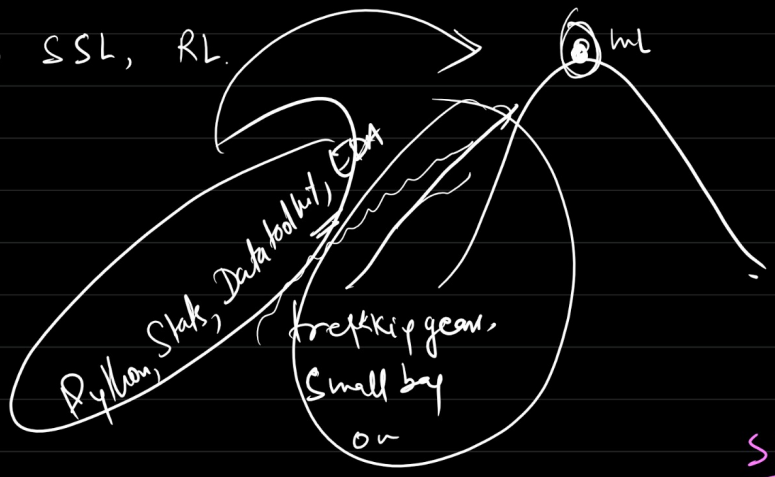
Till now

→ Intro to ML

→ SL, USL, SSL, RL.

Agenda →

ML-2




ML

SL

| | $x_1$ | $x_2$ | $y$ |
|---|---|---|---|
| | # of hours studied | # marks in internal exam | Pass \| fail. |
| | 5 | 80 | 1 |
| | 3 | 70 | 0 |
| | 1 | 40 | 1 |
| | 2 | 45 | 0 |
| | 3 | 50 | 1 |

train

↓

mathematical Relationship

↓

How to verify?

→ Test the model on some other data

↓

Unseen data ←

Take this model to business team and say

↓

verify this model

reason (money, time)

train-test split

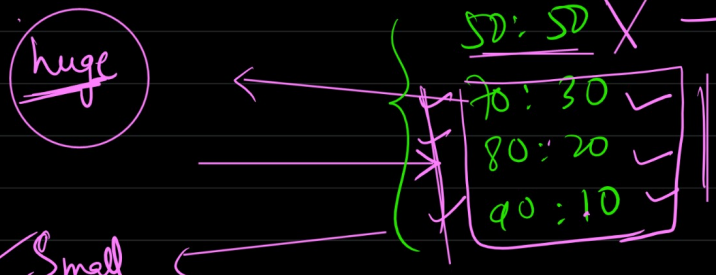
date

| | $x_1$ # of hours studied | $x_2$ # marks in internal exam | Pass / fail |
|---|---|---|---|
| train | 5 | 80 | 1 |
| | 3 | 70 | 0 |
| | 1 | 40 | 1 |
| | 2 | 45 | 0 |
| test | 3 | 50 | 1 |

4 : 1

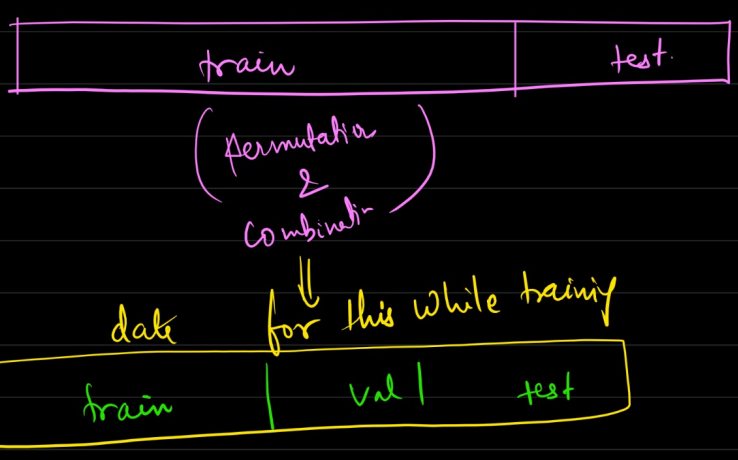
test → representative of Unseen data
→ Used to test model accuracy

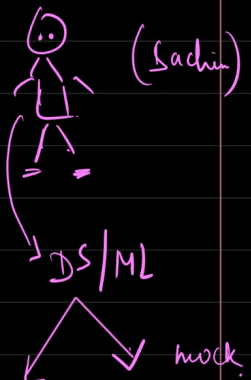
1000 → ratio : train : test split ?

