Decision Tree

Prof. Surya Prakash

Department of CSE, IIT Indore

Simple Health Dataset: Exercise & Diet vs. Heart Disease

Person	Exercise (hrs/week)	Diet Quality (1–10)	Heart Disease
1	0	3	Yes
2	1	4	Yes
3	2	5	Yes
4	3	6	No
5	4	7	No
6	5	8	No
7	3	4	Yes
8	4	6	No

Training Data

Test Person	Exercise (hrs/week)	Diet Quality (1–10)	Predicted Heart Disease
T1	2	5	
T2	4	7	

Test Data

Simple Health Dataset: Exercise & Diet *vs.* Heart Disease

Person	Exercise (hrs/week)	Diet Quality (1–10)	Heart Disease
1	**0**	3	Yes
2	**1**	4	Yes
3	**2**	5	Yes
4	3	6	No
5	4	7	No
6	5	8	No
7	3	4	Yes
8	4	6	No

Learning *(from training samples)*:

if Exercise < 3
    Heart Disease = Yes
else
    if Diet < 6
        Heart Disease = Yes
    else
        Heart Disease = No

Simple Health Dataset: Exercise & Diet vs. Heart Disease

Learning *(from training samples)*:

if Exercise < 3
    Heart Disease = Yes
else
    if Diet < 6
        Heart Disease = Yes
    else
        Heart Disease = No

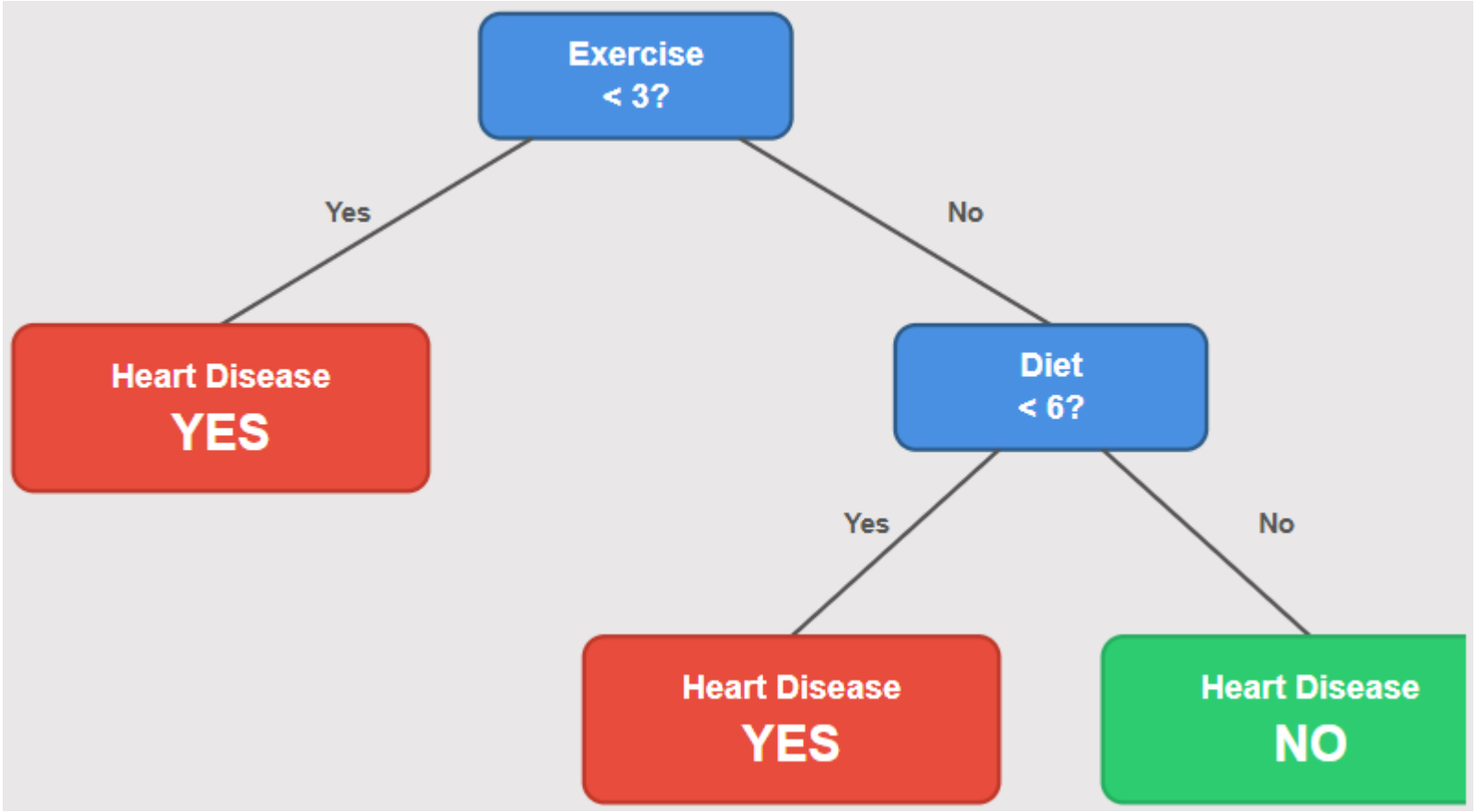
Test Person	Exercise (hrs/week)	Diet Quality (1–10)	Predicted Heart Disease
T1	2	5	**Yes** (low exercise)
T2	4	7	**No** (good exercise, good diet)

