Autumn 2012

KDD :
Data Mining project on Hepatitis C effects

Rym Barkaoui
Mohamed Seffar

Presentation :

Hepatitis C is an infectious disease affecting primarily the liver, caused by the hepatitis C virus (HCV).
The infection is often asymptomatic, it means that the patient experiences no symptoms, but chronic infection can lead to scarring of the liver and ultimately to cirrhosis, which is generally apparent after many years. In some cases, those with cirrhosis will go on to develop liver failure, liver cancer or life-threatening esophageal and gastric varices.

Learning the application domain:

Our application domain is medical. It is a study of several cases of hepatitis C. The purpose of applying KDD in this case is knowing from attributes like sex, age, anorexia if a person will die from hepatitis C or not.

Our dataset:

```
@relation 'hepatitis-domain'
@attribute 'AGE' integer
@attribute 'SEX' { 2, 1}
@attribute 'STEROID' { 1, 2}
@attribute 'ANTIVIRALS' { 2, 1}
@attribute 'FATIGUE' { 2, 1}
@attribute 'MALAISE' { 2, 1}
@attribute 'ANOREXIA' { 2, 1}
@attribute 'LIVER_BIG' { 1, 2}
@attribute 'LIVER_FIRM' { 2, 1}
@attribute 'SPLEEN_PALPABLE' { 2, 1}
@attribute 'SPIDERS' { 2, 1}
@attribute 'ASCITES' { 2, 1}
@attribute 'VARICES' { 2, 1}
@attribute 'BILIRUBIN' real
@attribute 'ALK_PHOSPHATE' integer
@attribute 'SGOT' integer
@attribute 'ALBUMIN' real
@attribute 'PROTIME' integer
@attribute 'HISTOLOGY' { 1, 2}
@attribute 'Class' { 'DIE', 'LIVE'}
@data
```

