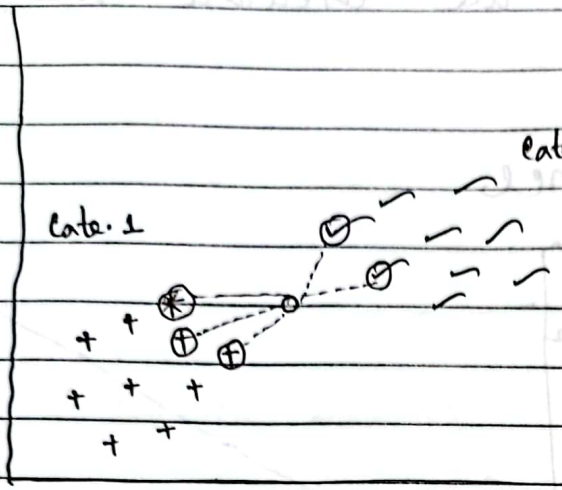
K- Nearest Neighbour. (KNN)



→ Classification Problem.

→ 2 Dimensional.

→ 2 Classification dataset
+, ~

→ 0 → new data point

→ 0 → new data point কোন ক্যাটেগরিতে যাবে that in our Problem.

Algorithm.

Step 1 : Choose the number K of neighbour.
By Default K → 5

Step 2: Nearest Neighbour point - from the new data Point.

Step 3: Calculate the "distance" of nearest neighbour.

Step 4: How many nearest neighbour belongs to each category [category 1 and 2]

Category-1 ⟶ 3 data point
Category-2 ⟶ 2 " "

} maximum ← new data point appears this category.

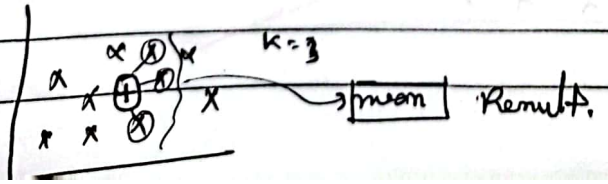
1. Choose the number K of Neighbour.
   Default k=5
2. Take the K nearest neighbours of the new data point, according to the Euclidean Distance
3. Among these K neighbors, count the number of data Points in each category.
4. Assign the new data point to the category where you counted the most neighbors.

Your Model is Ready.

[ML A-2]

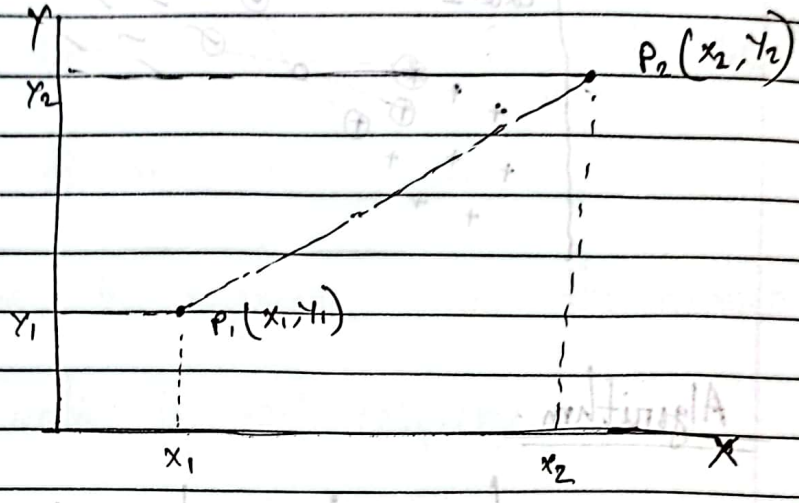
In Regression.

Select the nearest neighbors and their Mean is the output.



