Feature Engineering:

এটি হল মেশিন লার্নিং এর এমন একটি ধাপ, যেখানে ডাটা থেকে এমন ইনফরমেশন তৈরি করা হয় যেটা মডেলকে বুদ্ধিমান করে তোলে।

যেমন: শুধু জন্মসাল এর ডাটা প্রদান করলে মেশিন কিছু বুঝবেনা, কিন্তু যদি Age নামক একটা কলাম যোগ করি তাহলে এটি হবে নতুন একটি Feature.

Feature Engineering আর Feature Transformation একই।

→ মেশিনকে মানুষের মতো বুঝতে শেখায়।

→ সে যেটা দেখে চিন্তা করতে পারে।

→ চিন্তার উপর base করে predict করতে পারে।

Distance Based Algorithm

Feature Transformation includes:

→ Scaling and normalization
→ Logarithmic Transformation
→ polynomial features
→ Dimensionality Reduction
→ Encoding catagorical values
→ Handling missing values
→ Transforming text data.

Why is Feature Transformation important?

→ Model performance
→ Computational Efficiency
→ Improved Interpretability
→ Handling Non-linearity

Regression Algorithm :

Continuous value predict করতে ব্যবহার করা হয়।

① Linear Reg
② Rig "
③ Lasso "
④ Elastic Net "
(Rig + Lasso)

এছাড়া Tree based regression and Non-linear Algorithm আছে।

Clanification algorithm list:

কোনো ইনপুট ডেটা কোন গ্রুপের
পড়ে সেটা অনুমান করা,
(Yes/No, Cat/Dog), (Pass/Fail)
ইত্যাদি,

1. Logistic Regression
2. k-Nearest Neighbours
3. Naive Bayes

Tree based clanifiers,
Kernel based Algorithm.

Euclidean distance:

দুই পয়েন্টের মধ্যে সবচেয়ে কাছের সরল রেখার দূরত্ব (L2 distance). $d = \sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}$

→ ML এ data পয়েন্ট এর মিল বা অমিল বুঝানোর জন্য।

→ যত কম হবে, তত বেশি মিল

Manhattan distance : (L1 Norm)

দুটি পয়েন্টের মধ্যে সোজা দিকে চলার দূরত্ব। একে Taxicab distance ও বলে।

* কিন্তু diagonal চলা যায় না,

উদাহরণ: x, $(2,3)$

y, $(5,7)$

$\therefore$ M.D. $= |5-2| + |7-3|$

$= 3 + 4 = 7$

→ Robotics: রোবট কোন দিকে চলবে

→ Image processing: দুটি pixels মধ্যে দূরত্ব,

→ path finding games: যেমন গ্রিড ভিত্তিক game.

Minkowski distance :

একটা আলোচ্য বিষয় যা নিয়ে E.D. ও M.D. - দুইবার ব্যবহার হয়।

Suppose
if $p = 1 \rightarrow$ MD [এখন Formula ও বলব MD এর Formula বলে বলব]
if $p = 2 \rightarrow$ ED [এখন]

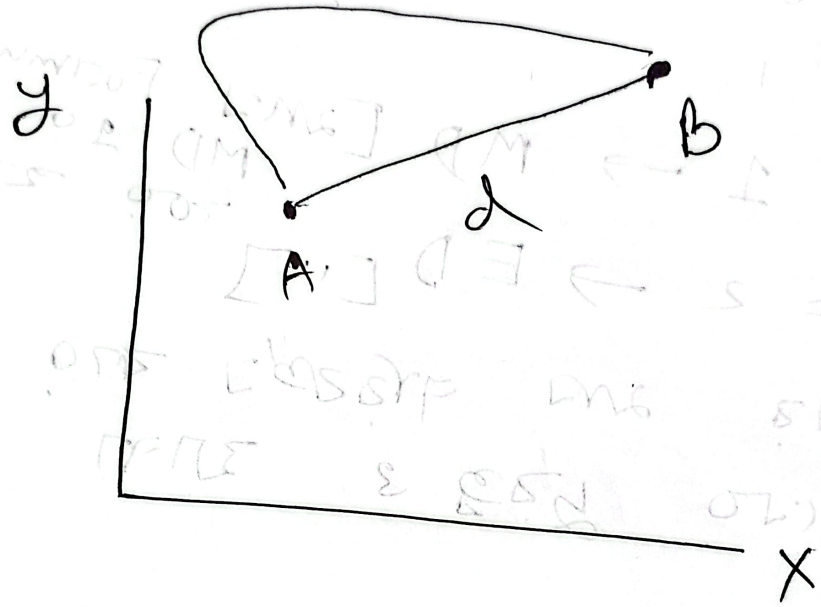
p এর মান পরিবর্তন করে বিভিন্ন ধরনের দূরত্ব ও মান বের করা যায়।

