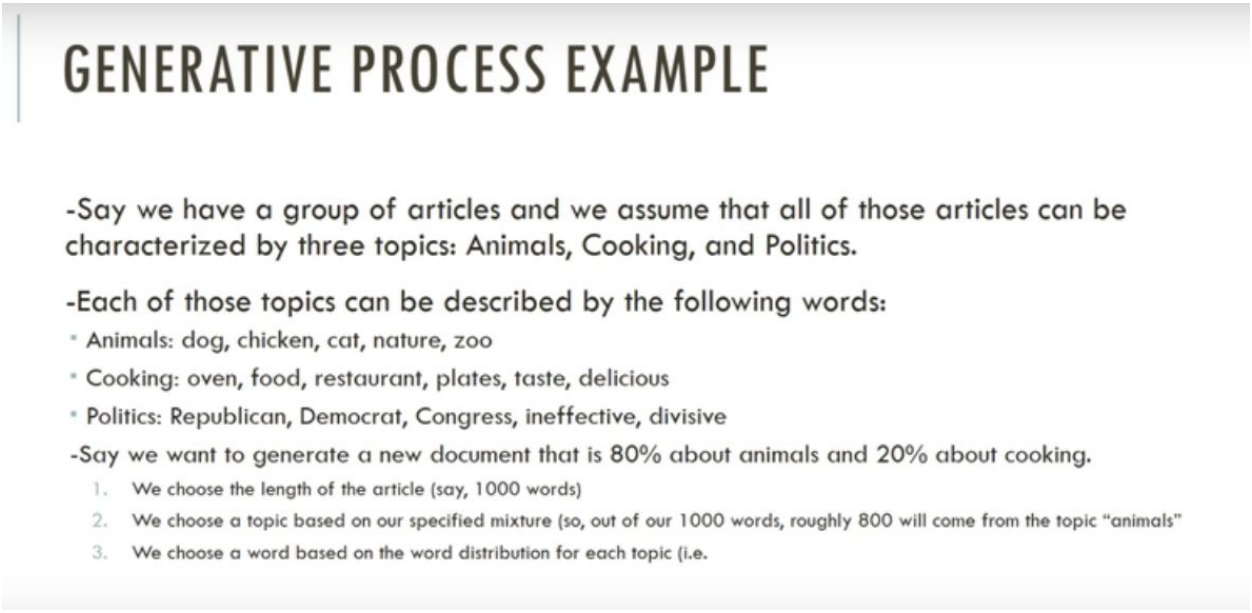
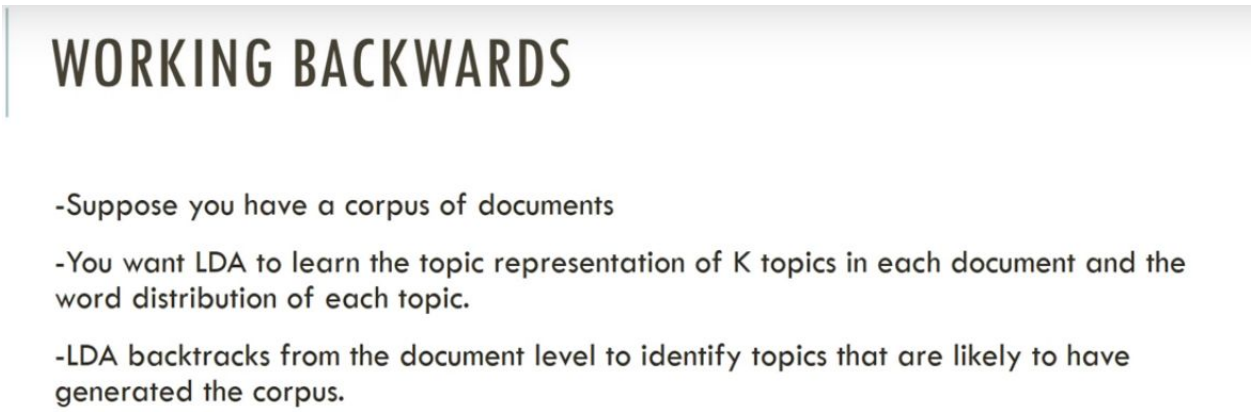
Machine Learning pipeline

LDA (https://www.youtube.com/watch?v=DWJYZq_fQ2A)
https://www.youtube.com/watch?v=fCmIceNqVog





WORKING BACKWARDS (CONT.)

1. Randomly assign each word in each document to one of the K topics.
2. For each document d:
   - Assume that all topic assignments except for the current one are correct.
   - Calculate two proportions:
     1. Proportion of words in document d that are currently assigned to topic t = p(topic t | document d)
     2. Proportion of assignments to topic t over all documents that come from this word w = p(word w | topic t)
   - Multiply those two proportions and assign w a new topic based on that probability. p(topic t | document d) * p(word w | topic t)
3. Eventually we'll reach a steady state where assignments make sense

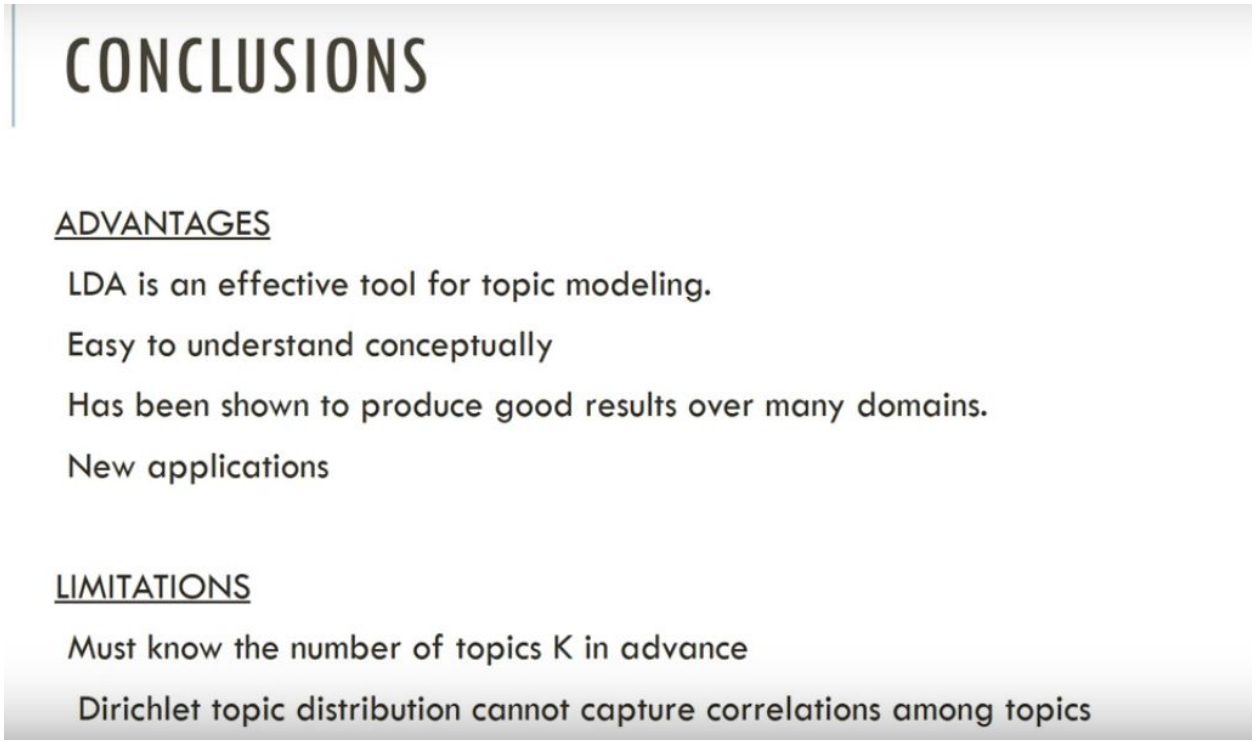
CONCLUSIONS

Documents are probability distributions over latent topics.

Topics are probability distributions over words.