Test Data

Learning Represented in the form of a Tree

Learning *(from training samples)*:

```
if  Exercise < 3
      Heart Disease = Yes
else
   if Diet < 6
      Heart Disease = Yes
   else
      Heart Disease = No
```



Heart Disease Decision Tree

Another Observation

Person	Exercise (hrs/week)	Diet Quality (1–10)	Heart Disease
1	0	3	Yes
2	1	4	Yes
3	2	5	Yes
4	3	6	No
5	4	7	No
6	5	8	No
7	3	4	Yes
8	4	6	No

Learning *(from training samples)*:

if Diet < 6

Heart Disease = Yes

else

Heart Disease = No

Learning Represented in the form of a Tree

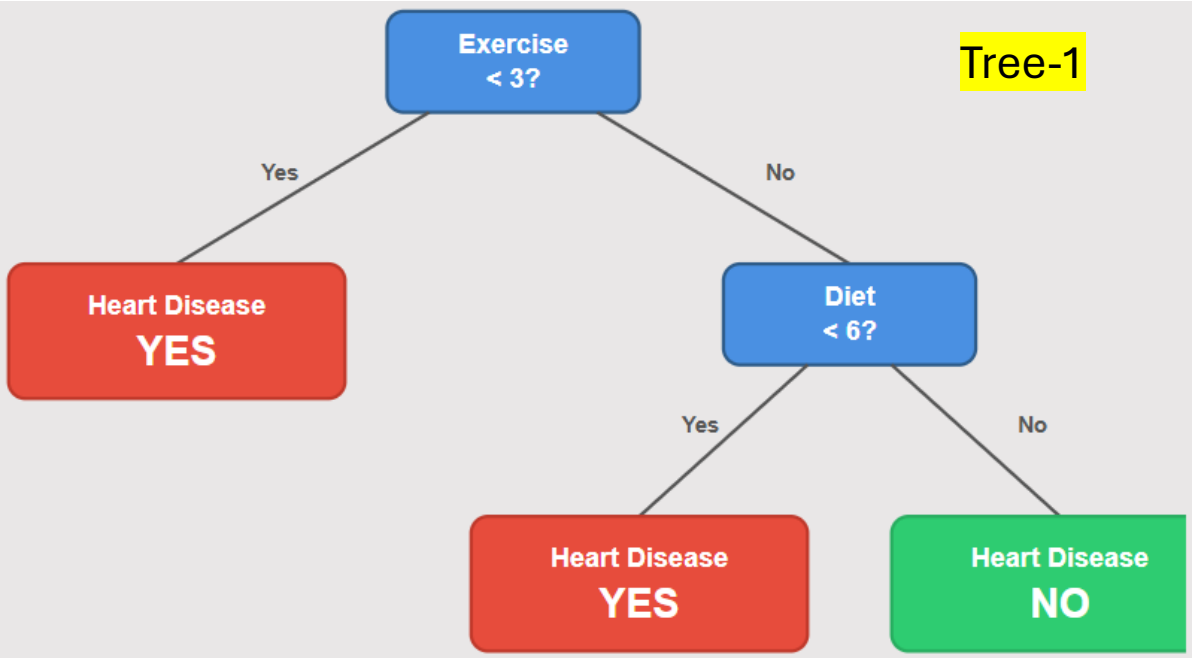
Learning *(from training samples)*:

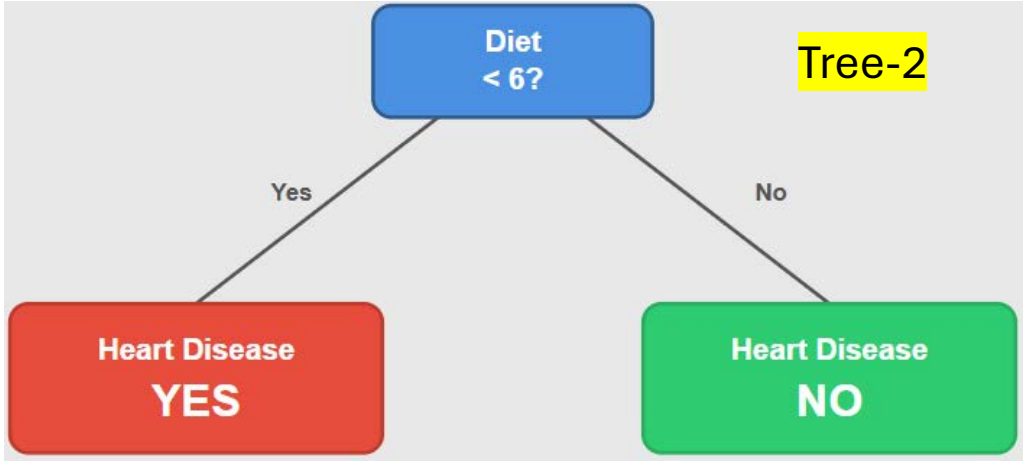
if Diet < 6
    Heart Disease = Yes
else
    Heart Disease = No

Heart Disease Decision Tree

Test Person	Exercise (hrs/week)	Diet Quality (1–10)	Predicted Heart Disease
T1	2	5	**Yes** (low exercise)
T2	4	7	**No** (good exercise, good diet)

Two Decision Trees – which one is better?





Heart Disease Decision Trees

- Both can perform classification, but which one is better?
- Tree-2 is better as it requires less inference time
- Decision tree building algorithm helps to get the best tree

