Scenario

We want to predict if someone has heart disease (**Yes** or **No**) based on two features:

1. **Age Group**: Young, Middle-Aged, or Old
2. **Cholesterol Level**: High or Low

Dataset

X1 X2 Y

Age Group	Cholesterol Level	Heart Disease
Old	High	Yes
Middle-Aged	High	Yes
Middle-Aged	Low	No
Young	Low	No

X1,X2 - Features or Predictors and Y - Target or output

X1,X2 have categorical values

New Case

- A person is **Middle-Aged** with **High Cholesterol**.
- What's the probability they have heart disease (**Yes**)?

Naive Assumption

Naive Bayes assumes that:

1. The probability of **Age Group** and **Cholesterol Level** are independent.
    - This means whether a person is Middle-Aged has no influence on whether they have High Cholesterol.

Steps

Step 1: Calculate Probabilities

- $P(\text{Yes}) = \frac{2}{4} = 0.5$ P(Yes) = Number of Yes category values in target / Total number of values in target
- $P(\text{No}) = \frac{2}{4} = 0.5$ P(No) = Number of No category values in target / Total number of values in target

P(Middle - Aged) = 2/4 = 0.5

P(High Cholestrol) = 2/4 = 0.5

Likelihoods:

For **Yes**:

- $P(\text{Middle-Aged} \mid \text{Yes}) = \frac{1}{2} = 0.5$ P(Middle-Aged | Yes) = Number of matching Middle Aged- Yes category values in dataset / Total number of yes catrgory values in target
- $P(\text{High Cholesterol} \mid \text{Yes}) = \frac{2}{2} = 1.0$ P(Middle-Aged | Yes) = Number of matching High- Yes category values in dataset / Total number of yes catrgory values in target

For **No**:

- $P(\text{Middle-Aged} \mid \text{No}) = \frac{1}{2} = 0.5$
- $P(\text{High Cholesterol} \mid \text{No}) = \frac{0}{2} = 0$

Step 2: Apply Naive Bayes Formula

$$P(\text{Yes} \mid \text{Features}) = P(\text{Yes}) \cdot P(\text{Middle-Aged} \mid \text{Yes}) \cdot P(\text{High Cholesterol} \mid \text{Yes}) \quad |$$

$$\text{P(Middle-Aged) .P(High Cholesterol)} = 0.5 \cdot 0.5 \cdot 1.0 \,/\, 0.5 \,.\, 0.5 = 1.0$$

$$P(\text{No} \mid \text{Features}) = P(\text{No}) \cdot P(\text{Middle-Aged} \mid \text{No}) \cdot P(\text{High Cholesterol} \mid \text{No}) \,/\, \text{P(Middle-Aged)}$$

$$\text{.P(High Cholesterol)} = 0.5 \cdot 0.5 \cdot 0 \,/\, 0.5 \,.\, 0.5 = 0$$

Step 3: Decision

- $P(\text{Yes} \mid \text{Features}) = 1.0$
- $P(\text{No} \mid \text{Features}) = 0$

Since $P(\text{Yes} \mid \text{Features}) > P(\text{No} \mid \text{Features})$, **the model predicts Heart Disease = Yes**.

Key Point

The **naive assumption** here is that being **Middle-Aged** and having **High Cholesterol** are independent, even though they might be related in real life.