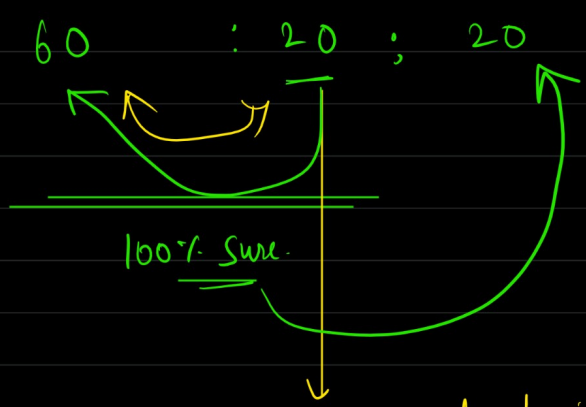



ML → learn patterns → train data

train > test data






Batsman → V.K. (trained)

R.D → Tuned → Hyperparameter tuning ⇓ T-20 world cup

parameter — mathematical relationship

Student

eg: Studying from (NCERT) ⇓ training

Solve sample papers ⇓ Hyperparameter tuning

Board exam ⇓ testing of model.

eg: guitar tuning

Key takeaway.

→ train — test

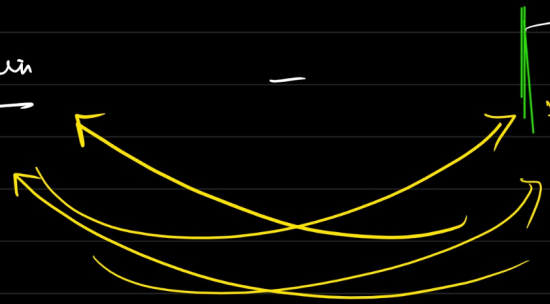
→ train — val — test

60 : 20 : 20

* validation data will act as test while training

Hyper parameter tuning ??

train — Test (Unseen data)



→ Ever board exam papers are given to students → Jailed because you are cheating.

* data leakage

↓

You have passed test information to the training of the model ⇒ model will perform well on both train and test.

⇓

One day you will deploy it.

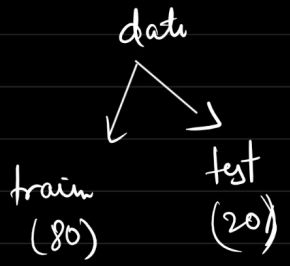
⇓

Models performs worst in real world scenario

⇓

huge loss ⇒ you are fired

* Machine learning framework




* framework → set of rules
  - → a predefined process.
  - → set of tools
  - → Guideline
  - → Standardized Process

| # Area of house | # of rooms | Price of house |
|---|---|---|
| 1100 | 2 | 2.1 |
| — | — | — |
| — | — | — |
| — | — | — |
| 1200 | 3 | ? |

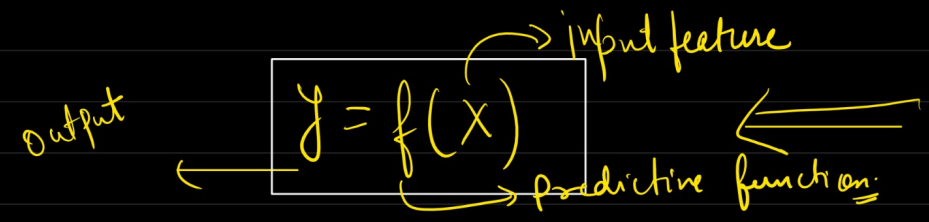
train 80 → training of model

test 20

New data (test)

