Lung Cancer Prediction

Data preprocessing

1. 对源文件进行修改

- 录入错误
- 格式错误

2. 计算横轴缺失比例

缺失__以下，AB组剩余的样本数量

MissingRate	A_left	B_left
0	0	0
0.05	2519	2481
0.1	3789	3738
0.15	3962	3796
0.2	3986	3796
0.25	4004	3796
0.3	4011	3796
0.35	4012	3796
0.4	4030	3796
0.45	4069	3798
0.5	4123	3817
0.55	4182	4085
0.6	4185	4087
0.65	4187	4087
0.7	4188	4087
0.75	4189	4087
0.8	4193	4087
0.85	4213	4090
0.9	4222	4095
0.95	4642	4099
1	4642	4099

最终保留横轴缺失20%以下的样本

3. 计算纵轴缺失比例

	A	B	Delete
A=Case, B=Control	0	0	
Origin	0	0	
A=male B=female	0	0	
Age	0	0	
Pattern	0.218992	1	
TNM	0.752799	1	Y
Smoke	0.025409	0	
RBC	0.108958	0.003414	
HGB	0.10702	0.003414	
HCT	0.10745	0.003414	
MCV	0.107666	0.003414	
MCH	0.107666	0.003414	
MCHC	0.107666	0.003414	
RDW-CV	0.107666	0.003414	
RDW-SD	0.107881	0.003414	
PLT	0.10745	0.003414	
PCT	0.480835	1	Y
MPV	0.483204	1	Y
P-LCR	0.483635	1	Y
PDW	0.483204	1	Y
WBC	0.108958	0.003414	
NEUT%	0.107881	0.003658	
LYM%	0.11025	0.003414	
MONO%	0.110896	0.003901	
EO%	0.112834	0.006096	
BASO%	0.120155	0.030237	
AC%	0.994186	1	Y
NEUT	0.11068	0.003658	
LYMPH	0.109604	0.003414	
MONO	0.111111	0.003901	

	A	B	Delete
EO	0.113695	0.006096	
BASO	0.118432	0.030237	
TBIL	0.125969	0.073397	
DBIL	0.126615	0.073397	
IBIL	0.1264	0.073397	
ALT	0.125108	0.073397	
AST	0.125108	0.073397	
AST:ALT	0.125754	0.073397	
TP	0.125754	0.073397	
ALB	0.125538	0.073397	
GLO	0.125754	0.073397	
A/G	0.125754	0.073397	
GLU	0.130706	0.073397	
BUN	0.125538	0.073397	
Cr	0.125754	0.073397	
Cystatin-C	0.255168	0.073397	
UA	0.126184	0.073397	
TG	0.140181	0.073397	
TCH	0.140181	0.073397	
HDL-C	0.140181	0.073641	
LDL-C	0.140181	0.073397	
ALP	0.127476	0.073397	
GGT	0.125969	0.073397	
CK	0.143842	0.08754	
LDH-L	0.143196	0.08754	
HBDH	0.144272	0.08754	
Na	0.164944	0.90612	Y
K	0.164729	0.90612	Y
Cl	0.164944	0.90612	Y
CO2Cp	0.165805	0.90612	Y

	A	B	Delete
AG	0.166236	0.90612	Y
Beta-HB	0.61391	0.906364	Y
Ca	0.2177	0.90612	Y
Mg	0.216839	0.90612	Y
P	0.218346	0.90612	Y
T-CEA	0.314815	0.003414	
T-CA199	0.628338	0.183858	
T-CA125	0.572567	0.65228	
T-CYFRA21-1	0.382429	0.448183	
T-NSE	0.419251	0.546696	
T-ESR	0.726529	1	Y
T-CRP	0.795004	1	Y

