Assignment 10

Node Purity in Decision Trees

Let us assume we are building a decision tree to differentiate between sick and healthy patients. The tree has three nodes. Each node has 30 data points that belong to the two classes: Sick and Healthy as follows:

* Node 1 contains 30 data points
  + 17 points belong to class Sick
  + 13 points belong to class Healthy

|  | Node 1 |
| --- | --- |
| Sick | 17 |
| Healthy | 13 |

* Node 2 contains 30 data points
  + 12 points belong to class Sick
  + 18 points belong to class Healthy

|  | Node 2 |
| --- | --- |
| Sick | 12 |
| Healthy | 18 |

* Node 3 contains 30 data points
  + 15 points belong to class Sick
  + 15 points belong to class Healthy

|  | Node 3 |
| --- | --- |
| Sick | 15 |
| Healthy | 15 |

1. Which node is the purest? Explain.
   1. Node 2 is the purest. A pure node is one with most of its data points belonging to the same group. Node two has a 6-patient majority of healthy patients over sick patients. Node 1 only has a 4-patient majority of sick patients, making Node 2 more pure.
2. Which node is the most impure? Explain.
   1. Node 3 is the most impure. An impure node is one where both options appear with equal probability, and Node 3 has an equal number of sick and healthy patients.