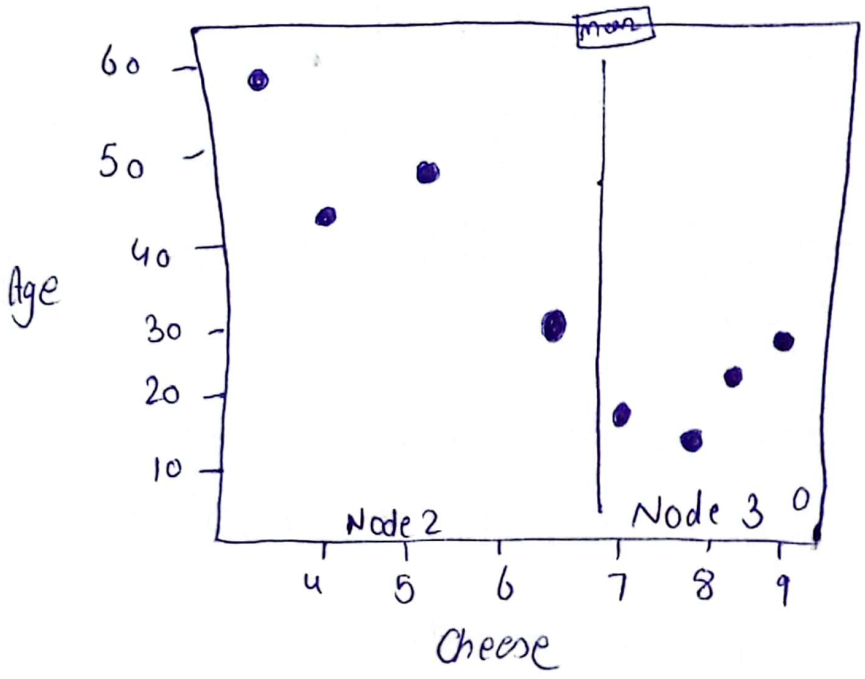
Problem → 2

Kdtree on Pizza Outlet

→ Split regions at mean value of observations

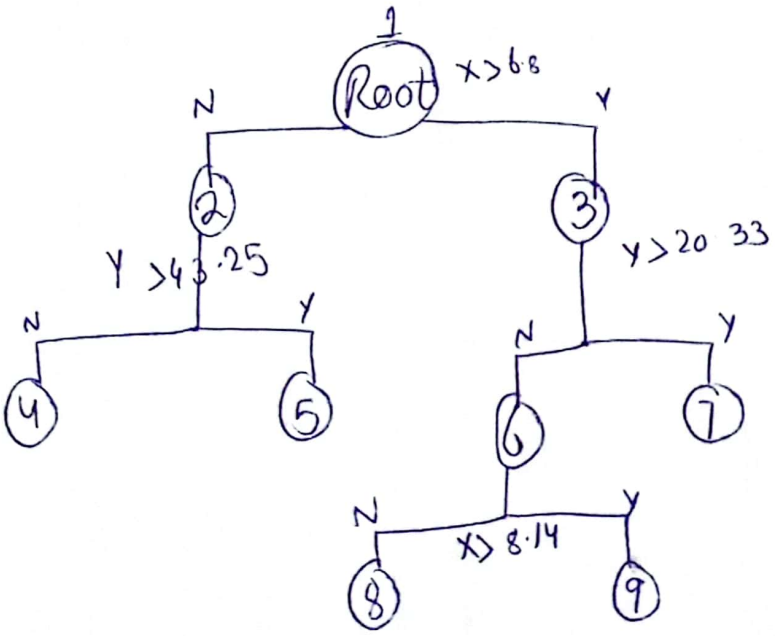
First cut: find Mean of X-coordinates (Cheese content)