1. 30,2,1,2,2,2,2,1,2,2,2,2,2,1,85,18,4,?,1,'LIVE'
2. 50,1,1,2,1,2,2,1,2,2,2,2,2,0.9,135,42,3.5,?,1,'LIVE'
3. 78,1,2,2,1,2,2,2,2,2,2,2,2,0.7,96,32,4,?,1,'LIVE'
4. 31,1,?,1,2,2,2,2,2,2,2,2,2,0.7,46,52,4,80,1,'LIVE'
5. 34,1,2,2,2,2,2,2,2,2,2,2,2,1,?,200,4,?,1,'LIVE'
6. 34,1,2,2,2,2,2,2,2,2,2,2,2,0.9,95,28,4,75,1,'LIVE'
7. 51,1,1,2,1,2,1,2,2,1,1,2,2,?,?,?,?,?,1,'DIE'
8. 23,1,2,2,2,2,2,2,2,2,2,2,2,1,?,?,?,?,1,'LIVE'
9. 39,1,2,2,1,2,2,2,1,2,2,2,2,0.7,?,48,4.4,?,1,'LIVE'
10. 30,1,2,2,2,2,2,2,2,2,2,2,2,1,?,120,3.9,?,1,'LIVE'
11. 39,1,1,1,2,2,2,1,1,2,2,2,2,1.3,78,30,4.4,85,1,'LIVE'
12. 32,1,2,1,1,2,2,2,1,2,1,2,2,1,59,249,3.7,54,1,'LIVE'
13. 41,1,2,1,1,2,2,2,1,2,2,2,2,0.9,81,60,3.9,52,1,'LIVE'
14. 30,1,2,2,1,2,2,2,1,2,2,2,2,2.2,57,144,4.9,78,1,'LIVE'
15. 47,1,1,1,2,2,2,2,2,2,2,2,2,?,?,60,?,?,1,'LIVE'
16. 38,1,1,2,1,1,1,2,2,2,2,1,2,2,72,89,2.9,46,1,'LIVE'
17. 66,1,2,2,1,2,2,2,2,2,2,2,2,1.2,102,53,4.3,?,1,'LIVE'
18. 40,1,1,2,1,2,2,2,1,2,2,2,2,0.6,62,166,4,63,1,'LIVE'
19. 38,1,2,2,2,2,2,2,2,2,2,2,2,0.7,53,42,4.1,85,2,'LIVE'

```
20.     38,1,1,1,2,2,2,1,1,2,2,2,2,0.7,70,28,4.2,62,1,'LIVE'
21.     22,2,2,1,1,2,2,2,2,2,2,2,2,0.9,48,20,4.2,64,1,'LIVE'
22.     27,1,2,2,1,1,1,1,1,1,1,2,2,1.2,133,98,4.1,39,1,'LIVE'
23.     31,1,2,2,2,2,2,2,2,2,2,2,2,1,85,20,4,100,1,'LIVE'
24.     42,1,2,2,2,2,2,2,2,2,2,2,2,0.9,60,63,4.7,47,1,'LIVE'
25.     25,2,1,1,2,2,2,2,2,2,2,2,2,0.4,45,18,4.3,70,1,'LIVE'
26.     27,1,1,2,1,1,2,2,2,2,2,2,2,0.8,95,46,3.8,100,1,'LIVE'
27.     49,1,1,1,1,1,1,2,1,2,1,2,2,0.6,85,48,3.7,?,1,'LIVE'
28.     58,2,2,2,1,2,2,2,1,2,1,2,2,1.4,175,55,2.7,36,1,'LIVE'
29.     61,1,1,2,1,2,2,1,1,2,2,2,2,1.3,78,25,3.8,100,1,'LIVE'
30.     39,1,1,1,1,1,1,2,2,1,2,2,2,2,2.3,280,98,3.8,40,1,'DIE'
31.     51,1,1,1,1,1,1,2,2,2,2,2,2,2,1,78,58,4.6,52,1,'LIVE'
32.     62,1,1,2,1,1,2,?,?,2,2,2,2,1,?,60,?,?,1,'DIE'
33.     41,2,2,1,1,1,1,1,2,2,2,2,2,2,0.7,81,53,5,74,1,'LIVE'
34.     26,2,1,2,2,2,2,2,1,2,2,2,2,2,0.5,135,29,3.8,60,1,'LIVE'
35.     35,1,2,2,1,2,2,2,2,2,2,2,2,2,0.9,58,92,4.3,73,1,'LIVE'
36.     37,1,2,2,1,2,2,2,2,2,1,2,2,0.6,67,28,4.2,?,1,'DIE'
37.     23,1,2,2,1,1,1,2,2,1,2,2,2,1.3,194,150,4.1,90,1,'LIVE'
38.     20,2,1,2,1,1,1,1,1,1,1,2,2,2.3,150,68,3.9,?,1,'LIVE'
39.     42,1,1,2,2,2,2,2,2,2,2,2,2,1,85,14,4,100,1,'LIVE'
40.     65,1,2,2,1,1,2,2,1,1,1,1,2,0.3,180,53,2.9,74,2,'LIVE'
41.     52,1,1,1,2,2,2,2,2,2,2,2,2,0.7,75,55,4,21,1,'LIVE'
42.     23,1,2,2,2,2,2,?,?,?,?,?,?,4.6,56,16,4.6,?,1,'LIVE'
43.     33,1,2,2,2,2,2,2,2,2,2,2,2,1,46,90,4.4,60,1,'LIVE'
44.     56,1,1,2,1,2,2,2,2,2,2,2,2,0.7,71,18,4.4,100,1,'LIVE'
45.     34,1,2,2,2,2,2,2,2,2,2,2,2,?,?,86,?,?,1,'LIVE'
46.     28,1,2,2,1,1,2,2,2,2,2,2,2,0.7,74,110,4.4,?,1,'LIVE'
47.     37,1,1,2,2,2,2,2,1,2,1,2,2,0.6,80,80,3.8,?,1,'LIVE'
48.     28,2,2,2,1,1,2,2,1,2,2,2,2,1.8,191,420,3.3,46,1,'LIVE'
49.     36,1,1,2,2,2,2,2,2,1,2,2,2,0.8,85,44,4.2,85,1,'LIVE'
50.     38,1,2,1,1,1,1,2,2,2,1,2,2,0.7,125,65,4.2,77,1,'LIVE'
51.     39,1,1,2,2,2,2,2,2,2,2,2,2,0.9,85,60,4,?,1,'LIVE'
52.     39,1,2,2,2,2,2,2,2,2,2,2,2,1,85,20,4,?,1,'LIVE'
53.     44,1,2,2,2,2,2,2,2,2,2,2,2,0.6,110,145,4.4,70,1,'LIVE'
54.     40,1,2,1,1,2,2,2,1,1,2,2,2,1.2,85,31,4,100,1,'LIVE'
55.     30,1,2,2,1,2,2,2,2,2,2,2,2,0.7,50,78,4.2,74,1,'LIVE'
56.     37,1,1,2,1,1,1,2,2,2,2,2,2,0.8,92,59,?,?,1,'LIVE'
57.     34,1,1,2,?,?,?,?,?,?,?,?,?,?,?,?,?,?,1,'LIVE'
58.     30,1,2,1,2,2,2,2,2,2,2,2,2,0.7,52,38,3.9,52,1,'LIVE'
59.     64,1,2,1,1,1,2,1,1,2,2,2,2,1,80,38,4.3,74,1,'LIVE'
60.     45,2,1,2,1,1,2,2,2,1,2,2,2,1,85,75,?,?,1,'LIVE'
61.     37,1,2,2,2,2,2,2,2,2,2,2,2,0.7,26,58,4.5,100,1,'LIVE'
62.     32,1,2,2,2,2,2,2,2,2,2,2,2,0.7,102,64,4,90,1,'LIVE'
63.     32,1,2,2,1,1,1,2,2,2,1,2,1,3.5,215,54,3.4,29,1,'LIVE'
64.     36,1,1,2,2,2,2,1,1,1,2,2,2,0.7,164,44,3.1,41,1,'LIVE'
65.     49,1,2,2,1,1,2,2,2,2,2,2,2,0.8,103,43,3.5,66,1,'LIVE'
66.     27,1,2,2,2,2,2,2,2,2,2,2,2,0.8,?,38,4.2,?,1,'LIVE'
67.     56,1,1,2,2,2,2,2,2,2,2,2,2,0.7,62,33,3,?,1,'LIVE'
68.     57,1,2,2,1,1,1,2,2,2,1,1,2,4.1,?,48,2.6,73,1,'DIE'
69.     39,1,2,2,1,2,2,2,2,2,2,2,2,1,34,15,4,54,1,'LIVE'
70.     44,1,1,2,1,1,2,2,2,2,2,2,2,1.6,68,68,3.7,?,1,'LIVE'
71.     24,1,2,2,2,2,2,2,2,2,2,2,2,0.8,82,39,4.3,?,1,'LIVE'
72.     34,1,1,2,1,1,2,1,1,2,1,2,2,2.8,127,182,?,?,1,'DIE'
73.     51,1,2,2,1,1,1,?,?,?,?,?,?,0.9,76,271,4.4,?,1,'LIVE'
74.     36,1,1,2,1,1,1,2,1,2,2,2,2,1,?,45,4,57,1,'LIVE'
75.     50,1,2,2,2,2,2,2,2,2,2,2,2,1.5,100,100,5.3,?,1,'LIVE'
```

```
76.     32,1,1,1,1,1,2,2,2,2,2,2,2,1,55,45,4.1,56,1,'LIVE'
77.     58,1,2,2,1,2,2,1,1,1,1,2,2,2,167,242,3.3,?,1,'DIE'
78.     34,2,1,1,2,2,2,2,1,2,2,2,2,2,0.6,30,24,4,76,1,'LIVE'
79.     34,1,1,2,1,2,2,1,1,2,1,2,2,1,72,46,4.4,57,1,'LIVE'
80.     28,1,2,2,2,2,2,2,2,2,2,2,2,2,0.7,85,31,4.9,?,1,'LIVE'
81.     23,1,2,2,1,1,1,2,2,2,2,2,2,2,0.8,?,14,4.8,?,1,'LIVE'
82.     36,1,2,2,2,2,2,2,2,2,2,2,2,2,0.7,62,224,4.2,100,1,'LIVE'
83.     30,1,1,2,2,2,2,2,2,2,2,2,2,2,0.7,100,31,4,100,1,'LIVE'
84.     67,2,1,2,1,1,2,2,2,?,?,?,?,1.5,179,69,2.9,?,1,'LIVE'
85.     62,2,2,2,1,1,2,2,1,2,1,2,2,1.3,141,156,3.9,58,1,'LIVE'
86.     28,1,1,2,1,1,1,2,1,2,2,2,2,2,1.6,44,123,4,46,1,'LIVE'
87.     44,1,1,2,1,1,2,2,2,1,2,2,1,0.9,135,55,?,41,2,'DIE'
88.     30,1,2,2,1,1,1,2,1,2,1,1,1,2.5,165,64,2.8,?,2,'DIE'
89.     38,1,1,2,1,1,1,2,1,2,1,1,1,1.2,118,16,2.8,?,2,'DIE'
90.     38,1,1,2,1,1,1,1,1,2,2,2,2,0.6,76,18,4.4,84,2,'LIVE'
91.     50,2,1,2,1,2,2,1,1,1,1,2,2,0.9,230,117,3.4,41,2,'LIVE'
92.     42,1,1,2,1,1,1,2,2,1,1,2,1,4.6,?,55,3.3,?,2,'DIE'
93.     33,1,2,2,2,2,2,?,?,2,2,2,2,1,?,60,4,?,2,'LIVE'
94.     52,1,1,2,2,2,2,2,2,2,2,2,2,2,1.5,?,69,2.9,?,2,'LIVE'
95.     59,1,1,2,1,1,2,2,1,1,1,2,2,1.5,107,157,3.6,38,2,'DIE'
96.     40,1,1,1,1,1,1,1,1,2,2,2,2,0.6,40,69,4.2,67,2,'LIVE'
97.     30,1,1,2,1,1,2,2,1,2,1,2,2,0.8,147,128,3.9,100,2,'LIVE'
98.     44,1,1,2,1,1,2,1,1,2,1,2,2,3,114,65,3.5,?,2,'LIVE'
99.     47,1,2,2,2,2,2,2,2,2,1,2,1,2,84,23,4.2,66,2,'DIE'
100.    60,1,1,2,1,2,2,1,1,1,1,2,2,?,?,40,?,?,2,'LIVE'
101.    48,1,1,2,1,1,2,2,1,2,1,1,1,4.8,123,157,2.7,31,2,'DIE'
102.    22,1,2,2,2,2,2,2,2,2,2,2,2,2,0.7,?,24,?,?,2,'LIVE'
103.    27,1,1,2,1,2,2,2,1,2,2,2,2,2.4,168,227,3,66,2,'LIVE'
104.    51,1,1,2,1,1,1,2,1,1,1,2,1,4.6,215,269,3.9,51,2,'LIVE'
105.    47,1,2,2,1,1,2,2,1,2,2,1,1,1.7,86,20,2.1,46,2,'DIE'
106.    25,1,2,2,2,2,2,2,2,2,2,2,2,2,0.6,?,34,6.4,?,2,'LIVE'
107.    35,1,1,2,1,2,2,?,?,1,1,1,2,1.5,138,58,2.6,?,2,'DIE'
108.    45,1,1,2,1,1,1,2,2,2,2,2,2,2.3,?,648,?,?,2,'LIVE'
109.    54,1,1,1,2,2,2,1,1,2,2,2,2,1,155,225,3.6,67,2,'LIVE'
110.    33,1,1,2,1,1,2,2,2,2,2,1,2,0.7,63,80,3,31,2,'DIE'
111.    7,1,2,2,2,2,2,2,1,1,2,2,2,0.7,256,25,4.2,?,2,'LIVE'
112.    42,1,1,1,1,1,1,2,2,2,2,1,2,2,0.5,62,68,3.8,29,2,'DIE'
113.    52,1,1,2,1,2,2,2,2,2,2,2,2,2,1,85,30,4,?,2,'LIVE'
114.    45,1,1,2,1,2,2,2,2,1,1,2,2,2,1.2,81,65,3,?,1,'LIVE'
115.    36,1,1,2,2,2,2,2,2,2,2,2,2,2,1.1,141,75,3.3,?,2,'LIVE'
116.    69,2,2,2,1,2,2,2,2,2,2,2,2,2,3.2,119,136,?,?,2,'LIVE'
117.    24,1,1,2,1,2,2,2,2,2,2,2,2,2,1,?,34,4.1,?,2,'LIVE'
118.    50,1,2,2,2,2,2,2,2,2,2,2,2,2,1,139,81,3.9,62,2,'LIVE'
119.    61,1,1,2,1,1,2,?,?,2,1,2,2,?,?,?,?,?,2,'DIE'
120.    54,1,2,2,1,2,2,1,1,2,2,2,2,2,3.2,85,28,3.8,?,2,'LIVE'
121.    56,1,1,2,1,1,1,1,1,2,1,2,2,2.9,90,153,4,?,2,'DIE'
122.    20,1,1,2,1,1,1,2,2,2,1,1,2,1,160,118,2.9,23,2,'LIVE'
123.    42,1,2,2,2,2,2,2,2,2,1,2,2,2,1.5,85,40,?,?,2,'LIVE'
124.    37,1,1,2,1,2,2,2,2,2,1,2,2,0.9,?,231,4.3,?,2,'LIVE'
125.    50,1,2,2,2,2,2,2,2,1,1,1,2,2,1,85,75,4,72,2,'LIVE'
126.    34,2,2,2,1,1,1,1,1,2,1,2,2,0.7,70,24,4.1,100,2,'LIVE'
127.    28,1,2,2,1,1,1,?,?,2,1,1,2,1,?,20,4,?,2,'LIVE'
128.    50,1,2,2,1,2,2,2,2,1,1,2,1,1,2.8,155,75,2.4,32,2,'DIE'
129.    54,1,1,2,1,1,1,2,2,2,2,2,1,2,1.2,85,92,3.1,66,2,'LIVE'
130.    57,1,1,2,1,1,1,2,2,2,2,1,1,2,4.6,82,55,3.3,30,2,'DIE'
131.    54,1,2,2,2,2,2,2,2,2,2,2,2,2,1,85,30,4.5,0,2,'LIVE'
```

```
132.    31,1,1,2,1,1,1,2,2,1,2,2,2,8,?,101,2.2,?,2,'DIE'
133.    48,1,2,2,1,1,1,2,1,2,1,2,2,2,158,278,3.8,?,2,'LIVE'
134.    72,1,2,1,1,2,2,2,1,2,2,2,2,1,115,52,3.4,50,2,'LIVE'
135.    38,1,1,2,2,2,2,2,1,2,2,2,2,2,0.4,243,49,3.8,90,2,'DIE'
136.    25,1,2,2,1,2,2,1,1,1,1,1,1,1.3,181,181,4.5,57,2,'LIVE'
137.    51,1,2,2,2,2,2,1,1,2,1,2,2,0.8,?,33,4.5,?,2,'LIVE'
138.    38,1,2,2,2,2,2,2,1,2,1,2,1,1.6,130,140,3.5,56,2,'LIVE'
139.    47,1,2,2,1,1,2,2,1,2,1,1,1,1,166,30,2.6,31,2,'DIE'
140.    45,1,2,1,2,2,2,2,2,2,2,2,2,1.3,85,44,4.2,85,2,'LIVE'
141.    36,1,1,2,1,1,1,1,1,2,1,2,1,1.7,295,60,2.7,?,2,'LIVE'
142.    54,1,1,2,1,1,2,?,?,1,2,1,2,3.9,120,28,3.5,43,2,'DIE'
143.    51,1,2,2,1,2,2,2,1,1,1,2,1,1,?,20,3,63,2,'LIVE'
144.    49,1,1,2,1,1,2,2,2,1,1,2,2,1.4,85,70,3.5,35,2,'DIE'
145.    45,1,2,2,1,1,1,2,2,2,1,1,2,1.9,?,114,2.4,?,2,'DIE'
146.    31,1,1,2,1,2,2,2,2,2,2,2,2,2,1.2,75,173,4.2,54,2,'LIVE'
147.    41,1,2,2,1,2,2,2,1,1,1,2,1,4.2,65,120,3.4,?,2,'DIE'
148.    70,1,1,2,1,1,1,?,?,?,?,?,?,1.7,109,528,2.8,35,2,'DIE'
149.    20,1,1,2,2,2,2,2,?,2,2,2,2,2,0.9,89,152,4,?,2,'LIVE'
150.    36,1,2,2,2,2,2,2,2,2,2,2,2,2,0.6,120,30,4,?,2,'LIVE'
151.    46,1,2,2,1,1,1,2,2,2,1,1,1,7.6,?,242,3.3,50,2,'DIE'
152.    44,1,2,2,1,2,2,2,1,2,2,2,2,2,0.9,126,142,4.3,?,2,'LIVE'
153.    61,1,1,2,1,1,2,1,1,2,1,2,2,2,0.8,75,20,4.1,?,2,'LIVE'
154.    53,2,1,2,1,2,2,2,2,2,1,1,2,1,1.5,81,19,4.1,48,2,'LIVE'
155.    43,1,2,2,1,2,2,2,2,2,1,1,1,2,1.2,100,19,3.1,42,2,'DIE'
```

