

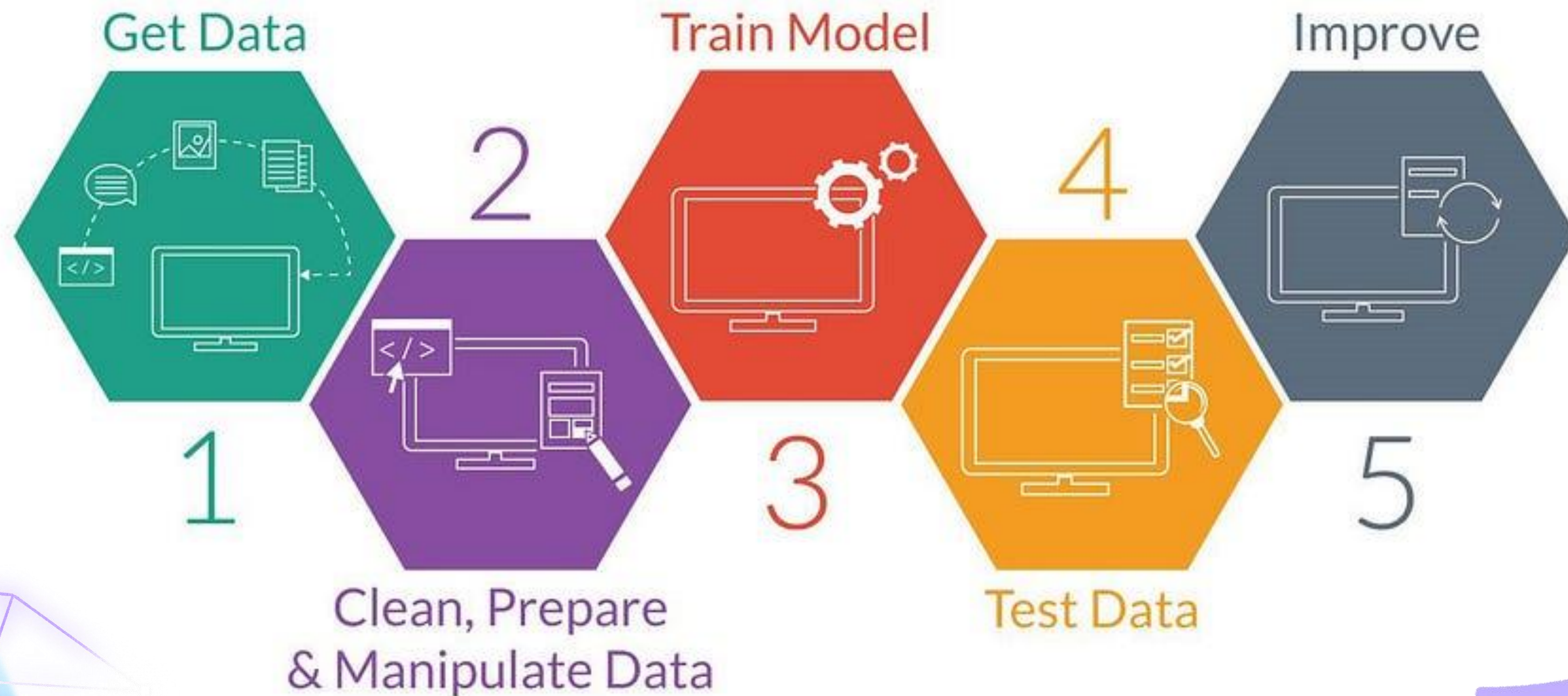
MACHINE LEARNING

DATA

PREPROCESSING

By Kittithat Kongtaworn (P'Junior)