最终纵轴缺失15%以上的特征被移除

4. 空值的处理

按照是否患病，性别和年龄分组后，计算中位数填充。

5.异常值处理

样本中多是远超正常范围多个量级的极大异常值。

3倍上下四分位确定上下界，超出范围的值被替换为空值。

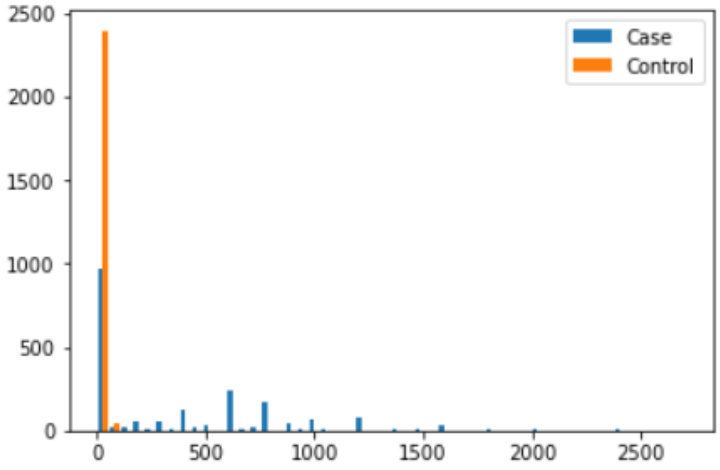
处理后，对各个特征的AB两组做t-test，MONO%与AST未呈现显著差异。

Feature Selection

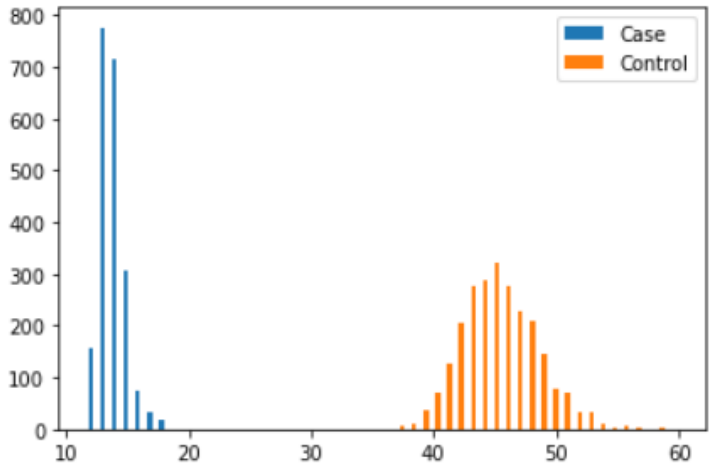
1.卡方检验与互信息计算出特征与是否患病的相关性

一些排名较高的特征：

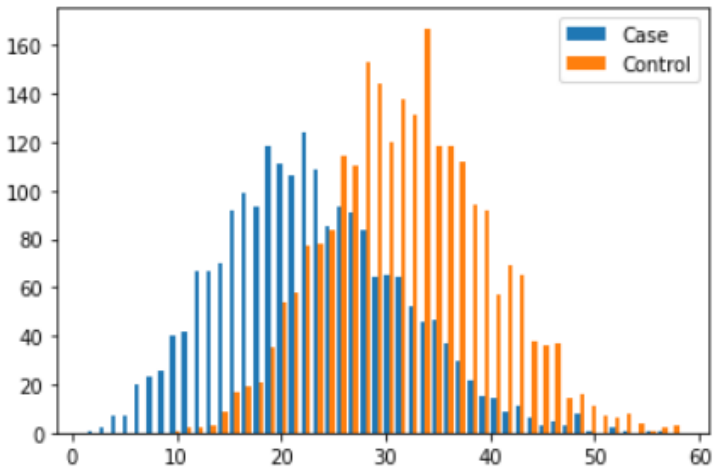
```
Smoke:  Case    Control
均值: 384.18330134357007  3.2531697341513293
方差: 469.46421181305783  13.476115258521245
Max:    2700.0  100.0
上四分位:   600.0   0.0
中位数:    200.0   0.0
下四分位:   0.0 0.0
Min:    0.0 0.0
偏度: 1.264115074068413   5.401275907448914
峰度: 1.4719263979967194  32.17128796478296
P值: 3.3767764625706554e-301
```




```
RDW-SD: Case    Control
均值: 13.898464491362779  45.27218813905938
方差: 1.0157509150604598  3.126233229229989
Max:    18.3    59.9
上四分位:   14.4    47.3
中位数:    13.8    45.0
下四分位:   13.2    43.1
Min:    11.7    36.5
偏度: 1.0421144946220882  0.4602331343785443
峰度: 1.8716885588793812  0.5421042937797984
P值: 0.0
```

