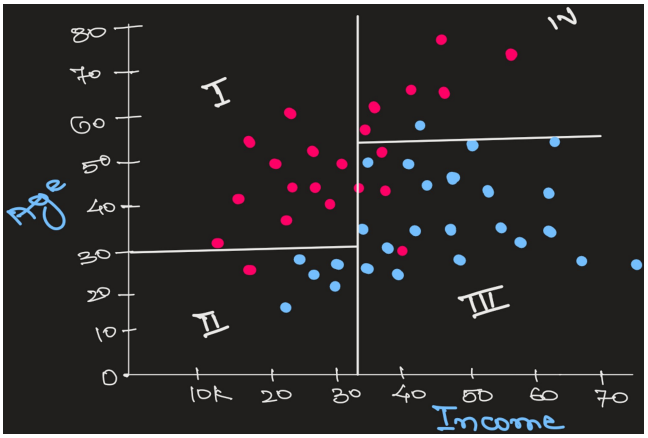
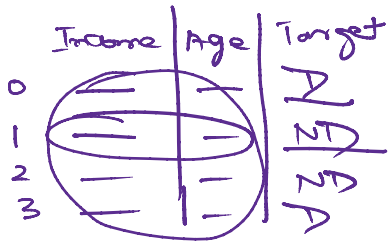
Decision Trees

18 December 2024 06:58

• Defaulters • Non Defaulters



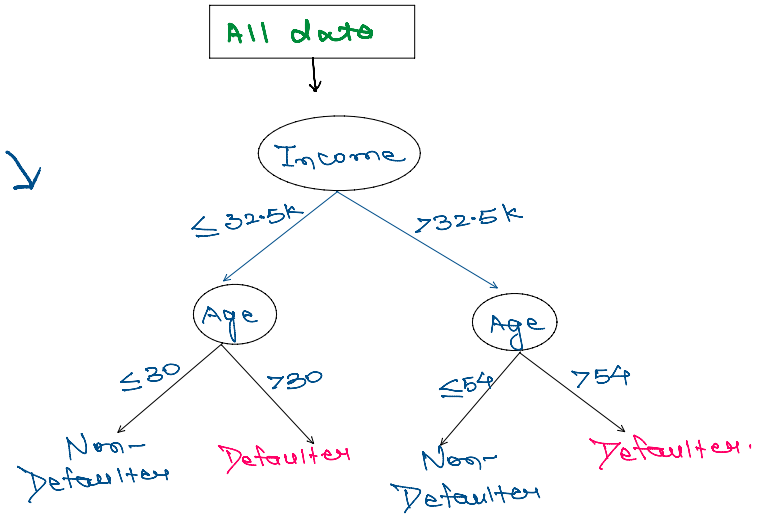


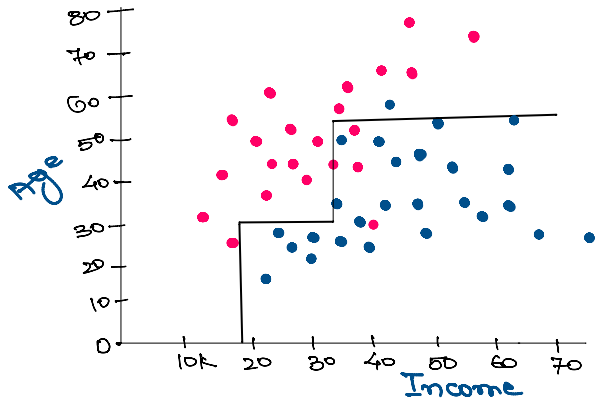
Aim: Using factors like income and Age of the customer can we identify whether a customer is going to default on loan repayment or not.

Let's build the Decision Tree using above graph →

When building a Decision Trees, we start with a single node which has the complete data then we start to split it on the basis of best split features.

The best split is the one where the data is split in two groups which are most homogeneous (most points of same category together).

