12/01

Encoding:

- One Hot Encoding
- Label Encoding

OHE :- → object (string)

Country	
Spain	0
Germany	1
France	2
Spain	
France	

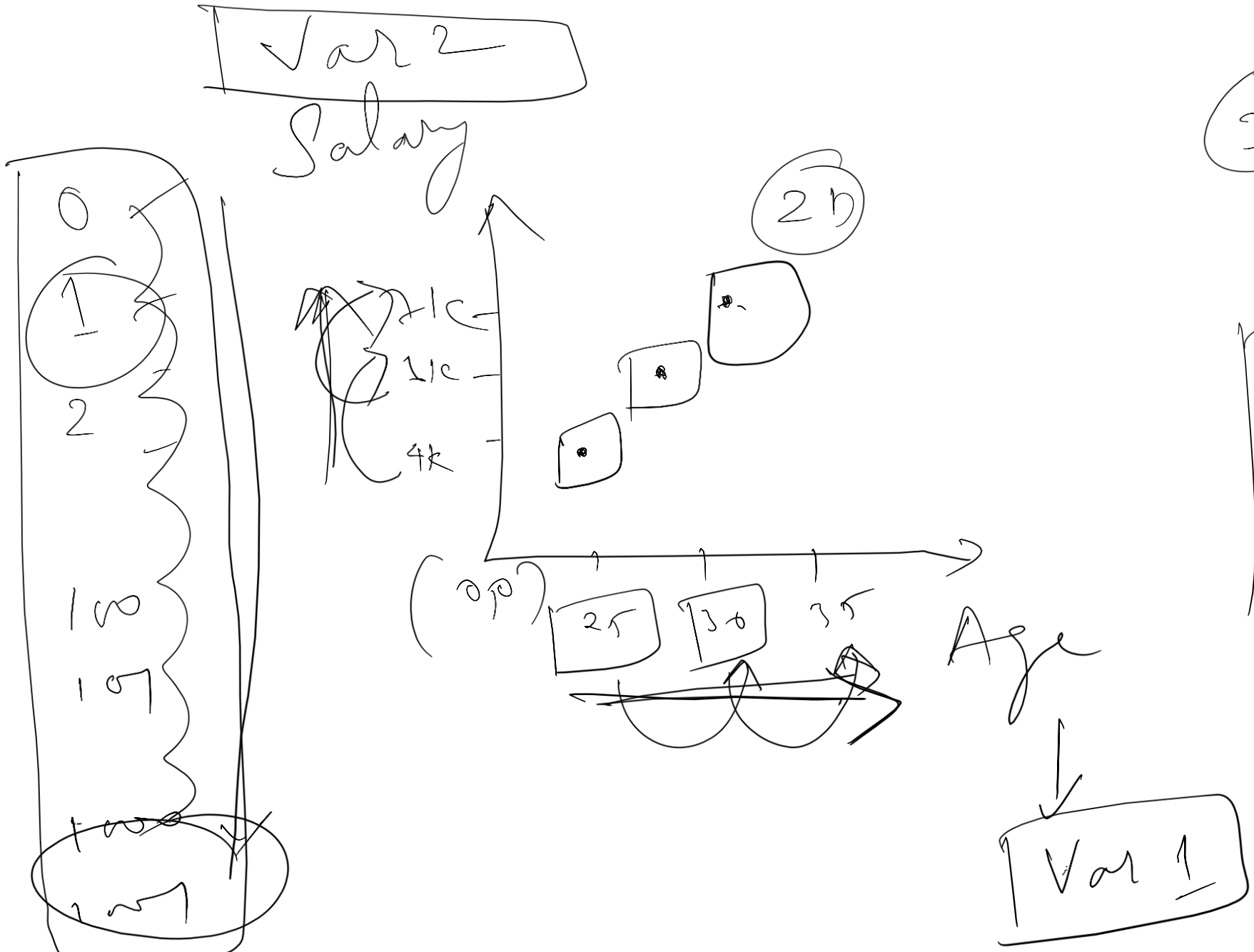
Age → int ✓

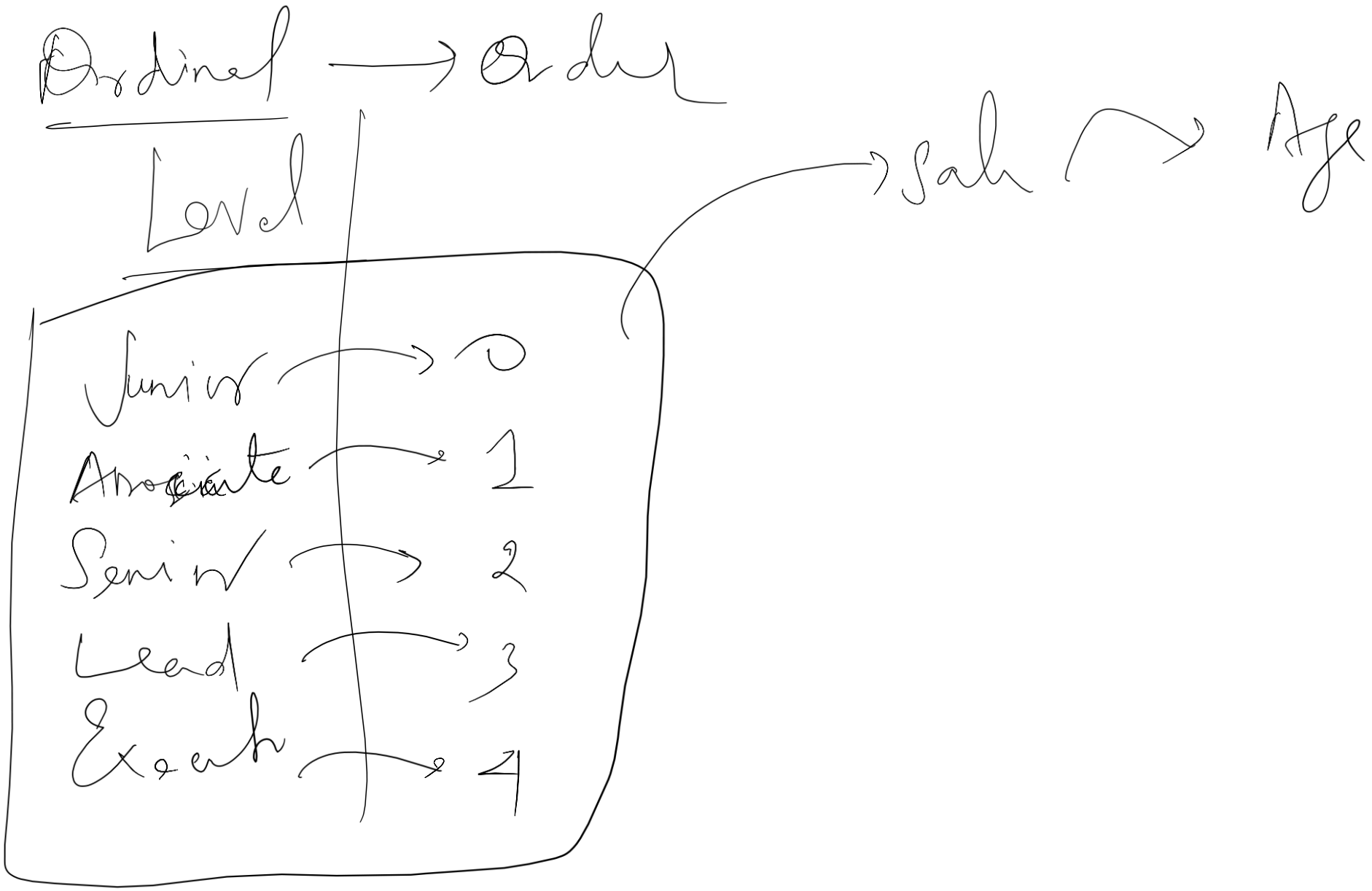
Salary → int ✓

