Assignment 1: Decision Trees

$$H(OK) = -\frac{5}{11}\log_2\left(\frac{5}{11}\right) - \frac{6}{11}\log_2\left(\frac{6}{11}\right) = 0.994$$

Calculating Information gain of Type

Type Classes: Easy, Some Difficulty, Advanced

~~$H(OK|Easy) = -\frac{3}{11}\log_2\left(\frac{3}{11}\right) - \frac{8}{11}\log\left(\frac{8}{11}\right)$~~

$H(OK$

$$H(OK|Easy) = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log\left(\frac{2}{3}\right)$$

$$= 0.918$$

$$H(OK|Some\ Difficulty) = -\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right)$$

$$H(OK|Advanced) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right) = 0.918$$

$$= 0.971$$

$$H(OK|Type) = \frac{3}{11}(0.918) - \frac{3}{11}(0.918) - \frac{5}{11}(0.971)$$

$$= 0.942$$

$$Gain(OK|Type) = 0.994 - 0.94 = 0.052$$

Distance Types : Short distance, within, far

$H(OK|\text{Short distance}) = -\frac{3}{5}\log\left(\frac{3}{5}\right) + -\frac{2}{5}\log\left(\frac{2}{5}\right) \approx 0.971$

$H(OK|\text{within}) = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1$

$H(OK|\text{far}) = -\frac{0}{2}\log_2\left(\frac{0}{2}\right) - \frac{2}{2}\log\left(\frac{2}{2}\right) = 0$

$H(OK|\text{Distance}) = -\frac{5}{11}(.971) - \frac{4}{11}(1) = -0.805$

$IG(OK, \text{Distance}) = 0.994 + -0.805 = 0.189$

Direction Types = West, South, North

$H(OK|\text{West}) = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1$

$H(OK|\text{South}) = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1$

$H(OK|\text{North}) = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right) = 0.918$

$H(OK|\text{Direction}) = -\frac{4}{11} - \frac{4}{11} - \frac{3}{11}(0.918) = -0.977$

$IG(OK|\text{Direction}) = 0.994 + -0.977 = 0.016$

Restriction Types: none, flat terrain, wheelchair access.

$H(OK|\text{none}) = 1$

$H(OK|\text{flat}) = 0.918$

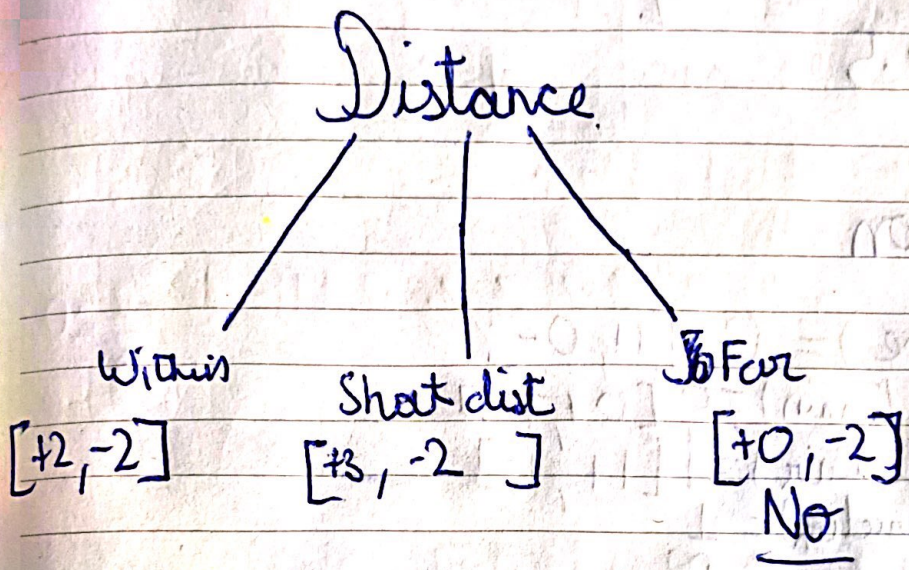
$H(OK|\text{wheelchair}) = 1$

$H(OK|\text{Restriction}) = -\frac{4}{11} - \frac{3}{11}(0.918) - \frac{4}{11} = -0.977$

$IG(OK|\text{Restriction}) = 0.994 + -0.977 = 0.016$

Column	Information gain
Type	0.052
Distance	0.189 → Highest information gain
Direction	0.016
Restriction	0.016

We will hence split our dataset on Distance



We now will need to repeat this process for the Within & short dist branches. We don't continue this for the Far branch because it is always a No.