Handling missing data's using the Mean/mode Imputation:

We replace the missing data with the mean when the attribute is numeric, when not we use the most frequent value in our sampling. These values are available on WEKA.

Our new data set:

@relation 'hepatitis-domain'
@attribute 'AGE' integer
@attribute 'SEX' { 2, 1}
@attribute 'STEROID' { 1, 2}
@attribute 'ANTIVIRALS' { 2, 1}
@attribute 'FATIGUE' { 2, 1}
@attribute 'MALAISE' { 2, 1}
@attribute 'ANOREXIA' { 2, 1}
@attribute 'LIVER_BIG' { 1, 2}
@attribute 'LIVER_FIRM' { 2, 1}
@attribute 'SPLEEN_PALPABLE' { 2, 1}
@attribute 'SPIDERS' { 2, 1}
@attribute 'ASCITES' { 2, 1}
@attribute 'VARICES' { 2, 1}
@attribute 'BILIRUBIN' real
@attribute 'ALK_PHOSPHATE' integer
@attribute 'SGOT' integer
@attribute 'ALBUMIN' real
@attribute 'PROTIME' integer
@attribute 'HISTOLOGY' { 1, 2}
@attribute 'Class' { 'DIE', 'LIVE'}
@data
30,2,1,2,2,2,2,1,2,2,2,2,2,1,85,18,4,61.852,1,'LIVE'
50,1,1,2,1,2,2,1,2,2,2,2,2,0.9,135,42,3.5,1,1,'LIVE'
78,1,2,2,1,2,2,2,2,2,2,2,2,0.7,96,32,4,61.852,1,'LIVE'
31,1,2,1,2,2,2,2,2,2,2,2,2,0.7,46,52,4,80,1,'LIVE'
34,1,2,2,2,2,2,2,2,2,2,2,2,1,105.325,200,4,61.852,1,'LIVE'
34,1,2,2,2,2,2,2,2,2,2,2,2,0.9,95,28,4,75,1,'LIVE'
51,1,1,2,1,2,1,2,2,1,1,2,2,1.428,105.325,85.894,3.817,61.852,1,'DIE'
23,1,2,2,2,2,2,2,2,2,2,2,2,1,105.325,85.894,3.817,61.852,1,'LIVE'
39,1,2,2,1,2,2,2,1,2,2,2,2,0.7,105.325,48,4.4,61.852,1,'LIVE'
30,1,2,2,2,2,2,2,2,2,2,2,2,1,105.325,120,3.9,61.852,1,'LIVE'
39,1,1,1,2,2,2,1,1,2,2,2,2,1.3,78,30,4.4,85,1,'LIVE'
32,1,2,1,1,2,2,2,1,2,1,2,2,1,59,249,3.7,54,1,'LIVE'
41,1,2,1,1,2,2,2,1,2,2,2,2,0.9,81,60,3.9,52,1,'LIVE'
30,1,2,2,1,2,2,2,1,2,2,2,2,2.2,57,144,4.9,78,1,'LIVE'
47,1,1,1,2,2,2,2,2,2,2,2,2,1.428,105.325,60,3.817,61.852,1,'LIVE'
38,1,1,2,1,1,1,2,2,2,2,1,2,2,72,89,2.9,46,1,'LIVE'

