12/01

Encoding:

- One Hot Encoding
- Label Encoding

OHE :- → object (string)

Country	
Spain	0
Germany	1
France	2
Spain	
France	

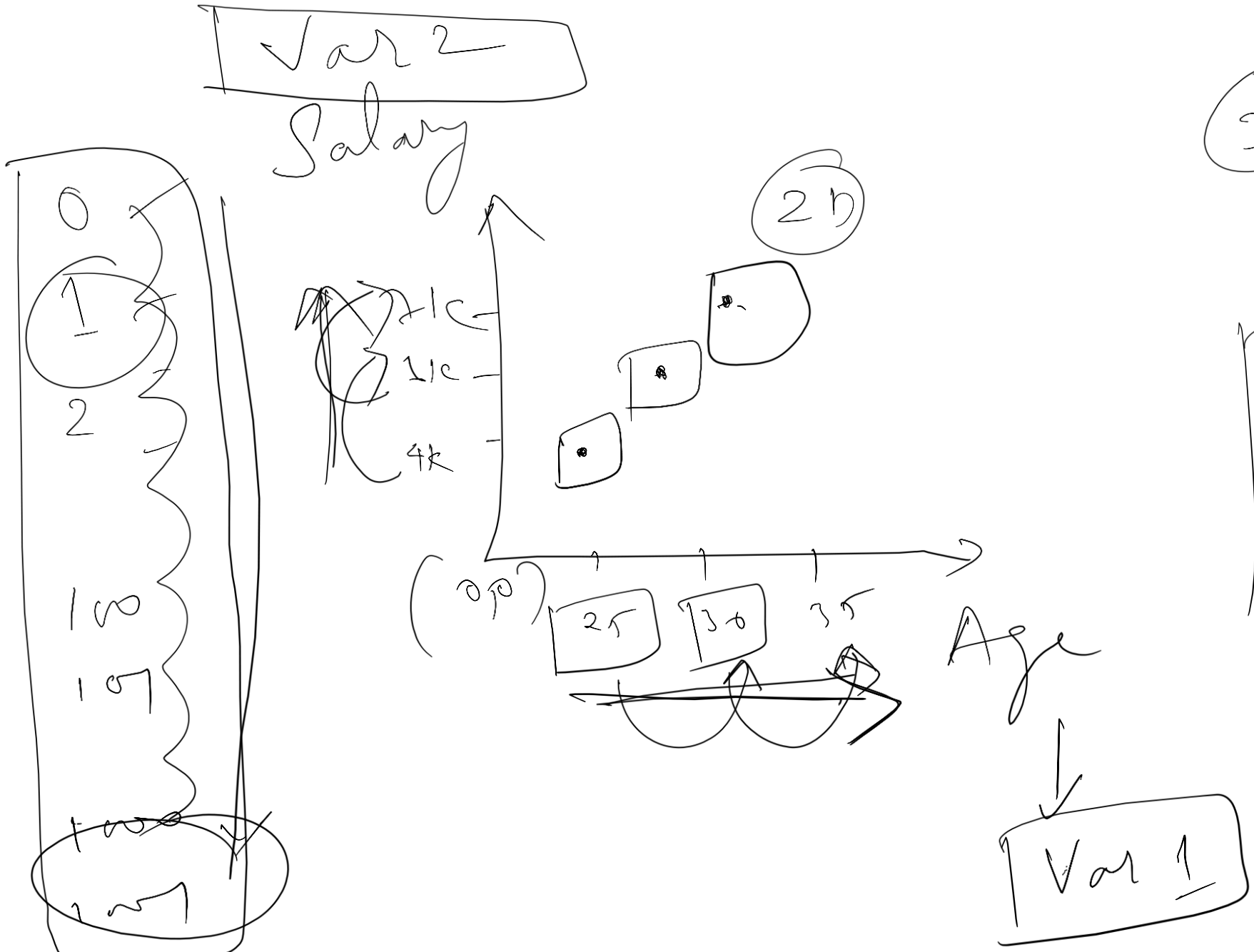
Age → int ✓

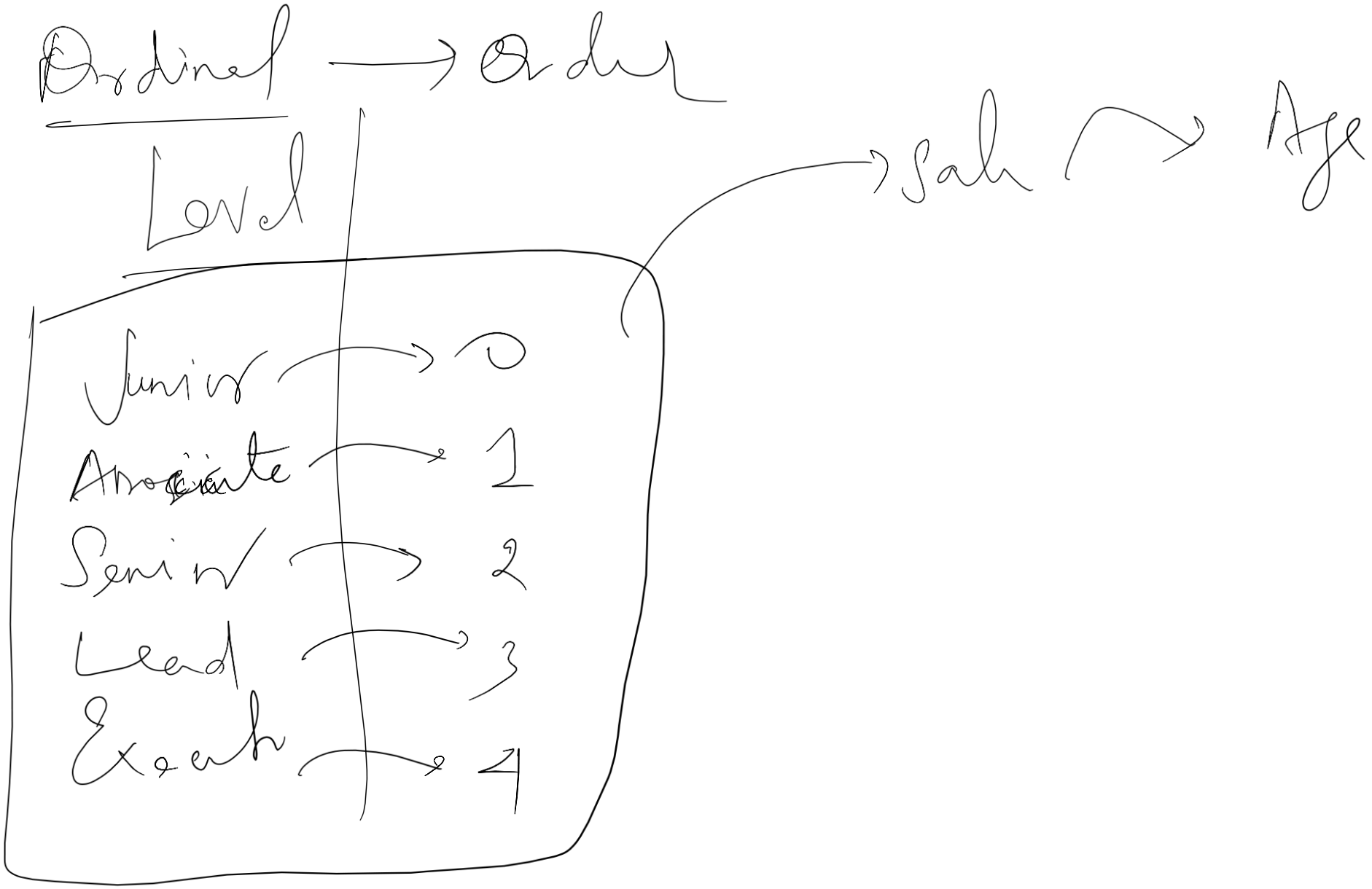
Salary → int ✓

S → 0

G → 1

F → 2


Var 2
Salary
2D
3D
Age
Sal
25
30
35
Age
