The data is initially in a raw form. You need to extract features from this data before supplying it to the algorithm. This process is called feature engineering. When you have relevant features, the complexity of the algorithms reduces. Then, even if a non-ideal algorithm is used, results come out to be accurate.

Feature engineering primarily has two goals:

- Prepare the suitable input data set to be compatible with the machine learning algorithm constraints.
- Enhance the performance of machine learning models.

Some of the techniques used for feature engineering include Imputation, Binning, Outliers Handling, Log transform, grouping operations, One-Hot encoding, Feature split, Scaling, Extracting date.

81. Differentiate between Statistical Modeling and Machine Learning?

Machine learning models are about making accurate predictions about the situations, like Foot Fall in restaurants, Stock-Price, etc. where-as, Statistical models are designed for inference about the relationships between variables, as What drives the sales in a restaurant, is it food or Ambience.

82. Differentiate between Boosting and Bagging?

Bagging and Boosting are variants of Ensemble Techniques.

Bootstrap Aggregation or bagging is a method that is used to reduce the variance for algorithms having very high variance. Decision trees are a particular family of classifiers which are susceptible to having high bias.

Decision trees have a lot of sensitiveness to the type of data they are trained on. Hence generalization of results is often much more complex to achieve in them despite very high fine-tuning. The results vary greatly if the training data is changed in decision trees.

Hence bagging is utilised where multiple decision trees are made which are trained on samples of the original data and the final result is the average of all these individual models.

Boosting is the process of using an n-weak classifier system for prediction such that every weak classifier compensates for the weaknesses of its classifiers. By weak classifier, we imply a classifier which performs poorly on a given data set.

It's evident that boosting is not an algorithm rather it's a process. Weak classifiers used are generally **logistic regression**, shallow decision trees etc.

There are many algorithms which make use of boosting processes but two of them are mainly used: Adaboost and Gradient Boosting and XGBoost.

83. What is the significance of Gamma and Regularization in SVM?

The gamma defines influence. Low values meaning 'far' and high values meaning 'close'. If gamma is too large, the radius of the area of influence of the support vectors only includes the support vector itself and no amount of regularization with C will be able to prevent overfitting. If gamma is very small, the model is too constrained and cannot capture the complexity of the data.

The regularization parameter (λ) serves as a degree of importance that is given to miss-classifications. This can be used to draw the tradeoff with OverFitting.

1. Why was Machine Learning Introduced?

The simplest answer is to make our lives easier. In the early days of “intelligent” applications, many systems used hardcoded rules of “if” and “else” decisions to process data or adjust the user input. Think of a spam filter whose job is to move the appropriate incoming email messages to a spam folder.

But with the machine learning algorithms, we are given ample information for the data to learn and identify the patterns from the data.

Unlike the normal problems we don't need to write the new rules for each problem in machine learning, we just need to use the same workflow but with a different dataset.

Let's talk about Alan Turing, in his 1950 paper, “Computing Machinery and Intelligence”, Alan asked, “Can machines think?”

Full paper [here](#)

The paper describes the “Imitation Game”, which includes three participants -

- Human acting as a judge,
- Another human, and
- A computer is an attempt to convince the judge that it is human.

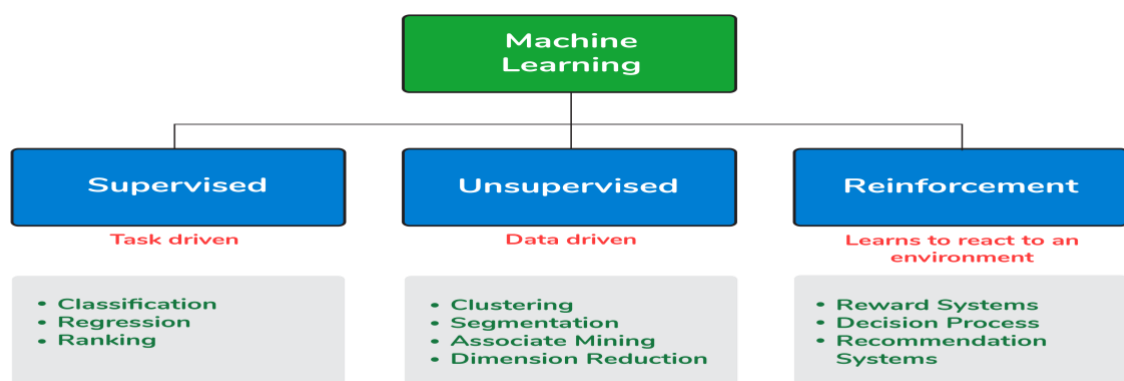
The judge asks the other two participants to talk. While they respond the judge needs to decide which response came from the computer. If the judge could not tell the difference the computer won the game.

The test continues today as an annual competition in artificial intelligence. The aim is simple enough: convince the judge that they are chatting to a human instead of a computer chatbot program.

2. What are Different Types of Machine Learning algorithms?

There are various types of machine learning algorithms. Here is the list of them in a broad category based on:

- Whether they are trained with human supervision (Supervised, unsupervised, reinforcement learning)
- The criteria in the below diagram are not exclusive, we can combine them any way we like.



3. What is Supervised Learning?

Supervised learning is a machine learning algorithm of inferring a function from labeled training data. The training data consists of a set of training examples.

Example: 01

Knowing the height and weight identifying the gender of the person. Below are the popular supervised learning algorithms.

- Support Vector Machines
- Regression
- Naive Bayes
- Decision Trees
- K-nearest Neighbour Algorithm and Neural Networks.

Example: 02

If you build a T-shirt classifier, the labels will be “this is an S, this is an M and this is L”, based on showing the classifier examples of S, M, and L.

4. What is Unsupervised Learning?

Unsupervised learning is also a type of machine learning algorithm used to find patterns on the set of data given. In this, we don't have any dependent variable or label to predict.

Unsupervised Learning Algorithms:

- Clustering,
- Anomaly Detection,
- Neural Networks and Latent Variable Models.

Example:

In the same example, a T-shirt clustering will categorize as “collar style and V neck style”, “crew neck style” and “sleeve types”.

5. What is ‘Naive’ in a Naive Bayes?

The Naive Bayes method is a supervised learning algorithm, it is naive since it makes assumptions by applying Bayes' theorem that all attributes are independent of each other.

Bayes' theorem states the following relationship, given class variable y and dependent vector x_1 through x_n :

$$P(y_i | x_1, \dots, x_n) = P(y_i)P(x_1, \dots, x_n | y_i)P(x_1, \dots, x_n)$$

Using the naive conditional independence assumption that each x_i is independent: for all i this relationship is simplified to:

$$P(x_i | y_i, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = P(x_i | y_i)$$

Since, $P(x_1, \dots, x_n)$ is a constant given the input, we can use the following classification rule:

$P(y_i | x_1, \dots, x_n) = P(y_i) \prod_{n=1}^n P(x_i | y_i)P(x_1, \dots, x_n)$ and we can also use Maximum A Posteriori (MAP) estimation to estimate $P(y_i)$ and $P(y_i | x_i)$ the former is then the relative frequency of class y_i in the training set.

$$P(y_i | x_1, \dots, x_n) \propto P(y_i) \prod_{n=1}^n P(x_i | y_i)$$

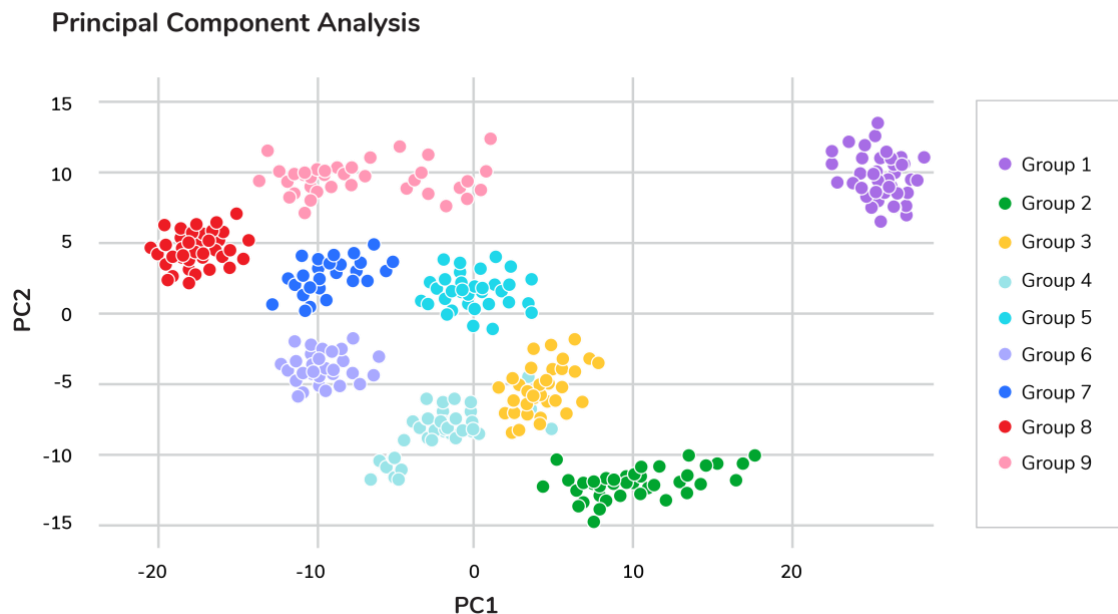
$$y = \arg \max P(y_i) \prod_{n=1}^n P(x_i | y_i)$$

The different naive Bayes classifiers mainly differ by the assumptions they make regarding the distribution of $P(y_i | x_i)$: can be Bernoulli, binomial, Gaussian, and so on.

