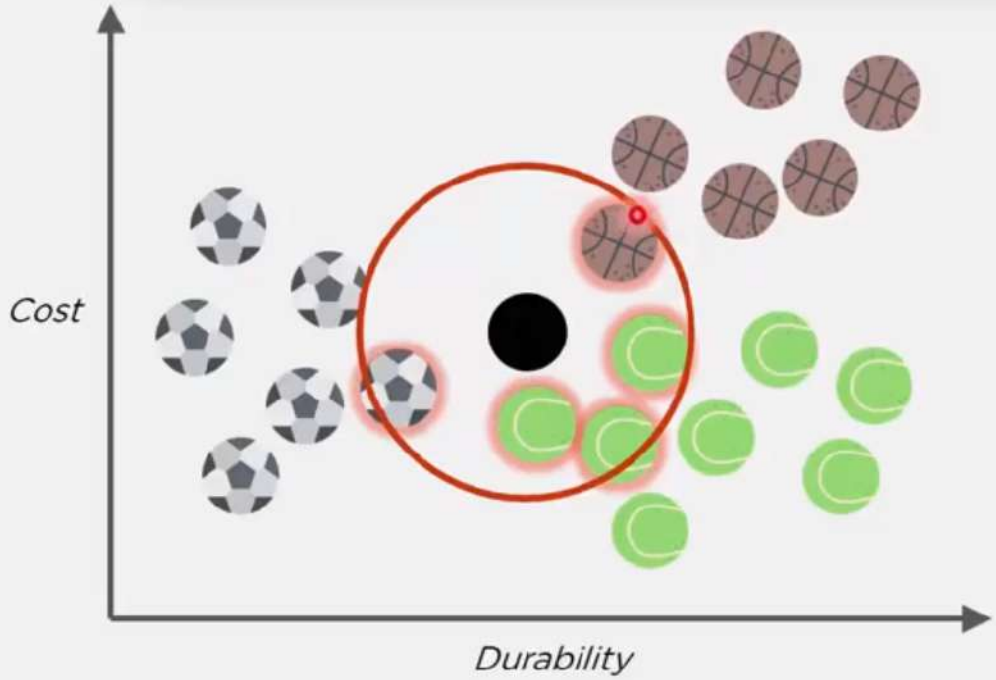
30 Machine Learning Interview Questions

Explain K Nearest Neighbor algorithm.



Let K = 5

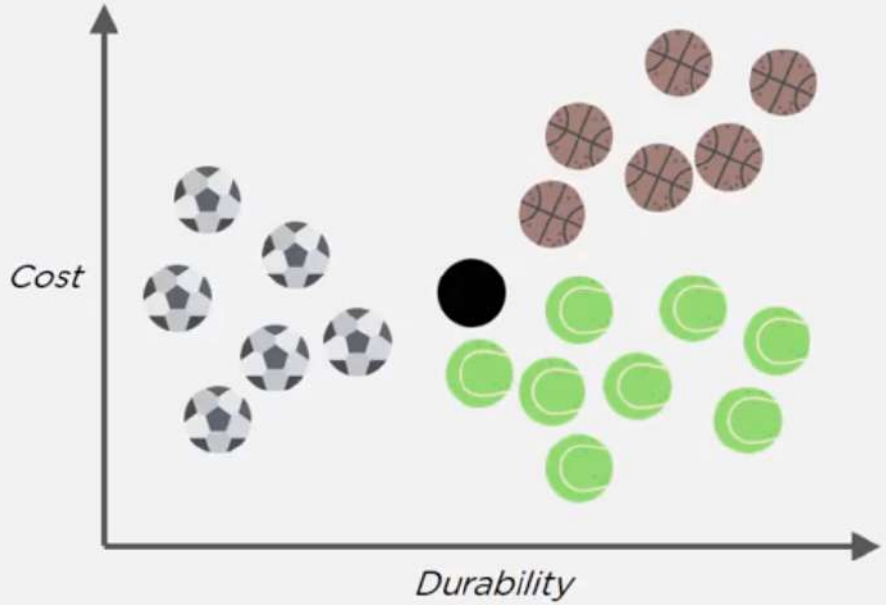
So these glowing ones are the closest 5 balls to our new data point (black ball)

Here, there are 3 tennis balls and one each of basketball and football

Thus we will classify the black ball as a tennis ball

30 Machine Learning Interview Questions

Explain K Nearest Neighbor algorithm.



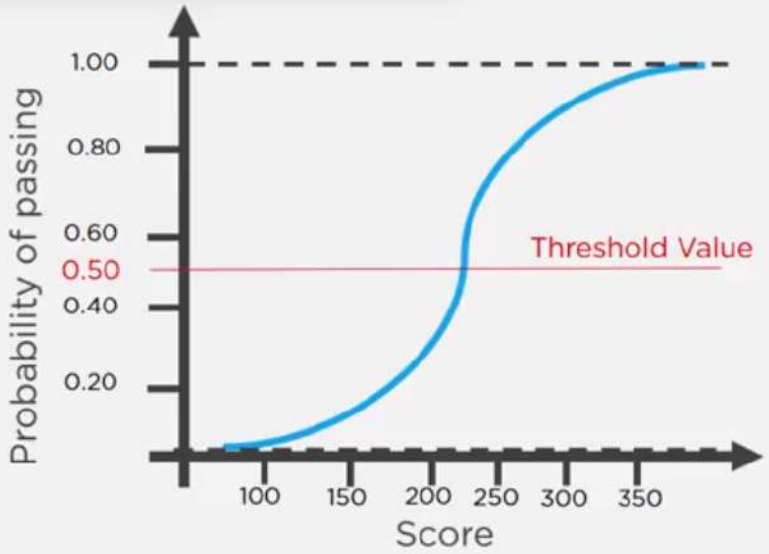
K Nearest Neighbors algorithm works in a way that a new data point is assigned to a neighboring group it is most similar to.

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 to.

29 Machine Learning Interview Questions

Briefly explain Logistic Regression.

- Logistic Regression is a classification algorithm, used to predict a binary outcomes for a given set of independent variables
- Output of a logistic regression is either a 0 or 1
- It has a threshold value which is generally 0.5
- Any value above 0.5 is considered as 1 and any point below 0.5 is considered as 0



Regression model created based on the performance of past participants

28 Machine Learning Interview Questions

What is pruning in decision trees and how is it done?

There is a popular pruning algorithm called **Reduced error pruning**

- ❑ Starting at the leaves, each node is replaced with its most popular class
- ❑ If the prediction accuracy is not affected then the change is kept
- ❑ Reduced error pruning has the advantage of **simplicity and speed**

28 Machine Learning Interview Questions

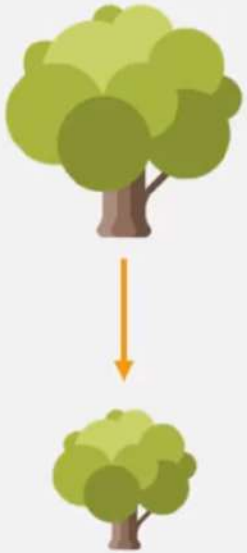
What is pruning in decision trees and how is it done?

Pruning can occur in a

- ❑ Top- down fashion
  - o A top down pruning will traverse nodes and trim subtrees starting at the root
- ❑ Bottom-up fashion
  - o A bottom up pruning will start at the leaf nodes

28 Machine Learning Interview Questions

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

27 Machine Learning Interview Questions

Define precision and recall.

Recall is the ratio of a number of events you can recall to the number of total events

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

If you can recall all 10 events, then, your recall ratio is 1.0 (100%)

100%

If you can recall 7 events, your recall ratio is 0.7 (70%)

70%

27 Machine Learning Interview Questions

Define precision and recall.

