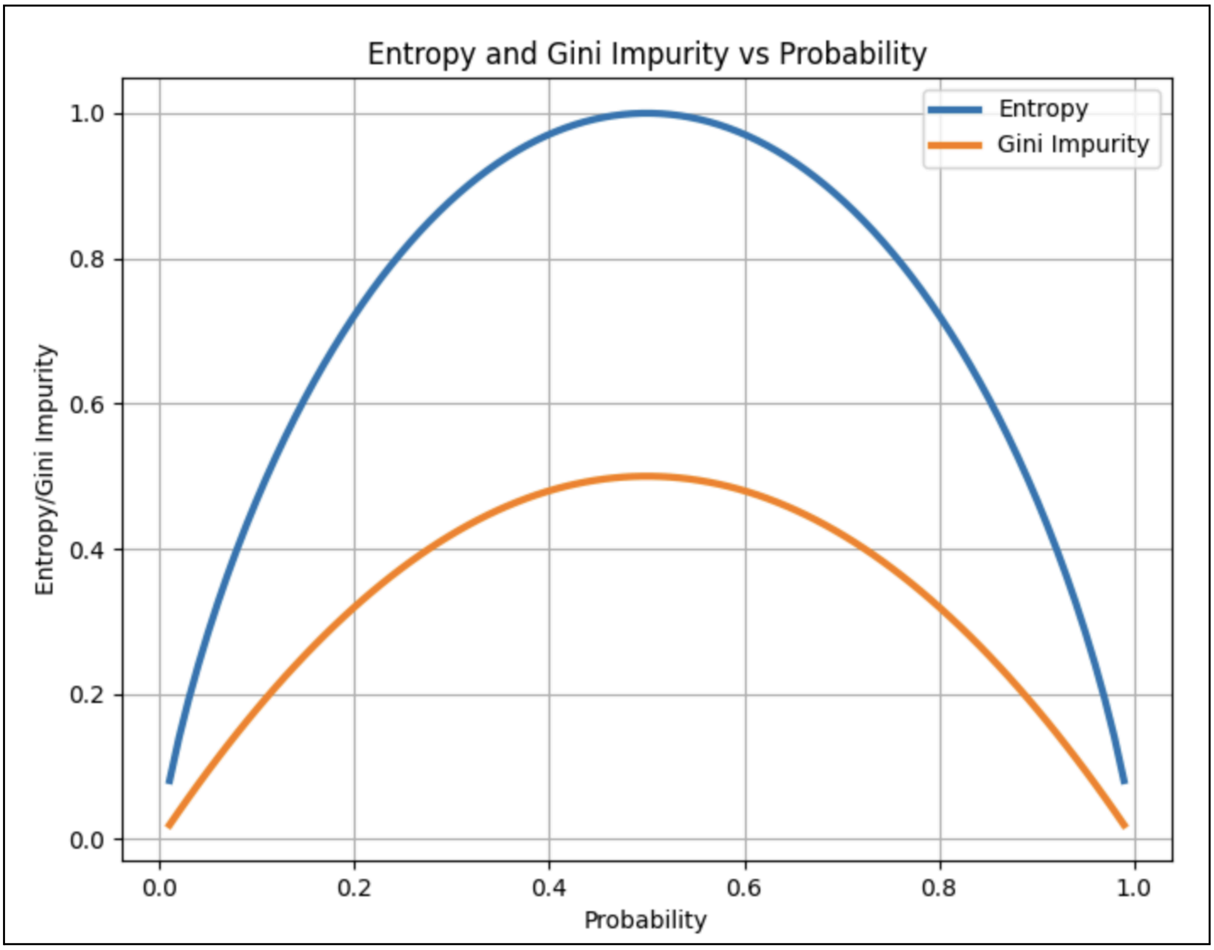
Homework 7 - CS 5805

Name : Jyothi Sevakula

Q1.



Q2.

