tutort

Top 21 Machine Learning Interview Questions & Answers

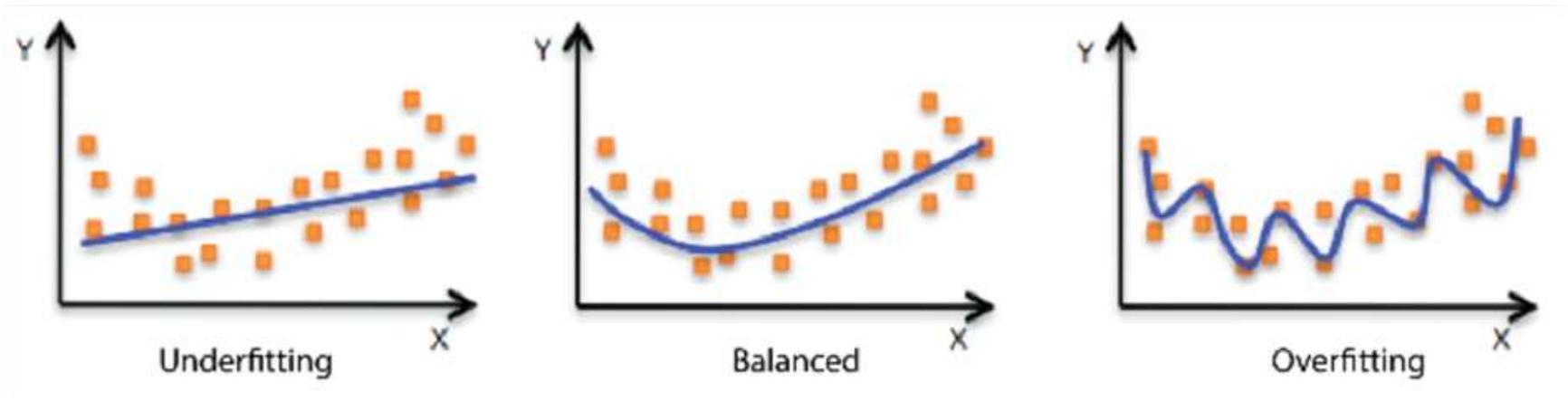
tutort academy

tutort

Question #1

What is the meaning of Overfitting and why overfitting occurs in Machine Learning?

When machine learning algorithms are constructed, they leverage a sample dataset to train the model. However, when the model trains for too long on sample data or when the model is too complex, it can start to learn the “noise,” or irrelevant information, within the dataset. When the model memorizes the noise and fits too closely to the training set, the model becomes “overfitted,” and it is unable to generalize well to new data. If a model cannot generalize well to new data, then it will not be able to perform the classification or prediction tasks that it was intended for. So, the model become a low bias and high variance model



tutort

Question #2

How to detect whether Overfitting has happened or not?

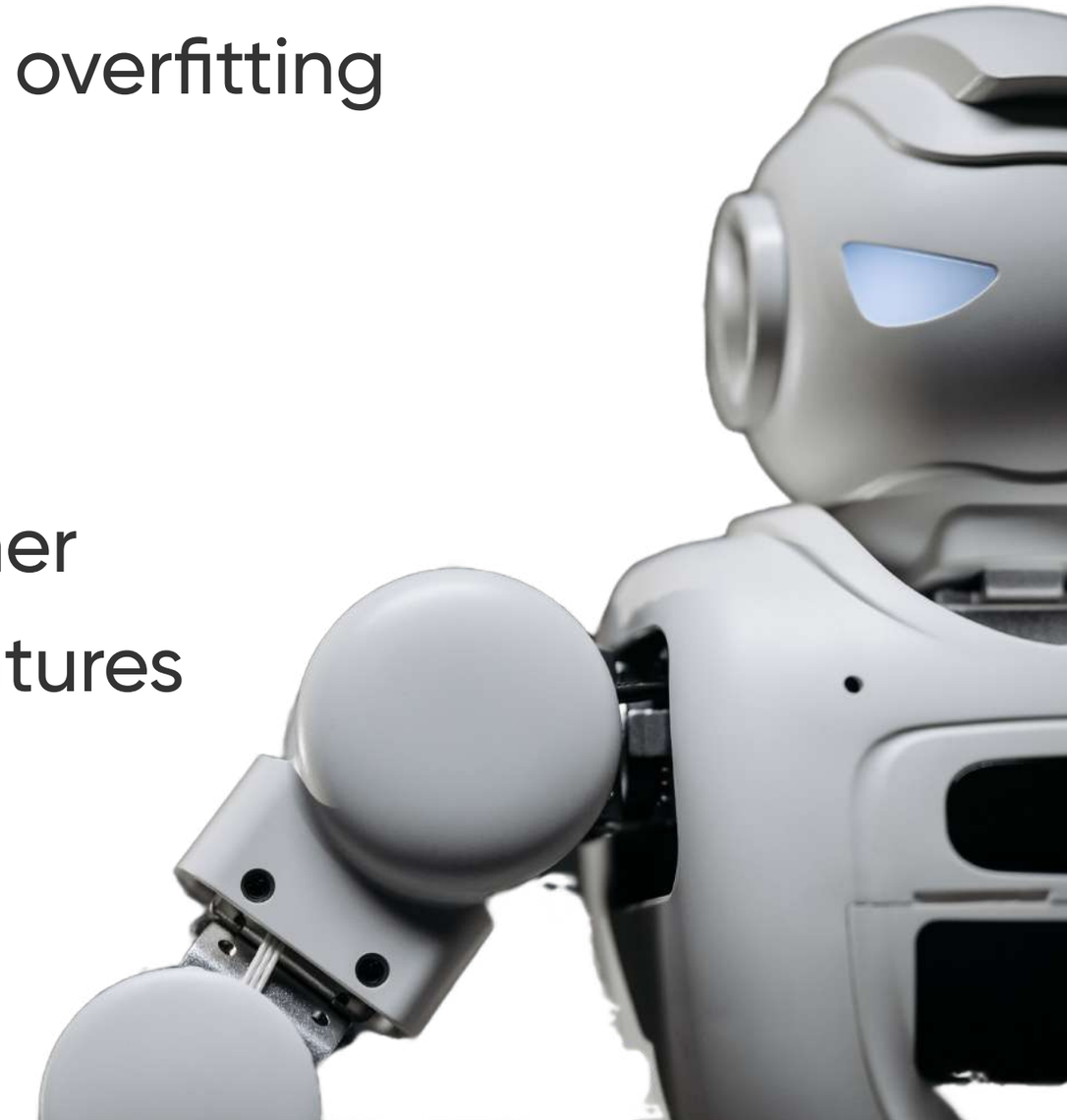
Overfitting is a low bias high variance problem, which means , if the machine learning model is doing exceptionally well in the training phase but fails badly with test set or with any new incoming data.
Example : A model has 95% training accuracy but 70% test accuracy implies that it is suffering from overfitting problem.