Precision is the ratio of a number of events you can correctly recall to a number all events you recall (mix of correct and wrong recalls)

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

In any 10 events, if you answer 10 times in which 8 events are correct and 2 events are wrong

26 Machine Learning Interview Questions

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

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

We need to find a balance of Bias and Variance so as to minimize the total error

High Bias
Low Variance

High Variance
Low Bias

26 Machine Learning Interview Questions

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

- Essentially, if you make the model more complex and add more variables, you'll lose bias but gain some variance — in order to get the optimally reduced amount of error, you'll have to tradeoff bias and variance
- You don't want either high bias or high variance in your model

25 Machine Learning Interview Questions

What is bias and variance in a Machine Learning model?

Variance refers to the amount the target model will change when trained with different training data

For a good model, 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

High Variance
Low Bias

25 Machine Learning Interview Questions

What is bias and variance in a Machine Learning model?

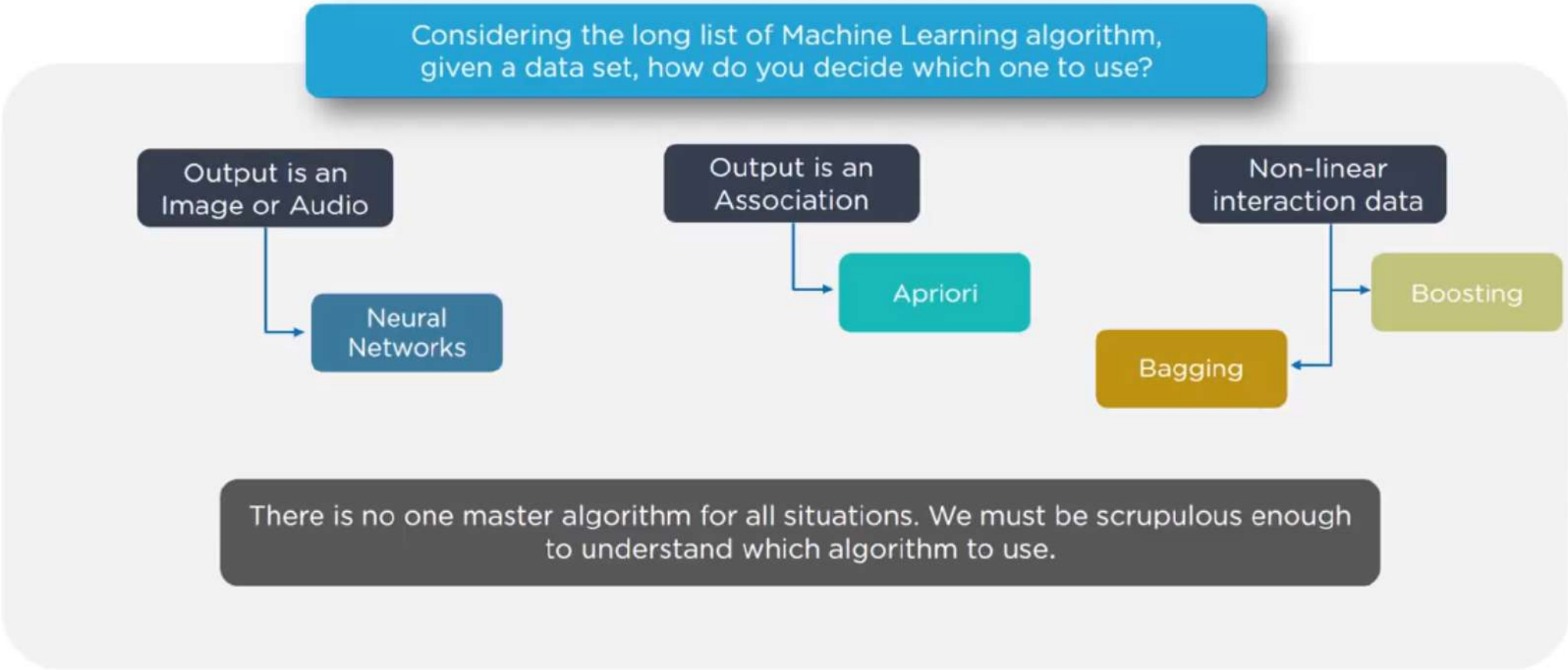
Bias in a Machine Learning model occurs when the predicted values are farther from the actual values

Low bias indicates a model where the prediction values are very close to the actual ones

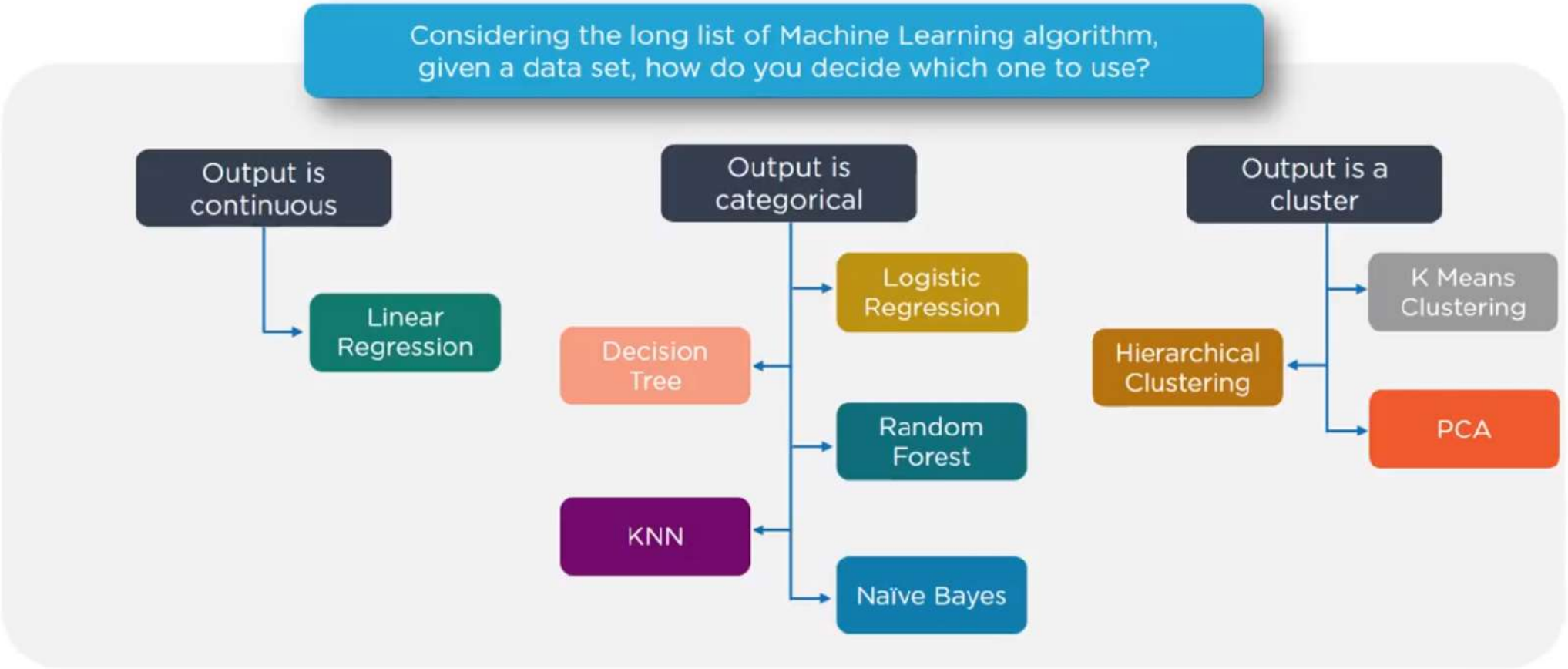
Underfitting: High bias can cause an algorithm to miss the relevant relations between features and target outputs

