中国科学技术大学
University of Science and Technology of China

Lab 4

乳腺癌数据集分析

姓名：高茂航

学号：PB22061161

日期：2024.5.16

Report 4

1 Task1

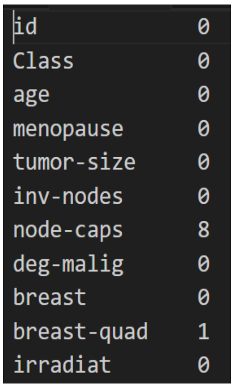
1.1 Algorithm Description

使用 $df.isnull().sum()$ 判断有缺失项的列，使用 $df.dropna()$ 删除有缺失项的行，使用 $df.value\_counts()$ 显示一列中的数据分布，快速找出异常项后用 $df.replace()$ 替换为正常值，等函数对数据集进行处理。

对于 Q3，遍历 excel 文件每行的 Description 列的所有元素放到一个列表中，以把 df 的值替换为索引，但是处理时发现 deg-malig 是整数类型而非字符串，因此一开始报错了，后来将这列转为字符串类型后就可以了。此外，还碰到一个问题是 node-caps 和 irradiat 这两列的值都是 yes/no，一开始在替换时会把 node-caps 的索引赋给 irradiat，因此在替换到 irradiat 时要特别找到第二个 yes/no 的索引。

1.2 Results

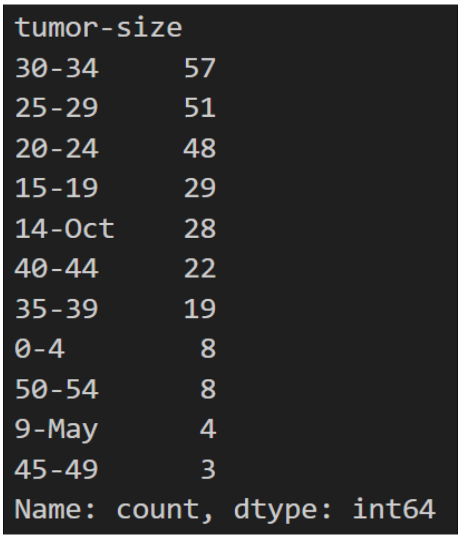
1.2.1 Q1



```
id             0
Class          0
age            0
menopause      0
tumor-size     0
inv-nodes      0
node-caps      8
deg-malig      0
breast         0
breast-quad    1
irradiat       0
```

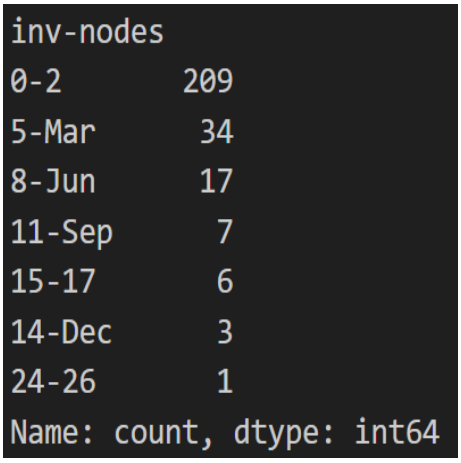
从结果可看出只有 node-caps 列和 breast-quad 列有缺失项，因此删除了这两行。

1.2.2 Q2



```
tumor-size
30-34     57
25-29     51
20-24     48
15-19     29
14-Oct    28
40-44     22
35-39     19
0-4        8
50-54      8
9-May      4
45-49      3
Name: count, dtype: int64
```

可看出 tumor-size 列有异常值’14-Oct’、’9-May’，需要将其替换为正常值。



```
inv-nodes
0-2        209
5-Mar       34
8-Jun       17
11-Sep       7
15-17        6
14-Dec       3
24-26        1
Name: count, dtype: int64
```