Q2)

Outlook

Outlook	Yes	No	
Sunny	2	3	5
overcast	4	0	4
rain	3	2	5
	9	5	14 ✓

$E(\text{tennis}) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right)$

(parent)

$= 0.407 + 0.53057 = \underline{\underline{0.937}}$

$E(\text{sunny}) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)$

$= 0.5275 + 0.4405 = \underline{\underline{0.968}}$

$E(\text{overcast}) = -\frac{4}{4}\log_2\left(\frac{4}{4}\right) - 0$

$= 0$

$E(\text{Rain}) = -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right) = 0.968$

weighted entropy

$= \frac{0.968(5) + 0 + 0.968(5)}{14}$

$= 0.691$

Information gain (outlook)

$= 0.937 - 0.691$

$\boxed{IG(\text{tennis}) = 0.246}$

Temp

Temp	Yes	No	
Hot	2	2	4
mild	4	2	6
cool	3	1	4
	~~4~~	5	14

~~$E(\text{Temp}) = -\frac{9}{4}$~~

$E(\text{hot}) = -\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right) = 1$

$E(\text{mild}) = -\frac{4}{6}\log_2\left(\frac{4}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right) = 0.918$

$E(\text{cool}) = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) = 0.811$

weighted entropy

$= \frac{1 \times 4 + 0.918 \times 6 + 0.811 \times 4}{14}$

$= 0.91$

$IG(\text{Temp}) = 0.94 - 0.91$

$= 0.03$

$\boxed{IG(\text{Temp}) = 0.03}$

Humidity

Humidity	Yes	No	
High	3	4	7
Normal	6	1	7
	9	5	14

$E(\text{high}) = -\frac{3}{7}\log_2\left(\frac{3}{7}\right) - \frac{4}{7}\log_2\left(\frac{4}{7}\right) = 0.986$

$E(\text{normal}) = -\frac{6}{7}\log_2\left(\frac{6}{7}\right) - \frac{1}{7}\log_2\left(\frac{1}{7}\right) = 0.5917$

weighted entropy

$= \frac{0.986 \times 7 + 0.5917 \times 7}{14}$

$= 0.78885$

$IG(\text{Humidity}) = 0.94 - 0.78885$

$= 0.1485$

$\boxed{IG(\text{Humidity}) = 0.1485}$

wind

	Yes	NO	
weak	6	2	8
strong	3	3	6
	9	5	14

$E(weak) = -\frac{6}{8}\log_2\left(\frac{6}{8}\right) - \frac{2}{8}\log_2\left(\frac{2}{8}\right)$

$= 0.811$

$E(strong) = -\frac{3}{6}\log_2\left(\frac{3}{6}\right) \times 2 = 1$

→ weighted entropy =

$\frac{0.811 \times 8 + 1 \times 6}{14} = 0.892$

$IG(wind) = 0.937 - 0.892$

$= 0.045$

$\boxed{IG(window) = 0.045}$

~~E(wind)~~

On comparing all IG values, outlook has highest information gain so it will be the root node.

	IG
outlook	0.246
Temp	0.03
humidity	0.1485
wind	0.045

→ we pick this as root node.

→ let's see outlook on <u>sunny</u>

Temp

	Yes	NO	
Hot	0	2	2
mild	1	1	2
cold	1	0	1
	2	3	5

$E(sunny|Temp=hot) = 0 - \frac{2}{2}\log_2(1) = 0$

$E(Sunny|Temp=mild) = -\frac{1}{2}\log\left(\frac{1}{2}\right) \times 2 = 1$

$E(Sunny|Temp=cold) = \frac{1}{1}\log_2\left(\frac{1}{1}\right) = 0$

$E(Sunny|Temp) = 0 \times \frac{2}{5} + 1 \times \frac{2}{5} + 0 \times \frac{1}{5} = 0.4$

$IG(Sunny|Temp) = 0.971 - 0.4 = 0.571$

$IG(Sunny|Temp) = 0.571$

Humidity (outlook→sunny)

	Yes	NO	
high	0	3	3
normal	2	0	2
	2	3	5

$E(Sunny|Hum=high) = -\frac{3}{3}\log(1) = 0$

$E(Sunny|Hum=normal) = -\frac{2}{2}\log(1) = 0$

$E(sunny|Hum) = 0$

$IG(Sunny|humidity) = 0.971 - 0$

$= 0.971$

wind (outlook→sunny)

	Yes	NO	
strong	1	1	2
weak	1	2	3
	2	3	5

$E(Sunny|wind=strong) = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) \times 2 = 1$

$E(sunny|wind=weak) = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)$

$= 0.918$

~~IG(wind)~~

$IG(Sunny|wind) = 0.971 - \left(\frac{2}{5} \times 1 + \frac{2}{5} \times 0.918\right)$

$= 0.420$

on compaing all IG above we can see humidity is the highest info gain for outlook=sunny.