6. What is PCA? When do you use it?

Principal component analysis (PCA) is most commonly used for dimension reduction.

In this case, PCA measures the variation in each variable (or column in the table). If there is little variation, it throws the variable out, as illustrated in the figure below:



Principal component analysis (PCA)

Thus making the dataset easier to visualize. PCA is used in finance, neuroscience, and pharmacology.

It is very useful as a preprocessing step, especially when there are linear correlations between features.

7. Explain SVM Algorithm in Detail

A Support Vector Machine (SVM) is a very powerful and versatile supervised machine learning model, capable of performing linear or non-linear classification, regression, and even outlier detection.

Suppose we have given some data points that each belong to one of two classes, and the goal is to separate two classes based on a set of examples.

In SVM, a data point is viewed as a p -dimensional vector (a list of p numbers), and we wanted to know whether we can separate such points with a $(p-1)$ -dimensional hyperplane. This is called a linear classifier.

There are many hyperplanes that classify the data. To choose the best hyperplane that represents the largest separation or margin between the two classes.

If such a hyperplane exists, it is known as a maximum-margin hyperplane and the linear classifier it defines is known as a maximum margin classifier. The best hyperplane that divides the data in H_3

We have data $(x_1, y_1), \dots, (x_n, y_n)$, and different features (x_{i1}, \dots, x_{ip}) , and y_i is either 1 or -1.

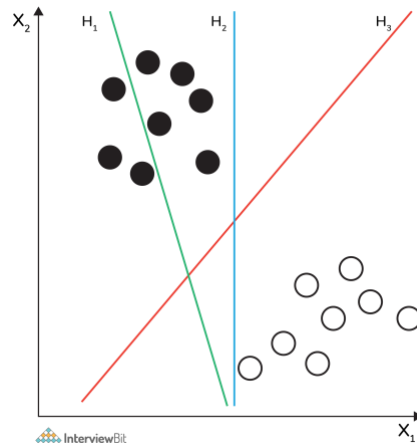
The equation of the hyperplane H_3 is the set of points satisfying:

$$w \cdot x - b = 0$$

Where w is the normal vector of the hyperplane. The parameter $b/\|w\|$ determines the offset of the hyperplane from the origin along the normal vector w

So for each i , either x_i is in the hyperplane of 1 or -1. Basically, x satisfies:

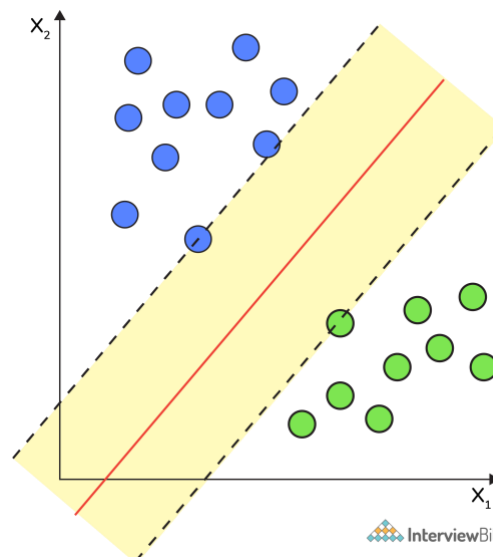
$$w \cdot x_i - b = 1 \quad \text{or} \quad w \cdot x_i - b = -1$$



8. What are Support Vectors in SVM?

A Support Vector Machine (SVM) is an algorithm that tries to fit a line (or plane or hyperplane) between the different classes that maximizes the distance from the line to the points of the classes.

In this way, it tries to find a robust separation between the classes. The Support Vectors are the points of the edge of the dividing hyperplane as in the below figure.



9. What are Different Kernels in SVM?

There are six types of kernels in SVM:

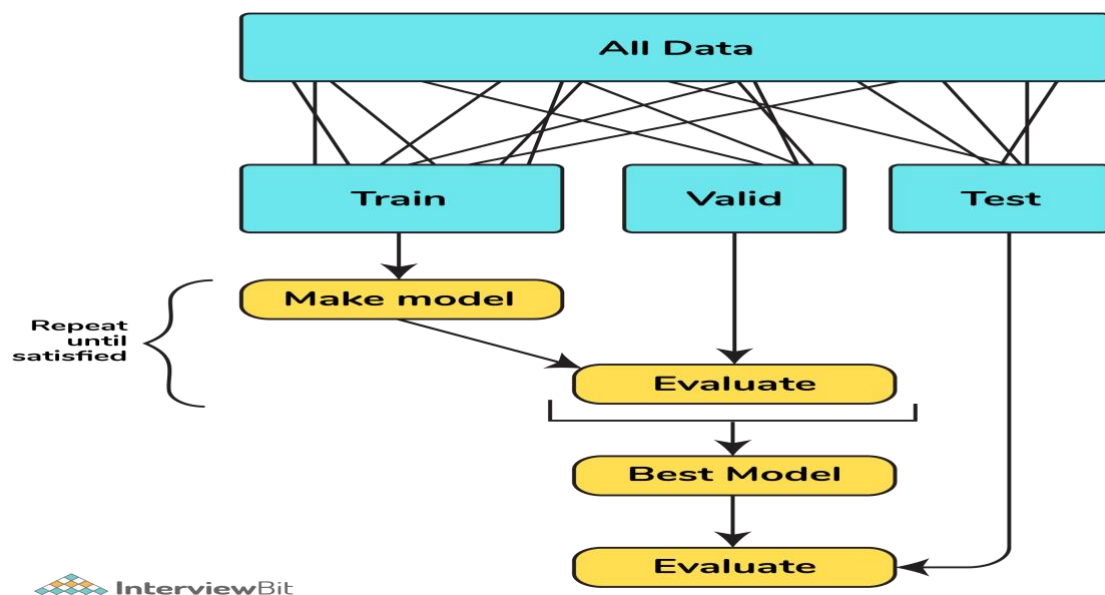
- Linear kernel - used when data is linearly separable.
- Polynomial kernel - When you have discrete data that has no natural notion of smoothness.

- Radial basis kernel - Create a decision boundary able to do a much better job of separating two classes than the linear kernel.
- Sigmoid kernel - used as an activation function for neural networks.

10. What is Cross-Validation?

Cross-validation is a method of splitting all your data into three parts: training, testing, and validation data. Data is split into k subsets, and the model has trained on $k-1$ of those datasets.

The last subset is held for testing. This is done for each of the subsets. This is k -fold cross-validation. Finally, the scores from all the k -folds are averaged to produce the final score.



Cross-validation

11. What is Bias in Machine Learning?

Bias in data tells us there is inconsistency in data. The inconsistency may occur for several reasons which are not mutually exclusive.

For example, a tech giant like Amazon to speed the hiring process they build one engine where they are going to give 100 resumes, it will spit out the top five, and hire those.

When the company realized the software was not producing gender-neutral results it was tweaked to remove this bias.

12. Explain the Difference Between Classification and Regression?

Classification is used to produce discrete results, classification is used to classify data into some specific categories.

For example, classifying emails into spam and non-spam categories.

Whereas, regression deals with continuous data.

For example, predicting stock prices at a certain point in time.

Classification is used to predict the output into a group of classes.

For example, Is it Hot or Cold tomorrow?

Whereas, regression is used to predict the relationship that data represents.

For example, What is the temperature tomorrow?

Advanced Machine Learning Questions

13. What is F1 score? How would you use it?

Let's have a look at this table before directly jumping into the F1 score.

Prediction	Predicted Yes	Predicted No
Actual Yes	True Positive (TP)	False Negative (FN)
Actual No	False Positive (FP)	True Negative (TN)

In binary classification we consider the F1 score to be a measure of the model's accuracy. The F1 score is a weighted average of precision and recall scores.

$$F1 = 2TP / 2TP + FP + FN$$

We see scores for F1 between 0 and 1, where 0 is the worst score and 1 is the best score. The F1 score is typically used in information retrieval to see how well a model retrieves relevant results and our model is performing.

