Health-insurance risk dataset with **3 numeric columns and 1 categorical**, and the target **Illness (Yes/No)**.

- Illness risk tends to increase with **age**
- Smoking is a strong risk factor
- Income varies but is not always directly tied to illness (realistic noise)
- Gender is included as a demographic feature (often present in real datasets)

Example-1

Initial weights: $w_i = 1/5 = 0.2$

Row No.	Gender	Age	Income (₹/year)	Smoking (cigs/day)	Illness	Sample Weights
1	Male	38	420000	0	No	1/5
2	Female	52	360000	5	Yes	1/5
3	Male	45	780000	0	No	1/5
4	Female	29	300000	12	Yes	1/5
5	Male	61	500000	8	Yes	1/5

Data and Encoding

Row	Gender	Age	Income	Smoking	Illness	y
1	M	38	420000	0	No	−1
2	F	52	360000	5	Yes	+1
3	M	45	780000	0	No	−1
4	F	29	300000	12	Yes	+1
5	M	61	500000	8	Yes	+1

Justification: AdaBoost works with $y \in \{-1, +1\}$so that the weight update formula becomes easy.

Choose the best stump h_1

AdaBoost picks the weak learner with **minimum weighted error**.

We test a few sensible stumps (because weak learners are simple):

Candidate stump on Smoking (very reasonable medically)

Rule A: "If Smoking ≥ 1 → Yes, else No"

In ± 1form:

$$h(x) = \begin{cases} +1 & \text{if Smoking} \geq 1 \\ -1 & \text{if Smoking} = 0 \end{cases}$$

Check predictions:

Row	Smoking	Y(Actual)	Prediction	Check
1	0	−1	−1	✓
2	5	+1	+1	✓
3	0	−1	−1	✓
4	12	+1	+1	✓
5	8	+1	+1	✓

Zero mistakes.

Weighted error: $\varepsilon_1 = 0$

Justification: Smoking is a very strong risk factor, and in this dataset it perfectly separates classes.

What happens in AdaBoost if $\varepsilon_1 = 0$?

$$\alpha_1 = \frac{1}{2}\ln\left(\frac{1-\varepsilon_1}{\varepsilon_1}\right)$$

This becomes ln (∞)→ **infinite**, meaning the learner is perfect and AdaBoost **stops** (no need to add more learners).

So technically, the "complete" AdaBoost solution ends in **one round**: the stump is already a perfect classifier.

Final Model (after Round 1)

$$H(x) = \text{sign}(\alpha_1 h_1(x)) \Rightarrow H(x) = h_1(x)$$

Final decision

If Smoking (cigs/day) ≥ 1 ⇒ Illness = Yes

If Smoking = 0 ⇒ Illness = No

This predicts all 5 rows correctly.

Data (with noise) + Encoding

Target encoding (AdaBoost standard):

- **Illness = Yes** $\rightarrow y = +1$
- **Illness = No** $\rightarrow y = -1$

Initial weights: $w_i = 1/5 = 0.2$

Row	Gender	Age	Income (₹/yr)	Smoking	Illness	y	w_i
1	M	38	420000	0	No	−1	0.2
2	F	52	360000	5	Yes	+1	0.2
3	M	45	780000	0	**Yes** *(noise)*	+1	0.2
4	F	29	300000	12	Yes	+1	0.2
5	M	61	500000	8	Yes	+1	0.2

Justification (noise): In real life, illness can occur even for a non-smoker (genetics, comorbidities, environment), so Smoking won't be a perfect separator anymore.

AdaBoost setup

For each weak learner $h_t(x)$(a simple rule/stump):

1. **Weighted error** $\varepsilon_t = \sum_{i=1}^{n} w_i \cdot \mathbf{1}(y_i \neq h_t(x_i))$
2. **Model weight** $\alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$
3. **Update sample weights** $w_i \leftarrow w_i \cdot e^{-\alpha_t y_i h_t(x_i)}$

Equivalent intuition:

- Correct ($y_i = h_t$) $\Rightarrow$ multiply by $e^{-\alpha_t}$(decrease)
- Wrong ($y_i \neq h_t$) $\Rightarrow$ multiply by $e^{+\alpha_t}$(increase)

  Then normalize so weights sum to 1.

