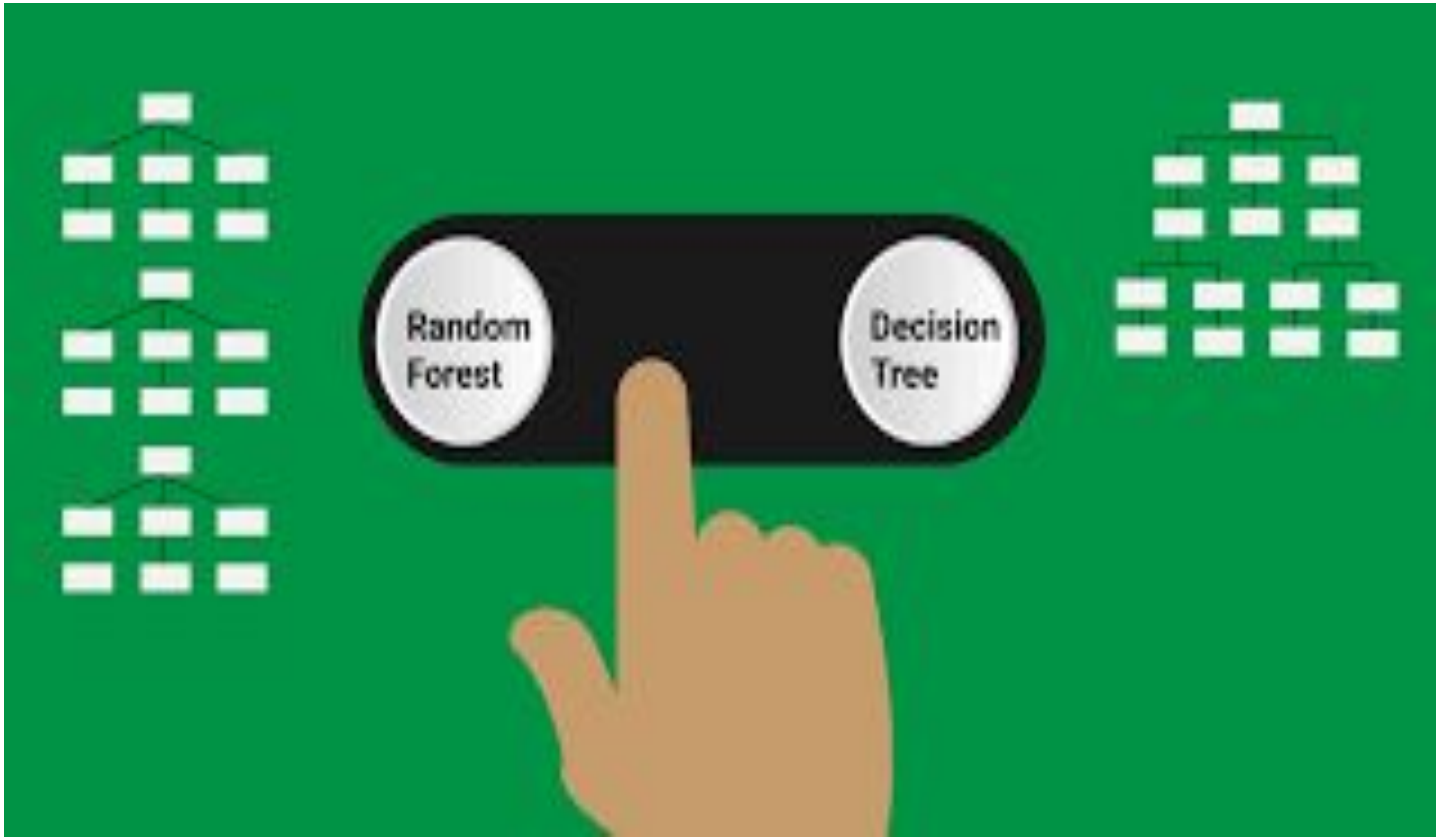
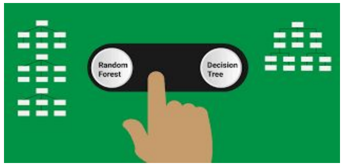
Decision Trees and Random Forest