⇓

✓ Understanding the pattern

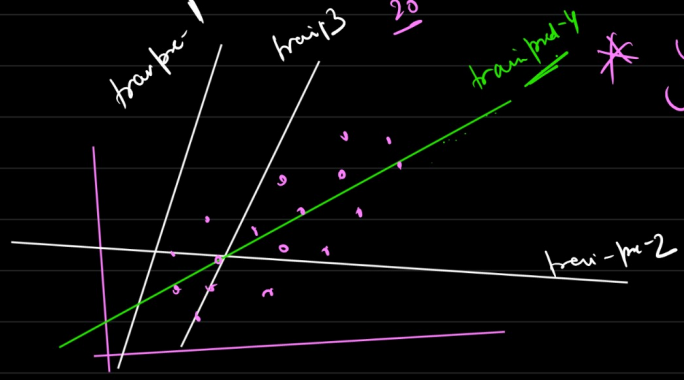



* Model training → given the data estimate the prediction function by minimizing error → ML algos
  - ↓ evaluation metrics

| # Area of house | # of rooms | Price of house ($y_{actual}$) | $y_{train\ prediction}$ |
|---|---|---|---|
| 1100 | 2 | 2.1 | 1.8 |
| — | — | — | 2.1 |
| — | — | — | 2.3 |
| — | — | — | 2.4 |
| 1200 | 3 | ? | |

$y = f(x)$ ← training 80 { exposed data

testing { 20




* you want to learn the optimal relationship.

$y = f(x)$

* Key takeaway
  - → train | test.
  - → $y = f(x)$

## Overfitting & Underfitting

* Overfitting

train → Model is trained ⇒ Accuracy is high (95%)

test → Model is tested ⇒ Accuracy is

Viswa Sir (10 Questions)

training { Ajay (memorised 10 Questions) | Vijay (Understood the concept & practiced the

(unseen)

low. ⇓ (65%)

Exam → Who will perform better in this exam?
test

* Model performs well on train data but worse on test data.

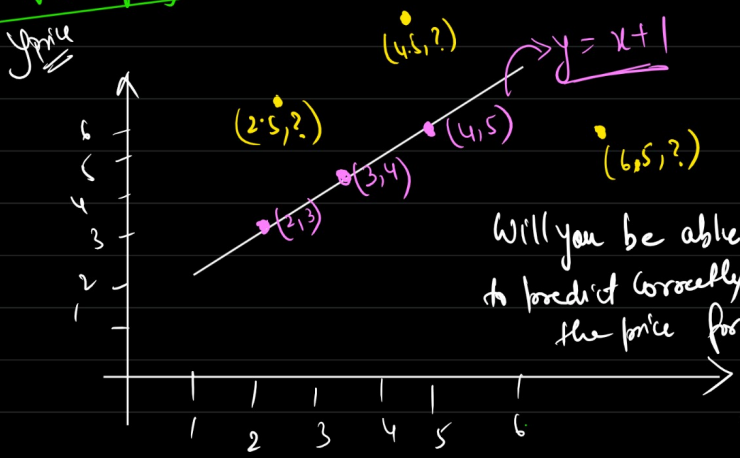
## * Underfitting

train → accuracy ↓ 50%
test → accuracy ↓ 40%
} Underfitting

## * Generalised model

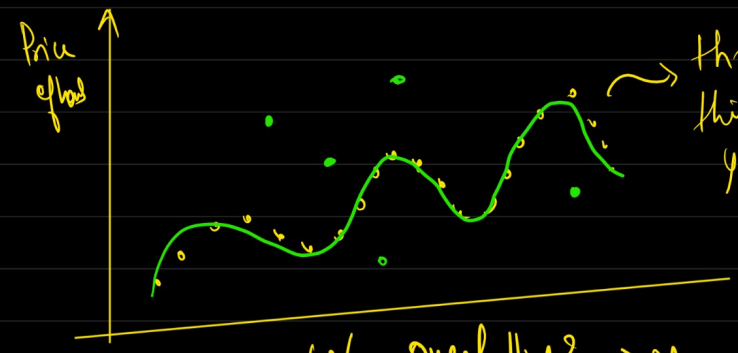
train → Accuracy ↑ 85%.
test → Accuracy ↑ 84%.

Expl. of overfitting




| # No of rooms (x) | Price of house (y) (min) |
|---|---|
| 2 | 3 |
| 3 | 4 |
| 4 | 5 |

→ train




→ through this curve you memorised the data

→ Why ??

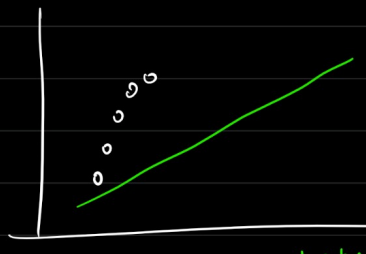
They don't fall in the pattern

* we have not learned the pattern well.

