Andrew Ruder
HW 2
CISC 5790
Due: 3/4/2024

1a. (K-Value, Accuracy scores) without Normalization

(1, 0.7522816166883963),
(5, 0.7548891786179922),
(11, 0.7648848326814428),
(21, 0.7466318991742721),
(41, 0.7522816166883963),
(61, 0.7375054324206867),
(81, 0.7266405910473707),
(101, 0.7288135593220338),
(201, 0.7314211212516297),
(401, 0.7196870925684485)

1b. (K-Value, Accuracy scores) with Z-Score Normalization

(1, 0.8231203824424164),
(5, 0.8322468491960018),
(11, 0.8748370273794003),
(21, 0.8709256844850065),
(41, 0.8704910908300739),
(61, 0.8700564971751412),
(81, 0.8696219035202086),
(101, 0.8639721860060843),
(201, 0.8461538461538461),
(401, 0.8144285093437635)

1c.

```
t 1   ['spam', 'spam', 'spam', 'spam', 'spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam']
t 2   ['spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'no-spam', 'no-spam', 'no-spam']
t 3   ['spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam']
t 4   ['spam', 'spam', 'spam', 'spam', 'no-spam', 'no-spam', 'spam', 'spam', 'spam', 'spam']
t 5   ['spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam']
t 6   ['spam', 'spam', 'spam', 'no-spam', 'no-spam', 'spam', 'spam', 'spam', 'spam', 'spam']
t 7   ['spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam']
t 8   ['spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam']
t 9   ['spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam']
t 10   ['spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam']
t 11   ['spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam']
t 12   ['spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam']
t 13   ['spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam']
t 14   ['no-spam', 'spam', 'spam', 'spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam']
t 15   ['spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam']
t 16   ['spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam']
t 17   ['spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam']
t 18   ['spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'no-spam', 'no-spam', 'no-spam']
t 19   ['spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam']
t 20   ['no-spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam']
t 21   ['spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam']
t 22   ['spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam']
t 23   ['spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam']
t 24   ['no-spam', 'no-spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam']
t 25   ['spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam']
```

```
t 26  ['spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam']
t 27  ['spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam']
t 28  ['spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam']
t 29  ['spam', 'spam', 'spam', 'no-spam', 'spam', 'spam', 'spam', 'spam', 'no-spam', 'no-spam']
t 30  ['spam', 'spam', 'spam', 'spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam']
t 31  ['spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam']
t 32  ['spam', 'spam', 'spam', 'spam', 'no-spam', 'spam', 'spam', 'spam', 'no-spam', 'no-spam']
t 33  ['spam', 'spam', 'spam', 'spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam']
t 34  ['spam', 'spam', 'no-spam', 'spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam']
t 35  ['spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam']
t 36  ['spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam']
t 37  ['spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam']
t 38  ['spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam']
t 39  ['spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam']
t 40  ['no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam']
t 41  ['no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam']
t 42  ['spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'no-spam', 'no-spam']
t 43  ['no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam']
t 44  ['no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam', 'no-spam']
t 45  ['spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam']
t 46  ['spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam']
t 47  ['spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam']
t 48  ['spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam']
t 49  ['spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam']
t 50  ['spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam', 'spam']
```

1d. Based on the results from a and b, we can conclude that Z-score normalization increases the prediction accuracy of the KNN model

Top H:4 L:6 Entropy $= -\frac{4}{10}\log(\frac{4}{10}) - \frac{6}{10}(\log\frac{6}{10}) = \boxed{0.971}$

Education Info gain: High Shool , College

High Shool	College
H:1 L:4	H: 3 L:2
$= -\frac{1}{5}\log(\frac{1}{5}) - \frac{4}{5}\log(\frac{4}{5})$	$= -\frac{3}{5}\log(\frac{3}{5}) - \frac{2}{5}\log(\frac{2}{5})$
$= 0.722$	$= 0.971$

