Assignment 3

Durbasmriti Saha
EE708: FUNDAMENTALS OF DATA SCIENCE AND MACHINE INTELLIGENCE

March 21, 2025

1)a) The Gini index for a dataset is given by:

$$\text{Gini} = 1 - \sum p_i^2$$

where p_i is the proportion of class i in the dataset.
given in the question, total sample = 200 , positive sample = 120, negative sample = 80.
The probabilities: $p_pos = \frac{120}{200} = 0.6$
$p_neg = \frac{80}{200} = 0.4$
Gini(before splitting) = $1 - (0.6^2 + 0.4^2)$ = 1- (0.36 + 0.16) = 1-0.52 = 0.48

b) After split The subsets:
Left subset: positive = 50, negative = 10, total = 60
Gini(left) = $1 - \left(\frac{50}{60}\right)^2 - \left(\frac{10}{60}\right)^2$ = 1 - $\frac{2600}{3600}$ = 1-0.7222 = 0.2778
Right subset : positive = 70, negative = 70, total = 140
Gini(right) = $1 - \left(\frac{70}{140}\right)^2 - \left(\frac{70}{140}\right)^2$ =$1 - (0.5^2 + 0.5^2)$ = 1-(0.25+0.25)= 0.5
Gini split (weighted) = $\frac{60}{200}Gini(left)+\frac{140}{200}Gini(right) = \frac{60}{200}*0.2778+\frac{140}{200}*0.5 = 0.0833+0.35$ = 0.4333
The weighted Gini index is lower than the previous Gini index, this means that split improves purity.

2)a) We know,

$$\text{SSE} = \sum (y_i - \bar{y})^2$$

here, y =[10,12,15,18,21,25,28,30] $\bar{y} = \frac{10+12+15+18+21+25+28+30}{8} = \frac{159}{8} = 19.875$

$$\text{SSE}_{\text{total}} = (10 - 19.875)^2 + (12 - 19.875)^2 + (15 - 19.875)^2 + (18 - 19.875)^2$$

$$+(21 - 19.875)^2 + (25 - 19.875)^2 + (28 - 19.875)^2 + (30 - 19.875)^2$$

$$+(21 - 19.875)^2 + (25 - 19.875)^2 + (28 - 19.875)^2 + (30 - 19.875)^2$$

$$= 97.0156 + 61.0156 + 23.9063 + 3.5156 + 1.2656 + 26.3906 + 67.1406 + 102.2656$$

$$= 97.0156 + 61.0156 + 23.9063 + 3.5156 + 1.2656 + 26.3906 + 67.1406 + 102.2656$$

$$= 382.875$$

The possible split points will be midpoints between consecutive x_1 values 1.5,2.5,3.5,4.5,5.5,6.5,7.5

split at $x_1 = 1.5$:
$x_{\text{left}} = 1$, $x_{\text{right}} = 2,3,4,5,6,7,8$, $\bar{y}_{\text{right}} = \frac{12+15+18+21+25+28+30}{7} = 21.2857$, $\bar{y}_{\text{left}} = 0$, $SSE_{\text{left}} = 0$,

$$\begin{aligned}
\text{SSE}_{\text{right}} &= (12-21.2857)^2 + (15-21.2857)^2 + (18-21.2857)^2 + (21-21.2857)^2 \\
&\quad + (25-21.2857)^2 + (28-21.2857)^2 + (30-21.2857)^2 \\
&= 86.898 + 39.505 + 10.805 + 0.081 + 13.795 + 44.711 + 74.288 \\
&= 269.083
\end{aligned}$$

$$\begin{aligned}
\text{SSE}_{\text{split}} &= \frac{1}{8} \times 0 + \frac{7}{8} \times 269.083 \\
&= 0 + 235.44 \\
&= 235.44
\end{aligned}$$

split at $x_1 = 2.5$:
$x_{\text{left}} = 1,2$, $x_{\text{right}} = 3,4,5,6,7,8$, $\bar{y}_{\text{right}} = \frac{15+18+21+25+28+30}{6} = 22.8333$, $\bar{y}_{\text{left}} = \frac{10+12}{2} = 11$,

$$\text{SSE}_{\text{left}} = (10-11)^2 + (12-11)^2 = 2$$

,

$$\begin{aligned}
\text{SSE}_{\text{right}} &= (15-22.8333)^2 + (18-22.8333)^2 + (21-22.8333)^2 + (28-22.8333)^2 \\
&\quad + (25-22.8333)^2 + (30-22.8333)^2
\end{aligned}$$