Agenda

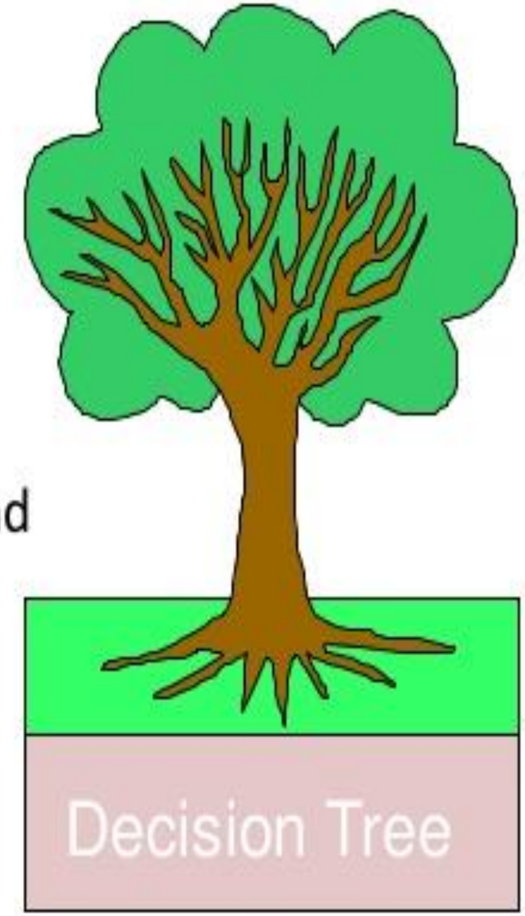


- Introduction to Decision Trees
- Homogenity
- Entropy
- Information Gain
- Gini Index
- Implementing Decision Trees
- Implementing Random Forest

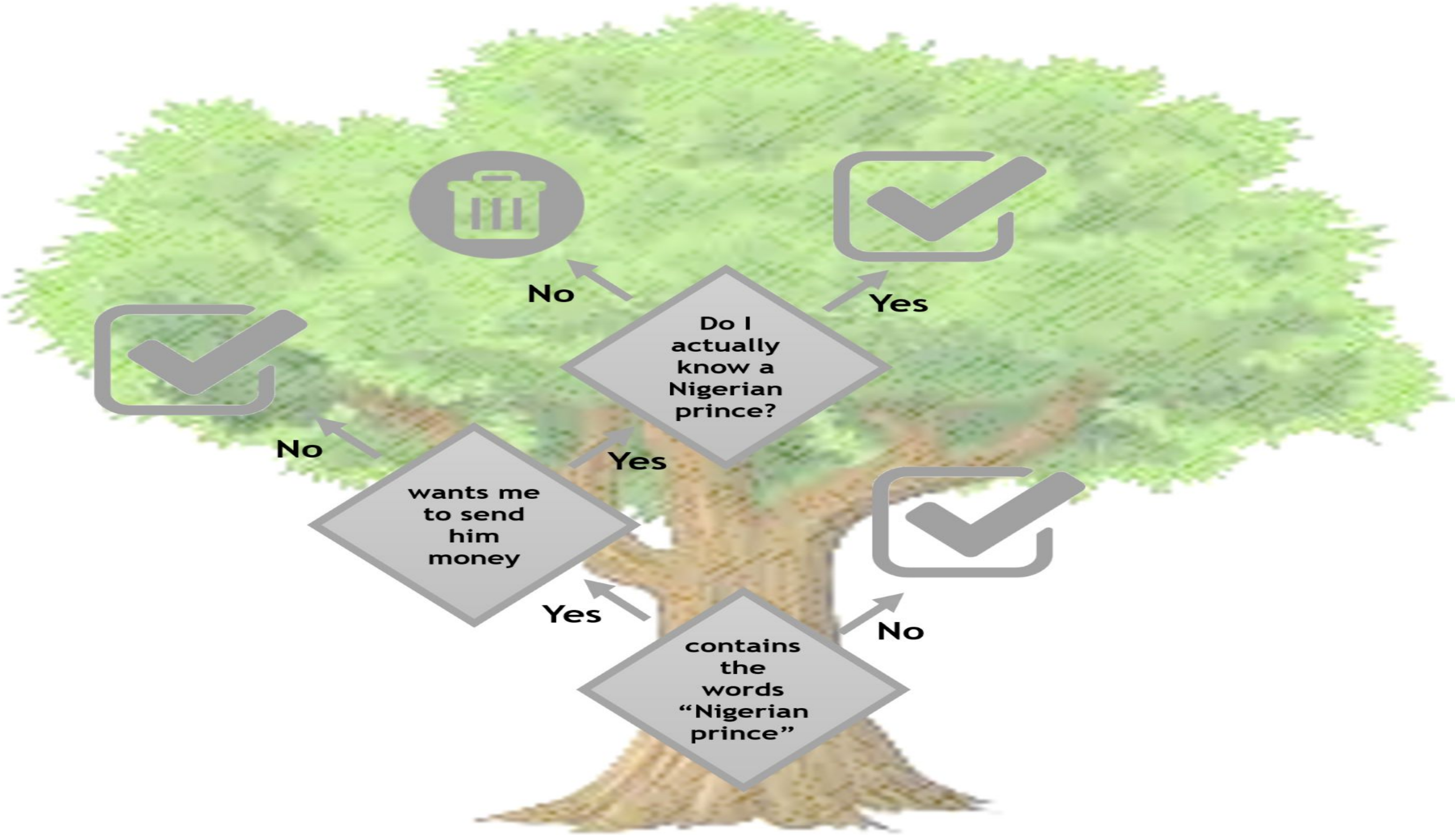
What is a Decision Tree?