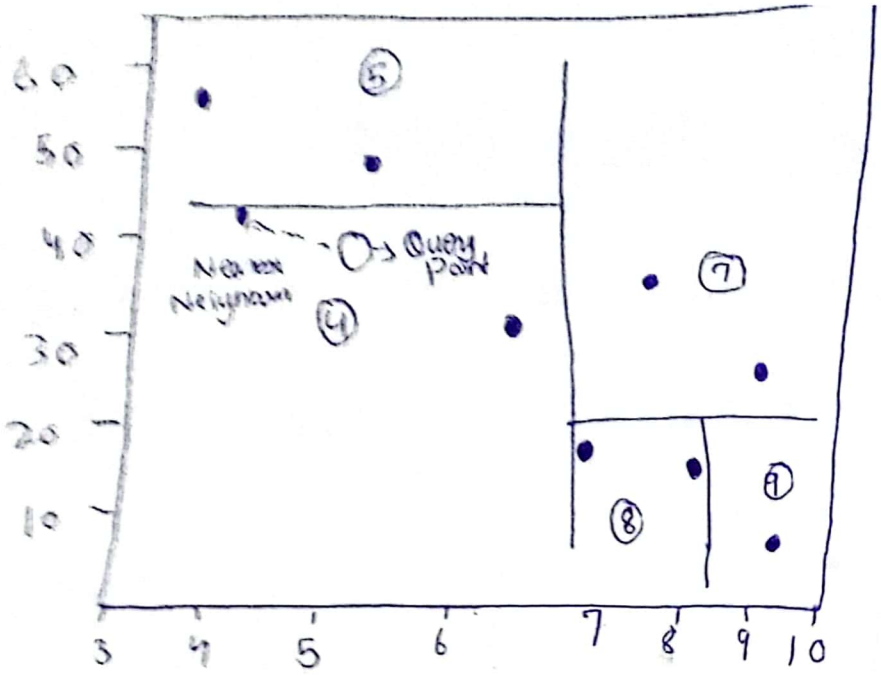


Cheese Content

6.2
8.4
3.3
5
9
1
9.2
7.6

Max = 9.2
Min = 3.3
Mean => 6.77



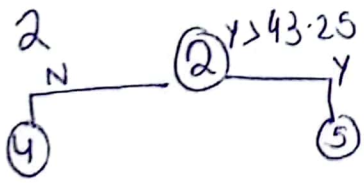


There are fewer than 3 points in each region

If new query point comes, find region traverse it lies in 4th region, find Nearest Neighbours

→ Query point might be not actual

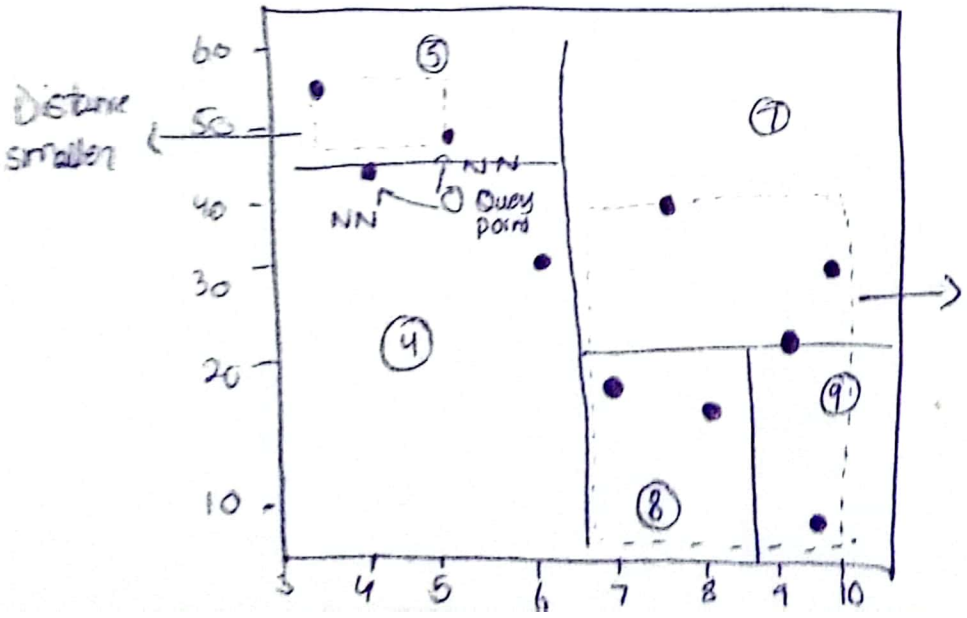
Traverse to node 2

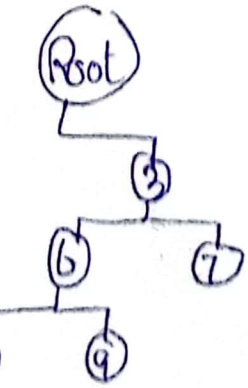


Result

Output
Harry | 46 | 7 | Pizza Hut

Node 5





Distance greater from new nearest neighbour
It is final point

Problem # 3

a) Rule accuracy after R1 has been discovered, none covered by R1 discarded.