可看出 inv-nodes 列有异常值’5-Mar’、’8-Jun’、’11-Sep’、’14-Dec’，需要将其替换为正常值。

1.2.3 Q3

$\{0 :' Class = no - recurrence - events', 1 :' Class = recurrence - events', 2 :' age = 10 - 19', 3 :' age = 20 - 29', 4 :' age = 30 - 39', 5 :' age = 40 - 49', 6 :' age = 50 - 59', 7 :' age = 60 - 69', 8 :' age = 70 - 79', 9 :' age = 80 - 89', 10 :' age = 90 - 99', 11 :' menopause = lt40', 12 :' menopause = ge40', 13 :' menopause = premeno', 14 :' tumor - size = 0 - 4', 15 :' tumor - size = 5 - 9', 16 :' tumor - size = 10 - 14', 17 :' tumor - size = 15 - 19', 18 :' tumor - size = 20 - 24', 19 :' tumor - size = 25 - 29', 20 :' tumor - size = 30 - 34', 21 :' tumor - size = 35 - 39', 22 :' tumor - size = 40 - 44', 23 :' tumor - size = 45 - 49', 24 :' tumor - size = 50 - 54', 25 :' tumor - size = 55 - 59', 26 :' inv - nodes = 0 - 2', 27 :' inv - nodes = 3 - 5', 28 :' inv - nodes = 6 - 8', 29 :' inv - nodes = 9 - 11', 30 :' inv - nodes = 12 - 14', 31 :' inv - nodes = 15 - 17', 32 :' inv - nodes = 18 - 20', 33 :' inv - nodes = 21 - 23', 34 :' inv - nodes = 24 - 26', 35 :' inv - nodes = 27 - 29', 36 :' inv - nodes = 30 - 32', 37 :' inv - nodes = 33 - 35', 38 :' inv - nodes = 36 - 39', 39 :' node - caps = yes', 40 :' node - caps = no', 41 :' deg - malig = 1', 42 :' deg - malig = 2', 43 :' deg - malig = 3', 44 :' breast = left', 45 :' breast = right', 46 :' breast - quad = left_up', 47 :' breast - quad = left_low', 48 :' breast - quad = right_up', 49 :' breast - quad = right_low', 50 :' breast - quad = central', 51 :' irradiat = yes', 52 :' irradiat = no'\}$

2 Task2

2.1 Algorithm Description

使用 Aprior 算法依次算出各个频繁项集，并根据关联规则算出置信度和提升度。具体过程是：先定义计算项集支持度的函数，然后定义产生候选项集的函数，最后写产生频繁项集的函数，依次产生各频繁项集。产生关联规则的思路主要是先找出所有含 0 的频繁项，在此基础上产生对应项去掉 0 后的集合以及只含有 0 的集合，就能计算它们的支持度，进而计算置信度和提升度。

本实验中再次使用了匿名函数，加深了对其的理解。

2.2 Results

2.2.1 Q1

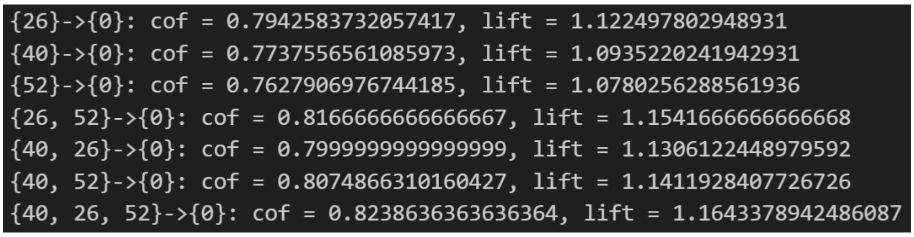
Frequent 1-itemsets: [{0}, {12}, {13}, {26}, {40}, {42}, {44}, {45}, {52}]

Frequent 2-itemsets: [{40, 13}, {26, 13}, {40, 44}, {26, 44}, {0, 26}, {26, 52}, {0, 40}, {0, 52}, {40, 52}, {44, 52}, {52, 13}, {40, 26}]

Frequent 3-itemsets: [{0, 26, 52}, {0, 26, 40}, {0, 40, 52}, {40, 26, 52}]

Frequent 4-itemsets: [{0, 40, 26, 52}]

2.2.2 Q2



```
{26}->{0}: cof = 0.7942583732057417, lift = 1.122497802948931
{40}->{0}: cof = 0.7737556561085973, lift = 1.0935220241942931
{52}->{0}: cof = 0.7627906976744185, lift = 1.0780256288561936
{26, 52}->{0}: cof = 0.8166666666666667, lift = 1.1541666666666668
{40, 26}->{0}: cof = 0.7999999999999999, lift = 1.1306122448979592
{40, 52}->{0}: cof = 0.8074866310160427, lift = 1.1411928407726726
{40, 26, 52}->{0}: cof = 0.8238636363636364, lift = 1.1643378942486087
```

2.3 Conclusion

无结节帽、受侵淋巴结数目范围在 0-2 且未进行放疗与不复发强关联，我们可以认为出这部分患者不易复发乳腺癌。