14. Define Precision and Recall?

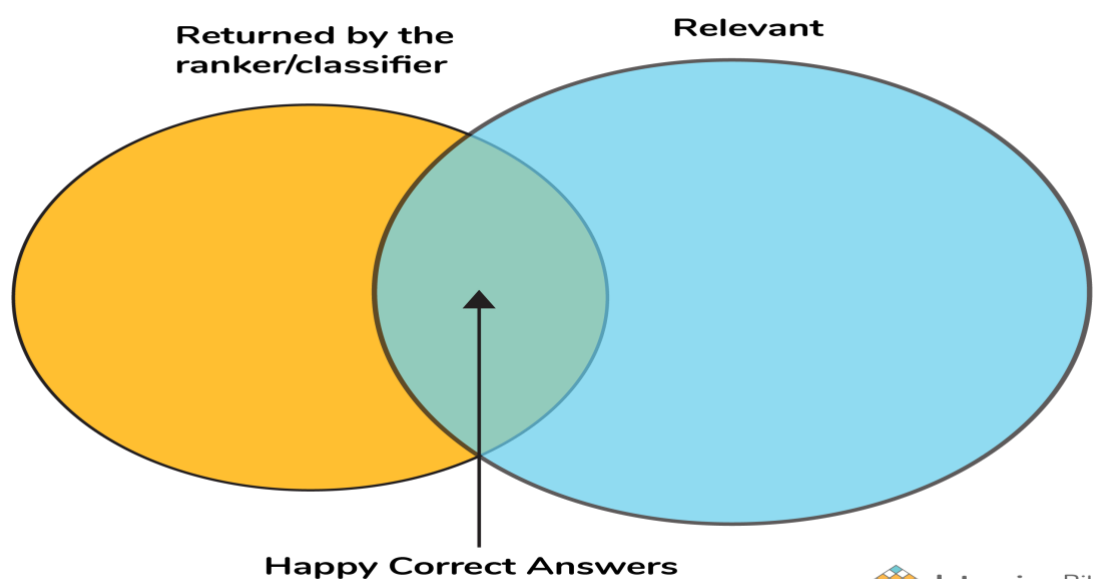
Precision and recall are ways of monitoring the power of machine learning implementation. But they often used at the same time.

Precision answers the question, "Out of the items that the classifier predicted to be relevant, how many are truly relevant?"

Whereas, recall answers the question, "Out of all the items that are truly relevant, how many are found by the classifier?"

In general, the meaning of precision is the fact of being exact and accurate. So the same will go in our machine learning model as well. If you have a set of items that your model needs to predict to be relevant. How many items are truly relevant?

The below figure shows the Venn diagram that precision and recall.



Mathematically, precision and recall can be defined as the following:

precision = # happy correct answers/# total items returned by ranker

recall = # happy correct answers/# total relevant answers

15. How to Tackle Overfitting and Underfitting?

Overfitting means the model fitted to training **data too well**, in this case, we need to resample the data and estimate the model accuracy using techniques like k-fold cross-validation.

Whereas for the Underfitting case we are **not able to understand** or capture the patterns from the data, in this case, we need to change the algorithms, or we need to feed more data points to the model.

16. What is a Neural Network?

It is a simplified model of the human brain. Much like the brain, it has neurons that activate when encountering something similar.

The different neurons are connected via connections that help information flow from one neuron to another.

17. What are Loss Function and Cost Functions? Explain the key Difference Between them?

When calculating loss we consider only a single data point, then we use the term loss function.

Whereas, when calculating the sum of error for multiple data then we use the cost function. There is no major difference.

In other words, the loss function is to capture the difference between the actual and predicted values for a single record whereas cost functions aggregate the difference for the entire training dataset.

The Most commonly used loss functions are Mean-squared error and Hinge loss.

Mean-Squared Error(MSE): In simple words, we can say how our model predicted values against the actual values.

$$MSE = \sqrt{(\text{predicted value} - \text{actual value})^2}$$

Hinge loss: It is used to train the machine learning classifier, which is

$$L(y) = \max(0, 1 - yy)$$

Where $y = -1$ or 1 indicating two classes and y represents the output form of the classifier. The most common cost function represents the total cost as the sum of the fixed costs and the variable costs in the equation $y = mx + b$

18. What is Ensemble learning?

Ensemble learning is a method that combines multiple machine learning models to create more powerful models.

There are many reasons for a model to be different. Few reasons are:

- Different Population
- Different Hypothesis
- Different modeling techniques

When working with the model's training and testing data, we will experience an error. This error might be bias, variance, and irreducible error.

Now the model should always have a balance between bias and variance, which we call a bias-variance trade-off.

This ensemble learning is a way to perform this trade-off.

There are many ensemble techniques available but when aggregating multiple models there are two general methods:

- Bagging, a native method: take the training set and generate new training sets off of it.
- Boosting, a more elegant method: similar to bagging, boosting is used to optimize the best weighting scheme for a training set.

19. How do you make sure which Machine Learning Algorithm to use?

It completely depends on the dataset we have. If the data is discrete we use SVM. If the dataset is continuous we use linear regression.

So there is no specific way that lets us know which ML algorithm to use, it all depends on the exploratory data analysis (EDA).

EDA is like "interviewing" the dataset; As part of our interview we do the following:

- Classify our variables as continuous, categorical, and so forth.
- Summarize our variables using descriptive statistics.
- Visualize our variables using charts.

Based on the above observations select one best-fit algorithm for a particular dataset.

20. How to Handle Outlier Values?

An Outlier is an observation in the dataset that is far away from other observations in the dataset. Tools used to discover outliers are

- Box plot
- Z-score
- Scatter plot, etc.

Typically, we need to follow three simple strategies to handle outliers:

- We can drop them.
- We can mark them as outliers and include them as a feature.
- Likewise, we can transform the feature to reduce the effect of the outlier.

21. What is a Random Forest? How does it work?

Random forest is a versatile machine learning method capable of performing both regression and classification tasks.

Like bagging and boosting, random forest works by combining a set of other tree models. Random forest builds a tree from a random sample of the columns in the test data.

Here's are the steps how a random forest creates the trees:

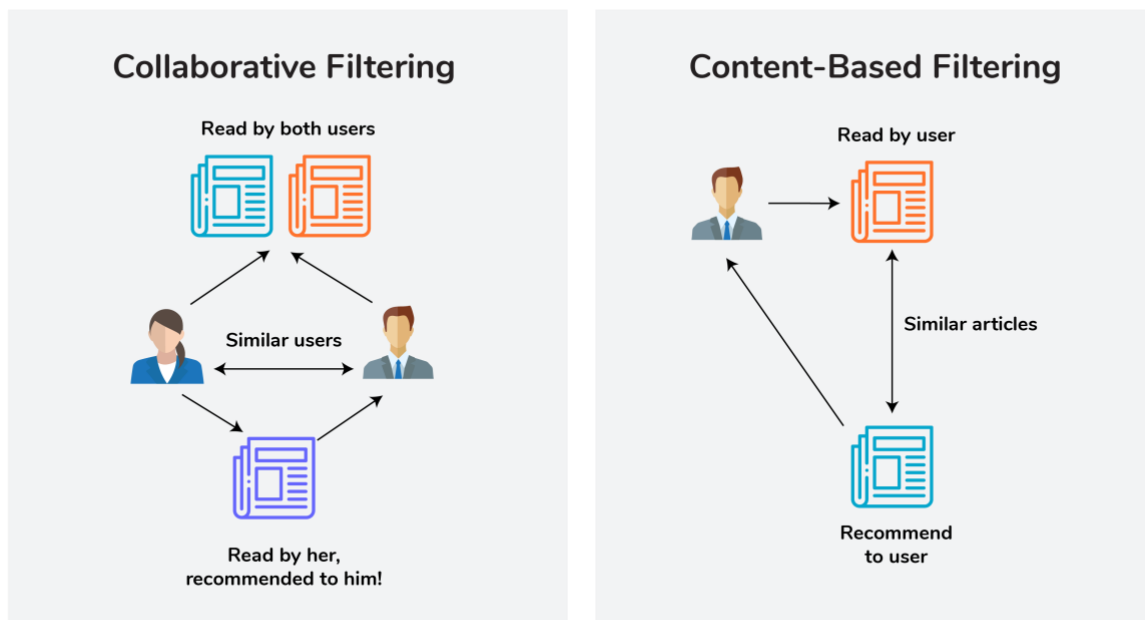
- Take a sample size from the training data.
- Begin with a single node.
- Run the following algorithm, from the start node:

- If the number of observations is less than node size then stop.
- Select random variables.
- Find the variable that does the “best” job of splitting the observations.
- Split the observations into two nodes.
- Call step `a` on each of these nodes.