(outlook=sunny)	IG
Temp	0.571
Hum	0.971
wind	0.420

→ we pick this child

For outlook=sunny we can say it will first branch/divide to humidity

Humidity (outlook → sunny)

high → No
Normal → yes } direct labels

→ let's see outlook on overcast

E(overcast) = 0
↓
All labels are yes

→ let's see outlook on rain

wind	Yes	No	
weak	3	0	3
strong	0	2	2
	3	2	5

Temp	Yes	No	
mild	2	1	3
cool	1	1	2
	3	2	5

$E(Rain \mid wind = weak)$

$= -\frac{3}{3}\log_2(1) = 0$

$E(Rain \mid wind = strong)$

$= -\frac{2}{2}\log_2(1) = 0.$

$E(Rain \mid wind) = \frac{3}{5}\times 0 + \frac{3}{5}\times 0$

$IG(Rain \mid wind) = 0.968 - 0 = 0.968$

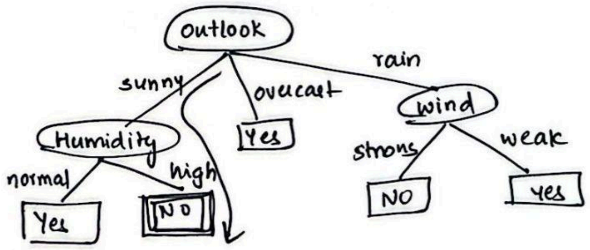
$E(rain \mid Temp = mild) = 0.9183$

$E(rain \mid Temp = cool) = 1$

$E(rain \mid temp) = 0.95$

$IG = 0.968 - 0.95 = 0.018$

∴ we choose wind as it has highest IG. for outlook=rain



∴ For given record, outlook=sunny, Temp=cool, humidity=high, wind=strong

∴ It results in "NO" therefore the player does not play Tennis

Q3.

Q3) Gini Impurity approach:-

Outlook	Yes	No	
sunny	2	3	5
overcast	4	0	4
rain	3	2	5
	9	5	14

Gini(parent | outlook=sunny)

$\cancel{\text{Gini(sunny)}} = 1-\left(\left(\frac{2}{5}\right)^2+\left(\frac{3}{5}\right)^2\right) = 0.48$

Gini(parent | outlook=overcast)

$= 1-\left(\frac{4}{4}\right)^2 = 0$

Gini(parent | outlook=rain) $= 1-\left(\left(\frac{3}{5}\right)^2+\left(\frac{2}{5}\right)^2\right)$

$= 0.48$

Gini(parent | outlook)

$= \frac{0.48\times5+0\times4+0.48\times5}{14} = 0.342$

$\boxed{\text{Gini(parent | outlook)} = 0.342}$

Temp	Yes	No	
Hot	2	2	4
mild	4	2	6
cool	3	1	4
	9	5	14

Gini(parent | Temp=hot) $= 1-\left(\left(\frac{1}{2}\right)^2\times2\right) = 0.5$

Gini(parent | Temp=mild) $= 1-\left(\left(\frac{4}{6}\right)^2+\left(\frac{2}{6}\right)^2\right) = 0.445$

Gini(parent | Temp=cool) $= 1-\left(\left(\frac{3}{4}\right)^2+\left(\frac{1}{4}\right)^2\right) = 0.375$

Gini(parent | Temp) $= \frac{0.5\times4+0.445\times6+0.375\times4}{14}$

$\boxed{\text{Gini(parent | Temp)} = 0.441}$

Humidity	Yes	No	
High	3	4	7
Normal	6	1	7
	9	5	14

Gini (parent | hum=high) $= 1-\left(\left(\frac{3}{7}\right)^2+\left(\frac{4}{7}\right)^2\right)$

$= 0.49$

Gini (parent | hum=normal) $= 1-\left(\left(\frac{6}{7}\right)^2+\left(\frac{1}{7}\right)^2\right)$

$= 0.245$

Gini(parent | humidity) $= \frac{0.49\times7+0.245\times7}{14}$

$\boxed{\text{Gini(pa | humidity)} \; 0.3675}$

Wind	Yes	No	
weak	6	2	8
strong	3	3	6
	9	5	14

Gini(parent | wind=weak) $= 1-\left(\left(\frac{6}{8}\right)^2+\left(\frac{2}{8}\right)^2\right)$