Var 1


Ordinal
Order
Level
Sal
Age
Junior
0
Associate
1
Senior
2
Lead
3
Exeuth
4

(OHE)

	① Country_France	② Country_Germany	③ Country_Spain
0	0	0	1
1	0	1	0
2	1	0	0
3	0	0	1
4	1	0	0

Age	Country	Salary
25	S	
30	F	
35	9	

df.shape

(5, 2)

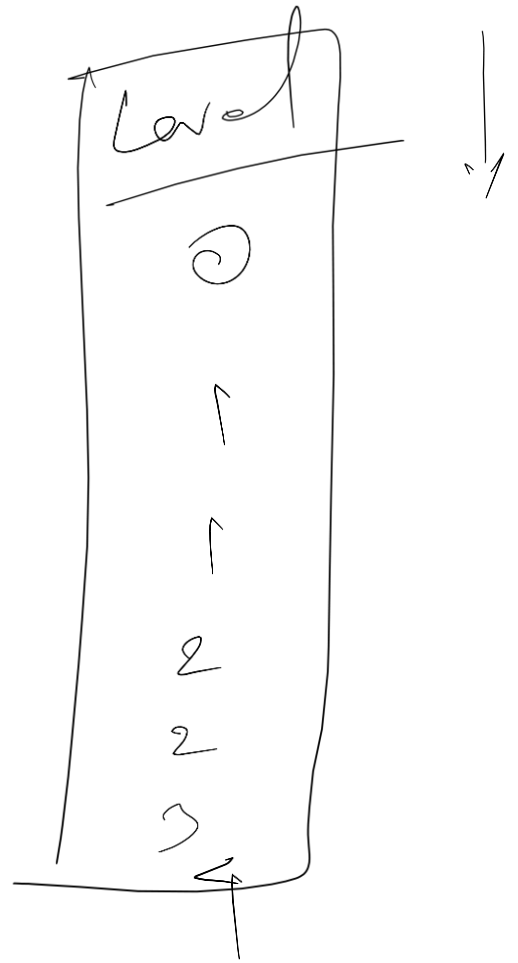
~~3 × 2~~

3 × (3, 2)