Justification: Misclassified points become more important so the next stump focuses on them.

ROUND 1 — Choose h_1

Stump1 (Smoking)

$$h_1(x) = \begin{cases} +1 & \text{if Smoking} \geq 1 \\ -1 & \text{if Smoking} = 0 \end{cases}$$

Predictions

Row	Smoking	y	$h_1(x)$	Check
1	0	−1	−1	✓
2	5	+1	+1	✓
3	0	+1	−1	X
4	12	+1	+1	✓
5	8	+1	+1	✓

Only **Row-3** wrong.

Step 1: Weighted error $\varepsilon_1 = w_3 = 0.2$

Step 2: Alpha $\alpha_1 = \frac{1}{2}\ln\left(\frac{1-0.2}{0.2}\right) = \frac{1}{2}\ln(4) = 0.6931$

Useful values:

- $e^{+\alpha_1} = e^{0.6931} \approx 2$
- $e^{-\alpha_1} \approx 0.5$

Step 3: Update weights (before normalization)

- Correct rows (1,2,4,5): $0.2 \times 0.5 = 0.1$
- Wrong row (3): $0.2 \times 2 = 0.4$

Sum = $0.1 + 0.1 + 0.4 + 0.1 + 0.1 = 0.8$

Normalize

- Correct: $0.1/0.8 = 0.125$
- Row-3: $0.4/0.8 = 0.5$

So after Round-1:

Row	Update weights	New weight w_i (after normalization)
1	0.1	0.125
2	0.1	0.125
3	0.4	0.500
4	0.1	0.125
5	0.1	0.125
Total	0.8	1

Justification: The only mistake was Row-3, so AdaBoost increases its weight heavily (now 0.5).

ROUND 2 — Choose h_2 to fix the hard point (Row-3)

Row-3 is **non-smoker but ill**. Smoking stump can't catch it. We need another feature.

Row	Gender	Age	Income (₹/yr)	Smoking	Illness	y	*New weight w_i*
1	M	38	420000	0	No	−1	0.125
2	F	52	360000	5	Yes	+1	0.125
3	M	45	780000	0	**Yes** *(noise)*	+1	0.500
4	F	29	300000	12	Yes	+1	0.125
5	M	61	500000	8	Yes	+1	0.125

Row-3 has **Age = 45**, which is relatively high compared to Row-4 (29).

Stump on Age: $h_2(x) = \begin{cases} +1 & \text{if Age} \geq 45 \\ -1 & \text{if Age} < 45 \end{cases}$

Predictions

Row	Age	y	$h_2(x)$	Check
1	38	−1	−1	✓
2	52	+1	+1	✓
3	45	+1	+1	✓
4	29	+1	−1	X
5	61	+1	+1	✓

Only **Row-4** is wrong.

Step 1: Weighted error (use Round-1 weights) $\varepsilon_2 = w_4 = 0.125$

Step 2: Alpha $\alpha_2 = \frac{1}{2}\ln\left(\frac{1-0.125}{0.125}\right) = \frac{1}{2}\ln(7) \approx \frac{1}{2}(1.9459) = 0.9730$

Useful values:

- $e^{+\alpha_2} \approx 2.646$
- $e^{-\alpha_2} \approx 0.378$

Step 3: Update weights (before normalization)

Current weights entering Round-2:

- $w_1 = w_2 = w_4 = w_5 = 0.125, w_3 = 0.5$

Update:

- Correct rows (1,2,3,5): multiply by 0.378
  - Row1: $0.125 \times 0.378 = 0.04725$
  - Row2: $0.125 \times 0.378 = 0.04725$
  - Row3: $0.5 \times 0.378 = 0.189$
  - Row5: $0.125 \times 0.378 = 0.04725$
- Wrong row (4): multiply by 2.646
  - Row4: $0.125 \times 2.646 = 0.33075$

Sum:

$$S = 0.04725(3) + 0.189 + 0.33075 = 0.14175 + 0.189 + 0.33075 = 0.6615$$

Normalize

- Rows 1,2,5: $0.04725/0.6615 \approx 0.0714$
- Row 3: $0.189/0.6615 \approx 0.2857$
- Row 4: $0.33075/0.6615 = 0.5$

Weights after Round-2:

Row	**New weight** w_i
1	0.0714
2	0.0714
3	0.2857
4	0.5000
5	0.0714

Justification: Now Row-4 became the "hard" case (young but ill), so its weight shoots up to 0.5.