22. What is Collaborative Filtering? And Content-Based Filtering?

Collaborative filtering is a proven technique for personalized content recommendations. Collaborative filtering is a type of recommendation system that predicts new content by matching the interests of the individual user with the preferences of many users.

Content-based recommender systems are focused only on the preferences of the user. New recommendations are made to the user from similar content according to the user’s previous choices.



Collaborative Filtering and Content-Based Filtering

23. What is Clustering?

Clustering is the process of grouping a set of objects into a number of groups. Objects should be similar to one another within the same cluster and dissimilar to those in other clusters.

A few types of clustering are:

- Hierarchical clustering
- K means clustering
- Density-based clustering
- Fuzzy clustering, etc.

24. How can you select K for K-means Clustering?

There are two kinds of methods that include direct methods and statistical testing methods:

- Direct methods: It contains elbow and silhouette
- Statistical testing methods: It has gap statistics.

The silhouette is the most frequently used while determining the optimal value of k.

25. What are Recommender Systems?

A recommendation engine is a system used to predict users' interests and recommend products that are quite likely interesting for them.

Data required for recommender systems stems from explicit user ratings after watching a film or listening to a song, from implicit search engine queries and purchase histories, or from other knowledge about the users/items themselves.

26. How do check the Normality of a dataset?

Visually, we can use plots. A few of the normality checks are as follows:

- Shapiro-Wilk Test
- Anderson-Darling Test
- Martinez-Iglewicz Test
- Kolmogorov-Smirnov Test
- D'Agostino Skewness Test

27. Can logistic regression use for more than 2 classes?

No, by default logistic regression is a binary classifier, so it cannot be applied to more than 2 classes. However, it can be extended for solving multi-class classification problems (**multinomial logistic regression**)

28. Explain Correlation and Covariance?

Correlation is used for measuring and also for estimating the quantitative relationship between two variables. Correlation measures how strongly two variables are related. Examples like, income and expenditure, demand and supply, etc.

Covariance is a simple way to measure the correlation between two variables. The problem with covariance is that they are hard to compare without normalization.

29. What is P-value?

P-values are used to make a decision about a hypothesis test. P-value is the minimum significant level at which you can reject the null hypothesis. The lower the p-value, the more likely you reject the null hypothesis.

30. What are Parametric and Non-Parametric Models?

Parametric models will have limited parameters and to predict new data, you only need to know the parameter of the model.

Non-Parametric models have no limits in taking a number of parameters, allowing for more flexibility and to predict new data. You need to know the state of the data and model parameters.

31. What is Reinforcement Learning?

Reinforcement learning is different from the other types of learning like supervised and unsupervised. In reinforcement learning, we are given neither data nor labels. Our learning is based on the rewards given to the agent by the environment.

32. Difference Between Sigmoid and Softmax functions?

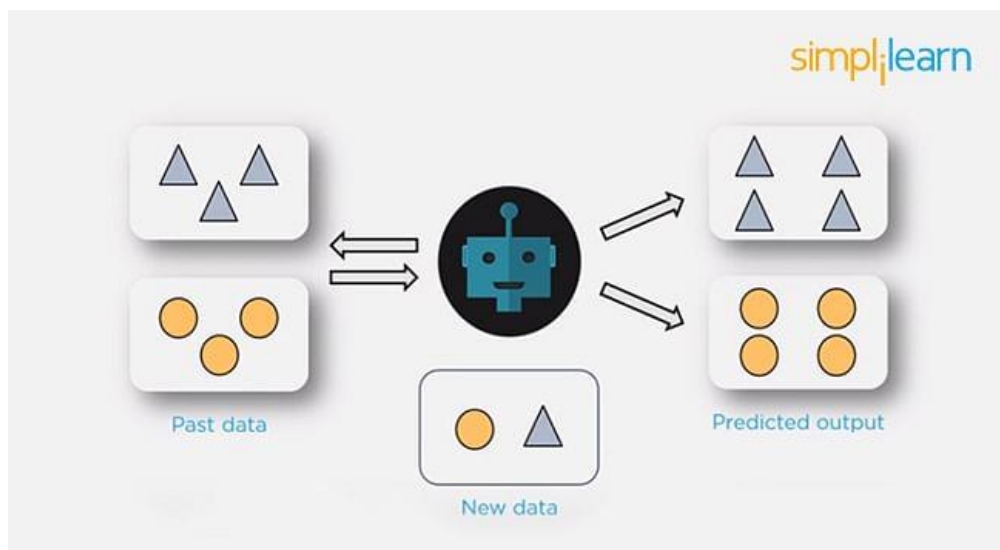
The sigmoid function is used for binary classification. The probabilities sum needs to be 1. Whereas, Softmax function is used for multi-classification. The probabilities sum will be 1.

1. What Are the Different Types of Machine Learning?

There are three types of machine learning:

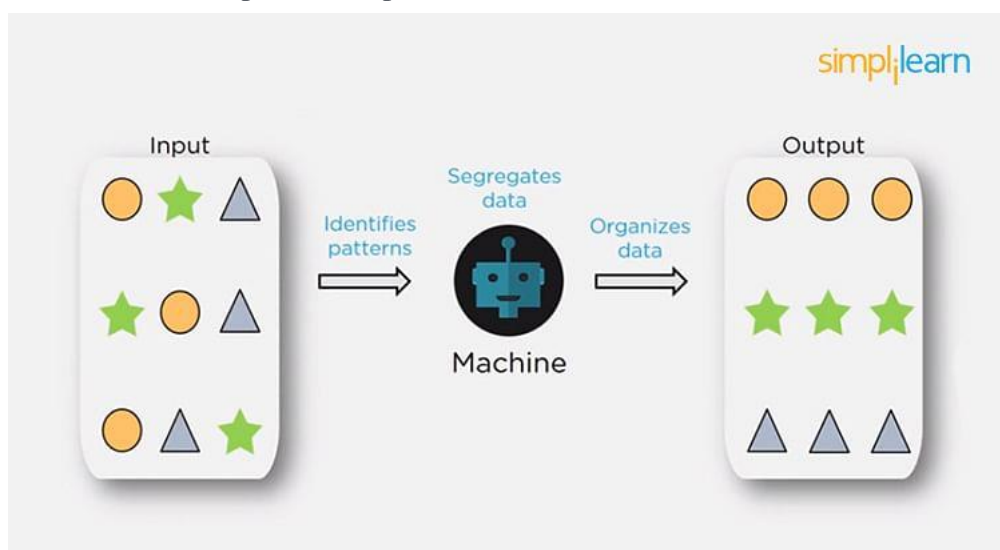
Supervised Learning

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.



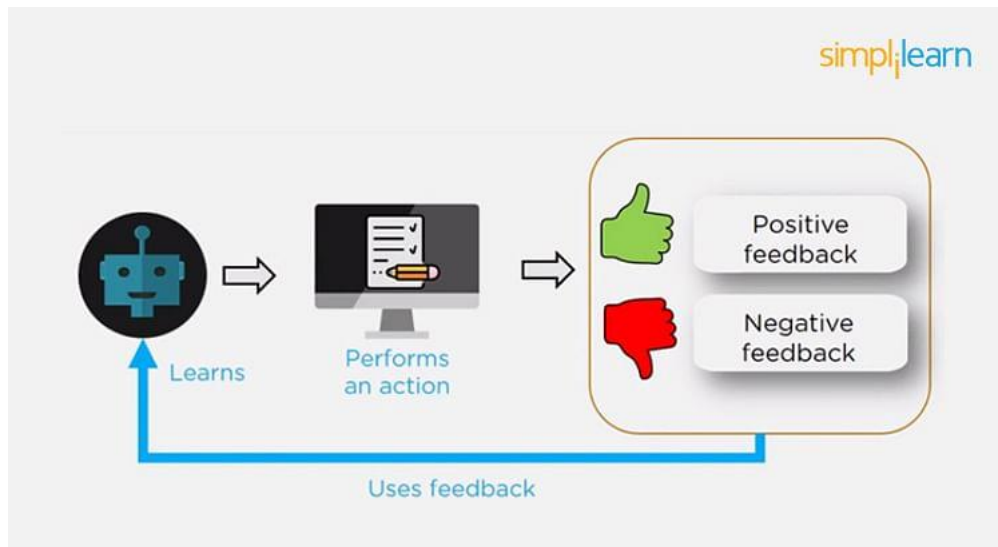
Unsupervised Learning

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.

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

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?