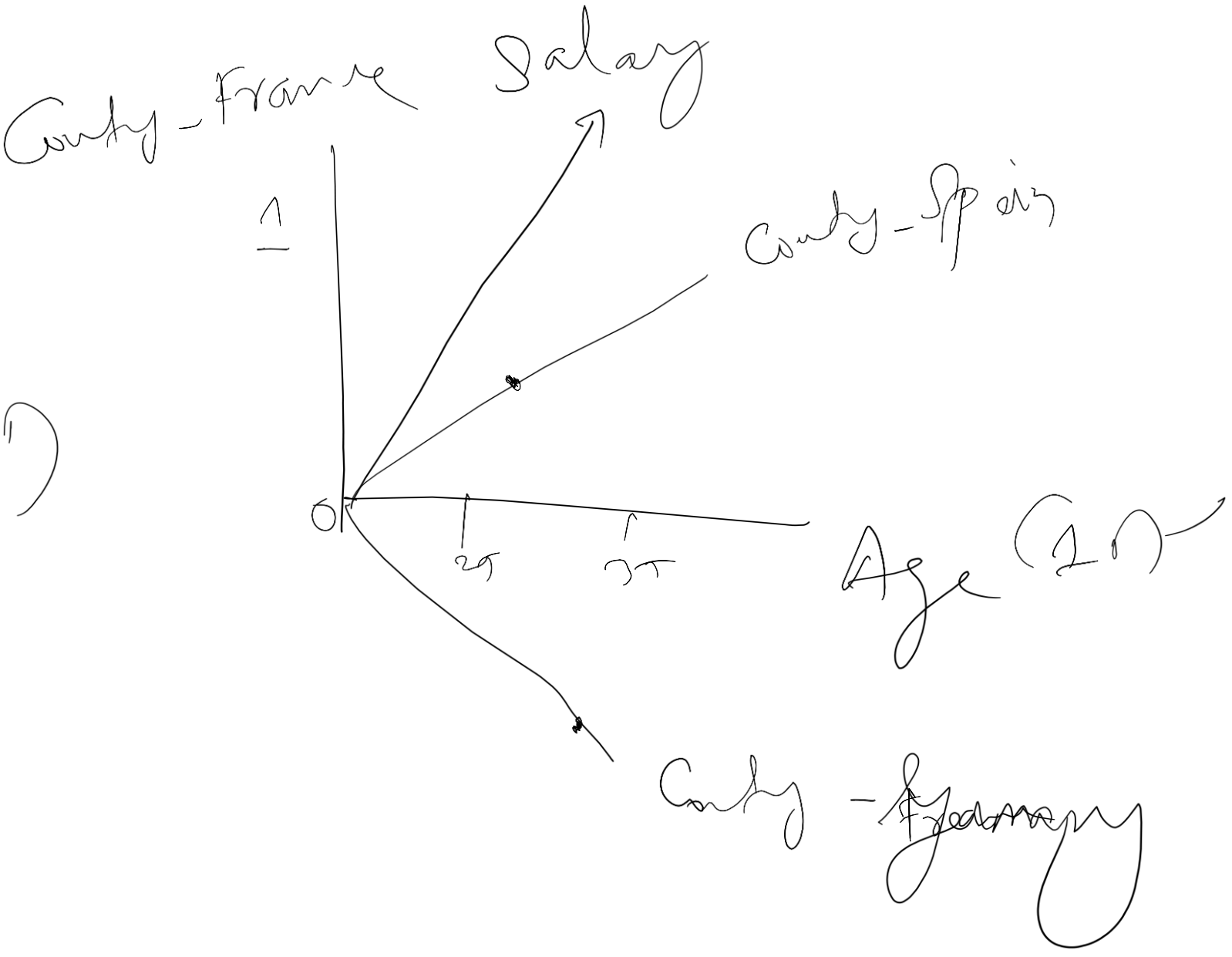
①	②	③	④

(5, 4)

Label Encoding :

