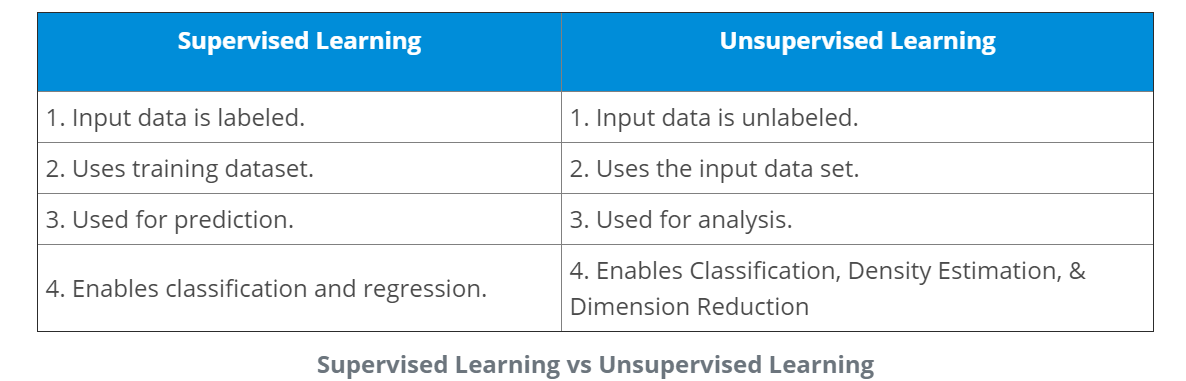
**Top Data Science Interview Questions**

### ****Q1: What is Data Science? Also, list the differences between supervised and unsupervised learning.****

Answer: Data Science is a blend of various tools, algorithms, and machine learning principles with the goal to discover hidden patterns from the raw data. How is this different from what statisticians have been doing for years?

The answer lies in the difference between explaining and predicting.

### 



#### Supervised Machine learning:

Supervised machine learning requires training labeled data.

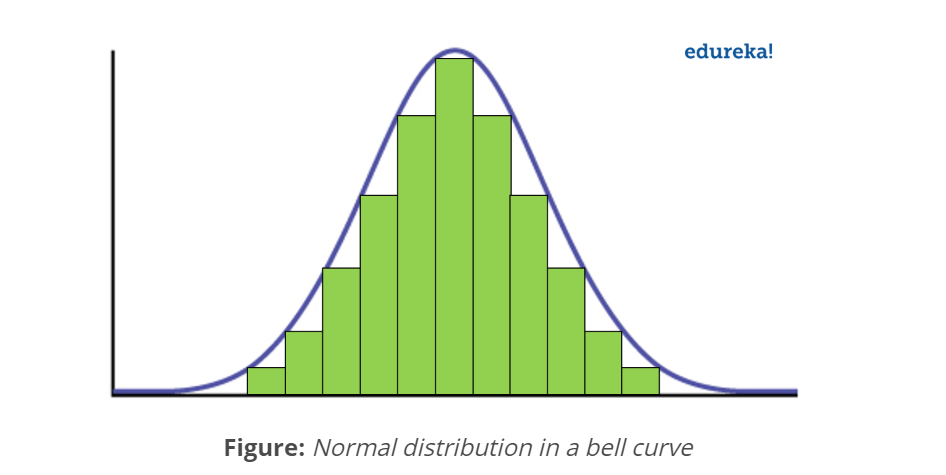
#### Unsupervised Machine learning:

Unsupervised machine learning doesn’t required labeled data.

### ****Q2: What do you understand by the term Normal Distribution?****

Answer: Data is usually distributed in different ways with a bias to the left or to the right or it can all be jumbled up.

However, there are chances that data is distributed around a central value without any bias to the left or right and reaches normal distribution in the form of a bell-shaped curve.



The random variables are distributed in the form of a symmetrical bell-shaped curve.

Properties of Normal Distribution:

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

### ****Q3: What do you understand by statistical power of sensitivity and how do you calculate it?****

Sensitivity is commonly used to validate the accuracy of a classifier (Logistic, SVM, Random Forest etc.).

Sensitivity is nothing but “Predicted True events/ Total events”. True events here are the events which were true and model also predicted them as true. Calculation of seasonality is pretty straightforward.

***Seasonality= (True Positives) / (Positives in Actual Dependent Variable)***

\*where true positives are positive events which are correctly classified as positives.

### ****Q4: What are the differences between overfitting and underfitting?****

In statistics and machine learning, one of the most common tasks is to fit a *model* to a set of training data, so as to be able to make reliable predictions on general untrained data.

In *overfitting*, a statistical model describes random error or noise instead of the underlying relationship. Overfitting occurs when a model is excessively complex, such as having too many parameters relative to the number of observations. A model that has been overfit has poor predictive performance, as it overreacts to minor fluctuations in the training data.

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

### ****Q5: Python or R – Which one would you prefer for text analytics?****

We will prefer Python because of the following reasons:

* Python would be the best option because it has Pandas library that provides easy to use data structures and high-performance data analysis tools.
* R is more suitable for machine learning than just text analysis.
* Python performs faster for all types of text analytics.

### ****Q6: How does data cleaning plays a vital role in analysis?****

Data cleaning can help in analysis because:

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

### ****Q7: Differentiate between univariate, bivariate and multivariate analysis.****

***Univariate*** *analyses* are descriptive statistical analysis techniques which can be differentiated based on the number of variables involved at a given point of time. For example, the pie charts of sales based on territory involve only one variable and can the analysis can be referred to as univariate analysis.

*The****bivariate****analysis* attempts to understand the difference between two variables at a time as in a scatterplot. For example, analyzing the volume of sale and spending can be considered as an example of bivariate analysis.

***Multivariate analysis*** deals with the study of more than two variables to understand the effect of variables on the responses.

### ****Q8: What is Cluster Sampling?****

*Cluster sampling* is a technique used when it becomes difficult to study the target population spread across a wide area and simple random sampling cannot be applied. Cluster Sample is a probability sample where each sampling unit is a collection or cluster of elements.

For eg., A researcher wants to survey the academic performance of high school students in Japan. He can divide the entire population of Japan into different clusters (cities). Then the researcher selects a number of clusters depending on his research through simple or systematic random sampling.

Let’s continue our Data Science Interview Questions blog with some more statistics questions.

### ****Q9: What is Systematic Sampling?****

Systematic sampling is a statistical technique where elements are selected from an ordered sampling frame. In systematic sampling, the list is progressed in a circular manner so once you reach the end of the list, it is progressed from the top again. The best example of systematic sampling is equal probability method.

### ****Q10: What are Eigenvectors and Eigenvalues?****

*Eigenvectors* are used for understanding linear transformations. In data analysis, we usually calculate the eigenvectors for a correlation or covariance matrix. Eigenvectors are the directions along which a particular linear transformation acts by flipping, compressing or stretching.

*Eigenvalue* can be referred to as the strength of the transformation in the direction of eigenvector or the factor by which the compression occurs.

### ****Q11: Can you cite some examples where a false positive is important than a false negative?****

Let us first understand what false positives and false negatives are.

* False Positives are the cases where you wrongly classified a non-event as an event a.k.a Type I error.
* False Negatives are the cases where you wrongly classify events as non-events, a.k.a Type II error.

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

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

### ****Q12: Can you cite some examples where a false negative important than a false positive?****

*Example 1*: Assume there is an airport ‘A’ which has received high-security threats and based on certain characteristics they identify whether a particular passenger can be a threat or not. Due to a shortage of staff, they decide to scan passengers being predicted as risk positives by their predictive model. What will happen if a true threat customer is being flagged as non-threat by airport model?

*Example 2*: What if Jury or judge decides to make a criminal go free?

*Example 3*: What if you rejected to marry a very good person based on your predictive model and you happen to meet him/her after a few years and realize that you had a false negative?

### ****Q13: Can you cite some examples where both false positive and false negatives are equally important?****

In the Banking industry giving loans is the primary source of making money but at the same time if your repayment rate is not good you will not make any profit, rather you will risk huge losses.

Banks don’t want to lose good customers and at the same point in time, they don’t want to acquire bad customers. In this scenario, both the false positives and false negatives become very important to measure.

### ****Q14: Can you explain the difference between a Validation Set and a Test Set?****

A*Validation set* can be considered as a part of the training set as it is used for parameter selection and to avoid overfitting of the model being built.

On the other hand, a *Test Set* is used for testing or evaluating the performance of a trained machine learning model.

In simple terms, the differences can be summarized as; training set is to fit the parameters i.e. weights and test set is to assess the performance of the model i.e. evaluating the predictive power and generalization.

### ****Q15: Explain cross-validation.****

Cross-validation is a model validation technique for evaluating how the outcomes of statistical analysis will generalize to an Independent dataset. Mainly used in backgrounds where the objective is forecast and one wants to estimate how accurately a model will accomplish in practice.

The goal of cross-validation is to term a data set to test the model in the training phase (i.e. validation data set) in order to limit problems like overfitting and get an insight on how the model will generalize to an independent data set.

### ****Q16: What is Machine Learning?****

Machine Learning explores the study and construction of algorithms that can learn from and make predictions on data. Closely related to computational statistics. Used to devise complex models and algorithms that lend themselves to a prediction which in commercial use is known as predictive analytics.



### ****Q17: What is logistic regression? State an example when you have used logistic regression recently.****

Logistic Regression often referred as logit model is a technique to predict the binary outcome from a linear combination of predictor variables.

For example, if you want to predict whether a particular political leader will win the election or not. In this case, the outcome of prediction is binary i.e. 0 or 1 (Win/Lose). The predictor variables here would be the amount of money spent for election campaigning of a particular candidate, the amount of time spent in campaigning, etc.

