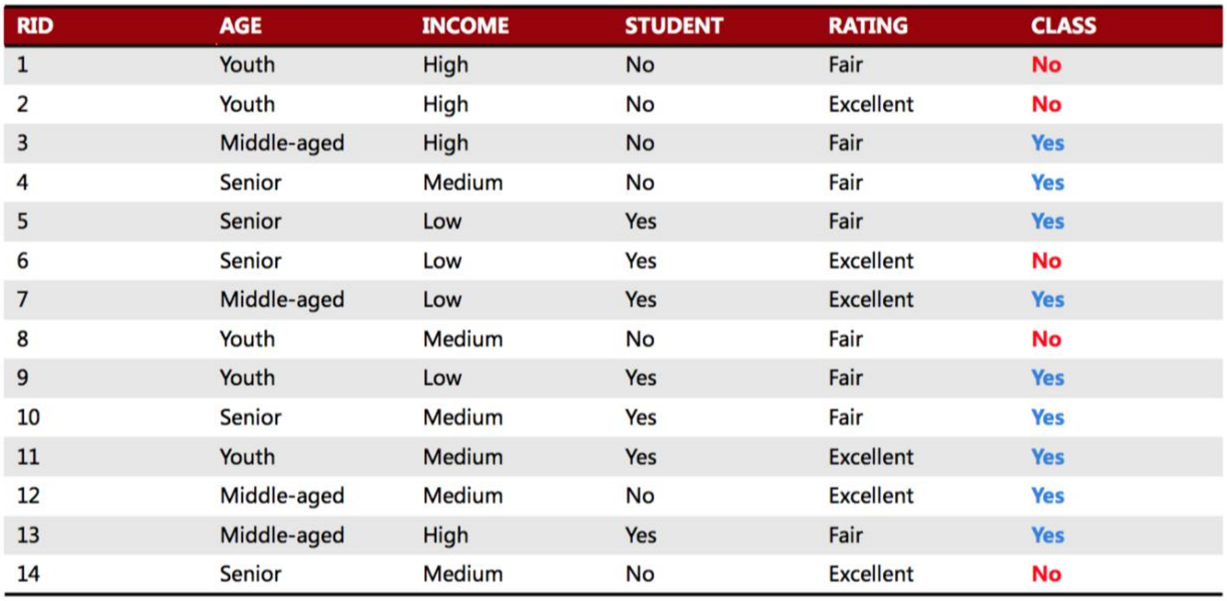
INFS4203/7203 Data Mining

Tutorial 4: Classification and Clustering

1. The provided dataset is a collection of observations that describe if a person decided to buy a computer.
    a. Each observed person is described based on the attributes of **Age, Income, Student status (Student),** and **Credit Rating (Rating)**.
    b. Each person's response to buy a computer was recorded and labeled **Yes** if they bought a computer, and **No** if they did not. The responses to buy a computer are the labels that describe the **Class** category.

   Using the provided dataset, construct a decision tree that will determine if an observed person will be classified **Yes** or **No** to buy a computer. Use the GINI index based splitting criterion to construct the decision tree.



RID	AGE	INCOME	STUDENT	RATING	CLASS
1	Youth	High	No	Fair	No
2	Youth	High	No	Excellent	No
3	Middle-aged	High	No	Fair	Yes
4	Senior	Medium	No	Fair	Yes
5	Senior	Low	Yes	Fair	Yes
6	Senior	Low	Yes	Excellent	No
7	Middle-aged	Low	Yes	Excellent	Yes
8	Youth	Medium	No	Fair	No
9	Youth	Low	Yes	Fair	Yes
10	Senior	Medium	Yes	Fair	Yes
11	Youth	Medium	Yes	Excellent	Yes
12	Middle-aged	Medium	No	Excellent	Yes
13	Middle-aged	High	Yes	Fair	Yes
14	Senior	Medium	No	Excellent	No

2. Suppose the data mining task is to cluster the following measurements of the variable *age* into **three** groups. Age = {18, 22, 25, 42, 27, 43, 33, 35, 56, 28}.

    a. For each initial centroid of {22, 35, 43} and {18, 27, 35}:
        i. Use *k-means* algorithm to show the clustering procedures **step by step;**
        ii. Calculate corresponding **SSE** values.