66,1,2,2,1,2,2,2,2,2,2,2,2,1.2,102,53,4.3,61.852,1,'LIVE'
40,1,1,2,1,2,2,2,1,2,2,2,2,0.6,62,166,4,63,1,'LIVE'
38,1,2,2,2,2,2,2,2,2,2,2,2,0.7,53,42,4.1,85,2,'LIVE'
38,1,1,1,2,2,2,1,1,2,2,2,2,0.7,70,28,4.2,62,1,'LIVE'
22,2,2,1,1,2,2,2,2,2,2,2,2,0.9,48,20,4.2,64,1,'LIVE'
27,1,2,2,1,1,1,1,1,1,1,2,2,1.2,133,98,4.1,39,1,'LIVE'
31,1,2,2,2,2,2,2,2,2,2,2,2,1,85,20,4,100,1,'LIVE'
42,1,2,2,2,2,2,2,2,2,2,2,2,0.9,60,63,4.7,47,1,'LIVE'
25,2,1,1,2,2,2,2,2,2,2,2,2,0.4,45,18,4.3,70,1,'LIVE'
27,1,1,2,1,1,2,2,2,2,2,2,2,0.8,95,46,3.8,100,1,'LIVE'
49,1,1,1,1,1,1,2,1,2,1,2,2,0.6,85,48,3.7,61.852,1,'LIVE'
58,2,2,2,1,2,2,2,1,2,1,2,2,1.4,175,55,2.7,36,1,'LIVE'
61,1,1,2,1,2,2,1,1,2,2,2,2,1.3,78,25,3.8,100,1,'LIVE'
51,1,1,1,1,1,2,2,2,2,2,2,2,1,78,58,4.6,52,1,'LIVE'
39,1,1,1,1,1,2,2,1,2,2,2,2,2.3,280,98,3.8,40,1,'DIE'
62,1,1,2,1,1,2,2,2,2,2,2,2,1,105.325,60,3.817,61.852,1,'DIE'
41,2,2,1,1,1,1,2,2,2,2,2,2,0.7,81,53,5,74,1,'LIVE'
26,2,1,2,2,2,2,2,1,2,2,2,2,0.5,135,29,3.8,60,1,'LIVE'
35,1,2,2,1,2,2,2,2,2,2,2,2,0.9,58,92,4.3,73,1,'LIVE'
37,1,2,2,1,2,2,2,2,2,1,2,2,0.6,67,28,4.2,61.852,1,'DIE'
23,1,2,2,1,1,1,2,2,1,2,2,2,1.3,194,150,4.1,90,1,'LIVE'
20,2,1,2,1,1,1,1,1,1,1,2,2,2.3,150,68,3.9,61.852,1,'LIVE'
42,1,1,2,2,2,2,2,2,2,2,2,2,1,85,14,4,100,1,'LIVE'
65,1,2,2,1,1,2,2,1,1,1,1,2,0.3,180,53,2.9,74,2,'LIVE'
52,1,1,1,2,2,2,2,2,2,2,2,2,0.7,75,55,4,21,1,'LIVE'
23,1,2,2,2,2,2,2,2,2,2,2,2,4.6,56,16,4.6,61.852,1,'LIVE'
33,1,2,2,2,2,2,2,2,2,2,2,2,1,46,90,4.4,60,1,'LIVE'
56,1,1,2,1,2,2,2,2,2,2,2,2,0.7,71,18,4.4,100,1,'LIVE'
34,1,2,2,2,2,2,2,2,2,2,2,2,1.428,105.325,86,3.817,61.852,1,'LIVE'
28,1,2,2,1,1,2,2,2,2,2,2,2,0.7,74,110,4.4,61.852,1,'LIVE'
37,1,1,2,2,2,2,2,1,2,1,2,2,0.6,80,80,3.8,61.852,1,'LIVE'
28,2,2,2,1,1,2,2,1,2,2,2,2,1.8,191,420,3.3,46,1,'LIVE'
36,1,1,2,2,2,2,2,2,1,2,2,2,0.8,85,44,4.2,85,1,'LIVE'
38,1,2,1,1,1,1,2,2,2,1,2,2,0.7,125,65,4.2,77,1,'LIVE'
39,1,1,2,2,2,2,2,2,2,2,2,2,0.9,85,60,4,61.852,1,'LIVE'
39,1,2,2,2,2,2,2,2,2,2,2,2,1,85,20,4,61.852,1,'LIVE'
44,1,2,2,2,2,2,2,2,2,2,2,2,0.6,110,145,4.4,70,1,'LIVE'
40,1,2,1,1,2,2,2,1,1,2,2,2,1.2,85,31,4,100,1,'LIVE'
30,1,2,2,1,2,2,2,2,2,2,2,2,0.7,50,78,4.2,74,1,'LIVE'
37,1,1,2,1,1,1,2,2,2,2,2,2,0.8,92,59,3.817,61.852,1,'LIVE'
34,1,1,2,1,2,2,2,2,2,2,2,2,1.428,105.325,85.894,3.817,61.852,1,'LIVE'
30,1,2,1,2,2,2,2,2,2,2,2,2,0.7,52,38,3.9,52,1,'LIVE'
64,1,2,1,1,1,2,1,1,2,2,2,2,1,80,38,4.3,74,1,'LIVE'
45,2,1,2,1,1,2,2,2,1,2,2,2,1,85,75,3.817,61.852,1,'LIVE'
37,1,2,2,2,2,2,2,2,2,2,2,2,0.7,26,58,4.5,100,1,'LIVE'
32,1,2,2,2,2,2,2,2,2,2,2,2,0.7,102,64,4,90,1,'LIVE'
32,1,2,2,1,1,1,2,2,2,1,2,1,3.5,215,54,3.4,29,1,'LIVE'
36,1,1,2,2,2,2,1,1,1,2,2,2,0.7,164,44,3.1,41,1,'LIVE'
49,1,2,2,1,1,2,2,2,2,2,2,2,0.8,103,43,3.5,66,1,'LIVE'