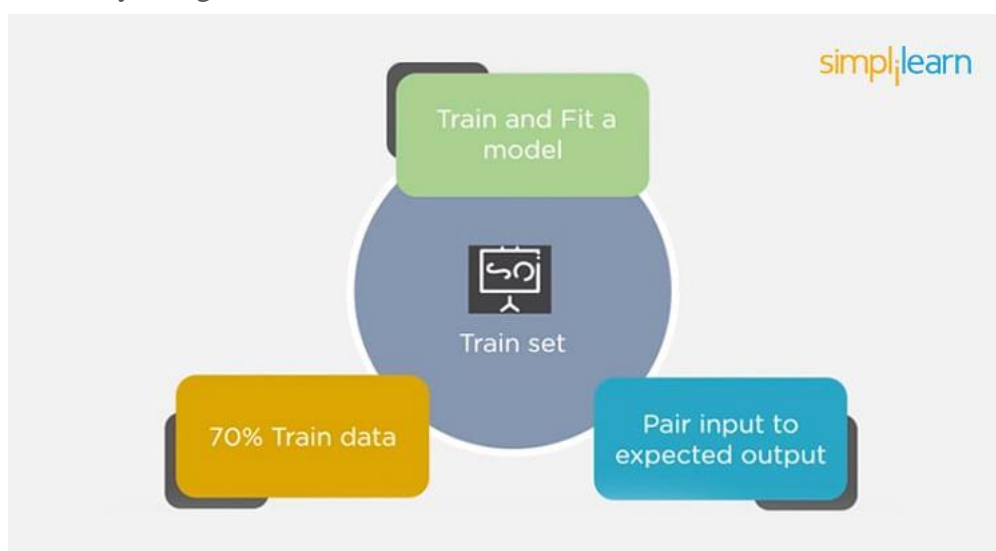
There is a three-step process followed to create a model:

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

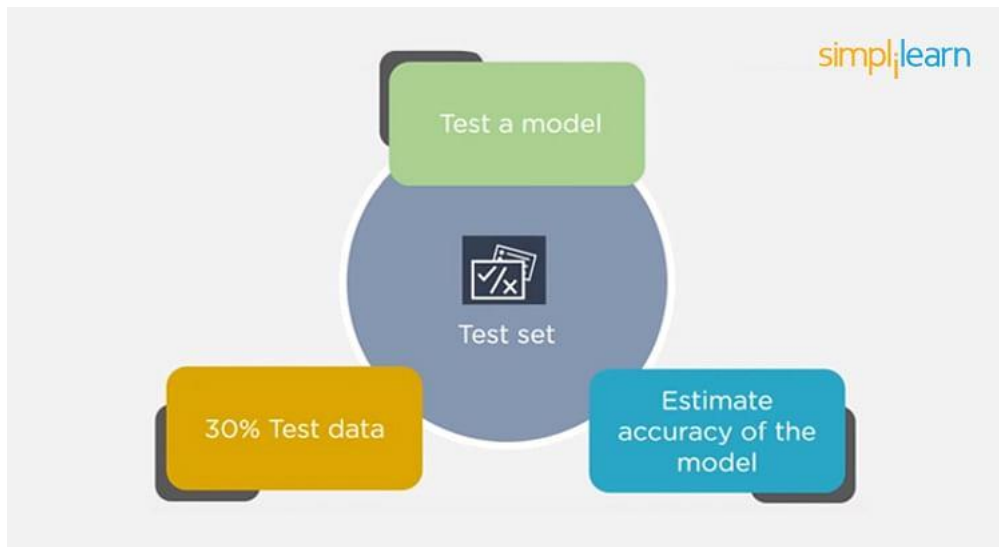
Training Set	Test Set
<p>The training set is examples given to the model to analyze and learn</p> <p>70% of the total data is typically taken as the training dataset</p> <p>This is labeled data used to train the model</p>	<p>The test set is used to test the accuracy of the hypothesis generated by the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

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.



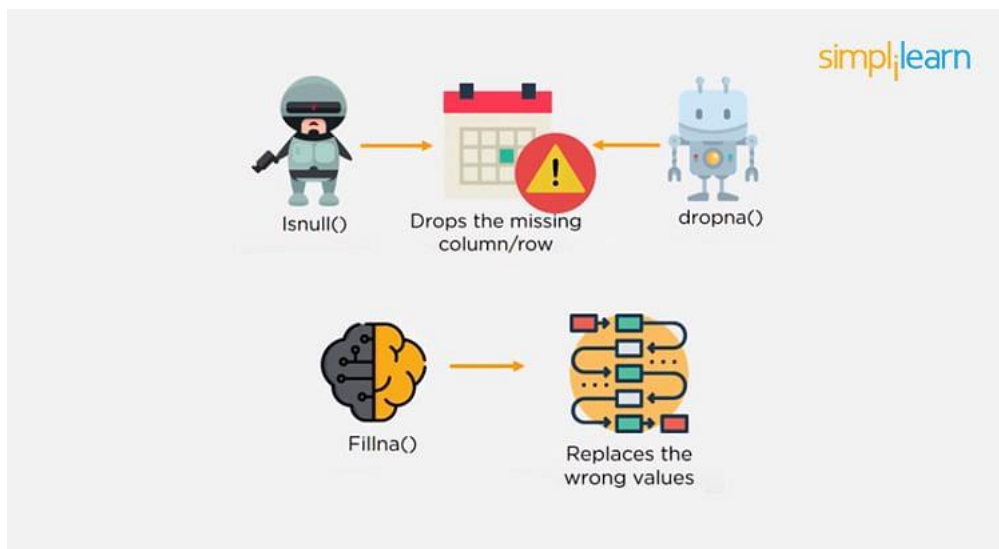
Regarding the question of how to split the data into a training set and test set, there is no fixed rule, and the ratio can vary based on individual preferences.

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.

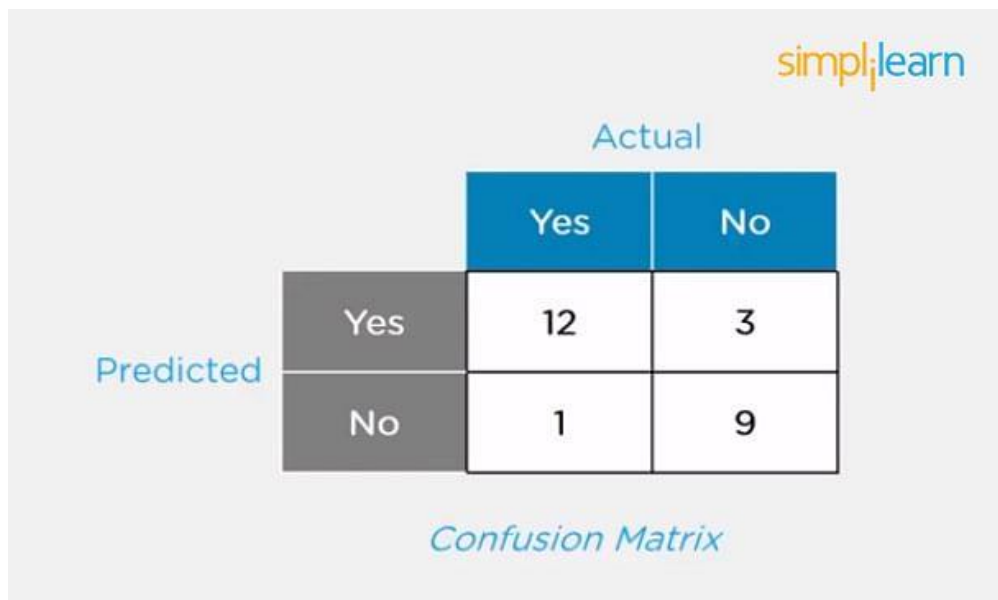
A [confusion matrix](#) (or error matrix) is a specific table that is used to measure the performance of an algorithm. It is mostly used in supervised learning; in unsupervised learning, it's called the matching matrix.

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:

$$\text{Total Yes} = 12 + 1 = 13$$

$$\text{Total No} = 3 + 9 = 12$$

Similarly, for predicted values:

$$\text{Total Yes} = 12 + 3 = 15$$

$$\text{Total No} = 1 + 9 = 10$$

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

$$= (12+9) / 25$$

$$= 21 / 25$$

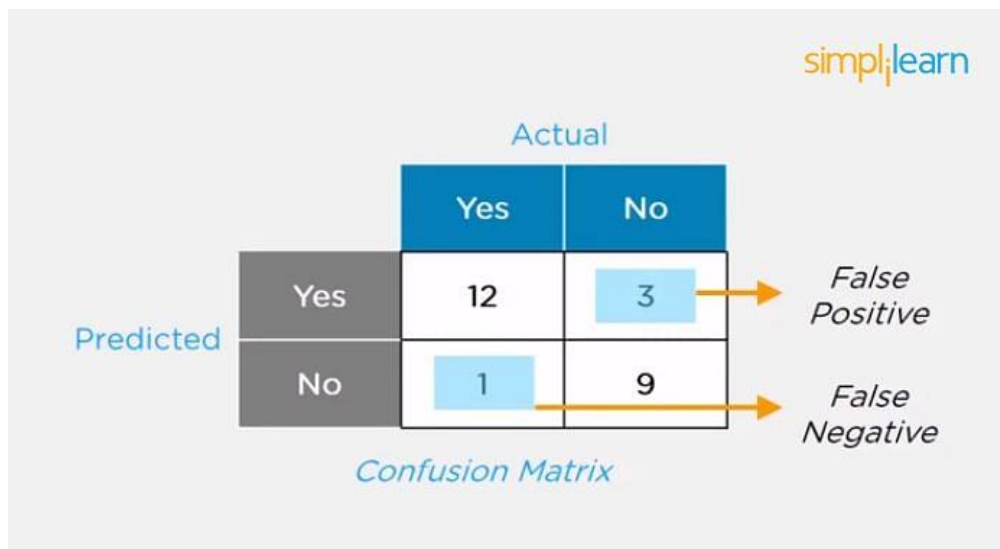
$$= 84\%$$

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.