ML Workflow

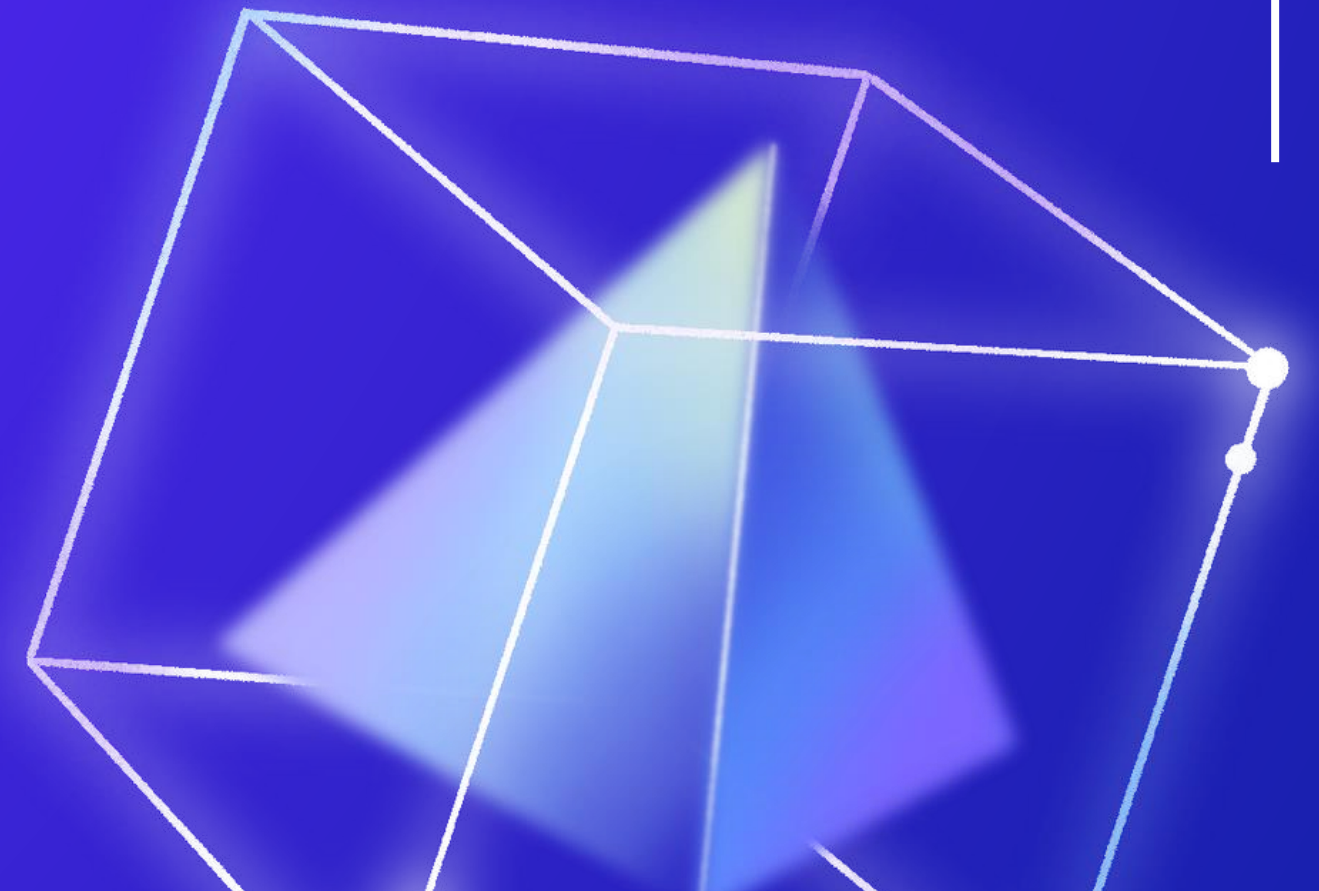


Ref: <https://medium.com/nerd-for-tech/the-ideal-workflow-for-your-machine-learning-project-9df1a7125b17>