Entopy $= \frac{5}{10}(0.722) + \frac{5}{10}(0.971)$

$= 0.8465$ ✓

Edu Infogain $= 0.971 - 0.8465 = \boxed{0.1245}$

Career Info gain: Management Service

Management	Service
H: 3 L:2	H:1 L:4
$= -\frac{3}{5}\log(\frac{3}{5}) - \frac{2}{5}\log(\frac{2}{5})$	$= -\frac{1}{5}\log(\frac{1}{5}) - \frac{4}{5}\log(\frac{4}{5})$
$= 0.971$	$= 0.722$

Entropy $= \frac{5}{10}(0.971) + \frac{5}{10}(0.722) = 0.8465$

career Infogain $= 0.971 - 0.8465 = \boxed{0.1245}$

Exp Info gain: <3 3-10 >10

<3	3-10	>10
H:1 L:2	H:1 L:2	H:2 L:2
$= -\frac{1}{3}\log(\frac{1}{3}) - \frac{2}{3}\log(\frac{2}{3})$	$= -\frac{1}{3}\log(\frac{1}{3}) - \frac{2}{3}\log(\frac{2}{3})$	$-\frac{2}{4}\log(\frac{2}{4}) - \frac{2}{4}\log(\frac{2}{4})$
$= 0.918$	0.918	1

Entropy $= \frac{3}{10}(0.918) + \frac{3}{10}(0.918) + \frac{4}{10}(1) = 0.9508$

Exp Info gain $= 0.971 - 9508 = \boxed{0.0202}$

Edu Highschool H:1 L:4 Entropy $= -\frac{1}{5}\log(\frac{1}{5}) - \frac{4}{5}\log(\frac{4}{5}) = \boxed{0.722}$

Career Info gain management service

management	service
H:1 L:2	H:0 L:2
$= -\frac{1}{3}\log(\frac{1}{3}) - \frac{2}{3}\log(\frac{2}{3})$	$-0\log(0) - 1\log(1)$
$= 0.918$	$= 0$

Entropy $= \frac{3}{5}(0.918) = 0.5508$

Edu Info gain $= 0.722 - 0.5508 = \boxed{0.1712}$

Exp Info gain: <3 3-10 >10

<3	3-10	>10
H:0 L:1	H:0 L:2	H:1 L:1
0	0	1

Entropy $= \frac{1}{5}(0) + \frac{2}{5}(0) + \frac{2}{5}(1) = 0.4$

Exp Info gain $= 0.722 - 0.4 = \boxed{0.322}$ ✓

Edu - College H: 3 L: 2 Entropy $= -\frac{3}{5}\log(\frac{3}{5}) - \frac{2}{5}\log(\frac{2}{5}) = 0.97$

Career Infogain: Management Service

H: 2 L: 0 H: 1 L: 2

0 $= -\frac{1}{3}\log(\frac{1}{3}) - \frac{2}{3}\log(\frac{2}{3})$

$= 0.918$

Entropy $= \frac{2}{5}(0) + \frac{3}{5}(0.918) = 0.5508$

Career Infogain $= 0.971 - 0.5508 = \boxed{0.4202}$

Exp Infogain: < 3 3-10 > 10

H: 1 L: 1 H: 1 L: 0 H: 1 L: 1

1 0 1

Entropy $= \frac{2}{5}(1) + \frac{1}{5}(0) + \frac{2}{5}(1) = 0.8$

Exp Infogain $= 0.971 - 0.8 = \boxed{0.171}$

Edu - Highschool → Exp > 10 H: 1 L: 1 Entropy = 1

Career Infogain: Management Service

H: 1 L: 0 H: 0 L: 1

0 0

Entropy = 0

Career Infogain $= 1 - 0 = \boxed{1}$

Edu - College → Career - Service H: 1 L: 2 Entropy $= -\frac{1}{3}\log(\frac{1}{3}) - \frac{2}{3}\log(\frac{2}{3})$

