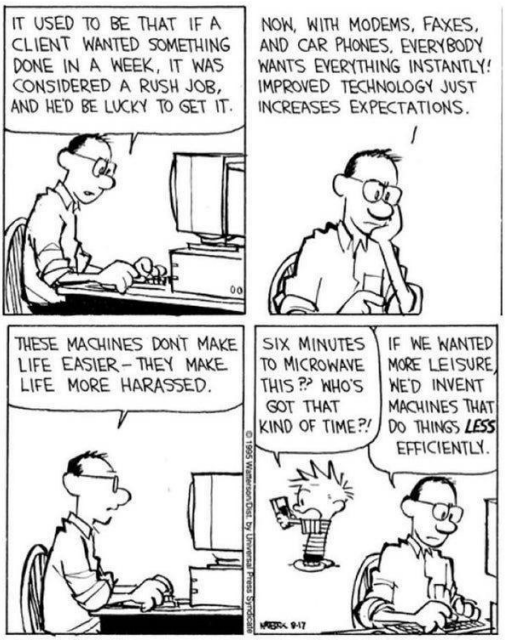
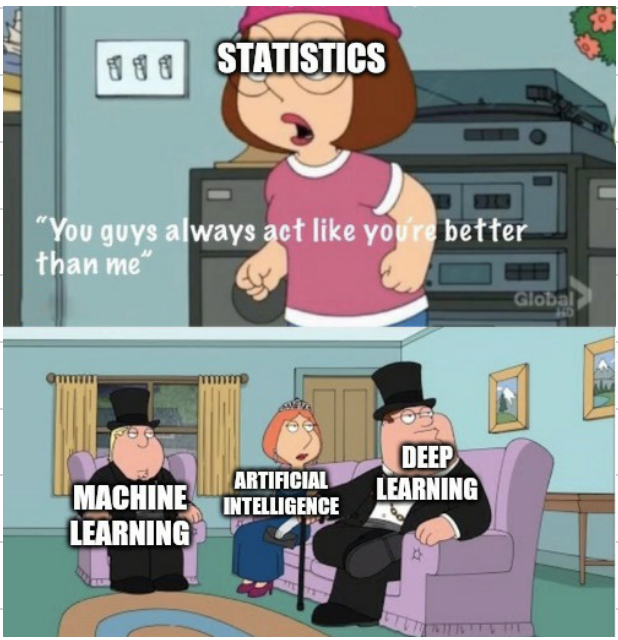
Session -1

KNN -1

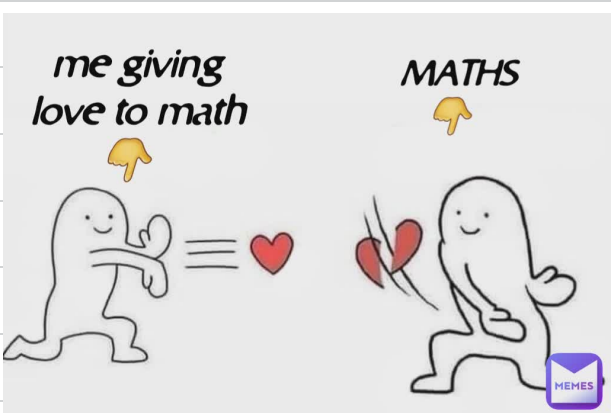
Aug 25, 2025





AGENDA

1. KNN
2. SMOTE



ML models we're going to learn in this Module

1. KNN → Classification
2. Decision trees → C, R
3. Random Forest → C, R
4. Light GBM / XGBoost → C, R
5. Naive Bayes → C
6. SVM → C, R

When to use which ML model ?