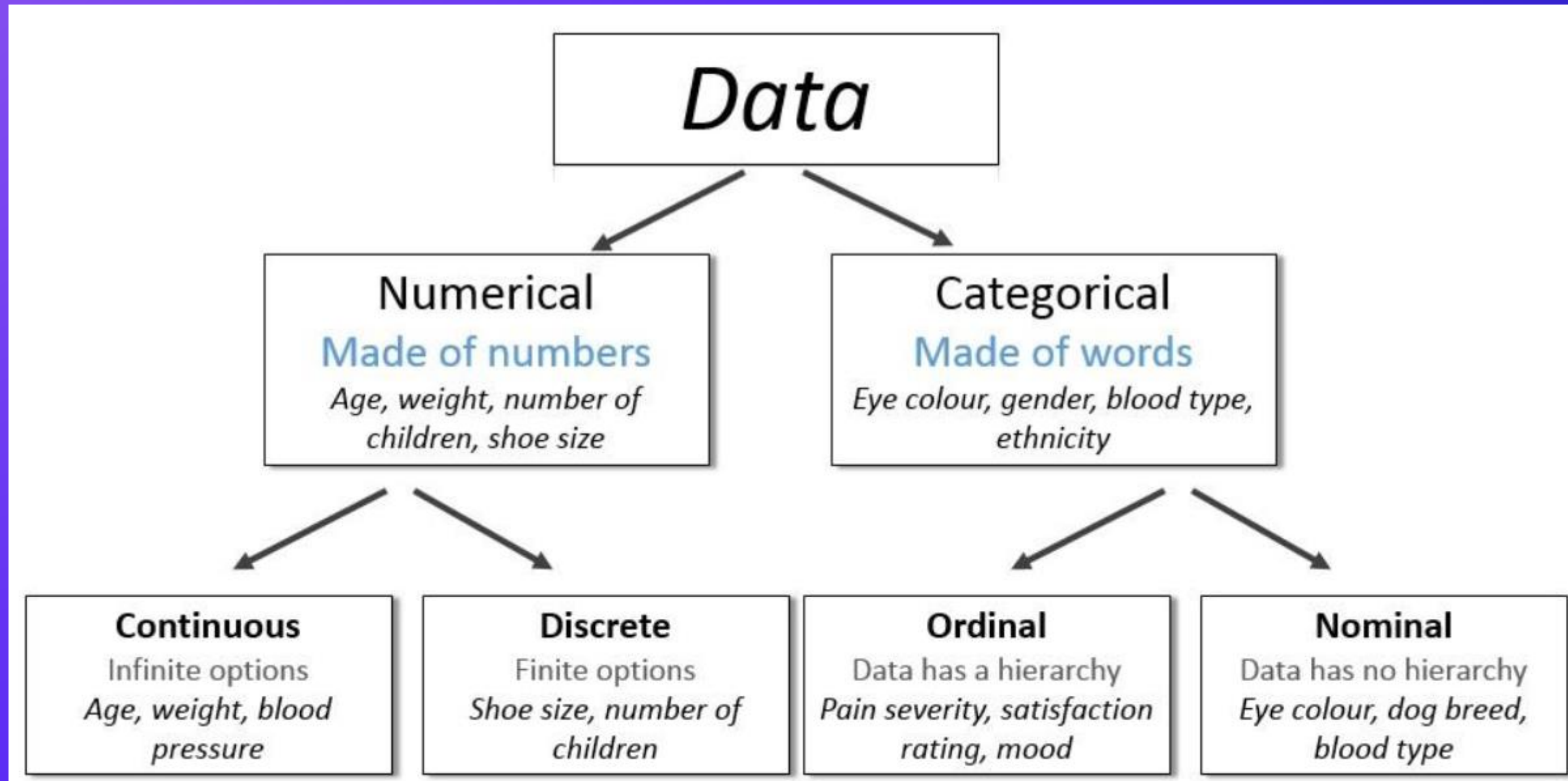


TABLE OF CONTENTS

- One-pot encoding
- Data splitting
- Missing value
- Data Normalization