S → 0

G → 1

F → 2


Var 2
Salary
2D
Age
Var 1
3D


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

(OHE)

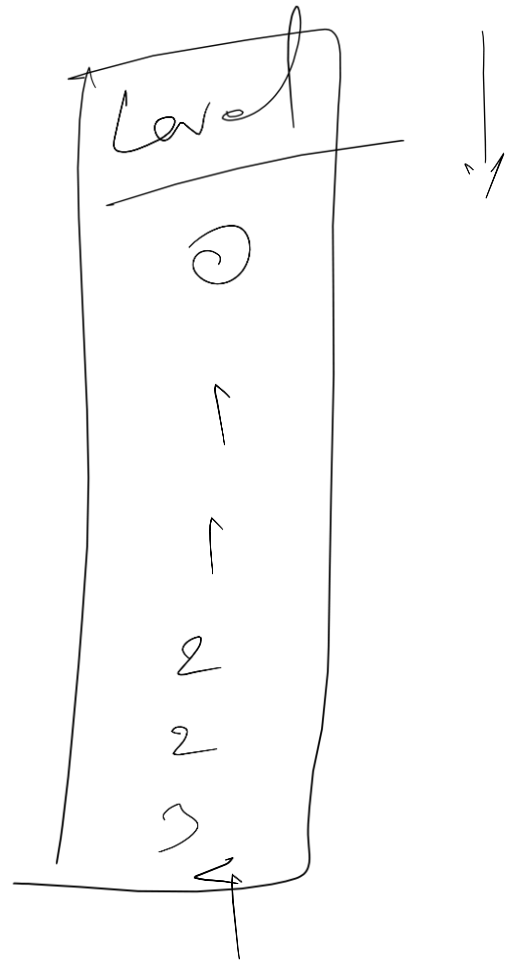
	Country_France	Country_Germany	Country_Spain
0	0	0	1
1	0	1	0
2	1	0	0
3	0	0	1
4	1	0	0