Question #3

What is the method to avoid overfitting?

There are couple of ways to tackle overfitting problem:

- 🎯 Cross-validation
- 🎯 Train with more data
- 🎯 Remove features which are either highly correlated with other features
- 🎯 Regularization
- 🎯 Ensembling Technique



tutort

Question #4

Differentiate supervised and unsupervised machine learning

Supervised and unsupervised learning have one key difference. Supervised learning uses labeled datasets, whereas unsupervised learning uses unlabeled datasets. By “labeled” we mean that the data is already tagged with the right answer.

Question #5

How is KNN different from k-means?

KNN is a supervised machine learning algorithm which is used for both classification and Regresison problem.

K-means is an unsupervised machine learning technique which is used to find pattern in the data which are not labeled.

- 🎯 K in KNN refers to number of neighbours.
- 🎯 K in K-means refers to number of clusters
- 🎯 KNN is also called a Lazy Learner
- 🎯 K-Means is also known as Eager Learner

tutort

Question #6

How does Machine Learning differ from Deep Learning?

Machine learning and deep learning are both types of AI. In short, machine learning is AI that can automatically adapt with minimal human interference. Deep learning is a subset of machine learning that uses artificial neural networks to mimic the learning process of the human brain.

Machine Learning

- A subset of AI
- Can train on smaller datasets
- Requires more human intervention to correct and learn.
- Shorter training and lower accuracy.
- Makes simpler, linear correlations.
- Can train on a CPU (Central processing unit)

Deep Learning

- A subset of Machine Learning
- Requires large amounts of data
- Learns on its own from environment & past mistakes.
- Longer training and higher accuracy
- Makes non-linear, complex correlations.
- Needs a specialized GPU to train

tutort

Question #7

What are the different types of algorithm methods in Machine Learning?

There are basically 3 types of algorithms present is Machine Learning:

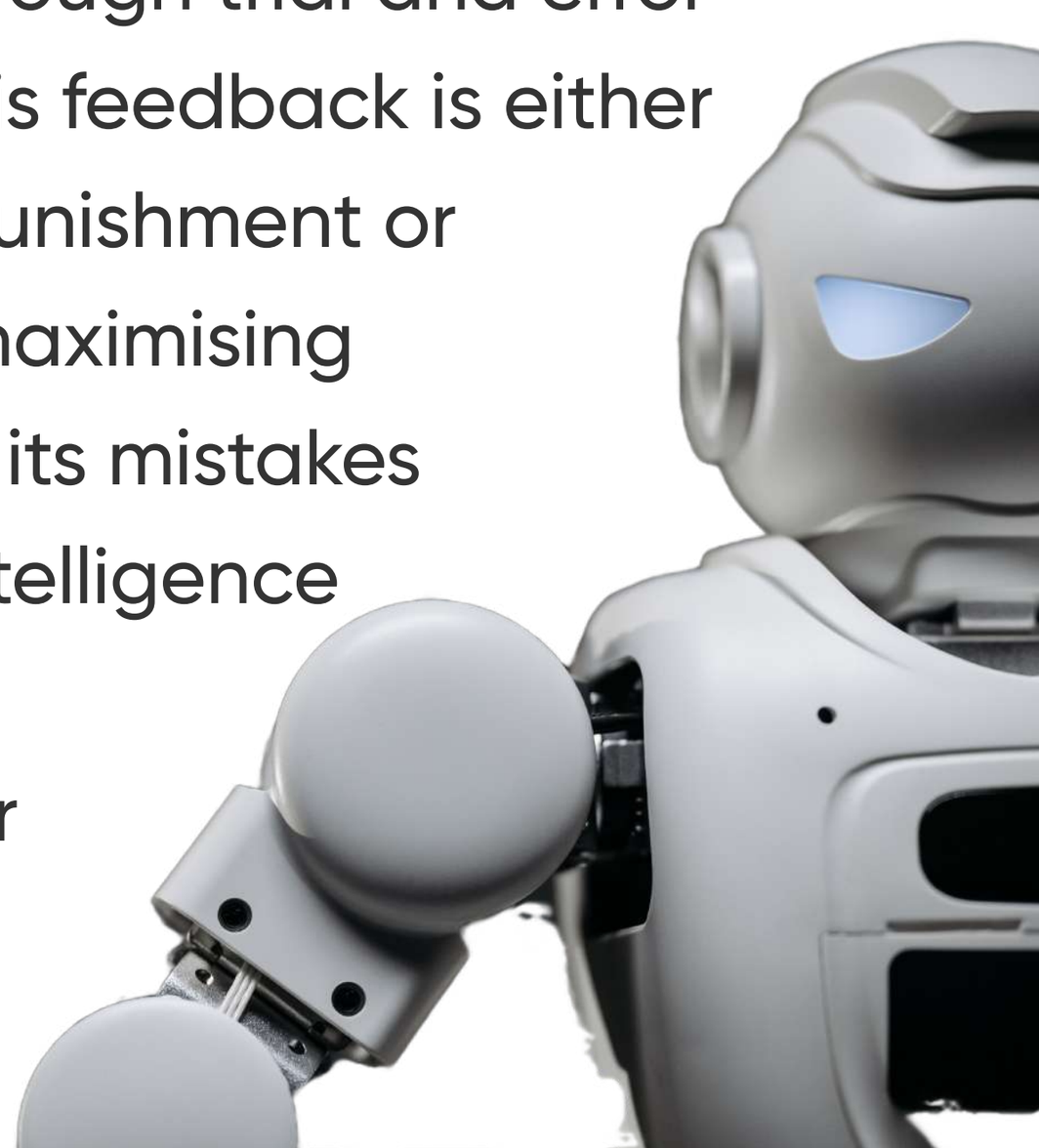
Supervised, UnSupervised and Reinforcement Learning.

