**KNN**

1. Định nghĩa

**K-Nearest Neighbors (KNN)** là một thuật toán học máy **không tham số** (*non-parametric*) và **dựa trên ví dụ** (*instance-based*), được sử dụng cho cả bài toán **phân loại** và **hồi quy**. [1]

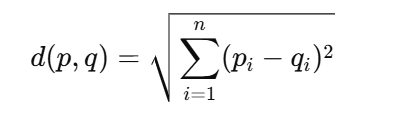
(Nguồn [1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York: Springer, 2009.)

Đối với bài toán **hồi quy**, KNN dự đoán giá trị đầu ra cho một điểm dữ liệu mới bằng cách tính **trung bình (hoặc trung vị)** của các giá trị mục tiêu (*target*) của **K láng giềng gần nhất** trong không gian đặc trưng.

Trong bài toán cụ thể này, ta sử dụng **K = 5**, tức là mô hình sẽ dự đoán giá trị đầu ra dựa trên trung bình của **5 điểm dữ liệu gần nhất**.

### **Khoảng cách Euclid**

Độ gần giữa các điểm được xác định bằng **khoảng cách Euclid**, với công thức:

[2]

Trong đó:

* p và q là hai điểm dữ liệu trong không gian đặc trưng.
* n là số lượng đặc trưng (*features*).

(Nguồn [2] "Khoảng cách Euclid," *Scholarship Wiki*, [Online]. Available:<https://wiki.scholarship.edu.vn/Kho%E1%BA%A3ng_c%C3%A1ch_Euclid>. [Accessed: 24-May-2025] )

**Evaluate metric**

| KNN | Train - test - validate | MAE | MSE | RMSE | MAPE | R² |
| --- | --- | --- | --- | --- | --- | --- |
| 75 - 15 -10 | 287.92 | 189633.04 | 435.47 | 2.56% | 0.86 |
| 70 - 20 -10 | 293.96 | 194234.23 | 440.72 | 2.62% | 0.86 |
| 65 - 25 - 10 | **291.42** | **157227.14** | **396.52** | **0.82%** | **0.89** |

**Mô hình KNN với tỷ lệ chia 65% Train – 25% Validation – 10% Test cho kết quả tốt nhất**, với độ chính xác cao (R² = 0.89), sai số tương đối thấp, và phân phối dữ liệu hợp lý giúp cải thiện khả năng tổng quát hóa.

