DataMining Project2 林嘉源 P76051072

1. Introduction to my dataset

My dataset simulate a imaginary country and 100(number of data: **M = 100)** civilian. This project try to predict whether people can live longer than the HALE (health-adjusted life expectancy) of not.

In my simulation, 100 civilian are sampled, and the HALE of this country is set to 76 years old. Each row in my dataset can be viewed as the basic information about a person in that country.

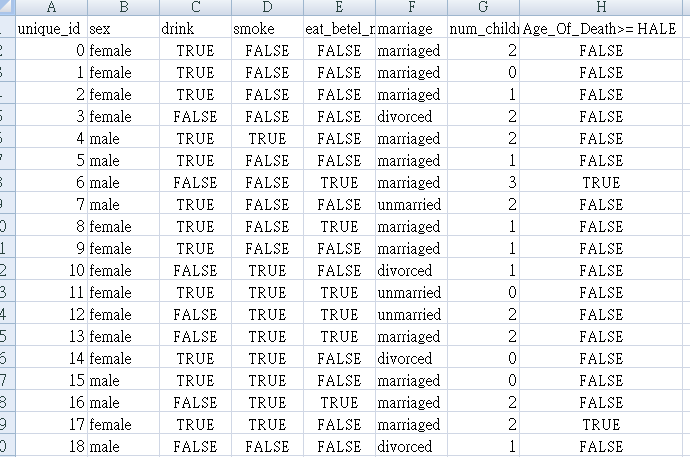
**The features** are listed as following (number of attributes: **K = 6**):

1. **Sex**: female: 0, male: 1
2. **Drink**: drinking alcohol frequently or not: yes/ no
3. **Smoke**: taking cigarettes frequently or not: yes/ no
4. **Eat\_betel\_nuts**: 很常吃檳榔嗎? yes/ no
5. **Marriage**: Marriage: 2, Divorced: 1, Unmarried: 0
6. **Num\_children**: the number of children this person has. Here I don’t want to talk about why an unmarried person has children. Maybe, he/she is an unmarried father/mother.

**My Target**: their age of death is longer than HALE of this country (= 76) or not: yes/ no

Note:

1. the attached **GenerateData.py** can generate my simulated dataset.
2. The split ratio of training and testing are 7:3.

