Machine Learning Process:

1 Data Preparation:

1 → Data Collection:

Kaggle, UCI, Hugging Face, Google, Amazon, Microsoft, Awesome public, CV

2 → Data preprocessing:

① Data cleaning ⇒ Undesired data (none Value)

→ remove row from dataset.

→ replace them by Column average.

② Data encoding ⇒ ML require numerical input

→ Convert Categorical Value to numerical.

→ target must also be Converted to numerical if it's Categorical (nominal) Values.

1 binary features (gender – sign(+,-) ...) ⇒ 0,1

2 Feature has more than 2 values

② one-hot encoding

Country	X_1	X_2 ...
Egypt	1	0
UK	2	2
US	3	1
US	4	4

⇒

Egypt	UK	US	X_1	X_2 ...
1	0	0	1	0
0	1	0	2	2
0	0	1	3	1
0	0	1	4	4

① replace them by numbers

egypt : 0

UK : 1

US : 2

③ Data Normalization ⇒ different ranges for each features.

$X_1 \rightarrow 0 - 10$	$X_2 \rightarrow 0 - 100000$

→ make them have the same range

new range : $X_{new} = \frac{X_{old} - X_{min}}{X_{max} - X_{min}}$ $(0 \rightarrow 1)$

④ Feature selection/dimensionality reduction

⇒ large number of features.

1 → Feature selection : select the most important features and drop the ~~rest~~ others.

2 → Dimensinality reduction : calculating new features from the original.

⑤ Data spliting ⇒