27,1,2,2,2,2,2,2,2,2,2,2,2,2,0.8,105.325,38,4.2,61.852,1,'LIVE'
56,1,1,2,2,2,2,2,2,2,2,2,2,2,0.7,62,33,3,61.852,1,'LIVE'
57,1,2,2,1,1,1,2,2,2,1,1,2,4.1,105.325,48,2.6,73,1,'DIE'
39,1,2,2,1,2,2,2,2,2,2,2,2,1,34,15,4,54,1,'LIVE'
44,1,1,2,1,1,2,2,2,2,2,2,2,2,1.6,68,68,3.7,61.852,1,'LIVE'
24,1,2,2,2,2,2,2,2,2,2,2,2,2,0.8,82,39,4.3,61.852,1,'LIVE'
34,1,1,2,1,1,2,1,1,2,1,2,2,2.8,127,182,3.817,61.852,1,'DIE'
51,1,2,2,1,1,1,2,2,2,2,2,2,2,0.9,76,271,4.4,61.852,1,'LIVE'
36,1,1,2,1,1,1,2,1,2,2,2,2,1,105.325,45,4,57,1,'LIVE'
50,1,2,2,2,2,2,2,2,2,2,2,2,2,1.5,100,100,5.3,61.852,1,'LIVE'
32,1,1,1,1,1,2,2,2,2,2,2,2,1,55,45,4.1,56,1,'LIVE'
58,1,2,2,1,2,2,1,1,1,1,2,2,2,167,242,3.3,61.852,1,'DIE'
34,2,1,1,2,2,2,2,1,2,2,2,2,0.6,30,24,4,76,1,'LIVE'
34,1,1,2,1,2,2,1,1,2,1,2,2,1,72,46,4.4,57,1,'LIVE'
28,1,2,2,2,2,2,2,2,2,2,2,2,2,0.7,85,31,4.9,61.852,1,'LIVE'
23,1,2,2,1,1,1,2,2,2,2,2,2,0.8,105.325,14,4.8,61.852,1,'LIVE'
36,1,2,2,2,2,2,2,2,2,2,2,2,2,0.7,62,224,4.2,100,1,'LIVE'
30,1,1,2,2,2,2,2,2,2,2,2,2,2,0.7,100,31,4,100,1,'LIVE'
67,2,1,2,1,1,2,2,2,2,2,2,2,1.5,179,69,2.9,61.852,1,'LIVE'
62,2,2,2,1,1,2,2,1,2,1,2,2,1.3,141,156,3.9,58,1,'LIVE'
28,1,1,2,1,1,1,2,1,2,2,2,2,1.6,44,123,4,46,1,'LIVE'
44,1,1,2,1,1,2,2,2,1,2,2,1,0.9,135,55,61.852,41,2,'DIE'
30,1,2,2,1,1,1,2,1,2,1,1,1,2.5,165,64,2.8,61.852,2,'DIE'
38,1,1,2,1,1,1,2,1,2,1,1,1,1.2,118,16,2.8,61.852,2,'DIE'
38,1,1,2,1,1,1,1,1,2,2,2,2,0.6,76,18,4.4,84,2,'LIVE'
50,2,1,2,1,2,2,1,1,1,1,2,2,0.9,230,117,3.4,41,2,'LIVE'
42,1,1,2,1,1,1,2,2,1,1,2,1,4.6,105.325,55,3.3,61.852,2,'DIE'
33,1,2,2,2,2,2,2,2,2,2,2,2,1,85.894,60,4,61.852,2,'LIVE'
52,1,1,2,2,2,2,2,2,2,2,2,2,2,1.5,105.325,69,2.9,61.852,2,'LIVE'
59,1,1,2,1,1,2,2,1,1,1,2,2,1.5,107,157,3.6,38,2,'DIE'
40,1,1,1,1,1,1,1,1,2,2,2,2,0.6,40,69,4.2,67,2,'LIVE'
30,1,1,2,1,1,2,2,1,2,1,2,2,0.8,147,128,3.9,100,2,'LIVE'
44,1,1,2,1,1,2,1,1,2,1,2,2,3,114,65,3.5,61.852,2,'LIVE'
47,1,2,2,2,2,2,2,2,2,2,1,2,1,2,84,23,4.2,66,2,'DIE'
60,1,1,2,1,2,2,1,1,1,1,2,2,1.428,105.325,40,3.817,61.852,2,'LIVE'
48,1,1,2,1,1,2,2,1,2,1,1,1,4.8,123,157,2.7,31,2,'DIE'
22,1,2,2,2,2,2,2,2,2,2,2,2,2,0.7,105.325,24,3.817,61.852,2,'LIVE'
27,1,1,2,1,2,2,2,1,2,2,2,2,2.4,168,227,3,66,2,'LIVE'
51,1,1,2,1,1,1,2,1,1,1,2,1,4.6,215,269,3.9,51,2,'LIVE'
47,1,2,2,1,1,2,2,1,2,2,1,1,1.7,86,20,2.1,46,2,'DIE'
25,1,2,2,2,2,2,2,2,2,2,2,2,2,0.6,105.325,34,6.4,61.852,2,'LIVE'
35,1,1,2,1,2,2,2,2,1,1,1,2,1.5,138,58,2.6,61.852,2,'DIE'
45,1,1,2,1,1,1,2,2,2,2,2,2,2,2.3,105.325,648,3.817,61.852,2,'LIVE'
54,1,1,1,2,2,2,1,1,2,2,2,2,1,155,225,3.6,67,2,'LIVE'
33,1,1,2,1,1,2,2,2,2,2,1,2,0.7,63,80,3,31,2,'DIE'
7,1,2,2,2,2,2,2,1,1,2,2,2,0.7,256,25,4.2,61.852,2,'LIVE'
42,1,1,1,1,1,2,2,2,2,1,2,2,0.5,62,68,3.8,29,2,'DIE'
52,1,1,2,1,2,2,2,2,2,2,2,2,1,85,30,4,61.852,2,'LIVE'
45,1,1,2,1,2,2,2,1,1,2,2,2,1.2,81,65,3,61.852,1,'LIVE'

