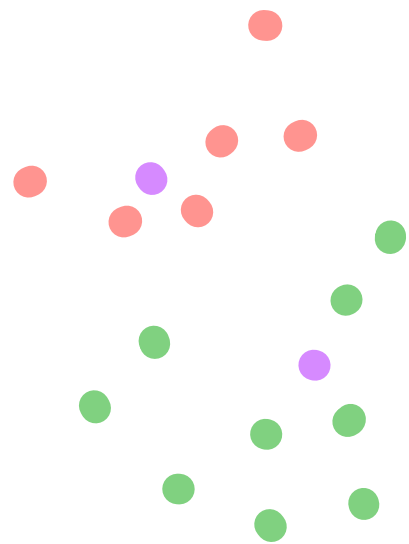
KNN-2

KNN

Bias - Variance Tradeoff

100 → 60
100 → 40



$k = 1$

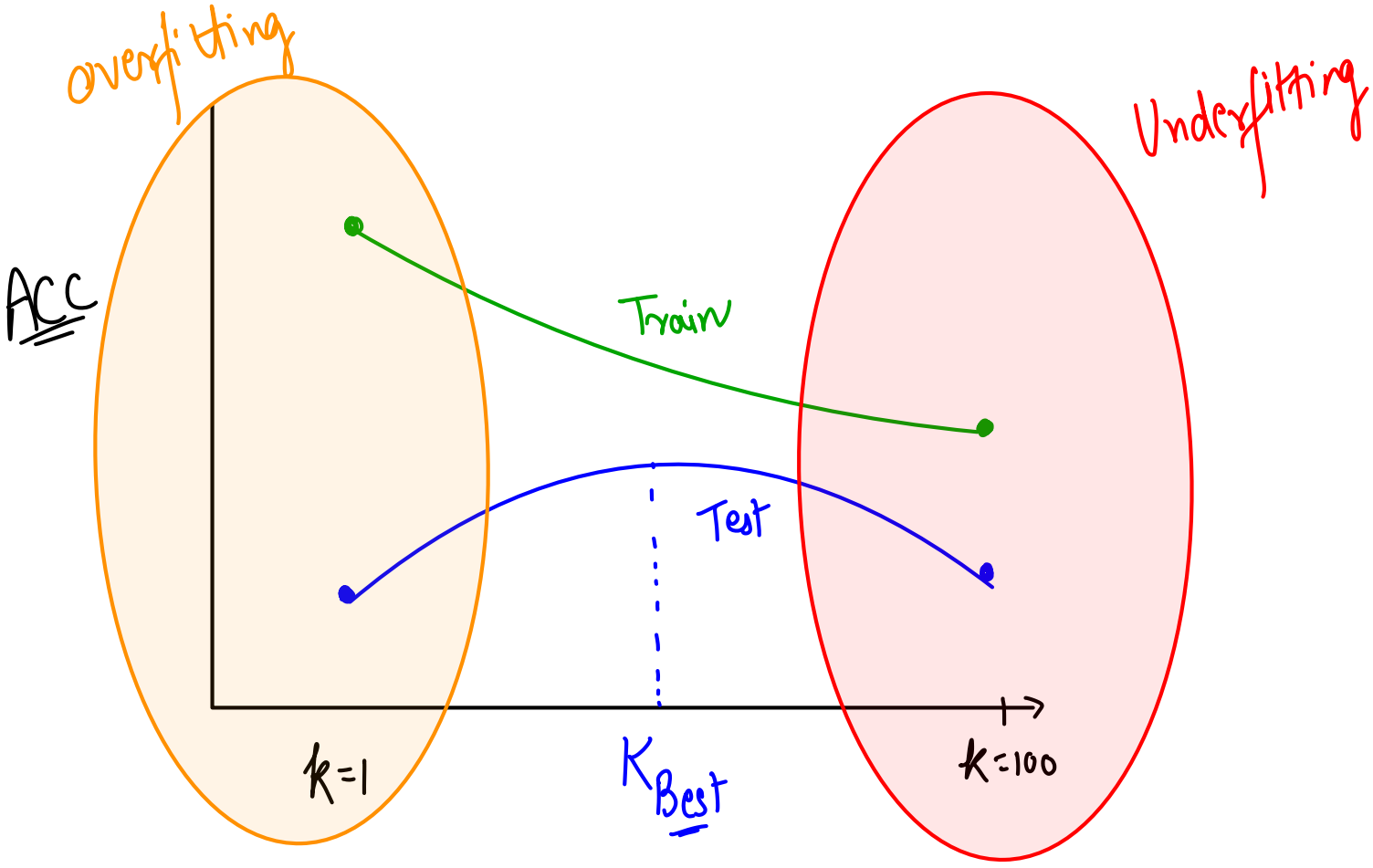
Overfitting

high variance

$k = 100$

Underfitting

high bias

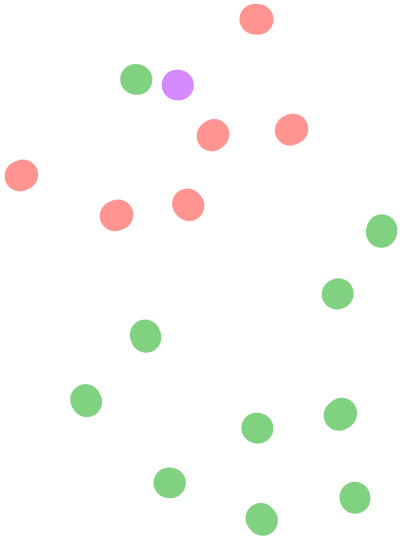

overfitting
Underfitting
ACC
Train
Test
k=1
K Best
k=100

Impact of Outlier

$k=1$ Impact of Outlier ↑↑↑
low

$k=7$ Impact of Outlier ↑
moderate

$k=100$ Impact of Outlier ↓↓↓
high



k ↑ Impact of Outlier ↓

Quiz time!

Quiz Ended!

what to say if model has high varaince and low bias ?

44 users have participated

A	Model overfits	80%
