ANSWERS

Exercise 1

i We start by calculating the entropy of the whole unsplit set and the information gain values that would be achieved for each of the variables if the set were split on that variable.

$$entropy(whole\_set) = -P(yes)log_2P(yes) - P(no)log_2P(no) =$$
$$-\frac{7}{12}log_2\frac{7}{12} - \frac{5}{12}log_2\frac{5}{12} = 0.98$$

Alternatively, entropy values for two-class outcomes can be looked up in the table headed "Entropy for subsets with various relative percentage combinations", found in the appendix of the lab sheet. This table gives 'ready-calculated' entropies for different proportions of the two classes in a set. The calculation above corresponds to the following line in the table:

Number of items	Breakdown		Entropy
	...		
12	5	7	0.98

From here on we will be looking up entropy values in the table rather than calculating them.

There are three variables on which the top level set could be split: `head shape`, `body shape` and `body colour`. For each of these we calculate the entropies of the variable values and then the information gain that would result if the set were split on that variable. We look up the table to find the entropy for the yes/no split ratio.

- `head shape`

  $entropy(head\ shape : square) = 0.99$ (9/12 instances, yes/no split: 5,4)

  $entropy(head\ shape : round) = 0.92$ (3/12 instances, yes/no split: 2,1)

  $IG(head\ shape) = entropy(whole\ set) - 9/12 \times 0.99 - 3/12 \times 0.92 = 0.01$

- `body shape`

  $entropy(body\ shape : rectangular) = 0.65$ (6/12 instances, yes/no split: 5,1)

  $entropy(body\ shape : oval) = 0.92$ (6/12 instances, yes/no split: 2,4)

  $IG(Temp) = entropy(whole\ set) - 6/12 \times 0.65 - 6/12 \times 0.92 = 0.2$

- `body colour`

  $entropy(body\ colour : white) = 0.97$ (10/12 instances, yes/no split: 6,4)

  $entropy(body\ colour : grey) = 1$ (2/12 instances, yes/no split: 1,1)

  $IG(Temp) = entropy(whole\ set) - 10/12 \times 0.97 - 2/12 \times 1 = 0.005$

The variable that should be used first for splitting the data is `body shape`, as it has the highest information gain, 0.2, with respect to the target variable `Likely to default on mortgage`.

After the first splitting, the tree looks like this:

```
            whole set
            /        \
body shape:          body shape:
rectangular              oval
    ?                     ?
```

ii Now we'll continue to split the set until we have a complete tree.

First let's look at the `rectangular` branch, which needs to be split further.

- `head shape`

  $entropy(head\ shape : square|rectangular) = 0$ (5/6 instances, yes/no split: 6,0)

  $entropy(head\ shape : round|rectangular) = 0$ (1/6 instances, yes/no split: 0,1)

- `body colour`

  $entropy(body\ colour : white|rectangular) = 0$ (5/6 instances, yes/no split: 6,0)

  $entropy(body\ colour : grey|rectangular) = 0$ (1/6 instances, yes/no split 0,1)

Note that when the split is between 0 and another number, the entropy is 0. With both atrributes above the information gain is maximal, the entropy being 0, so we pick one of the two attributes at random, let's say `head shape`, to split at the `body shape:rectangular` node. Since its entropy is 0, this attribute defines leaf nodes, with values for the target variable.

```
                    whole set
                    /        \
            body shape:      body shape:
            rectangular          oval
            /        \             ?
head shape:   head shape
   square        round
   [Yes]         [No]
```

Now let's look at the `oval` branch, which needs to be split further.