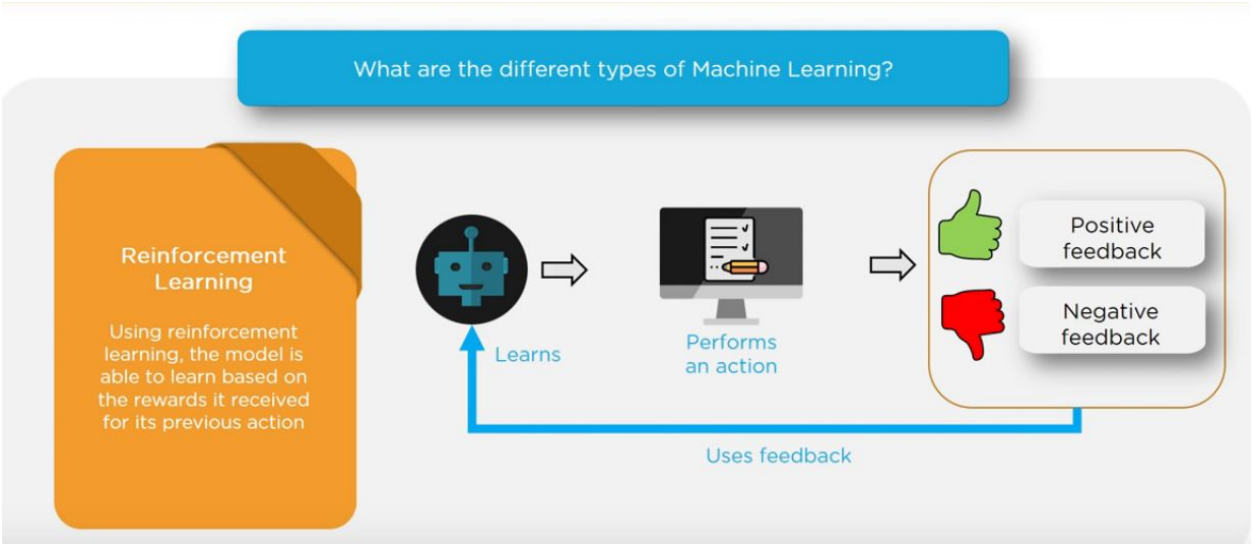
LDA takes a number of documents. It assumes that the words in each document are related. It then tries to figure out the "recipe" for how each document could have been created. We just need to tell the model how many topics to construct and it uses that "recipe" to generate topic and word distributions over a corpus. Based on that output, we can identify similar documents within the corpus.

Machine Learning pipeline



Machine Learning interview questions:

1. Type of machine learning
   Supervised, unsupervised, reinforcement



Machine Learning pipeline

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

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

Machine Learning pipeline

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

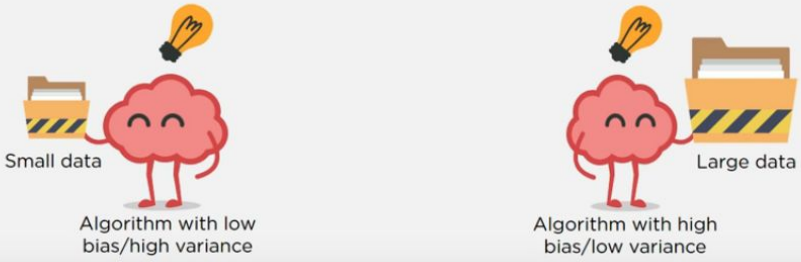
b) Fillna() will replace the wrong values with a placeholder value(0)

05 Machine Learning Interview Questions

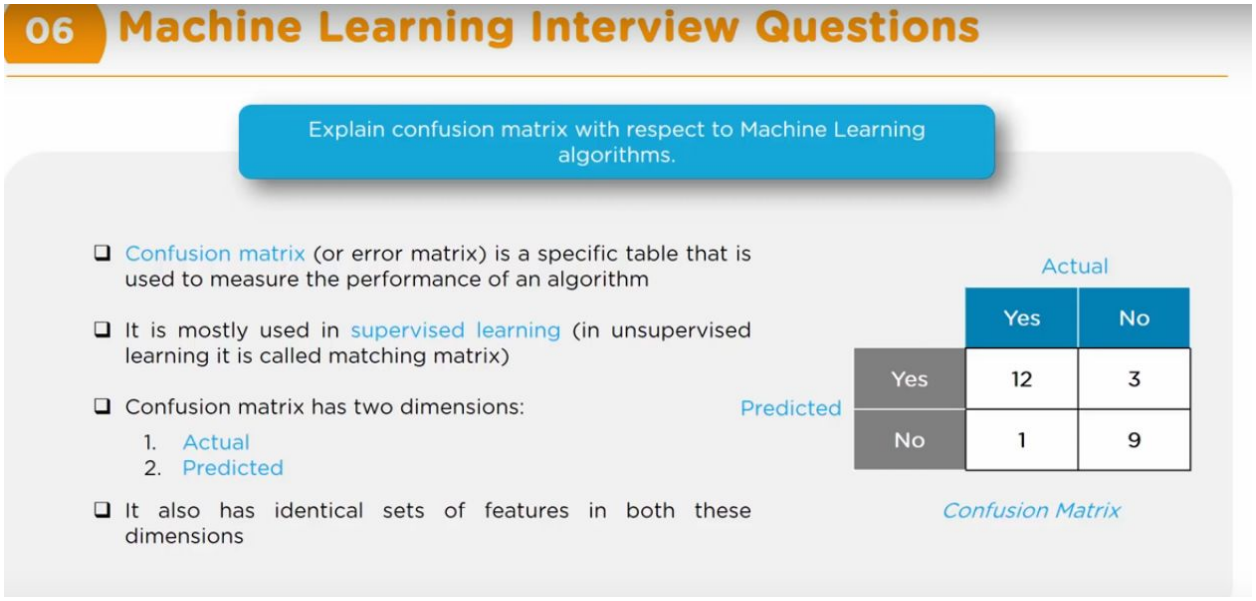
How can you choose a classifier based on training set size?

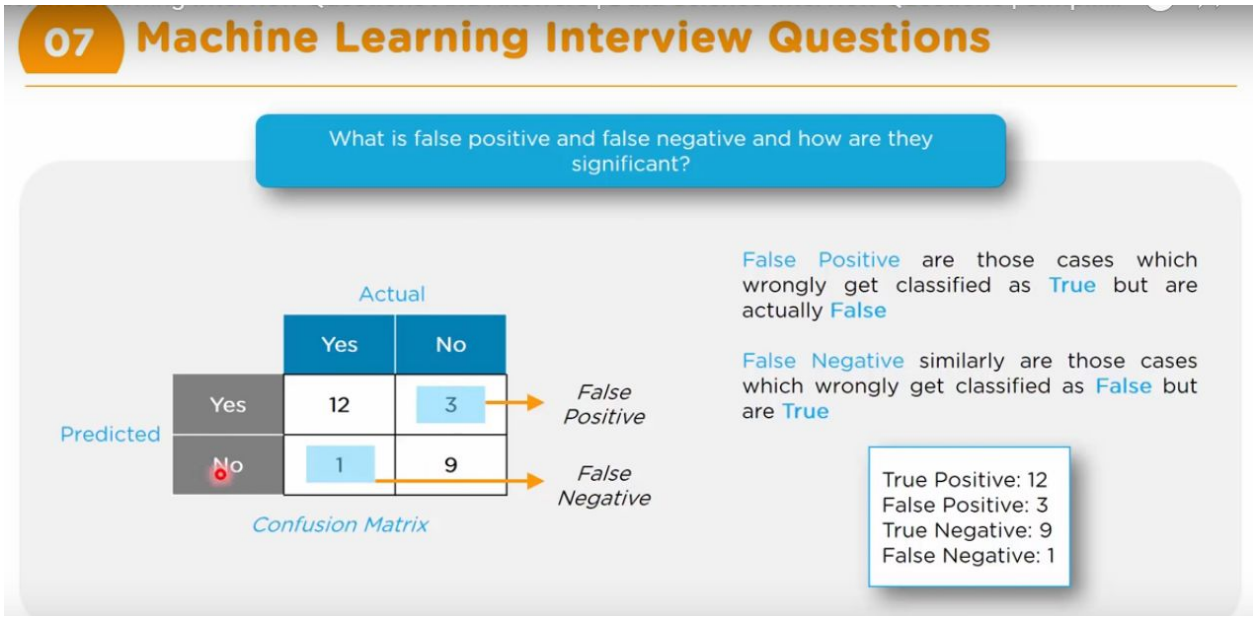
When the training set is small, a model that has a high bias and low variance seems to work better because they are less likely to overfit. For e.g. Naïve Bayes works best