1. Size / type of dataset
2. Latency / Throughput of ML model
3. Deployment hardware.
4. Thickness of your wallet

2022

GPT ↓ — decent

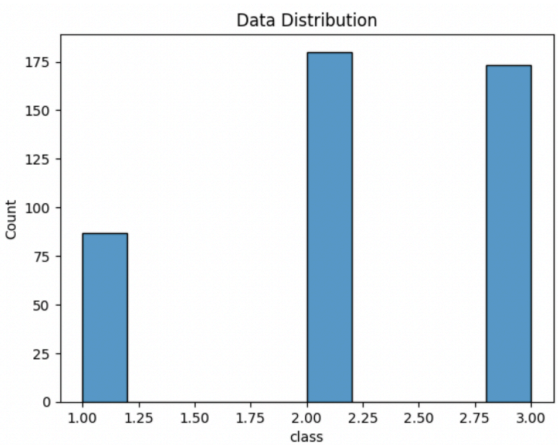
2023

↓

GPT ? acc ↑

Powerful ∝ $\frac{1}{\text{Speed.}}$

Latency = time taken for your computer to send request to server + time taken for server to generate response + time taken for their server to send response to your computer.

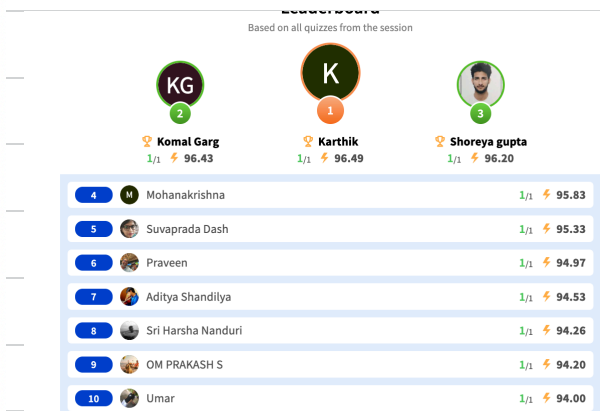


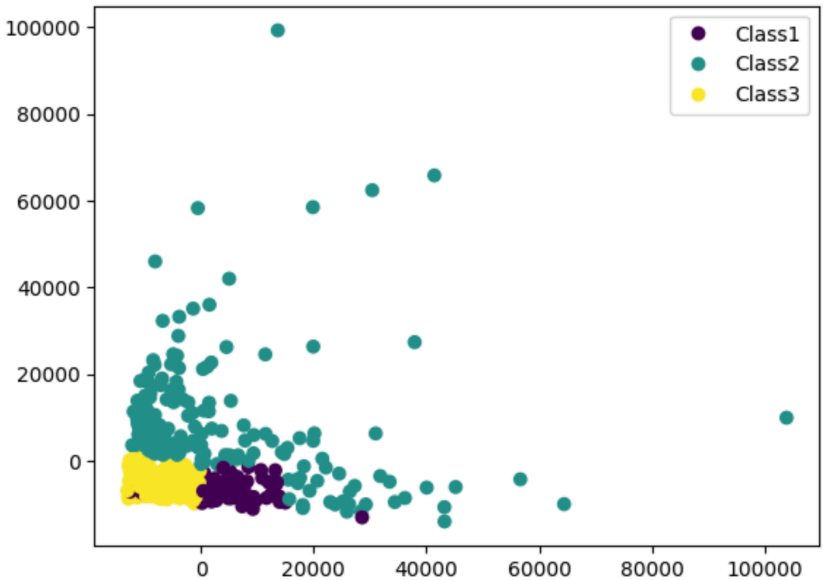
What can be said about the data ?

0 users have participated

A	Multi-class balanced data	0%
B	Multi-class imbalanced data	0%
C	Binary-class imbalanced data	0%
D	Binary-class balanced data	0%

End Quiz Now



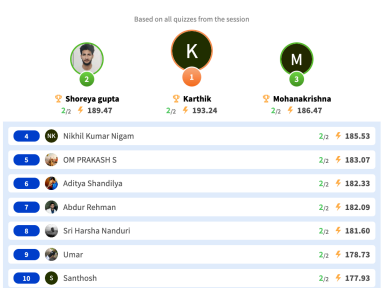


How will Logistic Regression handle non-linear,multi-class data ?