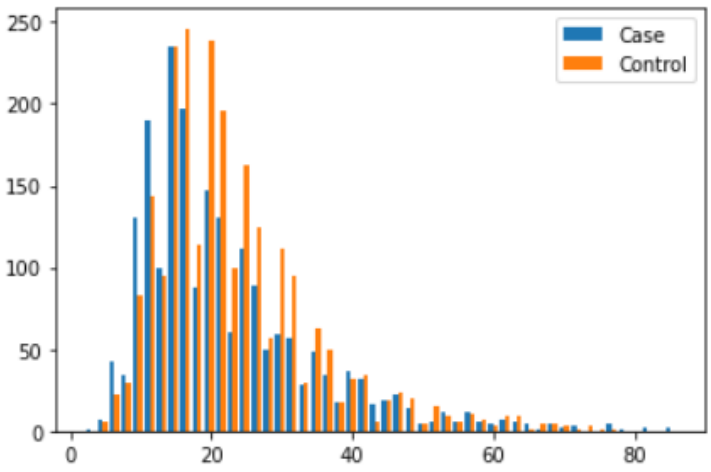


```
LYM%:   Case    Control
均值: 22.90830134357008   32.072269938650315
方差: 8.585865372757414   7.728787026177298
Max:    56.3    58.2
上四分位:   28.4    37.2
中位数:    22.2    31.9
下四分位:   16.8    26.6
Min:    1.3 9.5
偏度: 0.4090114870262905  0.1940300449555222
峰度: 0.07594256487085138 -0.09803415977083496
P值: 7.676649082903106e-272
```

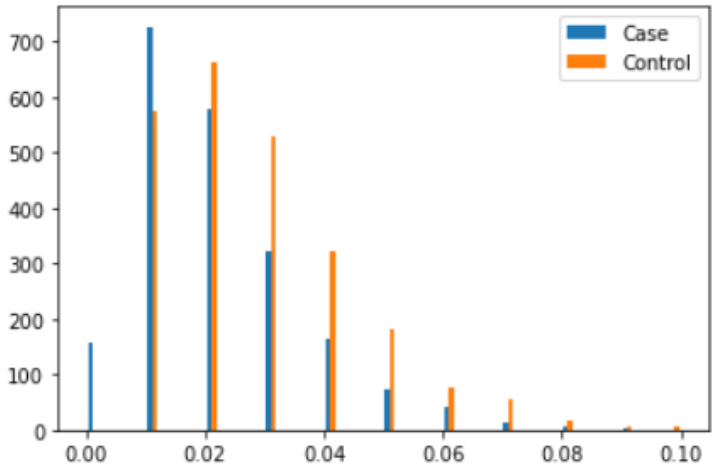


一些排名较低的特征：

```
ALT:    Case    Control
均值: 22.318618042226486  23.23885480572597
方差: 12.592365182257184  11.312198649917471
Max:    86.0    76.0
上四分位:   27.25   28.0
中位数:    19.0    21.0
下四分位:   14.0    15.0
Min:    2.0 5.0
偏度: 1.5903226609956571  1.4954990653765328
峰度: 3.065045140742479   2.8455378142431766
P值: 0.0096330943469289
```