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

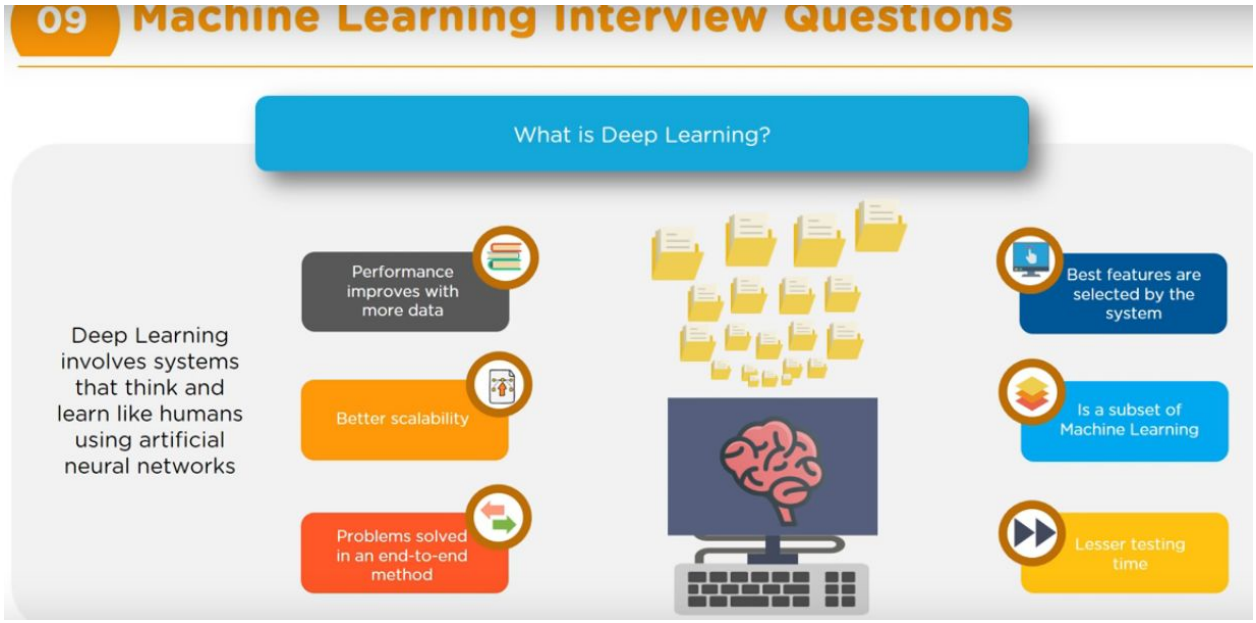


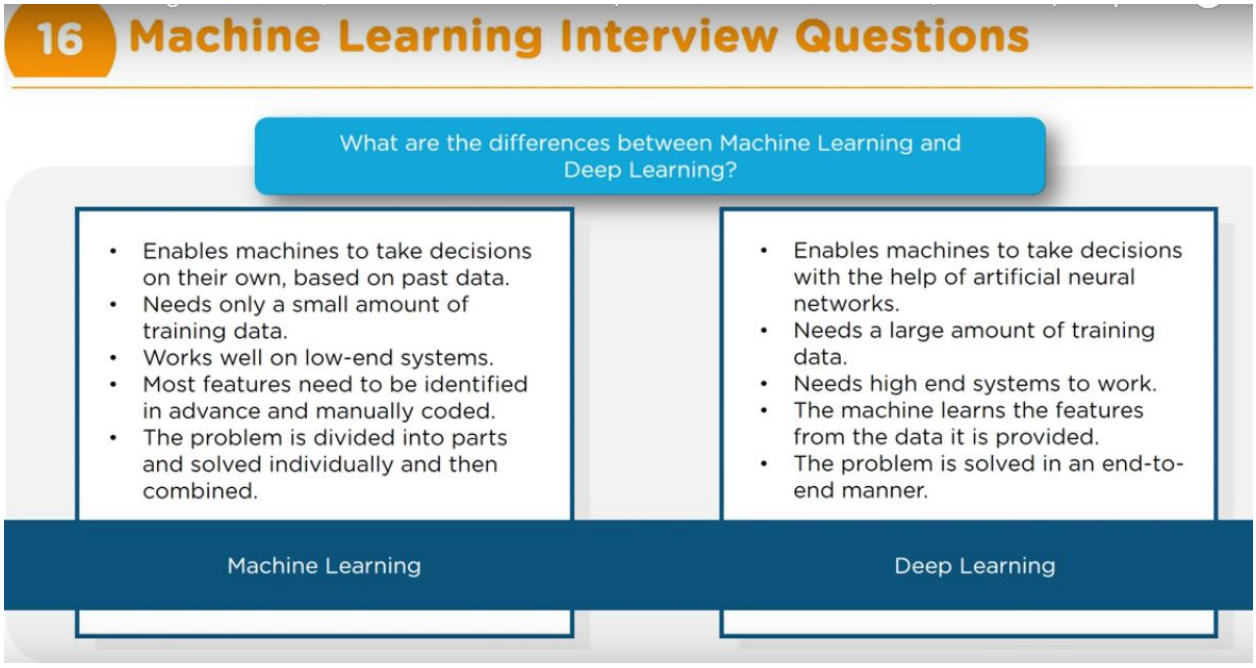
Machine Learning pipeline



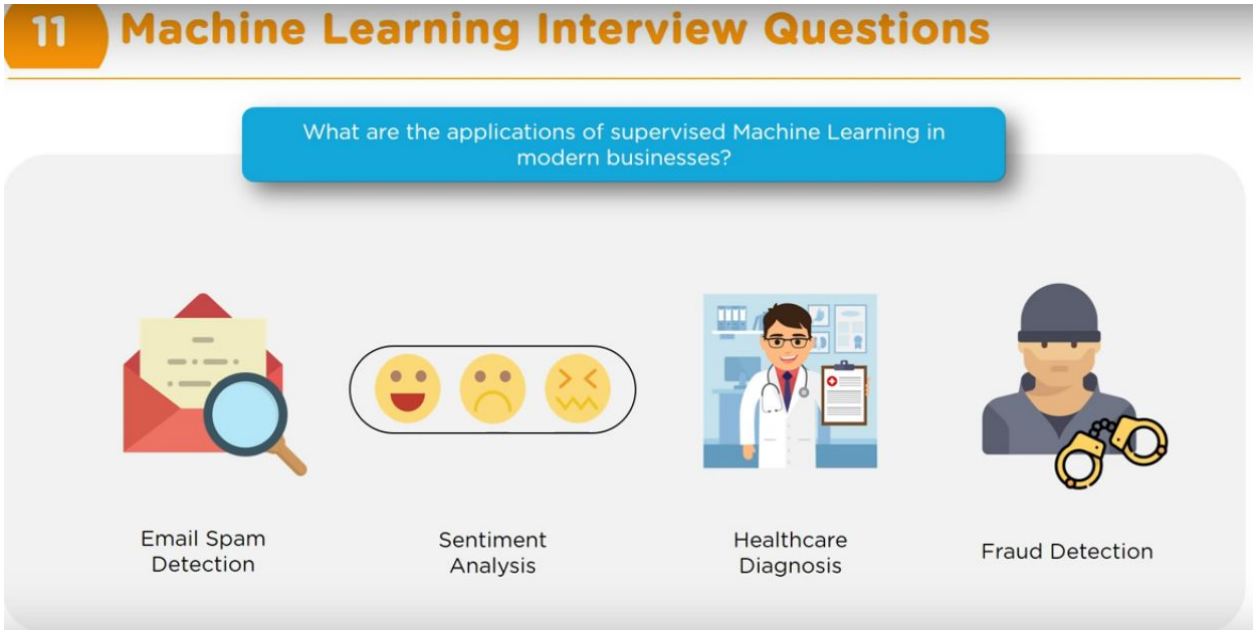


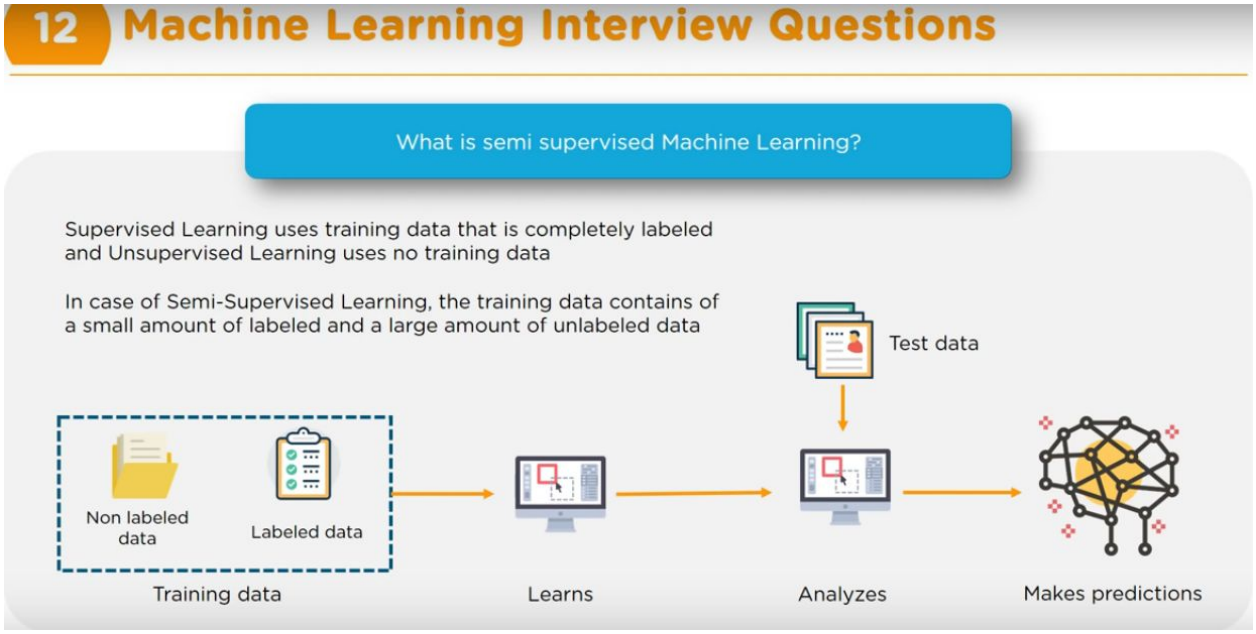
Machine Learning pipeline