$= 61.639+23.372+3.361+4.694+26.694+51.361 = 171.12$

$$\begin{aligned}
\text{SSE}_{\text{split}} &= \frac{2}{8} \times 2 + \frac{6}{8} \times 171.12 \\
&= 0.5 + 128.34 \\
&= 128.84
\end{aligned}$$

split at $x_1 = 3.5$:
$x_{\text{left}} = 1,2,3$, $x_{\text{right}} = 4,5,6,7,8$, $\bar{y}_{\text{right}} = \frac{18+21+25+28+30}{5} = 24.4$, $\bar{y}_{\text{left}} = \frac{10+12+15}{3} = 12.33$,

$$\text{SSE}_{\text{left}} = (10-12.33)^2 + (12-12.33)^2 + (15-12.33)^2 = 12.65$$

,

$$\begin{aligned}
\text{SSE}_{\text{right}} &= (18-24.4)^2 + (21-24.4)^2 + (28-24.4)^2 \\
&\quad + (25-24.4)^2 + (30-24.4)^2
\end{aligned}$$

$= 40.96+11.56+0.36+12.96+30.96 = 96.8$

$$\begin{aligned}
\text{SSE}_{\text{split}} &= \frac{3}{8} \times 12.65 + \frac{5}{8} \times 96.8 \\
&= 4.74375 + 60.5 \\
&= 65.24
\end{aligned}$$

split at $x_1 = 4.5$:
$x_{\text{left}} = 1,2,3,4$, $x_{\text{right}} = 5,6,7,8$, $\bar{y}_{\text{right}} = \frac{21+25+28+30}{4} = 26$, $\bar{y}_{\text{left}} = \frac{10+12+15+18}{4} = 13.75$,

$$\text{SSE}_{\text{left}} = (10-13.75)^2 + (12-13.75)^2 + (15-13.75)^2 + (18-13.75)^2 = 36.75$$

,

$$\text{SSE}_{\text{right}} = (21-26)^2 + (28-26)^2 + (25-26)^2 + (30-26)^2$$

$= 46$

$$\begin{aligned}\text{SSE}_{\text{split}} &= \frac{4}{8} \times 36.75 + \frac{4}{8} \times 46 \\ &= 18.375 + 23 \\ &= 41.375\end{aligned}$$

split at $x_1 = 5.5$:
$x_{\text{left}} = 1,2,3,4,5$, $x_{\text{right}} = 6,7,8$, $\bar{y}_{\text{right}} = \frac{25+28+30}{3} = 27.67$, $\bar{y}_{\text{left}} = \frac{10+12+15+18+21}{5} = 15.2$,

$$\text{SSE}_{\text{left}} = (10-15.2)^2 + (12-15.2)^2 + (15-15.2)^2 + (18-15.2)^2 + (21-15.2)^2 = 78.8$$

,

$$\text{SSE}_{\text{right}} = (28-27.67)^2 + (25-27.67)^2 + (30-27.67)^2$$

$= 12.67$

$$\begin{aligned}\text{SSE}_{\text{split}} &= \frac{5}{8} \times 78.8 + \frac{3}{8} \times 12.67 \\ &= 49.25 + 4.7513 \\ &= 54.0013\end{aligned}$$

split at $x_1 = 6.5$:
$x_{\text{left}} = 1,2,3,4,5,6$, $x_{\text{right}} = 7,8$, $\bar{y}_{\text{right}} = \frac{28+30}{2} = 29$, $\bar{y}_{\text{left}} = \frac{10+12+15+18+21+25}{6} = 16.83$,

$$\text{SSE}_{\text{left}} = (10-16.83)^2+(12-16.83)^2+(15-16.83)^2+(18-16.83)^2+(21-16.83)^2+(25-16.83)^2 = 158.83$$