$= 0.375$

Gini(parent | wind=strong) $= 1-\left(\left(\frac{3}{6}\right)^2\times2\right)$

$= 0.5$

Gini(parent | wind) = 0.428

parent	Gini
outlook	0.342
Temp	0.441
hum	0.3675
wind	0.428

less gini index so we choose "outlook" as root node

∴ lets see outlook on sunny,

Temp

Temp	yes	NO	
hot	0	2	2
mild	1	1	2
cold	1	0	1
	2	3	5

$Gini(Sunny|Temp=hot) = 1-\left(\left(\frac{2}{2}\right)^2\right)=0$

$Gini(Sunny|Temp=mild) = 1-\left(\left(\frac{1}{2}\right)^2+\left(\frac{1}{2}\right)^2\right)=0.5$

$Gini(Sunny|Temp=cold) = 1-\left(\left(\frac{1}{1}\right)^2\right)=0$

$Gini(sunny|Temp) = \frac{0.5\times 2}{5} = 0.2$

$\boxed{Gini(Sunny|Temp)=0.2}$

Humidity

Humidity	yes	NO	
high	0	3	3
normal	2	0	2
	2	3	5

$Gini(Sunny|humidity=high) = 1-1=0$

$Gini(sunny|humidity=normal) = 1-\left(\left(\frac{2}{2}\right)^2\right)=0$

$\boxed{Gini(Sunny|Humidity)=0}$

wind

wind	yes	NO	
strong	1	1	2
weak	1	2	3
	2	3	5

$Gini(sunny|wind=strong) = 1-\left(\left(\frac{1}{2}\right)^2+\left(\frac{1}{2}\right)^2\right)=0.5$

$Gini(Sunny|wind=weak) = 1-\left(\left(\frac{1}{3}\right)^2+\left(\frac{2}{3}\right)^2\right)=0.44$

$\boxed{Gini(sunny|wind)=0.466}$

outlook=Sunny	Gini
Temp	0.2
hum	0
wind	0.466

→ we choose humidity as child of outlook on sunny.

For humidity high → ~~yes~~ no, normal → yes } direct labels.

- let's see outlook on overcast,

all records has value yes } direct label.

let's see outlook on ~~wind~~ rain

wind	yes	no	
weak	3	0	3
strong	0	2	2
	3	2	5

Temp	yes	NO	
mild	2	1	3
cold	1	1	2
	3	2	5

$Gini(Rain|wind=weak) = 1-\left(\left(\frac{3}{3}\right)^2\right)=0$

$Gini(Rain|wind=strong) = 1-\left(\left(\frac{2}{2}\right)^2\right)=0$

$\boxed{Gini(Rain|wind)=0}$

$Gini(Rain|Temp_{=mild}) = 1-\left(\left(\frac{2}{3}\right)^2+\left(\frac{1}{3}\right)^2\right) = 0.445$

$Gini(Rain|Temp=cold) = 1-\left(\frac{1}{2}\right) = 0.5$

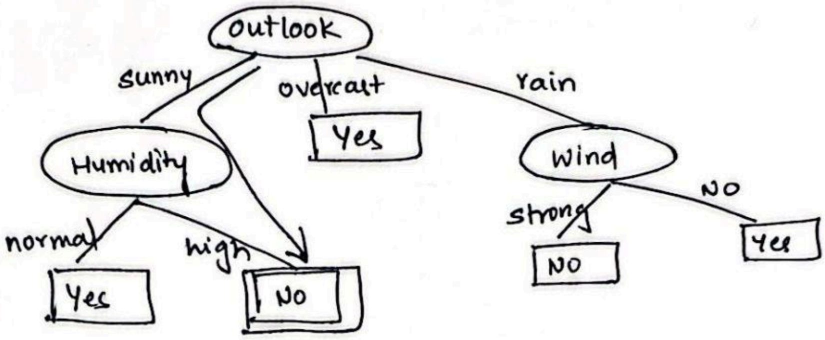
$Gini(Rain|Temp) = \frac{0.445\times 3+0.5\times 2}{5}$

