



# **SNS COLLEGE OF ENGINEERING**

Kurumbapalayam (Po), Coimbatore – 641 107

**An Autonomous Institution**

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A' Grade  
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

## **DEPARTMENT OF COMPUTER SCIENCE AND TECHNOLOGY**

**COURSE NAME :19CS407 DATA ANALYTICS WITH R**  
**II YEAR /IV SEMESTER**

**Unit 1- Introduction**

**Topic : A Short Taxonomy of Data Analytics**





# Data





# Data



- ✓ Data, in the information age, are a large set of bits encoding numbers, texts, images, sounds, videos, and so on.
- ✓ When we add information, giving a meaning to them, these data become knowledge.
- ✓ But before data become knowledge, typically, they pass through several steps where they are still referred to as data, despite being a bit more organized;



# Example



Table 1.1 Data set of our private contact list.

Contact	Age	Educational level	Company
Andrew	55	1.0	Good
Bernhard	43	2.0	Good
Carolina	37	5.0	Bad
Dennis	82	3.0	Good
Eve	23	3.2	Bad
Fred	46	5.0	Good
Gwyneth	38	4.2	Bad
Hayden	50	4.0	Bad
Irene	29	4.5	Bad
James	42	4.1	Good
Kevin	35	4.5	Bad
Lea	38	2.5	Good
Marcus	31	4.8	Bad
Nigel	71	2.3	Good

- Each column represents a characteristic of the data and each row represents an occurrence of the data.
- A column is referred to as an attribute or, with the same meaning, a feature, while a row is referred to as an instance, or with the same meaning, an object



# Example



**Table 1.2** Family relations between contacts.

Friend	Father	Mother	Sister
Eve	Andrew	Hayden	Irene
Irene	Andrew	Hayden	Eve

- Data sets represented by several tables, making clear the relations between these tables, are called relational data sets.
- This information is easily handled using relational databases.



# Instance or Object

- we intend to characterize people in our private contact list. Each member is, in this case, an instance or object. It corresponds to a row of the table

# Attribute or Feature

- Attributes, also called features, are characteristics of the instances
- contact, age, education level and company are four different attributes



# A Short Taxonomy of Data Analytics



- A natural taxonomy that exists in data analytics is:
  - Descriptive analytics:** summarize or condense data to extract patterns
  - Predictive analytics:** extract models from data to be used for future predictions.
- In descriptive analytics tasks, the result of a given method or technique, is obtained directly by applying an algorithm to the data.
- The result can be a statistic, such as an average, a plot, or a set of groups with similar instances, among other things



# A Short Taxonomy of Data Analytics



## Method or technique

- method or technique is a systematic procedure that allows us to achieve an intended goal.
- A method shows how to perform a given task. But in order to use a language closer to the language computers can understand, it is necessary to describe the method/technique through an algorithm

## Algorithm

- An algorithm is a self-contained, step-by-step set of instructions easily understandable by humans, allowing the implementation of a given method.
- They are self-contained in order to be easily translated to an arbitrary programming language





# A Short Taxonomy of Data Analytics



## Example

The method to obtain the average age of my contacts uses the ages of each (we could use other methods, such as using the number of contacts for each different age)

---

**Algorithm** An algorithm to calculate the average age of our contacts

---

- 1: INPUT:  $A$ : a vector of size  $N$  with the ages of all contacts.
  - 2:  $S \leftarrow 0$  ▷ Initialize the sum  $S$  to zero
  - 3: **for**  $i = 1$  **to**  $N$  **do** ▷ Iterate through all the elements of  $A$ .
  - 4:      $S \leftarrow S + A_i$  ▷ Add the current (*ith*) element of  $A$  to  $S$ .
  - 5:  $\bar{A} \leftarrow S/N$  ▷ Divide the sum by the number  $N$  of contacts.
  - 6: **return**( $\bar{A}$ ) ▷ Return the result, i.e. the average age of the  $N$  contacts.
- 

For instance, the average could be expressed as:  $\bar{A} = \sum_{i=1}^N A_i / N$ .