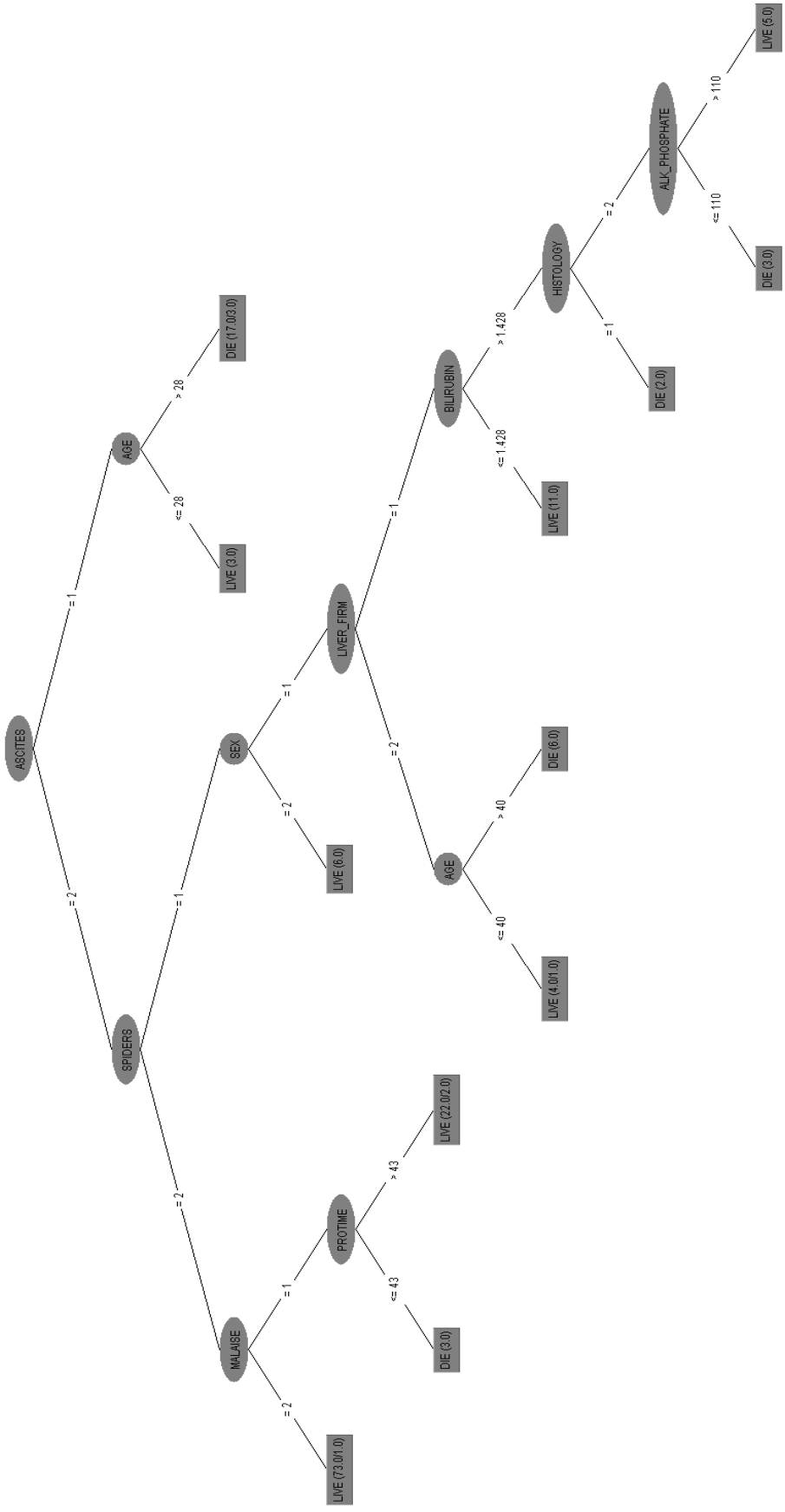
36,1,1,2,2,2,2,2,2,2,2,2,2,1.1,141,75,3.3,61.852,2,'LIVE'
69,2,2,2,1,2,2,2,2,2,2,2,2,3.2,119,136,3.817,61.852,2,'LIVE'
24,1,1,2,1,2,2,2,2,2,2,2,2,1,105.325,34,4.1,61.852,2,'LIVE'
50,1,2,2,2,2,2,2,2,2,2,2,2,1,139,81,3.9,62,2,'LIVE'
61,1,1,2,1,1,2,2,2,2,1,2,2,2,1.428,105.325,3.817,61.852,2,'DIE'
54,1,2,2,1,2,2,1,1,2,2,2,2,3.2,85,28,3.8,61.852,2,'LIVE'
56,1,1,2,1,1,1,1,1,2,1,2,2,2.9,90,153,4,61.852,2,'DIE'
20,1,1,2,1,1,1,2,2,2,1,1,2,1,160,118,2.9,23,2,'LIVE'
42,1,2,2,2,2,2,2,2,2,1,2,2,2,1.5,85,40,3.817,61.852,2,'LIVE'
37,1,1,2,1,2,2,2,2,2,1,2,2,0.9,105.325,231,4.3,61.852,2,'LIVE'
50,1,2,2,2,2,2,2,2,1,1,1,2,2,1,85,75,4,72,2,'LIVE'
34,2,2,2,1,1,1,1,1,2,1,2,2,0.7,70,24,4.1,100,2,'LIVE'
28,1,2,2,1,1,1,2,2,2,1,1,2,1,105.325,20,4,61.852,2,'LIVE'
50,1,2,2,1,2,2,2,1,1,2,1,1,2.8,155,75,2.4,32,2,'DIE'
54,1,1,2,1,1,2,2,2,2,2,1,2,1.2,85,92,3.1,66,2,'LIVE'
57,1,1,2,1,1,2,2,2,2,1,1,2,4.6,82,55,3.3,30,2,'DIE'
54,1,2,2,2,2,2,2,2,2,2,2,2,1,85,30,4.5,0,2,'LIVE'
31,1,1,2,1,1,1,2,2,1,2,2,2,8,105.325,101,2.2,61.852,2,'DIE'
48,1,2,2,1,1,1,2,1,2,1,2,2,2,158,278,3.8,61.852,2,'LIVE'
72,1,2,1,1,2,2,2,1,2,2,2,2,1,115,52,3.4,50,2,'LIVE'
38,1,1,2,2,2,2,2,1,2,2,2,2,0.4,243,49,3.8,90,2,'DIE'
25,1,2,2,1,2,2,1,1,1,1,1,1,1.3,181,181,4.5,57,2,'LIVE'
51,1,2,2,2,2,2,1,1,2,1,2,2,0.8,105.325,33,4.5,61.852,2,'LIVE'
38,1,2,2,2,2,2,2,1,2,1,2,1,1.6,130,140,3.5,56,2,'LIVE'
47,1,2,2,1,1,2,2,1,2,1,1,1,1,166,30,2.6,31,2,'DIE'
45,1,2,1,2,2,2,2,2,2,2,2,2,1.3,85,44,4.2,85,2,'LIVE'
36,1,1,2,1,1,1,1,1,2,1,2,1,1.7,295,60,2.7,61.852,2,'LIVE'
54,1,1,2,1,1,2,2,2,1,2,1,2,3.9,120,28,3.5,43,2,'DIE'
51,1,2,2,1,2,2,2,1,1,1,2,1,1,105.325,20,3,63,2,'LIVE'
49,1,1,2,1,1,2,2,2,1,1,2,2,1.4,85,70,3.5,35,2,'DIE'
45,1,2,2,1,1,1,2,2,2,1,1,2,1.9,1.428,114,2.4,61.852,2,'DIE'
31,1,1,2,1,2,2,2,2,2,2,2,2,1.2,75,173,4.2,54,2,'LIVE'
41,1,2,2,1,2,2,2,1,1,1,2,1,4.2,65,120,3.4,61.852,2,'DIE'
70,1,1,2,1,1,1,2,2,2,2,2,2,1.7,109,528,2.8,35,2,'DIE'
20,1,1,2,2,2,2,2,2,2,2,2,2,0.9,89,152,4,61.852,2,'LIVE'
36,1,2,2,2,2,2,2,2,2,2,2,2,0.6,120,30,4,61.852,2,'LIVE'
46,1,2,2,1,1,1,2,2,2,1,1,1,7.6,105.325,242,3.3,50,2,'DIE'
44,1,2,2,1,2,2,2,1,2,2,2,2,0.9,126,142,4.3,61.852,2,'LIVE'
61,1,1,2,1,1,2,1,1,2,1,2,2,0.8,75,20,4.1,61.852,2,'LIVE'
53,2,1,2,1,2,2,2,2,1,1,2,1,1.5,81,19,4.1,48,2,'LIVE'
43,1,2,2,1,2,2,2,2,1,1,1,2,1.2,100,19,3.1,42,2,'DIE'

Data mining goals:

Based on a set of data, we want to know if a new patient is likely to die or to live based on his "parameters".
Algorithm Decision:
Considering our goal and our type of data we chose to apply the J48 algorithm that would eventually give us a decision tree that is very useful and is quite easy to interpret. Decision trees can represent diverse types of data. The simplest and most familiar is numerical data... For example, here we have many medical attributes that are either integer or real and that perfectly fit the algorithm we decided to use.
Once in WEKA we put the class, which is whether the patient is going to die or to live and then we have the following results.

```
Number of Leaves  :     12

Size of the tree :      23


Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         115               74.1935 %
Incorrectly Classified Instances        40               25.8065 %
Kappa statistic                          0.2124
Mean absolute error                      0.2706
Root mean squared error                  0.4812
Relative absolute error                 81.9501 %
Root relative squared error            118.8407 %
Coverage of cases (0.95 level)          85.8065 %
Mean rel. region size (0.95 level)      70.9677 %
Total Number of Instances              155

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC    ROC Area  PRC Area  Class
                 0.375    0.163    0.375      0.375   0.375      0.212  0.594     0.313     DIE
                 0.837    0.625    0.837      0.837   0.837      0.212  0.594     0.801     LIVE
Weighted Avg.    0.742    0.53     0.742      0.742   0.742      0.212  0.594     0.7

=== Confusion Matrix ===

   a   b   <-- classified as
  12  20 |   a = DIE
  20 103 |   b = LIVE
```

We notice that 115 out of 165 instances were correctly classified (74.1935%) the rate were higher while using missing data which may be explained by the fact that in the medical domain we never use the mean/mode method to replace data. We obtain the following tree.
The tree works the following way: if ABSCITES value is 1 and the patient is more than 28 then he will most likely die (17/20 which is 85% that is why we have only 75% on the instances that are well classified because some examples do not respect the tree for example here 3 patients that had ABSCITES =1 and had more than 28 years didn't die.


ASCITES
= 1
= 2
AGE
> 28
<= 28
DIE (17.0/3.0)
LIVE (3.0)
SPIDERS
= 1
= 2
SEX
= 1
= 2
LIVER_FIRM
= 1
= 2
LIVE (6.0)
BILIRUBIN
> 1.428
<= 1.428
AGE
> 40
<= 40
HISTOLOGY
= 2
= 1
LIVE (11.0)
DIE (6.0)
LIVE (4.0/1.0)
DIE (2.0)
ALK_PHOSPHATE
> 110
<= 110
LIVE (5.0)
DIE (3.0)
MALAISE
= 1
= 2
PROTIME
> 43
<= 43
LIVE (22.0/2.0)
DIE (3.0)
LIVE (73.0/1.0)

Evaluation :