B	Model underfits	20%

Quiz time!

Quiz Ended!

k→ hyperparameter, then what data to use for hyperparamter tuning ?

34 users have participated

A	validation	82%
B	training	6%
C	test	3%
D	entire data	9%

Quiz time!

Quiz Ended!

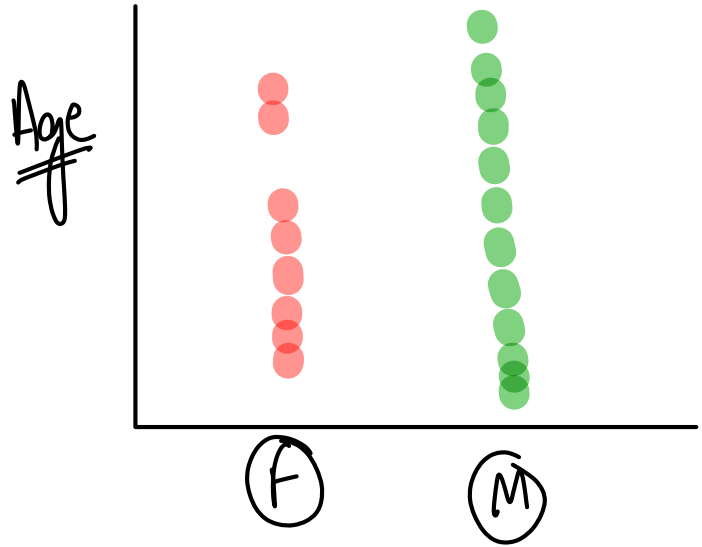
k-NN algorithm does more computation on test time rather than train time.

37 users have participated

A	TRUE	65%
B	FALSE	35%

KNN for Categorical features

gender	Blood group	Risk of heart diseases
		Y / N



Solution ⇒ Encoding

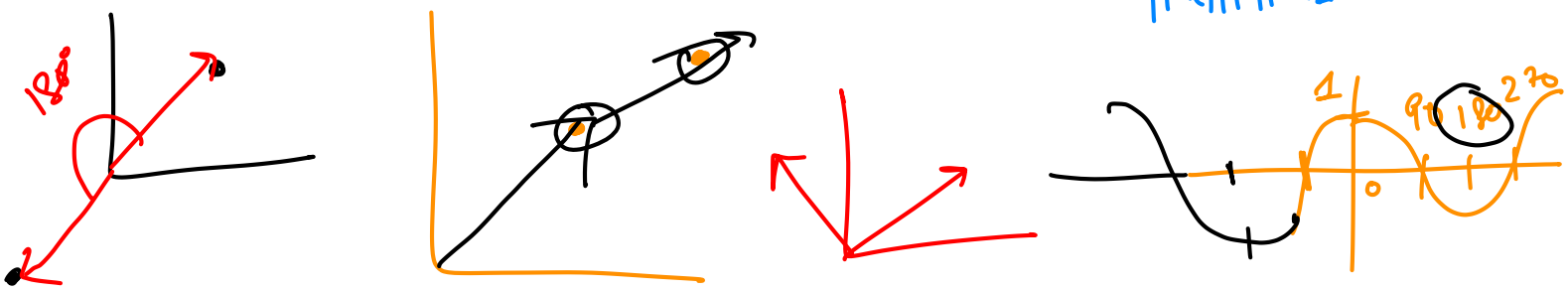
- → OHE → ✗
- → Label → ✗
- → Target → ✓