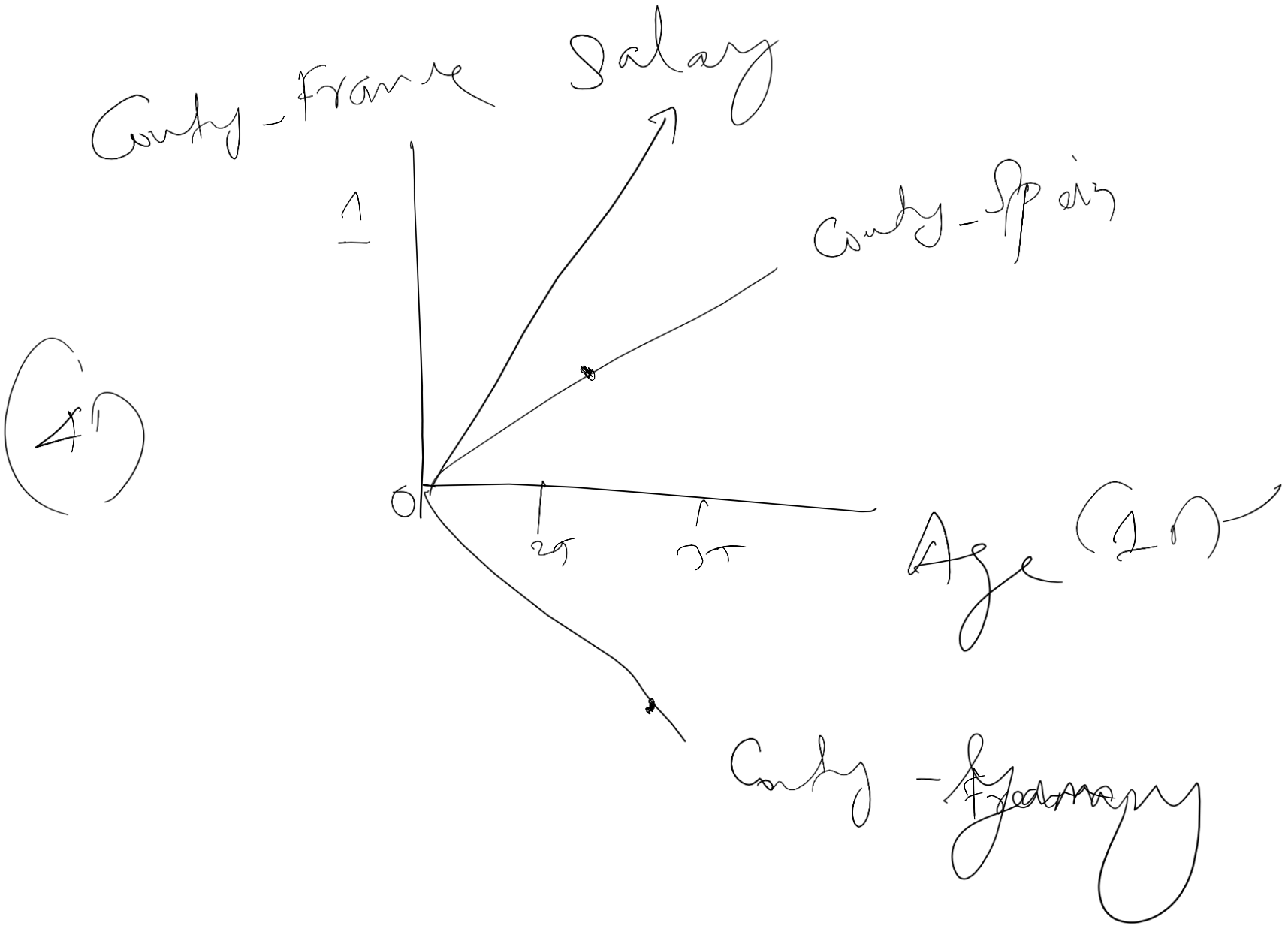
Age	Country_Italy
25	0
37	1
29	0

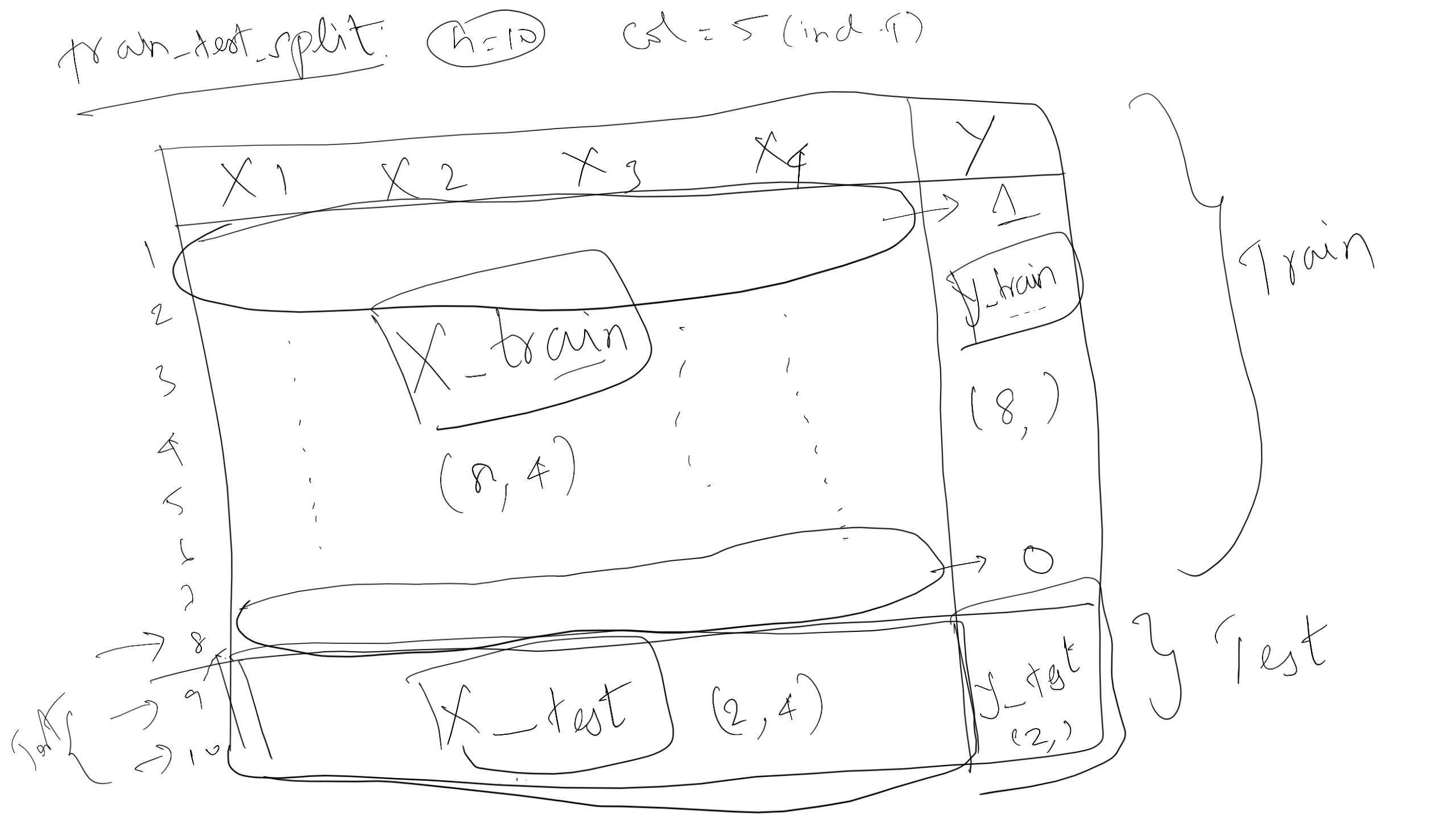
0	0	0	1
0	1	0	0
0	0	1	0
1	0	0	0
0	0	0	1

(5, 4)

Label Encoding :




Salary
Country - France
Country - Spain
Country - Germany
Age
0
25
35


train_test_split
test_size = 0.2
random_state = 42
X1
X2
X3
X4
Y
1
X_train
(8, 4)
y_train
(8,)
Train
0
X_test
(2, 4)
y_test
(2,)
Test