High Bias
Low Variance

24 Machine Learning Interview Questions



24 Machine Learning Interview Questions

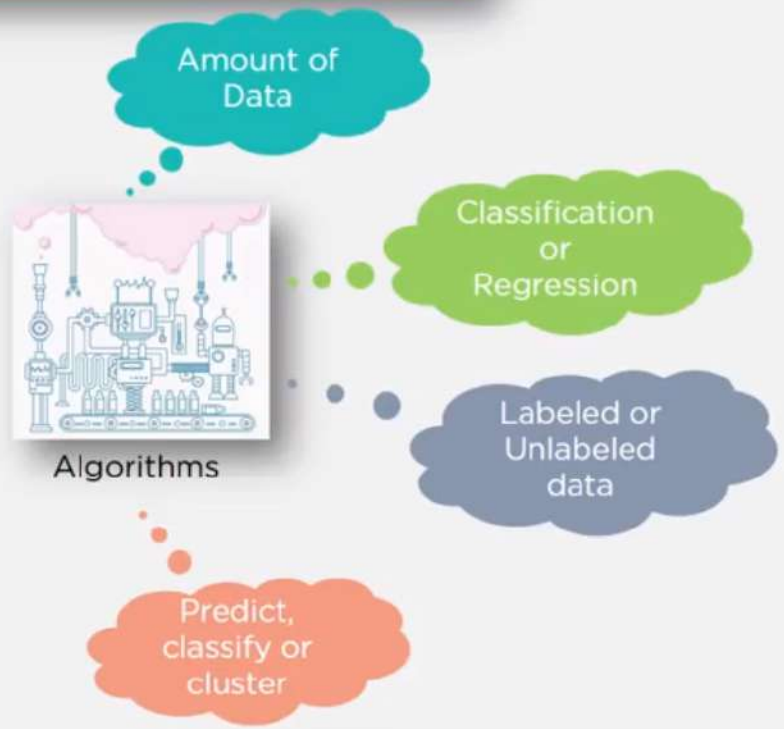


24 Machine Learning Interview Questions

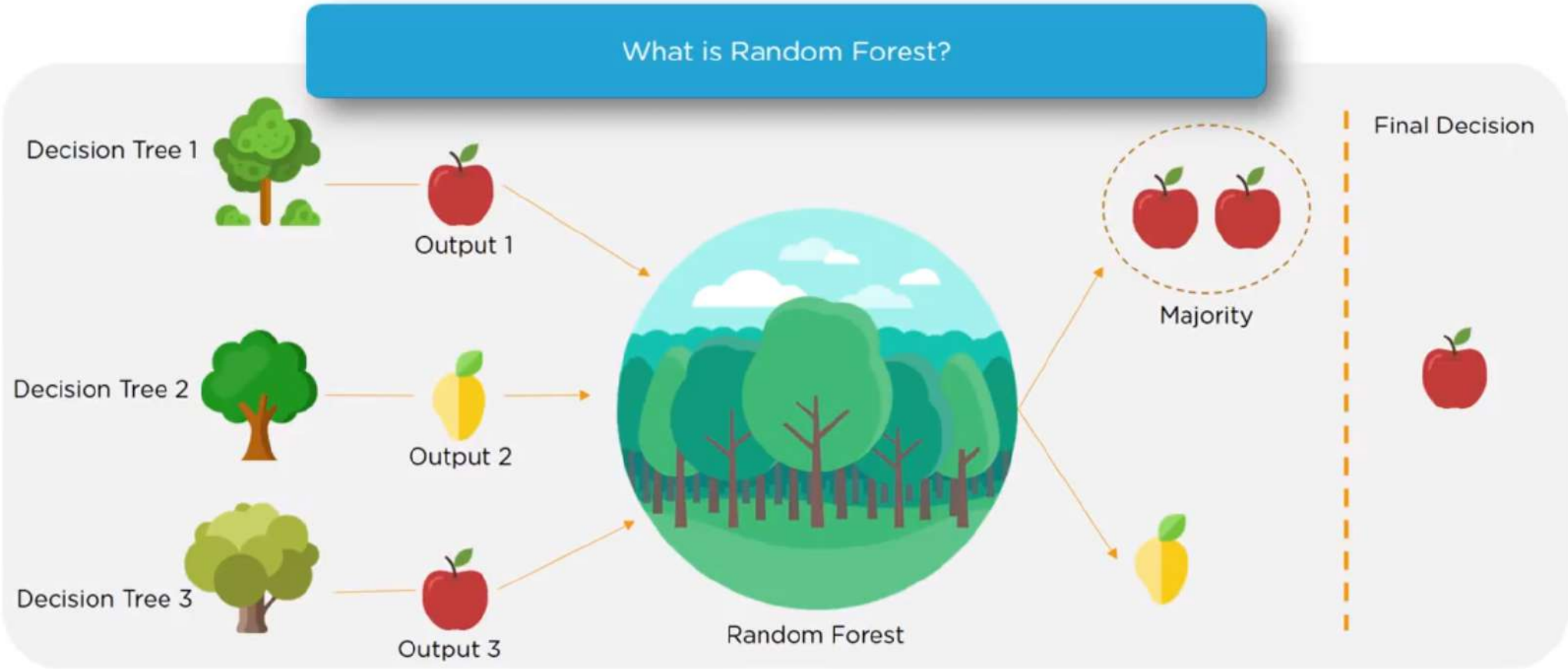
Considering the long list of Machine Learning algorithm, given a data set, how do you decide which one to use?

Choosing an algorithm depends on following questions:

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



23 Machine Learning Interview Questions



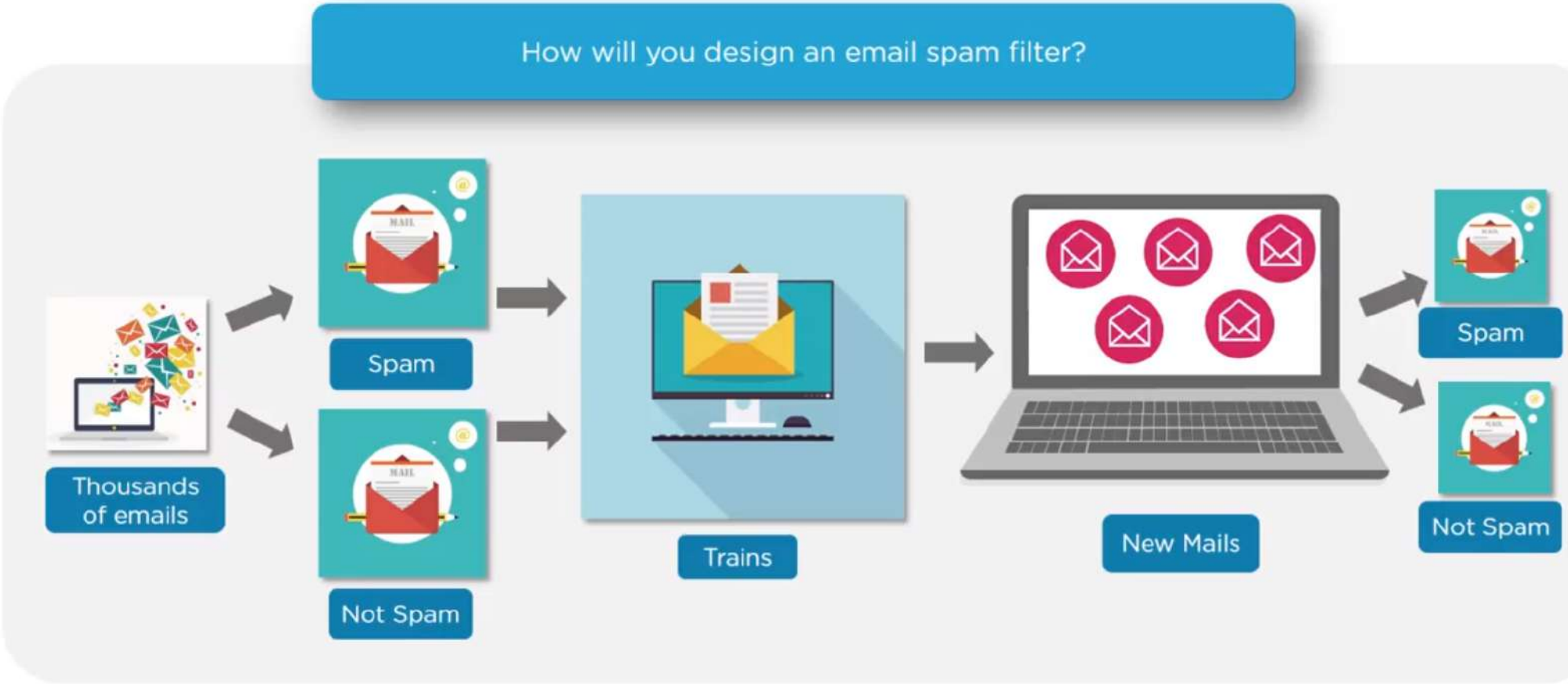
23 Machine Learning Interview Questions

What is Random Forest?