Splitting for Within {2, 3, 7, 11}

$H(OK) = -\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2}) = 1$

Type:

$H(OK|Easy) = 0$ [0+, 1-]

$H(OK|Some\ Difficulty) = 0$ [1+, 0-]

$H(OK|Advanced) = 1$ [1+, 1-]

$H(OK|Type) = \frac{2}{4} \times 1 = \frac{2}{4}$

$IG(OK, Type) = 1 - \frac{2}{4} = 0.5$

Direction

$H(OK|West) = 0$ [1+, 0-]
$H(OK|South) = 0$ [1+, 0-]
$H(OK|North) = 1$ [1+, 1-]
$H(OK|Direction) = \frac{1}{2}$

$IG = 1 + -\frac{1}{2} = 0.5$

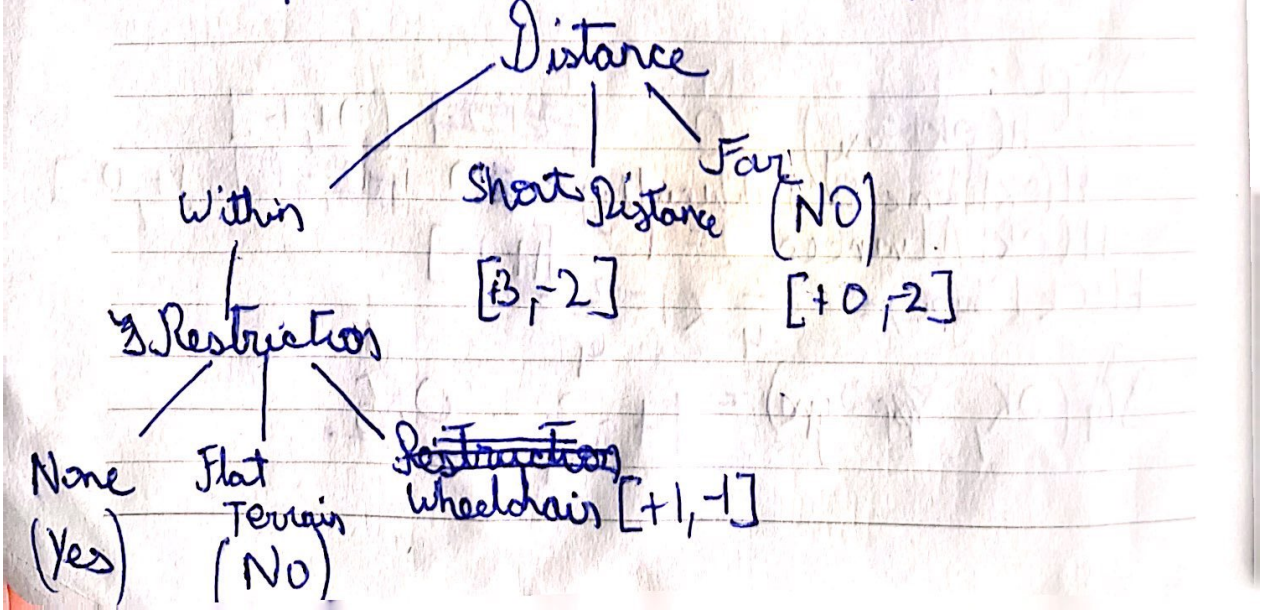
Restriction

$H(OK|none) = 0$ [1+, 0-]
$H(OK|flat\ terrain) = 0$ [0+, 1-]
$H(OK|wheelchair) = 1$ [1+, 1-]
$H(OK|Restriction) = \frac{1}{2}$

$IG(Restriction, OK) = 1 - \frac{1}{2} = 0.5$

Column	IG
Type	0.5
Direc	0.5
Restriction	0.5

All our information gain is the same so we can split on any feature. We will split it on Restriction.



Splitting for Wheelchair: {Y7, Y11}

H(OK) = 1 [1+, 1-]

Type

H(OK | Easy) = 0 [+0, -1]
H(OK | Advanced) = 0 [+1, 0]

Gr(OK, Type) = 1 - 0 = 1

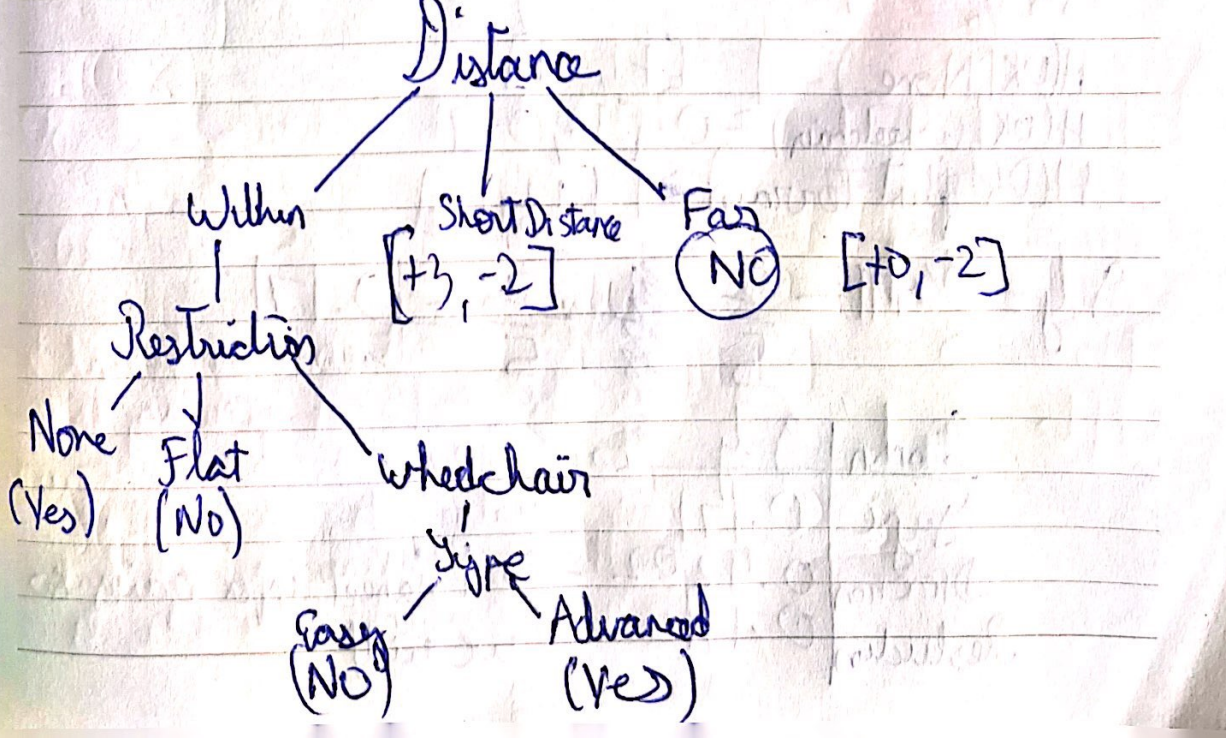
Direction

H(OK | North) = 1/2 [1+, 1-]

Gr(Direction, OK) = 1 - 1/2 = 0

Column	Gr
Type	1
Direction	0

We will split on Type as it has high Gr.



We will now split on short distance: [T, F, T, F, T]

$H(OK) = 0.971$

Type (OK)

$H(OK|Easy) = 0$ [1+, 0−]
$H(OK|Advanced) = 1$ [1+, 1−]
$H(OK|Some\ diff) = 1$ [1+, 0−]

$IG = 0.971 - \frac{2}{5} - \frac{2}{5} = 0.171$

② Direction

$H(OK|West) = 0.918$ [1+, 2−]
$H(OK|South) = 0$ [2+, 0−]

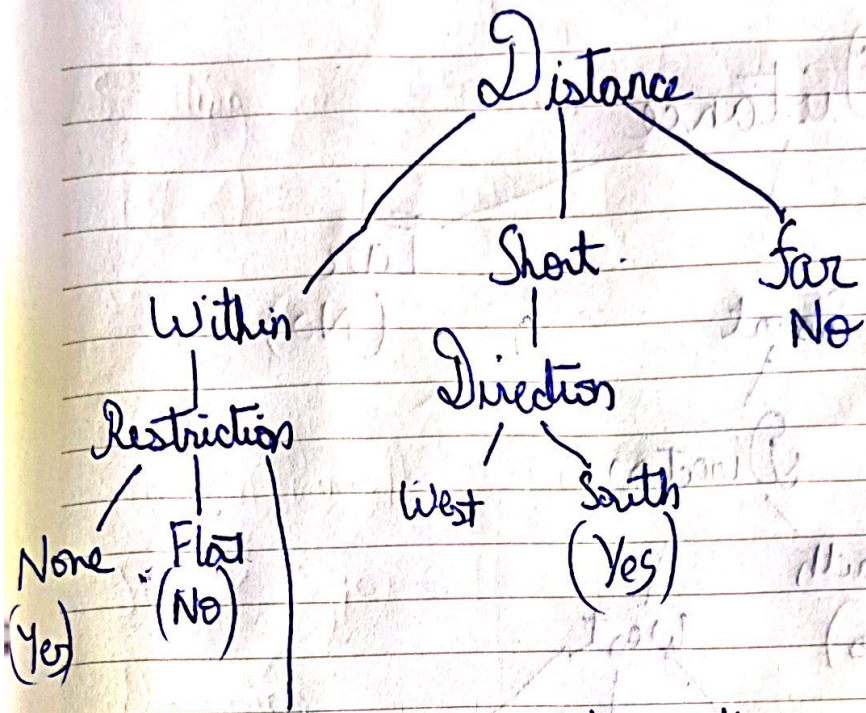
$IG(OK, Direction) = 0.971 - \frac{3 \times 0.918}{5} = 0.4202$

Restriction

$H(OK|None) = 1$ [+1, −1]
$H(OK|Wheelchair) = 0$ [1+, 0−]
$H(OK|flat\ terrain) = 1$ [1+, 1−]

$IG = 0.971 - \frac{2}{5} - \frac{2}{5} = 0.171$

Columns	IG
Type	0.171
Direction	0.4202 → Highest info gain so we will
Restriction	0.171



We will not split on West
{Y1, Y8, Y9}
H(Ok) = 0.918

Type:
H(Ok | Easy) = 0
H(Ok | Some Difficulty) = 0
H(Ok | Advanced) = 0
Gf = 0.918

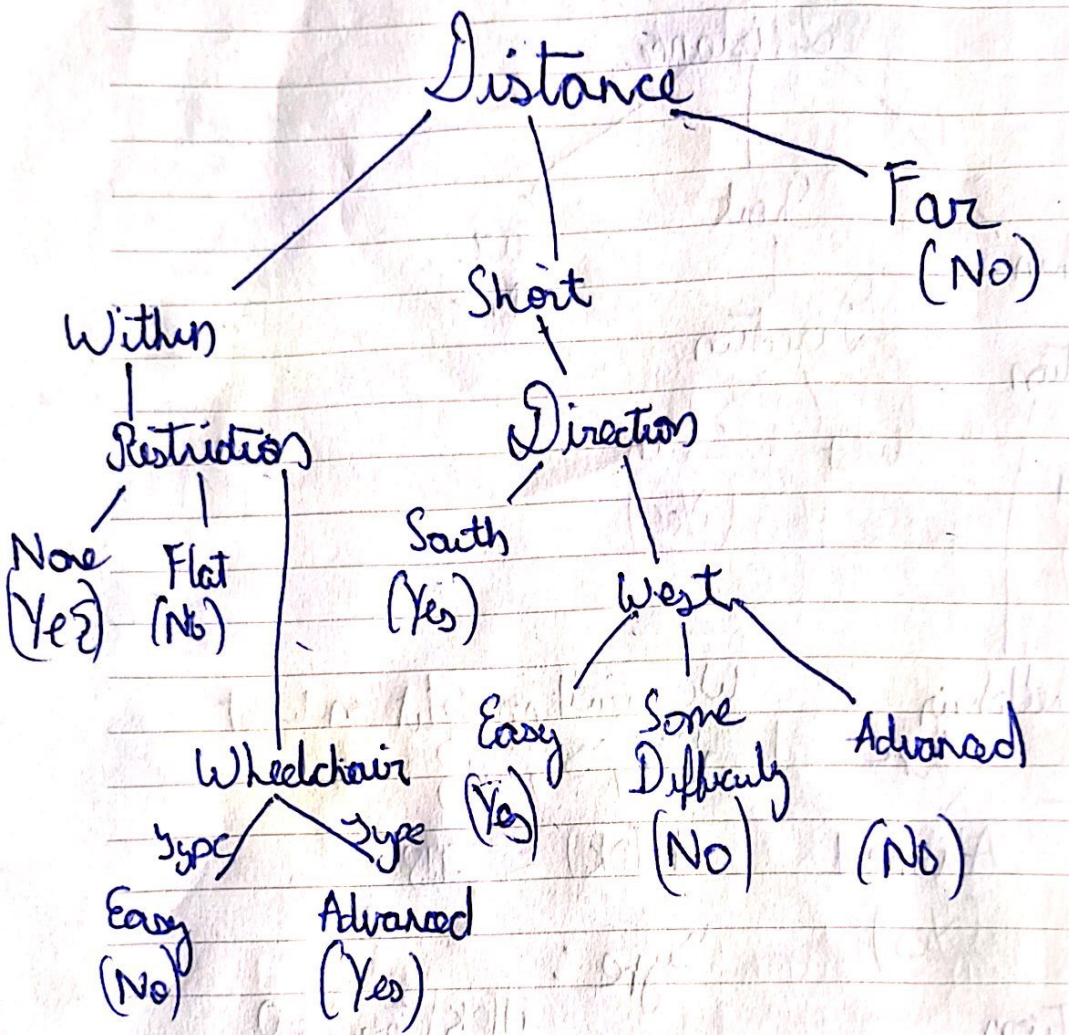
Restriction
H(Ok | None) = 1
H(Ok | flat terrain) = 0

$Gf = 0.918 - \frac{2}{3} = 0.251$

Columns	
Type	0.918 →
Restriction	0.251

We will now split on Type because it has highest information gain

Final Decision Tree



B) The training set accuracy of our decision tree is 100%

C) T12 = 1
T13 = 0
T14 = ~~Yes~~ 1
T15 = 0
T16 = ~~Yes~~ 1

D) Test Set Accuracy = $\frac{4}{5}$ = 80%

Test set Error = $\frac{1}{5}$ = 20%

This result is significant because the test set accuracy is less than 50%. The test set accuracy is fairly high. It is ideal as the 80% accuracy is high, but this is not perfect as the datasets

are not very large so the model has not been trained & tested on a large dataset so it not representative of the entire dataset.

Task 3

Discuss what will happen if you decide to change the splitting criterion. Explain the new splitting criterion and how it might change your decision tree.

Changing the splitting criterion to Gini impurity, the decision tree can massively change.

Information gain is biased towards attributes with many distinct values. Gini impurity on the other hand, is biased towards attributes with fewer values. This will result in the root of the decision tree changing (as it has a lot of distinct values) and hence the whole decision tree would change as we would be biased towards features with the fewest distinct values.

Explain whether your evaluation method can indicate whether your tree is over- or underfitting.

Cross validation can help us determine if our tree is overfit. Our tree is trained on k-1 folds(from the training dataset) and then tested on the validation fold. Then we perform the same step but using folds from our full dataset. If the validation set result in a high level of accuracy but the validation set results on the full dataset perform poorly, then we are able to deduce that our tree is over fit. If it performs poorly on both validation tests then we know that it is underfit.