0 users have participated

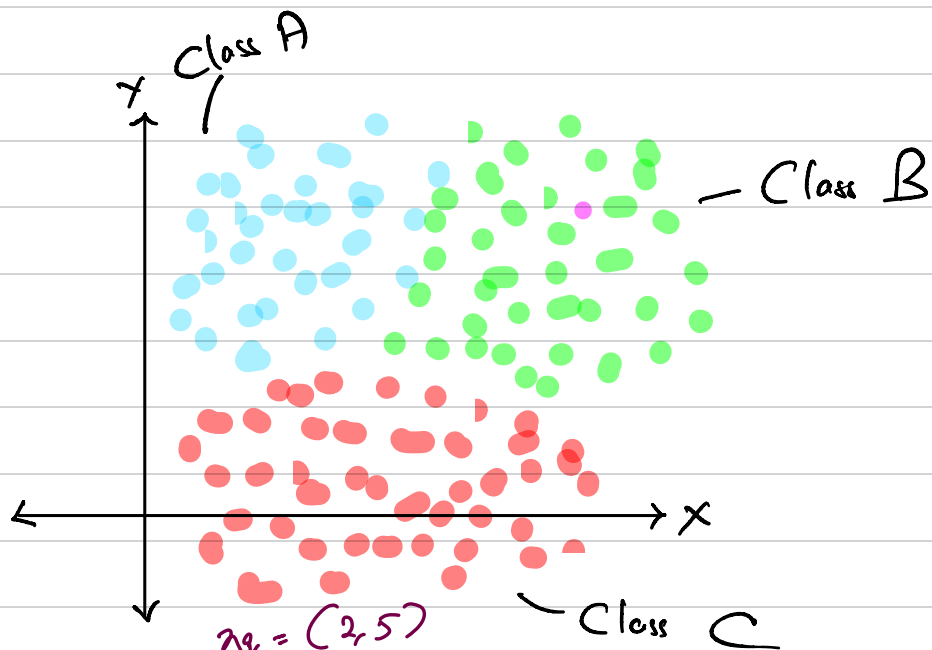
✓ A	Polynomial, OneVsRest	0%
B	Linear, OneVsRest	0%
C	OneVsRest, Polynomial	0%
D	OneVsRest, Linear	0%

End Quiz Now



How does KNN work??

K nearest neighbours



	f^1	f^2	y
X^1	3	6	1
X^2	6	4	1
X^3	8	2	3
X^4	7	5	3
X^5	1	4	2
X^6	2	2	2

=sqrt((3-2)^2 + (6-5)^2)

	f^1	f^2	y
X^1	3	6	1.41
X^2	6	4	3.00
X^3	8	2	6.48
X^4	7	5	5.00
X^5	1	4	1.41
X^6	2	2	2.00

	f^1	f^2	y	y
X^1	3	6	1	1.41
X^5	1	4	2	1.41
X^6	2	2	2	2.00
X^2	6	4	1	3.00
X^4	7	5	3	5.00
X^3	8	2	3	6.48

(1,.1)

(2,,1)

L2 Norm

$\sqrt{(1-2)^2+(1-1)^2}$

K-value = How many of my neighbors should I look at for deciding the class of my query point?

I'll do a voting on the class of my three most nearest neighbours.

$k=3$

All machine learning models are wrong, but some of them are useful.

Poll results!

background?

39 responses from 39 users

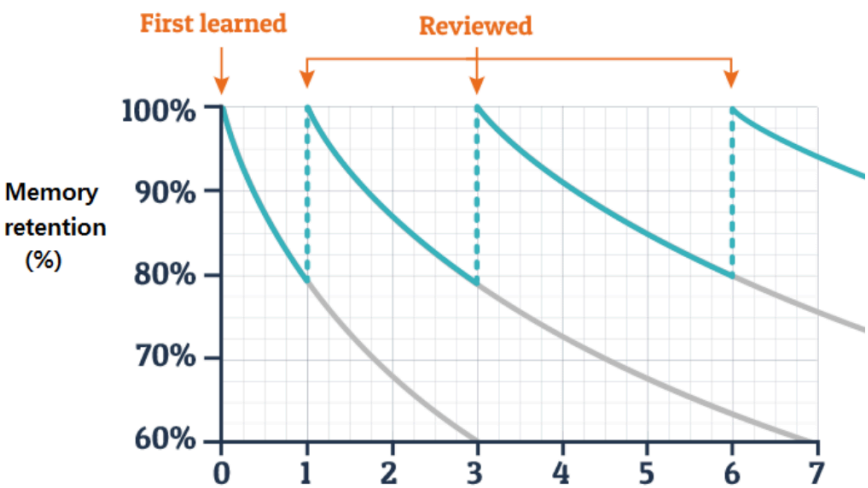
A	non-tech	51%
B	tech	49%

Poll results!