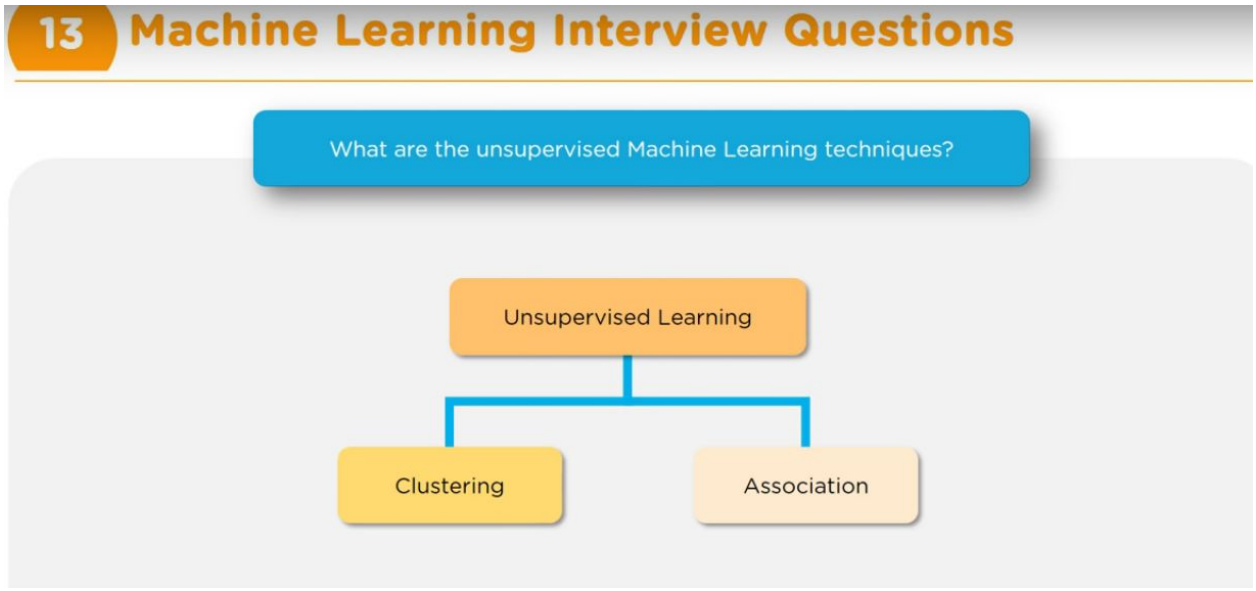


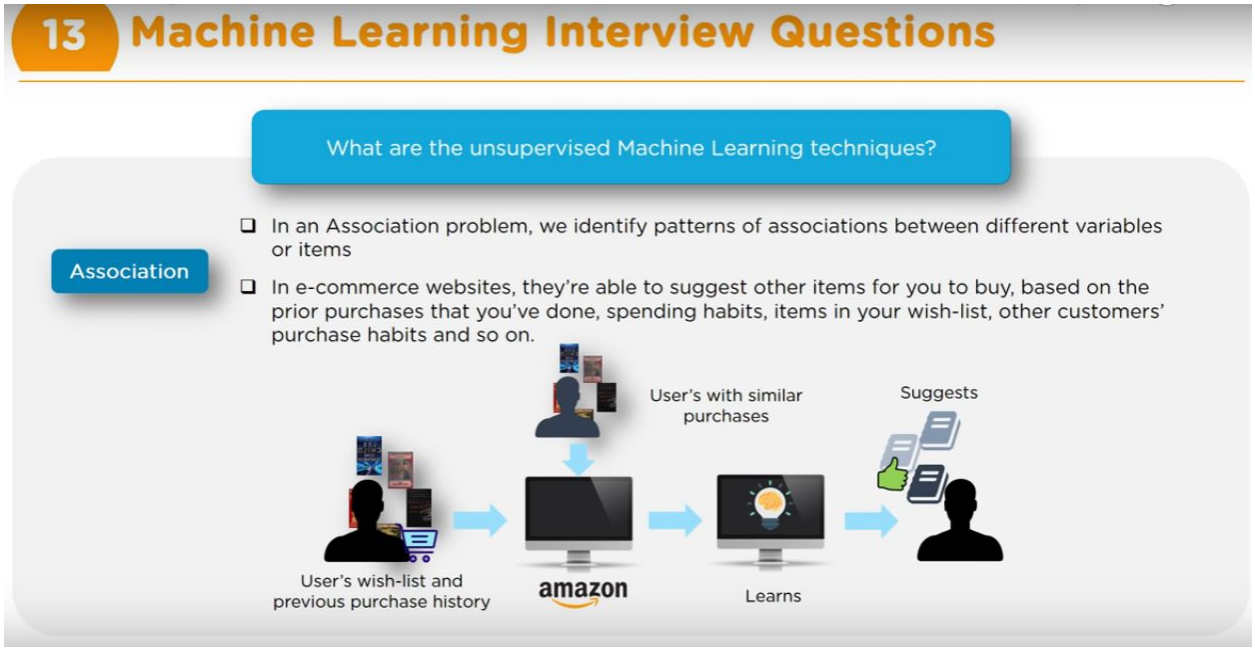
Machine Learning pipeline



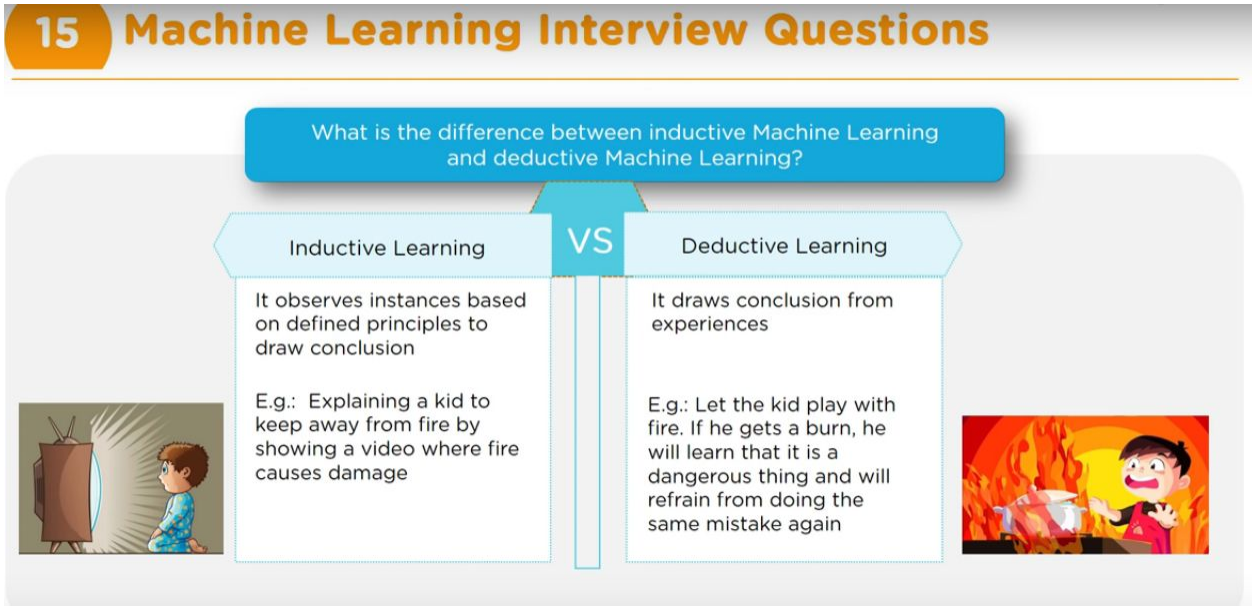


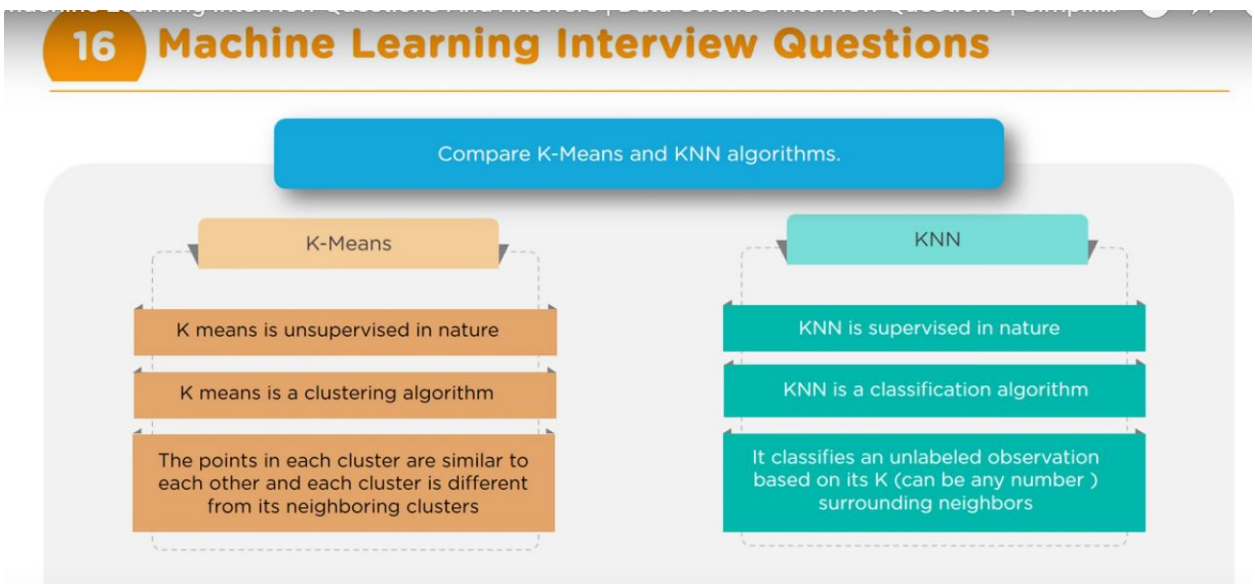
Machine Learning pipeline





Machine Learning pipeline





Machine Learning pipeline

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