Type of data

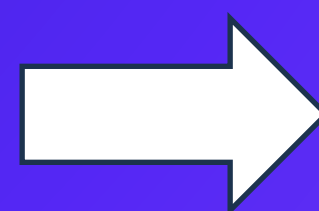


One-pot encoding

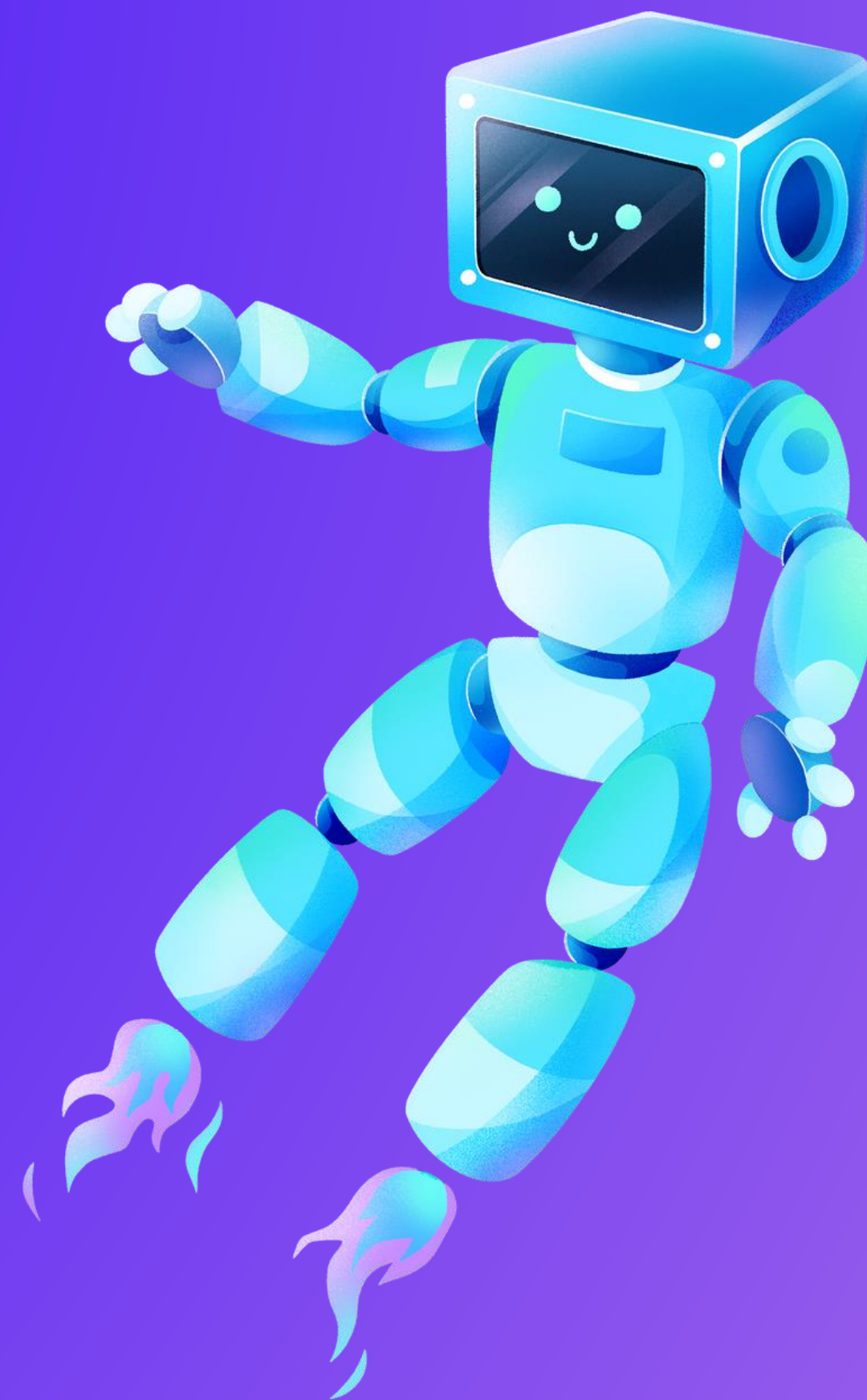
What is One-Pot encoding?