$= 0.918$

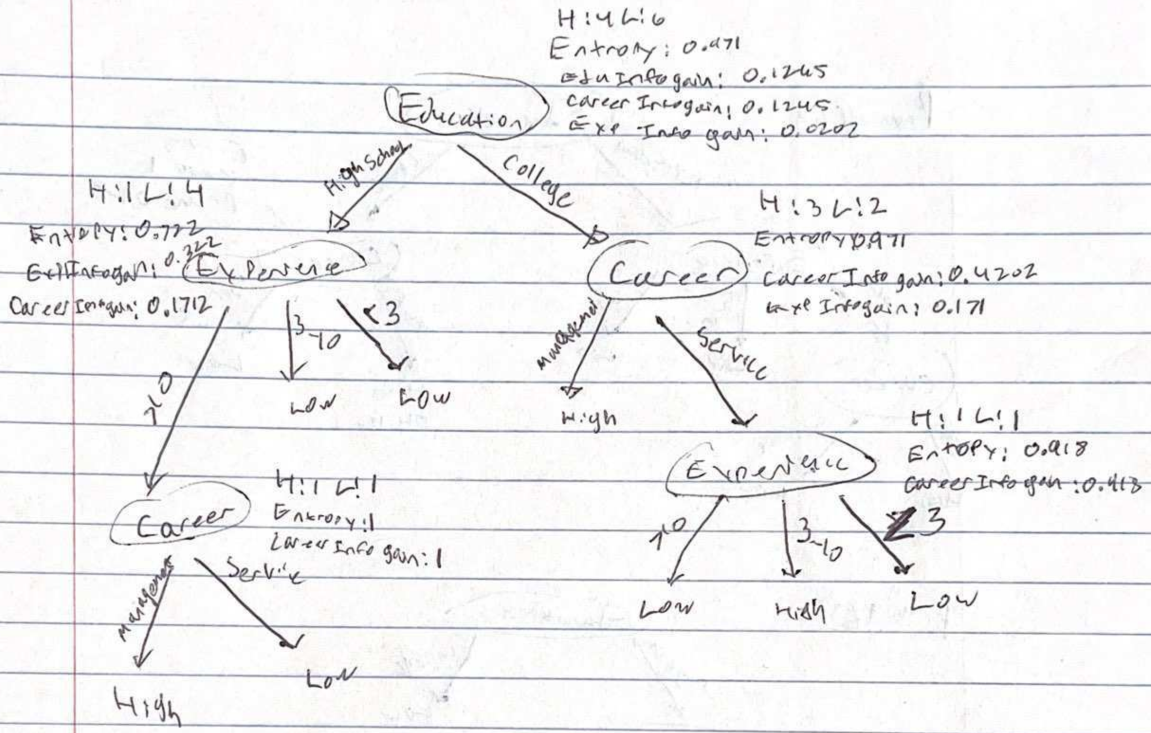
Exp Info gain < 3 3-10 > 10

H: 0 L: 1 H: 1 L: 0 H: 0 L: 1

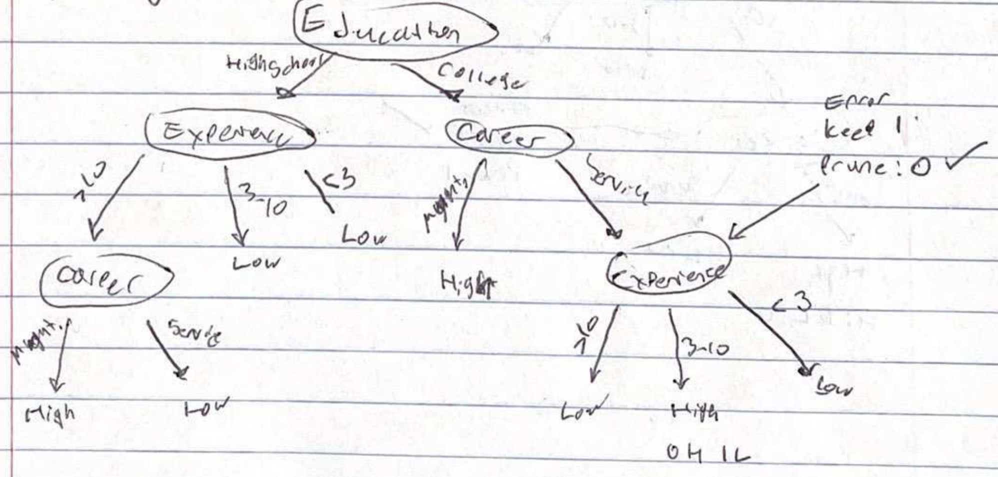
0 0 0

Entropy = 0

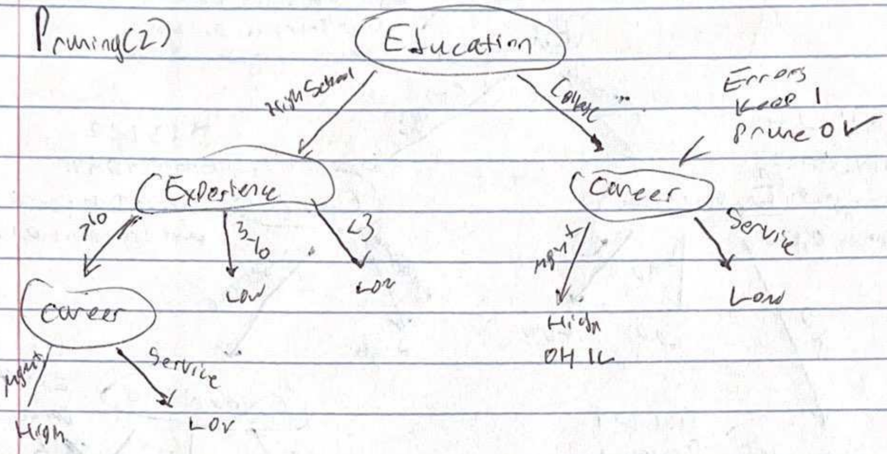
Exp Info gain $= 0.918 - 0 = \boxed{0.918}$


H:4 L:6
Entropy: 0.971
Edu Info gain: 0.1245
Career Info gain: 0.1245
Exp Info gain: 0.0202
Education
High School
College
H:1 L:4
Entropy: 0.722
Exp Info gain: 0.322
Career Info gain: 0.1712
Experience
>10
3-10
<3
Low
Low
H:3 L:2
Entropy: 0.971
Career Info gain: 0.4202