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

Machine Learning pipeline

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

21 Machine Learning Interview Questions

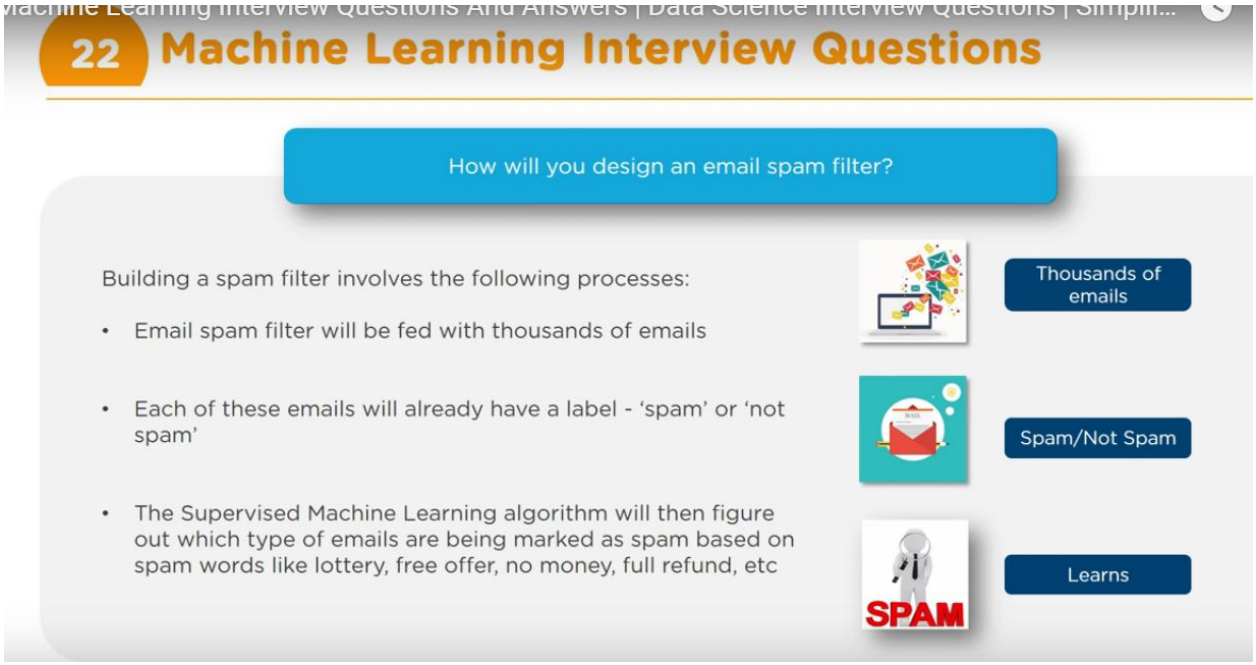
When will you use classification over regression?