$\boxed{Gini(Rain|Temp)=0.467}$

outlook=rain

	Gini
wind	0
Temp	0.467

we choose ~~rain~~ 'wind' as child of outlook on rain



For the record-test,

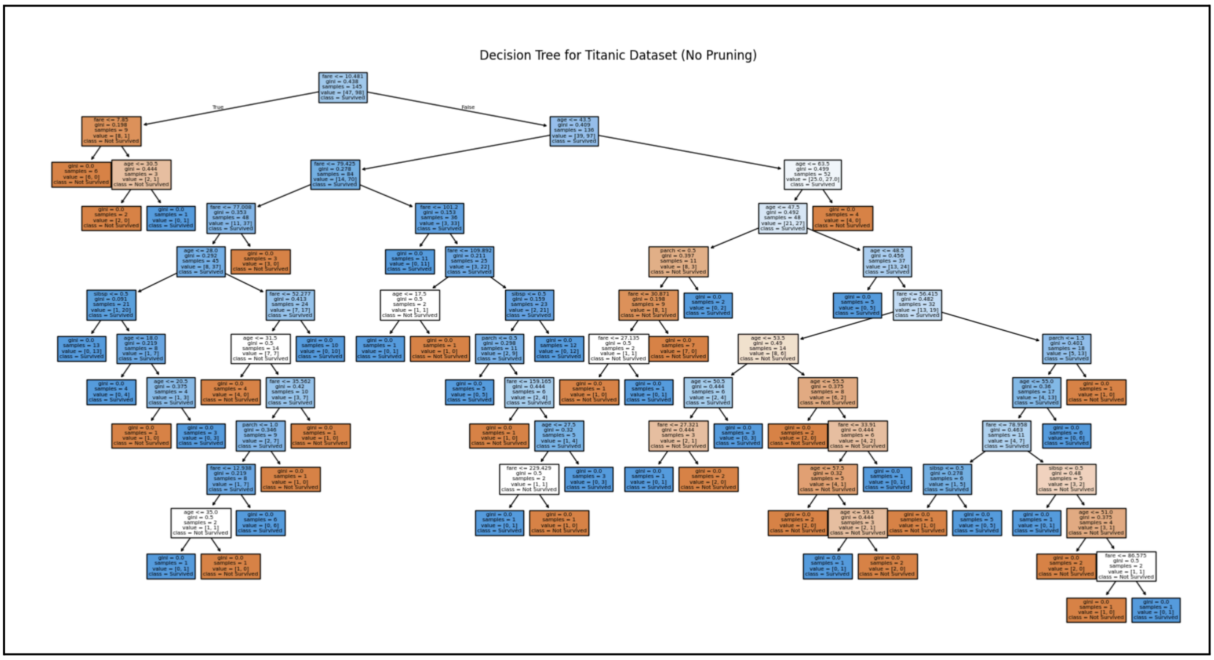
outlook=sunny, temp=cool, humidity = high & wind= strong.

$\therefore$ Based on above decision tree we get the player does not play tennis i.e "NO"

Q4.

```
Training Accuracy (No Pruning): 1.00
Test Accuracy (No Pruning): 0.62
Decision Tree Parameters: {'ccp_alpha': 0.0, 'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features': None, 'max_leaf_nodes': None, 'min_impur
```

Decision Tree Parameters: {'ccp_alpha': 0.0, 'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features': None, 'max_leaf_nodes': None, 'min_impurity_decrease': 0.0, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'monotonic_cst': None, 'random_state': 5805, 'splitter': 'best'}



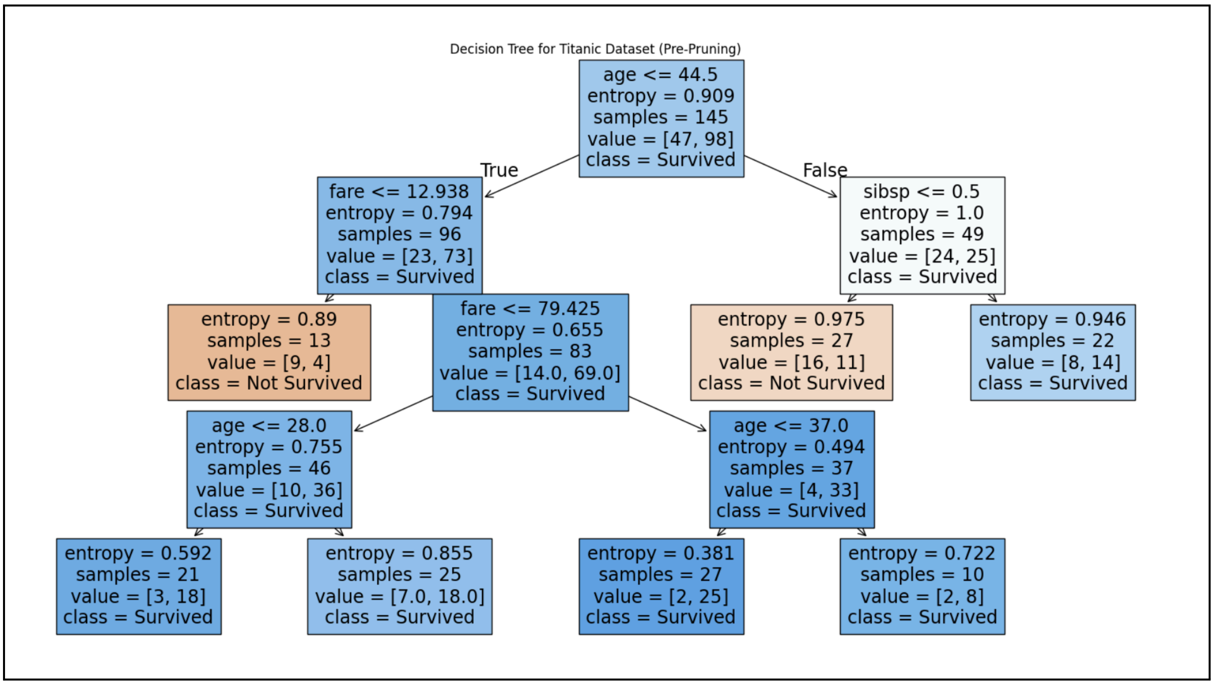
Comments → Overfitting the training data