Question #8

What do you understand by Reinforcement Learning technique?

Reinforcement learning (RL) is a subset of machine learning that allows an AI-driven system (sometimes referred to as an agent) to learn through trial and error using feedback from its actions. This feedback is either negative or positive, signalled as punishment or reward with, of course, the aim of maximising the reward function. RL learns from its mistakes and offers AI that mimics natural intelligence as closely as it is currently possible.

Example: Self Driven Car, Computer Games



tutort

Question #9

What is the trade-off between bias and variance?

The bias–variance tradeoff is a central problem in supervised learning. Ideally, one wants to choose a model that both accurately captures the regularities in its training data, but also generalizes well to unseen data. Unfortunately, it is typically impossible to do both simultaneously. High-variance learning methods may be able to represent their training set well but are at risk of overfitting to noisy or unrepresentative training data. In contrast, algorithms with high bias typically produce simpler models that may fail to capture important regularities (i.e. underfit) in the data.

This situation is known as ‘Bias Variance Trade off’ as we will have to trade any one of them to improve other.

tutort

Question #10

How do classification and regression differ?

Both Classification and Regression is a part of supervised Machine Learning algorithm.
However, Classification algorithm is used for predicting categorical variables such as ‘Yes/No’ or ‘Good/Better/Best’ etc. Regresison algorithm on the other hand is used predicting continuous variables such as Temperature/ Humidity etc.

Question #11

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

Linear Regression, Logistic Regression, Bagging Algorithms, Boosting Algorithms, Clustering.

tutort

Question #12

What do you mean by ensemble learning?

Ensemble learning is a general meta approach to machine learning that seeks better predictive performance by combining the predictions from multiple models.

Although there are a seemingly unlimited number of ensembles that you can develop for your predictive modeling problem, there are three methods that dominate the field of ensemble learning.

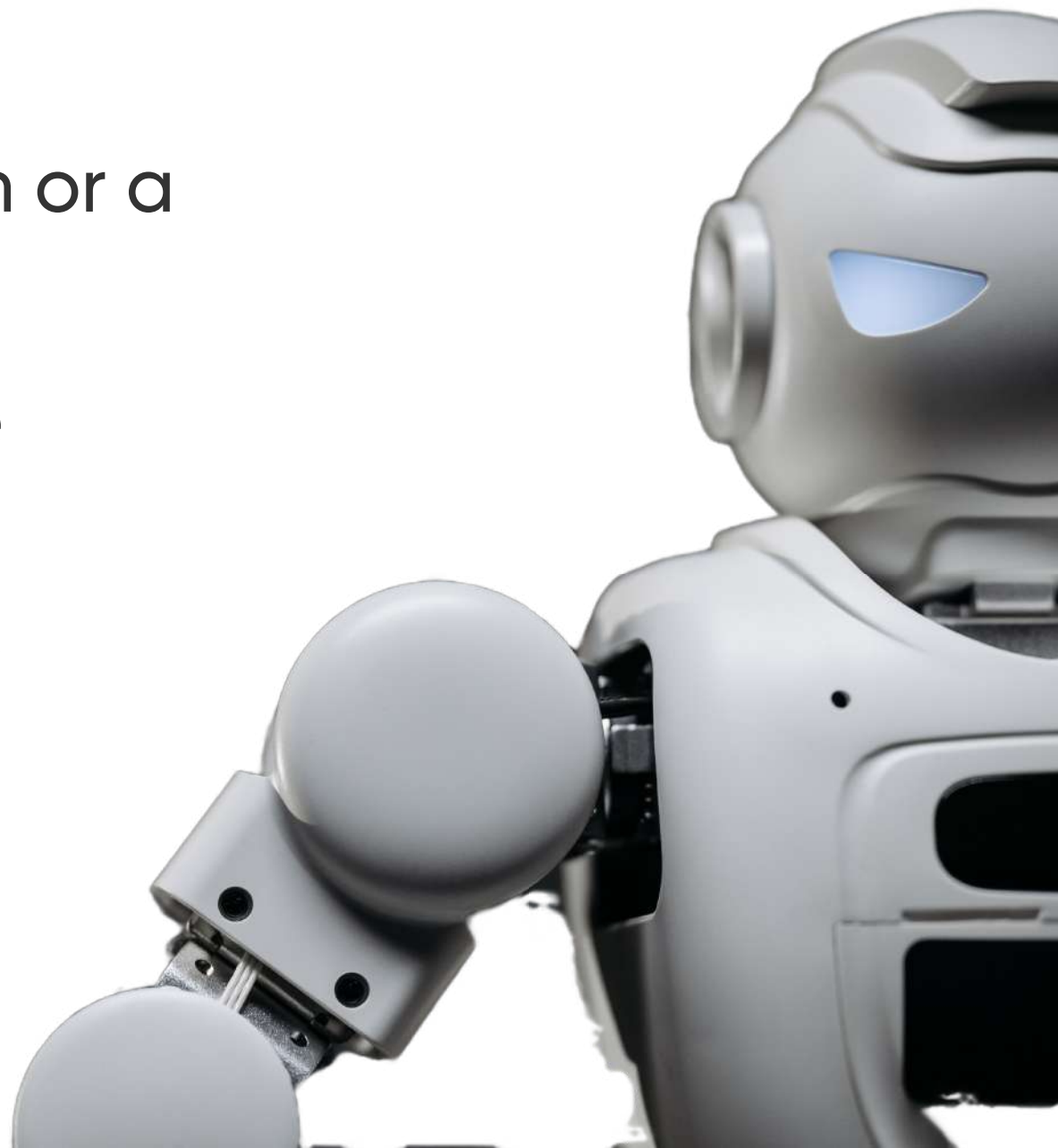
The three main classes of ensemble learning methods are bagging, stacking, and boosting.