- `head shape`

  $entropy(head\ shape : square|oval) = 0$ (4/6 instances, yes/no split: 0,4)

  $entropy(head\ shape : round|oval) = 0$ (2/6 instances, yes/no split: 2,0)

- `body colour`

  $entropy(body\ colour : white|oval) = 0.72$ (5/6 instances, yes/no split: 1,4)

  $entropy(body\ colour : grey|oval) = 0$ (1/6 instances, yes/no split: 1,0)

The attribute `head shape` has an entropy of 0 for `body shape:oval`. We pick it as the next splitting attribute. The decision tree now looks like this:

```
                    whole set
                    /        \
            body shape:      body shape:
           rectangular           oval
            /        \           /      \
head shape:   head shape  head shape:  head shape:
   square        round       square       round
   [Yes]         [No]        [No]         [Yes]
```

iii To get the prediction for a customer with `head shape:square, body shape:rectangular, body colour:gray` we navigate the derived decision tree along the following nodes:

- `whole set`

- `body shape:rectangular`
- `head shape:square`

The last listed node is a leaf node associated with the `Yes` class i.e. the tree predicts that the customer is likely to default on the mortgage.

Exercise 2

a) **Finding the best splitter attribute for the whole set**

For the whole set we read the entropy out of the table (split 5, 9).

$entropy(whole\_set) = 0.94$ (split 5, 9)

Now we look at the different attributes to see which one will provide the best Information Gain. We read the entropies from the table, based on the "No"/"Yes" split of the "Play" values in the subset.

(a) **Outlook**

$entropy(Outlook : Sunny) = 0.97$ (subset size: 5/14, split: 2, 3)
$entropy(Outlook : Overcast) = 0$ (subset size: 4/14, split: 4, 0)
$entropy(Outlook : Rainy) = 0.97$ (subset size: 5/14, split: 3, 2)
$IG(Outlook) = entropy(whole\_set) - 5/14 * 0.97 - 4/14 * 0 - 5/14 * 0.97 = 0.25$

(b) **Temp**

$entropy(Temp : Hot) = 1$ (subset size: 4/16, split: 2, 2)
$entropy(Temp : Mild) = 0.92$ (subset size: 6/14, split: 4, 2)
$entropy(Temp : Cool) = 0.81$ (subset size: 4/14, split: 3, 1)
$IG(Temp) = entropy(whole\_set) - 4/14 * 1 - 6/14 * 0.92 - 4/14 * 0.81 = 0.03$

(c) **Humidity**

$entropy(Humidity : High) = 0.99$ (subset size: 7/14, split: 3, 4)
$entropy(Humidity : Normal) = 0.59$ (subset size: 7/14, split: 6, 1)
$IG(Humidity) = entropy(whole\_set) - 7/14 * 0.99 - 7/14 * 0.59 = 0.15$

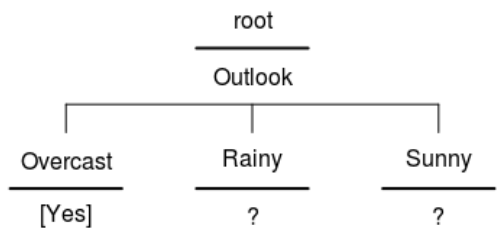
(d) **Windy**

$entropy(Windy : True) = 1$ (subset size: 6/14, split: 3, 3)
$entropy(Windy : False) = 0.81$ (subset size: 8/14, split: 6, 2)
$IG(Windy) = entropy(whole\_set) - 6/14 * 1 - 8/14 * 0.81 = 0.05$

The variable that should be used first for splitting the data is 'Outlook', as it has the highest information gain, 0.25, with respect to the target variable 'Play'.

b) **Building the rest of the tree**

After the first split, the tree looks like this:



The Overcast branch has an entropy of 0 and does not need to be split further.

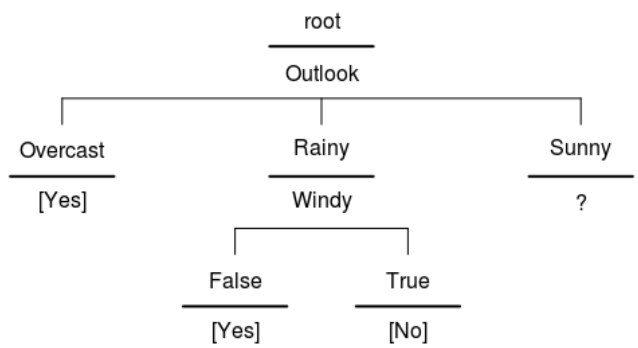
Now we look at the Rainy branch, which needs to be split further.