The training accuracy is 1.00 (100%), indicating that the model has perfectly fit the training data. However, the test accuracy is significantly lower at 0.62 (62%). This suggests that the model has ***overfitted the training data***, capturing noise and details specific to the training data that do not generalize well to unseen data. The model has likely created many branches that lead to an overly complex decision boundary, which does not align well with the real underlying patterns in the data.

Q5.

```
Best Parameters: {'criterion': 'entropy', 'max_depth': 4, 'max_features': 'sqrt', 'min_samples_leaf': 10, 'min_samples_split': 30, 'splitter': 'best'}
Training Accuracy (Pre-pruned Tree): 0.74
Test Accuracy (Pre-pruned Tree): 0.73
```

Best Parameters: {'criterion': 'entropy', 'max_depth': 4, 'max_features': 'sqrt', 'min_samples_leaf': 10, 'min_samples_split': 30, 'splitter': 'best'}

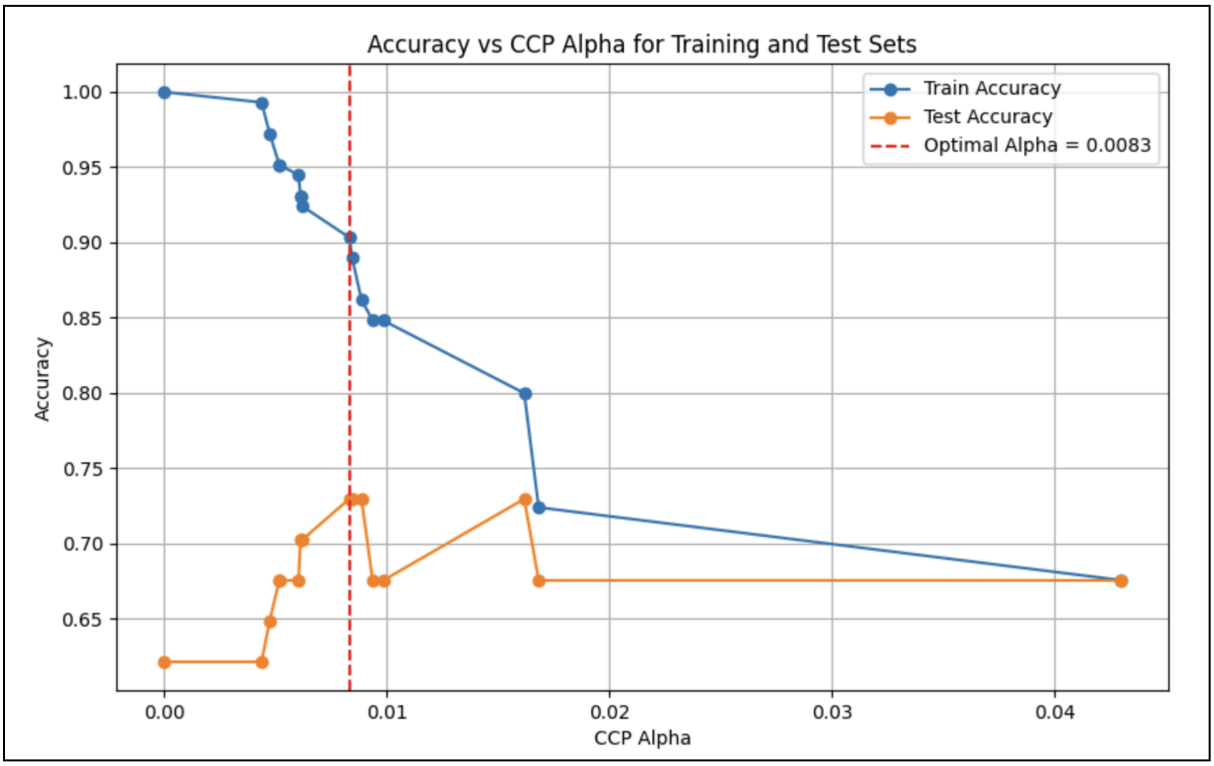