Random Forest is a Supervised Machine Learning Algorithm that is generally used for classification problems

Random forest operates by constructing multiple Decision Trees during training phase.
The Decision of the majority of the trees is chosen by the random forest as the final decision

22 Machine Learning Interview Questions



22 Machine Learning Interview Questions

How will you design an email spam filter?

Building a spam filter involves the following processes:

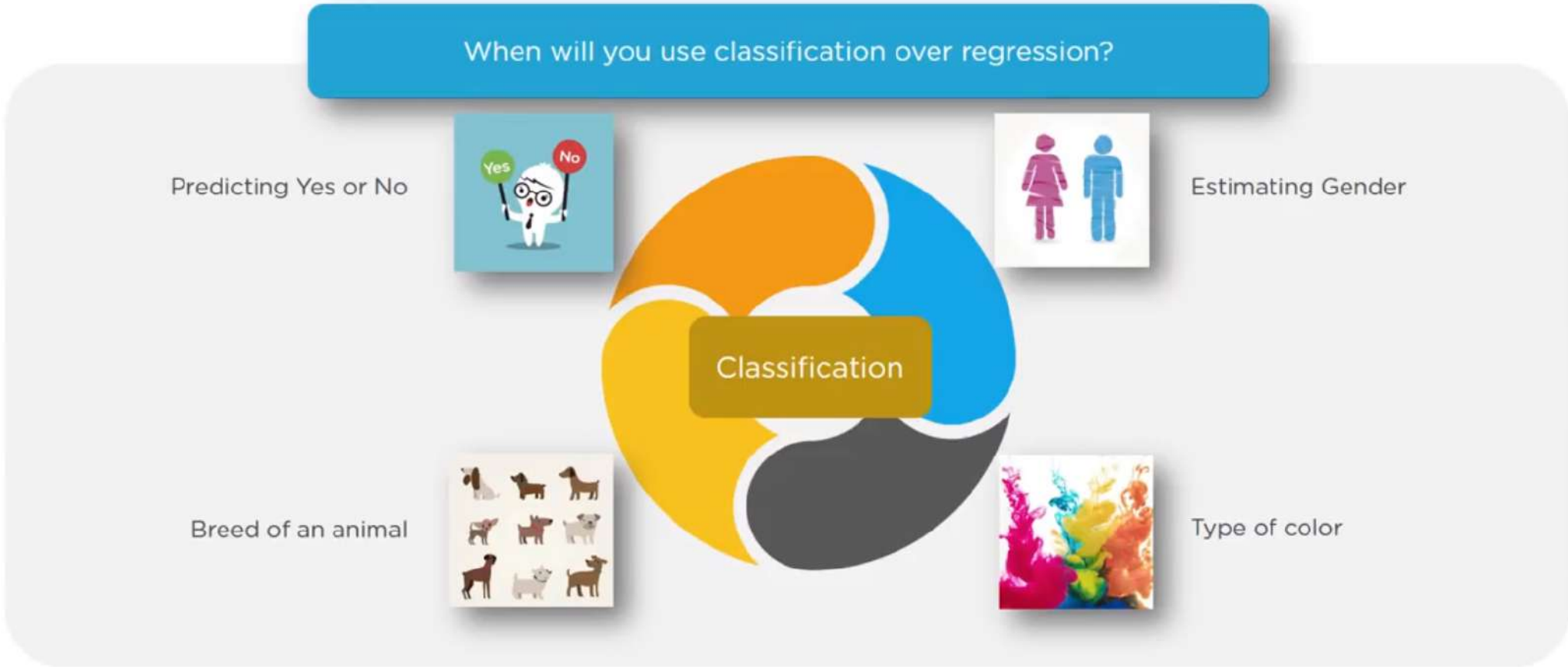
- Email spam filter will be fed with thousands of emails
- Each of these emails will already have a label - 'spam' or 'not spam'
- The Supervised Machine Learning algorithm will then figure out which type of emails are being marked as spam based on spam words like lottery, free offer, no money, full refund, etc

Thousands of emails

Spam/Not Spam

Learns

21 Machine Learning Interview Questions



21 Machine Learning Interview Questions

When will you use classification over regression?

Classification is used when your target variable is *Categorical* in nature. While Regression is used when your target variable is *Continuous* in nature. Both belong to the category of *Supervised Machine Learning Algorithms*.

Classification problems could be estimating the Gender of a person, the type of color, if the result is True or False, etc.

Regression problems could be estimating sale and price of a product, predicting sports score, amount of rainfall, etc.

20 Machine Learning Interview Questions

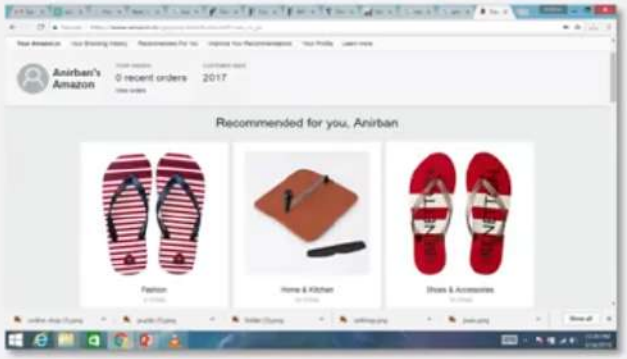
How is Amazon able to recommend other things to buy? How does it work?

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

Buy something from amazon

Association algorithm stores data and looks for patterns in it

Comes up with suggestions for the customers based on the patterns



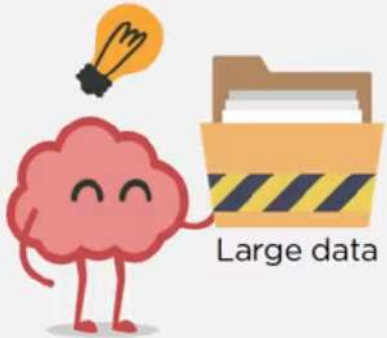
19 Machine Learning Interview Questions

How will you know which machine learning algorithm to choose for your classification problem?

- If accuracy is a concern, then one can test different algorithms and cross validate them.
- If the training dataset is small, one should use models that have low variance and high bias
- If the training dataset is large, one should use models with high variance and low bias.