Question #13

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

A standard approach to supervised Learning is to first understand :

- 🎯 Whether it is a regression problem or a classification problem
- 🎯 Collect the labeled data to fit the model
- 🎯 Improve the model's performance by Feature selection and Feature Engineering approach



tutort

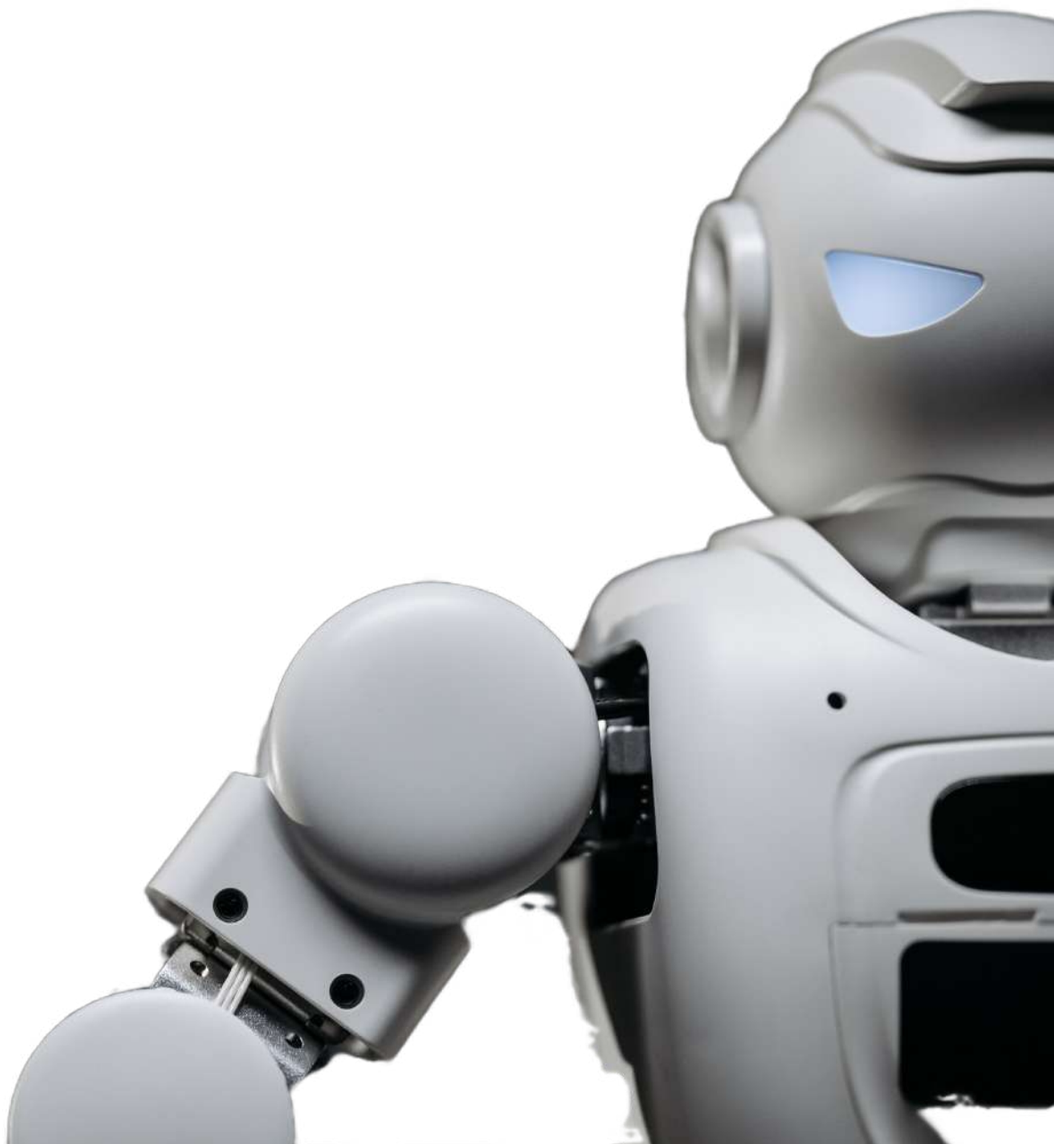
Question #14

What is a model selection in Machine Learning?

Model selection is the process of choosing one among many candidate models for a predictive modeling problem.

There may be many competing concerns when performing model selection beyond model performance, such as complexity, maintainability, and available resources.

