**Tóm tắt về thuật toán K-Nearest Neighbors**

* Thuật toán K-Nearest Neighbors (KNN) là thuật toán học máy có giám sát đơn giản trong Machine Learning, được ứng dụng rộng rãi vào các bài toán phân loại và hồi quy. Ý tưởng chính của thuật toán này chính là việc tập hợp các điểm dữ liệu gần nhau trong không gian thì sẽ tổng hợp thành 1 nhóm
* Thuật toán này nằm trong nhóm các thuật toán dựa trên bộ nhớ ( học lười/lazy learning ) do thuật toán này không học bất kì các đặc điểm, tính chất gì của bộ dữ liệu training. Thay vào đó, nó so sánh mối quan hệ đầu vào với lần lượt các bộ dữ liệu lưu trữ trong bộ nhớ và tính toán để đưa ra kết quả.
* Các thông số quan trọng của thuật toán bao gồm:
* K: số điểm dữ liệu lân cận (hàng xóm) gần nhất. Chọn K phù hợp để tránh bị nhiễu (K quá bé) hoặc tránh bỏ qua các chi tiết quan trọng ( K quá lớn)

## K quá nhỏ 🡪 (1) chỉ có 1 hàng xóm quyết định lớp🡪 Dễ ảnh hưởng bởi nhiễu/ ngoại lai. K quá lớn 🡪 chọn quá nhiều hàng xóm, kể cả hàng xóm xa 🡪Mất tính lân cận.

* Khoảng cách giữa các điểm dữ liệu tùy thuộc vào các thuộc tính của chúng (thông thường là KG Euclid)
* Cách thức thuật toán hoạt động:
* Lưu trữ các bộ dữ liệu huấn luyện
* Khi có đầu vào là 1 điểm mới cần dự đoán nhãn, ta tính khoảng cách từ điểm dữ liệu mới đến tất cả các điểm dữ liệu trong tập huấn luyện
* Lựa chọn K hàng xóm lân cận có khoảng cách gần nhất với điểm dữ liệu mới
* Tùy vào yêu cầu bài toán, ta đưa đầu ra:  
  Phân loại 🡪 gán điểm dữ liệu mới nhãn lớp xuất hiện nhiều nhất.

Hồi quy🡪 Tính giá trị trung bình của các điểm dữ liệu trong K hàng xóm gần nhất và gán giá trị đó cho điểm dữ liệu mới.

* **Ưu điểm**
* **Dễ hiểu và triển khai:** Thuật toán rất dễ nắm bắt và không yêu cầu nhiều kiến thức toán học phức tạp.
* **Linh hoạt:** Có thể áp dụng cho cả bài toán phân loại và hồi quy.
* **Không có giả định về phân phối dữ liệu:** thuật toán phi tham số, không yêu cầu dữ liệu phải tuân theo một phân phối xác suất cụ thể.
* Nhược điểm:
* Hiệu năng kém với tập dữ liệu lớn: việc tính khoảng cách giữa một điểm dữ liệu mới và tất cả các điểm dữ liệu trong tập huấn luyện có thể rất tốn kém về tài nguyên.
* Kết quả dễ bị dao động bởi nhiễu: Nếu dữ liệu có nhiều điểm nhiễu, kết quả có thể bị sai lệch so với thực tế chính xác