The diagram shows a confusion matrix with 'Actual' values as columns (Yes, No) and 'Predicted' values as rows (Yes, No). The cells contain counts: (Yes, Yes) = 12, (Yes, No) = 3, (No, Yes) = 1, and (No, No) = 9. An arrow points from the cell (Yes, No) to the label 'False Positive', and another arrow points from the cell (No, Yes) to the label 'False Negative'.

		Actual	
		Yes	No
Predicted	Yes	12	3
	No	1	9

Confusion Matrix

So, looking at the confusion matrix, we get:

$$\text{False-positive} = 3$$

$$\text{True positive} = 12$$

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:

$$\text{False Negative} = 1$$

$$\text{True Negative} = 9$$

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?

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	Enables machines to take decisions with the help of artificial neural networks It needs a large amount of training data

Most features need to be identified in advance and manually coded

The problem is divided into two parts and solved individually and then combined

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.

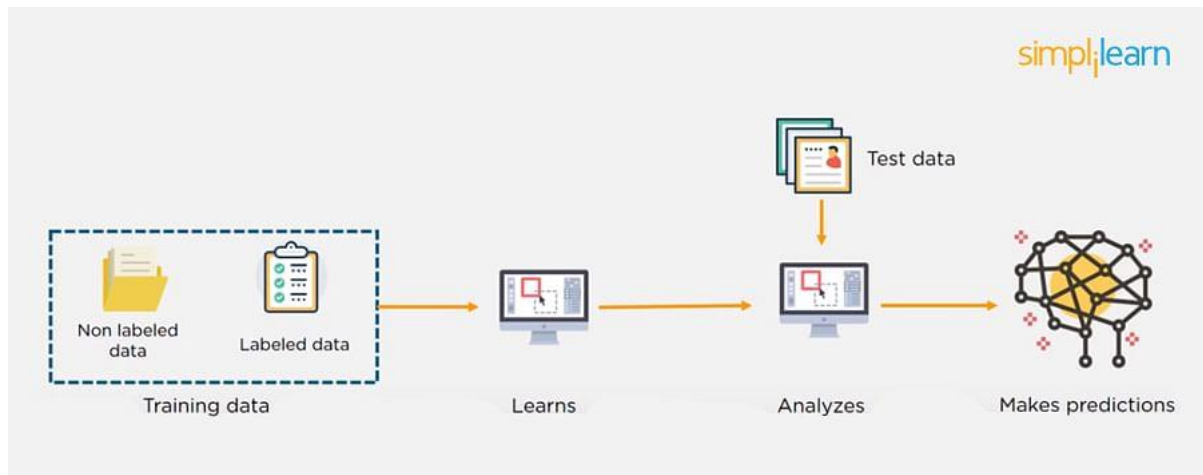
- **Fraud Detection**

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

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.

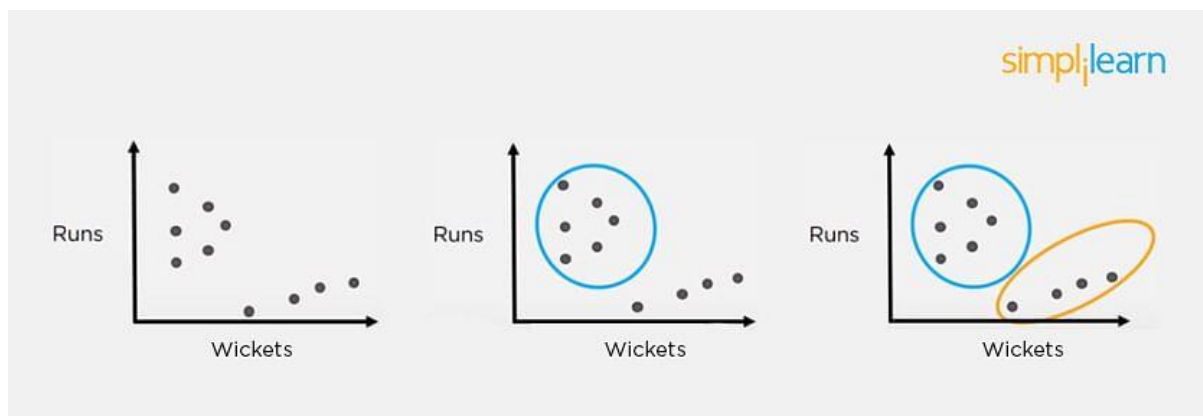


13. What Are Unsupervised Machine Learning Techniques?

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

Clustering

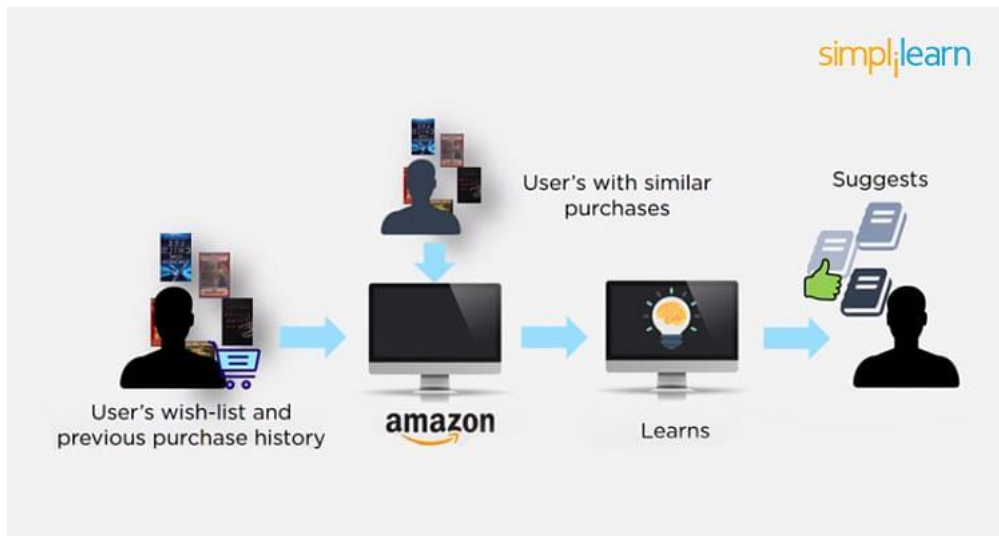
Clustering problems involve data to be divided into subsets. These subsets, also called clusters, contain data that are similar to each other. Different clusters reveal different details about the objects, unlike classification or regression.



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.



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
<p>It observes instances based on defined principles to draw a conclusion</p> <p>Example: Explaining to a child to keep away from the fire by showing a video where fire causes damage</p>	<p>It concludes experiences</p> <p>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</p>

16. Compare K-means and KNN Algorithms.

K-means	KNN
K-Means is unsupervised	KNN is supervised in nature

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 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.

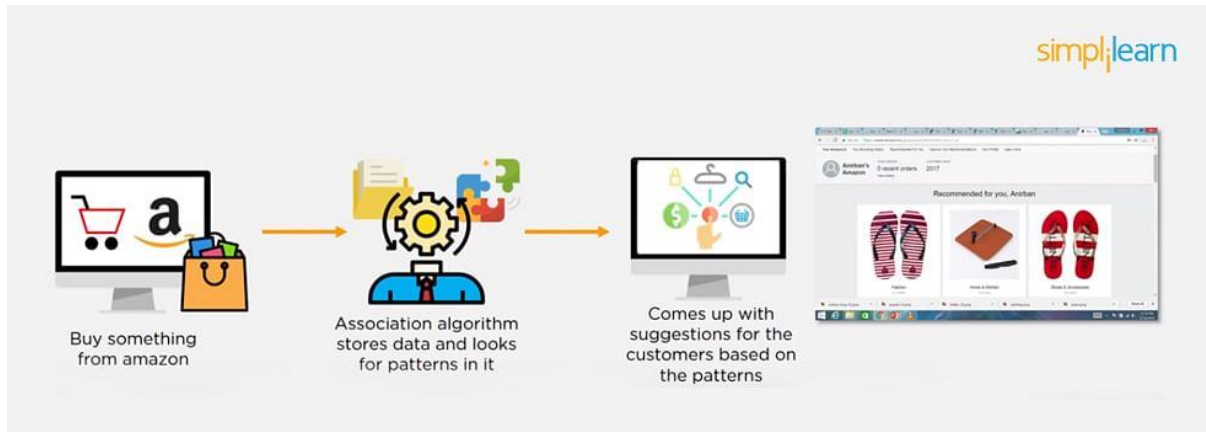
19. How Will You Know Which Machine Learning Algorithm to Choose for Your Classification Problem?