The two main classes of model selection techniques are: probabilistic measures and resampling methods.



tutort

Question #15

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

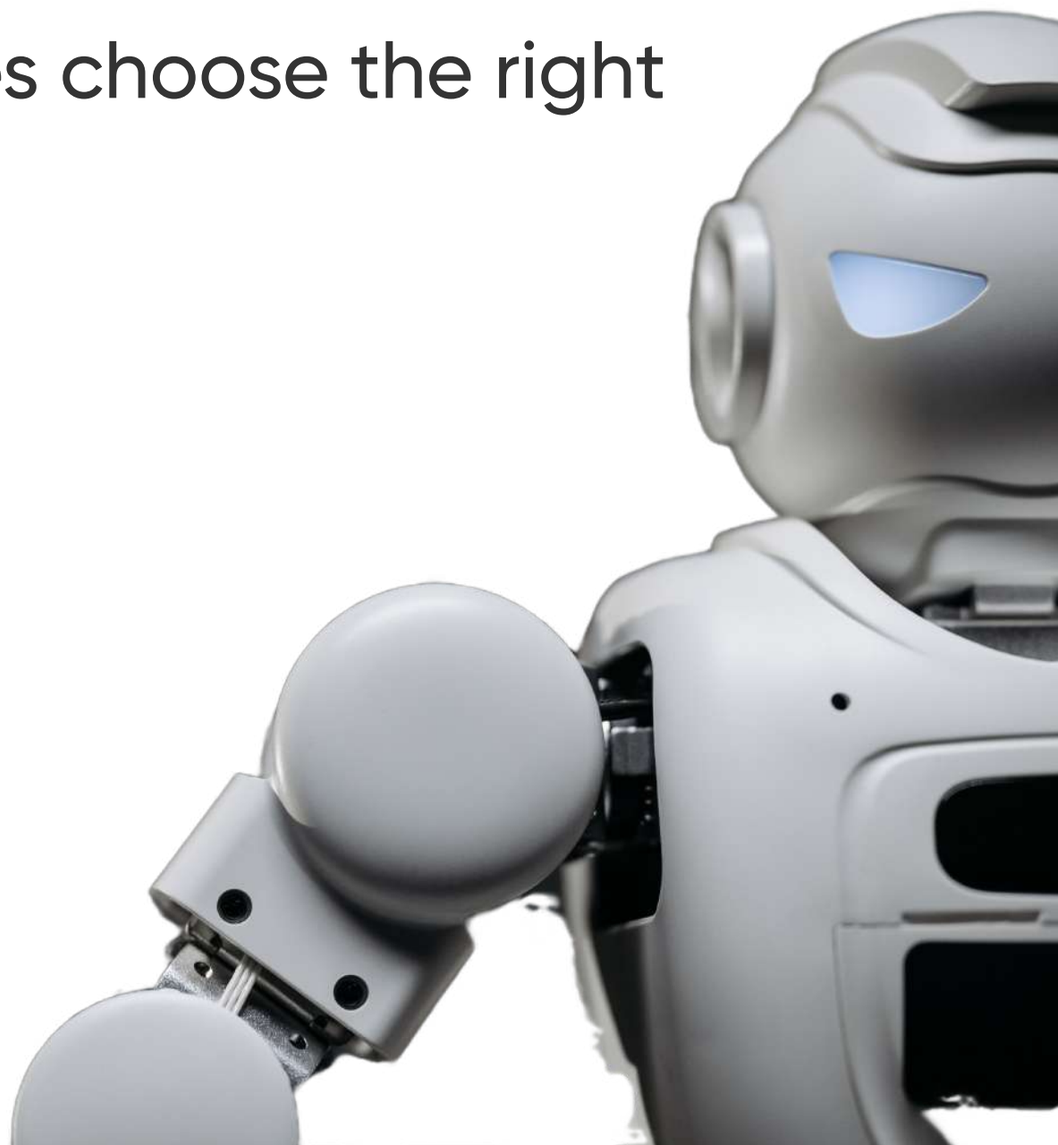
Hypothesis testing phases:

1. Build Null and alternate hypothesis
2. Select suitable sample data
3. Perform the respective test on the sample data and compare with the help of p-value

Model Building Steps:

1. Collect the data and understand what type of algorithm is suitable for solving the problem
2. Perform feature selection and feature engineering with proper EDA

Based on the problem type, build models and using different model selection techniques choose the right model for the problem

