☆ The technique we used in the last lecture is known as

Locality Sensitive Hashing (LSH)

☆ KNN Based Imputation (Imputation = Handling null values)

→ It is different from rebalancing the data. (SMOTE)

Length	Width	Weight	Type
25	10	100	1
30	15	350	1
5	1	15	0
40	10	250	1
50	20	700	1
45	15	300	1

← Imbalanced Data as there are more records belongs to class-1 than class-0.

∴ We use SMOTE.

len	width	weight	type
25	□	100	1
5	1	□	0
□	3	30	0
40	10	250	1
50	20	700	1
8	5	50	0

← This is not an Imbalanced Data but it has missing values. What can we do to them?

Ans- Next Page

Ans – (1) Remove those rows. – can be justified if our data is large enough & there is a less no. of null values.

(2) Replacing null values by some values. → Imputation

What should we replace the null values with?

(1) Zeroes – Not a right choice!

(2) Average of that column – Better than putting 0 but not the best option.

(3) Average of that column but only of that same class – seems the best option only because the data was small.

→ Let's consider dataset of height of the people from different areas of the world. Suppose we have height of 10,00,000 people from different parts of the globe, their weight and if they are diabetic or not (1 = diabetic, 0 = Not). Some of the values in height column are missing. Will it be the best choice to replace these values with average height of that class? No. It will make more sense if we replace null height by the average height of **Nearest Neighbors of that particular point**. This is known as **KNN Based Imputation.**