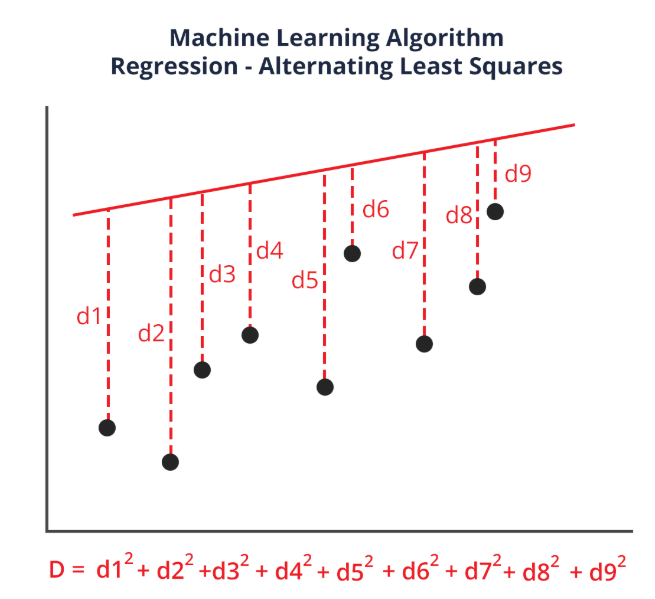
### ****Q18: What are Recommender Systems?****

[Recommender Systems](https://www.edureka.co/blog/videos/science-behind-product-recommendation-with-r-programming/) are a subclass of information filtering systems that are meant to predict the preferences or ratings that a user would give to a product. Recommender systems are widely used in movies, news, research articles, products, social tags, music, etc.

Examples include movie recommenders in IMDB, Netflix & BookMyShow, product recommenders in e-commerce sites like Amazon, eBay & Flipkart, YouTube video recommendations and game recommendations in Xbox.

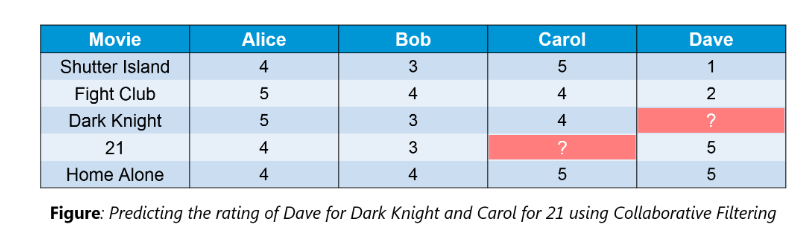
### ****Q19: What is Linear Regression?****

### [**Linear regression**](https://www.edureka.co/blog/linear-regression-in-python/) is a statistical technique where the score of a variable Y is predicted from the score of a second variable X. X is referred to as the predictor variable and Y as the criterion variable.



### ****Q20: What is Collaborative filtering?****

The process of filtering used by most of the recommender systems to find patterns or information by collaborating viewpoints, various data sources and multiple agents.



An example of collaborative filtering can be to predict the rating of a particular user based on his/her ratings for other movies and others’ ratings for all movies. This concept is widely used in recommending movies in IMDB, Netflix & BookMyShow, product recommenders in e-commerce sites like Amazon, eBay & Flipkart, YouTube video recommendations and game recommendations in Xbox.

### ****Q21: How can outlier values be treated?****

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

All extreme values are not outlier values. The most common ways to treat outlier values

1. To change the value and bring in within a range.
2. To just remove the value.

### ****Q22: What are the various steps involved in an analytics project?****

The following are the various steps involved in an analytics project:

1. Understand the Business problem
2. Explore the data and become familiar with it.
3. Prepare the data for modeling by detecting outliers, treating missing values, transforming variables, etc.
4. After data preparation, start running the model, analyze the result and tweak the approach. This is an iterative step until the best possible outcome is achieved.
5. Validate the model using a new data set.
6. Start implementing the model and track the result to analyze the performance of the model over the period of time.

### ****Q23: During analysis, how do you treat missing values?****

The extent of the missing values is identified after identifying the variables with missing values. If any patterns are identified the analyst has to concentrate on them as it could lead to interesting and meaningful business insights.

If there are no patterns identified, then the missing values can be substituted with mean or median values (imputation) or they can simply be ignored. Assigning a default value which can be mean, minimum or maximum value. Getting into the data is important.

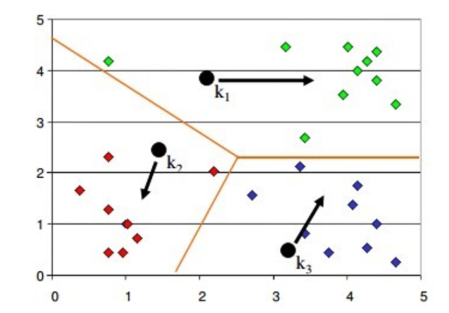
If it is a categorical variable, the default value is assigned. The missing value is assigned a default value. If you have a distribution of data coming, for normal distribution give the mean value.

If 80% of the values for a variable are missing then you can answer that you would be dropping the variable instead of treating the missing values.

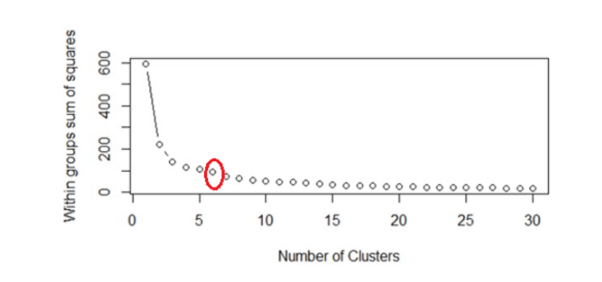
### ****Q24: How will you define the number of clusters in a clustering algorithm?****

Though the Clustering Algorithm is not specified, this question is mostly in reference to K-Means clustering where “K” defines the number of clusters. The objective of clustering is to group similar entities in a way that the entities within a group are similar to each other but the groups are different from each other.

For example, the following image shows three different groups.



Within Sum of squares is generally used to explain the homogeneity within a cluster. If you plot WSS for a range of number of clusters, you will get the plot shown below.



* The Graph is generally known as Elbow Curve.
* Red circled point in above graph i.e. Number of Cluster =6 is the point after which you don’t see any decrement in WSS.
* This point is known as the **bending** point and taken as K in K – Means.

This is the widely used approach but few data scientists also use Hierarchical clustering first to create dendrograms and identify the distinct groups from there.

Now that we have seen the Machine Learning Questions, Let’s continue our Data Science Interview Questions blog with some Probability questions.

### ****Q25: What do you mean by Deep Learning and Why has it become popular now?****

Deep Learning is nothing but a paradigm of machine learning which has shown incredible promise in recent years. This is because of the fact that Deep Learning shows a great analogy with the functioning of the human brain.

Now although Deep Learning has been around for many years, the major breakthroughs from these techniques came just in recent years. This is because of two main reasons:

* The increase in the amount of data generated through various sources
* The growth in hardware resources required to run these models

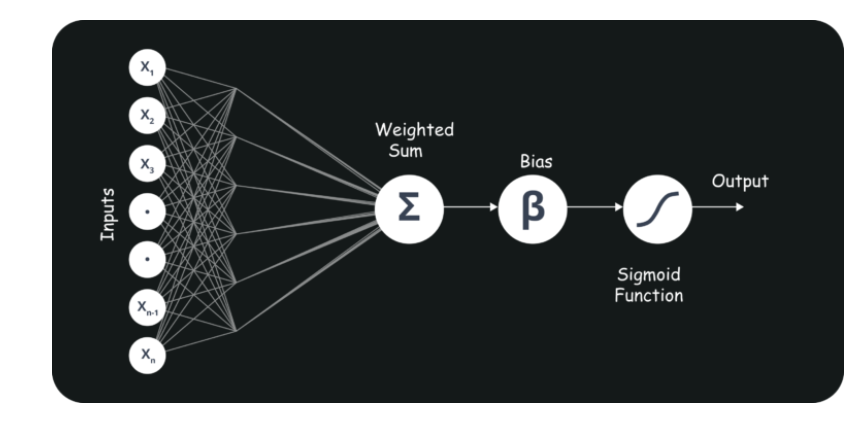
GPUs are multiple times faster and they help us build bigger and deeper deep learning models in comparatively less time than we required previously

### ****Q26: What are Artificial Neural Networks?****

Artificial Neural networks are a specific set of algorithms that have revolutionized machine learning. They are inspired by biological neural networks. [**Neural Networks**](https://www.edureka.co/blog/neural-network-tutorial/) can adapt to changing input so the network generates the best possible result without needing to redesign the output criteria.

### ****Q27: Describe the structure of Artificial Neural Networks?****

Artificial Neural Networks works on the same principle as a biological Neural Network. It consists of inputs which get processed with weighted sums and Bias, with the help of Activation Functions.



### ****Q28: Explain Gradient Descent.****

To Understand Gradient Descent, Let’s understand what is a Gradient first.

A **gradient** measures how much the output of a function changes if you change the inputs a little bit. It simply measures the change in all weights with regard to the change in error. You can also think of a gradient as the slope of a function.

Gradient Descent can be thought of climbing down to the bottom of a valley, instead of climbing up a hill.  This is because it is a minimization algorithm that minimizes a given function (**Activation Function**).

### ****Q29: What is Back Propagation and Explain it’s Working.****

[Backpropagation](https://www.edureka.co/blog/backpropagation/) is a training algorithm used for multilayer neural network. In this method, we move the error from an end of the network to all weights inside the network and thus allowing efficient computation of the gradient.

It has the following steps:

* Forward Propagation of Training Data
* Derivatives are computed using output and target
* Back Propagate for computing derivative of error wrt output activation
* Using previously calculated derivatives for output
* Update the Weights

### ****Q30: What are the variants of Back Propagation?****

* Stochastic Gradient Descent: We use only single training example for calculation of gradient and update parameters.
* Batch Gradient Descent: We calculate the gradient for the whole dataset and perform the update at each iteration.
* Mini-batch Gradient Descent: It’s one of the most popular optimization algorithms. It’s a variant of Stochastic Gradient Descent and here instead of single training example, mini-batch of samples is used.