If the examples for R_1 are not discarded then R_2 will be choosen because it has higher accuracy (70%) than R_3 (66.7%)

(Figure illustrates coverage of classification Rules R_1, R_2 & R_3 we have to determine which is best & worst rule)

All positive examples = 29
Negative examples = 21

Rule 1 covers = 12 positive, 3 negative

Rule 2 covers = 7 positive, 3 negative

Rule 3 covers = 8 positive, 4 negative

If Rule accuracy → $\frac{\text{Positive results}}{\text{Total results}}$

$$R1 = \frac{12}{15} = 0.8 = 80\%.$$

$$R2 = \frac{7}{10} = 0.7 = 70\%.$$

$$R3 = \frac{8}{12} = 0.6667 = 66.7\%.$$

Result R_1 is best than R_2 & R_3

→ when R_1 is discarded
we choose Rule 2
so Rule 2 > Rule 3
R_2 Higher accuracy than R_3
↓ 70% ↓ 66.7%.

(b) Rule accuracy after R1 discovered, where only positive examples covered by R1 discarded.

If positive examples covered by R1 are discarded, new accuracies for R_2 & R_3 approximately $R_2 = 70\%$, $R_3 = 60\%$ respectively.

$R_2 > R_3$

R_2 is preferred upon R_3.

c) R_1 discovered, both pos & neg examples by R_1 discarded.

If the positive and negative examples covered by R_1 discarded new accuracies R_2 and R_3 will be 70%, 75%.

$R_3 > R_2$.

R_3 is preffered over R_2

⇒ Foil's Info gains

R_1:

$$R_1 = 12\left(\log\left(\frac{12}{15}\right) - \log\left(\frac{29}{50}\right)\right)$$
$$= 12(\log(0.8) - \log(0.58))$$
$$= 12((-0.0969 - (-0.236))$$
$$= 12 \times 0.14 = 1.68$$

R_2:

$$R_2 = 7(\log\left(\frac{7}{10}\right) - \log\left(\frac{29}{50}\right))$$
$$= 7(\log(0.7) - \log(0.58))$$
$$= 7(-0.154 - (-0.236)) = 7(0.082) = 0.574$$

R_3:

$$R_3 = 8\left(\log\left(\frac{8}{12}\right) - \log\left(\frac{29}{50}\right)\right)$$
$$= 8(\log(0.667) - \log(0.58)$$
$$= 8(0.175 - (-0.236))$$
$$= 8(0.061)$$
$$= 0.488$$

$$\boxed{R_1 > R_2 > R_3}$$

Problem #5

Confusion Matrix

a) Total No. of Instances

Sum of all values

Roses	100	10	5
Daisies	15	85	20
Tulips	8	18	90
	123	113	115

$\Sigma = 351$

b) Accuracy

$$= \frac{\text{No. of correct predictions}}{\text{Total number of predictions}}$$

$$= \frac{Tp + TN}{\text{Total instances}} = \frac{100+85+90}{351} = \frac{275}{351}$$

$$= 0.7834 \approx 78.34\%$$

c) Sensitivity (Recall)

$$\frac{Tp}{Tp+FN} \Rightarrow \text{Roses} = \frac{100}{100+10+5} = \frac{100}{115} = 0.869 \approx 86.9\%$$

$$\text{Daisies} = \frac{85}{15+85+20} = \frac{85}{120} = 0.7083 \approx 70.83\%$$

$$\text{Tulips} = \frac{90}{8+18+90} = \frac{90}{116} = 0.7758 \approx 77.58\%$$

→ Roses highest sensitivity indicates best performance in correctly identifying instances of each class

→ Lower daises incorrectly identifying

d) specificity TN / (TN+FP) -> TN/N

Roses = 85+20+18+90 / 85+20+18+90+5 = 213/213 = 1.9565

Daisies = 100+8+90+5 / 100+8+90+5+15+20 = 203/203 = 0.8913

Tulips = 100+10+15+85 / 100+10+15+85+18 = 210/210 = 1.9275

Evaluate classifier's performance for individual classes

e) precession (1 indicates correctly identified)

$$\frac{TP}{TP+FP}$$

Roses - $\frac{100}{100+15+8} = \frac{100}{123} = 0.813 \approx 81.3\%$

Daisies = $\frac{85}{85+10+18} = \frac{85}{113} = 0.776$

$= \frac{90}{90+5+20} = \frac{90}{115} = 0.783$

Highest precesion Roses => fewer false positives