,

$$\text{SSE}_{\text{right}} = (28-29)^2 + (30-29)^2$$

$= 2$

$$\begin{aligned}\text{SSE}_{\text{split}} &= \frac{6}{8} \times 158.83 + \frac{2}{8} \times 2 \\ &= 119.1225 + 0.5 \\ &= 119.6225\end{aligned}$$

split at $x_1 = 7.5$:
$x_{\text{left}} = 1,2,3,4,5,6,7$, $x_{\text{right}} = 8$, $\bar{y}_{\text{right}} = \frac{30}{1} = 30$, $\bar{y}_{\text{left}} = \frac{10+12+15+18+21+25+28}{7} = 18.43$,

$$\text{SSE}_{\text{left}} = (10-18.43)^2+(12-18.43)^2+(15-18.43)^2+(18-18.43)^2+(21-18.43)^2+(25-18.43)^2+(25-18.43)^2 =$$

,

$$\text{SSE}_{\text{right}} = (30-0)^2$$

$= 0$

$$\begin{aligned}\text{SSE}_{\text{split}} &= \frac{7}{8} \times 265.71 + \frac{1}{8} \times 0 \\ &= 232.496 + 0 \\ &= 232.496\end{aligned}$$

The lowest SSE is at $x_1 = 4.5$ with SSE = 41.375
b) Regression Tree: $x_1 \leq 4.5$ then y = 13.75
$x_1 > 4.5$ then y = 26

3)a) Calculating distances of each point from every centroids: where D(ci) = distance from centroid i.
Point $(1, 2)$:

$$D((1,2), C_1(2,3)) = (1-2)^2 + (2-3)^2 = 1 + 1 = 2$$

$$D((1,2), C_2(5,8)) = (1-5)^2 + (2-8)^2 = 16 + 36 = 52$$

$$D((1,2), C_3(9,4)) = (1-9)^2 + (2-4)^2 = 64 + 4 = 68$$

$$\Rightarrow \text{Assigned to } C_1$$

Point $(3, 4)$

$$D((3,4), C_1(2,3)) = (3-2)^2 + (4-3)^2 = 1 + 1 = 2$$

$$D((3,4), C_2(5,8)) = (3-5)^2 + (4-8)^2 = 4 + 16 = 20$$

$$D((3,4), C_3(9,4)) = (3-9)^2 + (4-4)^2 = 36 + 0 = 36$$

$$\Rightarrow \text{Assigned to } C_1$$

Point $(6, 7)$

$$D((6,7), C_1(2,3)) = (6-2)^2 + (7-3)^2 = 16 + 16 = 32$$

$$D((6,7), C_2(5,8)) = (6-5)^2 + (7-8)^2 = 1 + 1 = 2$$

$$D((6,7), C_3(9,4)) = (6-9)^2 + (7-4)^2 = 9 + 9 = 18$$

$$\Rightarrow \text{Assigned to } C_2$$

Point $(8,3)$

$$D((8,3), C_1(2,3)) = (8-2)^2 + (3-3)^2 = 36 + 0 = 36$$

$$D((8,3), C_2(5,8)) = (8-5)^2 + (3-8)^2 = 9 + 25 = 34$$

$$D((8,3), C_3(9,4)) = (8-9)^2 + (3-4)^2 = 1 + 1 = 2$$

$$\Rightarrow \text{Assigned to } C_3$$

Point $(5,5)$

$$D((5,5), C_1(2,3)) = (5-2)^2 + (5-3)^2 = 9 + 4 = 13$$

$$D((5,5), C_2(5,8)) = (5-5)^2 + (5-8)^2 = 0 + 9 = 9$$

$$D((5,5), C_3(9,4)) = (5-9)^2 + (5-4)^2 = 16 + 1 = 17$$

$$\Rightarrow \text{Assigned to } C_2$$

Cluster 1(C1): New centroid $= C_1' = \left(\frac{1+3}{2}, \frac{2+4}{2}\right) = (2,3)$
Cluster 2(C2): New centroid $= C_2' = \left(\frac{6+5}{2}, \frac{7+5}{2}\right) = (5.5,6)$
Cluster 3(C3): New centroid $= C_1' = (8,3)$
b) Initial Distortion :
$D_{\text{initial}} = (2+2+2+2+9) = 17$
New Distortion: $D_{\text{new}} = (2+2+2.25+2+2.25) = 10.5$ Since the distortion has decreased from before, clustering has improved.

