REVIEW EXERCISE 08

Question 1. Roger Federer is one of the greatest tennis players since tennis have been invented. We want to learn a little about what makes Federer win or lose a match. To do so, we gathered data from games played by R. Federer, shown in the below table.

ID	Time	Match Type	Court Surface	Best Effort	Outcome
1	Morning	Master	Grass	True	Win
2	Afternoon	Grand slam	Clay	True	Win
3	Night	Friendly	Hard	False	Win
4	Afternoon	Friendly	Mixed	False	Lose
5	Afternoon	Master	Clay	True	Lose
6	Afternoon	Grand slam	Grass	True	Win
7	Afternoon	Grand slam	Hard	True	Win
8	Afternoon	Grand slam	Hard	True	Win
9	Morning	Master	Grass	True	Win
10	Afternoon	Grand slam	Clay	True	Lose
11	Night	Friendly	Hard	False	Win
12	Night	Master	Mixed	True	Lose
13	Afternoon	Master	Clay	True	Lose
14	Afternoon	Master	Grass	True	Win
15	Afternoon	Grand slam	Hard	True	Win
16	Afternoon	Grand slam	Clay	True	Win

a. Build a classification model using ID3 decision tree from the gathered data.

The entropy of the whole dataset

- $H(Dataset) = -11/16*\log_2 11/16 – 5/16*\log_2 5/16 = 0.896$ (11 Win – 5 Lose)

The information gain of the attribute Time

- $H(Time = Morning) = 0$ (2 Win – 0 Lose)

- $H(Time = Afternoon) = -7/11*\log_2 7/11 – 4/11*\log_2 4/11 = 0.946$

  (7 Win – 4 Lose)

- $H(Time = Night) = -2/3*\log_2 2/3 – 1/3*\log_2 1/3 = 0.918$ (2 Win – 1 Lose)

- $AE(Time) = 2/16*0 + 11/16*0.946 + 3/16*0.918 = 0.823$

- $IG(Time) = 0.896 – 0.823 = 0.073$

The information gain of the attribute Match Type

- $H(Match Type = Master) = 1$ (3 Win – 3 Lose)

- $H(\text{Match Type = Grand slam}) = -6/7*\log_2 6/7 - 1/7*\log_2 1/7 = 0.592$

<div align="right">(6 Win – 1 Lose)</div>

- $H(\text{Match Type = Friendly}) = -2/3*\log_2 2/3 - 1/3*\log_2 1/3 = 0.918$

<div align="right">(2 Win – 1 Lose)</div>

- $AE(\text{Match Type}) = 6/16*1 + 7/16*0.592 + 3/16*0.918 = 0.806$
- $IG(\text{Match Type}) = 0.896 - 0.806 = 0.09$

The information gain of the attribute Court Surface

- $H(\text{Court Surface = Grass}) = 0$ (4 Win – 0 Lose)
- $H(\text{Court Surface = Hard}) = 0$ (5 Win – 0 Lose)
- $H(\text{Court Surface = Clay}) = -2/5*\log_2(2/5) - 3/5*\log_2 3/5 = 0.971$

<div align="right">(2 Win – 3 Lose)</div>

- $H(\text{Court Surface = Mixed}) = 0$ (0 Win – 2 Lose)
- $AE(\text{Court Surface}) = 5/16*0.971 = 0.303$
- $IG(\text{Court Surface}) = 0.896 - 0.303 = 0.593$

The information gain of the attribute Best Effort

- $H(\text{Best Effort = True}) = -9/13*\log_2 9/13 - 4/13*\log_2 4/13 = 0.890$

<div align="right">(9 Win – 4 Lose)</div>

- $H(\text{Best Effort = False}) = -2/3*\log_2 2/3 - 1/3*\log_2 1/3 = 0.918$

<div align="right">(2 Win – 1 Lose)</div>

- $AE(\text{Best Effort}) = 13/16*0.890 + 3/16*0.918 = 0.895$
- $IG(\text{Best Effort}) = 0.001$

The root attribute will be Court Surface due to its largest IG.
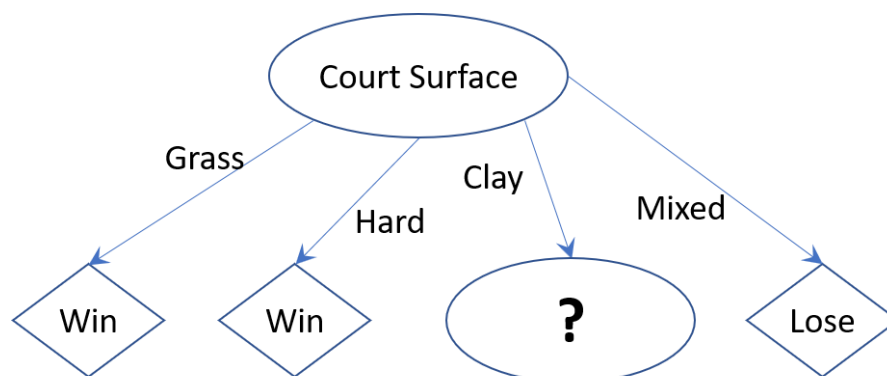
Repeat for every branch of the root attribute that has examples not fully classified into a single class