# A Short Taxonomy of Data Analytics



## Model

A model in data analytics is a generalization obtained from data that can be used afterwards to generate predictions for new given instances. It can be seen as a prototype that can be used to make predictions. Thus, model induction is a predictive task



# A Short Taxonomy of Data Analytics



## Model

A model in data analytics is a generalization obtained from data that can be used afterwards to generate predictions for new given instances. It can be seen as a prototype that can be used to make predictions. Thus, model induction is a predictive task



## Examples of Data

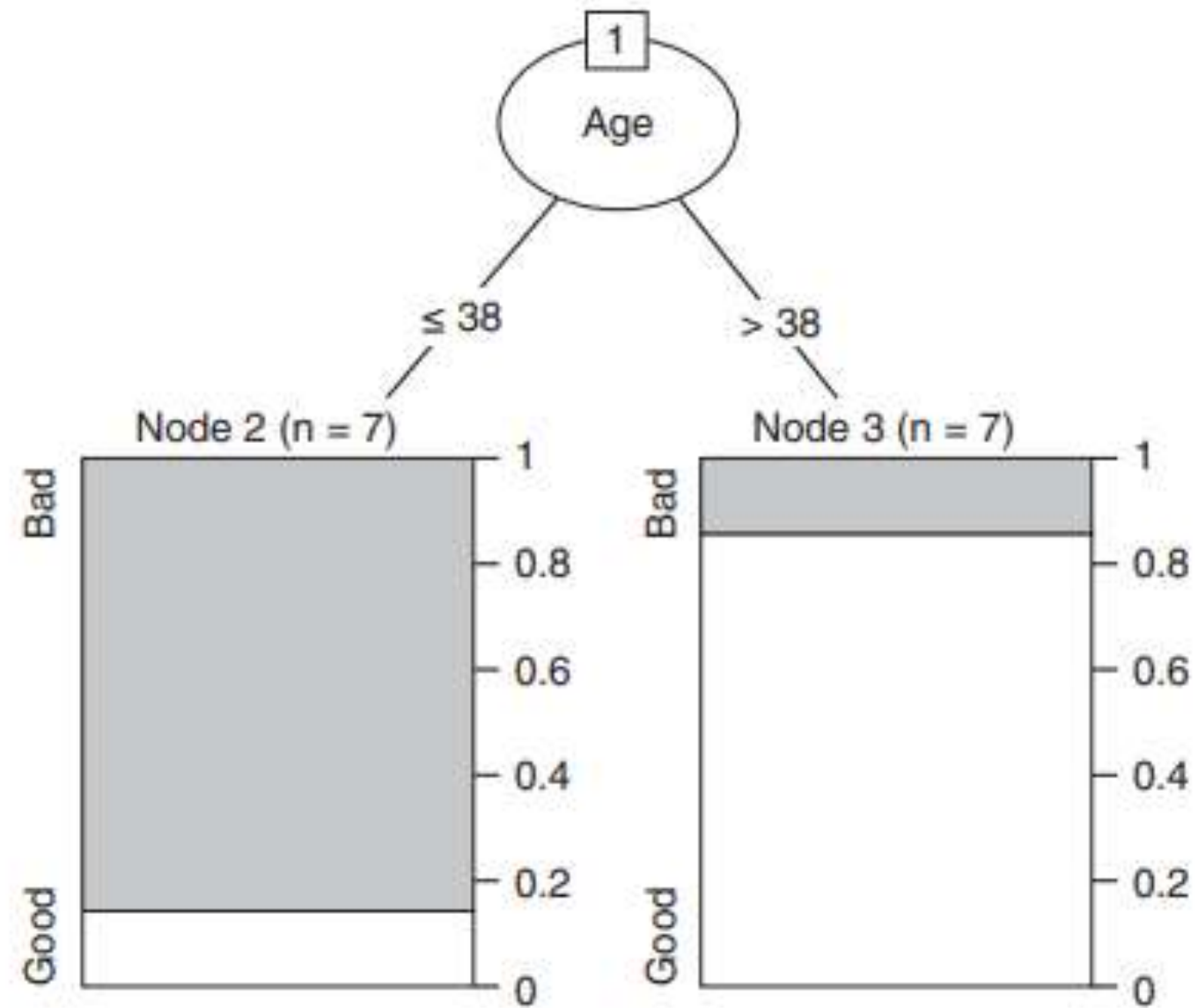
If we apply an algorithm for induction of decision trees to provide an explanation of who, among our contacts, is a good company, we obtain a model, called a decision tree

It can be seen that people older than 38 years are typically better company than those whose age is equal or less than 38 more than 80% of people aged 38 or less are bad company, while more than 80% of people older than 38 are good company.

This model could be used to predict whether a new contact is or not a good company. It would be enough to know the age of that new contact



# Examples of Data



A prediction model to classify someone as either good or bad company.



# Assessment 1



To design your own Algorithm and Model for any real world problem





# References



1. João Moreira, Andre Carvalho, Tomás Horvath – “A General Introduction to Data Analytics” – Wiley -2018

**Thank You**