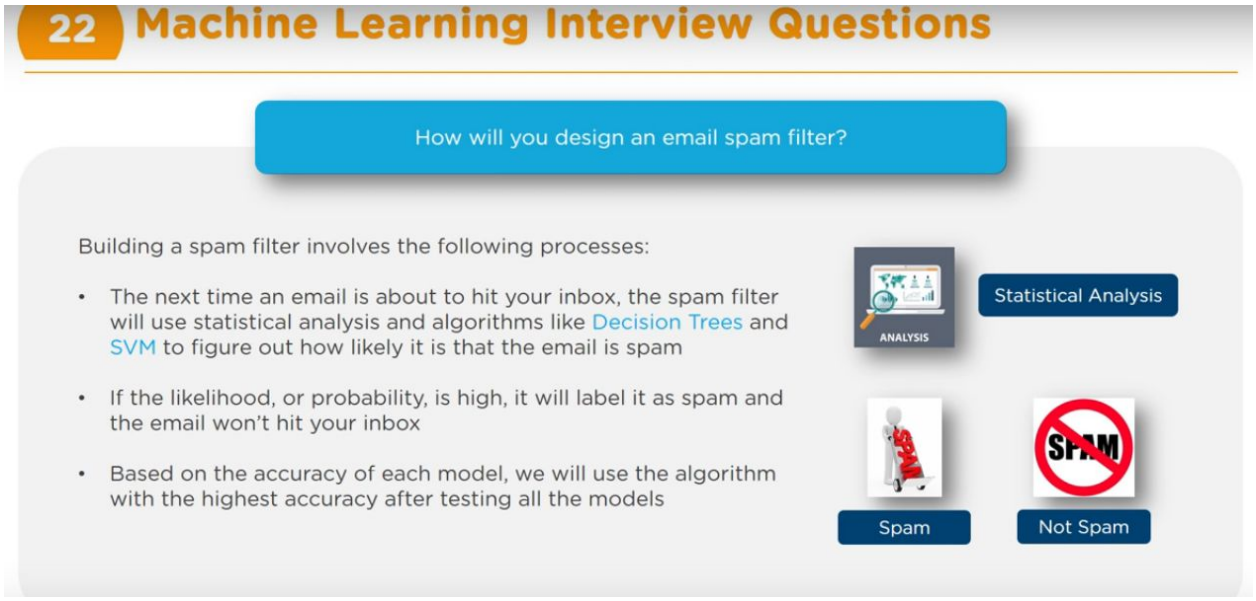
Classification is used when your target variable is *Categorical* in nature. While Regression is used when your target variable is *Continuous* in nature. Both belong to the category of *Supervised Machine Learning Algorithms*.

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

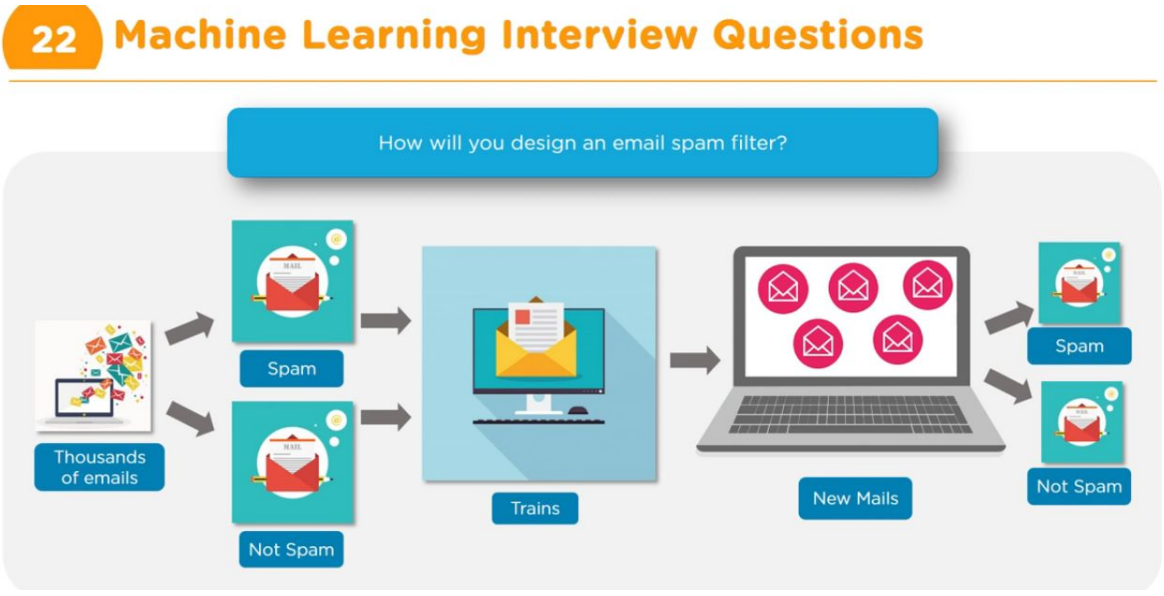
Regression problems could be estimating sale and price of a product, predicting sports score, amount of rainfall, etc.