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.



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:

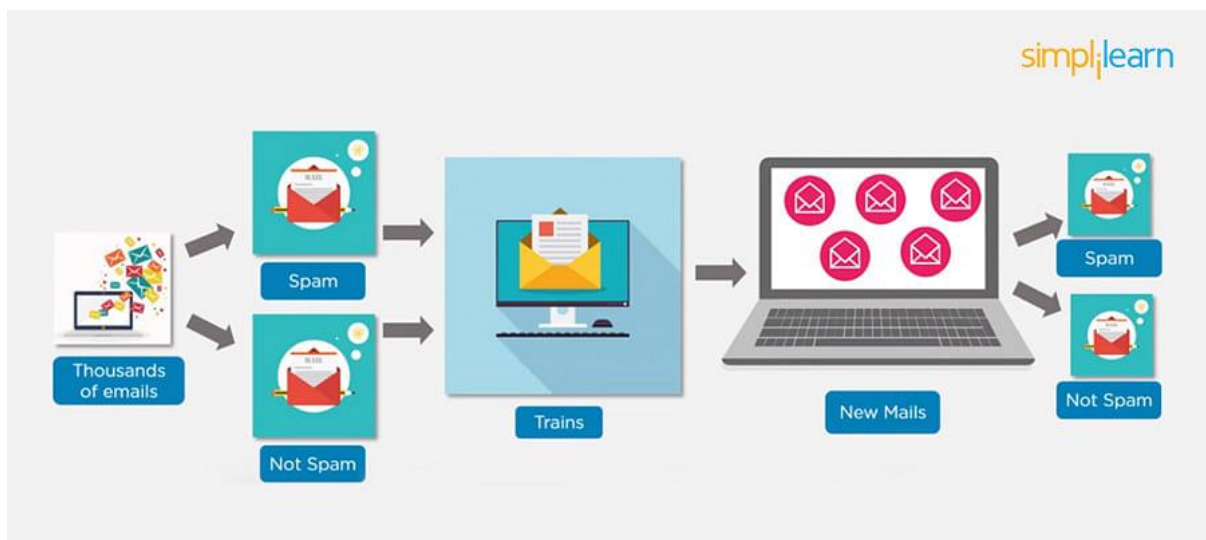
- Estimating sales and price of a product
- Predicting the score of a team
- Predicting the amount of rainfall

22. How Do You Design an Email Spam Filter?

Building a spam filter involves the following process:

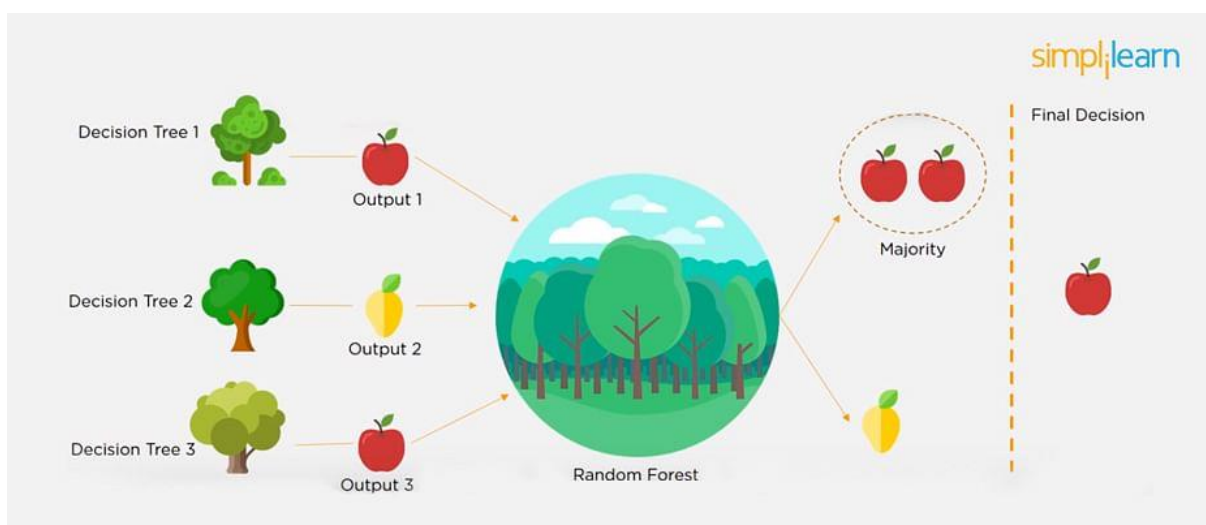
- The email spam filter will be fed with thousands of emails
- Each of these emails already has a label: 'spam' or 'not spam.'

- 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.

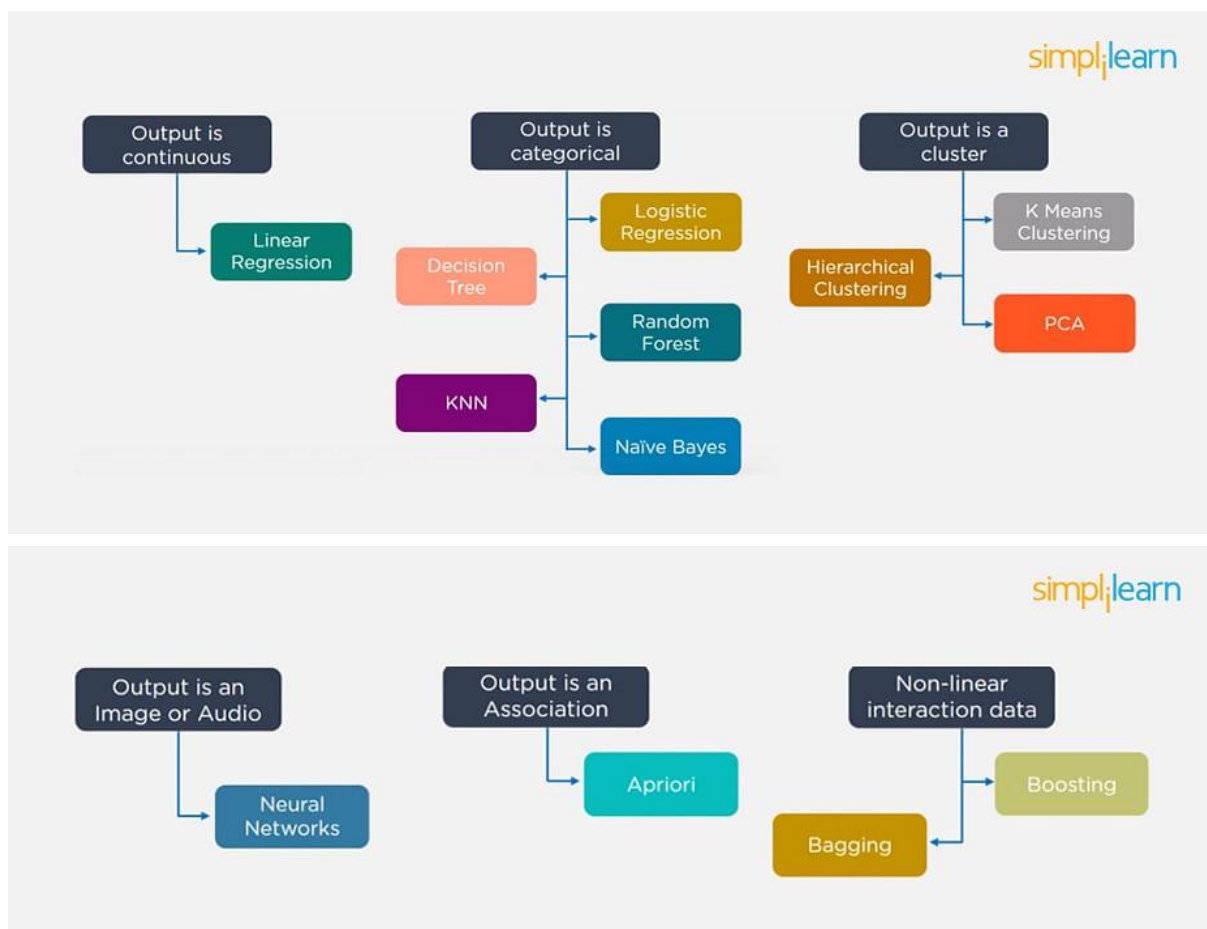


24. Considering a Long List of Machine Learning Algorithms, given a Data Set, How Do You Decide Which One to Use?

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).

