CIS 635 Data Mining   Name Sajal Shrestha

Homework 3

<center>Hand in sheet</center>

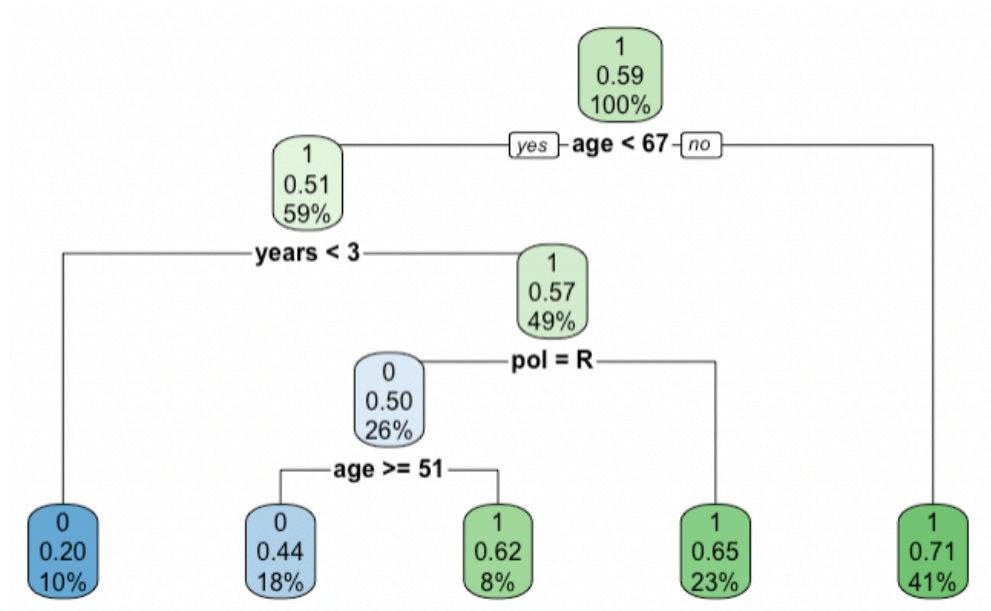Write in or copy and paste the answers for the following:

Part 1

1. gini (nominal)   left   right   both

    a) gender   0.278   0.375   0.3168

    b) ageGroup   0.444   0.408   0.419

2. gini (numeric)

    a) pulse   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0

3. which attribute   test

4. next step Split the attributes based on condition pulse less than 70

Part 2

1. class a {2031, f, 1, 66, 50, n}   b {2032, m, 3, 90 ,81 ,n}   c {2033, m, 2, 75 ,45 ,y}

2. Based on the decision tree, male patients are more likely to have the virus. The male patients in the age group 2 and 3 with test scores below 50 have the virus. Likewise, the virus is also present in male patients in age group two with a pulse rate above 60. On the other hand, male patients in age group three with pulse rates above 60 do not have the virus.

3. After selecting the best split for the attribute, we need to determine the type of attribute and the test condition for a split. The split can be a two-way or multi-way split. For instance, the test condition for the "test" attribute can be "test > 60". The best split is determined by nude purity with the test condition that produces the highest gain. Next, entropy is calculated. If the entropy is 0, It is a leaf node; else, it needs further splitting.

Part 3

1.


```
                                    ┌──────┐
                                    │  1   │
                                    │ 0.59 │
                                    │ 100% │
                                    └──────┘
                              ┌yes─ age < 67 ─no┐
                    ┌──────┐
                    │  1   │
                    │ 0.51 │
                    │ 59%  │
                    └──────┘
              ┌── years < 3 ──┐
                                  ┌──────┐
                                  │  1   │
                                  │ 0.57 │
                                  │ 49%  │
                                  └──────┘
                            ┌── pol = R ──┐
                       ┌──────┐
                       │  0   │
                       │ 0.50 │
                       │ 26%  │
                       └──────┘
                   ┌── age >= 51 ──┐
  ┌──────┐    ┌──────┐    ┌──────┐    ┌──────┐    ┌──────┐
  │  0   │    │  0   │    │  1   │    │  1   │    │  1   │
  │ 0.20 │    │ 0.44 │    │ 0.62 │    │ 0.65 │    │ 0.71 │
  │ 10%  │    │ 18%  │    │ 8%   │    │ 23%  │    │ 41%  │
  └──────┘    └──────┘    └──────┘    └──────┘    └──────┘
```

2. age

3. Senator's party can be crucial in predicting their approval rating. As observed in the decision tree, there is a 65% chance of voter's approval when the senator's party is democrat and more than three years in office for 23% of the data. Likewise, The chance is 62% when the senator's party is republican, and their age is over 51 years for 8% of the population.