### ****Q31: What are the different Deep Learning Frameworks?****

* Pytorch
* TensorFlow
* Microsoft Cognitive Toolkit
* Keras
* Caffe
* Chainer

### ****Q32: What is the role of Activation Function?****

### The Activation function is used to introduce non-linearity into the neural network helping it to learn more complex function. Without which the neural network would be only able to learn linear function which is a linear combination of its input data. An activation function is a function in an artificial neuron that delivers an output based on inputs

### ****Q33: What is an Auto-Encoder?****

### [**Autoencoders**](https://www.edureka.co/blog/autoencoders-tutorial/) are simple learning networks that aim to transform inputs into outputs with the minimum possible error. This means that we want the output to be as close to input as possible. We add a couple of layers between the input and the output, and the sizes of these layers are smaller than the input layer. The autoencoder receives unlabeled input which is then encoded to reconstruct the input.

### ****Q34: What is a Boltzmann Machine?****

Boltzmann machines have a simple learning algorithm that allows them to discover interesting features that represent complex regularities in the training data. The Boltzmann machine is basically used to optimize the weights and the quantity for the given problem. The learning algorithm is very slow in networks with many layers of feature detectors. “[**Restricted Boltzmann Machines**](https://www.edureka.co/blog/restricted-boltzmann-machine-tutorial/)**”** algorithm has a single layer of feature detectors which makes it faster than the rest.

### ****Q35: What is Selection Bias?****

Selection bias is a kind of error that occurs when the researcher decides who is going to be studied. It is usually associated with research where the selection of participants isn’t random. It is sometimes referred to as the selection effect. It is the distortion of statistical analysis, resulting from the method of collecting samples. If the selection bias is not taken into account, then some conclusions of the study may not be accurate.

The types of selection bias include:

1. **Sampling bias**: It is a systematic error due to a non-random sample of a population causing some members of the population to be less likely to be included than others resulting in a biased sample.
2. **Time interval**: A trial may be terminated early at an extreme value (often for ethical reasons), but the extreme value is likely to be reached by the variable with the largest variance, even if all variables have a similar mean.
3. **Data**: When specific subsets of data are chosen to support a conclusion or rejection of bad data on arbitrary grounds, instead of according to previously stated or generally agreed criteria.
4. **Attrition**: Attrition bias is a kind of selection bias caused by attrition (loss of participants) discounting trial subjects/tests that did not run to completion.

### Q36: What is bias, variance trade off?

#### Bias:

“Bias is error introduced in your model due to over simplification of machine learning algorithm.” It can lead to underfitting. When you train your model at that time model makes simplified assumptions to make the target function easier to understand.

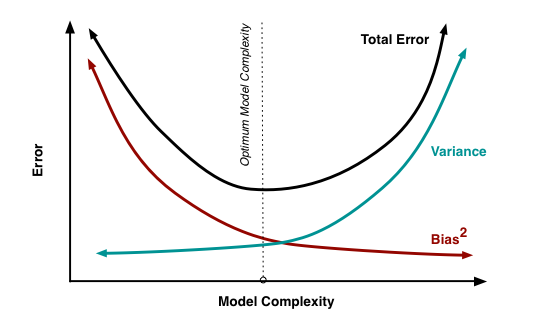
##### Low bias machine learning algorithms - Decision Trees, k-NN and SVM

##### Hight bias machine learning algorithms - Liear Regression, Logistic Regression

#### Variance:

“Variance is error introduced in your model due to complex machine learning algorithm, your model learns noise also from the training dataset and performs bad on test dataset.” It can lead high sensitivity and overfitting.

Normally, as you increase the complexity of your model, you will see a reduction in error due to lower bias in the model. However, this only happens till a particular point. As you continue to make your model more complex, you end up over-fitting your model and hence your model will start suffering from high variance.



#### Bias, Variance trade off:

The goal of any supervised machine learning algorithm is to have low bias and low variance to achive good prediction performance.

1. The k-nearest neighbors algorithm has low bias and high variance, but the trade-off can be changed by increasing the value of k which increases the number of neighbors that contribute to the prediction and in turn increases the bias of the model.
2. The support vector machine algorithm has low bias and high variance, but the trade-off can be changed by increasing the C parameter that influences the number of violations of the margin allowed in the training data which increases the bias but decreases the variance.

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

Increasing the bias will decrease the variance. Increasing the variance will decrease the bias.

### Q37: What are the important skills to have in Python with regard to data analysis?

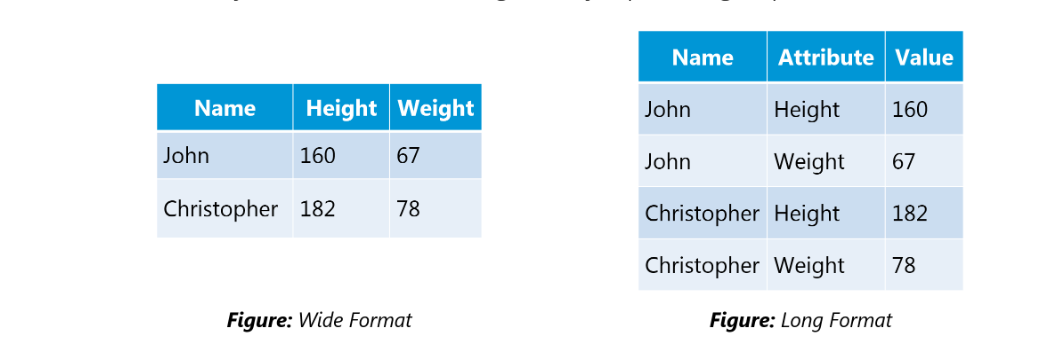
The following are some of the important skills to possess which will come handy when performing data analysis using Python.

* Good understanding of the built-in data types especially lists, dictionaries, tuples, and sets.
* Mastery of N-dimensional [NumPy Arrays](https://www.edureka.co/blog/python-numpy-tutorial/" \t "_blank).
* Mastery of [Pandas](https://www.edureka.co/blog/python-pandas-tutorial/) dataframes.
* Ability to perform element-wise vector and matrix operations on NumPy arrays.
* Knowing that you should use the Anaconda distribution and the conda package manager.
* Familiarity with [Scikit-learn](https://www.edureka.co/blog/scikit-learn-machine-learning/" \t "_blank). \*\***[Scikit-Learn Cheat Sheet](https://www.edureka.co/blog/cheatsheets/python-scikit-learn-cheat-sheet" \t "_blank)**\*\*
* Ability to write efficient list comprehensions instead of traditional for loops.
* Ability to write small, clean functions (important for any developer), preferably pure functions that don’t alter objects.
* Knowing how to profile the performance of a Python script and how to optimize bottlenecks.

The following will help to tackle any problem in data analytics and machine learning.

### Q38: What is the difference between “long” and “wide” format data?

In the **wide** format, a subject’s repeated responses will be in a single row, and each response is in a separate column. In the **long** format, each row is a one-time point per subject. You can recognize data in wide format by the fact that columns generally represent groups.



### Q39: What is exploding gradients?

“Exploding gradients are a problem where **large error gradients** accumulate and result in very large updates to neural network model weights during training.” At an extreme, the values of weights can become so large as to overflow and result in NAN values.

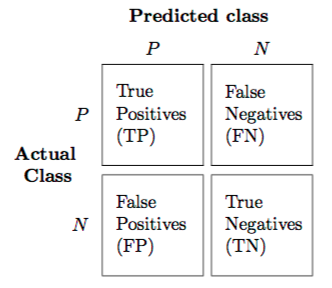
This has the effect of your model being unstable and unable to learn from your training data. Now let’s understand what the gradient is.

#### Gradient:

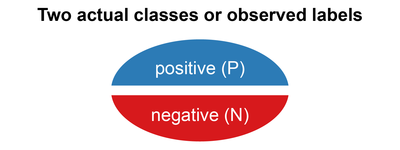
Gradient is the **direction and magnitude** calculated during training of a neural network that is used to update the network weights in the right direction and by the right amount.

### Q40: What is a confusion matrix?

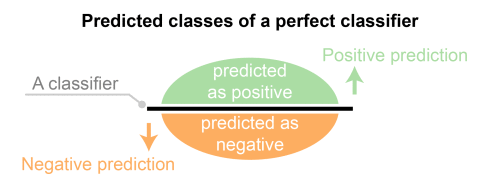
The confusion matrix is a 2X2 table that contains 4 outputs provided by the **binary classifier**. Various measures, such as error-rate, accuracy, specificity, sensitivity, precision and recall are derived from it. Confusion Matrix



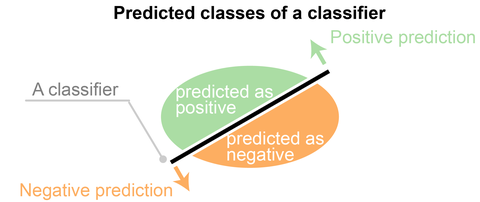
A dataset used for performance evaluation is called test dataset. It should contain the correct labels and predicted labels.



The predicted labels will exactly the same if the performance of a binary classifier is perfect.

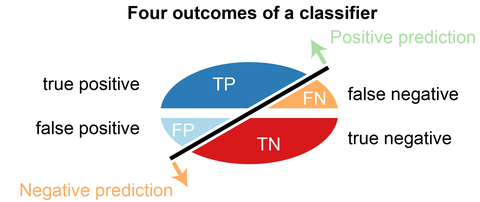


The predicted labels usually match with part of the observed labels in real world scenarios.