1000 → 1000 Columns ⇒ 1000 new features

⇒ " " dimensions.

✱ Euclidean will fail

Cosine Distance ⇒ $\cos\theta = \dfrac{x_1 \cdot x_2}{\|x_1\| \, \|x_2\|}$ → [-1, 1]



Quiz time!

Time Left: 13s

quiz (what do you think) if One Hot Encoding increases data dimension to (d=1000), will Eculidean Distance work ?

35 users have participated

A	Yes	20%
B	No	80%

End Quiz Now

Quiz time!

Quiz Ended!

Which of the following will be Manhattan Distance between the two data point A(1,1,3) and B(1,3,5)?

30 users have participated

A	1	7%
B	2	7%
C	4	83%
D	5	3%

Google KNN

[010]
[010110]
Qutub Minar (1)

[010]
[00001]
Red Fort (2)

[0 1 0]
[1 0 1 0 1 0]
Rashtrapati Bhawan, (3)

Marine Drive, Gateway of India,

[0 1 0 1 0 1]
[1 1 1] (4)

[1 0 1 0 1 0]
[1 1 1] (5)

key	value
Delhi	1, 2, 3
Mumbai	4, 5

Hashing Table { }

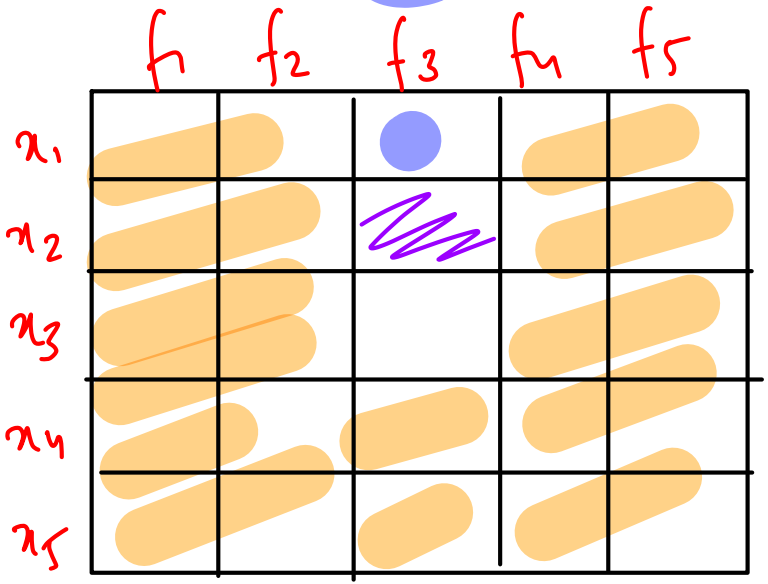
LSH ⇒ Locality Sensitive Hashing.

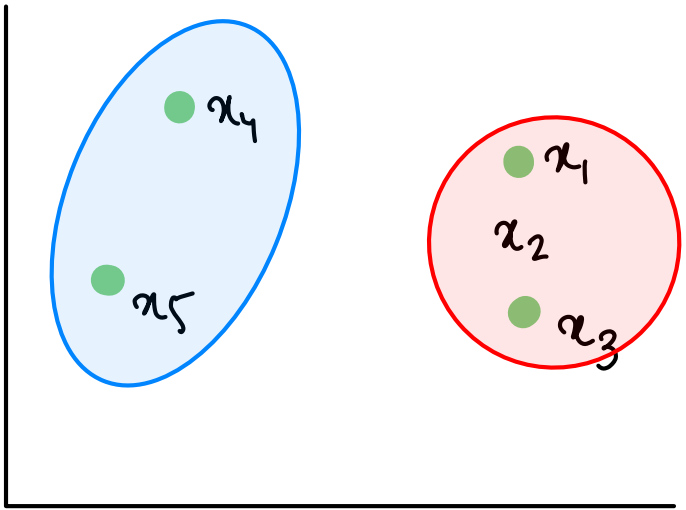
Missing data

median, mode, mean, fixed value,

KNN BASED IMPUTATION.