- A Visual Representation of Choices, Consequences, Probabilities, and Opportunities.
- A Way of Breaking Down Complicated Situations Down to Easier-to-Understand Scenarios.
- By applying

- Logic
- Likely Outcome
- Quantitative decision



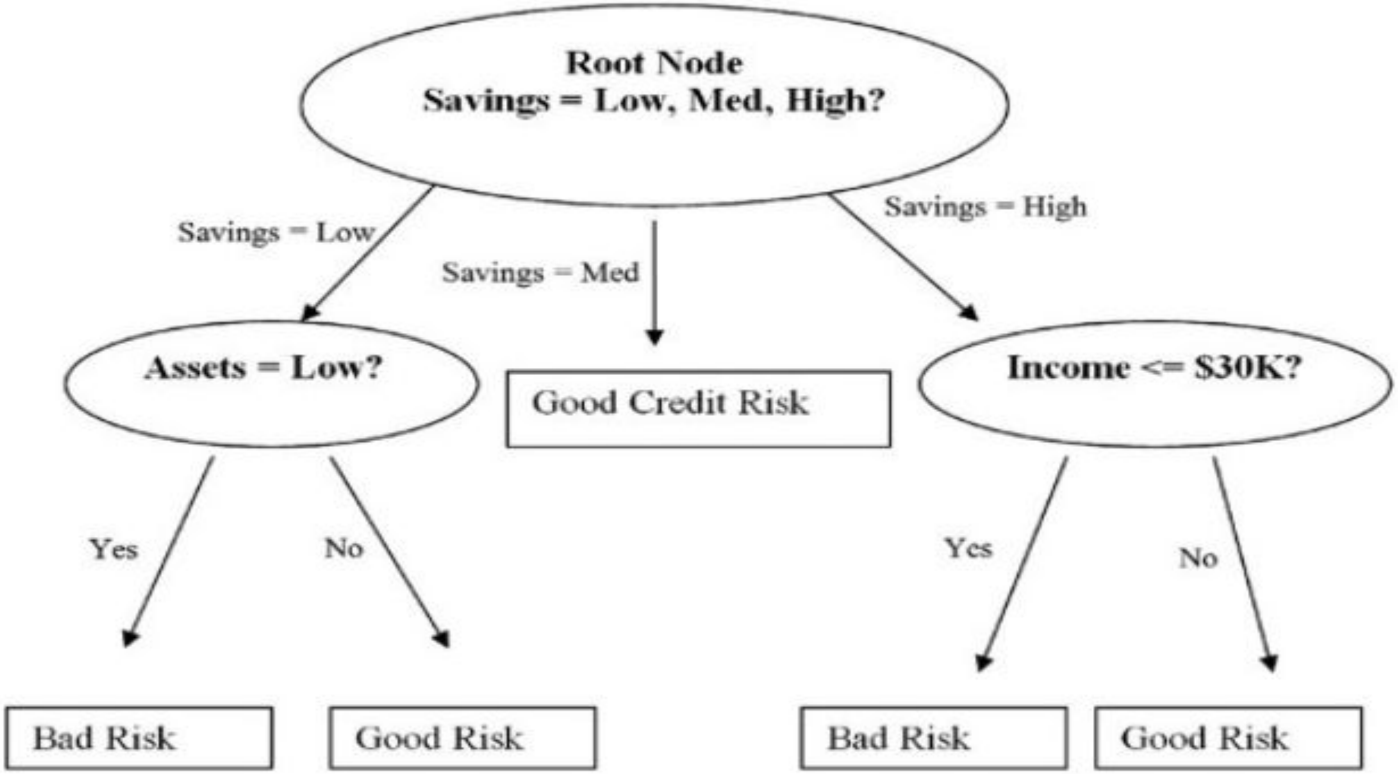
Introduction to Decision Trees


contains the words "Nigerian prince"
Yes
No
wants me to send him money
No
Yes
Do I actually know a Nigerian prince?
No
Yes

- A **decision tree** is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences
- A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

Decision Trees- Applications

Decision trees have a natural “if … then … else …” construction that makes it fit easily into a programmatic structure.