Exp Info gain: 0.171
Career
management
Service
High
H:1 L:1
Entropy: 0.918
Career Info gain: 0.918
Experience
>10
3-10
<3
Low
High
Low
Career
H:1 L:1
Entropy: 1
Career Info gain: 1
management
Service
High
Low

Pruning (1)

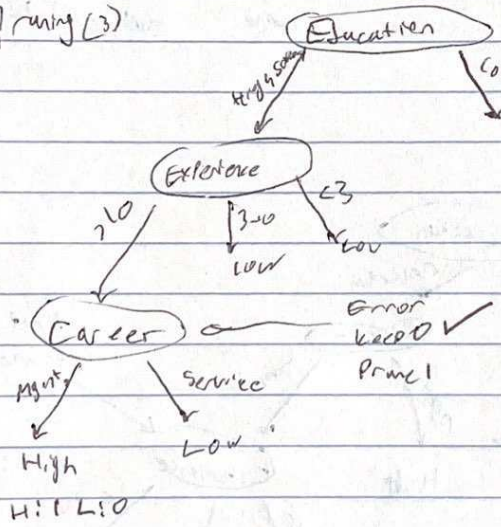

Education
Highschool
College
Experience
>10
3-10
<3
Low
Low
Career
mgmt.
Service
High
Low
Career
mgmt.
Service
High
Error
keep 1
prune: 0 ✓
Experience
>10
3-10
<3
Low
High
Low
0H 1L

Pruning(2)



Errors
Keep 1
Prune 0 ✓

Pruning(3)



Error
Keep 0 ✓
Prune 1

3.