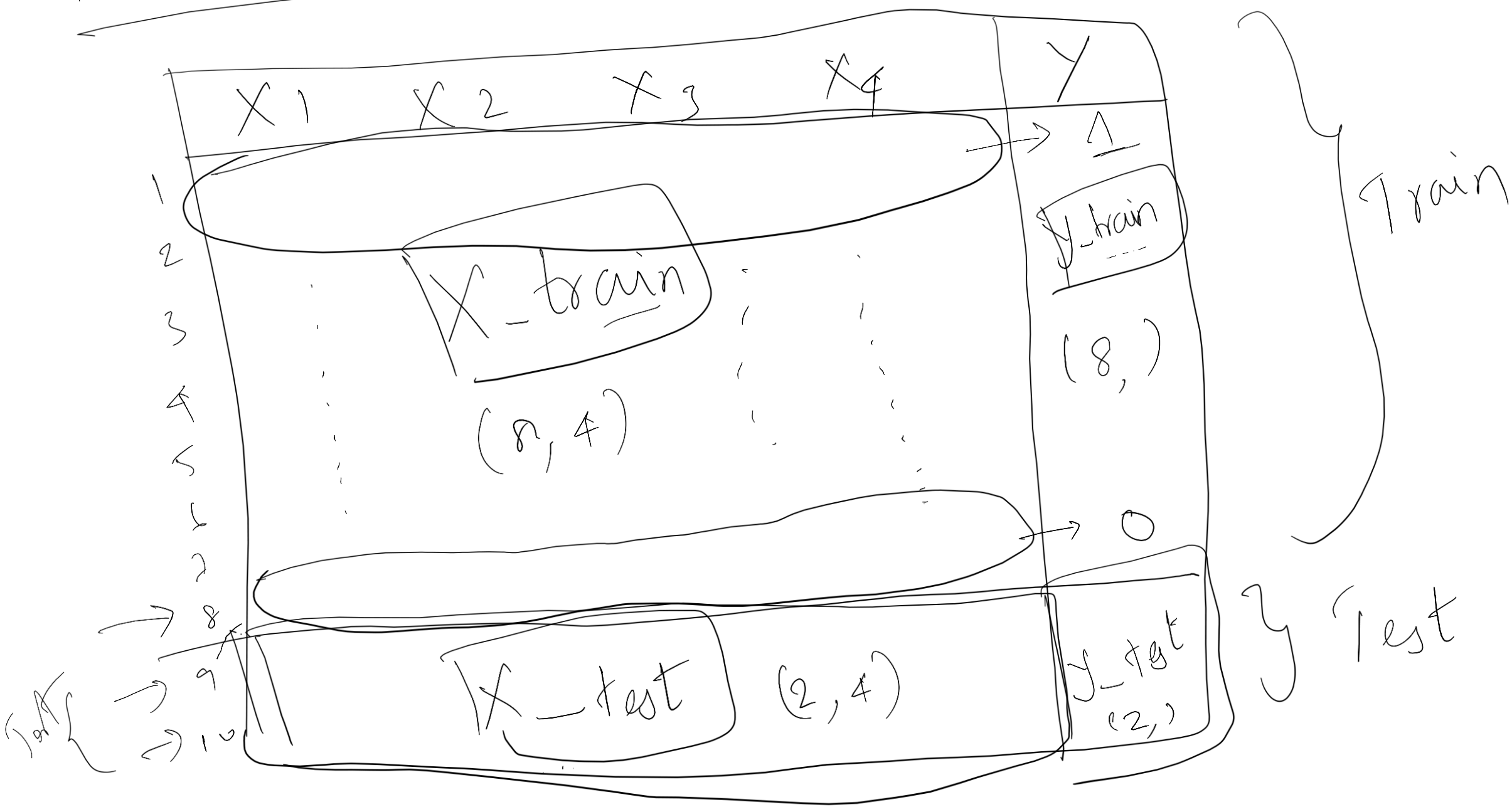


(4)



Salary
Country - France
1
Country - Spain
0
2π
3π
Age (10)
Country - Germany

train_test_split: (h=10) col = 5 (ind-1)



train_test_split → 0.2 < 0.8 (80%) train
0.2 (20%) test

X → Independent variable

y → Dependent " (Target/Response)