	location	menu	price
0	[1, 1, 1, 0]	[1, 1, 0]	[1, 0]
1	[0, 1, 0, 0]	[0, 0, 1]	[1, 0]
2	[1, 1, 0, 1]	[1, 1, 1]	[0, 1]

ตัวอย่างของการทำ one hot encoding



Binary value



One-pot encoding

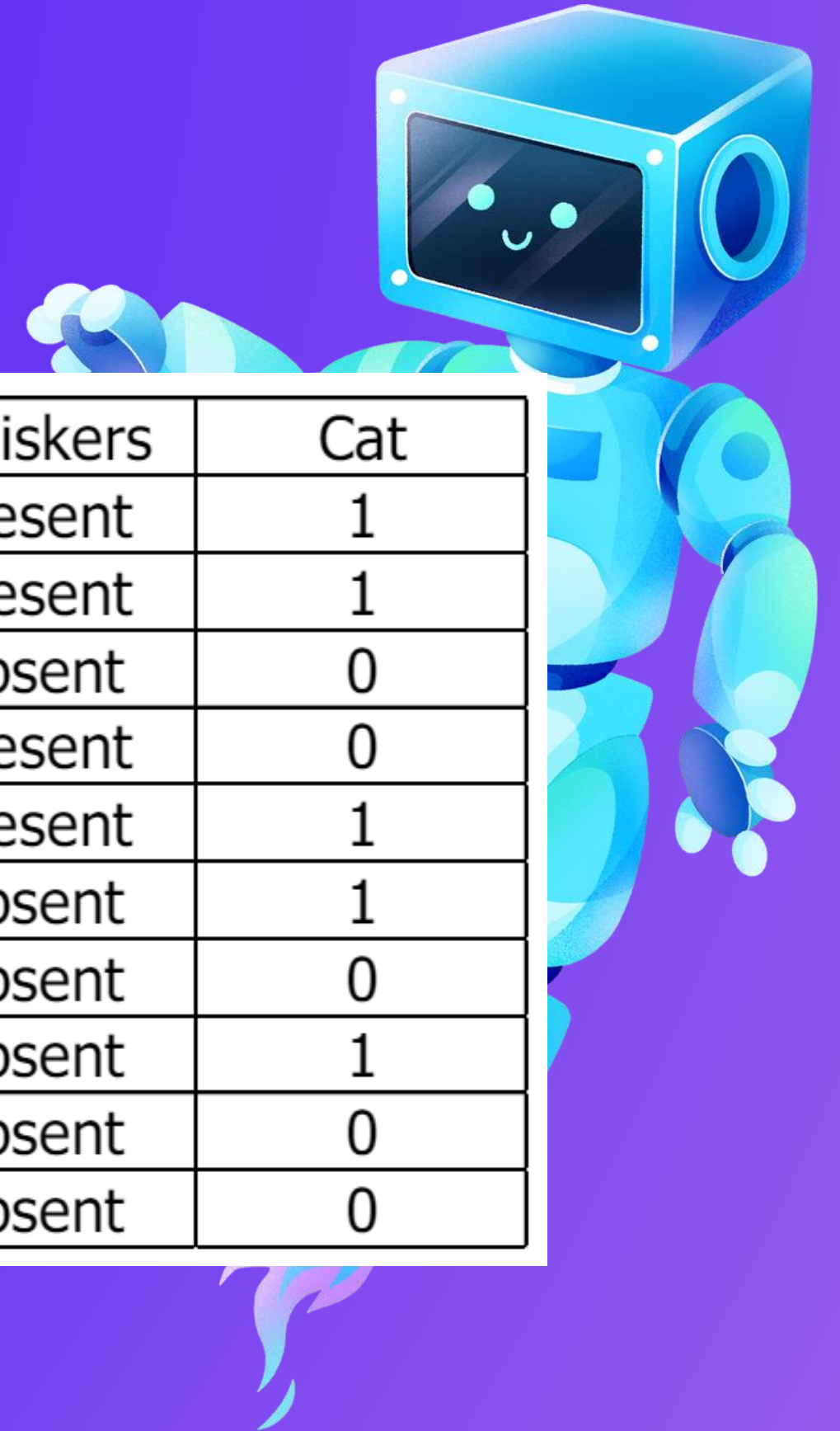


Ear Shape	Face shape	Whiskers	Cat
Pointy	Round	Present	1
Oval	Not round	Present	1
Oval	Round	Absent	0
Pointy	Not round	Present	0
Oval	Round	Present	1
Pointy	Round	Absent	1
Floppy	Not round	Absent	0
Oval	Round	Absent	1
Floppy	Round	Absent	0
Floppy	Round	Absent	0

