KNN Classifier

KNN হচ্ছে একটি Supervised Learning Algorithm, যা Classification এবং Regression– দুই ধরনের কাজে ব্যবহার করা যায়, তবে এটি বেশি ব্যবহৃত হয় Classification–এর জন্য।

KNN একটি **Instance-based** বা **Lazy learning** অ্যালগরিদম। এটি কোনো ট্রেনিংয়ের সময় কোনো মডেল তৈরি করে না। বরং যখন নতুন ডেটা আসে, তখন এটি ট্রেনিং ডেটাসেটের উপর ভিত্তি করে সিদ্ধান্ত নেয়।

কিভাবে কাজ করে?

1. **K মান নির্ধারণ:** প্রথমে একটি সংখ্যা নির্ধারণ করতে হয়, যাকে **K** বলা হয় (যেমন: K=3, K=5 ইত্যাদি)।

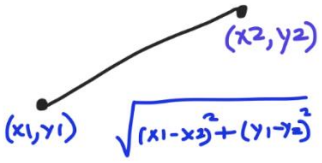
   Rule of Thumb

- $k = \sqrt{n}$ N হল ট্রেনিং উদাহরণের সংখ্যা।

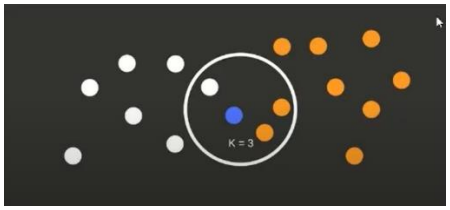
  উদাহরণ

  - যদি তোমার ১০০**টি** ট্রেনিং পয়েন্ট থাকে $(N = 100) \rightarrow \sqrt{100} = 10.$
  - যদি ১০০০**টি** পয়েন্ট $(N = 1000) \rightarrow \sqrt{1000} \approx 31.6.$

2. **Distance হিসাব:** নতুন (unseen) ডেটা পয়েন্ট থেকে ট্রেনিং ডেটার প্রতিটি পয়েন্টের দূরত্ব (Distance) হিসাব করা হয়। সাধারণত **Euclidean distance** ব্যবহৃত হয়:



3. **Kটি নিকটতম প্রতিবেশী নির্বাচন:** দূরত্বের ভিত্তিতে সবচেয়ে কাছের **Kটি ডেটা পয়েন্ট** নির্বাচন করা হয়।



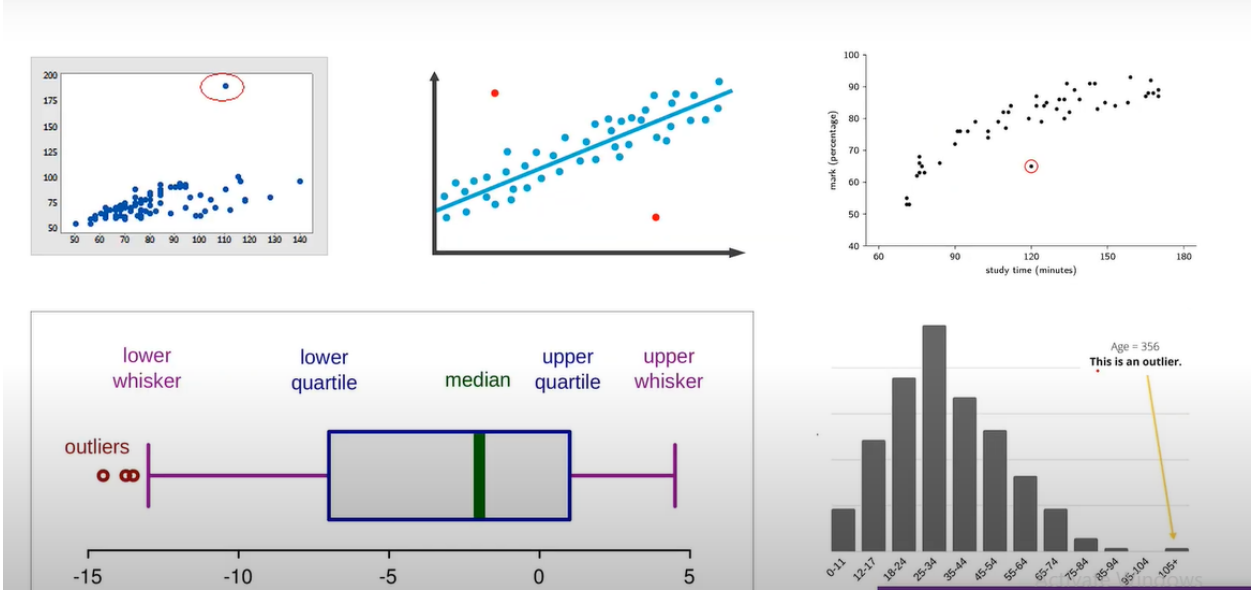
4. **ভোটিং বা গড়:**
    - **Classification:** Kটি প্রতিবেশীর মধ্যে যেই ক্লাস সবচেয়ে বেশি রয়েছে, সেটিই নতুন ডেটার ক্লাস হিসেবে গণ্য করা হয়।

KNN–এর সুবিধা:

- সহজ এবং ব্যাখ্যা করা সহজ।
- কোনো ট্রেনিং পর্যায় দরকার হয় না (Lazy learning)।
- নতুন ডেটায় ভালো কাজ করে যদি সঠিক K ও স্কেলিং করা হয়।

KNN–এর অসুবিধা:

- ডেটাসেট বড় হলে অনেক সময় নেয়।
- Outlier ও Irrelevant ফিচারে প্রভাব পড়ে।



- Feature scaling না করলে ফলাফল খারাপ হতে পারে।

  ডেটার প্রতিটি ফিচারের (feature/column) scale যেন সমান হয়।

  StandardScaler কী করে?

  StandardScaler প্রতিটি ফিচারকে **mean = 0 এবং standard deviation = 1** এ কনভার্ট করে।

  ➤ **ফর্মুলা:**

$$z = \frac{x - \mu}{\sigma}$$

যেখানে:

- x = মূল ডেটা ভ্যালু
- μ = ফিচারের গড় (mean)
- σ = ফিচারের স্ট্যান্ডার্ড ডিভিয়েশন (standard deviation)

🎯 এতে করে:

- ফিচারগুলো **centered** হয় (mean হয় 0)
- ফিচারগুলোর **স্কেল** সমান হয় (std হয় 1)

কেন স্কেল করা দরকার?

- অনেক ML অ্যালগরিদম (যেমনঃ KNN, SVM, Logistic Regression ইত্যাদি) ডেটা পয়েন্টের মধ্যে দূরত্ব (Distance) নির্ণয় করে কাজ করে। যদি ডেটার কোন ফিচার বড় স্কেলে থাকে আর আরেকটা ছোট স্কেলে, তাহলে বড় স্কেলের ফিচার অ্যালগরিদমকে প্রভাবিত করতে পারে। ফলে মডেল Misleading হতে পারে।

- `StandardScaler()` ফিচারগুলোর মানকে একক স্কেলে আনতে সাহায্য করে।
- এটি distance-based মডেলের পারফরম্যান্স উন্নত করে।
- শুধু ট্রেনিং ডেটায় fit করে, কিন্তু একই স্কেল টেস্ট ডেটায়ও প্রয়োগ করা উচিত।
- Imbalanced ডেটাসেটে তৈরি হতে পারে।

KNN ব্যবহারের ক্ষেত্র:

- Image Recognition
- Recommendation Systems
- Medical Diagnosis
- Text Classification