$entropy(Temp : Mild|Rainy) = 0.92$ (subset size: 3/5, split: 2, 1)
$entropy(Temp : Cool|Rainy) = 1$ (subset size: 2/5, split: 1, 1)

$entropy(Humidity : High|Rainy) = 1$ (subset size: 2/5, split: 1, 1)
$entropy(Humidity : Normal|Rainy) = 0.92$ (subset size: 3/5, split: 2, 1)

$entropy(Windy : True|Rainy) = 0$ (subset size: 2/5, split: 0, 2)
$entropy(Windy : False|Rainy) = 0$ (subset size: 3/5, split: 3, 0)

The IG doesn't need to be calculated as there is only one attribute with entropy of 0 and that is Windy. This means that this attribute provides the greatest information gain and we choose it to split the node 'Rainy'. The decision tree now looks like this:



The False and True branches both have entropy 0 and do not need to be split further.

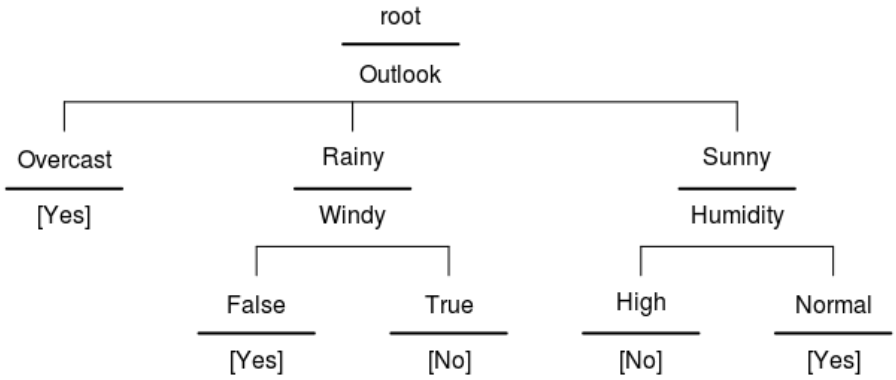
Finally, we look at the Sunny branch, which does need to be split further.

$entropy(Temp : Hot|Sunny) = 0$ (subset size: 2/5, split 2, 0)
$entropy(Temp : Mild|Sunny) = 1$ (subset size: 2/5, split 1, 1)
$entropy(Temp : Cool|Sunny) = 0$ (subset size: 1/5, split 1, 0)

$entropy(Humidity : High|Sunny) = 0$ (subset size: 3/5, split 3, 0)
$entropy(Humidity : Normal|Sunny) = 0$ (subset size: 2/5, split 2, 0)

$entropy(Windy : True|Sunny) = 1$ (subset size: 2/5, split 1, 1)
$entropy(Windy : False|Sunny) = 0.92$ (subset size: 3/5, split 2, 1)

When it's Sunny, the entropy for Humidity is 0 and non zero for the other attributes, so Humidity has the highest IG and should be chosen to split the node 'Sunny'. The decision tree finally looks like this:



c) **Predicting the outcome for a particular instance**

To predict the outcome for a particular instance, we traverse the tree using the attribute values of the instance. The given instance is:

```
Outlook:   Sunny
Temp:      Hot
Humidity:  Normal
Windy:     True
```

The first split in the tree is by `Outlook` and as the instance has the value `Sunny` for this attribute, at the first branching we take the rightmost branch. The next split is by `Humidity` and the given instance has the value `Normal`, which means that the traversal is along the rightmost branch again, arriving at a leaf node with a probability-1 outcome of "Yes".

For the given instance, our model predicts that the game will go ahead i.e. Play="Yes".