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

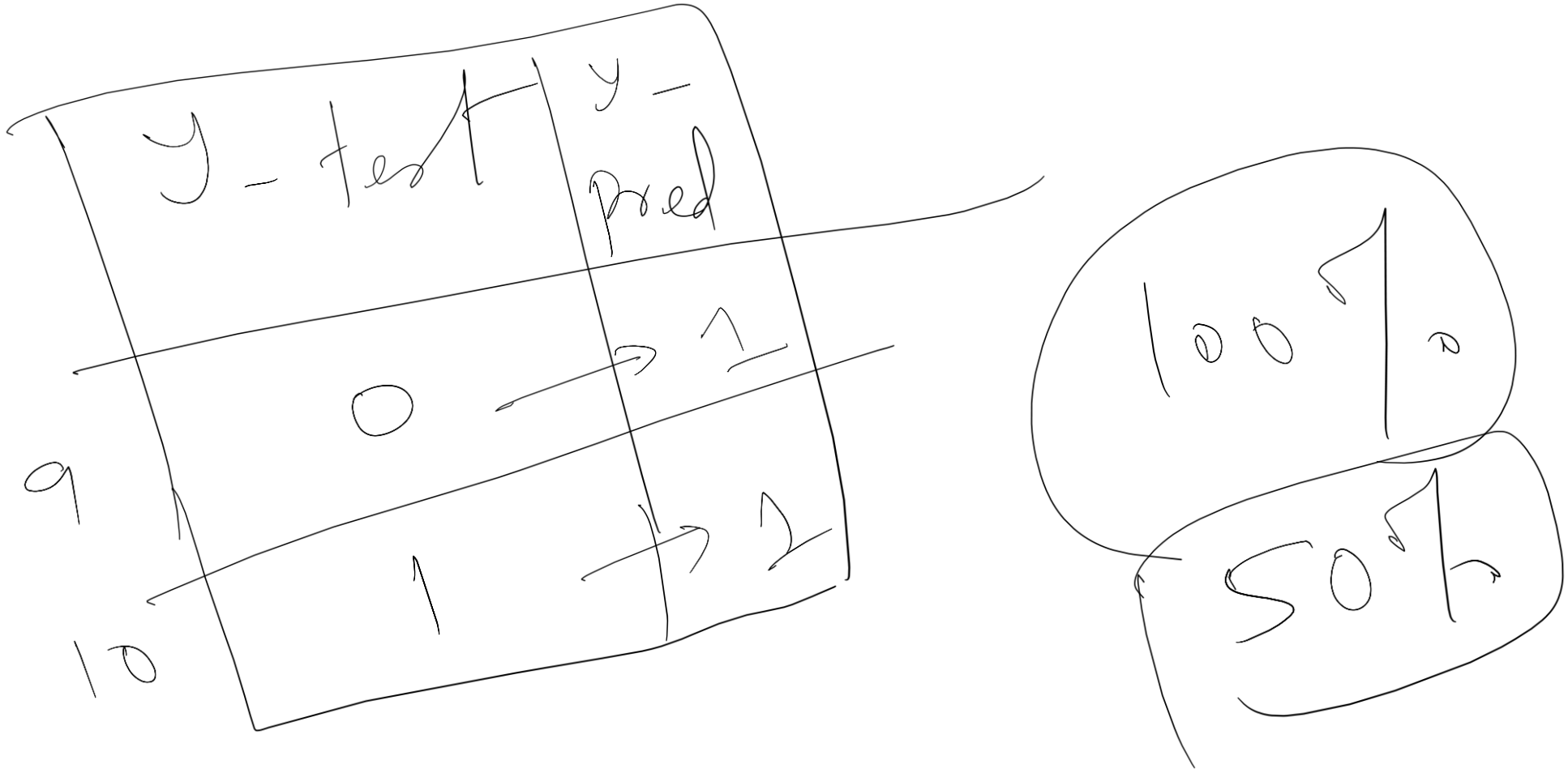
X_train ⟶ y_train Learnt

Testing / prediction

X_test ⟶ y_test (prediction)

9 → x_1, x_2, x_3, x_7 ⟶ 0 (y-pred.)

16 → x_1, x_2, x_3, x_4 ⟶ 1 (y-pred)


y - test
y - Pred
0 → 1
1 → 1
9
10
100%
50%

Feature Scaling :

→ Normalization → MinMax Scaling

→ Standardisation → Z-score Standardization

Min Max → (0, 1)