5)a) Given,

$$p(x) = \sum_{k=1}^{K} \pi_k p(x \mid k)$$

By Conditional Probabilty,

$$p(x_b \mid x_a) = \frac{p(x_a, x_b)}{p(x_a)}$$

substituting,

$$p(x_a, x_b) = \sum_{k=1}^{K} \pi_k p(x_a, x_b \mid k)$$

Similarly,

$$p(x_a) = \sum_{k=1}^{K} \pi_k p(x_a \mid k)$$

Thus the conditional density becomes:

$$p(x_b \mid x_a) = \frac{\sum_{k=1}^{K} \pi_k p(x_a, x_b \mid k)}{\sum_{j=1}^{K} \pi_j p(x_a \mid j)}$$

Using

$$p(x_a, x_b \mid k) = p(x_b \mid x_a, k) p(x_a \mid k)$$

Substitute,

$$p(x_b \mid x_a) = \frac{\sum_{k=1}^{K} \pi_k p(x_b \mid x_a, k) p(x_a \mid k)}{\sum_{j=1}^{K} \pi_j p(x_a \mid j)}$$

Rearranging,

$$p(x_b \mid x_a) = \sum_{k=1}^{K} \left(\sum_{j=1}^{K} \pi_j p(x_a \mid j) \pi_k p(x_a \mid k) \right) p(x_b \mid x_a, k)$$

Let new mixing coefficient $= \pi_k^*$,

$$\pi_k^*(x_a) = \sum_{j=1}^{K} \pi_j p(x_a \mid j) \pi_k p(x_a \mid k)$$

This conditional density becomes:

$$p(x_b \mid x_a) = \sum_{k=1}^{K} \pi_k^*(x_a) p(x_b \mid x_a, k)$$

It is in mixture model form proving that $p(x_b \mid x_a)$ is also a mixture distribution.

6)a)

$$p(x_n \mid \Theta) = \sum_{k=1}^{K} \pi_k N(x_n \mid \mu_k, \Sigma_k)$$

K is the number of Gaussian components.
π_k are the mixing coefficients,

$$\sum_{k=1}^{K} \pi_k = 1, \quad \pi_k > 0$$

$N(x_n \mid \mu_k, \Sigma_k)$ is the gaussian function with mean μ_k and covariance Σ_k

$$N(x \mid \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{d/2} |\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) \right)$$

Since $x_1, x_2, ...x_N$ are independent, the log-likelihood is:

$$\log p(X \mid \Theta) = \sum_{n=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k N(x_n \mid \mu_k, \Sigma_k) \right)$$

b) The complete data log-likelihood is :

$$\log p(X, Z \mid \Theta) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \log \left(\pi_k N(x_n \mid \mu_k, \Sigma_k) \right)$$

Expanding:

$$\log p(X, Z \mid \Theta) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \left[\log \pi_k + \log N(x_n \mid \mu_k, \Sigma_k) \right]$$

Using the constraint, $\sum_{k=1}^{K} \pi_k = 1$, the MLE extimate is : $\pi_k = \frac{N_k}{N}, \quad$ where $N_k = \sum_{n=1}^{N} z_{nk}$
Taking the derivative w.r.t μ_k and setting it to zero gives:

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} z_{nk} x_n$$

And Taking the derivative w.r.t Σ_k and setting it to zero:

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} z_{nk} (x_n - \mu_k)(x_n - \mu_k)^T$$

The MLE update rules:
1) Mixing Coefficients:

$$\pi_k = \frac{N_k}{N}, \quad N_k = \sum_{n=1}^{N} z_{nk}$$

2) Means:

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} z_{nk} x_n$$

3) Convariances:

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} z_{nk} (x_n - \mu_k)(x_n - \mu_k)^T$$

7)

8)