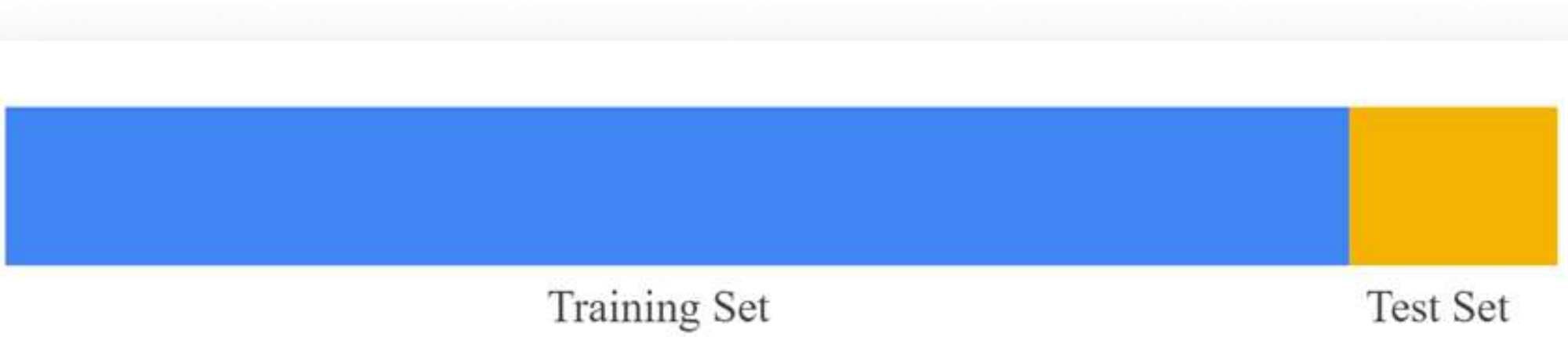


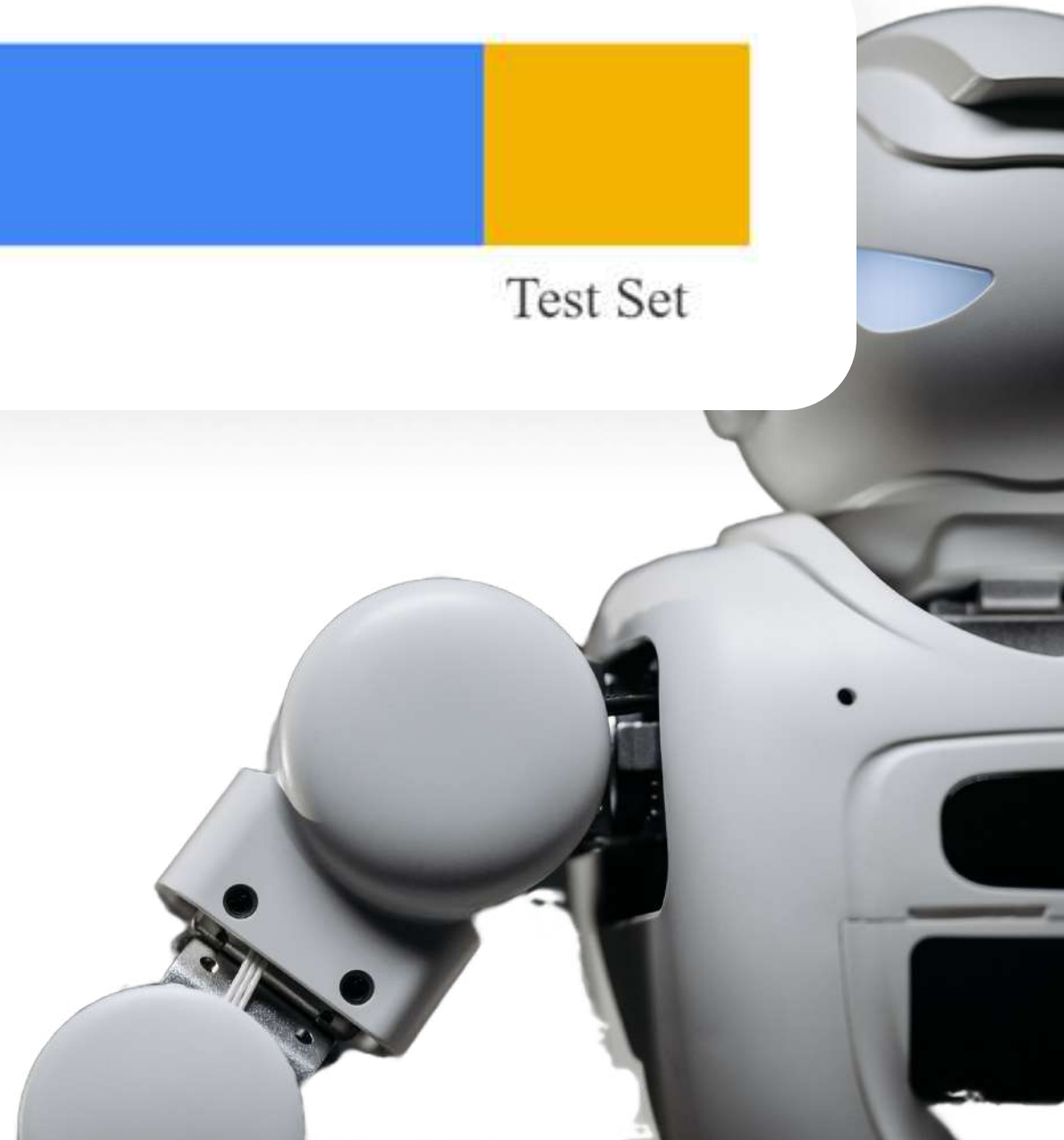
tutort

Question #16

Describe 'Training set' and 'training Test'.

To evaluate the performance of a machine learning model, the entire dataset is mainly divided into two parts. Majority of the data (70-80%) is used to train the model and then the remaining data is used as unseen data to test the performance of the model.
So, the data which is used for training purpose is known as train data and the remaining data is used as test data.
Sometimes, the entire dataset is divided into train , test and validation set.





tutort

Question #17

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

There are various ways to handle missing values in Machine Learning:

- 🎯 Delete the row which has missing value if the dataset is large and you have small numbers of missing records
- 🎯 Impute the missing values with mean or median depending on the distribution of the data for continuous variable
- 🎯 Impute missing value with mode for categorical features
- 🎯 Using forward and backward values to impute the missing value especially when handling time sensitive data. For example, stock data.
- 🎯 Use algorithms that supports missing values
- 🎯 Build a predictive model to handle and predict the missing values

tutort

Question #18

Describe Precision and Recall?

Precision: Precision measures what proportion of the positive predictions are correct.
Precision is defined by the following formula:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall: Recall measures what proportions of actual positives are identified correctly.
Recall is defined by the following formula:

$$\text{Recall} = \frac{TP}{TP + FN}$$

Example of Precision and Recall: Let's calculate precision and Recall of the following confusion matrix.

True Positives (TPs): 1	False Positives (FPs): 1
False Negatives (FNs): 8	True Negatives (TNs): 90

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{1}{1+1} = 0.5 \qquad \text{Recall} = \frac{TP}{TP + FN} = \frac{1}{1+8} = 0.11$$

tutort

Question #19

What do you understand by Decision Tree in Machine Learning?

A decision tree is a supervised machine learning model composed of a collection of "questions" organized hierarchically in the shape of a tree. The questions are usually called a condition, a split, or a test which aims at minimizing the impurity in the data. Each non-leaf node contains a condition, and each leaf node contains a prediction.

Decision Tree is a Greedy Algorithm and often suffers from low bias and high variance problem. However, proper pruning technique and hyperparameter tuning can solve the overfitting problem of Decision Tree.