Imputation: Finding the best possible guess.





$$= \frac{+}{2}$$

Quiz time!

Quiz Ended!

Select the true statements

s1- kNN is less time intensive when LSH is used

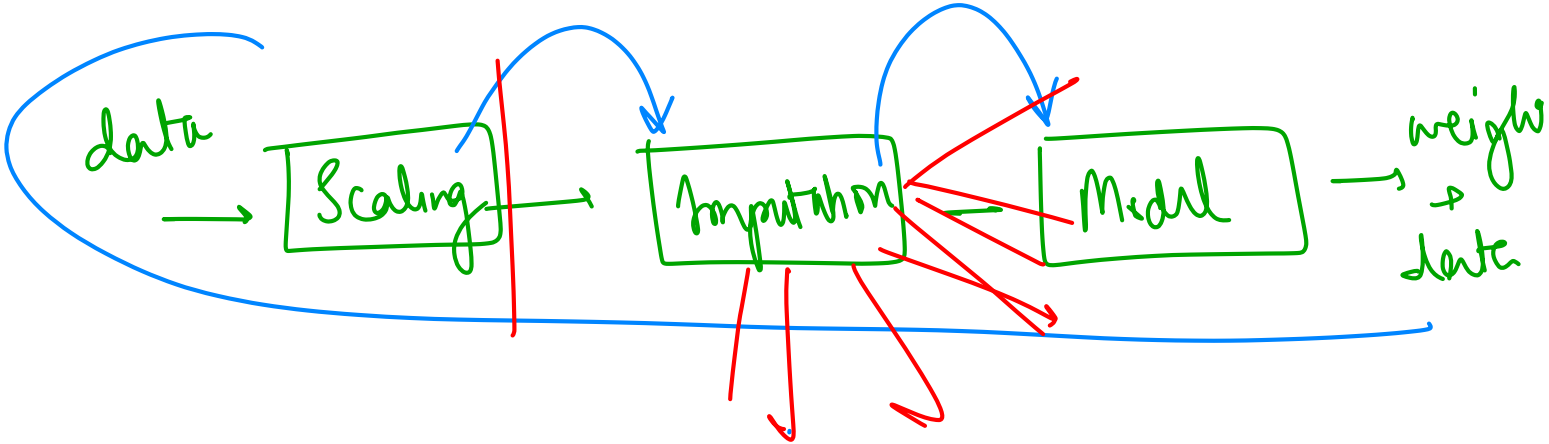
s2- k must be odd

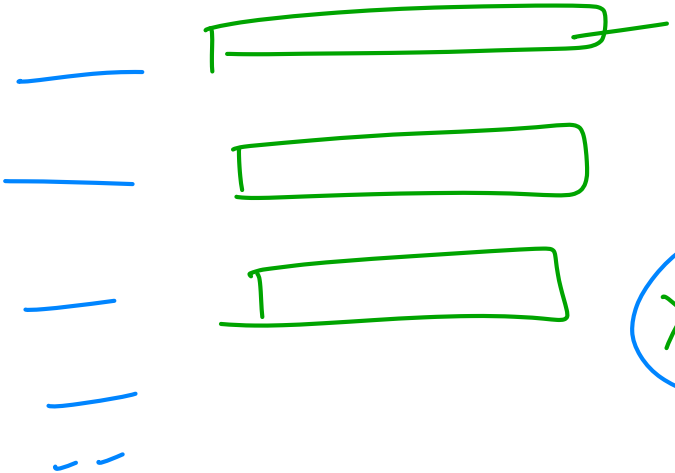
s3- kNN used for imputing

s4- For high dimension, euclidean not used

35 users have participated

A	s1	9%
B	s2	3%
C	s3	6%
D	all of the above	83%





scaler = StandardScaler ( )

X_sm = scaler.fit_transform(X)

X_final = KNNImputer(Xsm, k=5)

KNNClassifier(X_final)