Cross validate different algorithms

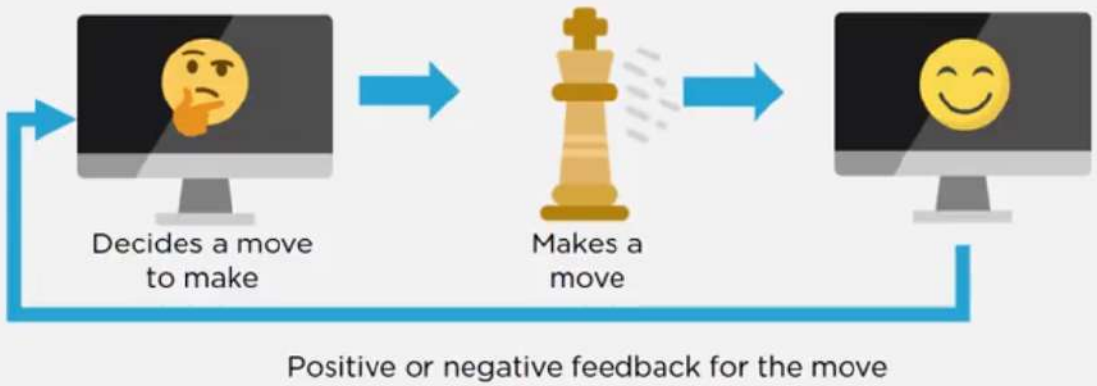
Algorithm with low bias/high variance



Algorithm with high bias/low variance

18 Machine Learning Interview Questions

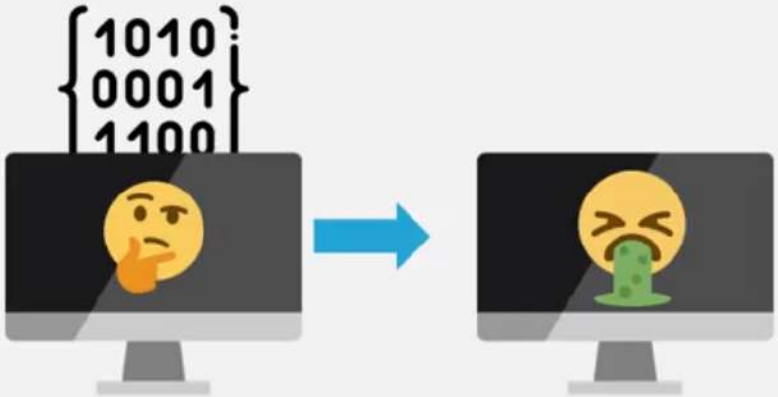
Can you explain how a system can play a game of chess using Reinforcement Learning?



- 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 mind for the next move it takes(learning)
- There's a reward for every correct decision the system takes and a punishment for a wrong one

18 Machine Learning Interview Questions

Can you explain how a system can play a game of chess using Reinforcement Learning?



Earlier, chess programs had to determine the best moves after a lot of research on numerous factors

Building a machine designed to play such games would require a lot of rules to be specified

17 Machine Learning Interview Questions

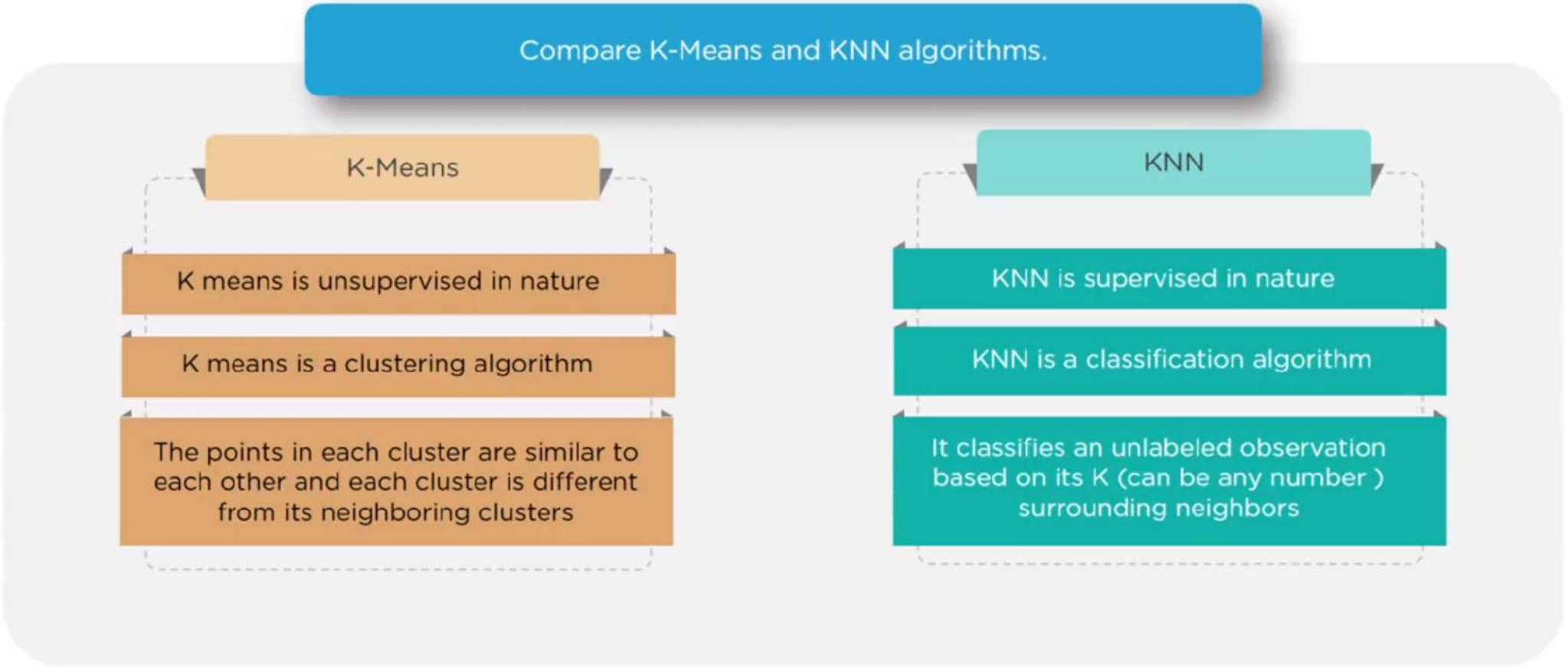
What is 'naive' in the Naive Bayes classifier?

- It's called "naive" because it makes assumptions that may or may not turn out to be correct
- The algorithm assumes 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, irrespective of other features and this assumption may or may not be right (E.g. Apple matches the description too).

Cherry

Apple

16 Machine Learning Interview Questions



15 Machine Learning Interview Questions

What is the difference between inductive Machine Learning and deductive Machine Learning?

Inductive Learning	VS	Deductive Learning
It observes instances based on defined principles to draw conclusion		It draws conclusion from experiences
E.g.: Explaining a kid to keep away from fire by showing a video where fire causes damage		E.g.: Let the kid play with fire. If he gets a burn, he will learn that it is a dangerous thing and will refrain from doing the same mistake again

14 Machine Learning Interview Questions

What is the difference between supervised and unsupervised Machine Learning?

Supervised Learning

Machine Learning model learns from the past input data and makes future prediction as output

Unsupervised Learning

Machine Learning model uses unlabeled input data and allows the algorithm to act on that information without guidance

13 Machine Learning Interview Questions

What are the unsupervised Machine Learning techniques?

Association

- In an Association problem, we identify patterns of associations between different variables or items
- In e-commerce websites, they're able to suggest other items for you to buy, based on the prior purchases that you've done, spending habits, items in your wish-list, other customers' purchase habits and so on.

13 Machine Learning Interview Questions

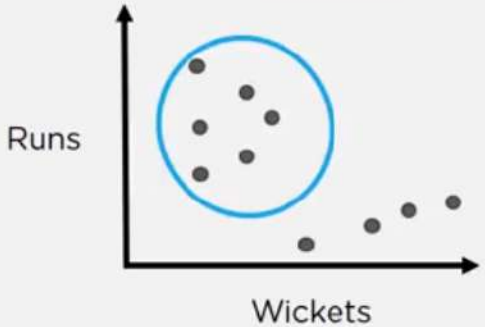
What are the unsupervised Machine Learning techniques?

Clustering

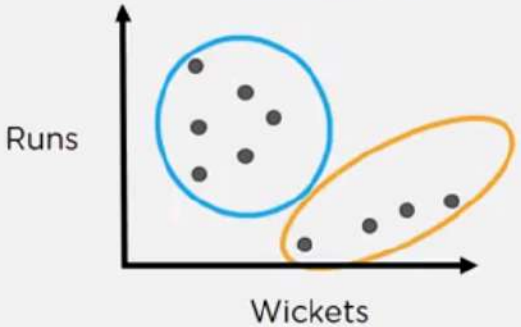
- Clustering problems involve data to be divided into subsets. These subsets, also called clusters contain data that are similar to each other.
- Different clusters reveal different details about the objects, making it different from classification or regression.

