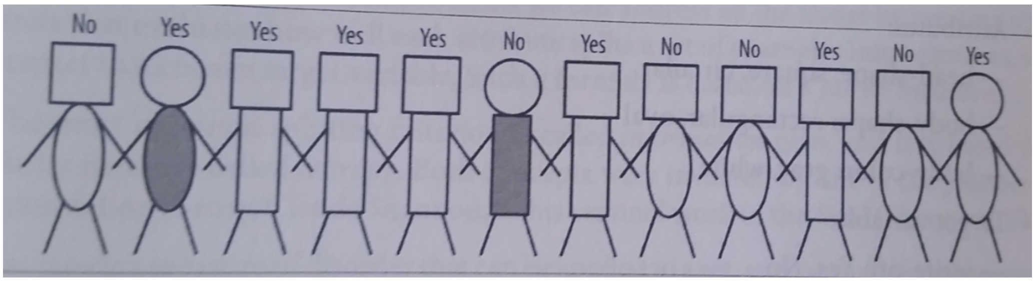
Exercise 1 - Classification Tree

The picture shows the customers of a bank, each representing a data instance in the bank's database. The customer attributes are abstracted into graphical elements: **head shape**, **body shape** and **body colour**. Above each customer, the known indicator of whether they defaulted on a loan is given as **Yes** (defaulted) or **No** (did not default).



Source: [DSB]

i Using *Information Gain* (IG) as the criterion, find which of the three attributes would be the best as the first splitting attribute in the creation of a classification tree that predicts default.

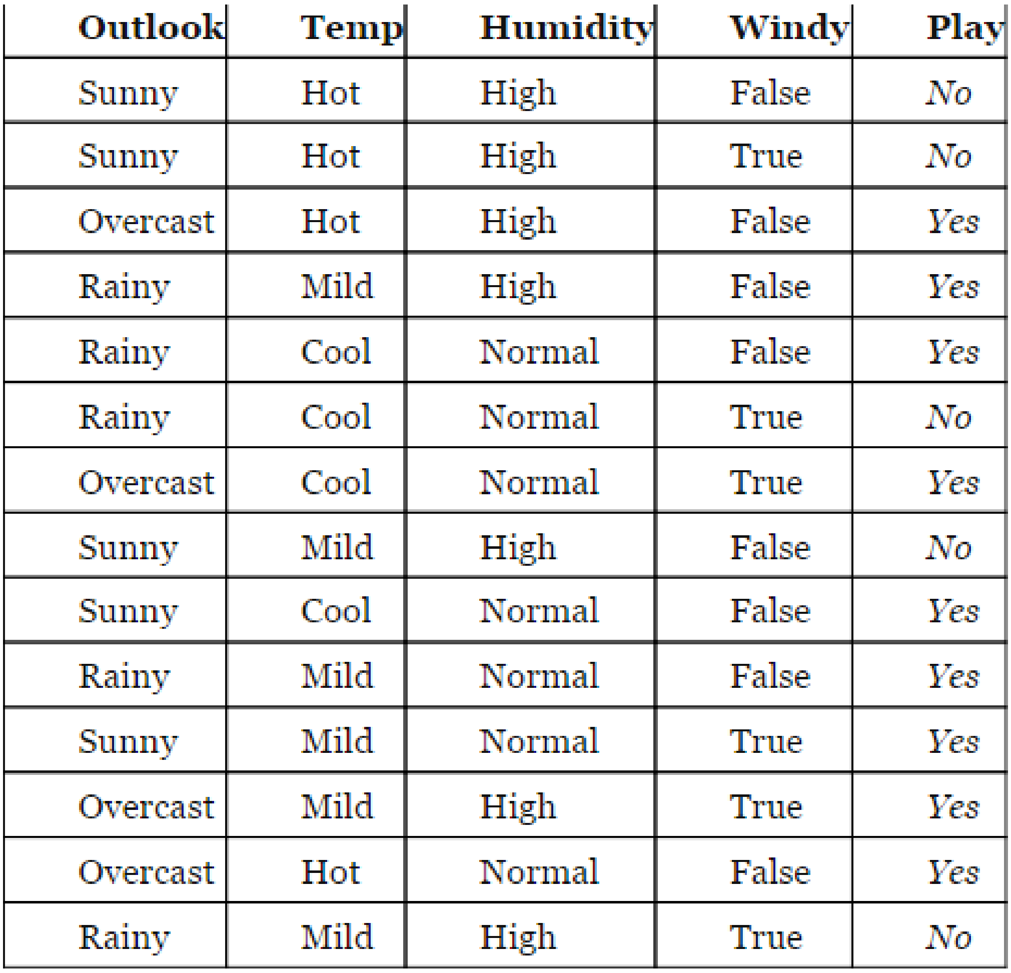
ii Continue the process of splitting the data set until the leaf nodes are homogeneous. Illustrate the resulting decision tree, marking the variables and values clearly.

iii Use your decision tree to predict if a customer with the following attribute values will default on their loan:

- head shape: square
- body shape: rectangular
- body colour: gray

Exercise 2 - Classification Tree

The following table shows data relating weather attributes with whether soccer games were played or not.



Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	*No*
Sunny	Hot	High	True	*No*
Overcast	Hot	High	False	*Yes*
Rainy	Mild	High	False	*Yes*
Rainy	Cool	Normal	False	*Yes*
Rainy	Cool	Normal	True	*No*
Overcast	Cool	Normal	True	*Yes*
Sunny	Mild	High	False	*No*
Sunny	Cool	Normal	False	*Yes*
Rainy	Mild	Normal	False	*Yes*
Sunny	Mild	Normal	True	*Yes*
Overcast	Mild	High	True	*Yes*
Overcast	Hot	Normal	False	*Yes*
Rainy	Mild	High	True	*No*

Source: [**DAMA**]

a) Using information gain (IG) as criterion, determine which of the four weather attributes (Outlook, Temp, Humidity, Windy) should be used as the first splitting variable for the data set in the table. (You can use the table at the end of this document, which lists entropy values for some probability combinations.)

b) Continue the process of splitting the data set given in section (b) in order to produce a complete decision tree. Illustrate the decision tree, clearly marking variables in the tree nodes and variable values in the tree branches.

c) Use the decision tree obtained in the previous section to predict whether a soccer game will be played when the weather attributes have the following values:

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

Entropy table

Number of items	Breakdown		Entropy
1	0	1	0.00
2	1	1	1.00
3	1	2	0.92
4	1	3	0.81
5	1	4	0.72
5	2	3	0.97
6	1	5	0.65
7	1	6	0.59
7	2	5	0.86
7	3	4	0.99
8	1	7	0.54
8	3	5	0.95
9	1	8	0.50
9	2	7	0.76
9	4	5	0.99
10	1	9	0.47
10	3	7	0.88
11	1	10	0.44
11	2	9	0.68
11	3	8	0.85
11	4	7	0.95
11	5	6	0.99
12	1	11	0.41
12	5	7	0.98
13	1	12	0.39
13	2	11	0.62
13	3	10	0.78
13	4	9	0.89
13	5	8	0.96
13	6	7	1.00
14	1	13	0.37
14	3	11	0.75
14	5	9	0.94
15	1	14	0.35
15	2	13	0.57
15	4	11	0.84
15	6	9	0.97
15	7	8	1.00

References

[**DSB**] *Data Science for Business: What you need to know about data mining and data-analytic thinking*, by Foster Provost and Tom Fawcett, O'Reilly Media, 2013.

[**DAMA**] *Data Analytics Made Accessible*, by Anil Maheshwari, Kindle Direct Publishing eBook, 2016.