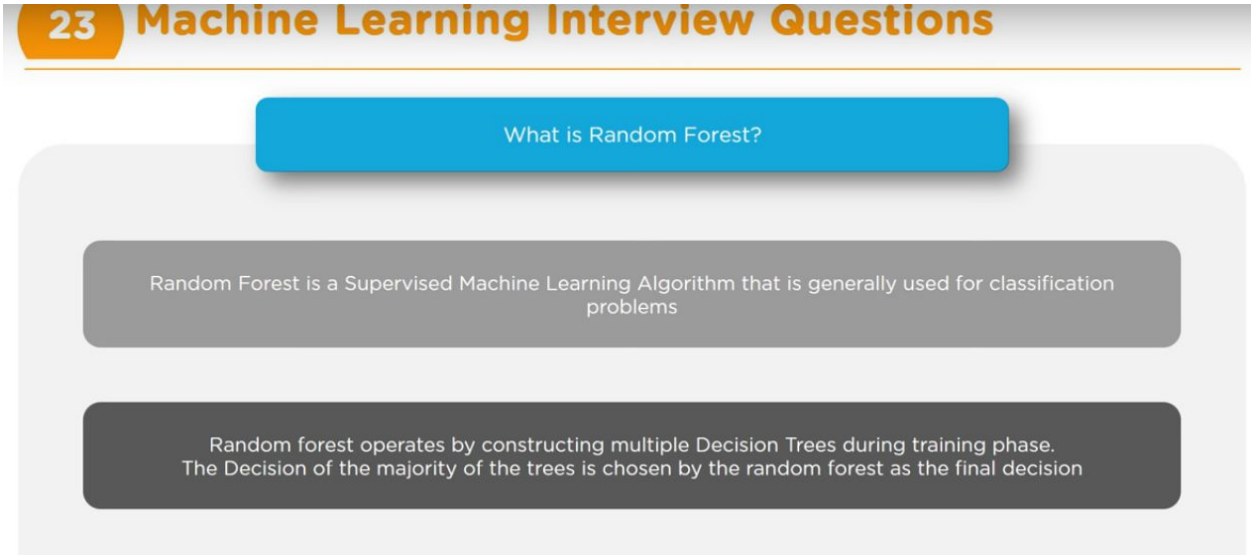
Machine Learning pipeline



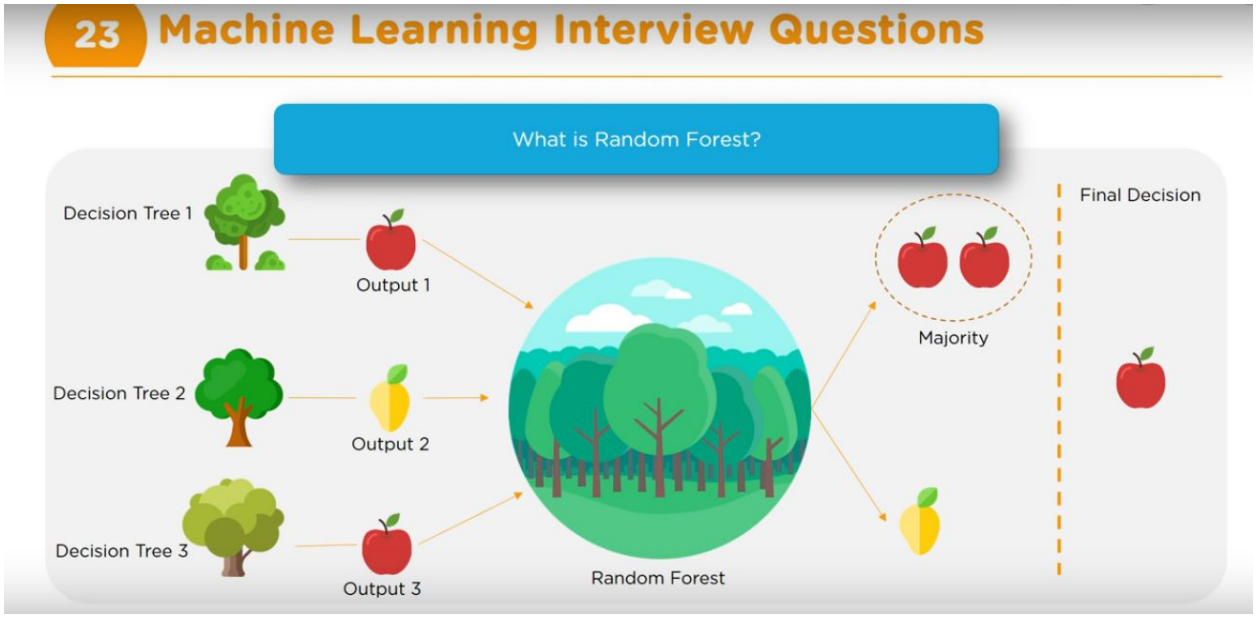


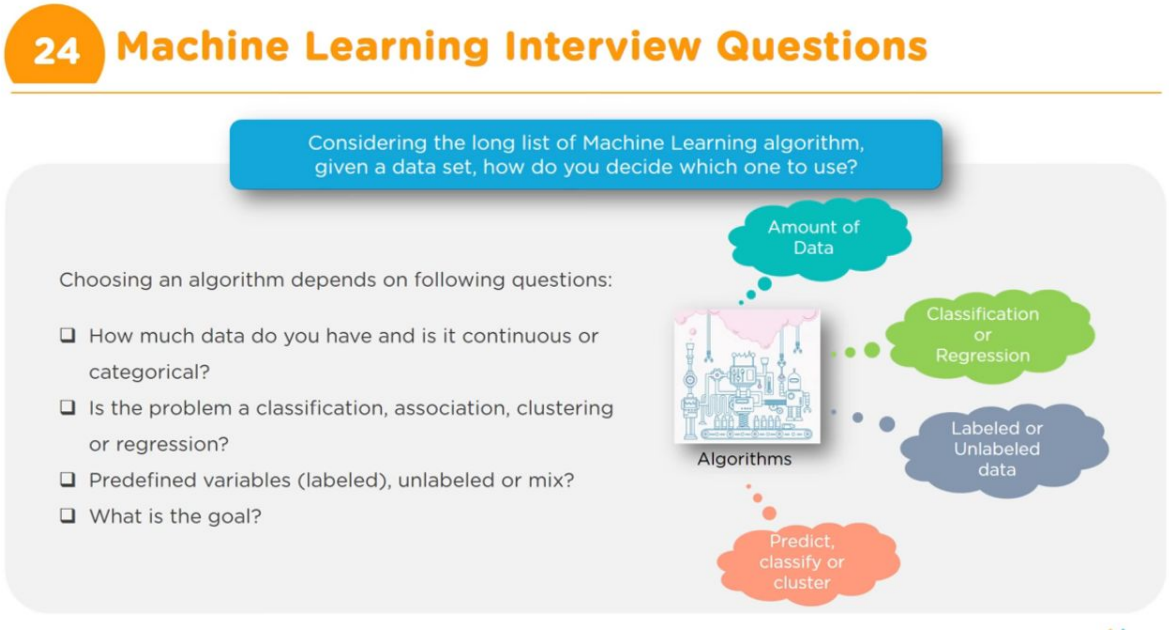
Machine Learning pipeline





Machine Learning pipeline





Machine Learning pipeline

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?

- Output is an Image or Audio → Neural Networks
- Output is an Association → Apriori
- Non-linear interaction data → Boosting, Bagging

There is no one master algorithm for all situations. We must be scrupulous enough to understand which algorithm to use.