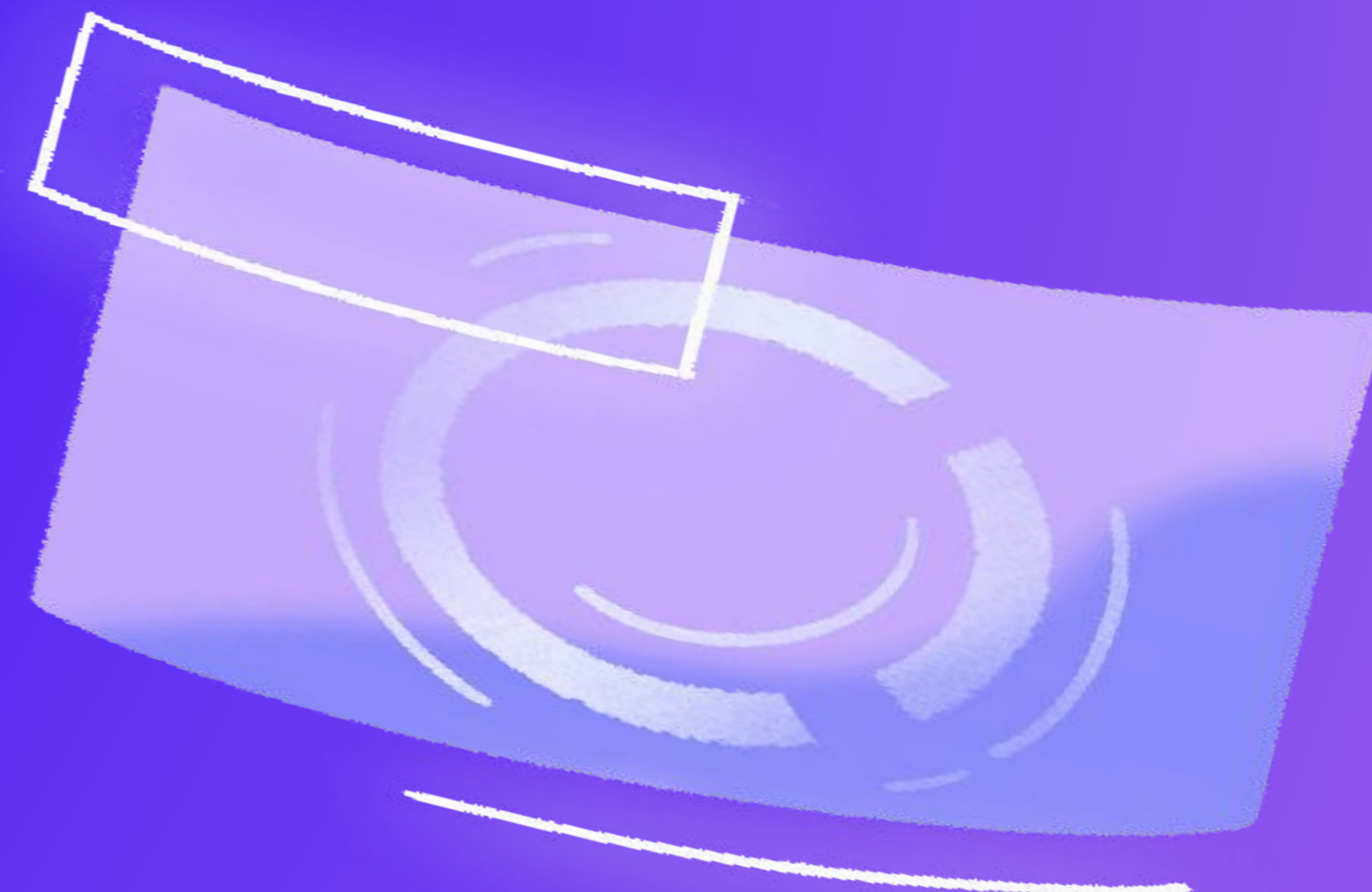
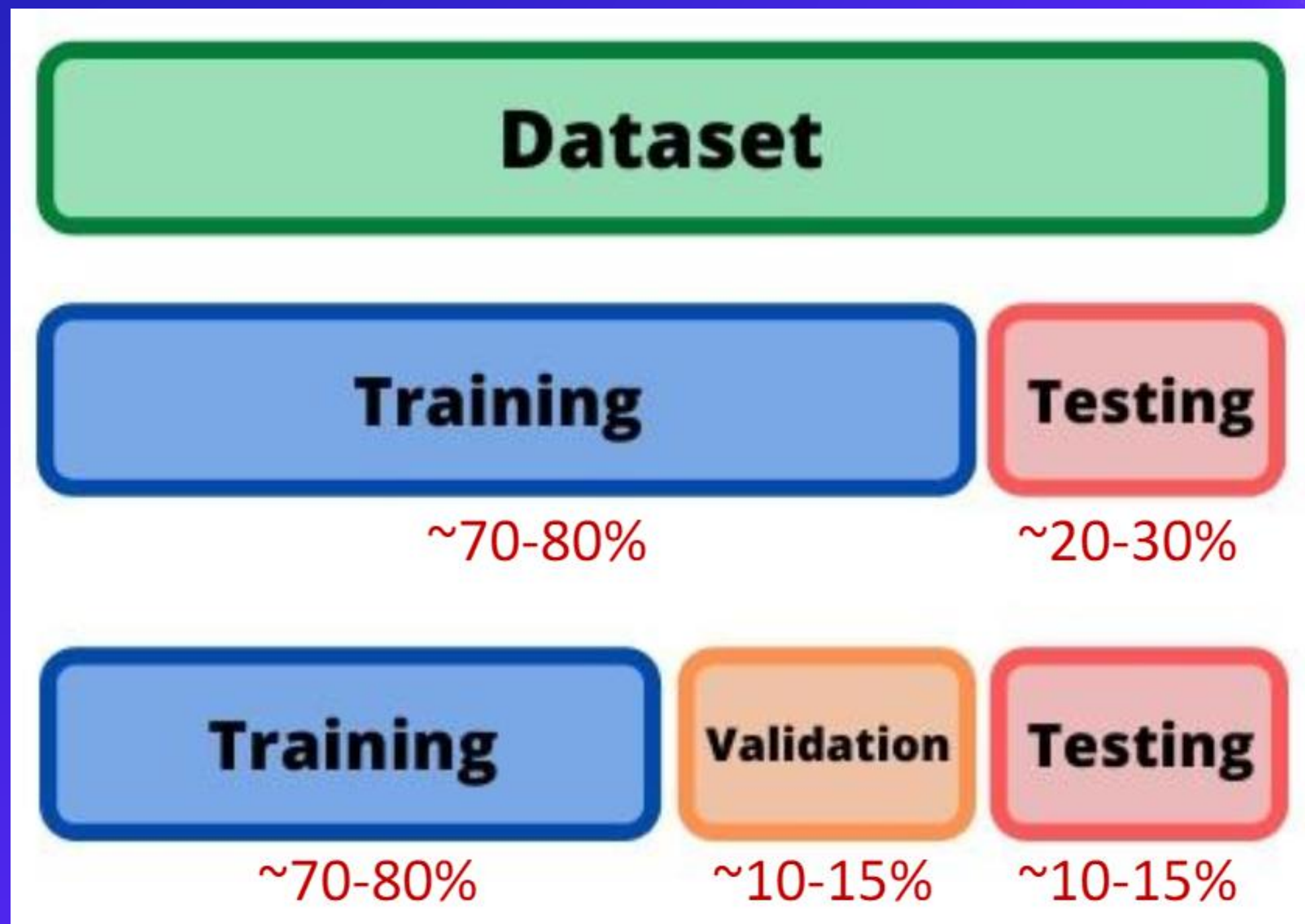
Ear Shape	Face shape	Whiskers	Cat
Pointy	Round	Present	1
Oval	Not round	Present	1
Oval	Round	Absent	0
Pointy	Not round	Present	0
Oval	Round	Present	1
Pointy	Round	Absent	1
Floppy	Not round	Absent	0
Oval	Round	Absent	1
Floppy	Round	Absent	0
Floppy	Round	Absent	0