✓ → Done

X_train, y_train → Learning process

X_test, y_test ⇒ prediction process

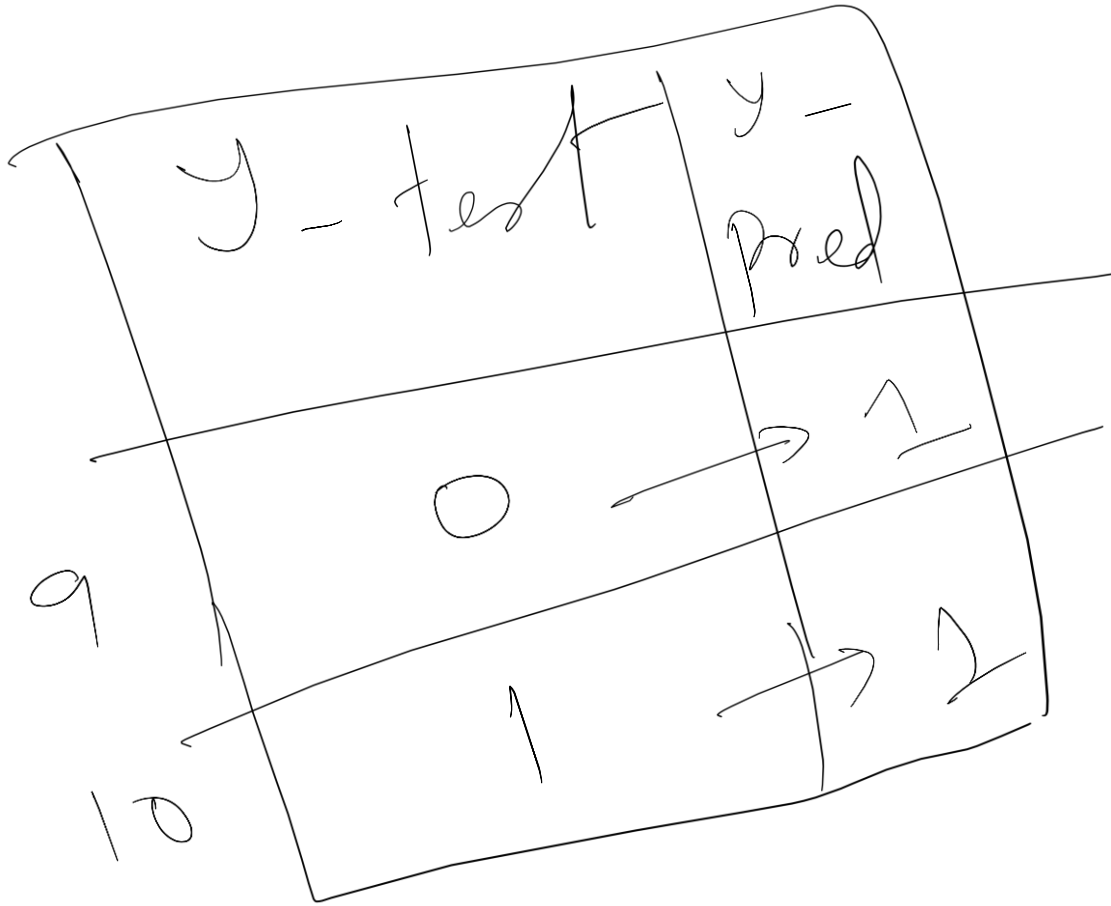
X_train ⟶ y_train Learnt-

Testing / prediction

X_test ⟶ y_test (prediction)

9 → x_1, x_2, x_3, x_4 ⟶ 0 (y-pred.)

10 → x_1, x_2, x_3, x_4 ⟶ 1 (y-pred)


y - test
y - Pred
0 → 1
1 → 2
9
10

100%
50%

Feature Scaling :

→ Normalization → MinMax Scaling

→ Standardisation → Z-score Standardization

Min Max ⟶ (0, 1)