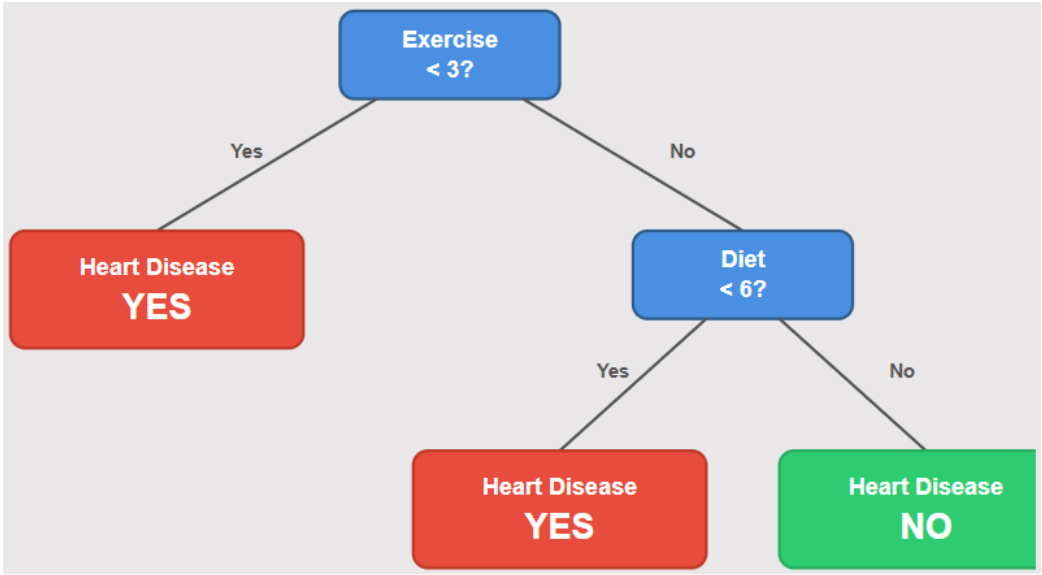
Introduction to Decision Trees

- **What is a Decision Tree?**
  - A decision tree is a flowchart-like structure used for decision making or classification.

- **Uses:**
  - Classification
  - Regression
  - Feature selection

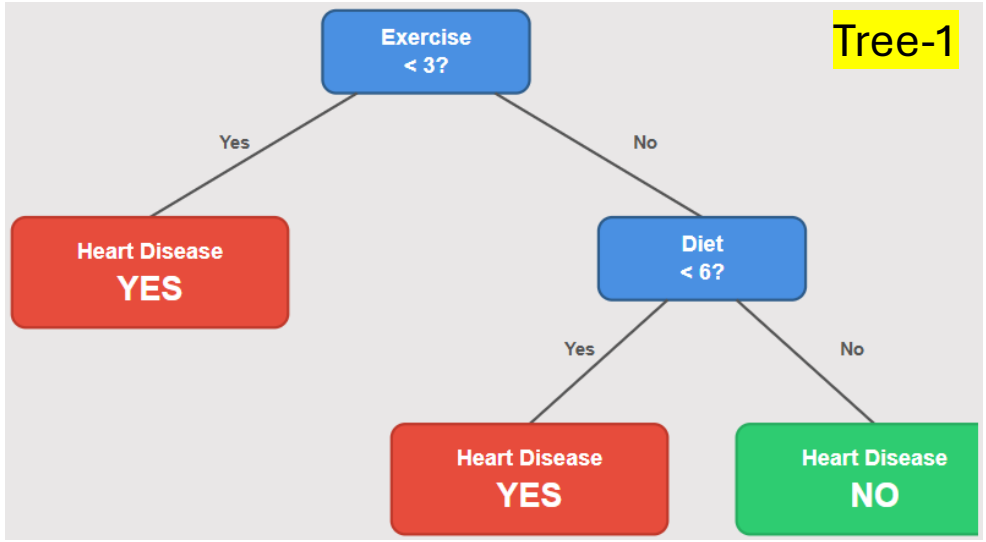
Structure of a Decision Tree

- **Root Node:**
  - Represents the entire dataset and the first decision point
- **Internal Nodes:**
  - Represent decisions based on attributes
- **Leaf Nodes:**
  - Represent the output class/label
- **Edges:**
  - Represent outcomes of the decision at a node



