UNSUPERVISED LEARNING – CASE STUDY: INCOME ANALYSIS ON CENSUS DATA

The data set (Adult data set) contains a sample of data (48842 examples) extracted in 1994 from the Census Bureau database.

Available fields are:

- age – Age of the individual
- workclass – Type of employment (Government, Private, Military, etc.)
- education – Highest level of education achieved for that individual
- education-num – Highest level of education in numerical form
- marital-status – Never-married, Separated, Widowed, etc.
- occupation – Exec-managerial, Farming-fishing, etc.
- relationship – Family relationship value (Husband, Father, Unmarried, etc.)
- race – Amer-Indian-Eskimo, Asian-Pac-Islander, Black, etc.
- sex – Male, Female
- capital-gain – Capital gains recorded (Income from investment sources, apart from salary)
- capital-loss – Capital loss recorded (Losses from investment sources)
- fnlwgt – The # of units in the target population that the responding unit represents
- hours-per-week – Hours worked per week
- native-country – Country of origin of the individual
- income – Annual income of the individual (small: <= 50K USD, large: otherwise)

Objective: Analyze the elements promoting high annual incomes (above 50K USD) based on census data

UNSUPERVISED LEARNING – CASE STUDY: INCOME ANALYSIS ON CENSUS DATA

Numeric fields must be converted into ordinal attributes.

Let's consider the following mapping:

- age →
  Levels: Young (0-25), Middle-aged (26-45), Senior (46-65) and Old (66+)
- hours-per-week →
  Levels: Part-time (0-25), Full-time (25-40), Over-time (40-60) and Workaholic (60+)
- capital-gain and capital-loss →
  Levels: None (0), Low (0 < median of the values greater zero) and High (>= median of the values greater zero)