They also are well suited to categorization problems where attributes or features are systematically checked to determine a final category.



Types of Decision Trees

Types of decision tree is based on the type of target variable we have. It can be of two types

Categorical Variable Decision Tree (classifiers)

- In the scenario of student problem, where the target variable was “Student will play cricket or not” i.e. YES or NO.

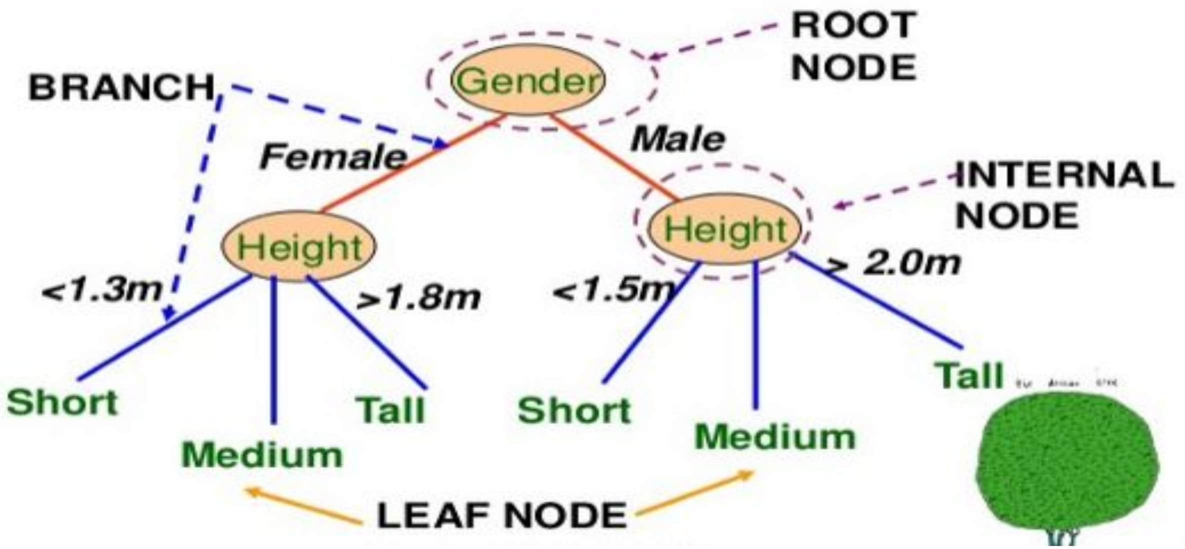
Continuous Variable Decision Tree (Regressors)

- Decision Tree has continuous target variable then it is called as Continuous Variable Decision Tree.