Here, we have our dataset with x and y coordinates that represent the runs scored and wickets taken by the players

We can see that this cluster has players with high runs and low wickets

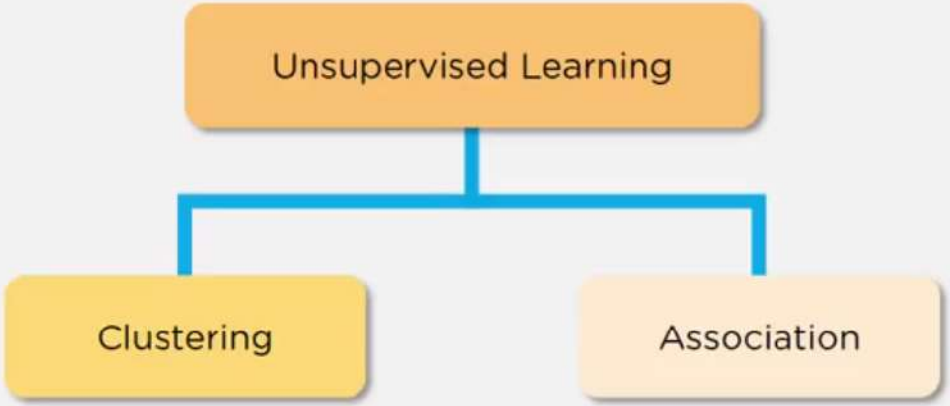


And here, we can see that this cluster has players with high wickets and low runs



13 Machine Learning Interview Questions

What are the unsupervised Machine Learning techniques?



12 Machine Learning Interview Questions

What is semi supervised Machine Learning?

Supervised Learning uses training data that is completely labeled and Unsupervised Learning uses no training data

In case of Semi-Supervised Learning, the training data contains of a small amount of labeled and a large amount of unlabeled data

11 Machine Learning Interview Questions

What are the applications of supervised Machine Learning in modern businesses?

Email Spam Detection

Sentiment Analysis

Healthcare Diagnosis

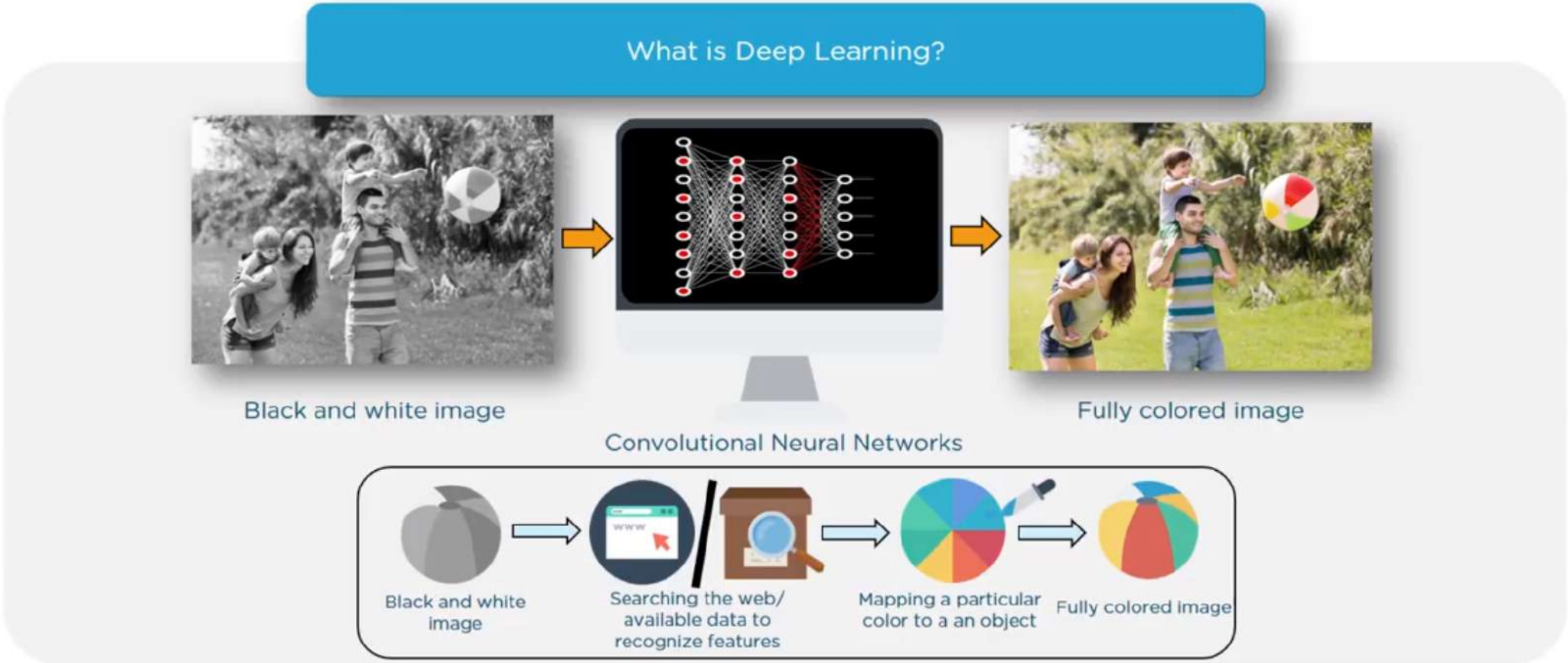
Fraud Detection

16 Machine Learning Interview Questions

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. • Needs only a small amount of training data. • Works well on low-end systems. • Most features need to be identified in advance and manually coded. • The problem is divided into parts and solved individually and then combined.	• Enables machines to take decisions with the help of artificial neural networks. • Needs a large amount of training data. • Needs high end systems to work. • The machine learns the features from the data it is provided. • The problem is solved in an end-to-end manner.

09 Machine Learning Interview Questions

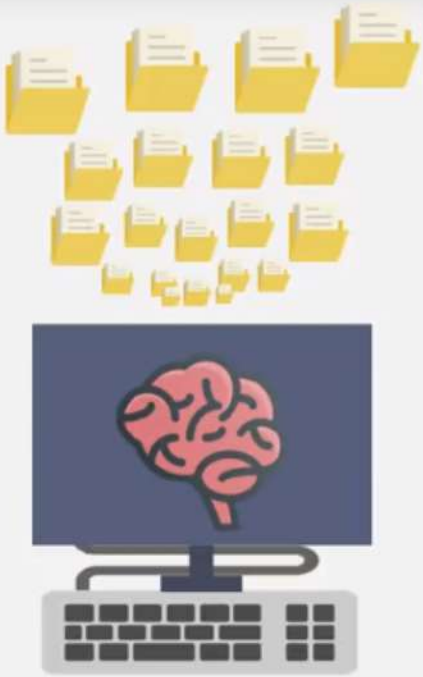


09 Machine Learning Interview Questions

What is Deep Learning?

Deep Learning involves systems that think and learn like humans using artificial neural networks

- Performance improves with more data
- Better scalability
- Problems solved in an end-to-end method



- Best features are selected by the system
- Is a subset of Machine Learning
- Lesser testing time

08 Machine Learning Interview Questions

What are the three stages to build a model in Machine Learning?

- ❑ Model Building : Choose the 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 apply the final model

Model Building

- Choose algorithm
- Train the model using Training dataset

Model Testing

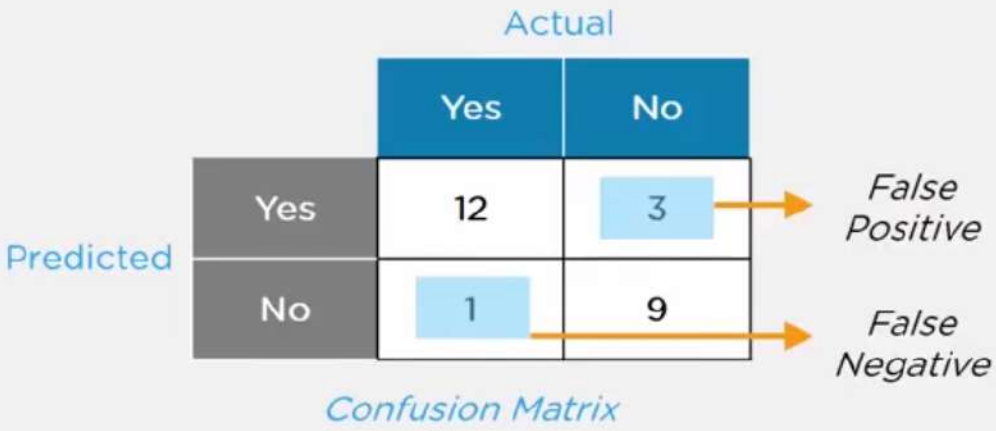
- Test the model with new data
- Check the accuracy of the model

Applying The Model

- Make required changes after testing
- Apply for real time projects

07 Machine Learning Interview Questions

What is false positive and false negative and how are they significant?