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

Machine Learning pipeline

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

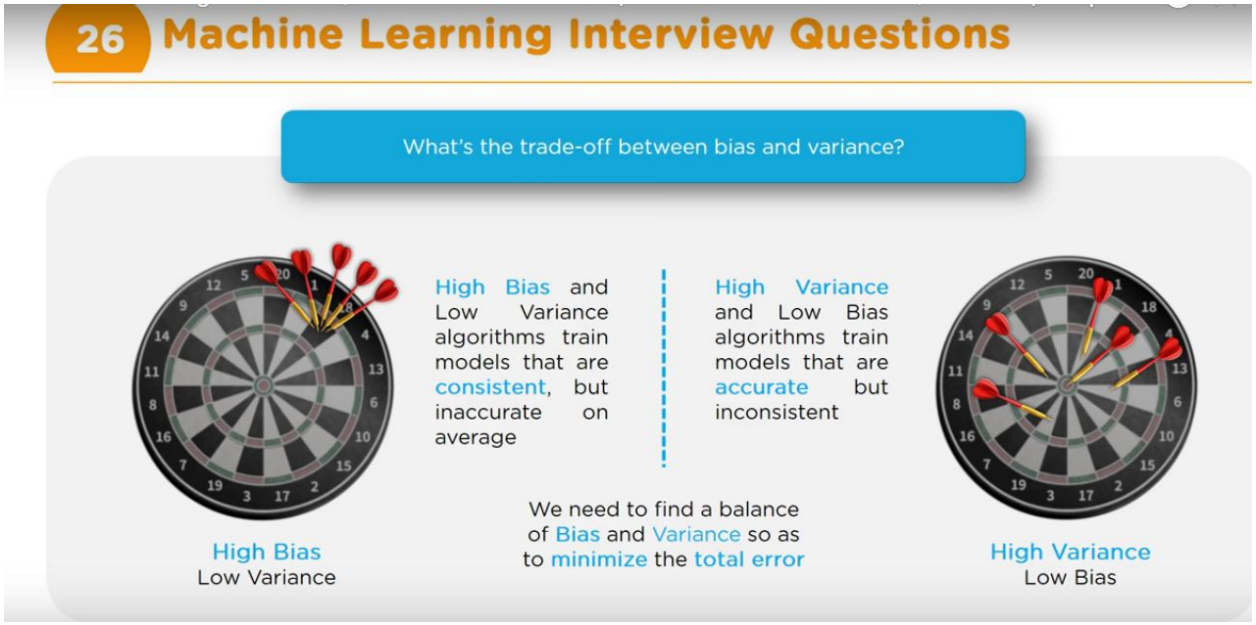
High Variance
Low Bias

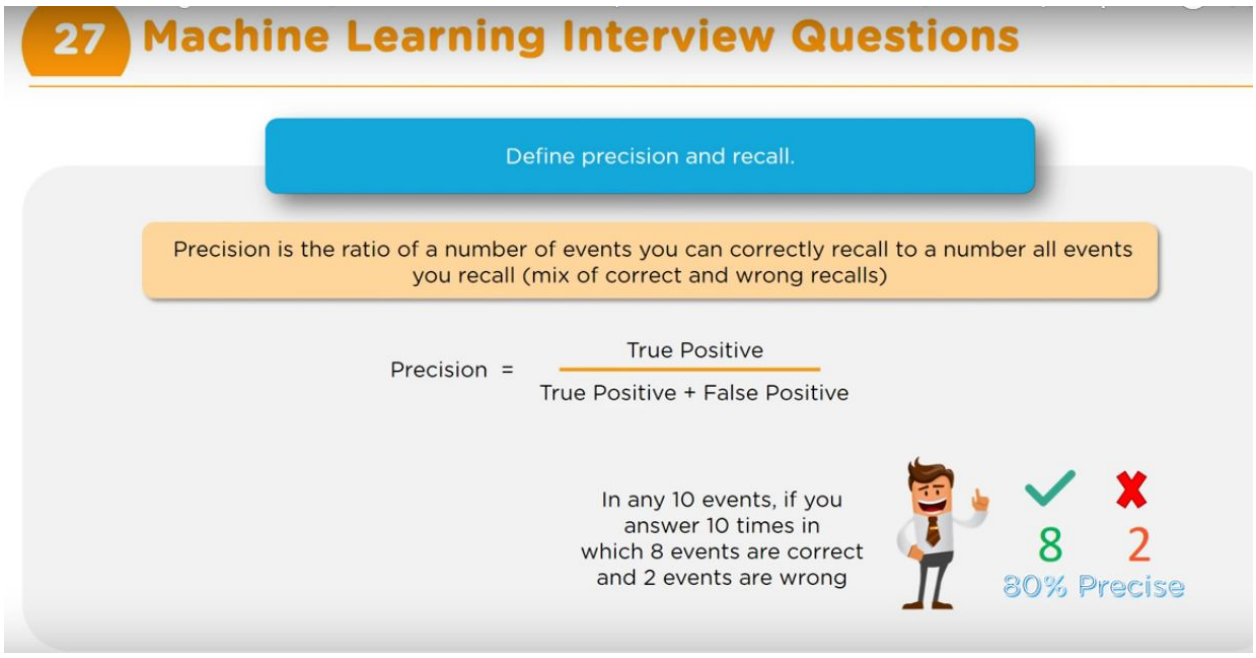
26 Machine Learning Interview Questions

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

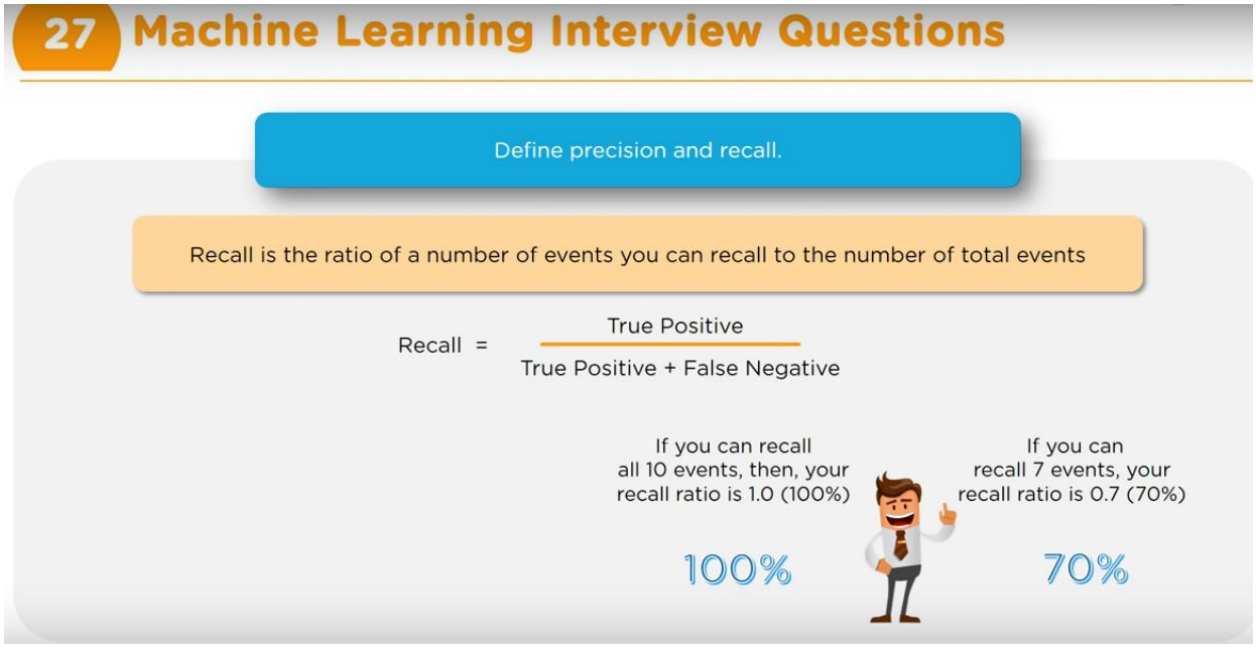
- Essentially, if you make the model more complex and add more variables, you'll lose bias but gain some variance — in order to get the optimally reduced amount of error, you'll have to tradeoff bias and variance
- You don't want either high bias or high variance in your model

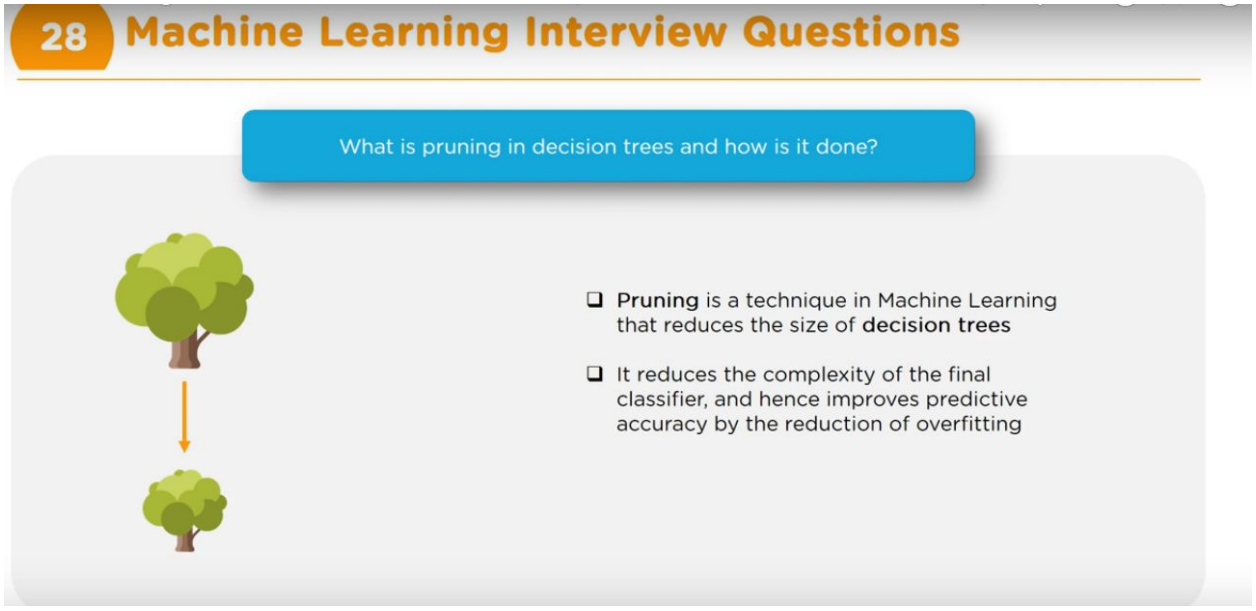
Machine Learning pipeline





Machine Learning pipeline





Machine Learning pipeline

28 Machine Learning Interview Questions

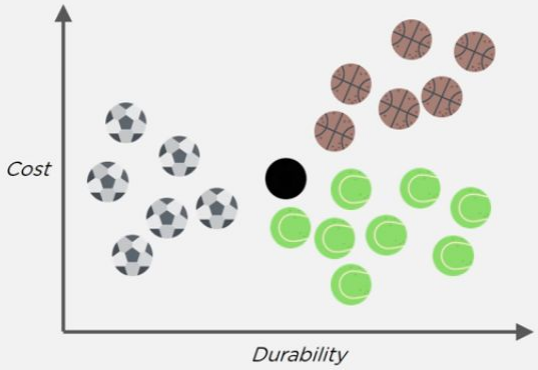
What is pruning in decision trees and how is it done?

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**

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.

Machine Learning pipeline

The following diagram shows an example pipeline: