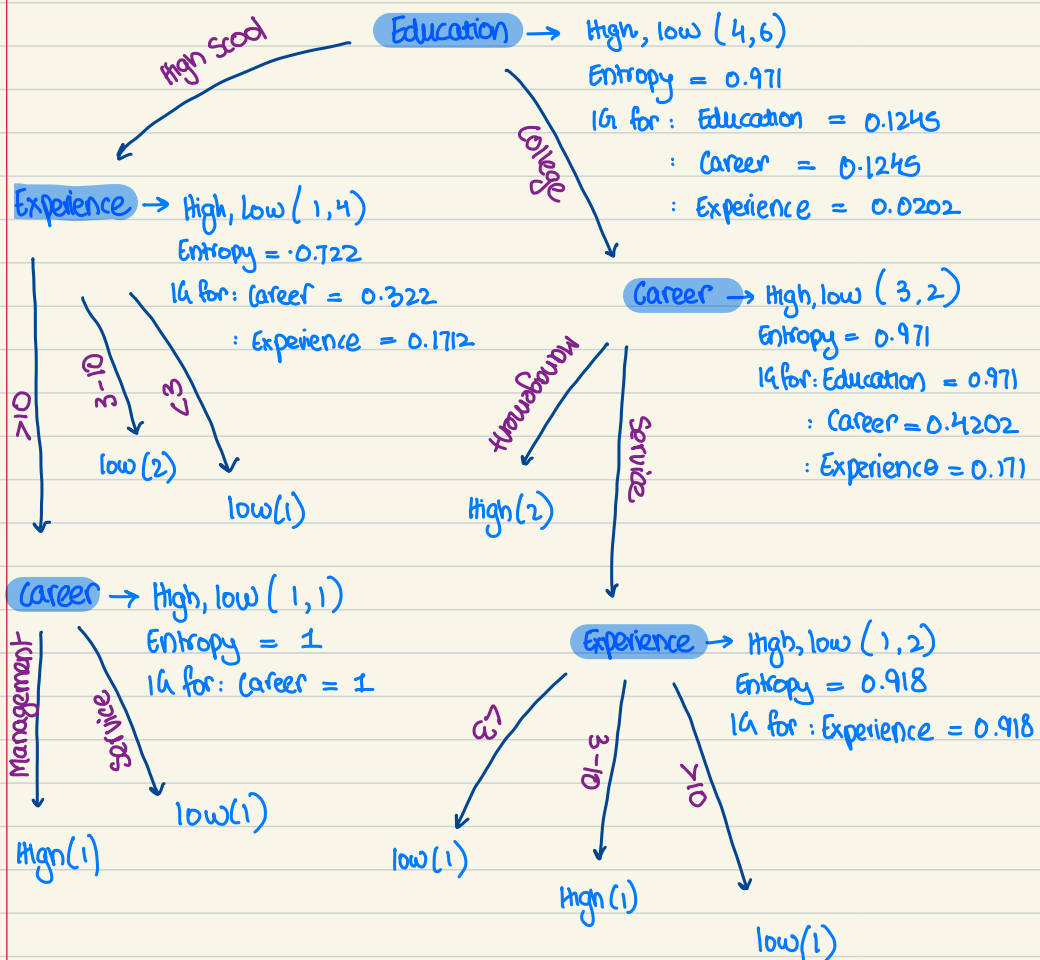
Homework 3

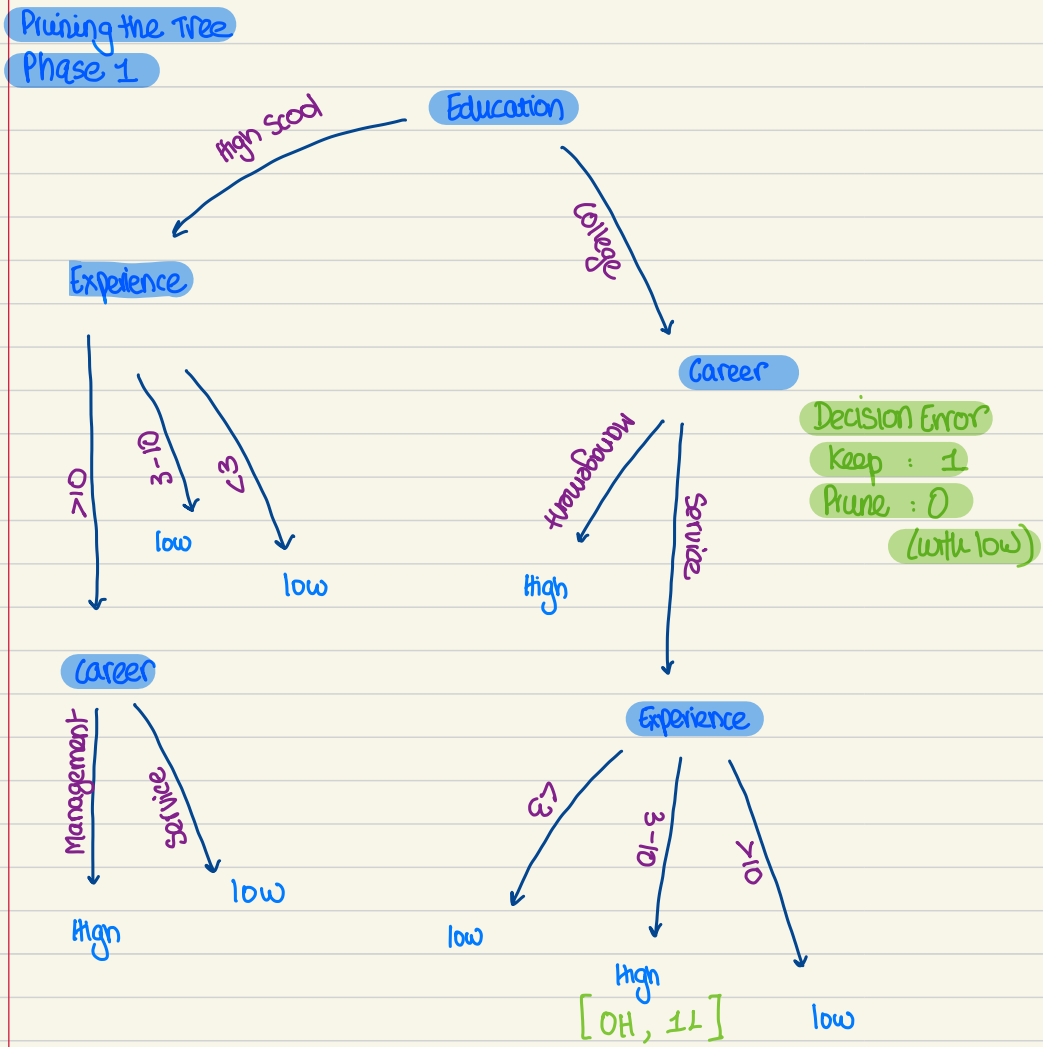
Due: 03/17/2022

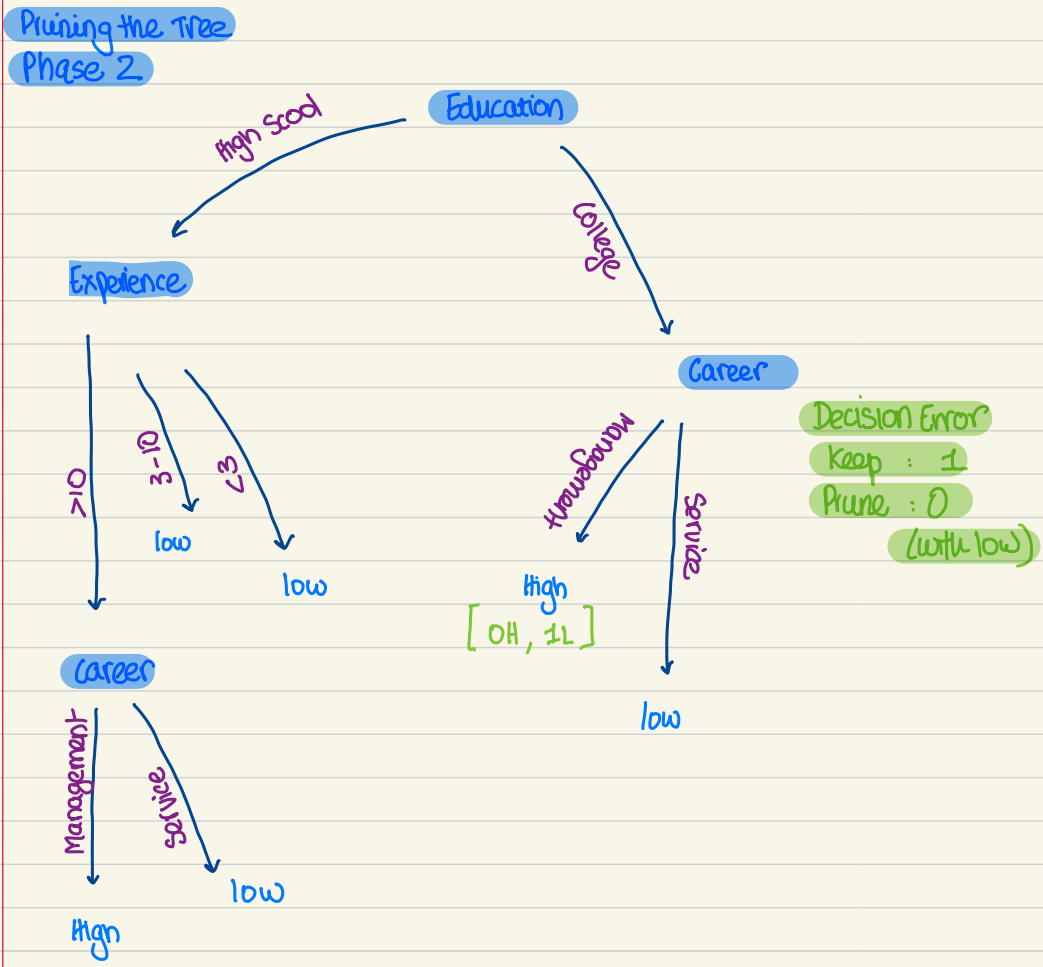
Devika Chandnani
A13405666

Question 1

- Decision Tree including: Number of highs & lows (salary)
  Base Entropy at each step
  Information Gain for each feature

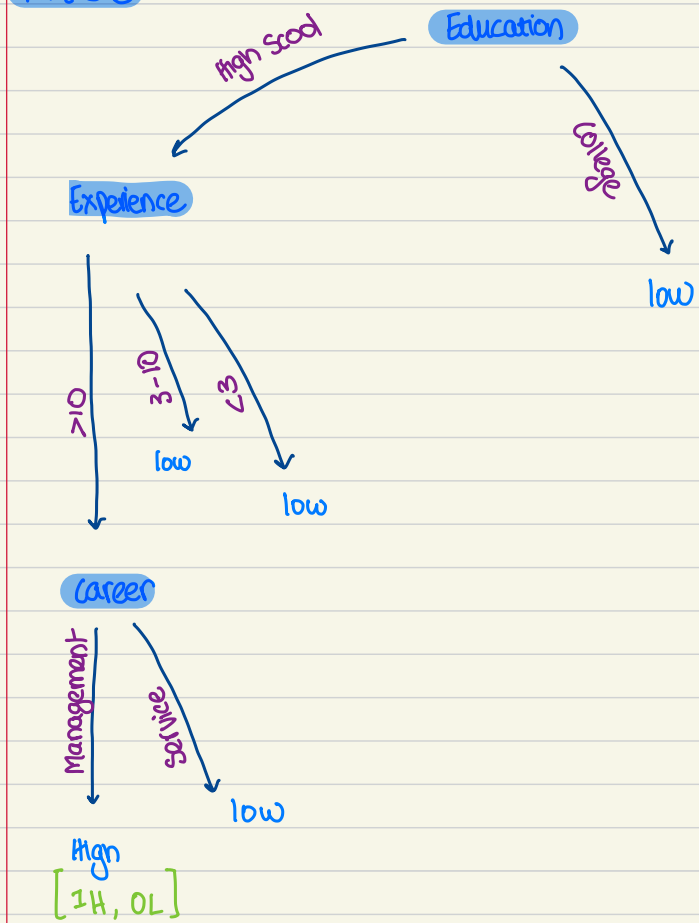



Pruning the Tree
Phase 1
Education
High Scool
College
Experience
Career
>10
3-10
<3
low
low
Management
Service
Career
Decision Error
Keep : 1
Prune : 0
(with low)
Management
Service
high
Experience
Career
high
low
<3
3-10
>10
low
high
[0H, 1L]
low


Pruning the Tree
Phase 2
Education
high scool
college
Experience
>10
3-10
<3
low
low
Career
management
service
High
[0H, 1L]
low
Decision Error
Keep : 1
Prune : 0
(with low)
Career
Management
service
low
High

Pruning the Tree

Phase 3



Question 2

→ **Instance 1** [Find out if the probability is low or high]

GIVEN high school, service, less than 3.

* $P(y = \text{low} \mid x = \text{high school, service, } < 3) = \frac{4}{6} \times \frac{4}{6} \times \frac{2}{6} \times \frac{6}{10}$

Add one smoothing $= \frac{5}{8} \times \frac{5}{8} \times \frac{3}{9} \times \frac{6}{10}$

$= \frac{450}{5760} \times 100 = 7.8\%$

* $P(y = \text{high} \mid x = \text{high school, service, } < 3) = \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{4}{10}$

Add one smoothing $= \frac{2}{6} \times \frac{2}{6} \times \frac{2}{7} \times \frac{4}{10}$

$= \frac{32}{2520} \times 100 = 1.27\%$

$P(y = \text{low}) > P(y = \text{high})$ given instance 1

∴ **Prediction = Low**

Question 2

→ **Instance 2** [Find out if the probability is low or high]

GIVEN college, retail, less than three

* $P(y = \text{low} \mid x = \text{college, retail, } <3) = \frac{2}{6} \times \frac{0}{6} \times \frac{2}{6} \times \frac{6}{10}$

Add one smoothing $= \frac{3}{8} \times \frac{1}{9} \times \frac{3}{9} \times \frac{6}{10}$

$= \frac{54}{6480} \times 100 = 0.83\%$

* $P(y = \text{high} \mid x = \text{college, retail, } <3) = \frac{3}{4} \times \frac{0}{4} \times \frac{1}{4} \times \frac{4}{10}$

Add one smoothing $= \frac{4}{6} \times \frac{1}{7} \times \frac{2}{7} \times \frac{4}{10}$

$= \frac{32}{2940} \times 100 = 1.1\%$

$P(y = \text{low}) < P(y = \text{high})$ given instance 1

∴ **Prediction = High**

Question 2

→ Instance 3 [Find out if the probability is low or high]

GIVEN graduate, service, 3 to 10

* $P(y = \text{low} \mid x = \text{graduate, service, 3 to 10}) = \frac{0}{6} \times \frac{4}{6} \times \frac{2}{0} \times \frac{6}{10}$

Add one smoothing $= \frac{1}{9} \times \frac{5}{8} \times \frac{3}{9} \times \frac{6}{10}$

$= \frac{90}{6480} \times 100 = 1.4\ \%$

* $P(y = \text{high} \mid x = \text{graduate, service, 3 to 10}) = \frac{0}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{4}{10}$

Add one smoothing $= \frac{1}{7} \times \frac{2}{6} \times \frac{2}{7} \times \frac{4}{10}$

$= \frac{16}{2940} \times 100 = 0.54\ \%$

$P(y = \text{low}) > P(y = \text{high})$ given instance 1

∴ Prediction = low

Question 3

a)

	FEATURE	|PCC|
4	f4	0.436922
13	f13	0.368269
14	f14	0.368224
16	f16	0.366025
7	f7	0.352141
22	f22	0.351350
26	f26	0.341043
1	f1	0.308811
20	f20	0.299049
31	f31	0.290783
34	f34	0.266093
2	f2	0.195732
28	f28	0.156904
25	f25	0.153096
19	f19	0.137636
17	f17	0.113945
32	f32	0.093174
8	f8	0.087773
0	f0	0.069795
10	f10	0.056876
21	f21	0.056605
11	f11	0.042117
33	f33	0.038810
6	f6	0.035295
15	f15	0.031478
35	f35	0.030855
29	f29	0.020829
18	f18	0.017931
27	f27	0.015606
9	f9	0.013005
3	f3	0.009214
30	f30	0.008955
24	f24	0.007780
23	f23	0.005508
12	f12	0.002179
5	f5	0.000098

→ Here is the output result listing the features from the highest absolute value to the lowest.

→ Taking the absolute value of PCC is meaningful to us because we are interested in the magnitude of the correlation without regard to the direction.

b) These are the 20 features with the highest accuracy of 0.9255 / 92.55 % accuracy level

```
The Following Selected Feature Set =  ['f4', 'f13', 'f14', 'f16', 'f7', 'f22', 'f26', 'f1', 'f20', 'f31', 'f34', 'f2', 'f28', 'f25', 'f19', 'f17', 'f32', 'f8', 'f0', 'f10']
Has an Accuracy Percentage of =  92.55
Has an Accuracy Count of =  783
```

Question 4

a) The selected feature set at each iteration is:

['f20']
['f20', 'f10']
['f20', 'f10', 'f19']
['f20', 'f10', 'f19', 'f8']
['f20', 'f10', 'f19', 'f8', 'f7']
['f20', 'f10', 'f19', 'f8', 'f7', 'f14']
['f20', 'f10', 'f19', 'f8', 'f7', 'f14', 'f2']
['f20', 'f10', 'f19', 'f8', 'f7', 'f14', 'f2', 'f4']
['f20', 'f10', 'f19', 'f8', 'f7', 'f14', 'f2', 'f4', 'f13']
['f20', 'f10', 'f19', 'f8', 'f7', 'f14', 'f2', 'f4', 'f13', 'f22']
['f20', 'f10', 'f19', 'f8', 'f7', 'f14', 'f2', 'f4', 'f13', 'f22', 'f25']
['f20', 'f10', 'f19', 'f8', 'f7', 'f14', 'f2', 'f4', 'f13', 'f22', 'f25', 'f16']

b) At the final iteration:

```
LOOVC Accuracy did not increase from the previous iteration 98.11
Final Selected Feature set is , ['f20', 'f10', 'f19', 'f8', 'f7', 'f14', 'f2', 'f4', 'f13', 'f22', 'f25', 'f16']
Final Accuracy with above feature set is  98.11
```