One-pot encoding

Ear Shape	Pointy ears	Floppy ears	Oval ears	Face shape	Whiskers	Cat
Pointy	1	0	0	Round	Present	1
Oval	0	0	1	Not round	Present	1
Oval	0	0	1	Round	Absent	0
Pointy	1	0	0	Not round	Present	0
Oval	0	0	1	Round	Present	1
Pointy	1	0	0	Round	Absent	1
Floppy	0	1	0	Not round	Absent	0
Oval	0	0	1	Round	Absent	1
Floppy	0	1	0	Round	Absent	0
Floppy	0	1	0	Round	Absent	0



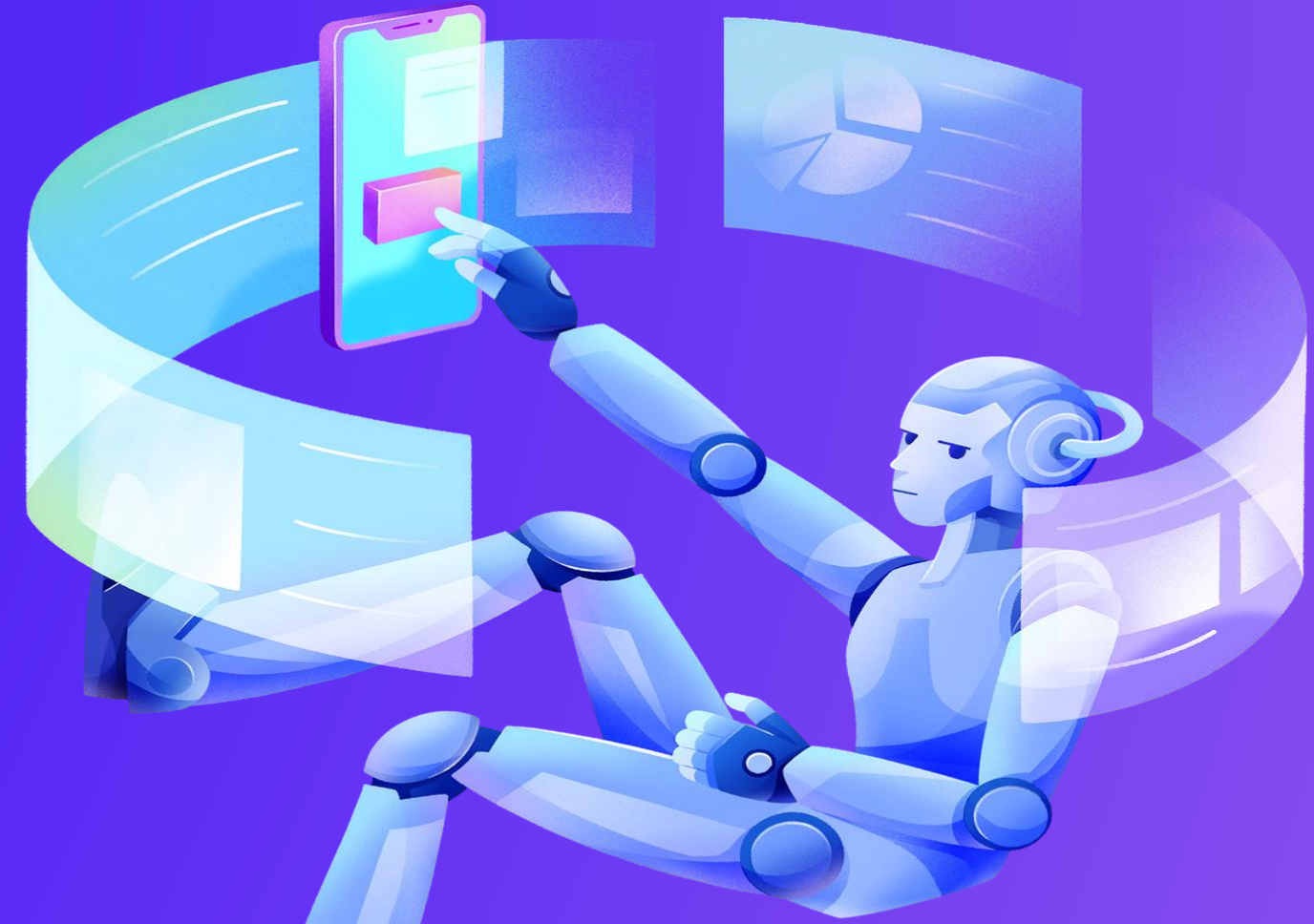
Data splitting



Training set for learning model
Testing set for test the model
Validation set for tuning the model