A binary classifier predicts all data instances of a test dataset as either positive or negative. This produces four outcomes-

1. True positive(TP) - Correct positive prediction
2. False positive(FP) - Incorrect positive prediction
3. True negative(TN) - Correct negative prediction
4. False negative(FN) - Incorrect negative prediction

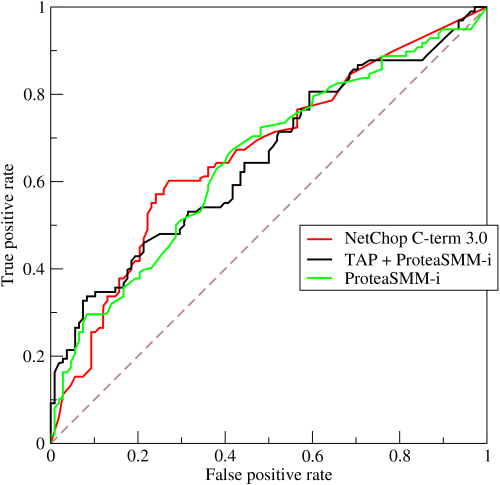


**Basic measures derived from the confusion matrix**

1. Error Rate = (FP+FN)/(P+N)
2. Accuracy = (TP+TN)/(P+N)
3. Sensitivity(Recall or True positive rate) = TP/P
4. Specificity(True negative rate) = TN/N
5. Precision(Positive predicted value) = TP/(TP+FP)
6. F-Score (Harmonic mean of precision and recall) = (1+b)(PREC.REC)/(b^2PREC+REC) where b is commonly 0.5, 1, 2.

### Q41: Explain how a ROC curve works?

The **ROC** curve is a graphical representation of the contrast between true positive rates and false positive rates at various thresholds. It is often used as a proxy for the trade-off between the sensitivity(true positive rate) and false positive rate.



### Q42: What is selection Bias?

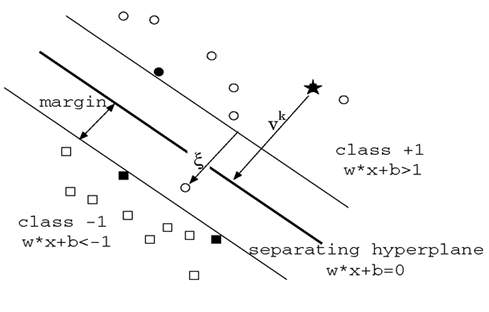
Selection bias occurs when sample obtained is not representative of the population intended to be analyzed.

### Q43: Explain SVM machine learning algorithm in detail.

SVM stands for support vector machine; it is a supervised machine learning algorithm which can be used for both **Regression and Classification**. If you have n features in your training dataset, SVM tries to plot it in n-dimensional space with the value of each feature being the value of a particular coordinate. SVM uses hyper planes to separate out different classes based on the provided kernel function.



### Q44: What are support vectors in SVM?



In the above diagram we see that the thinner lines mark the distance from the classifier to the closest data points called the support vectors (darkened data points). The distance between the two thin lines is called the margin.

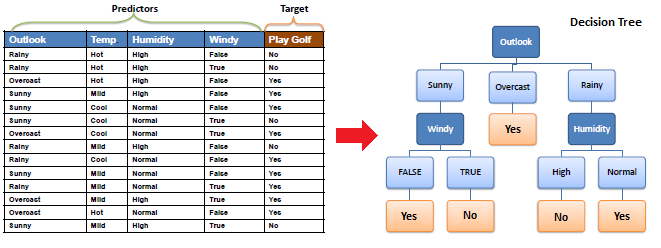
### Q45: What are the different kernels functions in SVM?

There are four types of kernels in SVM.

1. Linear Kernel
2. Polynomial kernel
3. Radial basis kernel
4. Sigmoid kernel

### Q46: Explain Decision Tree algorithm in detail.

Decision tree is a supervised machine learning algorithm mainly used for the **Regression and Classification**. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. Decision tree can handle both categorical and numerical data.

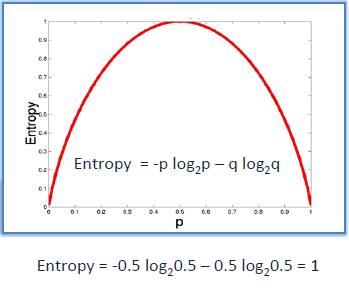


### Q47: What is Entropy and Information gain in Decision tree algorithm?

The core algorithm for building decision tree is called **ID3. ID3** uses **Entropy** and **Information Gain** to construct a decision tree.

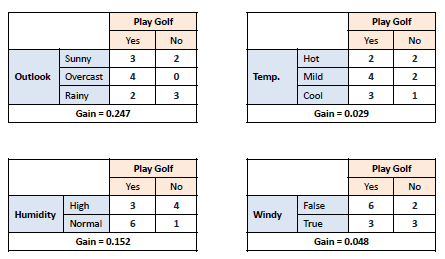
***Entropy***

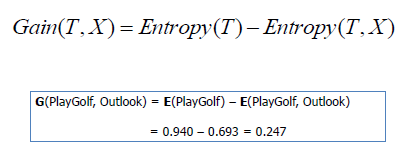
A decision tree is built top-down from a root node and involve partitioning of data into homogeneous subsets. **ID3** uses entropy to check the homogeneity of a sample. If the sample is completely homogeneous then entropy is zero and if the sample is an equally divided it has entropy of one.



***Information Gain***

The **Information Gain** is based on the decrease in entropy after a dataset is split on an attribute. Constructing a decision tree is all about finding attributes that returns the highest information gain.





### Q48: What is pruning in Decision Tree?

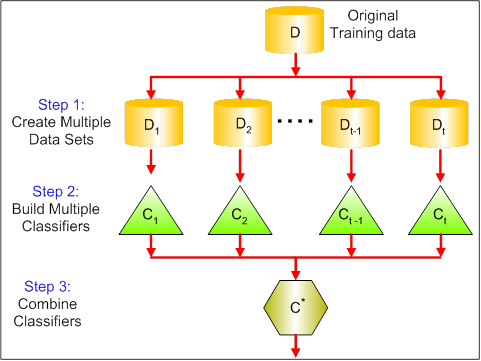
When we remove sub-nodes of a decision node, this process is called pruning or opposite process of splitting.

### Q49: What is Ensemble Learning?

Ensemble is the art of combining diverse set of learners (Individual models) together to improvise on the stability and predictive power of the model. Ensemble learning has many types but two more popular ensemble learning techniques are mentioned below.

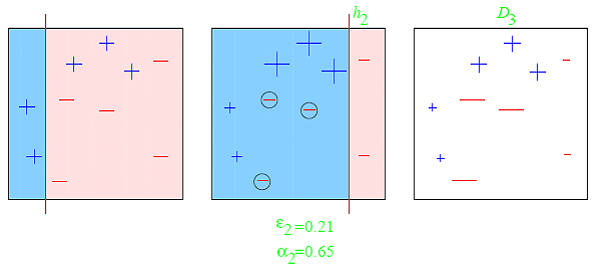
***Bagging***

Bagging tries to implement similar learners on small sample populations and then takes a mean of all the predictions. In generalized bagging, you can use different learners on different population. As you expect this helps us to reduce the variance error.



***Boosting***

Boosting is an iterative technique which adjust the weight of an observation based on the last classification. If an observation was classified incorrectly, it tries to increase the weight of this observation and vice versa. Boosting in general decreases the bias error and builds strong predictive models. However, they may overfit on the training data.



### Q50: What is Random Forest? How does it work?

Random forest is a versatile machine learning method capable of performing both regression and classification tasks. It is also used for dimensionality reduction, treats missing values, outlier values. It is a type of ensemble learning method, where a group of weak models combine to form a powerful model.

In Random Forest, we grow multiple trees as opposed to a single tree. To classify a new object based on attributes, each tree gives a classification. The forest chooses the classification having the **most votes** (Over all the trees in the forest) and in case of regression, it takes the **average** of outputs by different trees.

### Q51: What cross-validation technique would you use on a time series dataset.

Instead of using k-fold cross-validation, you should be aware to the fact that a time series is not randomly distributed data - It is inherently ordered by chronological order.

In case of time series data, you should use techniques like forward chaining – Where you will be model on past data then look at forward-facing data.

fold 1: training[1], test[2]

fold 1: training[1 2], test[3]

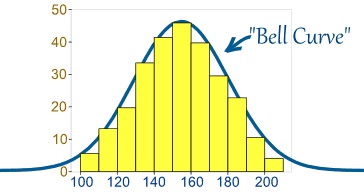
fold 1: training[1 2 3], test[4]

fold 1: training[1 2 3 4], test[5]

### Q52: What is logistic regression? Or State an example when you have used logistic regression recently.

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

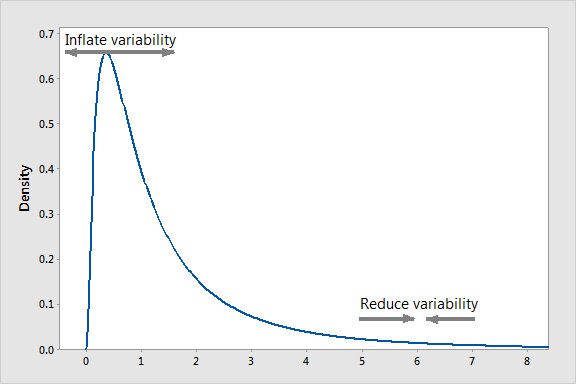
### Q53: What do you understand by the term Normal Distribution?



Data is usually distributed in different ways with a bias to the left or to the right or it can all be jumbled up. However, there are chances that data is distributed around a central value without any bias to the left or right and reaches normal distribution in the form of a bell shaped curve. The random variables are distributed in the form of a symmetrical bell shaped curve.

### Q54: What is a Box Cox Transformation?

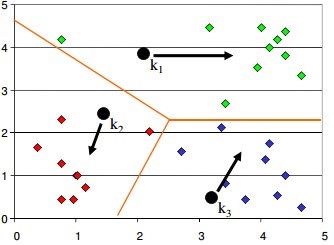
Dependent variable for a regression analysis might not satisfy one or more assumptions of an ordinary least squares regression. The residuals could either curve as the prediction increases or follow skewed distribution. In such scenarios, it is necessary to transform the response variable so that the data meets the required assumptions. A Box cox transformation is a statistical technique to transform non-normal dependent variables into a normal shape. If the given data is not normal then most of the statistical techniques assume normality. Applying a box cox transformation means that you can run a broader number of tests.