Why Overfitting? → More than required no of features
→ Overfitting Algorithms.
→ lesser no of train data

Underfitting

train & test accuracy ↓



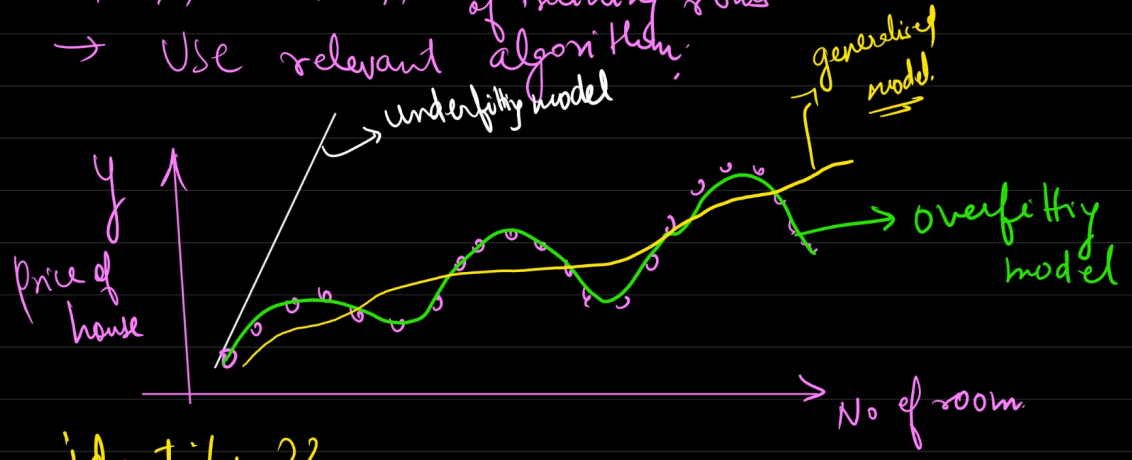
model has not learned well

Why underfitting

→ less no of required features
→ less no of train data
→ Underfitting algorithm

## Generalised model

→ give required no of features
→ " " " of training rows
→ Use relevant algorithm




## * How to identify ??

* Underfitting ≈ model is not performing well while training.
* Overfitting ≈ train acc ↑ 80% test ↓ 60% (difference more than 5%)
* generalised model ≈ train ↑ test ↑

65; 64 diff is ≈ range 5%

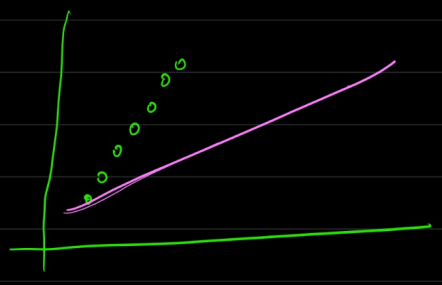
## Bias | Variance

* training error is also known as bias.
  High training error means high bias
* testing error is also known as Variance.
  High testing error means high variance

bias → assumption | prejudice

Why ??

Because you assumed the distribution of data wrongly



Train → Accuracy ⇓ 50% → High training error → High Bias

test → Accuracy ⇓ → High testing error → High Variance

→ Underfitting ⟹ High bias, High Variance

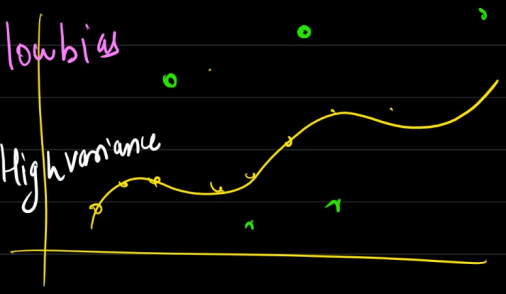
Variance → Why variance is analogous to High testing error.?

→ test data varies from the pattern, thats why high testing error.