Min → 1

Max → 5

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

✓ 1 → 0

✓ 3 → 0.5

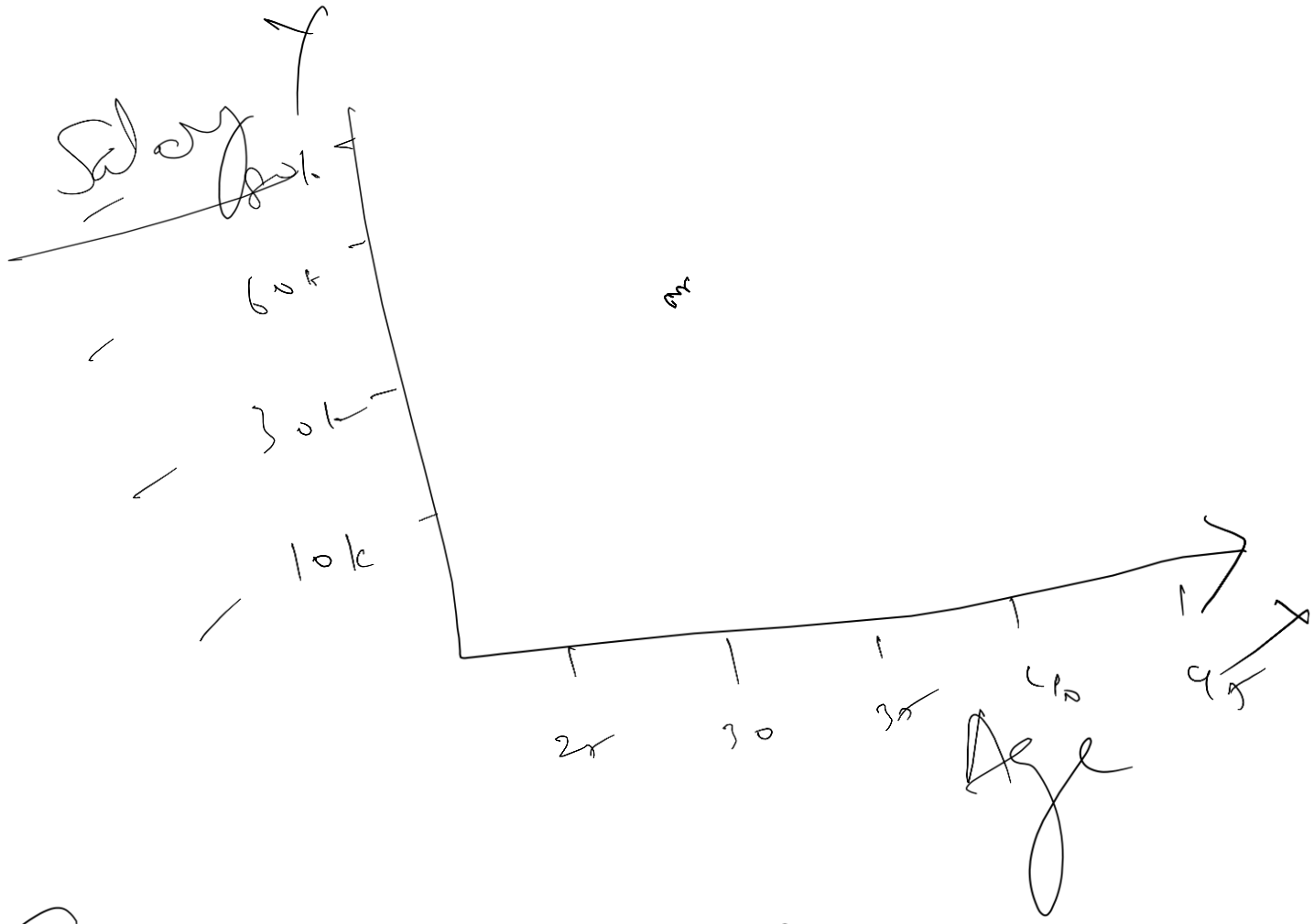
✓ 5 → 1

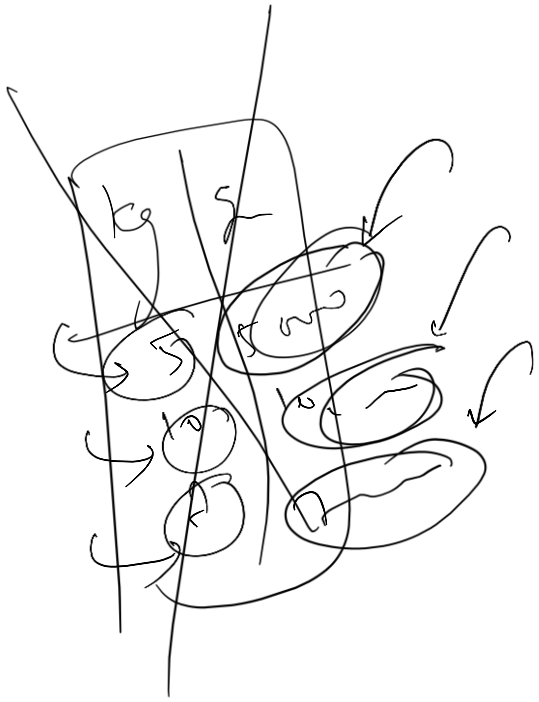
1 = 1 − 1 = 0

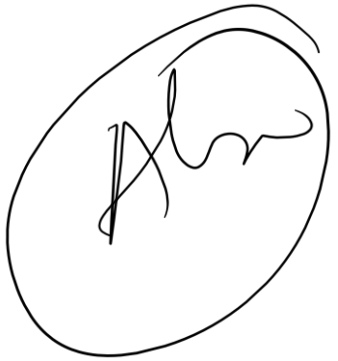
2 = $\frac{3-1}{5-1} = \frac{2}{4} = 0.5$