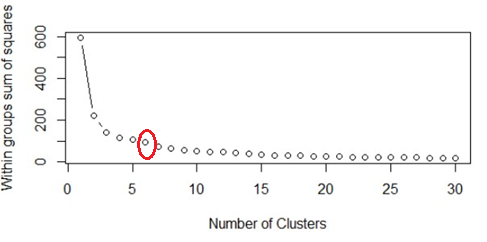
A Box Cox transformation is a way to transform non-normal dependent variables into a normal shape. Normality is an important assumption for many statistical techniques, if your data isn’t normal, applying a Box-Cox means that you are able to run a broader number of tests. The Box Cox transformation is named after statisticians **George Box** and **Sir David Roxbee Cox** who collaborated on a 1964 paper and developed the technique.

### Q55: How will you define the number of clusters in a clustering algorithm?

Though the Clustering Algorithm is not specified, this question will mostly be asked in reference to K-Means clustering where “K” defines the number of clusters. For example, the following image shows three different groups.



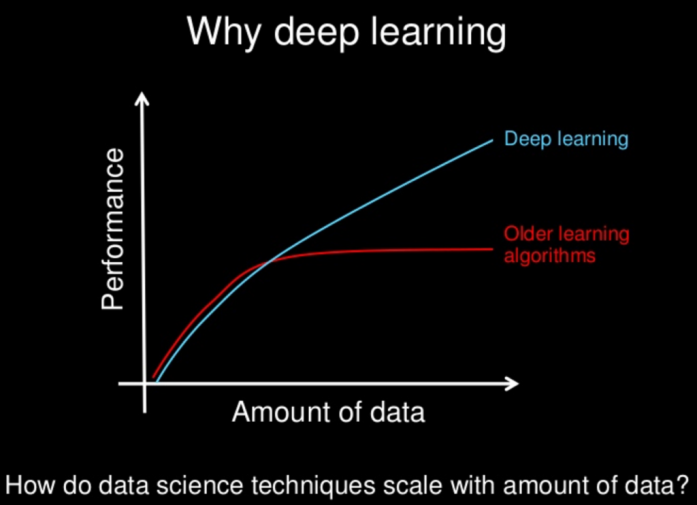
Within Sum of squares is generally used to explain the homogeneity within a cluster. If you plot WSS for a range of number of clusters, you will get the plot shown below. The Graph is generally known as Elbow Curve.



Red circled point in above graph i.e. Number of Cluster =6 is the point after which you don’t see any decrement in WSS. This point is known as bending point and taken as K in K – Means. This is the widely used approach but few data scientists also use Hierarchical clustering first to create dendograms and identify the distinct groups from there.

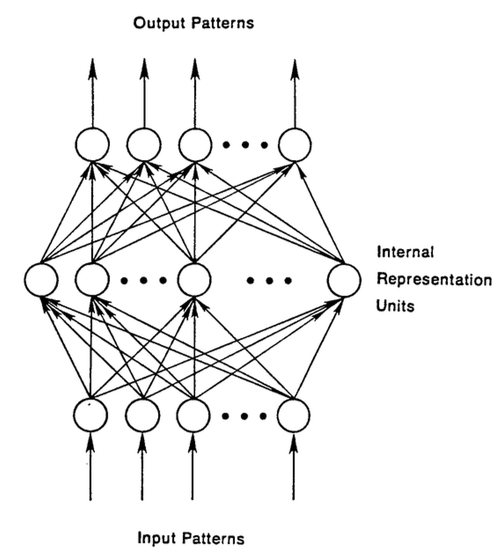
### Q56: What is deep learning?

Deep learning is subfield of machine learning inspired by structure and function of brain called artificial neural network. We have a lot numbers of algorithms under machine learning like linear regression, SVM, Neural network etc. and deep learning is just an extension of Neural networks. In neural nets we consider small number of hidden layers but when it comes to deep learning algorithms we consider a huge number of hidden layers to better understand the input output relationship.



### Q57: What are Recurrent Neural Networks (RNNs)?

Recurrent nets are type of artificial neural networks designed to recognize pattern from the sequence of data such as Time series, stock market and government agencies etc. To understand recurrent nets, first you have to understand the basics of feedforward nets. Both these networks RNN and feedforward named after the way they channel information through a series of mathematical operations performed at the nodes of the network. One feeds information through straight (never touching same node twice), while the other cycles it through loop, and the latter are called recurrent.

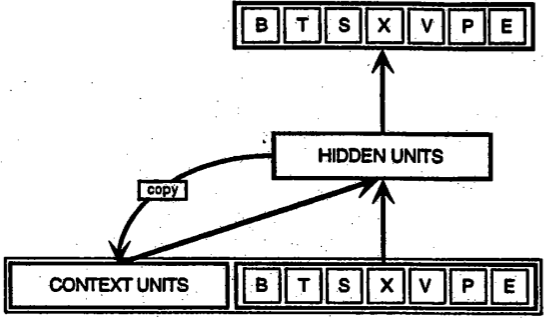


Recurrent networks on the other hand, take as their input not just the current input example they see, but also the what they have perceived previously in time. The BTSXPE at the bottom of the drawing represents the input example in the current moment, and CONTEXT UNIT represents the output of the previous moment. The decision a recurrent neural network reached at time t-1 affects the decision that it will reach one moment later at time t. So recurrent networks have two sources of input, the present and the recent past, which combine to determine how they respond to new data, much as we do in life.

The error they generate will return via backpropagation and be used to adjust their weights until error can’t go any lower. Remember, the purpose of recurrent nets is to accurately classify sequential input. We rely on the backpropagation of error and gradient descent to do so.

Backpropagation in feedforward networks moves backward from the final error through the outputs, weights and inputs of each hidden layer, assigning those weights responsibility for a portion of the error by calculating their partial derivatives – ∂E/∂w, or the relationship between their rates of change. Those derivatives are then used by our learning rule, gradient descent, to adjust the weights up or down, whichever direction decreases error.

Recurrent networks rely on an extension of backpropagation called backpropagation through time, or BPTT. Time, in this case, is simply expressed by a well-defined, ordered series of calculations linking one time step to the next, which is all backpropagation needs to work.



### ****Q58:**** What is the difference between machine learning and deep learning?

#### Machine learning:

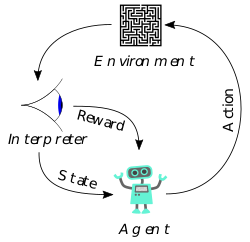
Machine learning is a field of computer science that gives computers the ability to learn without being explicitly programmed. Machine learning can be categorized in following three categories.

1. Supervised machine learning,
2. Unsupervised machine learning,
3. Reinforcement learning

#### Deep learning:

Deep Learning is a subfield of machine learning concerned with algorithms inspired by the structure and function of the brain called artificial neural networks.

**Reinforcement learning**



Reinforcement Learning is learning what to do and how to map situations to actions. The end result is to maximize the numerical reward signal. The learner is not told which action to take, but instead must discover which action will yield the maximum reward. Reinforcement learning is inspired by the learning of human beings; it is based on the reward/penalty mechanism.

### Q59: What is selection bias?

***Selection Bias***

Selection bias is the bias introduced by the selection of individuals, groups or data for analysis in such a way that proper randomization is not achieved, thereby ensuring that the sample obtained is not representative of the population intended to be analyzed. It is sometimes referred to as the selection effect. The phrase “selection bias” most often refers to the distortion of a statistical analysis, resulting from the method of collecting samples. If the selection bias is not taken into account, then some conclusions of the study may not be accurate.

### Q60: Explain what regularization is and why it is useful.

***Regularization***

Regularization is the process of adding tuning parameter to a model to induce smoothness in order to prevent overfitting. This is most often done by adding a constant multiple to an existing weight vector. This constant is often the L1(Lasso) or L2(ridge). The model predictions should then minimize the loss function calculated on the regularized training set.

### Q61: What is TF/IDF vectorization?

tf–idf is short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining. The tf-idf value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general.

### Q62: What are Recommender Systems?

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

### Q63: What is the difference between Regression and classification ML techniques?

Both Regression and classification machine learning techniques come under **Supervised machine learning algorithms**. In Supervised machine learning algorithm, we have to train the model using labeled dataset, While training we have to explicitly provide the correct labels and algorithm tries to learn the pattern from input to output. If our labels are discrete values then it will a classification problem, e.g. A, B etc. but if our labels are continuous values then it will be a regression problem, e.g. 1.23, 1.333 etc.

### Q64: If you are having 4GB RAM in your machine and you want to train your model on 10GB dataset. How would you go about this problem? Have you ever faced this kind of problem in your machine learning/data science experience so far?

First of all you have to ask which ML model you want to train.

**For Neural networks:** Batch size with Numpy array will work.

**Steps:**

1. Load the whole data in Numpy array. Numpy array has property to create mapping of complete dataset, it doesn’t load complete dataset in memory.
2. You can pass index to Numpy array to get required data.
3. Use this data to pass to neural network.
4. Have small batch size.

**For SVM:** Partial fit will work

**Steps:**

1. Divide one big dataset in small size datasets.
2. Use partial fit method of SVM; it requires subset of complete dataset.
3. Repeat step 2 for other subsets.

### Q65: What is p-value?

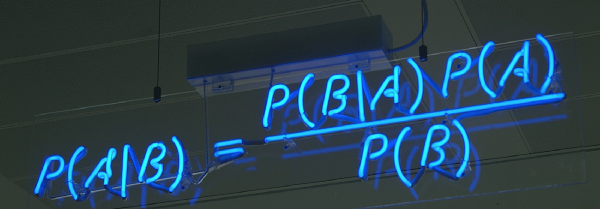
When you perform a hypothesis test in statistics, a p-value can help you determine the strength of your results. P-value is a number between 0 and 1. Based on the value it will denote the strength of the results. The claim which is on trial is called Null Hypothesis.

Low p-value (≤ 0.05) indicates strength against the null hypothesis which means we can reject the null Hypothesis. High p-value (≥ 0.05) indicates strength for the null hypothesis which means we can accept the null Hypothesis p-value of 0.05 indicates the Hypothesis could go either way. To put it in another way,

High P values: your data are likely with a true null. Low P values: your data are unlikely with a true null.

### Q66: What is ‘Naive’ in a Naive Bayes?

The Naive Bayes Algorithm is based on the Bayes Theorem. Bayes’ theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event.