$\text{Precision} = (\text{True Positive}) / (\text{True Positive} + \text{False Positive})$

Recall

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

$\text{Recall} = (\text{True Positive}) / (\text{True Positive} + \text{False Negative})$

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

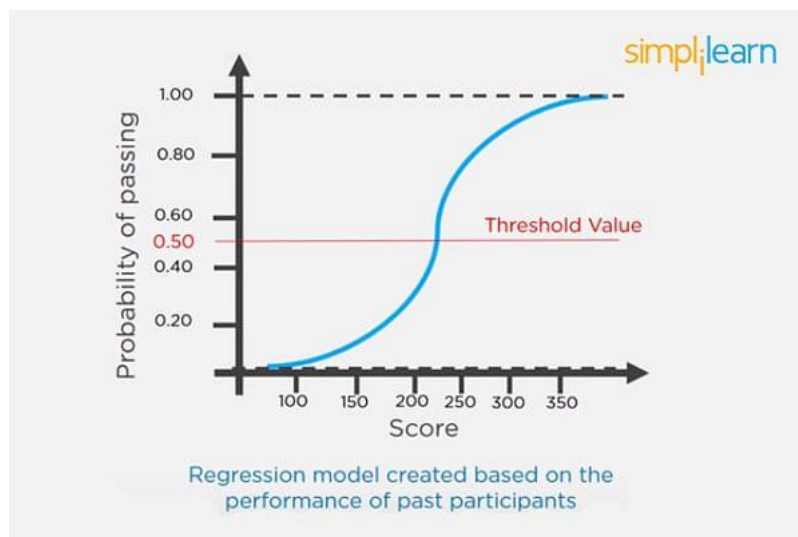
There is a popular pruning algorithm called reduced error pruning, in which:

- 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.



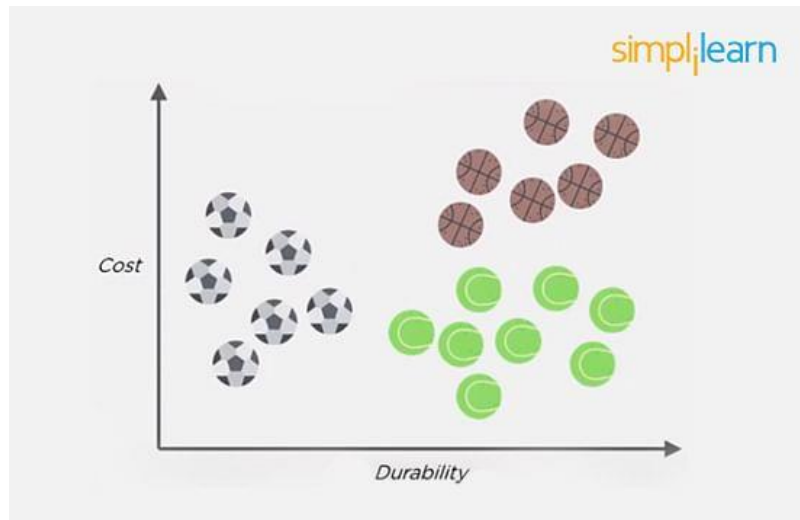
31. Explain the K Nearest Neighbor Algorithm.

[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.

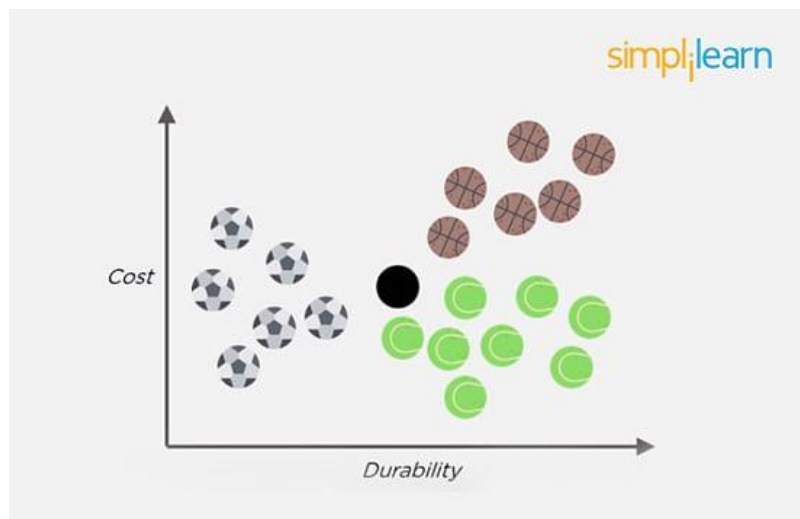
Let us classify an object using the following example. Consider there are three clusters:

- 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:

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

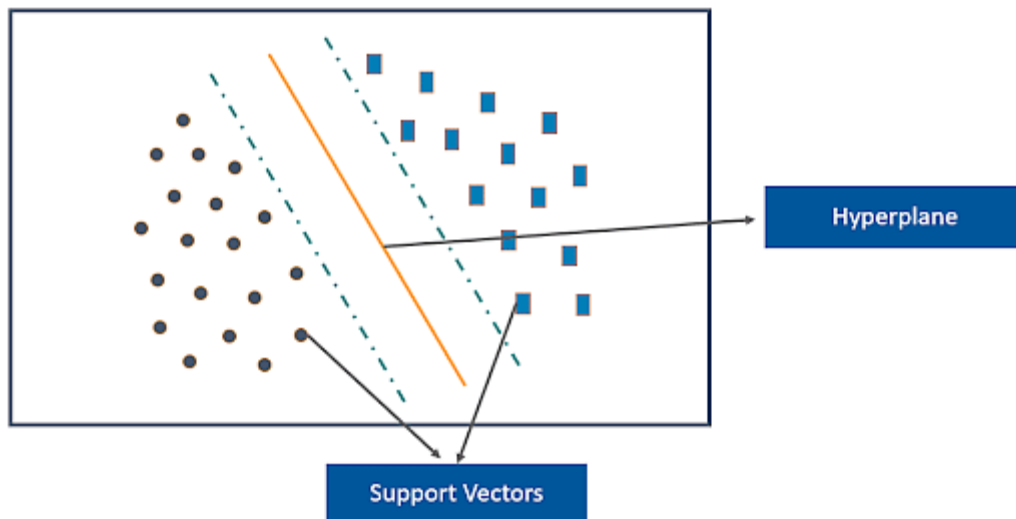
Covariance: Covariance tells us the direction of the linear relationship between two random variables. It can take any value between $-\infty$ and $+\infty$.

Formula to calculate Covariance:

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

39. What are Support Vectors in SVM?

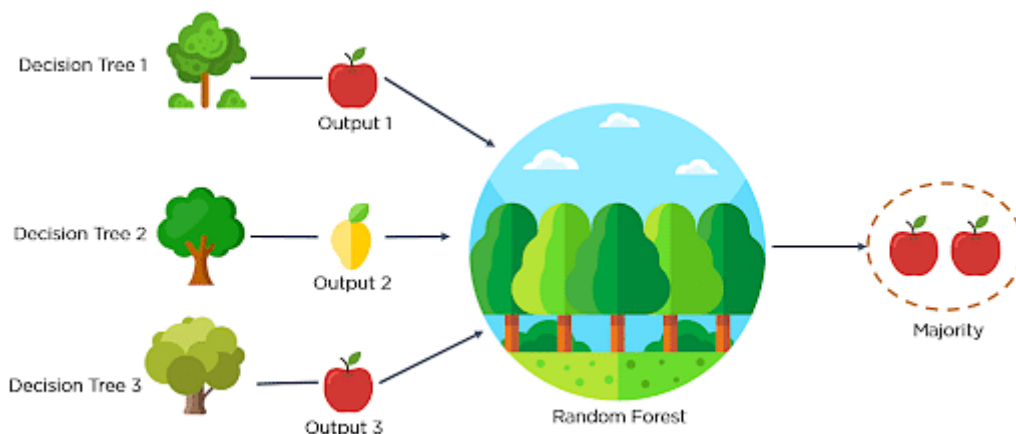
Support Vectors are data points that are nearest to the hyperplane. It influences the position and orientation of the hyperplane. Removing the support vectors will alter the position of the hyperplane. The support vectors help us build our support vector machine model.



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.

$$\text{IG} = 1 - \text{Entropy}$$
$$\text{Entropy} = - \sum p_i \log_2 p_i$$

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.