5 → 1

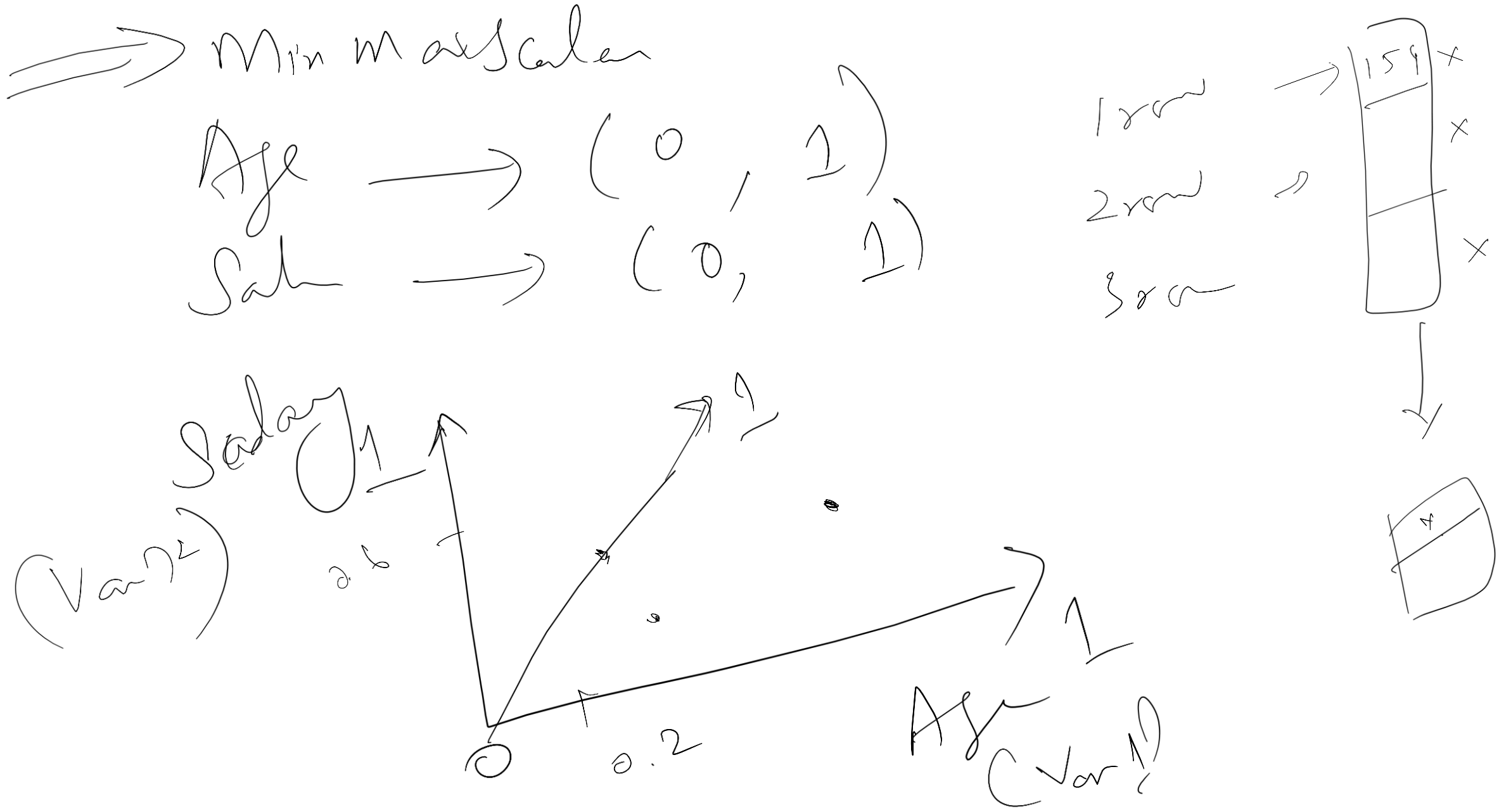
3 = $\frac{5-1}{4} = \frac{4}{4} = 1$ ✓







(Euclidean distance)

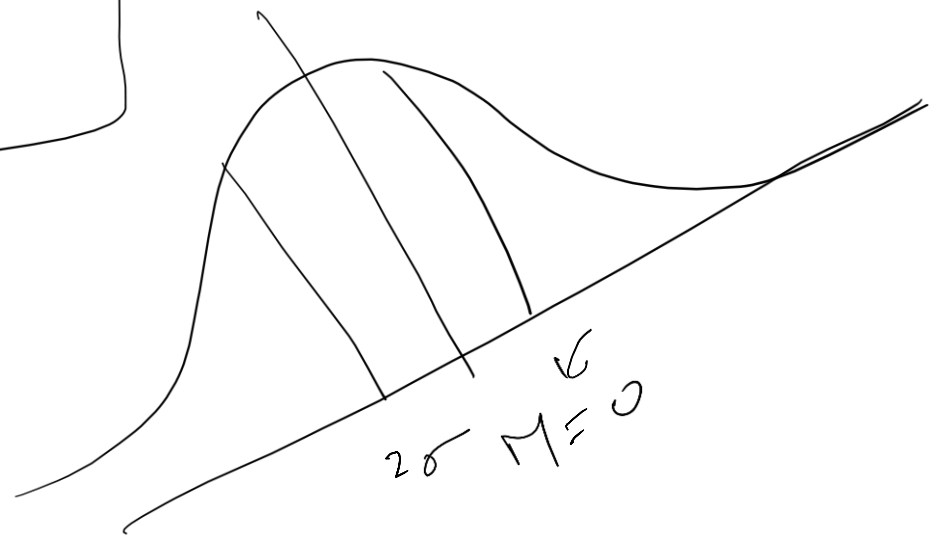

Min Max scaler
Age → (0, 1)
Sal → (0, 1)
1 row
2 row
3 row
159
X
X
X
Salary
1
0.6
0.2
0
1
1
Age
(Var1)
(Var2)