years of exp

39 responses from 39 users

A	<2	21%
B	2-5	36%
C	5-10	18%
D	10+	26%



2^0 2^1 2^2 2^3

1 2 4 8

9:03 — 10 min

9:15

9:15 - 9:30

Interview

Knn is a non-parametric algorithm.

Training phase { Load data into RAM }

Testing phase { 1. Computing L2 norm
2. Sorting.
3. Picking top K & Voting

Arrange the statements in correct order based on kNN algo

s1- find majority vote

s2- perform euclidean distance

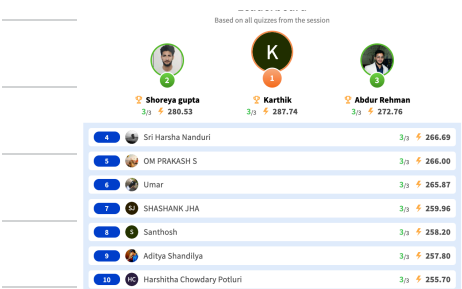
s3- sort and select k datapoints

s4- give class to x_q datapoint

0 users have participated

A	s2,s1,s3,s4	0%
B	s2,s3,s4,s3	0%
✓ C	s2,s3,s1,s4	0%
D	s2,s4,s3,s1	0%

End Quiz Now



POINTS TO REMEMBER

- kNN is a non parametric algorithm.
- kNN predicts class of test data [xq] on the basis of neighbourhood.

WORKING OF kNN:

- Find distance (xq and all training data)
- Sort distance
- Pick k nearest neighbors
- Majority rate of class prediction

If x1 at (4,0) , x2 at (0,1) and x3 (5,0) and xq at (0,0). Which is nearest point to xq ?

1 user has participated

A	x1	0%
✓ B	x2	100%
C	x3	0%

End Quiz Now

Based on all quizzes from the session

2	Shoreya gupta	4/4 366.66
1	Karthik	4/4 379.01
3	Abdur Rehman	4/4 362.69

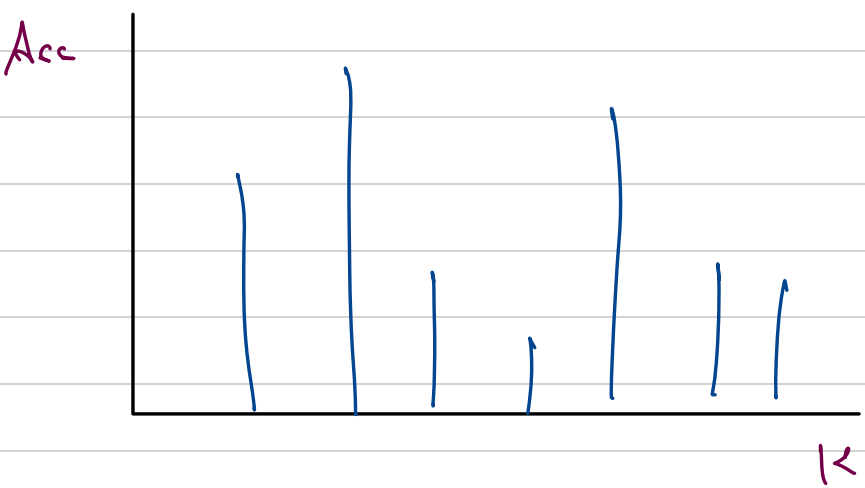
4	Sri Harsha Nanduri	4/4	357.16
5	OM PRAKASH S	4/4	355.33
6	Manjeet Singh	4/4	345.48
7	Santhosh	4/4	344.89
8	Aditya Shandilya	4/4	344.86
9	SHASHANK JHA	4/4	344.56
10	Umar	4/4	334.17

Wherever distance calculation is involved, you should always standardize your dataset.

