Machine Learning

K Nearest Neighbours Algorithm.

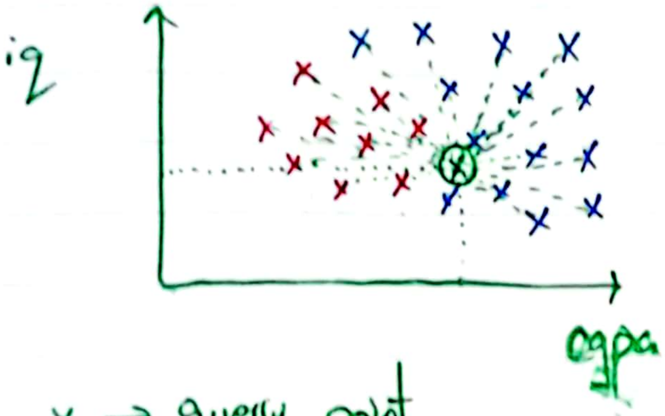
Let

	CGPA	iq	Placement
100	8	80	1
	7	70	0
	.	.	.
	.	.	.

Our goal is to build a machine learning model

⇓

cgpa | iq → prediction



x → 1

x → 0

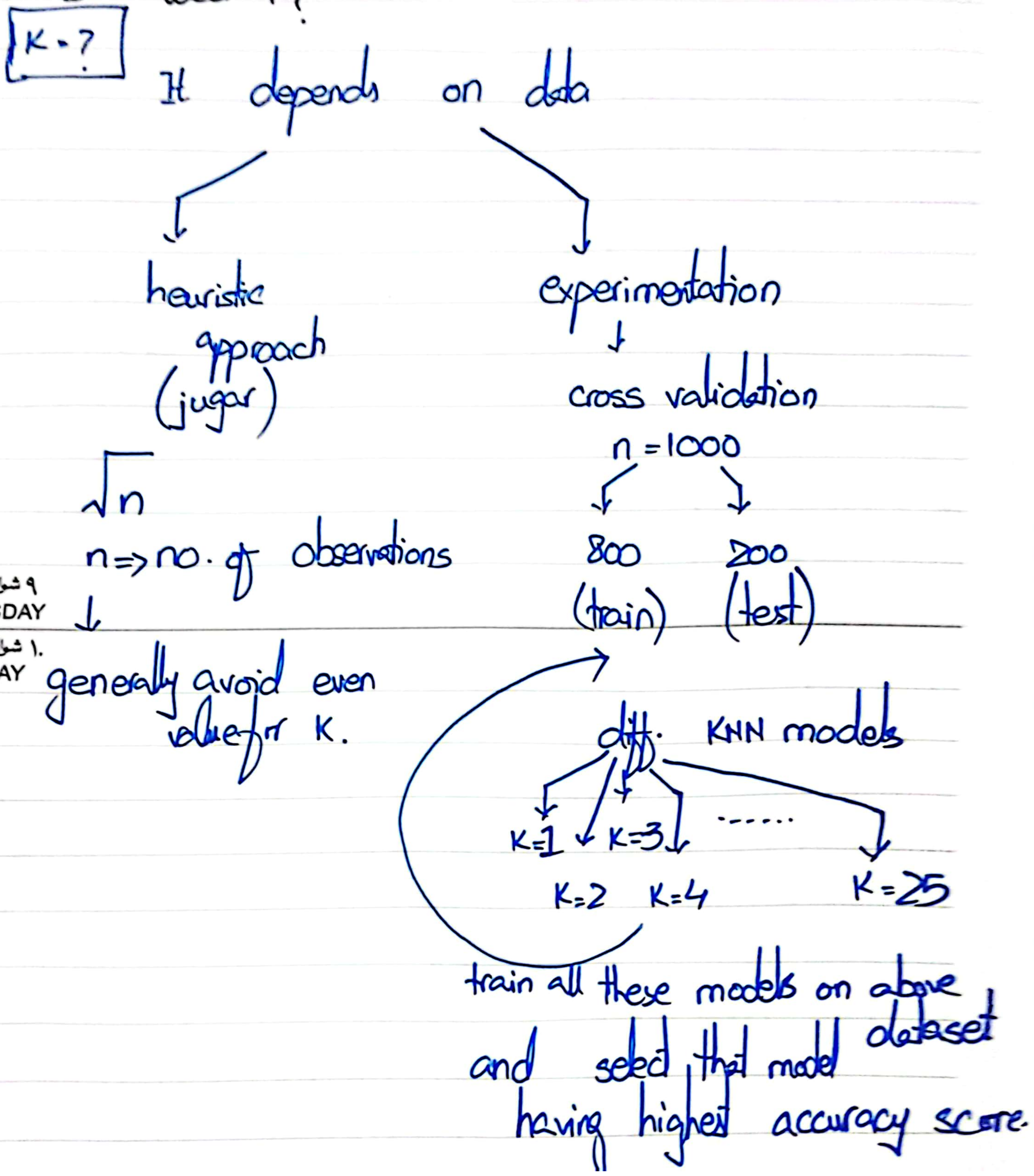
we decide a value of K let say we assume K = 3

3 nearest nei

x → query point

generally we find the euclidian distance.
we have calculated 100 distances.
then sort in ascending order
then majority count. 1 1 0 → 1

How to select K?



K = ?

It depends on data

heuristic approach (jugar)

$\sqrt{n}$

n => no. of observations

generally avoid even value for K.

experimentation

cross validation

n = 1000

800 (train)

200 (test)

diff. KNN models

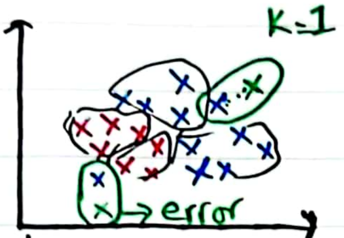
K=1 K=2 K=3 K=4 ...... K=25

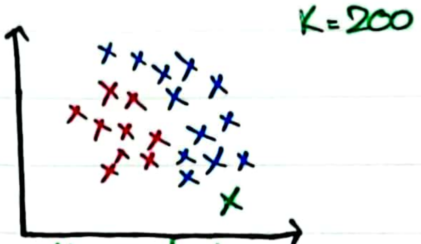
train all these models on above dataset and select that model having highest accuracy score.

Overfitting and Underfitting in KNN:

let we have data for 200 students.

Consider two scenario.





decision surface would be divided in multiple regions
overfitting → minor changes

all points lie in neighbor
blue > red
i-e result always blue

الجمعة - ١١ شوال
FRIDAY 13

السبت - ١٢ شوال
SATURDAY 14

Limitations:

1) large datasets → n = 5L , f=100 prediction slow
it is lazy learning technique.
2) High dimensional data
curse of dimension
3) Outliers
4) Non-homogenous scales
5) Imbalance dataset Yes-98%, No-2%
6) Inference and not for prediction.