False Positive are those cases which wrongly get classified as True but are actually False

False Negative similarly are those cases which wrongly get classified as False but are True

True Positive: 12
False Positive: 3
True Negative: 9
False Negative: 1

06 Machine Learning Interview Questions

Explain confusion matrix with respect to Machine Learning algorithms.

- 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 is called matching matrix)
- Confusion matrix has two dimensions:
  1. Actual
  2. Predicted
- It also has identical sets of features in both these dimensions

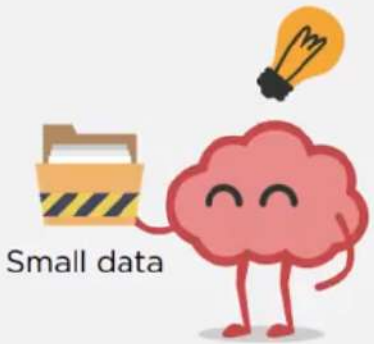
		Actual: Yes	Actual: No
Predicted	Yes	12	3
	No	1	9

Confusion Matrix

05 Machine Learning Interview Questions

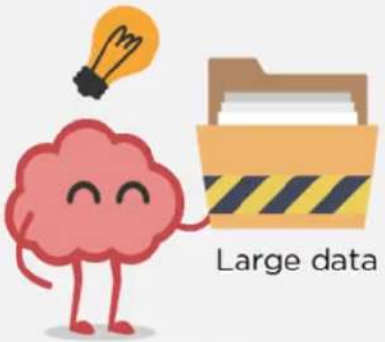
How can you choose a classifier based on training set size?

When the training set is small, a model that has a high bias and low variance seems to work better because they are less likely to overfit. For e.g. Naïve Bayes works best



Algorithm with low bias/high variance

When the training set is large, models with low bias and high variance tend to perform better as they work fine with complex relationships. E.g. Decision Tree



Algorithm with high bias/low variance

04 Machine Learning Interview Questions

How do you handle missing or corrupted data in a dataset?

The ways to handle missing / corrupted data is to drop those rows / columns or replace them completely with some other value

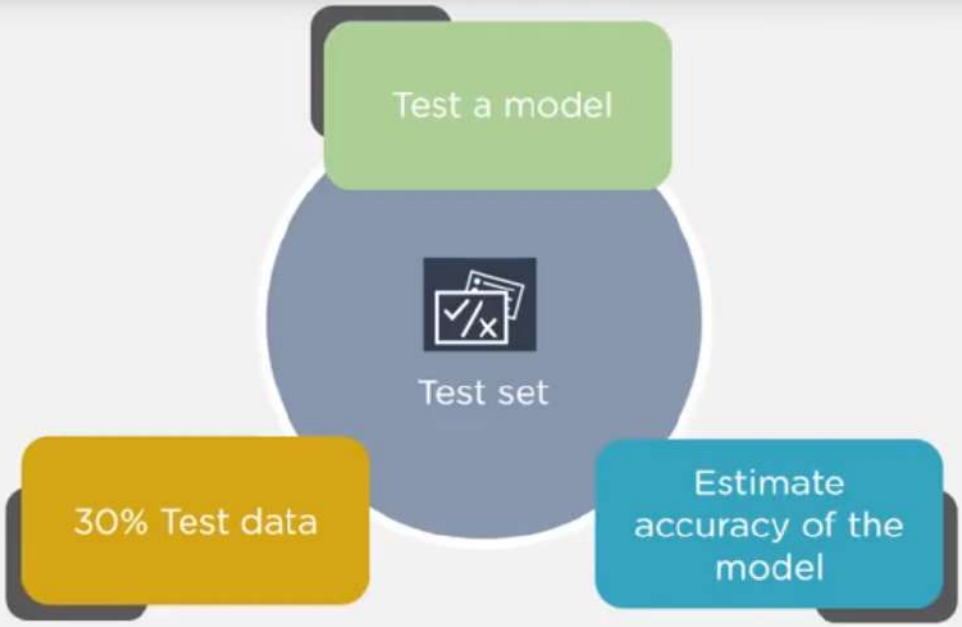
There are two useful methods in Panda:

a) Isnull() and dropna() will help finding the columns / rows with missing data and drop them

b) Fillna() will replace the wrong values with a placeholder value(0)

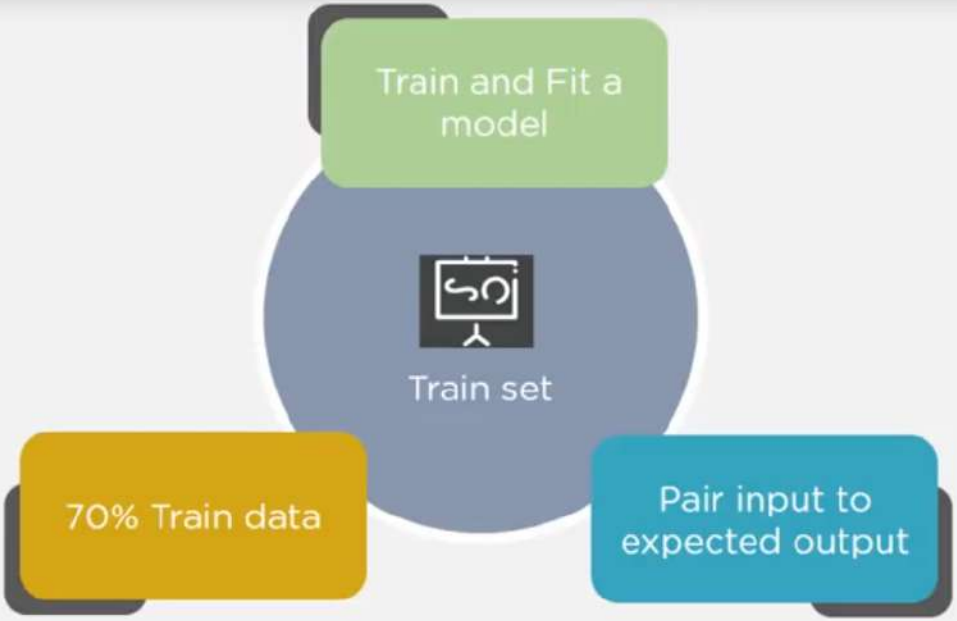
03 Machine Learning Interview Questions

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?



