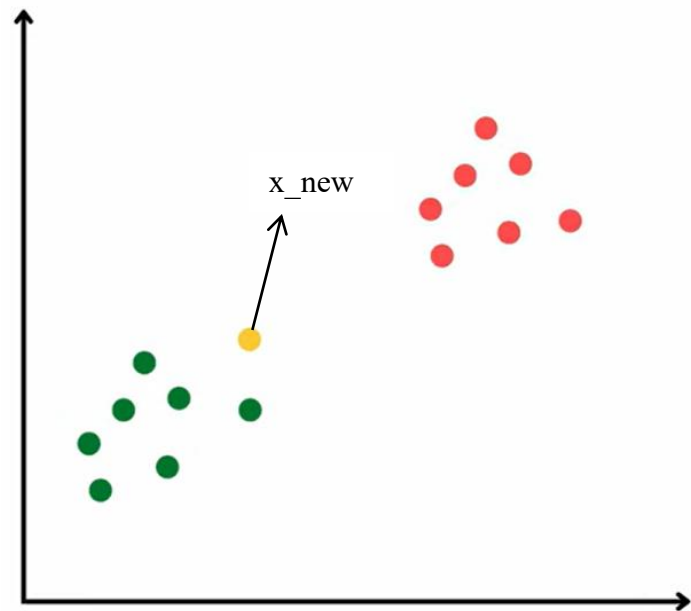


## K-Nearest Neighbors

K-Nearest Neighbors là một thuật toán dự đoán kết quả của dữ liệu dựa trên thông tin của những dữ liệu huấn luyện gần nó nhất.



*Tập dữ liệu với một điểm dữ liệu mới*

Mình cho thuật toán một điểm dữ liệu mới **x\_new**, nó sẽ:

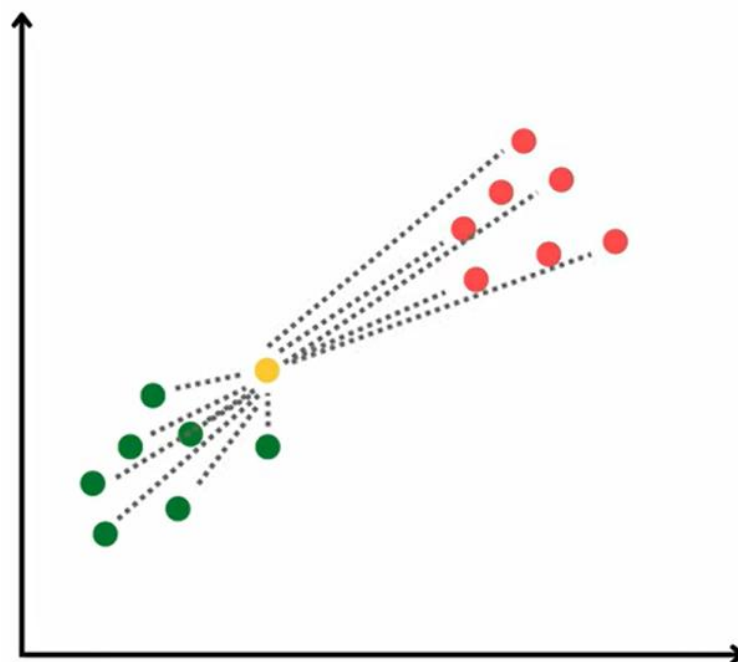
- Tính toán khoảng cách của điểm dữ liệu đó với các điểm dữ liệu khác có trong tập dữ liệu **x\_train**. Công thức tính khoảng cách thường có 3 loại cơ bản được sử dụng, tùy trường hợp mà mình sẽ lựa chọn cái phù hợp cho tập dữ liệu, với số lượng **n** mẫu dữ liệu:

➤ Khoảng cách Euclidean:  $\sqrt{\sum_{i=1}^n (x_{new_i} - x_{train_i})^2}$

➤ Khoảng cách Manhattan:  $\sum_{i=1}^n |x_{new_i} - x_{train_i}|$

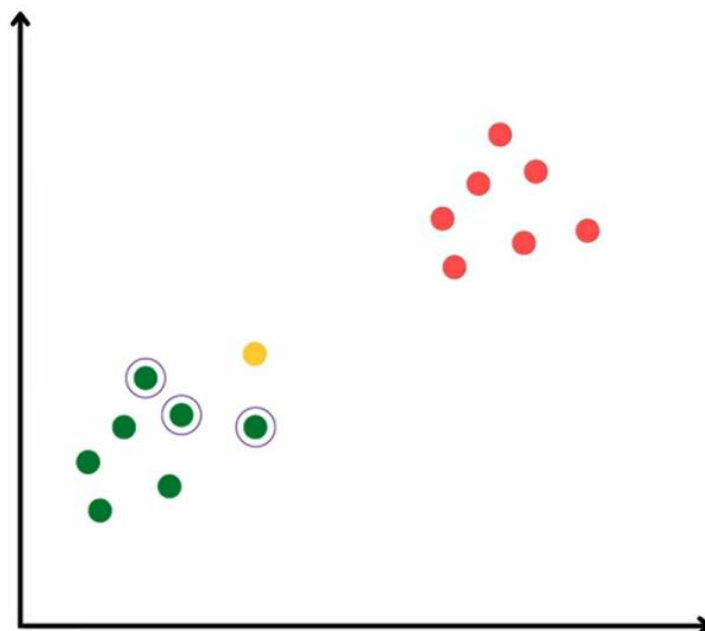
➤ Khoảng cách Minkowski:  $(\sum_{i=1}^n (|x_{new_i} - x_{train_i}|)^p)^{\frac{1}{p}}$

(với **p** là số dương)



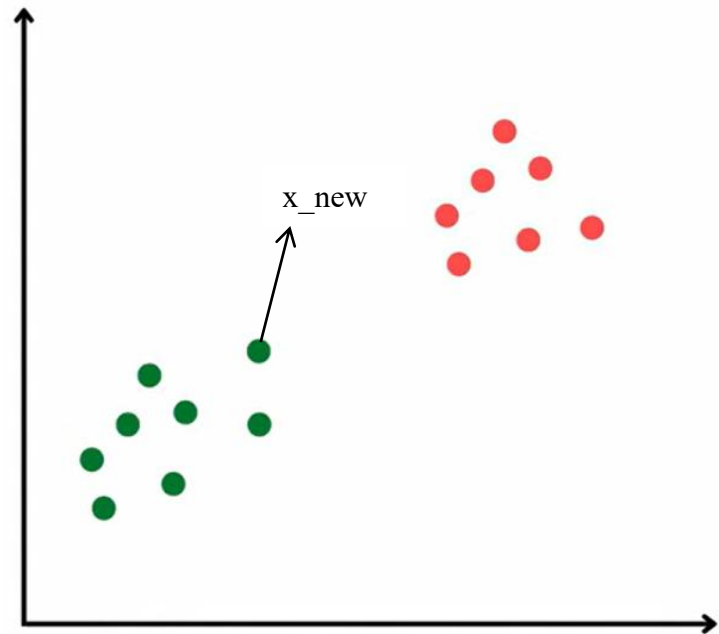
*Tính khoảng cách của điểm dữ liệu mới với tất các điểm dữ liệu khác.*

- Lấy số lượng **K** điểm dữ liệu trong tập có khoảng cách gần nhất với điểm dữ liệu mới.



*Tìm được 3 điểm dữ liệu màu xanh gần nhất với điểm mới ( $K = 3$  trong trường hợp này)*

- Kết quả:
  - Đối với dạng hồi quy (regression) của thuật toán: Mình có được giá trị trung bình từ **K** điểm dữ liệu.
  - Đối với dạng phân loại (classification) của thuật toán: Mình có được nhãn dữ liệu theo số đông trong **K** điểm dữ liệu xung quanh nó.



Theo số đông của 3 điểm dữ liệu gần nó,  **$x_{new}$**  thuộc về dữ liệu màu xanh  
(classification)