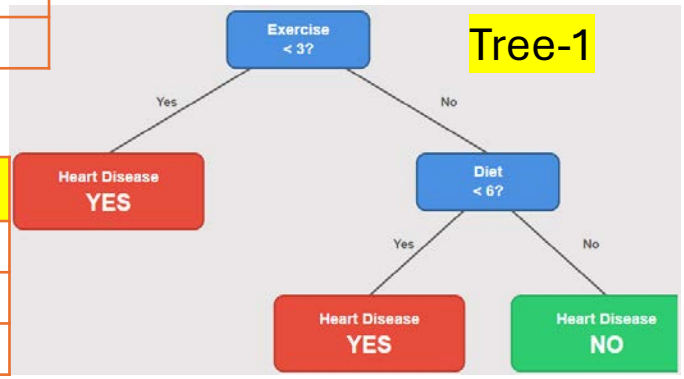
Building a Decision Tree

- The **main objective** in building a **Decision Tree** is to **select the best feature (attribute)** at each level (or node) to **split the data** so that the resulting subsets are as **pure** as possible.
- Purity of subsets?
  - This meaning the data points within each subset mostly belong to **one class**.



Building a Decision Tree

Person	Exercise (hrs/week)	Diet Quality (1–10)	Heart Disease
1	0	3	Yes
2	1	4	Yes
3	2	5	Yes
4	3	6	No
5	4	7	No
6	5	8	No
7	3	4	Yes
8	4	6	No



Person	Exercise (hrs/week)	Diet Quality (1–10)	Heart Disease
1	0	3	Yes
2	1	4	Yes
3	2	5	Yes

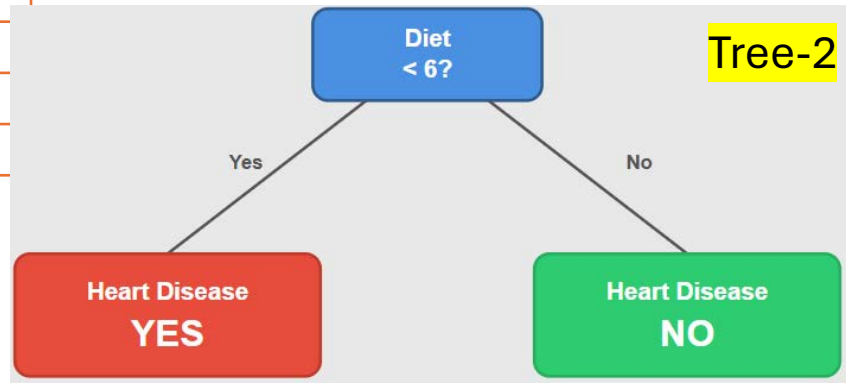
Person	Exercise (hrs/week)	Diet Quality (1–10)	Heart Disease
4	3	6	No
5	4	7	No
6	5	8	No
7	3	4	Yes
8	4	6	No

Person	Exercise (hrs/week)	Diet Quality (1–10)	Heart Disease
7	3	4	Yes

Person	Exercise (hrs/week)	Diet Quality (1–10)	Heart Disease
4	3	6	No
5	4	7	No
6	5	8	No
8	4	6	No

Building a Decision Tree

Person	Exercise (hrs/week)	Diet Quality (1–10)	Heart Disease
1	0	3	Yes
2	1	4	Yes
3	2	5	Yes
4	3	6	No
5	4	7	No
6	5	8	No
7	3	4	Yes
8	4	6	No



Person	Exercise (hrs/week)	Diet Quality (1–10)	Heart Disease
1	0	3	Yes
2	1	4	Yes
3	2	5	Yes
7	3	4	Yes

Person	Exercise (hrs/week)	Diet Quality (1–10)	Heart Disease
4	3	6	No
5	4	7	No
6	5	8	No
8	4	6	No

Building a Decision Tree

- **ID3: Iterative Dichotomiser 3:**
  - ID3 algorithm is specifically designed for **classification**, not regression.
  - The earlier versions (ID1 and ID2) were unpublished or internal prototypes.
  - ID3 was the first version to be formally published and widely adopted.

ID3 Algorithm

- Uses the concept of
  - Entropy
  - Information gain

End