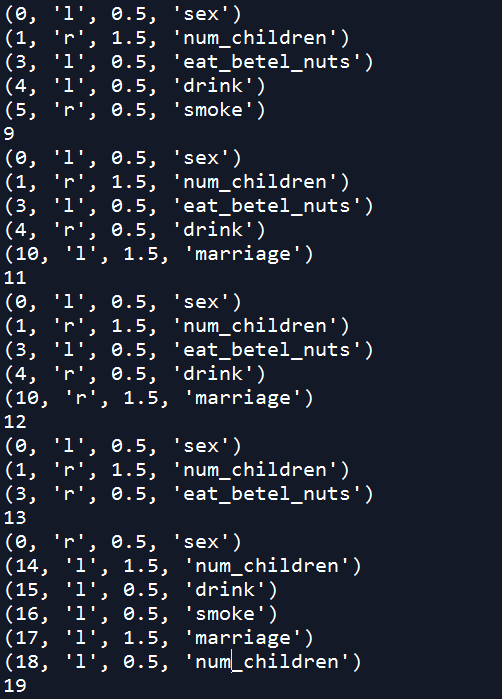
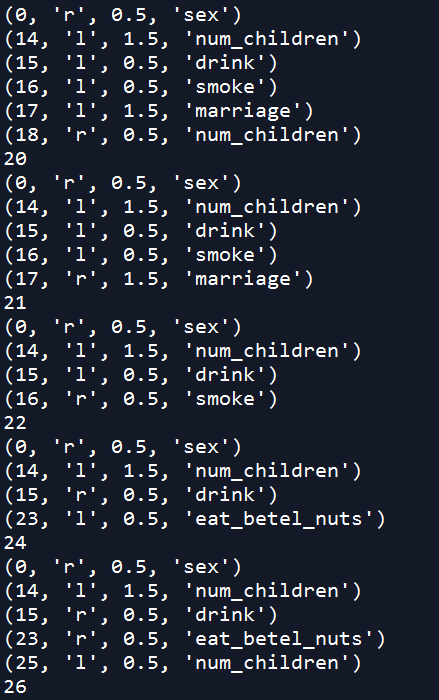
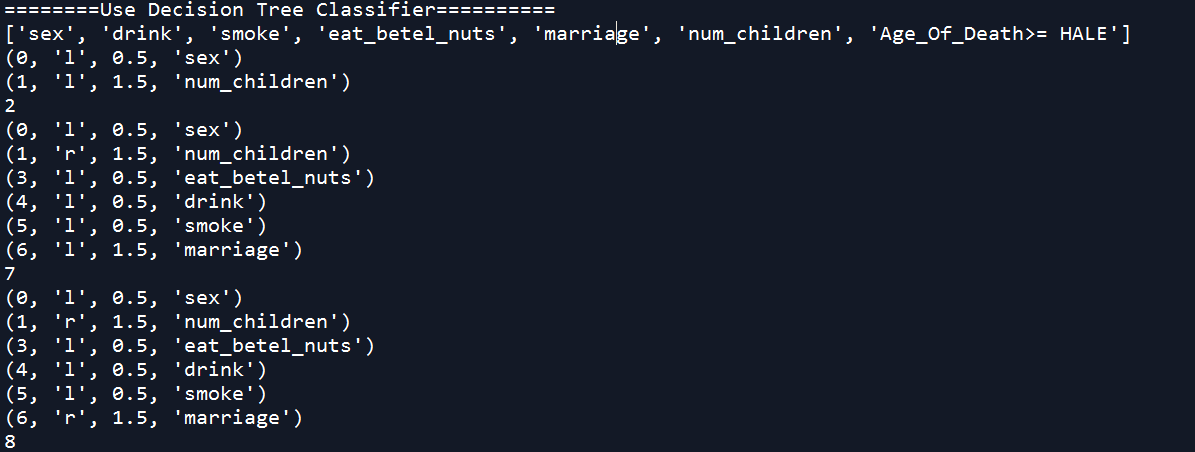


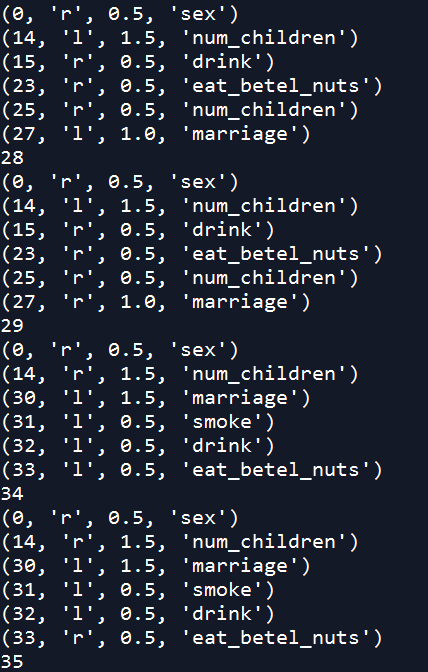
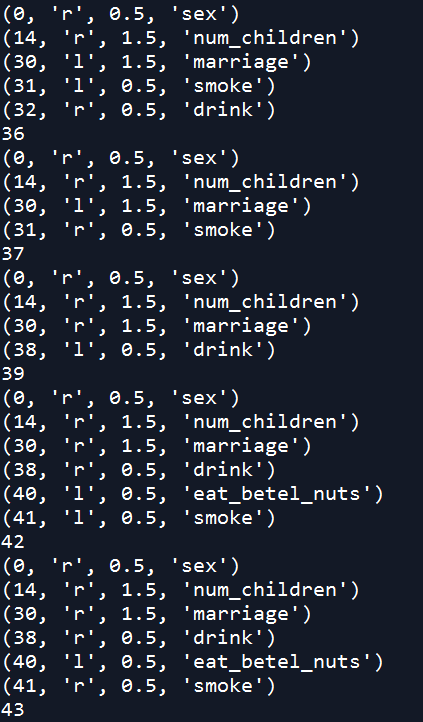
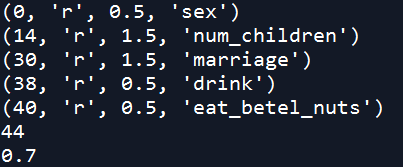
1. Experiment Result

In my experiment, I try decision tree classifier method first, and then apply the K-NN method to predict the result. Both are build in the python *sklearn* toolkit.