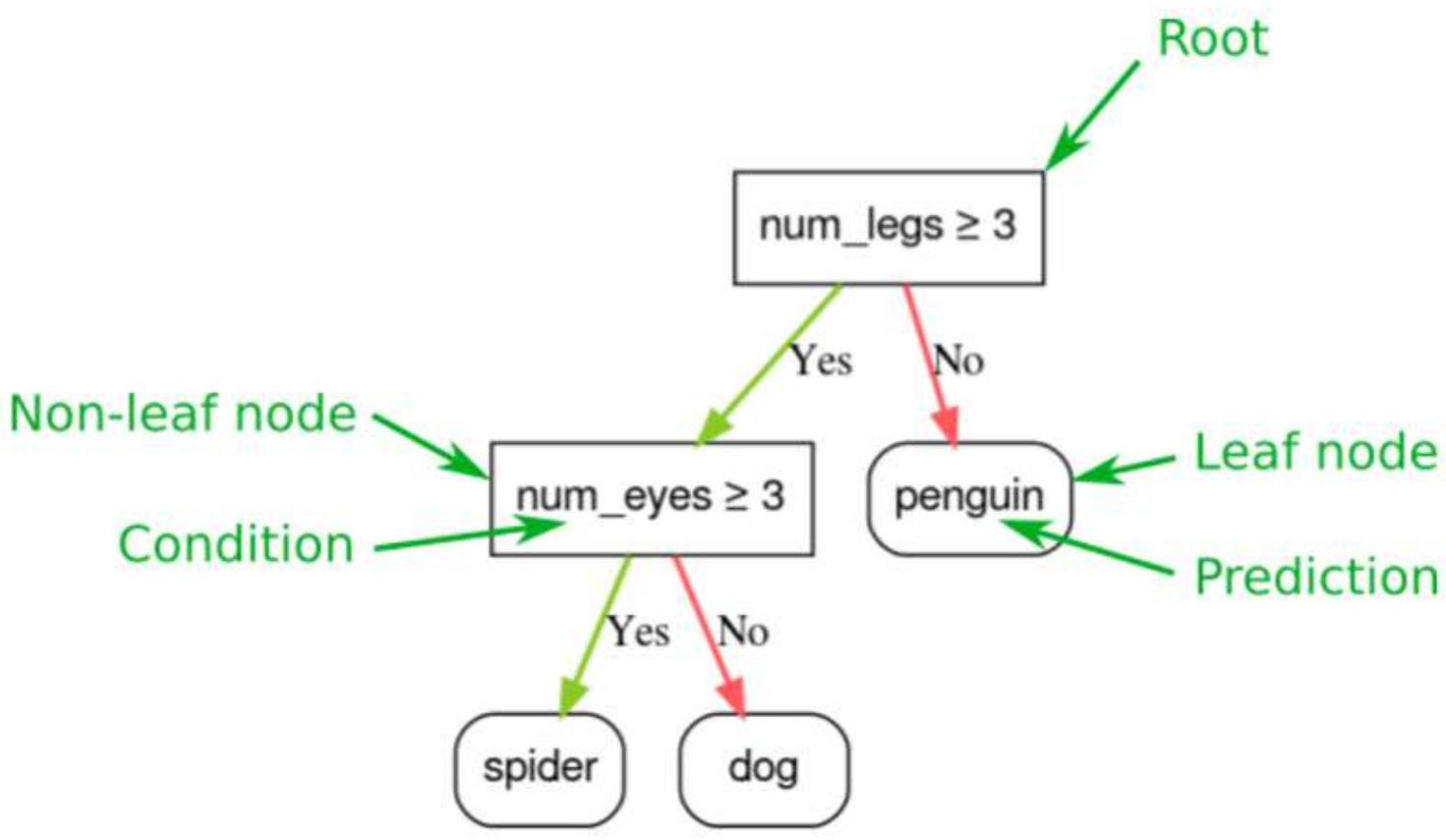


Figure 1. A simple classification decision tree. The legend in green is not part of the decision tree.