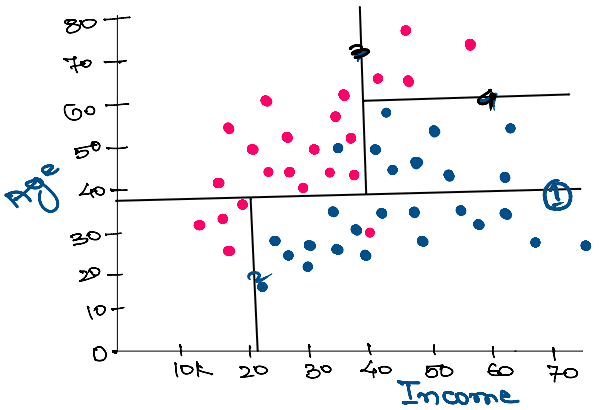




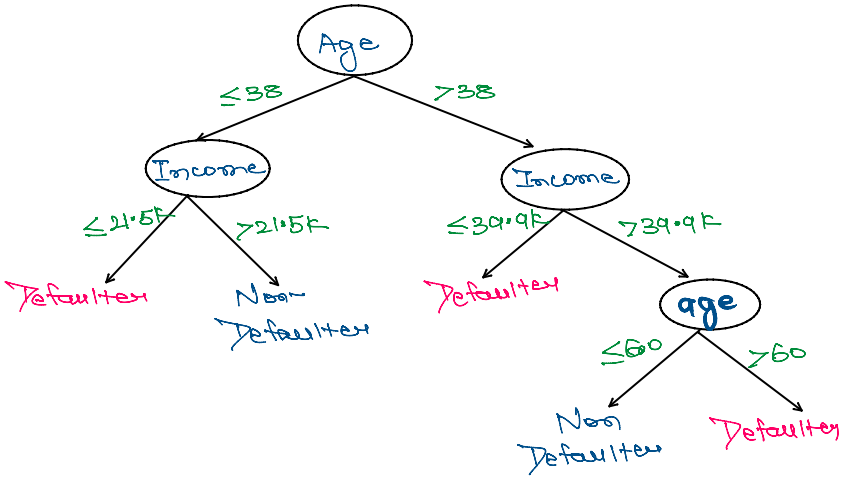
This is how our final decision boundary will look like.

** What if we split on the Age variable first?

Let's see



For this split, our tree will look like:



On same data, we see that we can split the tree in various ways, but how do we find the best split variable and best split values.

Also when to stop splitting?

Finding the best split →

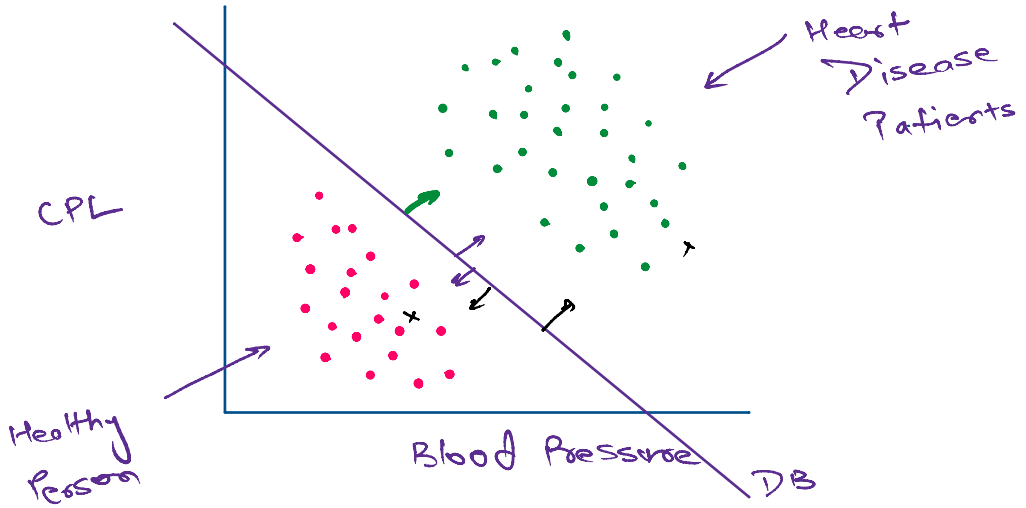
There are various ways to select the best split.