Consider the branch Court Surface = Clay

- $H(D_{\text{Court Surface = Clay}}) = 0.971$

The information gain of the attribute Time

- AE(Time, $D_{\text{Court Surface = Clay}}$) = 0.971

- IG(Time, $D_{\text{Court Surface = Clay}}$) = 0.971 – 0.971 = 0
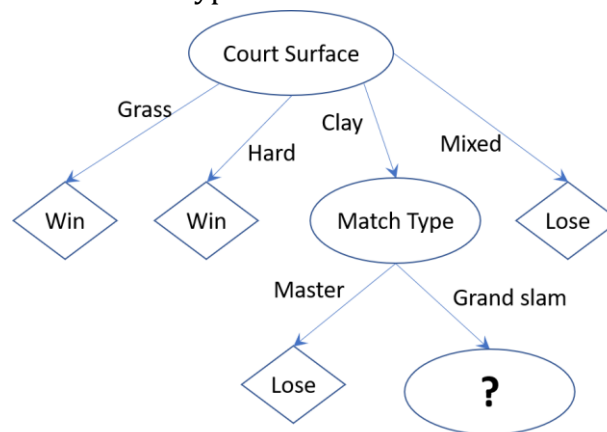
The information gain of the attribute Match Type

- H(Match Type = Master) = 0   (0 Win – 2 Lose)

- H(Match Type = Grand slam) = $-2/3*\log_2 2/3 - 1/3*\log_2 1/3$ = 0.918

                                                              (2 Win – 1 Lose)

- AE(Match Type) = $2/5*0 + 3/5*0.918$ = 0.551

- IG(Match Type) = 0.971 – 0.551 = 0.42

The information gain of the attribute Best Effort

- AE(Best Effort, $D_{\text{Court Surface = Clay}}$) = 0.971

- IG(Best Effort, $D_{\text{Court Surface = Clay}}$) = 0.971 – 0.971 = 0

Thus, the chose attribute is Match Type

Consider the branch Court Surface = Clay $\rightarrow$ Match Type = Grand slam.

For Time and Best Effort, each attribute has a single value. Thus, it is pointless to continue to grow the tree. Instead, we use the rule of Majority Voting.
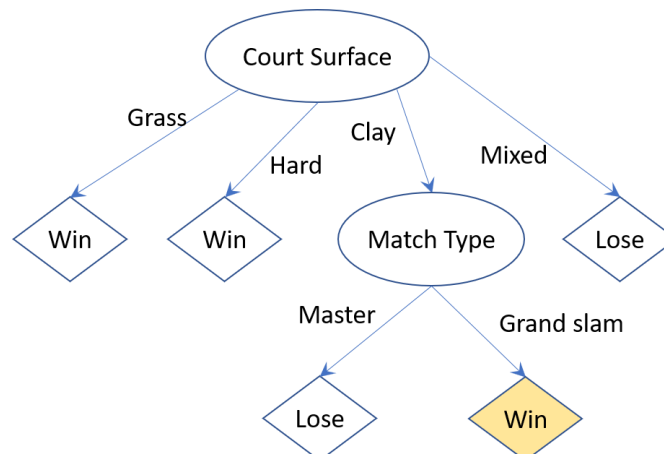
b. Knowing the conditions in which a tennis match takes place, we would like to predict whether R. Federer will win or lose the match, using ID3 decision tree in a.

Time	Match type	Court Surface	Best Effort	Outcome
Morning	Grand Slam	Grass	Yes	?
Afternoon	Friendly	Clay	No	?

The first example: From the root Court Surface, follow the branch Court Surface = Grass to the conclusion Outcome = Win

The second example: From the root Court Surface, follow the branch Court Surface = Clay to the attribute Match Type. We have no such branch Match Type = Friendly there. What should we do? Choose either of the following two solutions

- Conclude that Outcome cannot be decided due to the lack of training data
- Follow Rule 3 (Lecture 09, slide 40) to create a default value: Among 16 examples used to construct the root node, there are 3 examples that have Match Type = Friendly. Among those 3 examples, there are 2 Win and 1 Lose. Thus, the label created for Court Surface = Clay → Match Type = Friendly is Win.