**What is Naive?**

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

**Q67: What is the Central Limit Theorem and why is it important?**

* “Suppose that we are interested in estimating the average height among all people. Collecting data for every person in the world is impossible. While we can’t obtain a height measurement from everyone in the population, we can still sample some people. The question now becomes, what can we say about the average height of the entire population given a single sample. The Central Limit Theorem addresses this question exactly.” *Read more*[*here*](https://spin.atomicobject.com/2015/02/12/central-limit-theorem-intro/)*.*

**Q68: What is sampling? How many sampling methods do you know?**

* “Data sampling is a statistical analysis technique used to select, manipulate and analyze a representative subset of data points to identify patterns and trends in the larger data set being examined.” *Read the full answer*[*here*](https://searchbusinessanalytics.techtarget.com/definition/data-sampling)*.*

**Q69: What is the difference between type I vs type II error?**

* “A type I error occurs when the null hypothesis is true, but is rejected. A type II error occurs when the null hypothesis is false, but erroneously fails to be rejected.” *Read the full answer*[*here*](https://www.datasciencecentral.com/profiles/blogs/understanding-type-i-and-type-ii-errors)*.*

**Q70: What is linear regression? What do the terms p-value, coefficient, and r-squared value mean? What is the significance of each of these components?**

* A linear regression is a good tool for quick predictive analysis: for example, the price of a house depends on a myriad of factors, such as its size or its location. In order to see the relationship between these variables, we need to build a linear regression, which predicts the line of best fit between them and can help conclude whether or not these two factors have a positive or negative relationship. *Read more*[*here*](https://www.springboard.com/blog/linear-regression-in-python-a-tutorial/)*and*[*here.*](http://blog.minitab.com/blog/adventures-in-statistics/how-to-interpret-regression-analysis-results-p-values-and-coefficients)

**Q71: What are the assumptions required for linear regression?**

* There are four major assumptions: 1. There is a linear relationship between the dependent variables and the regressors, meaning the model you are creating actually fits the data, 2. The errors or residuals of the data are normally distributed and independent from each other, 3. There is minimal multicollinearity between explanatory variables, and 4. Homoscedasticity. This means the variance around the regression line is the same for all values of the predictor variable.

**Q72: What is a statistical interaction?**

* ”Basically, an interaction is when the effect of one factor (input variable) on the dependent variable (output variable) differs among levels of another factor.” *Read more*[*here*](http://icbseverywhere.com/blog/mini-lessons-tutorials-and-support-pages/statistical-interactions/)*.*

**Q73: What is an example of a data set with a non-Gaussian distribution?**

* “The Gaussian distribution is part of the Exponential family of distributions, but there are a lot more of them, with the same sort of ease of use, in many cases, and if the person doing the machine learning has a solid grounding in statistics, they can be utilized where appropriate.” *Read more*[*here*](https://www.quora.com/Most-machine-learning-datasets-are-in-Gaussian-distribution-Where-can-we-find-the-dataset-which-follows-Bernoulli-Poisson-gamma-beta-etc-distribution)*.*

**Q74: What is the Binomial Probability Formula?**

* “The binomial distribution consists of the probabilities of each of the possible numbers of successes on N trials for independent events that each have a probability of π (the Greek letter pi) of occurring.” *Read more*[*here*](http://onlinestatbook.com/2/probability/binomial.html)*.*

**Q75: What are two main components of the Hadoop framework? Explain how MapReduce works as simply as possible.**

* The Hadoop Distributed File System (HDFS), MapReduce, and YARN. *Read more*[*here*](https://www.quora.com/What-are-the-main-components-of-a-Hadoop-Application)*.*
* MapReduce is a programming model that enables distributed processing of large data sets on compute clusters of commodity hardware. Hadoop MapReduce first performs mapping which involves splitting a large file into pieces to make another set of data.” *Read more*[*here*](https://bigdata-madesimple.com/basic-components-of-hadoop-architecture-frameworks-used-for-data-science/)*.*

**Q76: What is the purpose of the group functions in SQL? Give some examples of group functions. Tell me the difference between an inner join, left join/right join, and union.**

* Group functions are necessary to get summary statistics of a data set. COUNT, MAX, MIN, AVG, SUM, and DISTINCT are all group functions.
* “In a Venn diagram the inner join is when both tables have a match, a left join is when there is a match in the left table and the right table is null, a right join is the opposite of a left join, and a full join is all of the data combined.” *Read more*[*here*](https://www.springboard.com/blog/joining-data-tables/)*.*

**Q77: In Python, how is memory managed?**

In Python, memory is managed in a private heap space. This means that all the objects and data structures will be located in a private heap. However, the programmer won’t be allowed to access this heap. Instead, the Python interpreter will handle it. At the same time, the core API will enable access to some Python tools for the programmer to start coding. The memory manager will allocate the heap space for the Python objects while the inbuilt garbage collector will recycle all the memory that’s not being used to boost available heap space. *Read more*[*here*](https://www.springboard.com/blog/python-interview-questions/)*.*

**Q78: What are the supported data types in Python?**

“Python’s built-in (or standard) data types can be grouped into several classes. Sticking to the hierarchy scheme used in the official Python documentation these are numeric types, sequences, sets and mappings.” *Read more*[*here*](https://www.quora.com/What-are-the-supported-data-types-in-Python)*.*

**Q79: What is the difference between a tuple and a list in Python?**

“Apart from tuples being immutable there is also a semantic distinction that should guide their usage.” *Read more*[*here*](https://stackoverflow.com/questions/626759/whats-the-difference-between-lists-and-tuples)*.*

**Q80: How is k-NN different from k-means clustering?**

k-NN, or k-nearest neighbors is a classification algorithm, where the k is an integer describing the number of neighboring data points that influence the classification of a given observation. K-means is a clustering algorithm, where the k is an integer describing the number of clusters to be created from the given data.

**Q81: Explain the 80/20 rule, and tell me about its importance in model validation.**

“People usually tend to start with a 80-20% split (80% training set – 20% test set) and split the training set once more into a 80-20% ratio to create the validation set.” *Read more*[*here*](https://www.beyondthelines.net/machine-learning/how-to-split-a-dataset/)*.*

**Q82: Explain what precision and recall are. How do they relate to the ROC curve?**

Recall describes what percentage of true positives are described as positive by the model. Precision describes what percent of positive predictions were correct. The ROC curve shows the relationship between model recall and specificity–specificity being a measure of the percent of true negatives being described as negative by the model. Recall, precision, and the ROC are measures used to identify how useful a given classification model is.*Read more*[*here*](http://www.kdnuggets.com/faq/precision-recall.html)*.*

**Q83: Explain the difference between L1 and L2 regularization methods.**

“A regression model that uses L1 regularization technique is called Lasso Regression and model which uses L2 is called Ridge Regression. The key difference between these two is the penalty term.” *Read more*[*here*](https://towardsdatascience.com/l1-and-l2-regularization-methods-ce25e7fc831c)*.*

**Q84: You are given a train data set having 1000 columns and 1 million rows. The data set is based on a classification problem. Your manager has asked you to reduce the dimension of this data so that model computation time can be reduced. Your machine has memory constraints. What would you do? (You are free to make practical assumptions.)**

Processing a high dimensional data on a limited memory machine is a strenuous task, your interviewer would be fully aware of that. Following are the methods you can use to tackle such situation:

1. Since we have lower RAM, we should close all other applications in our machine, including the web browser, so that most of the memory can be put to use.
2. We can randomly sample the data set. This means, we can create a smaller data set, let’s say, having 1000 variables and 300000 rows and do the computations.
3. To reduce dimensionality, we can separate the numerical and categorical variables and remove the correlated variables. For numerical variables, we’ll use correlation. For categorical variables, we’ll use chi-square test.
4. Also, we can use [PCA](https://www.analyticsvidhya.com/blog/2016/03/practical-guide-principal-component-analysis-python/) and pick the components which can explain the maximum variance in the data set.
5. Using online learning algorithms like Vowpal Wabbit (available in Python) is a possible option.
6. Building a linear model using Stochastic Gradient Descent is also helpful.
7. We can also apply our business understanding to estimate which all predictors can impact the response variable. But, this is an intuitive approach, failing to identify useful predictors might result in significant loss of information.

**Q85: Is rotation necessary in PCA? If yes, Why? What will happen if you don’t rotate the components?**

Yes, rotation (orthogonal) is necessary because it maximizes the difference between variance captured by the component. This makes the components easier to interpret. Not to forget, that’s the motive of doing PCA where, we aim to select fewer components (than features) which can explain the maximum variance in the data set. By doing rotation, the relative location of the components doesn’t change, it only changes the actual coordinates of the points.

If we don’t rotate the components, the effect of PCA will diminish and we’ll have to select more number of components to explain variance in the data set.

**Q:86: You are given a data set. The data set has missing values which spread along 1 standard deviation from the median. What percentage of data would remain unaffected? Why?**

This question has enough hints for you to start thinking! Since, the data is spread across median, let’s assume it’s a normal distribution. We know, in a normal distribution, ~68% of the data lies in 1 standard deviation from mean (or mode, median), which leaves ~32% of the data unaffected. Therefore, ~32% of the data would remain unaffected by missing values.

**Q87: You are given a data set on cancer detection. You’ve build a classification model and achieved an accuracy of 96%. Why shouldn’t you be happy with your model performance? What can you do about it?**

If you have worked on enough data sets, you should deduce that cancer detection results in imbalanced data. In an imbalanced data set, accuracy should not be used as a measure of performance because 96% (as given) might only be predicting majority class correctly, but our class of interest is minority class (4%) which is the people who actually got diagnosed with cancer. Hence, in order to evaluate model performance, we should use Sensitivity (True Positive Rate), Specificity (True Negative Rate), F measure to determine class wise performance of the classifier. If the minority class performance is found to to be poor, we can undertake the following steps:

1. We can use undersampling, oversampling or SMOTE to make the data balanced.
2. We can alter the prediction threshold value by doing [probability caliberation](https://www.analyticsvidhya.com/blog/2016/07/platt-scaling-isotonic-regression-minimize-logloss-error/) and finding a optimal threshold using AUC-ROC curve.
3. We can assign weight to classes such that the minority classes gets larger weight.
4. We can also use anomaly detection.