Finding the best split:

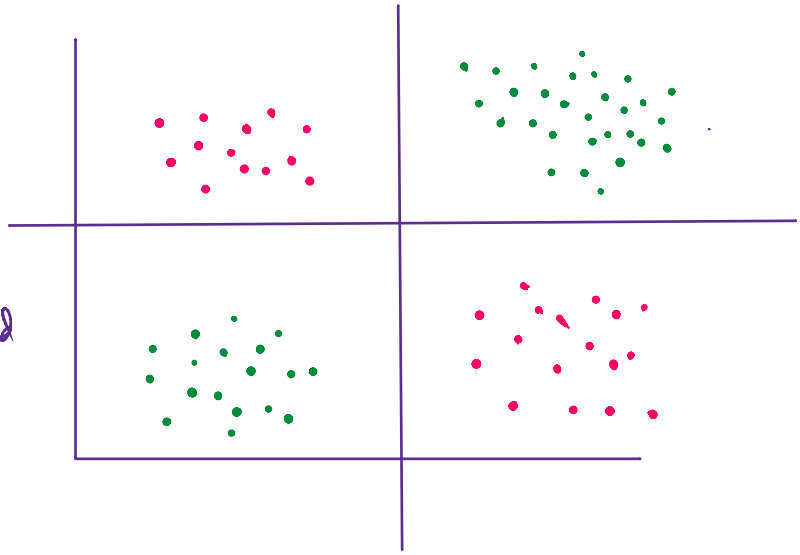
There are various ways to select the best split.

Commonly used approaches →

1. ID3 Algorithm (uses Entropy and Information gain).
2. CART Algorithm (uses Gini Impurity) → Evaluates all possible binary splits (exhaustive search).



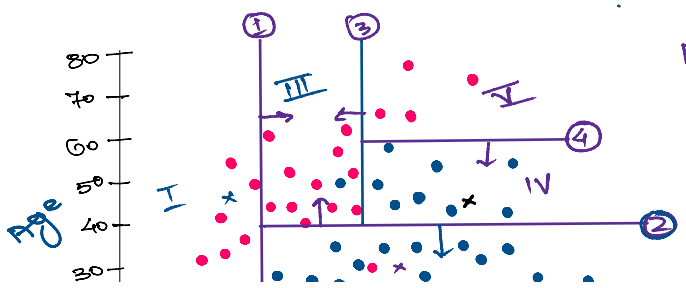
LR can't build more than one decision boundary if required