$$= \frac{X_i - \text{Min}(x)}{\text{Max}(x) - \text{Min}(x)}$$

Min → 1

Max → 5

✓ 1 ⟶ 0

✓ 3 ⟶ 0.5

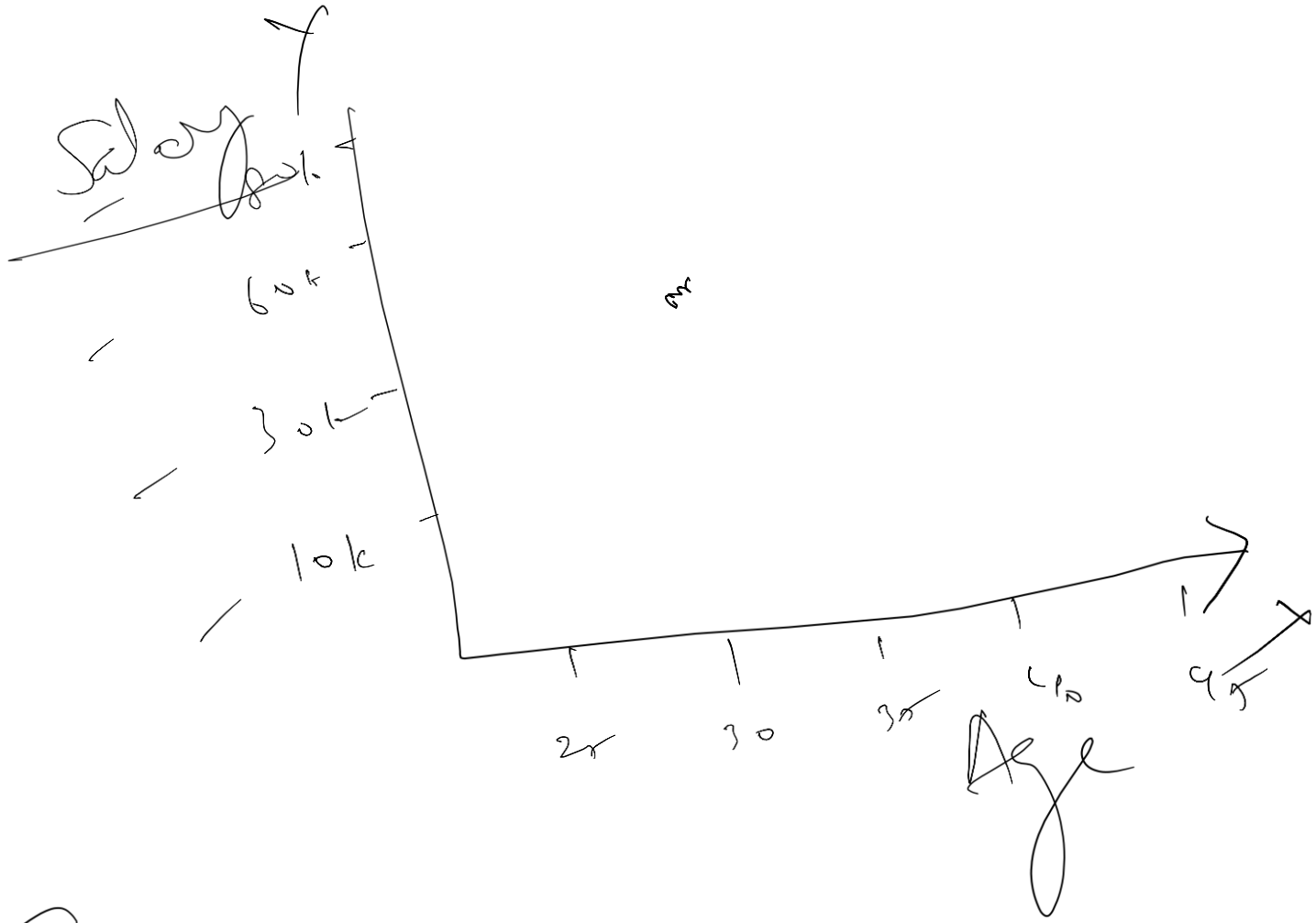
✓ 5 ⟶ 1

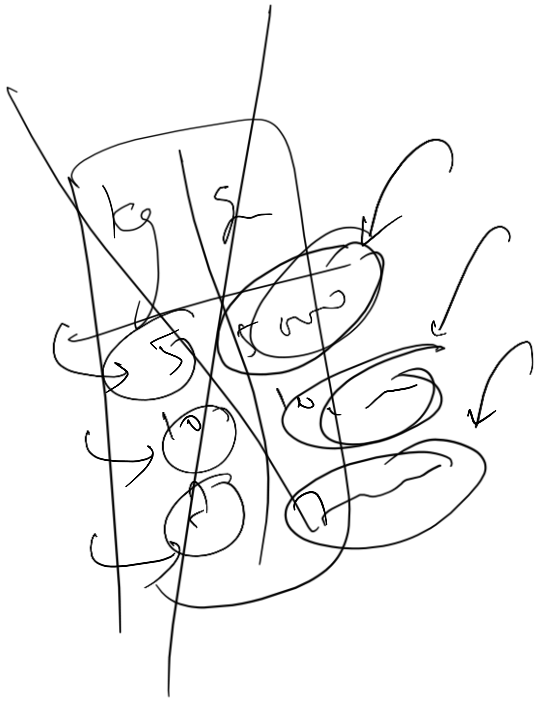
1 = 1 − 1 = 0

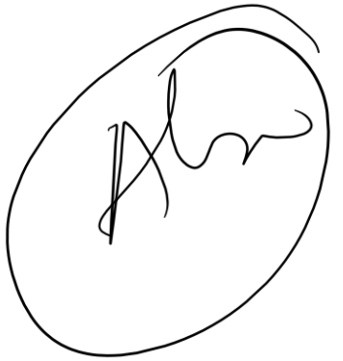
2 = $\frac{3-1}{5-1} = \frac{2}{4} =$ 0.5