Formula :

$$\text{Distance} = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

এখানে,
x_i , y_i হলো দুই বিন্দুর স্থানাঙ্ক।
p হলো অর্ডার যা ধরিব করব। (1, 2, 3 ... n)

→ Pattern Recognition
→ Clustering Algorithm
→ Distance tuning



Note: $p \neq 0$ is not valid in Minkowski distance.

Error: $|x_i - y_i|^0 = 1$

Error: $\frac{1}{p} = \frac{1}{0} =$ undefined

Hamming distance :

দুইটা সমান Length এর string এর মধ্যে কতগুলো position এ ভিন্নতা আছে, সেটা গুণে বের করা।

Str 1 : "Karim"

Str 2 : "Kasim"

ভিন্নতা 1টা ($r \neq s$)

∴ Hamming distance = 1

→ Error detection & correction

→ DNA / protein sequence তুলনা

→ Cryptography → ব্যবহৃত।

করতে হয় লাগে

Feature Transformation Techniques

1. Normalization
2. standardization
3. Log Transformation
4. Robust Scaler
5. Max absolute Scalar

1. Normalization (Min-max scaling)

Data কে 0 থেকে 1 এর মধ্যে আনে।

Formula:

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Ex:

Suppose, কোনো height 120 cm — 180 cm এর মধ্যে। কোনো value 150 cm। তাহলে,

$$\frac{150-120}{180-120} = \frac{30}{60} = 0.5$$

এখন, height = 0.5 হলে বলা যায় মাঝামাঝি আছে।

2. Standardization (Z-score scaling)

Data কে standard scale এ আনে যেখানে mean = 0, SD = 1 হয়।

Formula:

$$X_{scaled} = \frac{X - \mu}{\sigma}$$

Ex:

একটা ছাত্র GPA 3.5, ক্লাস GPA 3.0, SD = 0.5

$$\frac{3.5 - 3.0}{0.5} = 1.0$$

মানে, ছাত্র = 1 standard deviation above average" ভালো result।

3. Log Transformation:

যেসব Data অনেক বেশি বাঁকা (skewed), তাদের ঠিক করে।

Formula:

$$X_{transformed} = \log(X)$$

Ex:

যেমন salary 1,000 → log(1000) = 3

আরেকজনের 10,000 → log(10,000) = 4

বড় Data এর difference ছোট হয়ে গেলে সহজে তা বুঝতে পারি।

4. Robust Scaler:

outlier (কার্যকারী বড় / ছোট মান) থাকলে ভালো কাজ করে, Median এবং interquartile range ব্যবহার করে।

Ex:

ধরো, ১০ জনের income 50-70 হাজার, এবং ১ জনের income ৩ কোটি টাকা — এই ব্যক্তিদের মান যদি Model এ (সঠিক) Robust না করে, scaler ভাঙা করে।

5. Max Absolute Scaler:

সব Data কে একটা রেঞ্জে নিয়ে আসে যা [-1, 1] scale এ থাকে।

Formula:

$$X_{scaled} = \frac{X}{|X_{max}|}$$

Ex:

Feature : [-5, 0, 50, 100]

Max = 100

এখন সব value কে 100 দিয়ে ভাগ করলে হবে : [-0.5, 0, 0.5, 1]