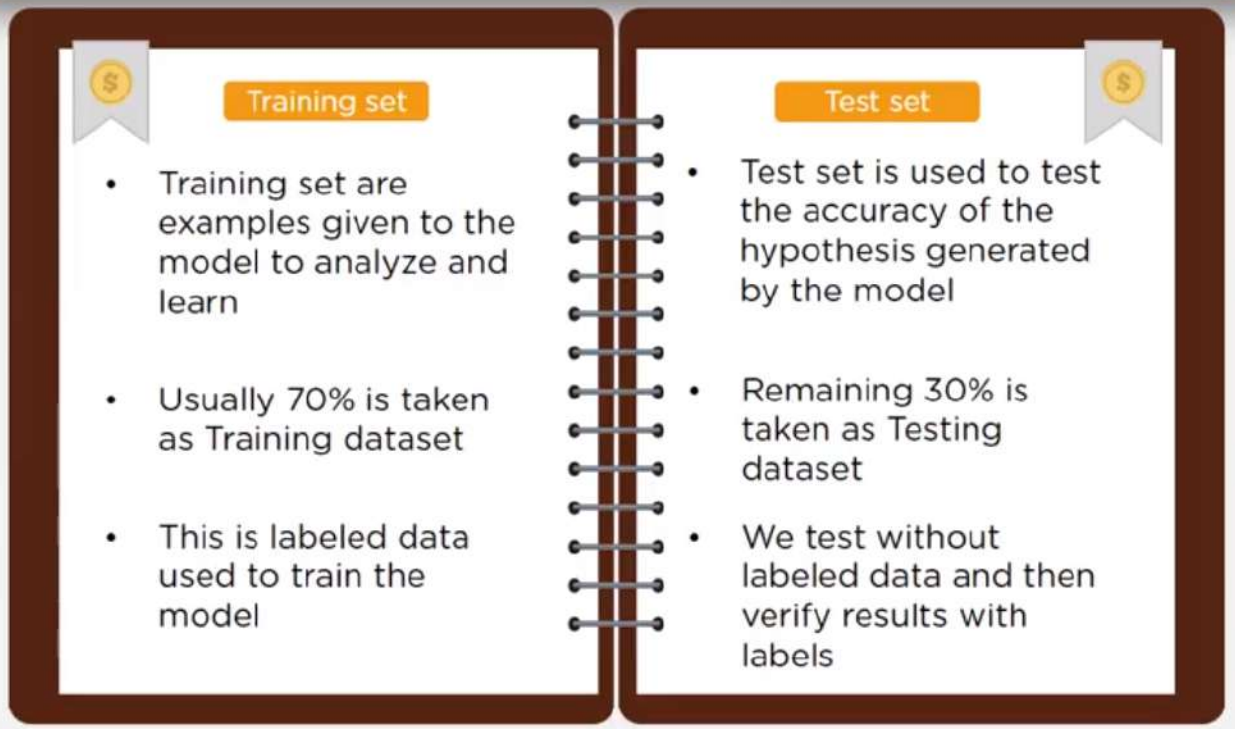
03 Machine Learning Interview Questions

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?



03 Machine Learning Interview Questions

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?



02 Machine Learning Interview Questions

What is overfitting? And how can you avoid it?

Simpler

Cross-validation

Punishing parameters

There are three main methods to avoid overfitting:

- ❑ Regularization: This involves a cost term for the features involved with the objective function
- ❑ Make 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

02 Machine Learning Interview Questions

What is overfitting? And how can you avoid it?

- Overfitting occurs when the model learns the training set too well. It takes up random fluctuations in the training data as concepts. These impact the model's ability to generalize and don't apply to new data

Training Dataset

Testing Dataset

Misclassified Data

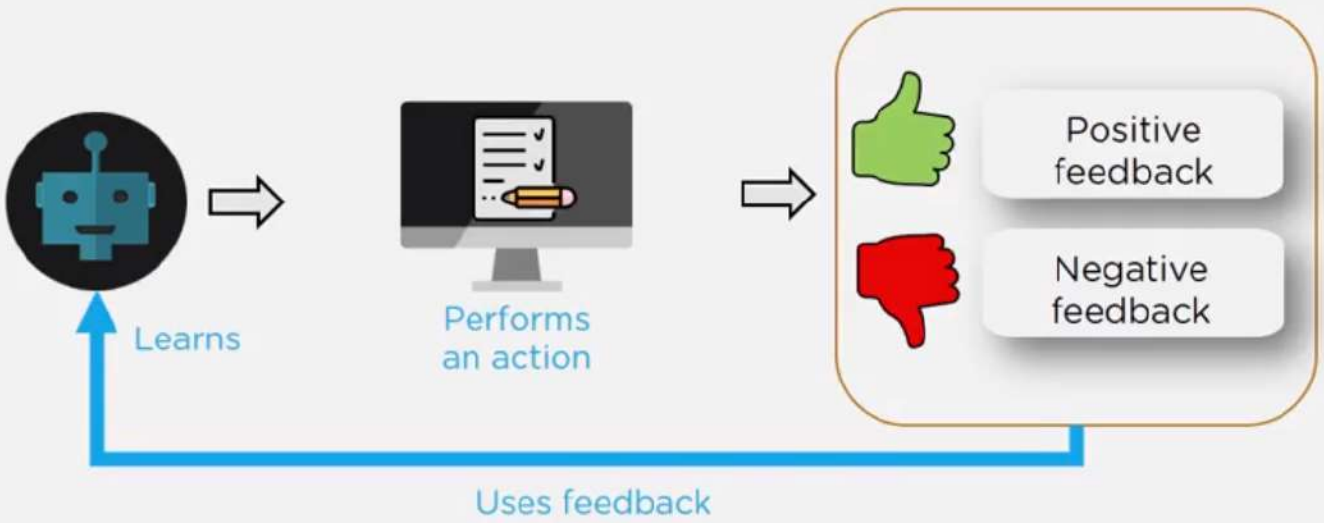
- Here the model is overfit to the training dataset and will give error when new testing dataset is introduced
- High loss and low accuracy is seen in the test dataset

01 Machine Learning Interview Questions

What are the different types of Machine Learning?

Reinforcement Learning

Using reinforcement learning, the model is able to learn based on the rewards it received for its previous action

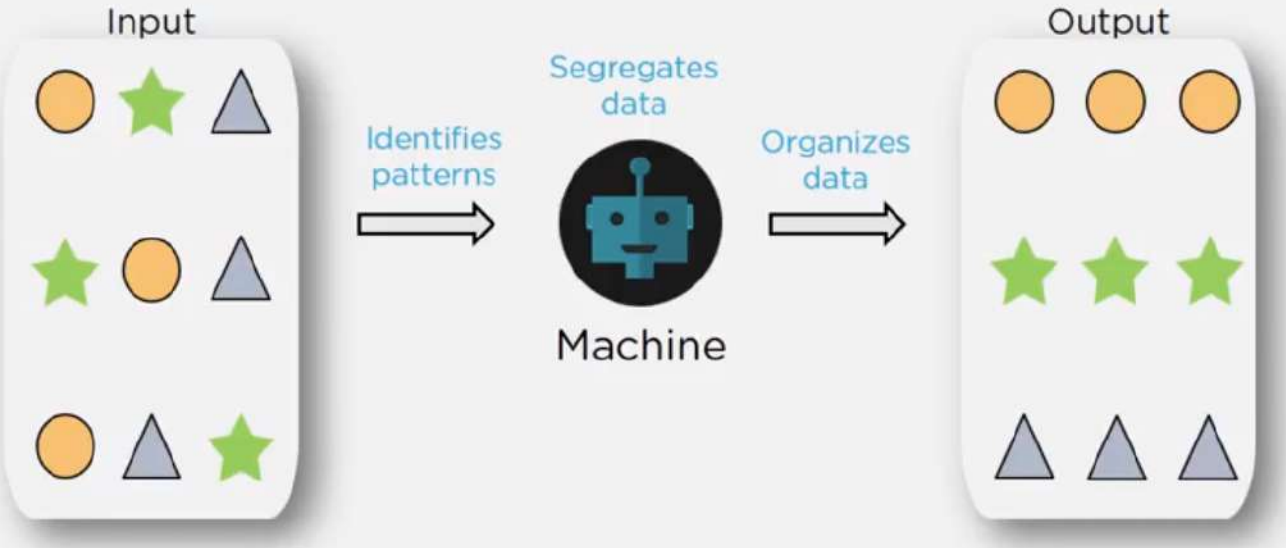


01 Machine Learning Interview Questions

What are the different types of Machine Learning?

Unsupervised Learning

Unsupervised learning, in which a model is able to identify patterns, anomalies and relationships in the input data



01 Machine Learning Interview Questions

What are the different types of Machine Learning?

Supervised Learning

In Supervised Machine Learning, a model makes predictions or takes decisions based on past data