ROUND 3 — Choose h_3 to fix Row-4

Row-4 has **Smoking = 12** (strongly suggests illness). So we reuse Smoking stump (it fixes Row-4 well).

$$h_3(x) = \begin{cases} +1 & \text{if Smoking} \geq 1 \\ -1 & \text{if Smoking} = 0 \end{cases}$$

Predictions (same as Round-1)

Only Row-3 (non-smoker but ill) will be wrong again.

Step 1: Weighted error (use Round-2 weights)

Row-3 is wrong, its weight is 0.2857:$\varepsilon_3 = 0.2857$

Step 2: Alpha$\alpha_3 = \frac{1}{2}\ln\left(\frac{1-\varepsilon_3}{\varepsilon_3}\right) = \frac{1}{2}\ln\left(\frac{0.7143}{0.2857}\right) = \frac{1}{2}\ln(2.5) \approx \frac{1}{2}(0.9163) = 0.4581$

Justification: This stump is weaker now (higher error), so it gets a smaller alpha than earlier rounds.

Final Strong Classifier after 3 rounds

AdaBoost combines stumps by weighted voting:

$$H(x) = \text{sign}(\alpha_1 h_1(x) + \alpha_2 h_2(x) + \alpha_3 h_3(x))$$

Here h_1 and h_3 are the *same* smoking rule, so we can combine:

- Smoking vote weight = $\alpha_1 + \alpha_3 = 0.6931 + 0.4581 = 1.1512$
- Age(≥45) vote weight = $\alpha_2 = 0.9730$

So:$H(x) = \text{sign}\big(1.1512\, h_{\text{smoke}}(x) + 0.9730\, h_{\text{age}\geq 45}(x)\big)$

Final predictions (showing the weighted vote)

Define:

- $h_{\text{smoke}} = +1$if Smoking≥1 else -1
- $h_{\text{age}\geq 45} = +1$if Age≥45 else -1

Row	h_{smoke}	$h_{\text{age}\geq 45}$	**Score** $= 1.1512h_s + 0.973h_a$	**Final** $H(x)$	**True** y
1	−1	−1	−2.1242	−1	✓ −1
2	+1	+1	+2.1242	+1	✓ +1
3	−1	+1	−0.1782	−1	X +1
4	+1	−1	+0.1782	+1	✓ +1
5	+1	+1	+2.1242	+1	✓ +1

After 3 rounds, AdaBoost correctly classifies **4 out of 5**.

Row-3 remains difficult because it's the "ill non-smoker" case; to fix it, we'd likely need:

- another feature (e.g., **BMI, chronic condition, family history**) or

- a different stump family (not just one-threshold rules), or
- more rounds and a stump that isolates Row-3 using Income + Age interactions.

Summary

- **Smoking** is a strong predictor (high alpha overall: 1.1512 across two rounds).
- **Age ≥ 45** helps catch illness even among non-smokers (alpha 0.9730).
- The noisy point (Row-3) shows why boosting is needed: one simple rule can't explain everything.

H.W: Continue with Round-4 and Round-5 using stumps on Income thresholds (e.g., Income ≥ 600000) and show whether the ensemble can eventually flip Row-3 without breaking others.

Example-2

Row No.	Gender	Age	Income (₹/year)	BMI	Illness	Sample Weights
1	Female	33	480000	22.1	No	1/5
2	Male	57	320000	29.5	Yes	1/5
3	Male	41	900000	24.0	No	1/5
4	Female	49	540000	31.2	Yes	1/5
5	Male	36	450000	27.8	No	1/5