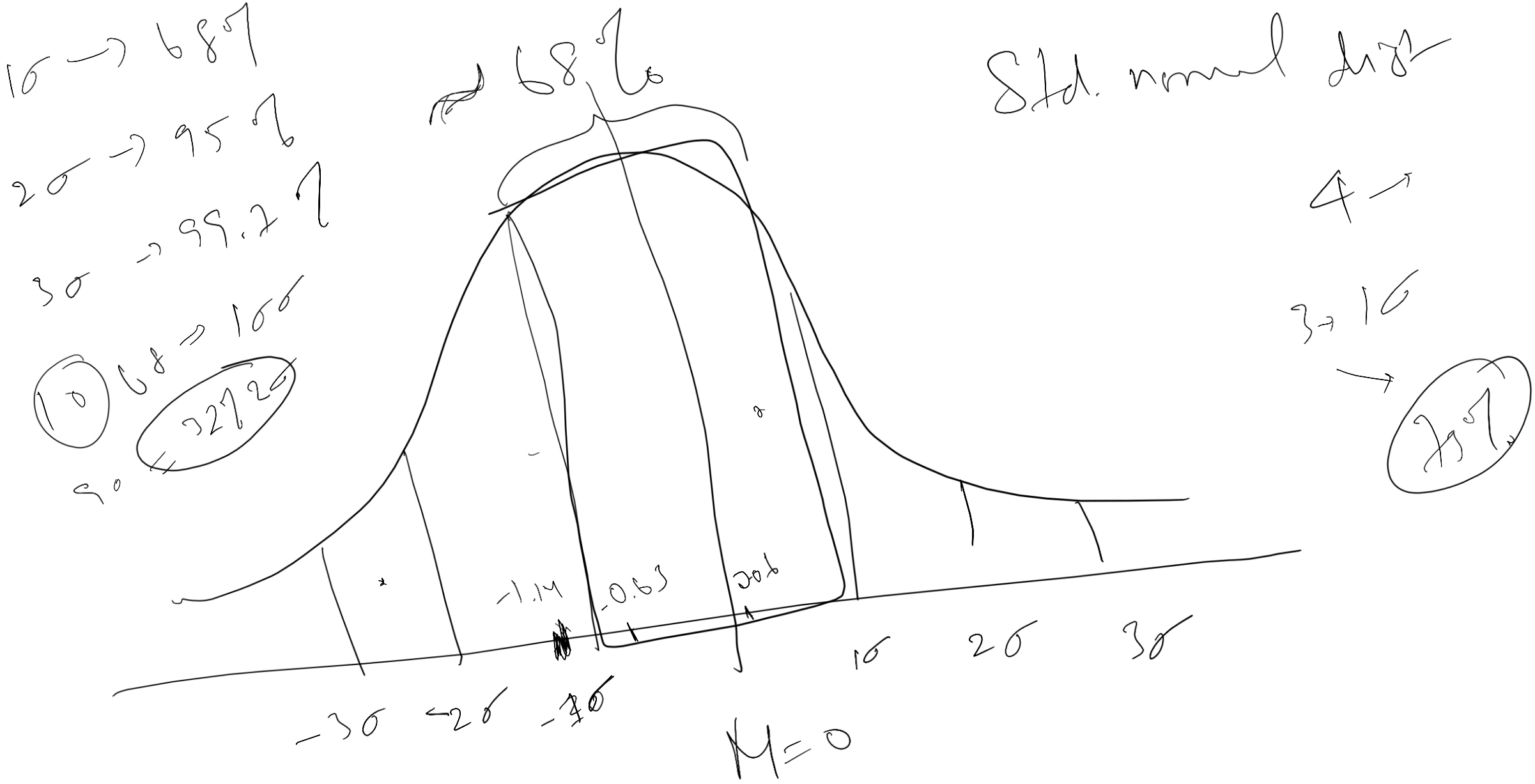
Normal dist → Standard Normal dist.

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

$\mu = 0$

$\sigma = 1$




1σ → 68%
2σ → 95%
3σ → 99.7%
≈ 68%
Std. normal dist
4 →
3 → 1σ
→ 99%
(1) 68 → 1σ
32%
50
-1.14
-0.63
30%
1σ
2σ
3σ
-3σ
-2σ
-1σ
μ=0