Train → trained → Accuracy ⇑ (low $T_{rain}E$) → low bias

Test → tested → Accuracy ⇓ → High variance

(High test error)



→ overfitting ( low bias , High Variance)

* Generalised model

(low bias, low variance)

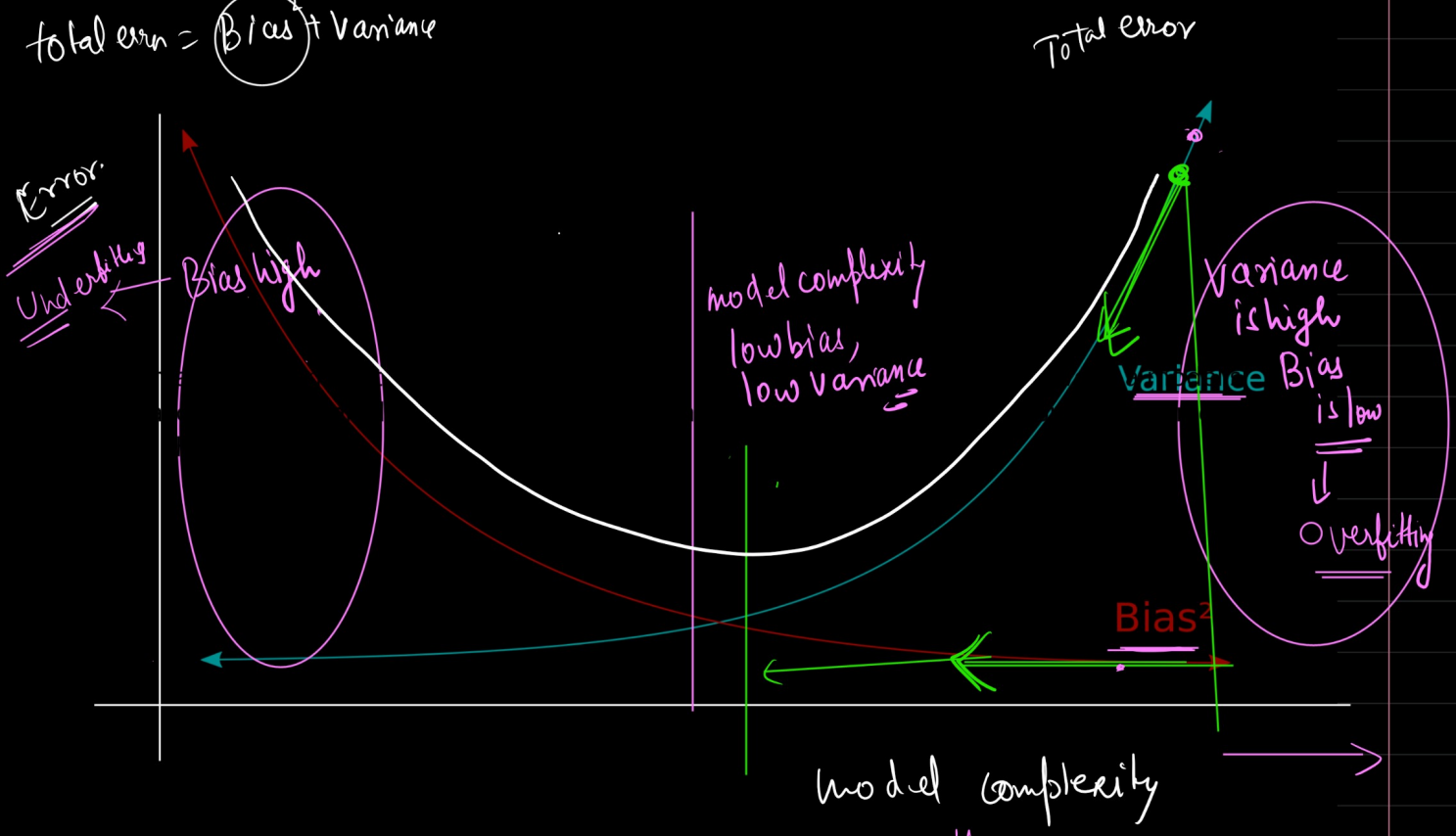
train acc ↑ (low bias)

test acc ↑ (low variance)

## Bias – Variance tradeoff

In statistics and machine learning, the bias–variance tradeoff describes the relationship between a model's complexity, the accuracy of its predictions, and how well it can make predictions on previously unseen data that were not used to train the model.

total error = $(Bias)^2$ + Variance




model complexity → You have used more than required features.

* In order to reduce Variance (reduce the testing error), introduce some bias in training of model - by selecting relevant features or selecting the right algorithms.

* In order to reduce high bias → Use some more dp's, more feature engineering, Use some other ML algorithms.