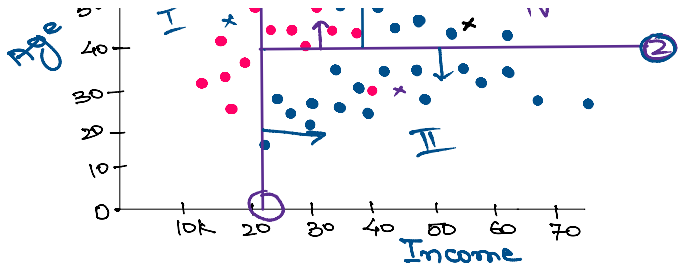
2D

5D

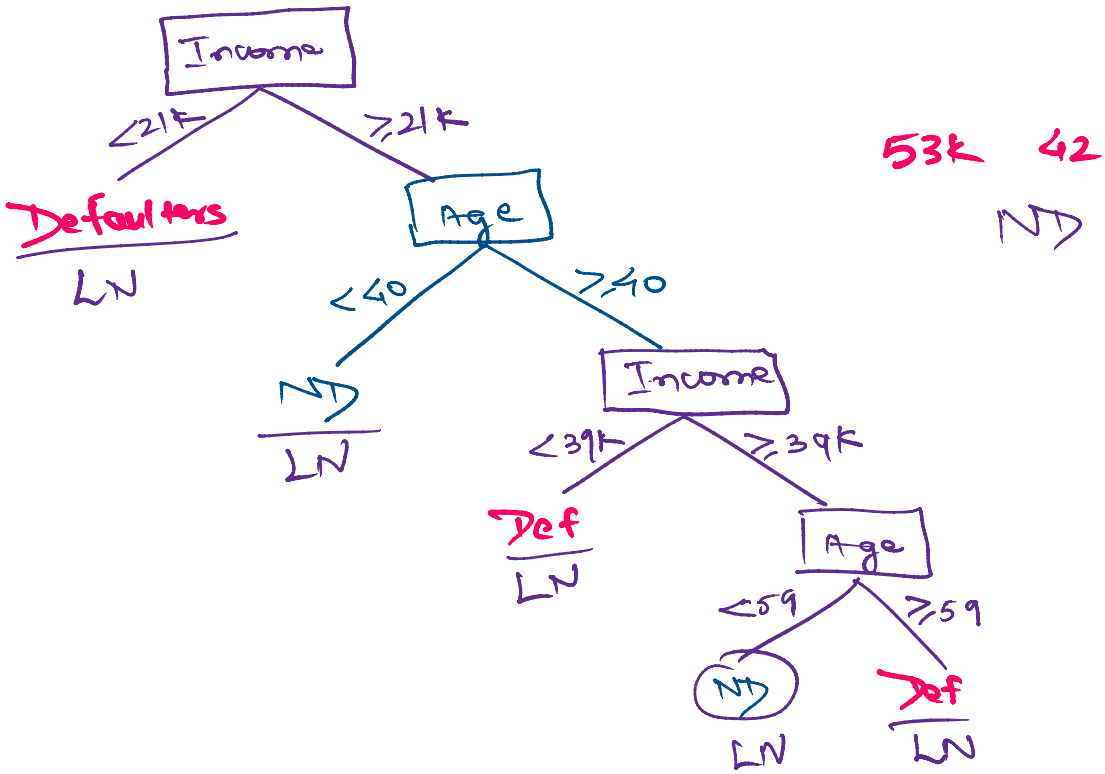
3D



More homogeneous

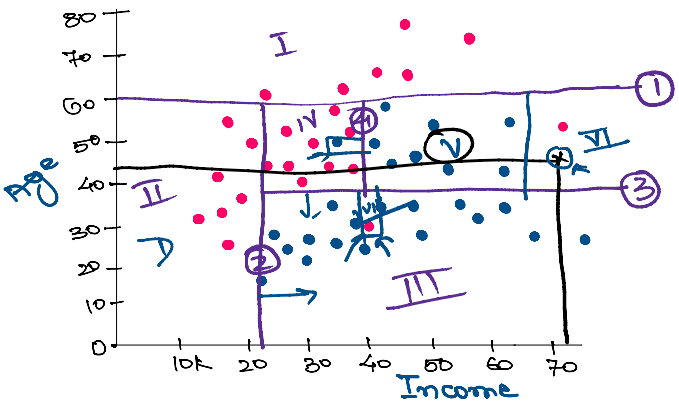


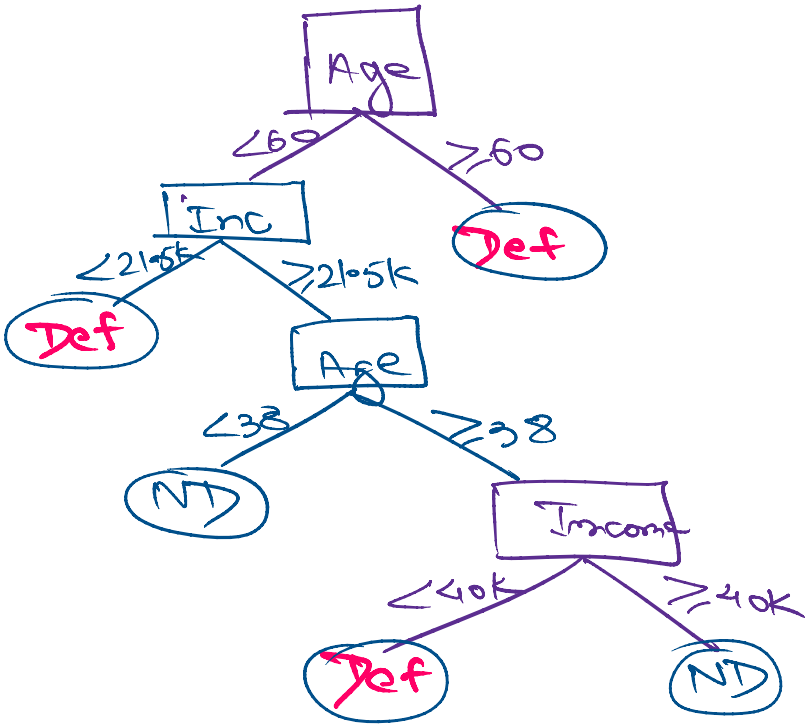
53k 42



53k 42

ND





53k (42) ND

CF ↔ TE ↓ TSE

① ID3 Algorithm ← old

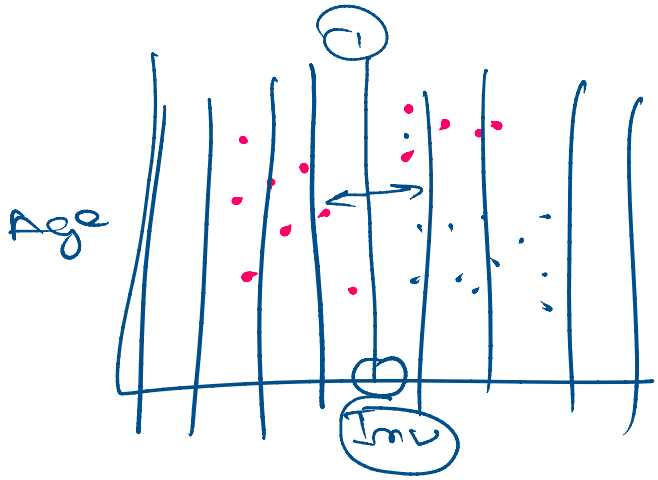
② CART Algorithm ← New