Decision Trees- Classification

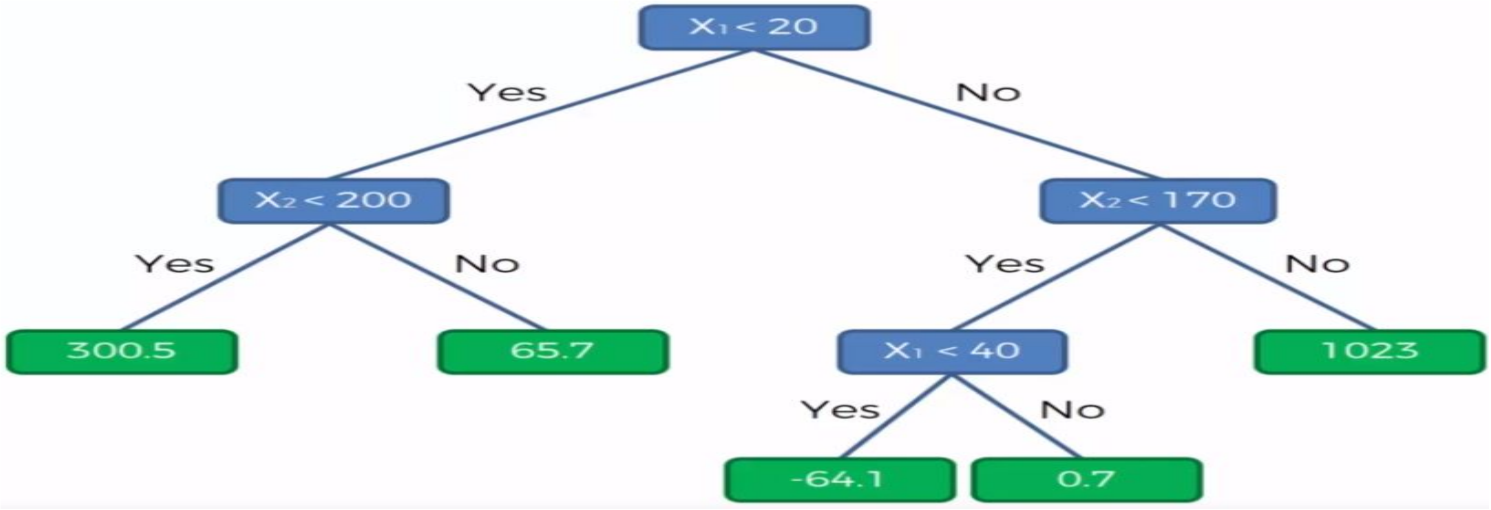
- Let’s say we have a sample of 30 students with three variables Gender (Boy/ Girl), Class (IX/ X) and Height (5 to 6 ft). 15 out of these 30 play cricket in leisure time.
- Now, we want to create a model to predict who will play cricket during leisure period?
- In this problem, we need to segregate students who play cricket in their leisure time based on highly significant input variable among all three.

Decision Tree Diagram

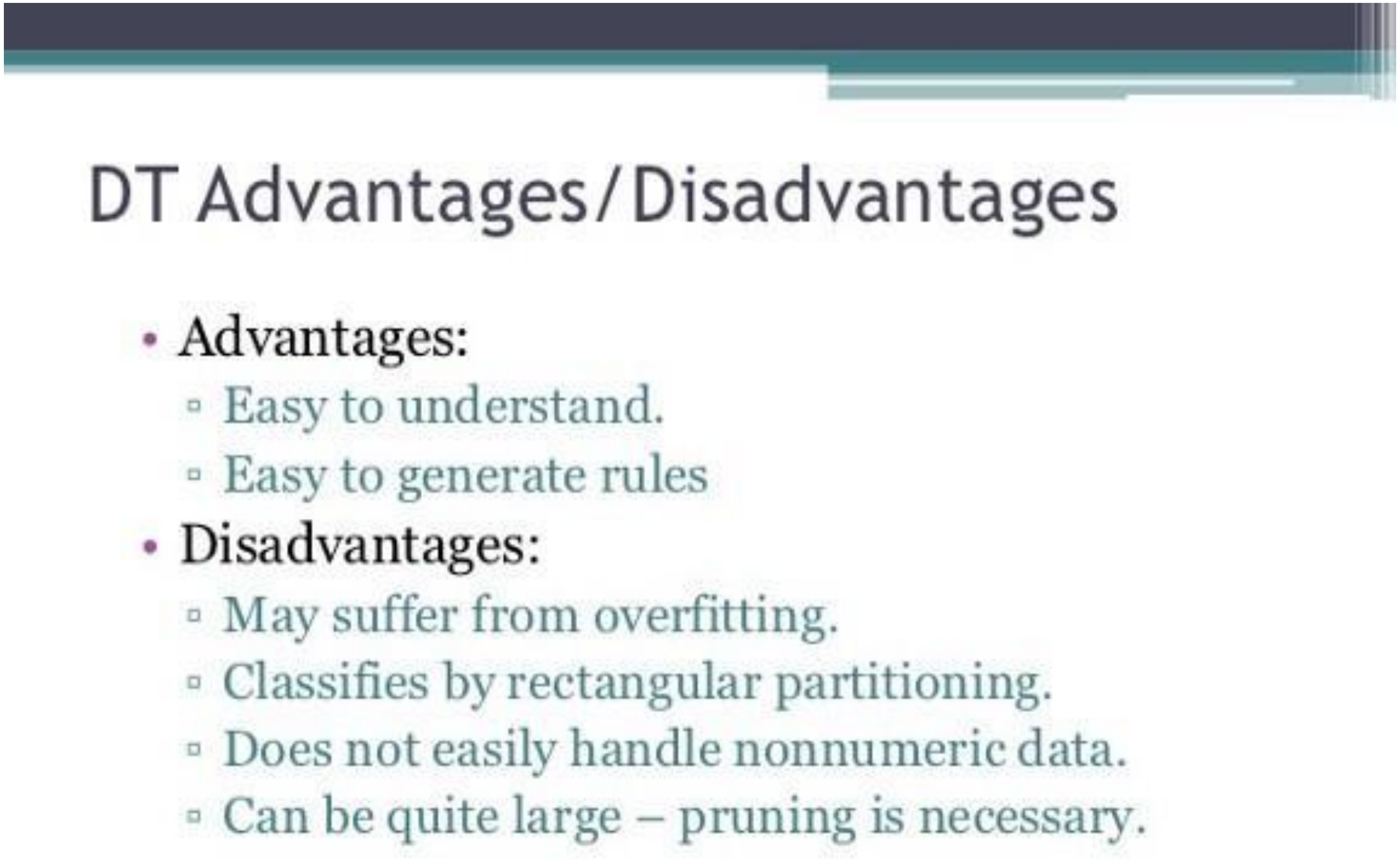


Decision Trees- Regression

- Let’s say we have a problem to predict whether a customer will pay his renewal premium with an insurance company (yes/ no).
- Here we know that income of customer is a significant variable but insurance company does not have income details for all customers.
- Now, as we know this is an important variable, then we can build a decision tree to predict customer income based on occupation, product and various other variables. In this case, we are predicting values for continuous variable.
- Another example below-