Missing value

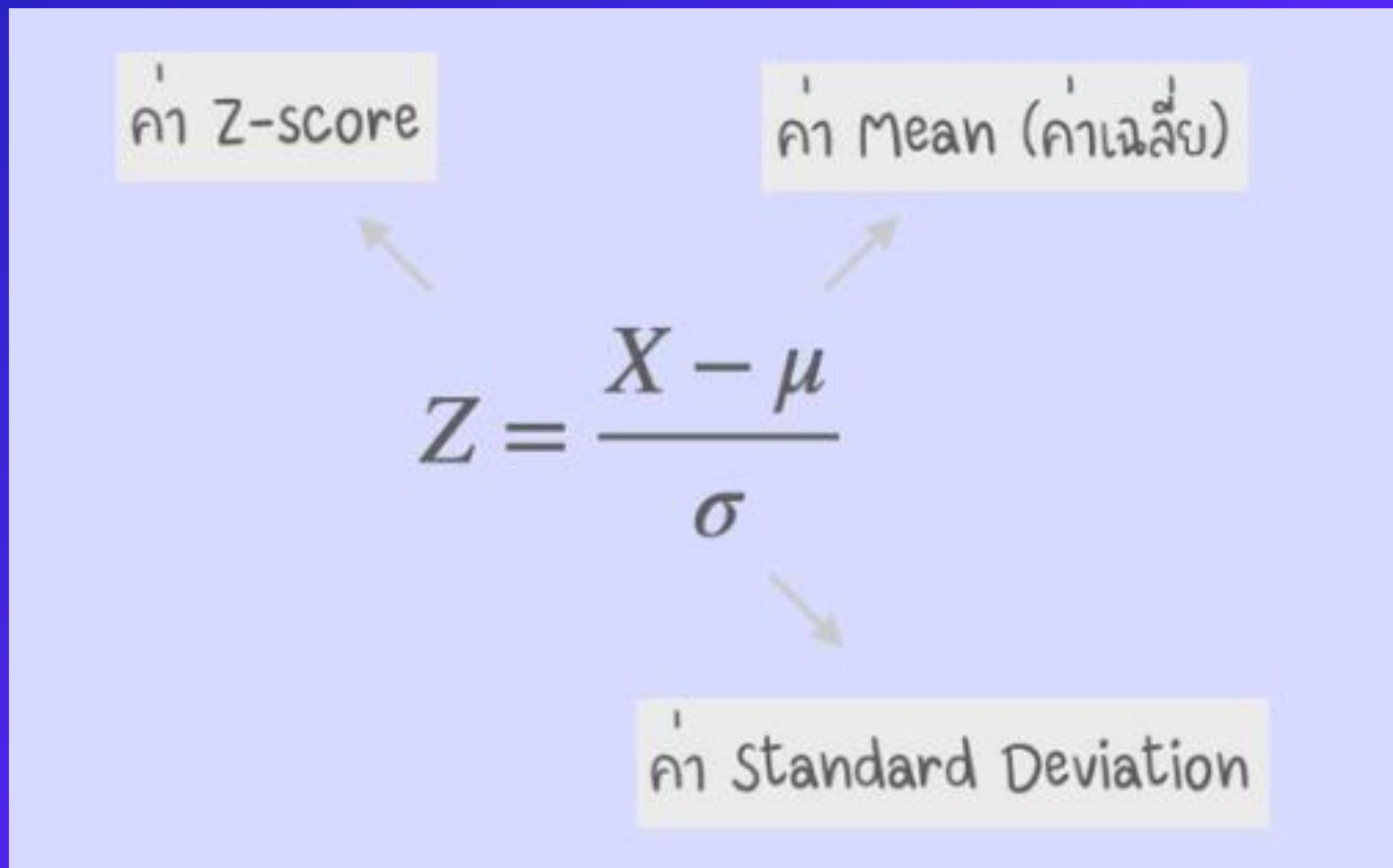
1. Removing rows
2. Predicting
3. Replace



CO2 Emissions (g/k	Make	Model	Engine Size (L)	Cylinders	Fuel Type	Fuel Consumption (Fuel Consumption (Fuel Consumption Comb (L/100 km)
205	AUDI	Q5 HYBRID	2	4	Z	9.8	7.9	8.9
209	AUDI	Q5 HYBRID QUATT	2	4	Z	9.8	7.9	8.9
244	AUDI	Q5 QUATTRO	2	4	Z	12	8.5	10.4
267	AUDI	Q5 QUATTRO	3	6	Z	13.2	9.2	11.4
244	AUDI	Q5 QUATTRO TDI	3	6	D	10.2	7.8	9
244	AUDI	Q5 TDI (modified)	3	6	D			
244	AUDI	Q5 TDI (modified)	3	6	D	10.2	7.8	9
253	AUDI	Q7	2	4	Z	11.9	9.6	10.8
258	AUDI	Q7	2	4	Z	12.2	9.5	11
258	AUDI	Q7	2	4	Z	12.2	9.5	11
304	AUDI	Q7	3	6	Z			
306	AUDI	Q7	3	6	Z	15.2	11	13.3
260	AUDI	Q7	3	6	Z	12.6	9.4	11.1
260	AUDI	Q7	3	6	Z	12.6	9.4	11.1
260	AUDI	Q7	3	6	Z	12.6	9.4	11.1
296	AUDI	Q7	3	6	Z	13.8	11.4	12.7
290	AUDI	Q7 TDI (modified)	3	6	D			

Data normalization

Z-score normalization



The diagram illustrates the Z-score normalization formula. It features a central equation $Z = \frac{X - \mu}{\sigma}$ with three arrows pointing to labels in Thai: 'ค่า Z-score' (Z-score value) for Z, 'ค่า Mean (ค่าเฉลี่ย)' (Mean value) for μ , and 'ค่า Standard Deviation' (Standard Deviation value) for σ .

$$Z = \frac{X - \mu}{\sigma}$$

ค่า Z-score

ค่า Mean (ค่าเฉลี่ย)

ค่า Standard Deviation



Data normalization

Min-Max scaling

ค่าที่ทำการ Scaled

ค่า Original

ค่า Min ใน Feature

$$X_i^{Scaled} = \frac{X_i - \min(X)}{\max(X) - \min(X)}$$

ค่า Max ใน Feature

ค่า Min ใน Feature



THANK YOU!