Instance 1:

$P(y=\text{low} \mid x=\text{Highschool}, \text{service}, <3) = P(x=\text{Highschool} \mid y=\text{low}) \times$
$P(x=\text{Service} \mid y=\text{low}) * P(x=<3 \mid y=\text{low}) * P(y=\text{low})$

$= \left(\frac{4}{6}\right)\left(\frac{4}{6}\right)\left(\frac{2}{6}\right)\left(\frac{6}{10}\right)$

Laplace Smoothing: $\left(\frac{4+1}{6+2}\right)\left(\frac{4+1}{6+2}\right)\left(\frac{2+1}{6+3}\right)\left(\frac{6}{10}\right) = \frac{450}{5760} = 0.078125$

$P(y=\text{high} \mid x=\text{highschool}, \text{service}, <3) = P(x=\text{highschool} \mid y=\text{low}) P(x=\text{service} \mid y=\text{low}) P(x<3 \mid y=\text{low}) P(y=\text{low})$

$= \left(\frac{1}{4}\right)\left(\frac{1}{4}\right)\left(\frac{1}{4}\right)\left(\frac{4}{10}\right)$

Laplace Smoothing: $\left(\frac{1+1}{4+2}\right)\left(\frac{1+1}{4+2}\right)\left(\frac{1+1}{4+3}\right)\left(\frac{4}{10}\right) = \frac{32}{2520} = 0.012698$

The P of Y being low given the features is higher than Y being high so the predicted class is $\boxed{\text{Low}}$

Instance 2.

$P(y=\text{low} \mid x=\text{college}, \text{cetail}, <3) = P(x=\text{college} \mid y=\text{low}) P(x=\text{retail} \mid y=\text{low}) P(x<3 \mid y=\text{low}) P(y=\text{low})$

$= \left(\frac{2}{6}\right)\left(\frac{0}{6}\right)\left(\frac{2}{6}\right)\left(\frac{6}{10}\right)$

Laplace smoothing: $\left(\frac{2+1}{6+2}\right)\left(\frac{0+1}{6+3}\right)\left(\frac{2+1}{6+3}\right)\left(\frac{6}{10}\right) = \frac{54}{6480} = 0.00833$

$P(y=\text{high} \mid x=\text{college}, \text{retail}, <3) = P(x=\text{college} \mid y=\text{high}) P(x=\text{retail} \mid y=\text{high}) P(x=<3 \mid y=\text{low}) P(y=\text{low})$

$= \left(\frac{3}{4}\right)\left(\frac{0}{4}\right)\left(\frac{1}{4}\right)\left(\frac{4}{10}\right)$

Laplace Smoothing: $\left(\frac{3+1}{4+2}\right)\left(\frac{0+1}{4+3}\right)\left(\frac{1+1}{4+3}\right)\left(\frac{4}{10}\right) = \frac{32}{2940} = 0.10884$

$P(y=\text{high} \mid x=\text{college}, \text{retail}, <3) > P(y=\text{low} \mid x=\text{college}, \text{retail}, <3)$

So we predict that y is $\boxed{\text{high}}$

Instance 3

$P(y=\text{Low} \mid x=\text{graduate}, \text{Service}, 3 \text{ to } 10) = P(x=\text{grad} \mid y=\text{low}) P(x=\text{service} \mid y=\text{low}) P(x=3\text{ to }10 \mid y=\text{low})$
$\cdot P(y=\text{low})$ $= \left(\frac{0}{6}\right)\left(\frac{4}{6}\right)\left(\frac{2}{6}\right)\left(\frac{6}{10}\right)$

Laplace Smoothing: $\left(\frac{0+1}{6+3}\right)\left(\frac{4+1}{6+2}\right)\left(\frac{2+1}{6+3}\right)\left(\frac{6}{10}\right) = \frac{90}{6480} = 0.01388$

$P(y=\text{high} \mid x=\text{graduate}, \text{service}, 3\text{ to }10) = P(x=\text{graduate} \mid y=\text{high}) P(x=\text{service} \mid y=\text{high}) P(x=3\text{ to }10 \mid y=\text{high}) P(y=\text{high})$

$= \left(\frac{0}{4}\right)\left(\frac{1}{4}\right)\left(\frac{1}{4}\right)\left(\frac{4}{10}\right)$

Laplace Smoothing: $\left(\frac{0+1}{4+3}\right)\left(\frac{1+1}{4+2}\right)\left(\frac{1+1}{4+3}\right)\left(\frac{4}{10}\right) = \frac{16}{2940} = 0.00544$

$0.01388 > 0.00544$

So the predicted class will be $\boxed{\text{Low}}$