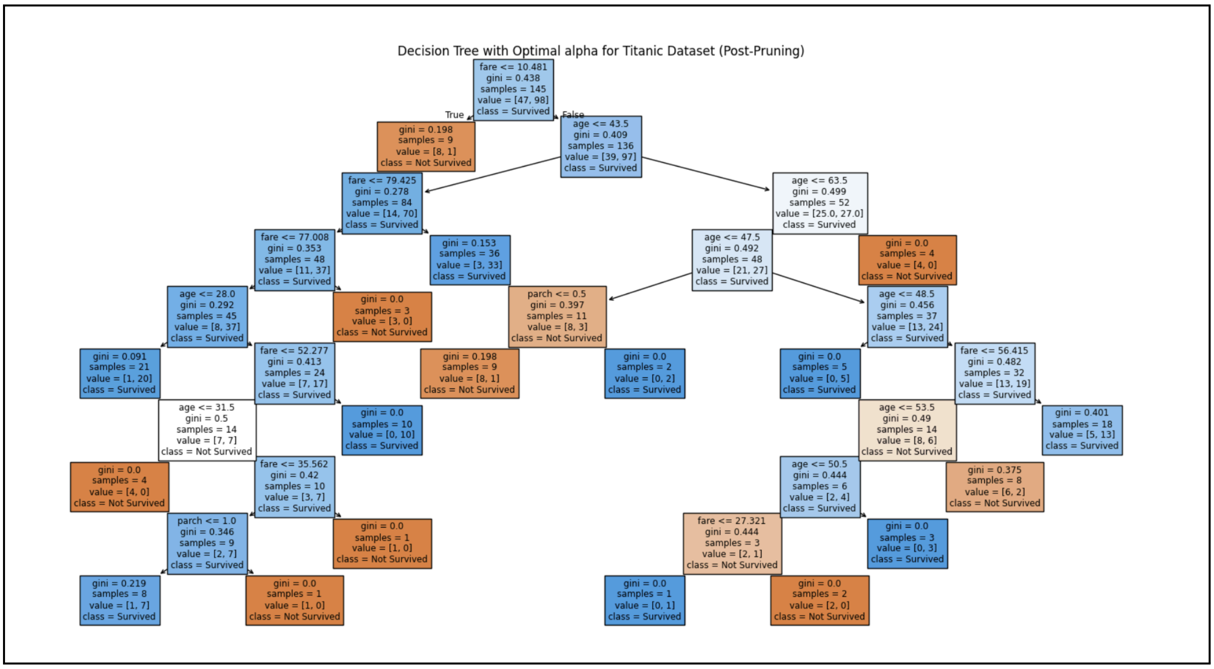
Comments → Improved model performance (reduced overfitting)

After applying pre-pruning using Grid Search with cross-validation.The training accuracy (0.74) and test accuracy (0.73) are very close. This suggests that the ***model is not overfitting***, as it performs similarly on both the training and test data. ***Pre-pruning improved the model's performance*** by reducing overfitting, leading to a more generalizable model. Although the training accuracy is slightly lower than in the no-pruning scenario, the consistency between training and test accuracies (0.74 and 0.73) indicates that the model is better suited for making reliable predictions on unseen data.

Q6.

```
Optimal CCP Alpha: 0.008314176245210173
Training Accuracy (Post-pruned Tree): 0.90
Test Accuracy (Post-pruned Tree): 0.73
```





Comments → Reduced overfitting when compared to no-pruned but less effective when compared to pre-pruned.