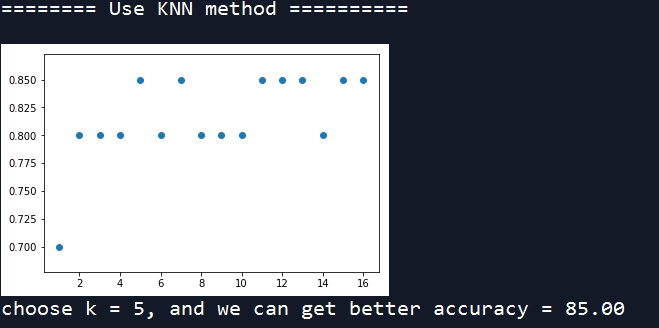
1. DecisionTreeClassifier:

This model classifies each row of data to be false (class-0: age of death < HALE) or true (class-1: age of death >= HALE).





My decision tree classifier accuracy is 0.7

1. KNeighborsClassifier:

My second experiment apply KNN-classifier to classify class-0 (age of death < HALE) or class-1 ( age of death >= HALE), and find that k=5will get a better accuracy (85%).

1. Analysis
2. Decision Tree Structure:

Target: ['sex', 'drink', 'smoke', 'eat\_betel\_nuts', 'marriage', 'num\_children', 'Age\_Of\_Death>= HALE']

Represent True

Represent False

Nodes with red word are viewed as yes (target)

Node-0:

sex<=0.5

Node-14:

Num\_children<=0.5

Node-40:

eat\_betel\_nuts<=0.5

Node-6:

marriage<=0.5

Node-18:

Num\_children<=0.5

Node-33:

eat\_betel\_nuts<=0.5

Node-32:

drink<=0.5

Node-25:

Num\_children<=0.5

Node-17:

marriage<=1.5

Node-10:

marriage<=1.5

Node-5:

smoke<=0.5

Node-16:

smoke<=0.5

Node-31:

smoke<=0.5

Node-38:

drink<=0.5

Node-1:

Num\_children<=1.5

Node-44

Node-43

Node-42

Node-41:

Smoke<=0.5

Node-39

Node-37

Node-36

Node-29

Node-34

Node-35

Node-27:

marriage<=1

Node-26

Node-28

Node-24

Node-30:

marriage<=1.5

Node-23:

Eat\_betel\_nuts<=0.5

Node-22

Node-21

Node-19

Node-20

Node-15:

drink<=0.5

Node-12

Node-3:

eat\_betel\_nuts<0.5

Node-11

Node-9

Node-8

Node-7

Node-2

Node-13

Node-4:

drink<=0.5

1. Discussion

My Right Rule are :



My right rule consists of the following terms:

1. The pow2 of the summation of sex, marriage and num\_children
2. Fix base term: 50
3. The bad habits, such as drink, smoke and eat\_betel\_nuts, may cause negative effects to one’s life expectancy
4. The random term: to make the model more complex

Let’s take a look at my decision tree, and we will find that the dominant term are sex, marriage and num\_children. In other word, the nodes which is closer to the root of the decision tree.

For bad habits like drink, smoke and eat\_betel\_nuts, you can easily discover that the less habits, the more likely to be our target.

My project apply decision tree classifier and knn classifier to test my dataset, and the knn classifier (accuracy = 85%) has better result than decision tree classifier (accuracy= 70%).

In my opinion, my right rule has this term:

That is, the behavior of summation of sex, marriage and num\_children is like the radius of a sphere, so knn classifier is more suitable to predict this model.