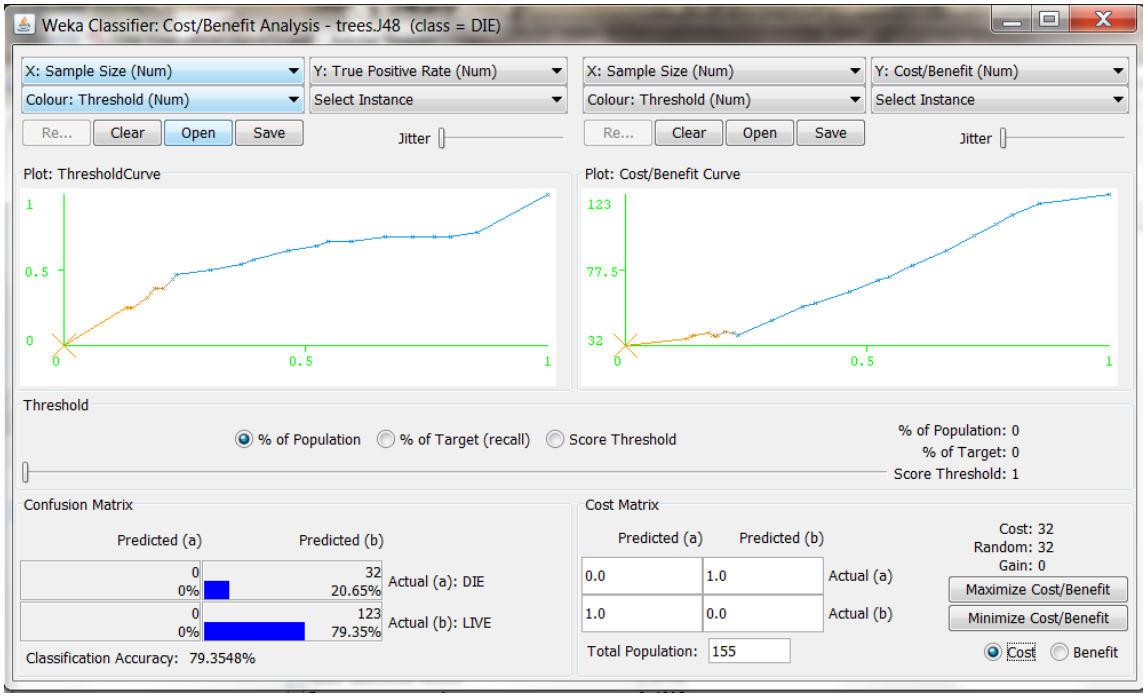
Once we obtain this tree we should ask ourselves how accurate is it?

Classification accuracy:

In our study we have 155 instances, out of these 155 instances, 110 were correctly classified which is about 74% of correctly classified instances, therefore we have 40 incorrectly classified instances (26%). This data along with many can be found in WEKA after running your algorithm here is the result:

```
Time taken to build model: 0.06 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         115               74.1935 %
Incorrectly Classified Instances        40               25.8065 %
Kappa statistic                          0.2124
Mean absolute error                      0.2706
Root mean squared error                  0.4812
Relative absolute error                 81.9501 %
Root relative squared error            118.8407 %
Coverage of cases (0.95 level)          85.8065 %
Mean rel. region size (0.95 level)      70.9677 %
Total Number of Instances              155

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
               0.375     0.163     0.375       0.375   0.375       0.212    0.594     0.313     DIE
               0.837     0.625     0.837       0.837   0.837       0.212    0.594     0.801     LIVE
Weighted Avg.  0.742     0.53      0.742       0.742   0.742       0.212    0.594     0.7

=== Confusion Matrix ===

   a   b   <-- classified as
  12  20 |   a = DIE
  20 103 |   b = LIVE
```

Cost & Benefit:



We can see here that we have unbalanced data, indeed our classes have very unequal frequency: DIE around 20% and live around 80%. How can we evaluate our classifier method?

A good approach since we have two classes would be to build a balanced set of data and to rerun the algorithm on this new set of data.
We randomly remove instances of the majority class, until the two classes are more or less balanced, after removing those data here is the new set:

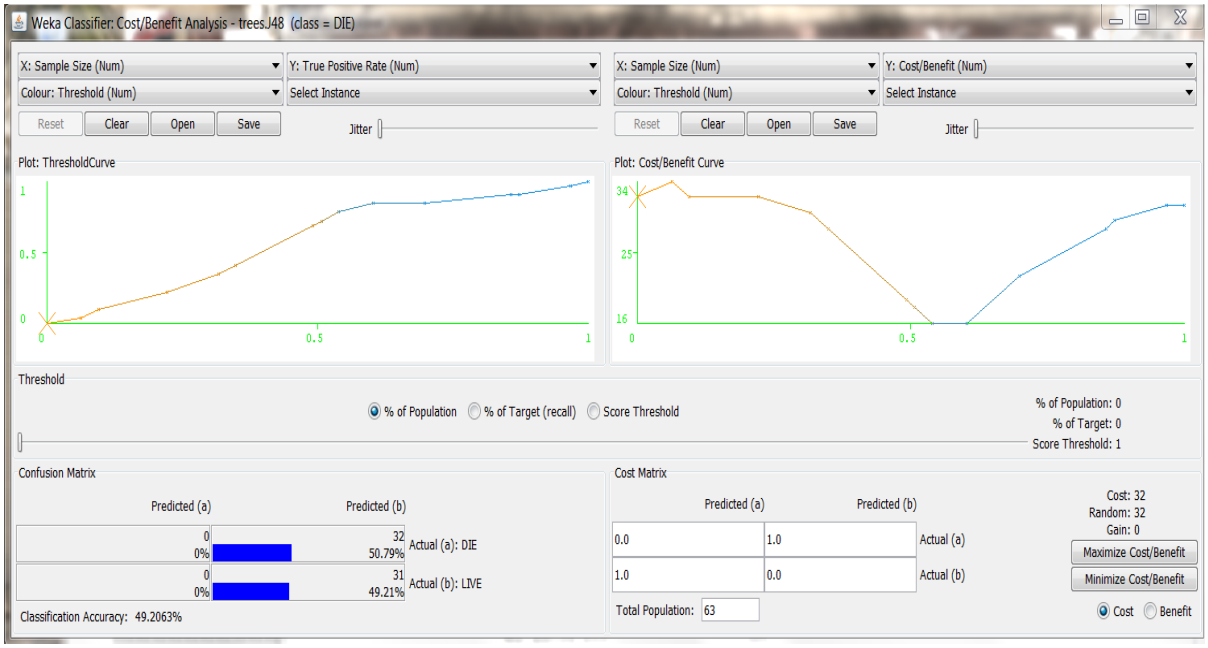
```
@relation 'hepatitis-domain'
@attribute 'AGE' integer
@attribute 'SEX' { 2, 1}
@attribute 'STEROID' { 1, 2}
@attribute 'ANTIVIRALS' { 2, 1}
@attribute 'FATIGUE' { 2, 1}
@attribute 'MALAISE' { 2, 1}
@attribute 'ANOREXIA' { 2, 1}
@attribute 'LIVER_BIG' { 1, 2}
@attribute 'LIVER_FIRM' { 2, 1}
@attribute 'SPLEEN_PALPABLE' { 2, 1}
@attribute 'SPIDERS' { 2, 1}
@attribute 'ASCITES' { 2, 1}
@attribute 'VARICES' { 2, 1}
@attribute 'BILIRUBIN' real
@attribute 'ALK_PHOSPHATE' integer
@attribute 'SGOT' integer
@attribute 'ALBUMIN' real
@attribute 'PROTIME' integer
@attribute 'HISTOLOGY' { 1, 2}
@attribute 'Class' { 'DIE', 'LIVE'}
```