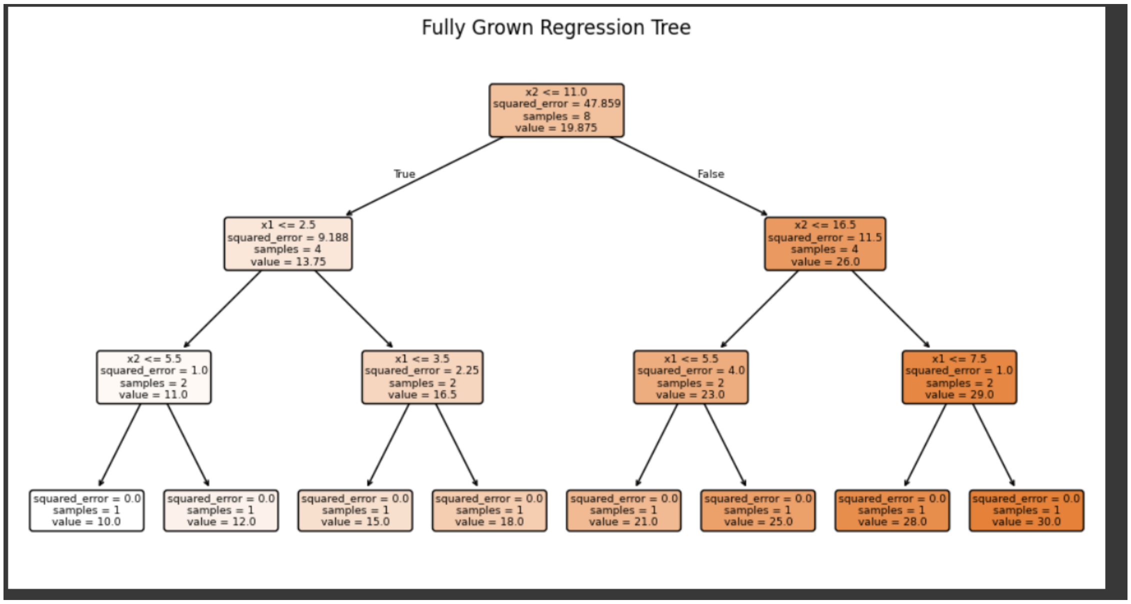

Figure 1: Fully grown regression tree

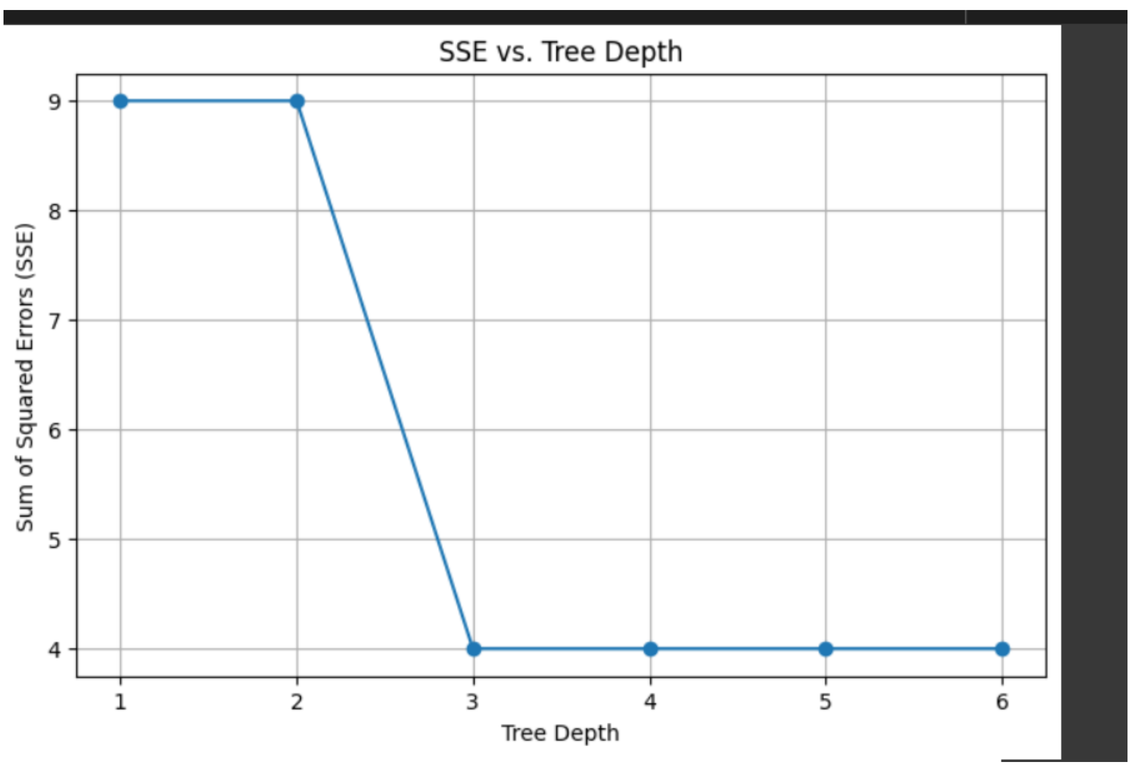