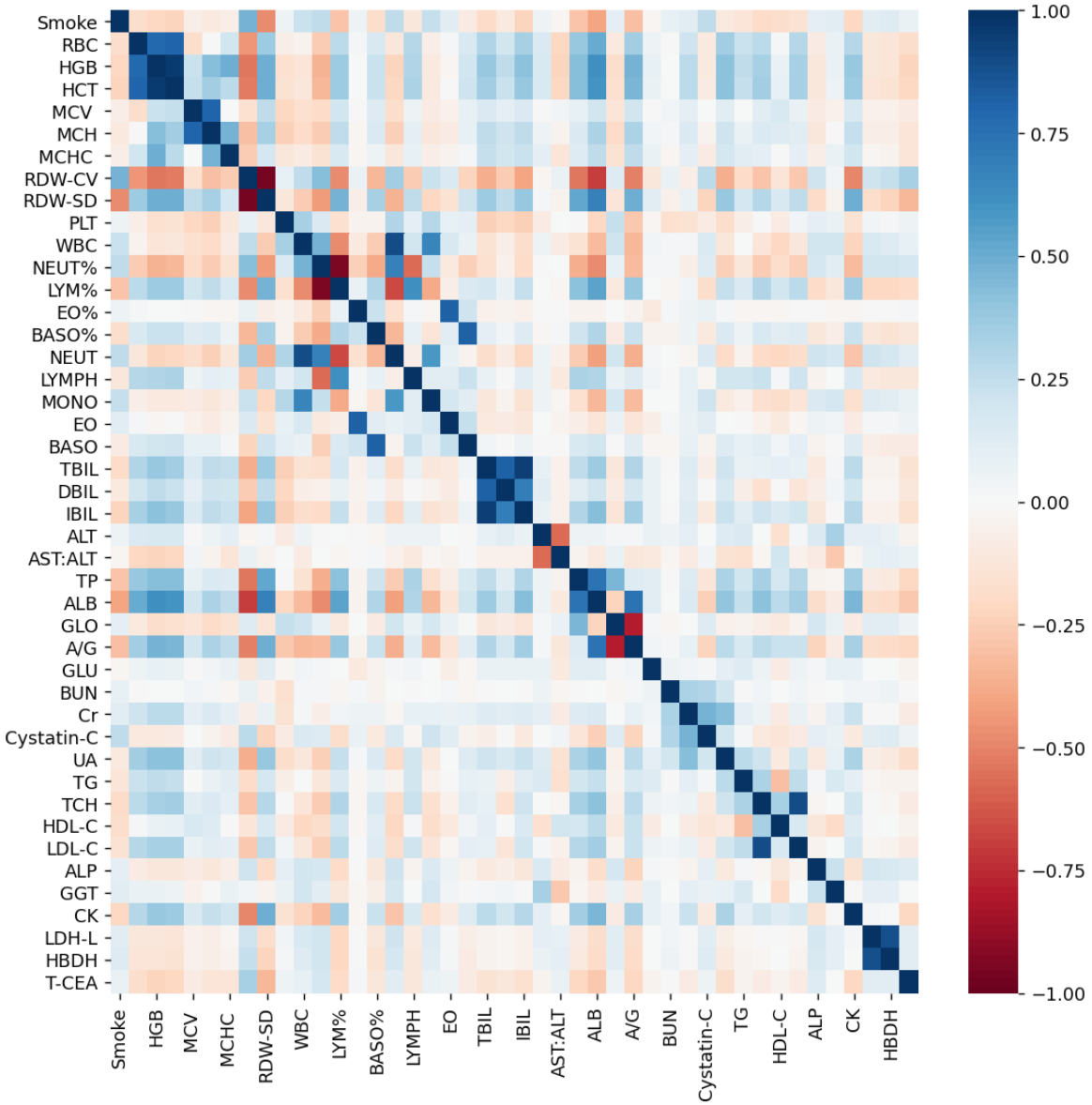


```
BASO:    Case     Control
均值: 0.02046065259117098 0.02805112474437657
方差: 0.014293369768990465     0.01641309765755486
Max:     0.09     0.1
上四分位:    0.03     0.04
中位数:     0.02     0.02
下四分位:    0.01     0.02
Min:     0.0 0.01
偏度: 1.1679952279304644   1.1323442108399424
峰度: 1.755305055029742    1.4176149399821463
P值: 3.999310517347449e-59
```



预计的处理方式：确定一个阈值，移除排名较低的特征，多次迭代

2. 相关性分析



预计的处理方式：在相关性高的特征中选择一个，进行后续分析

3. RFE与正则化

在具体训练时进一步筛选

问题：

特征的筛选，是否有技巧？

划分训练集时，使用原始数据还是处理后的数据？