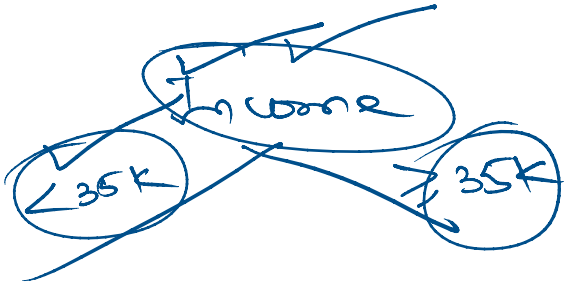
ID3 Algorithm
↓
Entropy &
Information
Gain

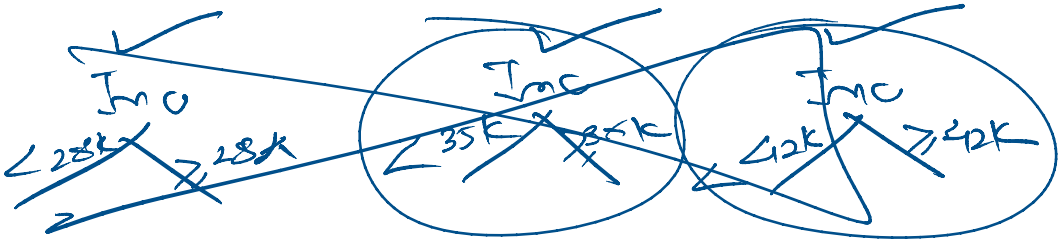
CART Algorithm
↓
→ Gini Impurity
Gini Index

CART: Classification & Regression Trees.

CART: Classification & Regression Trees.







GI

Powerful ⟶ overlearning

Powerful ⟶ overlearning

↓

overfitting