@data
30,2,1,2,2,2,2,1,2,2,2,2,2,1,85,18,4,61.852,1,'LIVE'
34,1,2,2,2,2,2,2,2,2,2,2,2,1,105.325,200,4,61.852,1,'LIVE'
34,1,2,2,2,2,2,2,2,2,2,2,2,0.9,95,28,4,75,1,'LIVE'
51,1,1,2,1,2,1,2,2,1,1,2,2,1.428,105.325,85.894,3.817,61.852,1,'DIE'
39,1,1,1,1,1,2,2,1,2,2,2,2,2.3,280,98,3.8,40,1,'DIE'
62,1,1,2,1,1,2,2,2,2,2,2,2,1,105.325,60,3.817,61.852,1,'DIE'
37,1,2,2,1,2,2,2,2,2,1,2,2,0.6,67,28,4.2,61.852,1,'DIE'
57,1,2,2,1,1,1,2,2,2,1,1,2,4.1,105.325,48,2.6,73,1,'DIE'
39,1,2,2,1,2,2,2,2,2,2,2,2,1,34,15,4,54,1,'LIVE'
44,1,1,2,1,1,2,2,2,2,2,2,2,1.6,68,68,3.7,61.852,1,'LIVE'
24,1,2,2,2,2,2,2,2,2,2,2,2,0.8,82,39,4.3,61.852,1,'LIVE'
34,1,1,2,1,1,2,1,1,2,1,2,2,2.8,127,182,3.817,61.852,1,'DIE'
51,1,2,2,1,1,1,2,2,2,2,2,2,0.9,76,271,4.4,61.852,1,'LIVE'
36,1,1,2,1,1,1,2,1,2,2,2,2,1,105.325,45,4,57,1,'LIVE'
50,1,2,2,2,2,2,2,2,2,2,2,2,1.5,100,100,5.3,61.852,1,'LIVE'
32,1,1,1,1,1,2,2,2,2,2,2,2,1,55,45,4.1,56,1,'LIVE'
58,1,2,2,1,2,2,1,1,1,1,2,2,2,167,242,3.3,61.852,1,'DIE'
34,2,1,1,2,2,2,2,1,2,2,2,2,0.6,30,24,4,76,1,'LIVE'
34,1,1,2,1,2,2,1,1,2,1,2,2,1,72,46,4.4,57,1,'LIVE'
28,1,2,2,2,2,2,2,2,2,2,2,2,0.7,85,31,4.9,61.852,1,'LIVE'
44,1,1,2,1,1,2,2,2,1,2,2,1,0.9,135,55,61.852,41,2,'DIE'
30,1,2,2,1,1,1,2,1,2,1,1,1,2.5,165,64,2.8,61.852,2,'DIE'
38,1,1,2,1,1,1,2,1,2,1,1,1,1.2,118,16,2.8,61.852,2,'DIE'
42,1,1,2,1,1,1,2,2,1,1,2,1,4.6,105.325,55,3.3,61.852,2,'DIE'
59,1,1,2,1,1,2,2,1,1,1,2,2,1.5,107,157,3.6,38,2,'DIE'
47,1,2,2,2,2,2,2,2,2,1,2,1,2,84,23,4.2,66,2,'DIE'
60,1,1,2,1,2,2,1,1,1,1,2,2,1.428,105.325,40,3.817,61.852,2,'LIVE'
48,1,1,2,1,1,2,2,1,2,1,1,1,4.8,123,157,2.7,31,2,'DIE'
47,1,2,2,1,1,2,2,1,2,2,1,1,1.7,86,20,2.1,46,2,'DIE'
25,1,2,2,2,2,2,2,2,2,2,2,2,0.6,105.325,34,6.4,61.852,2,'LIVE'
35,1,1,2,1,2,2,2,2,1,1,1,2,1.5,138,58,2.6,61.852,2,'DIE'
33,1,1,2,1,1,2,2,2,2,2,1,2,0.7,63,80,3,31,2,'DIE'
42,1,1,1,1,1,2,2,2,2,1,2,2,0.5,62,68,3.8,29,2,'DIE'
61,1,1,2,1,1,2,2,2,2,1,2,2,2,1.428,105.325,3.817,61.852,2,'DIE'
54,1,2,2,1,2,2,1,1,2,2,2,2,3.2,85,28,3.8,61.852,2,'LIVE'
56,1,1,2,1,1,1,1,1,2,1,2,2,2.9,90,153,4,61.852,2,'DIE'
20,1,1,2,1,1,1,2,2,2,1,1,2,1,160,118,2.9,23,2,'LIVE'
42,1,2,2,2,2,2,2,2,2,1,2,2,2,1.5,85,40,3.817,61.852,2,'LIVE'
37,1,1,2,1,2,2,2,2,2,1,2,2,0.9,105.325,231,4.3,61.852,2,'LIVE'
50,1,2,2,2,2,2,2,2,1,1,1,2,2,1,85,75,4,72,2,'LIVE'
34,2,2,2,1,1,1,1,1,2,1,2,2,0.7,70,24,4.1,100,2,'LIVE'
28,1,2,2,1,1,1,2,2,2,1,1,2,1,105.325,20,4,61.852,2,'LIVE'
50,1,2,2,1,2,2,2,2,1,1,2,1,1,2.8,155,75,2.4,32,2,'DIE'
54,1,1,2,1,1,2,2,2,2,2,1,2,1.2,85,92,3.1,66,2,'LIVE'
57,1,1,2,1,1,2,2,2,2,2,1,1,2,4.6,82,55,3.3,30,2,'DIE'
54,1,2,2,2,2,2,2,2,2,2,2,2,1,85,30,4.5,0,2,'LIVE'
31,1,1,2,1,1,1,2,2,1,2,2,2,8,105.325,101,2.2,61.852,2,'DIE'
48,1,2,2,1,1,1,2,1,2,1,2,2,2,158,278,3.8,61.852,2,'LIVE'

38,1,1,2,2,2,2,2,1,2,2,2,2,0.4,243,49,3.8,90,2,'DIE'
47,1,2,2,1,1,2,2,1,2,1,1,1,1,166,30,2.6,31,2,'DIE'
45,1,2,1,2,2,2,2,2,2,2,2,2,1.3,85,44,4.2,85,2,'LIVE'
36,1,1,2,1,1,1,1,1,2,1,2,1,1.7,295,60,2.7,61.852,2,'LIVE'
54,1,1,2,1,1,2,2,2,1,2,1,2,3.9,120,28,3.5,43,2,'DIE'
51,1,2,2,1,2,2,2,1,1,1,2,1,1,105.325,20,3,63,2,'LIVE'
49,1,1,2,1,1,2,2,2,1,1,2,2,1.4,85,70,3.5,35,2,'DIE'
45,1,2,2,1,1,1,2,2,2,1,1,2,1.9,1.428,114,2.4,61.852,2,'DIE'
31,1,1,2,1,2,2,2,2,2,2,2,2,1.2,75,173,4.2,54,2,'LIVE'
41,1,2,2,1,2,2,2,1,1,1,2,1,4.2,65,120,3.4,61.852,2,'DIE'
70,1,1,2,1,1,1,2,2,2,2,2,2,1.7,109,528,2.8,35,2,'DIE'
46,1,2,2,1,1,1,2,2,2,1,1,1,7.6,105.325,242,3.3,50,2,'DIE'
44,1,2,2,1,2,2,2,1,2,2,2,2,0.9,126,142,4.3,61.852,2,'LIVE'
61,1,1,2,1,1,2,1,1,2,1,2,2,0.8,75,20,4.1,61.852,2,'LIVE'
43,1,2,2,1,2,2,2,2,1,1,1,2,1.2,100,19,3.1,42,2,'DIE'

There are 31 instances of the class “live” and 32 of the class “die” these data’s are therefore way more balanced:

```
Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          45               71.4286 %
Incorrectly Classified Instances        18               28.5714 %
Kappa statistic                          0.4284
Mean absolute error                      0.3489
Root mean squared error                  0.4915
Relative absolute error                 69.7374 %
Root relative squared error             98.2402 %
Coverage of cases (0.95 level)          90.4762 %
Mean rel. region size (0.95 level)      90.4762 %
Total Number of Instances               63

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC    ROC Area  PRC Area  Class
                 0.719    0.29     0.719      0.719   0.719      0.428  0.663     0.601     DIE
                 0.71     0.281    0.71       0.71    0.71       0.428  0.663     0.655     LIVE
Weighted Avg.    0.714    0.286    0.714      0.714   0.714      0.428  0.663     0.627

=== Confusion Matrix ===

  a  b   <-- classified as
 23  9 |  a = DIE
  9 22 |  b = LIVE
```



We can see that our class is clearly more balanced nevertheless this has been done at the expense of classification accuracy that went form nearly 80% to 50%, which makes our work way less accurate therefore makes the tree less interesting to study.

Conclusion:

In this project we went from a set of data that contained patients that had the hepatitis C, first we learned about the application domain: the disease, how serious it was… Then we worked on our data, we handled the missing data with a mean mode method. Then we worked on the mining part, our data mining goal was starting from our set of data predict if a new input: a new patient was likely to die or not, we chose the decision tree technique with as a class attribute whether the subject was going to die or to live and we obtained some very coherent results. Finally, we tried to analyze the results, the accuracy of the mining thanks to some WEKA evaluation tools.