Decision Trees…

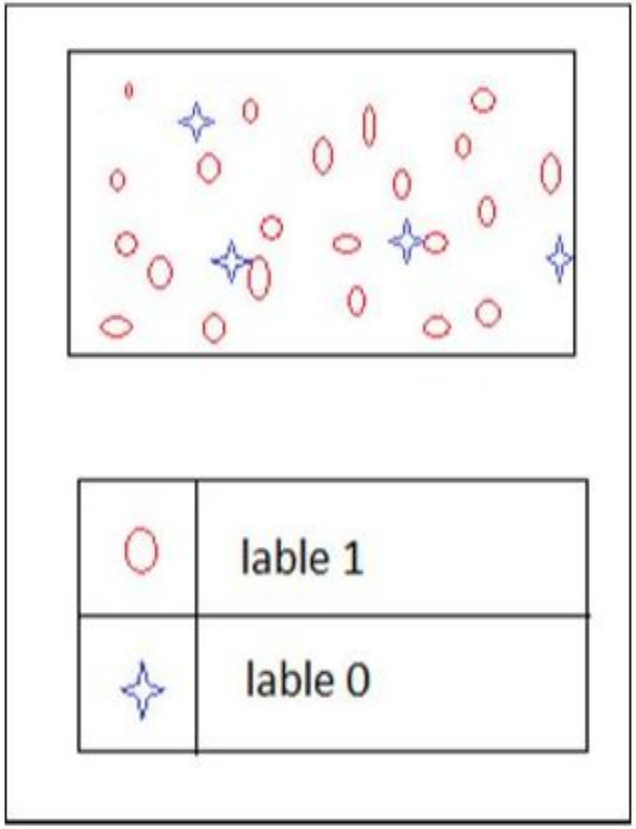


DT Advantages/Disadvantages

- Advantages:
  - Easy to understand.
  - Easy to generate rules
- Disadvantages:
  - May suffer from overfitting.
  - Classifies by rectangular partitioning.
  - Does not easily handle nonnumeric data.
  - Can be quite large – pruning is necessary.

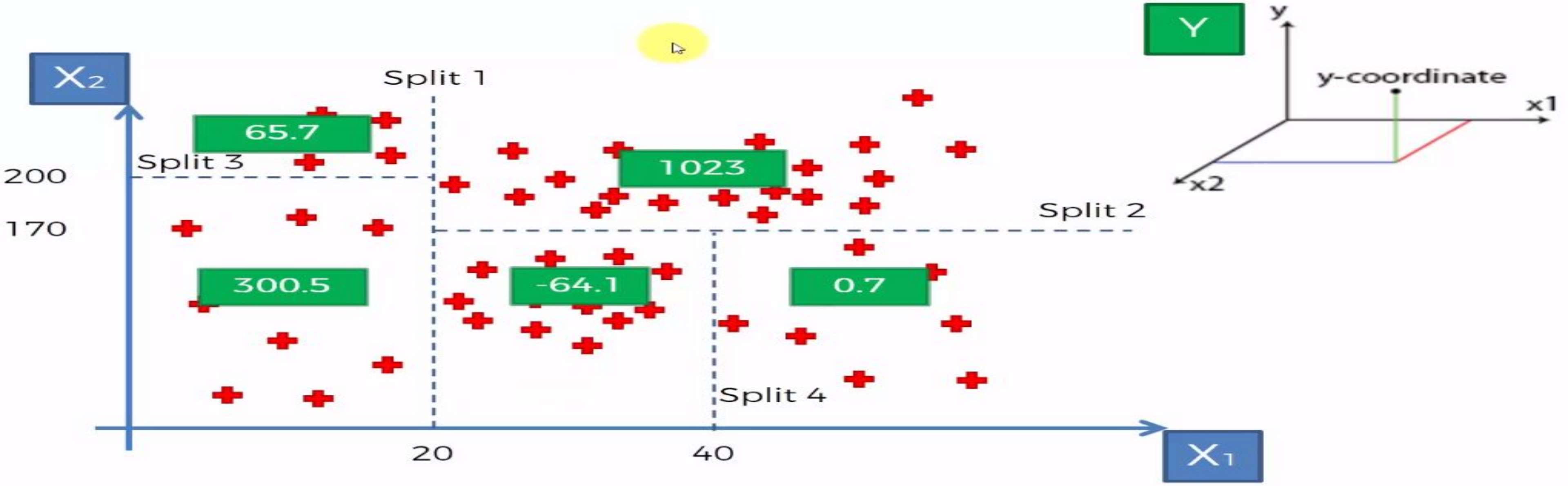
Homogeneity

- What Decision tree construction algorithm will try to do is to create a split in such a way that the homogeneity of different pieces must be as high as possible.



Homogeneity…

- If the previous fig denotes a branch after a split and H be its homogeneity.
- The decision tree checks whether H > some threshold then there is no further split.
- If the H < threshold then there will be further split. This process will be continued where the label is sufficiently homogeneous then a leaf is created.
- So we go step by step, picking an attribute and splitting the data such that the homogeneity increases after every split.
- We stop splitting when the resulting leaves are sufficiently homogeneous.
- There are various ways to quantify homogeneity, such as the Gini Index, Information gain, Entropy etc.



Q.1 Define Decision Trees?

Q.2 State some examples of Decision trees?

Q.3 What are the types of Decision trees?

Q.4 What are the advantages and disadvantages of Decision trees?

Q.5 Explain Homogenity ?