Figure 2: 8)SSE vs Tree Depth

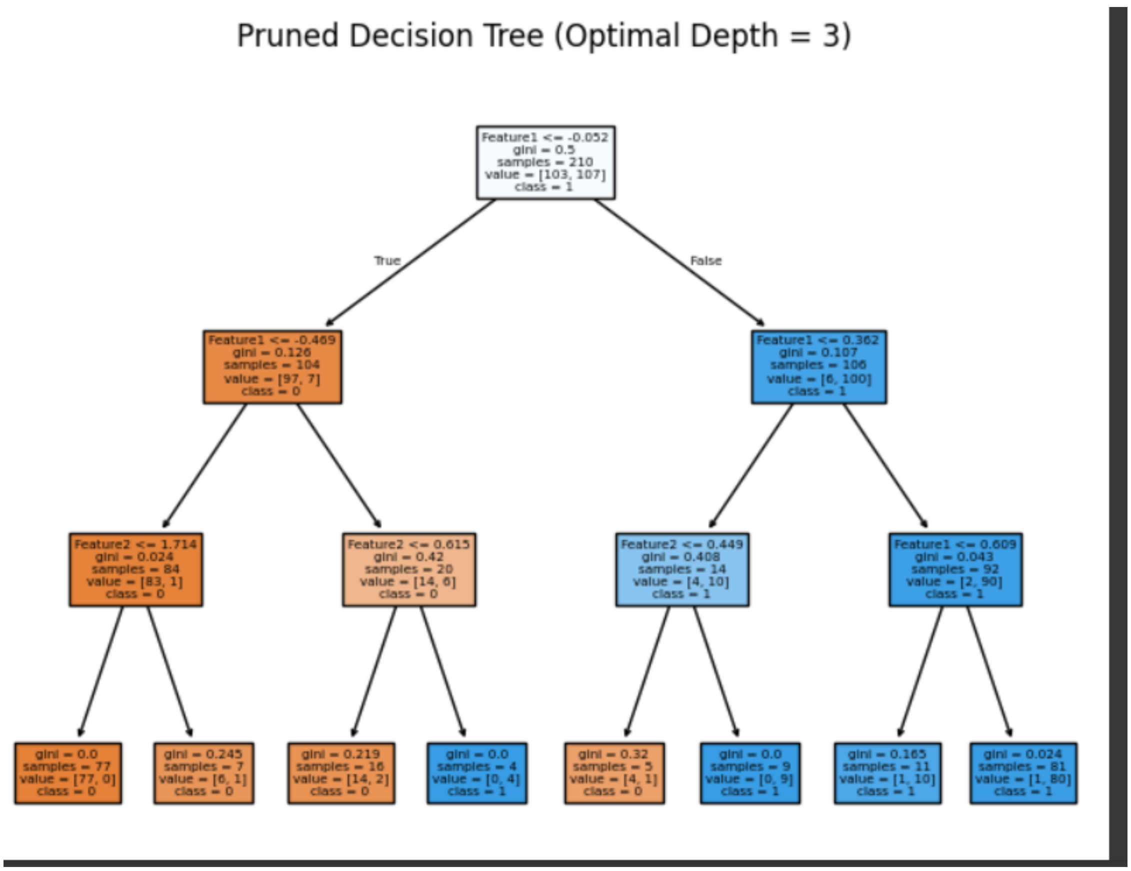

Figure 3: 8)Pruned Decision Tree