3 = $\frac{5-1}{4} = \frac{4}{4} =$ 1 ✓

5 ⟶ 1





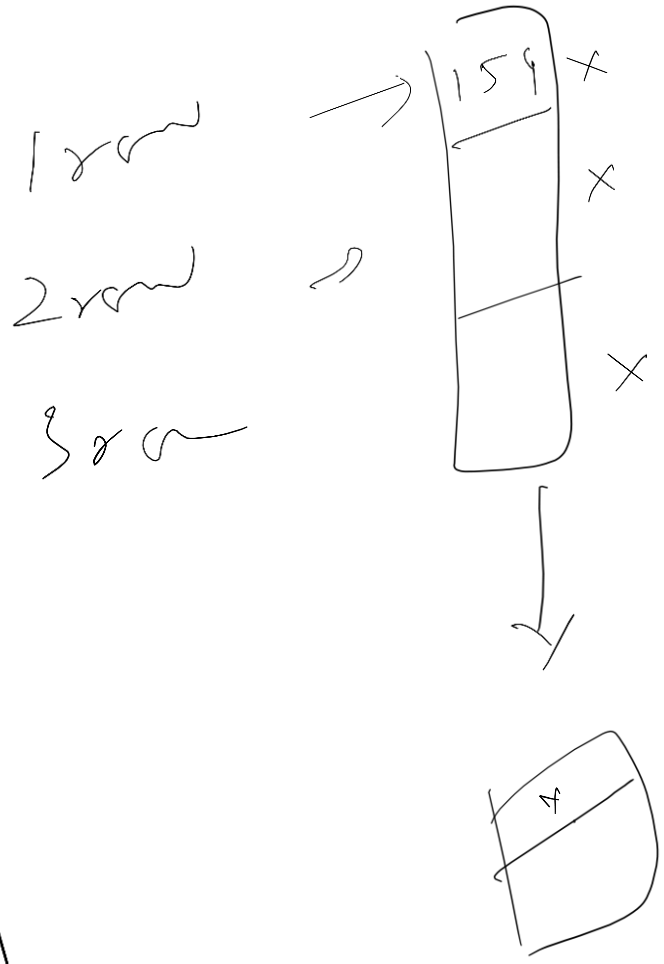


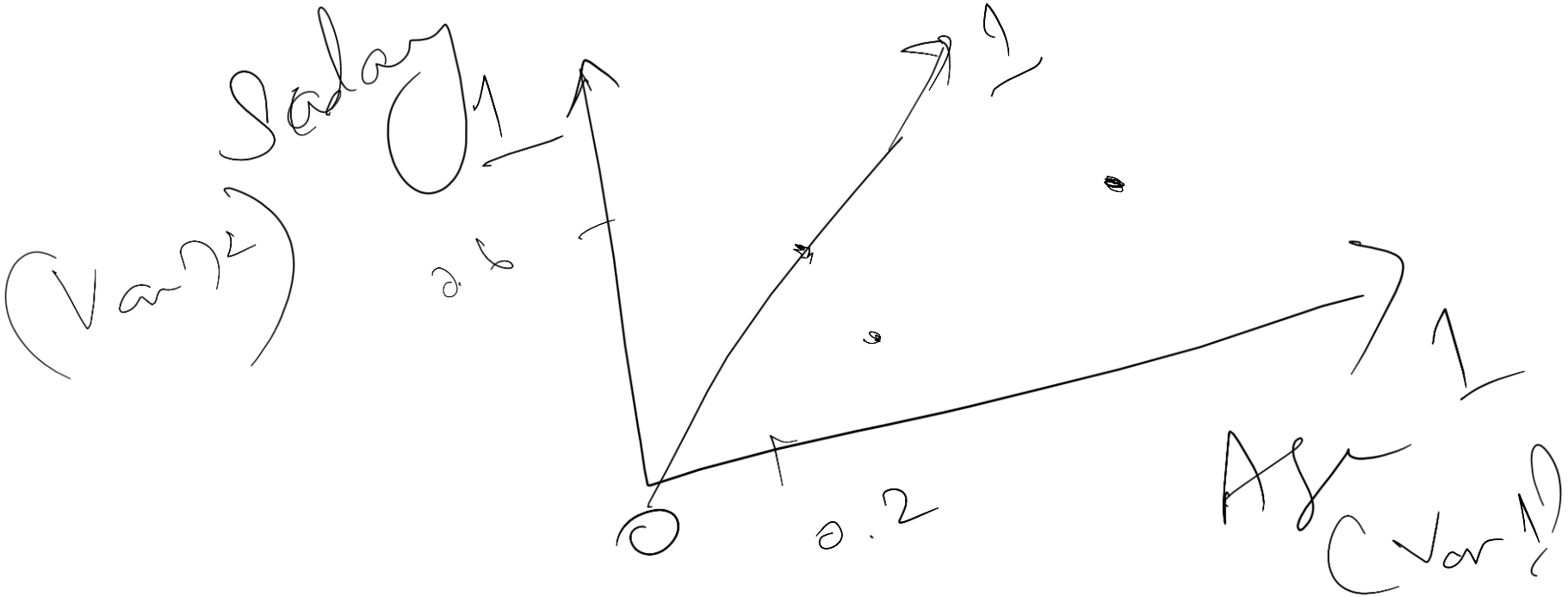
(Euclidean distance)

⟹ Min max scalar

Age ⟶ (0, 1)

Sal ⟶ (0, 1)



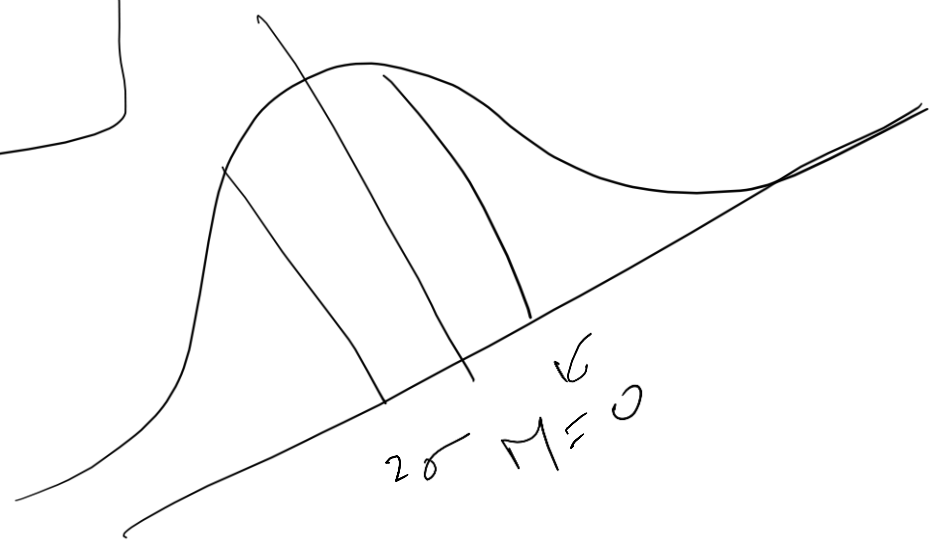


Normal dist → Standard Normal dist.