After applying post-pruning using optimum alpha in the cost complexity function.The training accuracy (0.9) and test accuracy (0.73). In short, ***Post-pruning reduced overfitting and improved test performance compared to the no-pruned model.*** However, ***it was slightly less effective at controlling overfitting compared to the pre-pruned model***, as indicated by the larger gap between training and test accuracies.

Comparison with Pre-Pruning:

- The pre-pruned model (training: 0.74, test: 0.73) achieved a better balance between training and test performance than the post-pruned model. With pre-pruning, the tree was constrained from growing too deep, leading to a simpler model that generalized better.
- In contrast, the post-pruning approach allows the tree to grow fully before pruning back some branches. Although post-pruning improves over the no-pruning model, it appears to be slightly less effective at controlling overfitting than pre-pruning in this case.

Comparison with No Pruning:

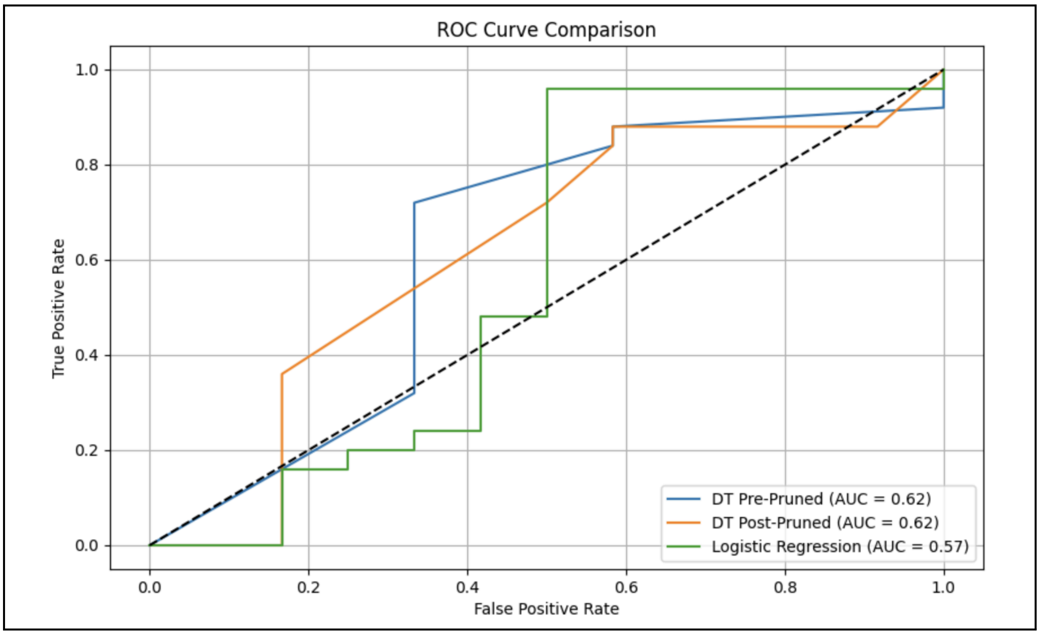
- Compared to the no-pruned model (training: 1.00, test: 0.62), post-pruning significantly reduces overfitting, improving the test accuracy from 0.62 to 0.73.
- The training accuracy has dropped from 1.00 to 0.90, showing that post-pruning effectively removes branches that capture noise or overly specific patterns in the training data.

Q7.

```
...
Training Accuracy (Logistic Regression): 0.73
Test Accuracy (Logistic Regression): 0.78
```

Q8.

Index	Model	Train Accuracy	Test Accuracy	Recall	AUC	Confusion Matrix
0	Pre-Pruning	0.74	0.73	0.88	0.62	[[5 7] [3 22]]
1	Post-Pruning	0.90	0.73	0.88	0.62	[[5 7] [3 22]]
2	Logistic Regression	0.73	0.78	0.96	0.57	[[5 7] [1 24]]



Comment: → Pre-pruning is the best classifier

Among DT Pre-pruned, DT Post-pruned and logistic regression comparing based on AUC, both DT Post-pruned and DT Pre-pruned have highest which is 0.62. Among post and pre pruned model, post-pruning model is little over-fitted on the training dataset hence the gap between train and test accuracy for post-pruning is higher. Hence, I would say Pre-pruning model classifier works best among all three model as the difference in test and train accuracy is very less when compared to others.(i.e 0.01) and also AUC is high.