Now how we calculate the distance

① Euclidian Distance



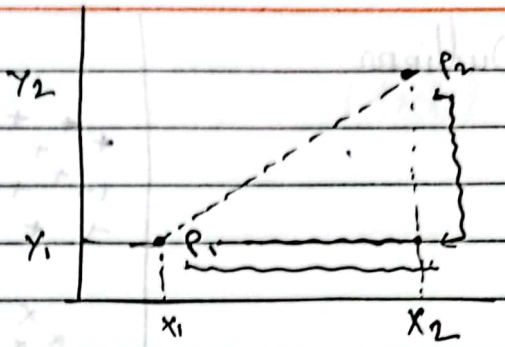
Distance between P_1 and P_2,

$$= \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

② Manhattan Distance formula, calculate as the sum of the absolute difference between the two vectors.

$$D(P, q) = \sum_{i=1}^{n} |P_i - q_i|$$

$|x_1 - x_2| + |y_1 - y_2|$



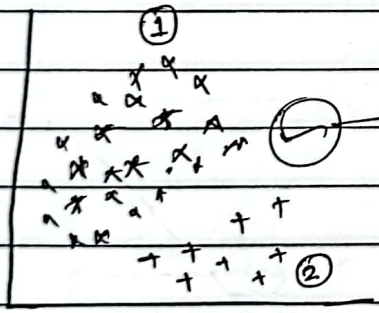
⇒ If a Imbalance Dataset,

→ 900 Yes

→ 100 No

will the KNN be biased with respect to the output.

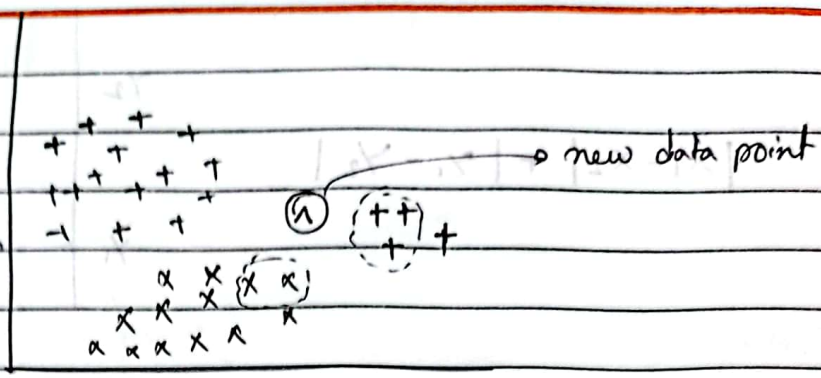
let's K = 150



So most of the neighbour point comes from (1) category.

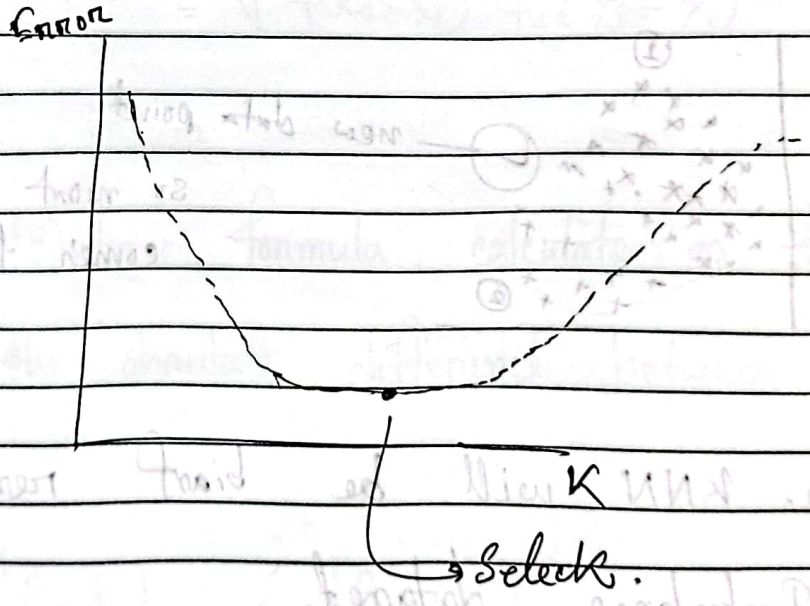
⇒ So, KNN will be biased respected to the Imbalance dataset.

→ Outliers

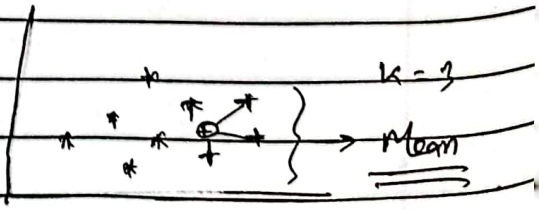


So because of outlier KNN will be biased or Impacted

How to Select K value.



K should be odd.



Regression.