Question #20

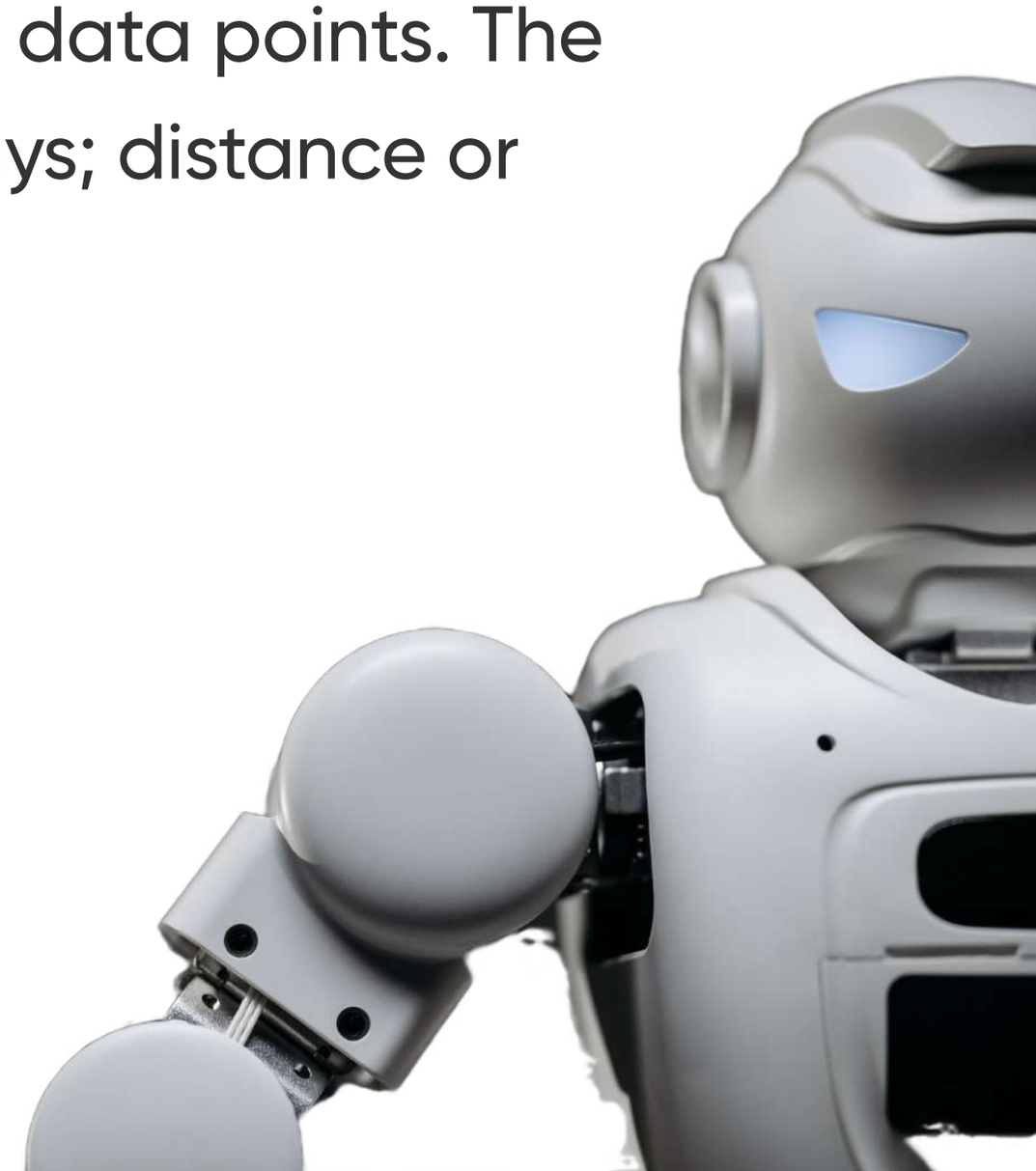
What are the functions of Supervised Learning?

The purpose of supervised algorithms are to understand the pattern in the data from a labeled given set of historical examples and after analyzing the pattern in the data, predict the same for any new data.

Question #21

What are the functions of Unsupervised Learning?

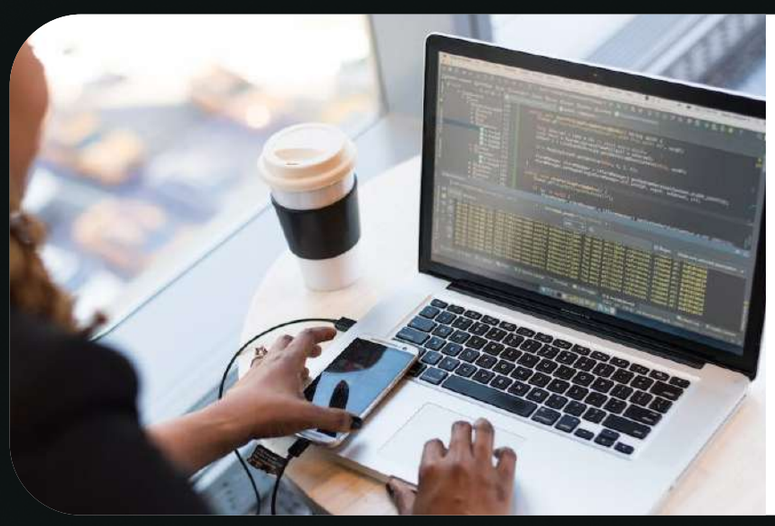
Unsupervised machine learning algorithm is used to identify and group together similar data points. The similarity is measured in various ways; distance or distribution or density etc.



tutort

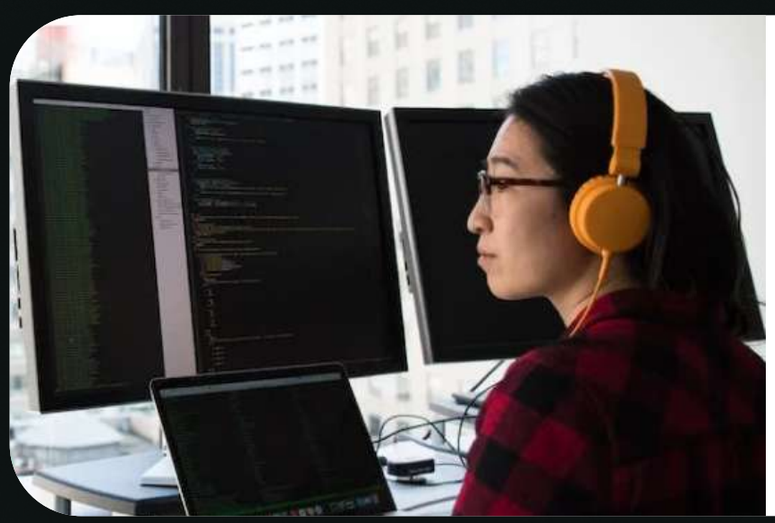
Kick Start your Data Science Career

Join Our



Full Stack Data Science Master's Course

Course Duration: 8 Months



Full Stack Artificial Intelligence and ML course

Course Duration: 10 Months

Full Stack DS & AI Masters program for Tech & Business Leaders

Course Duration: 10 Months

WhatsApp Now

Talk with our experts

Watch us on YouTube

tutort

Why Tutort Academy?

Profile Review by
Industry Experts

Session with
1:1 Mentorship

Real time projects from
Companies

100% guaranteed
Job Calls

750+ Students placed at:

Google
Morgan Stanley
EY Building a better working world
pwc
SOCIETE GENERALE
fractal INTELLIGENCE FOR IMAGINATION
CIKLUM
ClearCapital
NOBROKER

Explore more

www.tutort.net

tutort

Follow

tutort academy

for more such informative content.

www.tutort.net | +91-8712338901