**Q88: Why is naive Bayes so ‘naive’ ?**

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

**Q89: Explain prior probability, likelihood and marginal likelihood in context of naïve Bayes algorithm?**

Prior probability is nothing but, the proportion of dependent (binary) variable in the data set. It is the closest guess you can make about a class, without any further information. For example: In a data set, the dependent variable is binary (1 and 0). The proportion of 1 (spam) is 70% and 0 (not spam) is 30%. Hence, we can estimate that there are 70% chances that any new email would  be classified as spam.

Likelihood is the probability of classifying a given observation as 1 in presence of some other variable. For example: The probability that the word ‘FREE’ is used in previous spam message is likelihood. Marginal likelihood is, the probability that the word ‘FREE’ is used in any message.

**Q90: You are working on a time series data set. You manager has asked you to build a high accuracy model. You start with the decision tree algorithm, since you know it works fairly well on all kinds of data. Later, you tried a time series regression model and got higher accuracy than decision tree model. Can this happen? Why?**

Time series data is known to posses linearity. On the other hand, a decision tree algorithm is known to work best to detect non – linear interactions. The reason why decision tree failed to provide robust predictions because it couldn’t map the linear relationship as good as a regression model did. Therefore, we learned that, a linear regression model can provide robust prediction given the data set satisfies its [linearity assumptions](https://www.analyticsvidhya.com/blog/2016/07/deeper-regression-analysis-assumptions-plots-solutions/).

**Q91: You are assigned a new project which involves helping a food delivery company save more money. The problem is, company’s delivery team aren’t able to deliver food on time. As a result, their customers get unhappy. And, to keep them happy, they end up delivering food for free. Which machine learning algorithm can save them?**

You might have started hopping through the list of ML algorithms in your mind. But, wait! Such questions are asked to test your machine learning fundamentals.

This is not a machine learning problem. This is a route optimization problem. A machine learning problem consist of three things:

1. There exist a pattern.
2. You cannot solve it mathematically (even by writing exponential equations).
3. You have data on it.

Always look for these three factors to decide if machine learning is a tool to solve a particular problem.

**Q92: You came to know that your model is suffering from low bias and high variance. Which algorithm should you use to tackle it? Why?**

 Low bias occurs when the model’s predicted values are near to actual values. In other words, the model becomes flexible enough to mimic the training data distribution. While it sounds like great achievement, but not to forget, a flexible model has no generalization capabilities. It means, when this model is tested on an unseen data, it gives disappointing results.

In such situations, we can use bagging algorithm (like random forest) to tackle high variance problem. Bagging algorithms divides a data set into subsets made with repeated randomized sampling. Then, these samples are used to generate  a set of models using a single learning algorithm. Later, the model predictions are combined using voting (classification) or averaging (regression).

Also, to combat high variance, we can:

1. Use regularization technique, where higher model coefficients get penalized, and hence lowering model complexity.
2. Use top n features from variable importance chart. May be, with all the variable in the data set, the algorithm is having difficulty in finding the meaningful signal.

**Q93: You are given a data set. The data set contains many variables, some of which are highly correlated and you know about it. Your manager has asked you to run PCA. Would you remove correlated variables first? Why?**

Chances are, you might be tempted to say No, but that would be incorrect. Discarding correlated variables have a substantial effect on PCA because, in presence of correlated variables, the variance explained by a particular component gets inflated.

For example: You have 3 variables in a data set, of which 2 are correlated. If you run PCA on this data set, the first principal component would exhibit twice the variance than it would exhibit with uncorrelated variables. Also, adding correlated variables lets PCA put more importance on those variable, which is misleading.

**Q94: After spending several hours, you are now anxious to build a high accuracy model. As a result, you build 5 GBM models, thinking a boosting algorithm would do the magic. Unfortunately, neither of models could perform better than benchmark score. Finally, you decided to combine those models. Though, ensembled models are known to return high accuracy, but you are unfortunate. Where did you miss?**

As we know, ensemble learners are based on the idea of combining weak learners to create strong learners. But, these learners provide superior result when the combined models are uncorrelated. Since, we have used 5 GBM models and got no accuracy improvement, suggests that the models are correlated. The problem with correlated models is, all the models provide same information.

For example: If model 1 has classified User1122 as 1, there are high chances model 2 and model 3 would have done the same, even if its actual value is 0. Therefore, ensemble learners are built on the premise of combining weak uncorrelated models to obtain better predictions.

**Q95: How is kNN different from kmeans clustering?**

Don’t get mislead by ‘k’ in their names. You should know that the fundamental difference between both these algorithms is, kmeans is unsupervised in nature and kNN is supervised in nature. kmeans is a clustering algorithm. kNN is a classification (or regression) algorithm.

kmeans algorithm partitions a data set into clusters such that a cluster formed is homogeneous and the points in each cluster are close to each other. The algorithm tries to maintain enough separability between these clusters. Due to unsupervised nature, the clusters have no labels.

kNN algorithm tries to classify an unlabeled observation based on its k (can be any number ) surrounding neighbors. It is also known as lazy learner because it involves minimal training of model. Hence, it doesn’t use training data to make generalization on unseen data set.

**Q96: How is True Positive Rate and Recall related? Write the equation.**

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

**You have built a multiple regression model. Your model R² isn’t as good as you wanted. For improvement, your remove the intercept term, your model R² becomes 0.8 from 0.3. Is it possible? How?**

Yes, it is possible. We need to understand the significance of intercept term in a regression model. The intercept term shows model prediction without any independent variable i.e. mean prediction. The formula of  where y´ is predicted value.

When intercept term is present, R² value evaluates your model wrt. to the mean model. In absence of intercept term (ymean), the model can make no such evaluation, with large denominator,  equation’s value becomes smaller than actual, resulting in higher R².

**Q97: After analyzing the model, your manager has informed that your regression model is suffering from multicollinearity. How would you check if he’s true? Without losing any information, can you still build a better model?**

To check multicollinearity, we can create a correlation matrix to identify & remove variables having correlation above 75% (deciding a threshold is subjective). In addition, we can use calculate VIF (variance inflation factor) to check the presence of multicollinearity. VIF value <= 4 suggests no multicollinearity whereas a value of >= 10 implies serious multicollinearity. Also, we can use tolerance as an indicator of multicollinearity.

But, removing correlated variables might lead to loss of information. In order to retain those variables, we can use penalized regression models like ridge or lasso regression. Also, we can add some random noise in correlated variable so that the variables become different from each other. But, adding noise might affect the prediction accuracy, hence this approach should be carefully used.

**Q98: When is Ridge regression favorable over Lasso regression?**

You can quote ISLR’s authors Hastie, Tibshirani who asserted that, in presence of few variables with medium / large sized effect, use lasso regression. In presence of many variables with small / medium sized effect, use ridge regression.

Conceptually, we can say, lasso regression (L1) does both variable selection and parameter shrinkage, whereas Ridge regression only does parameter shrinkage and end up including all the coefficients in the model. In presence of correlated variables, ridge regression might be the preferred choice. Also, ridge regression works best in situations where the least square estimates have higher variance. Therefore, it depends on our model objective.

**Q99: Rise in global average temperature led to decrease in number of pirates around the world. Does that mean that decrease in number of pirates caused the climate change?**

After reading this question, you should have understood that this is a classic case of “causation and correlation”. No, we can’t conclude that decrease in number of pirates caused the climate change because there might be other factors (lurking or confounding variables) influencing this phenomenon.

Therefore, there might be a correlation between global average temperature and number of pirates, but based on this information we can’t say that pirated died because of rise in global average temperature.

**Q100: While working on a data set, how do you select important variables? Explain your methods.**

Following are the methods of variable selection you can use:

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

**Q101: What is the difference between covariance and correlation?**

Correlation is the standardized form of covariance.

Covariances are difficult to compare. For example: if we calculate the covariances of salary ($) and age (years), we’ll get different covariances which can’t be compared because of having unequal scales. To combat such situation, we calculate correlation to get a value between -1 and 1, irrespective of their respective scale.

**Q102: Is it possible capture the correlation between continuous and categorical variable? If yes, how?**

Answer: Yes, we can use ANCOVA (analysis of covariance) technique to capture association between continuous and categorical variables.

**Q103: Both being tree based algorithm, how is random forest different from Gradient boosting algorithm (GBM)?**

**Answer:** The fundamental difference is, random forest uses bagging technique to make predictions. GBM uses boosting techniques to make predictions.

In bagging technique, a data set is divided into n samples using randomized sampling. Then, using a single learning algorithm a model is build on all samples. Later, the resultant predictions are combined using voting or averaging. Bagging is done is parallel. In boosting, after the first round of predictions, the algorithm weighs misclassified predictions higher, such that they can be corrected in the succeeding round. This sequential process of giving higher weights to misclassified predictions continue until a stopping criterion is reached.

Random forest improves model accuracy by reducing variance (mainly). The trees grown are uncorrelated to maximize the decrease in variance. On the other hand, GBM improves accuracy my reducing both bias and variance in a model.

Know more: [Tree based modeling](https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/)

**Q104: Running a binary classification tree algorithm is the easy part. Do you know how does a tree splitting takes place i.e. how does the tree decide which variable to split at the root node and succeeding nodes?**

**Answer:** A classification trees makes decision based on Gini Index and Node Entropy. In simple words, the tree algorithm find the best possible feature which can divide the data set into purest possible children nodes.

Gini index says, if we select two items from a population at random then they must be of same class and probability for this is 1 if population is pure. We can calculate Gini as following:

1. Calculate Gini for sub-nodes, using formula sum of square of probability for success and failure (p^2+q^2).
2. Calculate Gini for split using weighted Gini score of each node of that split

Entropy is the measure of impurity as given by (for binary class):

Entropy, Decision Tree

Here p and q is probability of success and failure respectively in that node. Entropy is zero when a node is homogeneous. It is maximum when a both the classes are present in a node at 50% – 50%.  Lower entropy is desirable.