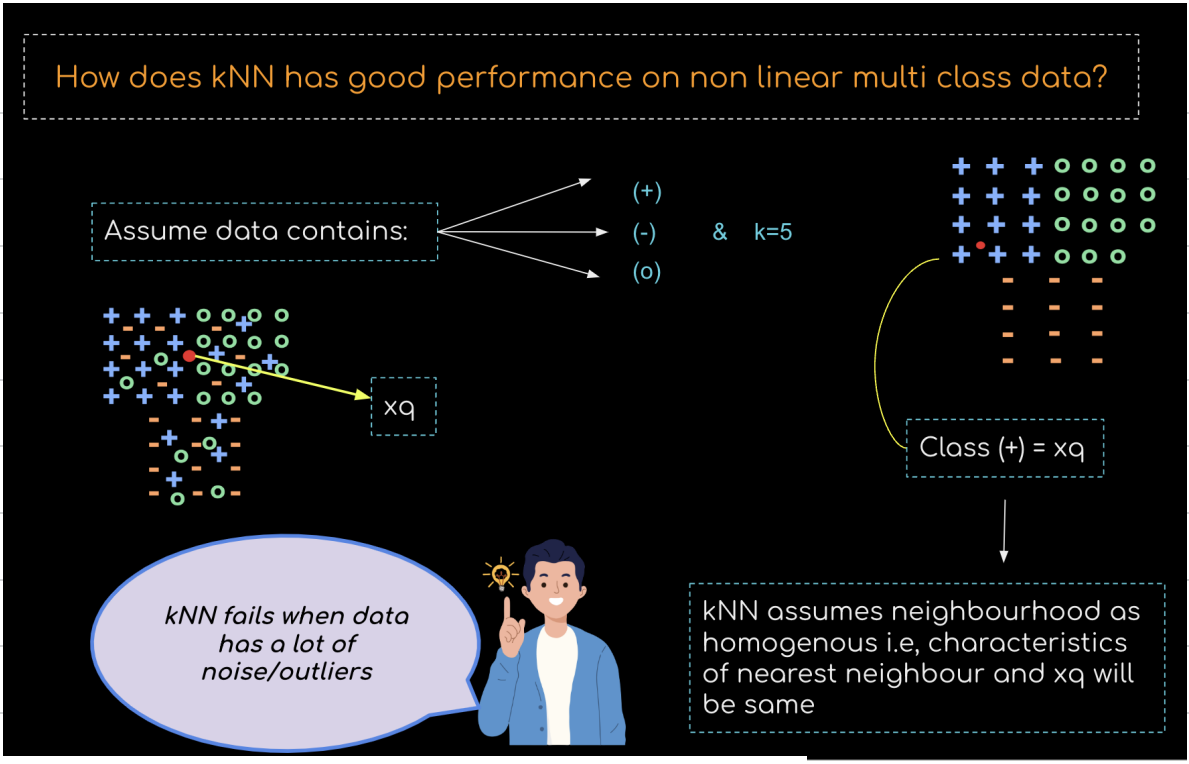
(60,000, 4) (40,000, 2)

A B

$\sqrt{(60k - 40k)^2 + (4-2)^2}$



1. How can the value of **k** impact underfitting and overfitting?
2. Time Complexity of KNN



how kNN is better than logistic regression. Select the correct option

1 user has participated

A	kNN has less time complexity	0%
B	kNN classifies data better	0%
C	kNN handles most noise/outlier	0%
✓ D	kNN handles multi-class problem	100%

End Quiz Now

Based on all quizzes from the session

2	Shoreya gupta	5/5 459.53
1	Karthik	5/5 477.06
3	OM PRAKASH S	5/5 447.70
4	Sri Harsha Nanduri	5/5 445.76
5	Manjeet Singh	5/5 428.62
6	Santhosh	5/5 415.89
7	RAHUL	5/5 411.76
8	Mohanakrishna	4/5 372.13
9	Abdur Rehman	4/5 362.69
10	Tanvi Singh	5/5 361.37

Closed Form → Linear Reg