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

$\mu = 0$

$\sigma = 1$



$1\sigma \to 68\%$

$2\sigma \to 95\%$

$3\sigma \to 99.7\%$

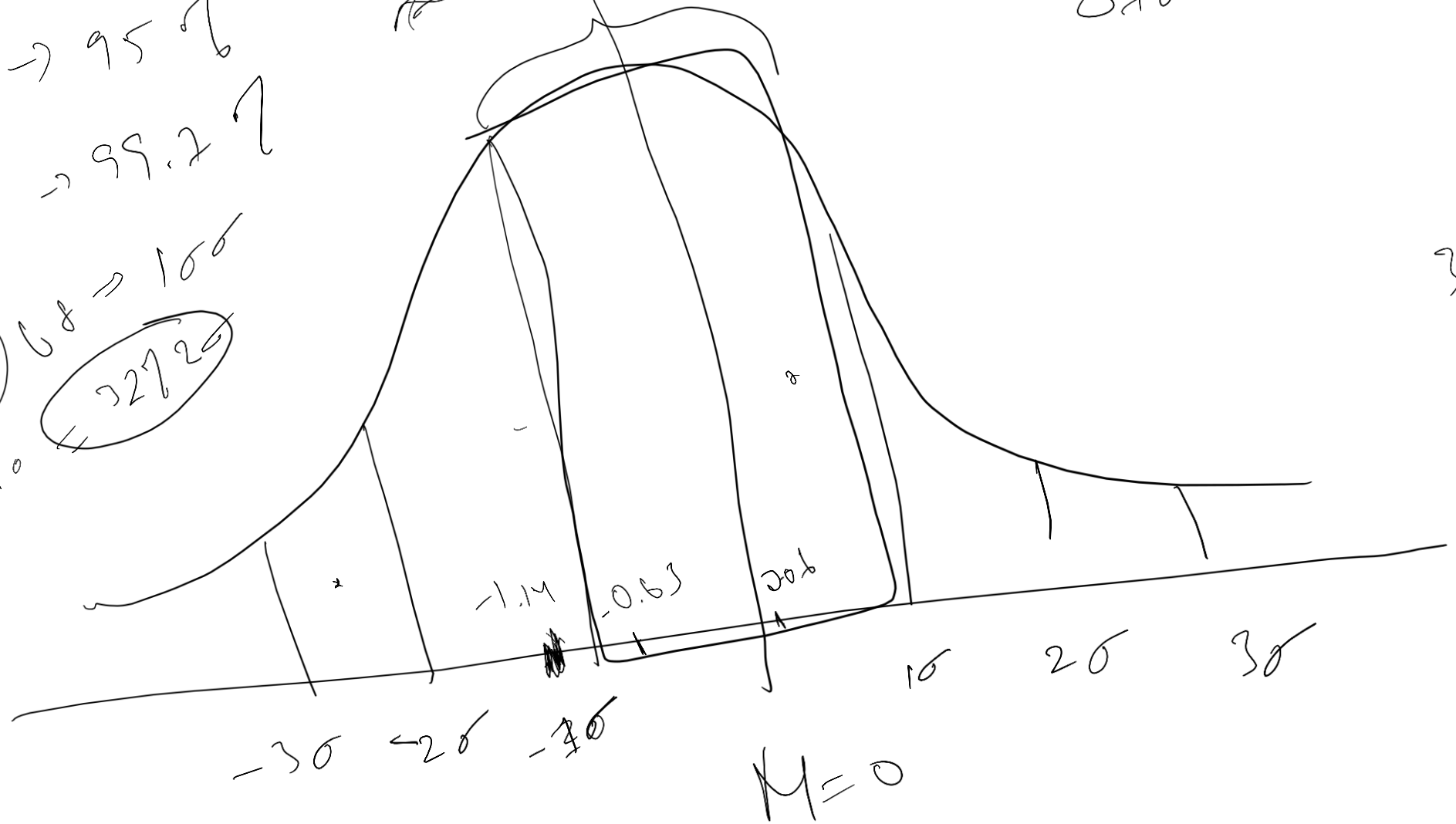
$10 \quad 68 \to 100$

$90 \quad 32\% \; 2\sigma$

$\approx 68\%$

Std. normal dist

$4 \to$

$3 \to 1\sigma$

$\to 73\%$