**Q105: You’ve built a random forest model with 10000 trees. You got delighted after getting training error as 0.00. But, the validation error is 34.23. What is going on? Haven’t you trained your model perfectly?**

**Answer:** The model has overfitted. Training error 0.00 means the classifier has mimiced the training data patterns to an extent, that they are not available in the unseen data. Hence, when this classifier was run on unseen sample, it couldn’t find those patterns and returned prediction with higher error. In random forest, it happens when we use larger number of trees than necessary. Hence, to avoid these situation, we should tune number of trees using cross validation.

**Q106:** **You’ve got a data set to work having p (no. of variable) > n (no. of observation). Why is OLS as bad option to work with? Which techniques would be best to use? Why?**

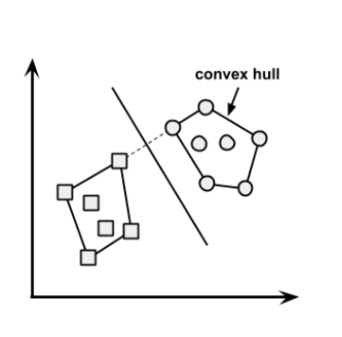
**Answer:** In such high dimensional data sets, we can’t use classical regression techniques, since their assumptions tend to fail. When p > n, we can no longer calculate a unique least square coefficient estimate, the variances become infinite, so OLS cannot be used at all.

To combat this situation, we can use penalized regression methods like lasso, LARS, ridge which can shrink the coefficients to reduce variance. Precisely, ridge regression works best in situations where the least square estimates have higher variance.

Among other methods include subset regression, forward stepwise regression.

**Q107: What is convex hull ? (Hint: Think SVM)**

**Answer:** In case of linearly separable data, convex hull represents the outer boundaries of the two group of data points. Once convex hull is created, we get maximum margin hyperplane (MMH) as a perpendicular bisector between two convex hulls. MMH is the line which attempts to create greatest separation between two groups.



**Q108: We know that one hot encoding increasing the dimensionality of a data set. But, label encoding doesn’t. How ?**

**Answer:** Don’t get baffled at this question. It’s a simple question asking the difference between the two.

Using one hot encoding, the dimensionality (a.k.a features) in a data set get increased because it creates a new variable for each level present in categorical variables. For example: let’s say we have a variable ‘color’. The variable has 3 levels namely Red, Blue and Green. One hot encoding ‘color’ variable will generate three new variables as Color.Red, Color.Blue and Color.Green containing 0 and 1 value.

In label encoding, the levels of a categorical variables gets encoded as 0 and 1, so no new variable is created. Label encoding is majorly used for binary variables.

**Q109: What cross validation technique would you use on time series data set? Is it k-fold or LOOCV?**

**Answer:** Neither.

In time series problem, k fold can be troublesome because there might be some pattern in year 4 or 5 which is not in year 3. Resampling the data set will separate these trends, and we might end up validation on past years, which is incorrect. Instead, we can use forward chaining strategy with 5 fold as shown below:

* fold 1 : training [1], test [2]
* fold 2 : training [1 2], test [3]
* fold 3 : training [1 2 3], test [4]
* fold 4 : training [1 2 3 4], test [5]
* fold 5 : training [1 2 3 4 5], test [6]

where 1,2,3,4,5,6 represents “year”.

**Q110: You are given a data set consisting of variables having more than 30% missing values? Let’s say, out of 50 variables, 8 variables have missing values higher than 30%. How will you deal with them?**

**Answer:** We can deal with them in the following ways:

1. Assign a unique category to missing values, who knows the missing values might decipher some trend
2. We can remove them blatantly.
3. Or, we can sensibly check their distribution with the target variable, and if found any pattern we’ll keep those missing values and assign them a new category while removing others.

**Q111: ‘People who bought this, also bought…’ recommendations seen on amazon is a result of which algorithm?**

**Answer:** The basic idea for this kind of recommendation engine comes from collaborative filtering.

Collaborative Filtering algorithm considers “User Behavior” for recommending items. They exploit behavior of other users and items in terms of transaction history, ratings, selection and purchase information. Other users behaviour and preferences over the items are used to recommend items to the new users. In this case, features of the items are not known.

Know more: [Recommender System](https://www.analyticsvidhya.com/blog/2015/10/recommendation-engines/)

**Q112: What do you understand by Type I vs Type II error ?**

**Answer:** Type I error is committed when the null hypothesis is true and we reject it, also known as a ‘False Positive’. Type II error is committed when the null hypothesis is false and we accept it, also known as ‘False Negative’.

In the context of confusion matrix, we can say Type I error occurs when we classify a value as positive (1) when it is actually negative (0). Type II error occurs when we classify a value as negative (0) when it is actually positive(1).

**Q113: You are working on a classification problem. For validation purposes, you’ve randomly sampled the training data set into train and validation. You are confident that your model will work incredibly well on unseen data since your validation accuracy is high. However, you get shocked after getting poor test accuracy. What went wrong?**

**Answer:** In case of classification problem, we should always use stratified sampling instead of random sampling. A random sampling doesn’t takes into consideration the proportion of target classes. On the contrary, stratified sampling helps to maintain the distribution of target variable in the resultant distributed samples also.

**Q114: You have been asked to evaluate a regression model based on R², adjusted R² and tolerance. What will be your criteria?**

**Answer:** Tolerance (1 / VIF) is used as an indicator of multicollinearity. It is an indicator of percent of variance in a predictor which cannot be accounted by other predictors. Large values of tolerance is desirable.

We will consider adjusted R² as opposed to R² to evaluate model fit because R² increases irrespective of improvement in prediction accuracy as we add more variables. But, adjusted R² would only increase if an additional variable improves the accuracy of model, otherwise stays same. It is difficult to commit a general threshold value for adjusted R² because it varies between data sets. For example: a gene mutation data set might result in lower adjusted R² and still provide fairly good predictions, as compared to a stock market data where lower adjusted R² implies that model is not good.

**Q115: In k-means or kNN, we use euclidean distance to calculate the distance between nearest neighbors. Why not manhattan distance ?**

**Answer:** We don’t use manhattan distance because it calculates distance horizontally or vertically only. It has dimension restrictions. On the other hand, euclidean metric can be used in any space to calculate distance. Since, the data points can be present in any dimension, euclidean distance is a more viable option.

Example: Think of a chess board, the movement made by a bishop or a rook is calculated by manhattan distance because of their respective vertical & horizontal movements.

**Explain machine learning to me like a 5 year old.**

**Answer:** It’s simple. It’s just like how babies learn to walk. Every time they fall down, they learn (unconsciously) & realize that their legs should be straight and not in a bend position. The next time they fall down, they feel pain. They cry. But, they learn ‘not to stand like that again’. In order to avoid that pain, they try harder. To succeed, they even seek support from the door or wall or anything near them, which helps them stand firm.

This is how a machine works & develops intuition from its environment.

Note: The interview is only trying to test if have the ability of explain complex concepts in simple terms.

**Q116: I know that a linear regression model is generally evaluated using Adjusted R² or F value. How would you evaluate a logistic regression model?**

**Answer:** We can use the following methods:

1. Since logistic regression is used to predict probabilities, we can use AUC-ROC curve along with confusion matrix to determine its performance.
2. Also, the analogous metric of adjusted R² in logistic regression is AIC. AIC is the measure of fit which penalizes model for the number of model coefficients. Therefore, we always prefer model with minimum AIC value.
3. Null Deviance indicates the response predicted by a model with nothing but an intercept. Lower the value, better the model. Residual deviance indicates the response predicted by a model on adding independent variables. Lower the value, better the model.

Know more: [Logistic Regression](https://www.analyticsvidhya.com/blog/2015/11/beginners-guide-on-logistic-regression-in-r/)

**Q117: Considering the long list of machine learning algorithm, given a data set, how do you decide which one to use?**

**Answer:** You should say, the choice of machine learning algorithm solely depends of the type of data. If you are given a data set which is exhibits linearity, then linear regression would be the best algorithm to use. If you given to work on images, audios, then neural network would help you to build a robust model.

If the data comprises of nonlinear interactions, then a boosting or bagging algorithm should be the choice. If the business requirement is to build a model which can be deployed, then we’ll use regression or a decision tree model (easy to interpret and explain) instead of black box algorithms like SVM, GBM etc.

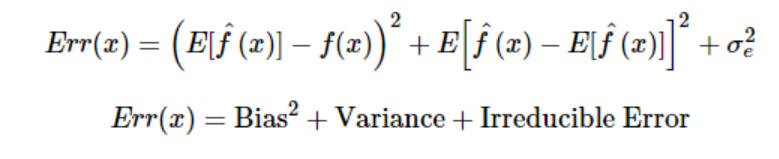
In short, there is no one master algorithm for all situations. We must be scrupulous enough to understand which algorithm to use.

**Q118: When does regularization becomes necessary in Machine Learning?**

**Answer:** Regularization becomes necessary when the model begins to ovefit / underfit. This technique introduces 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 hence reduce cost term. This helps to reduce model complexity so that the model can become better at predicting (generalizing).

**What do you understand by Bias Variance trade off?**

**Answer:**  The error emerging from any model can be broken down into three components mathematically. Following are these component :



**Bias error** is useful to quantify how much on an average are the predicted values different from the actual value. A high bias error means we have a under-performing model which keeps on missing important trends. **Variance**on the other side quantifies how are the prediction made on same observation different from each other. A high variance model will over-fit on your training population and perform badly on any observation beyond training.

**Q119: OLS is to linear regression. Maximum likelihood is to logistic regression. Explain the statement.**

**Answer:** OLS and Maximum likelihood are the methods used by the respective regression methods to approximate the unknown parameter (coefficient) value. In simple words,

Ordinary least square(OLS) is a method used in linear regression which approximates the parameters resulting in minimum distance between actual and predicted values. Maximum Likelihood helps in choosing the the values of parameters which maximizes the likelihood that the parameters are most likely to produce observed data.