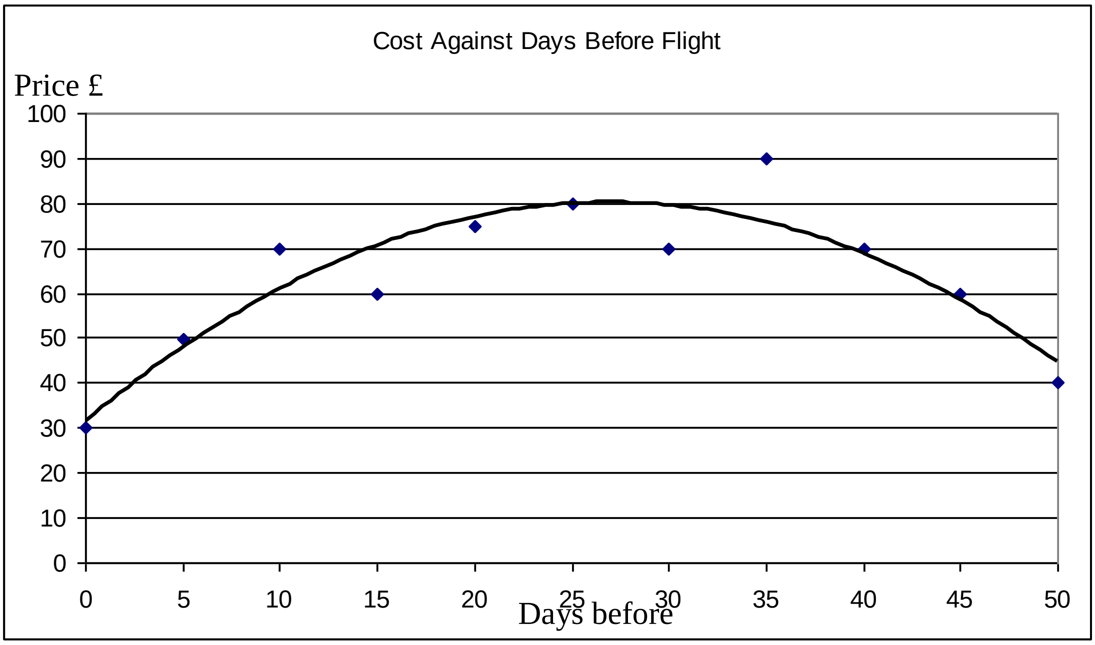
# T6 DM Tutorial on “Classification & Prediction”– Answers

## 1. Prediction

Take the data in the table below, plot each point in the empty graph below. Label the two axes correctly.

Draw a curve onto the chart that models the relationship between the number of days before a flight a ticket is bought and the cost of the flight

| Days before | Cost |
|---|---|
| 0 | 30 |
| 5 | 50 |
| 10 | 70 |
| 15 | 60 |
| 20 | 75 |
| 25 | 80 |
| 30 | 70 |
| 35 | 90 |
| 40 | 70 |
| 45 | 60 |
| 50 | 40 |

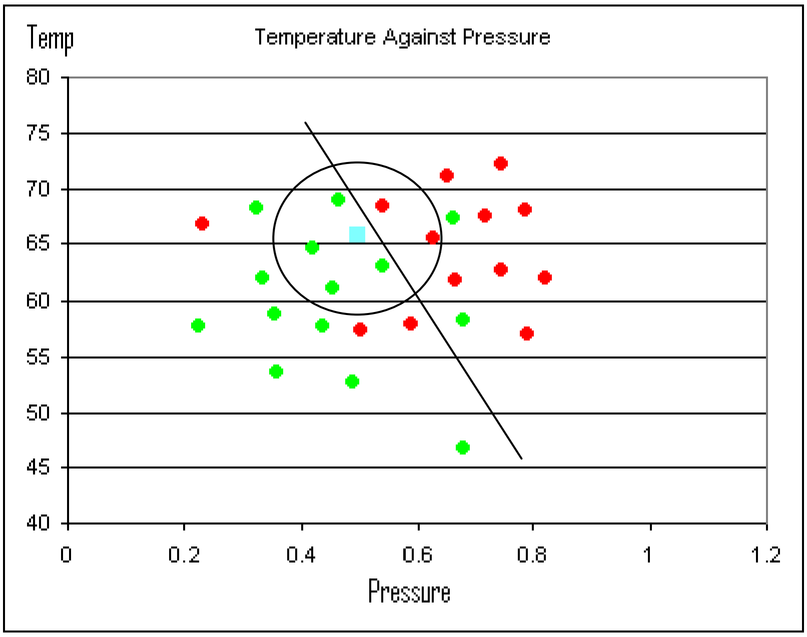


Now complete the table below, using the model you just drew above to make predictions for each day.

| Days before | Cost |
|---|---|
| 3 | 40 |
| 12 | 62 |
| 24 | 79 |
| 46 | 58 |
| 48 | 46 |

**2. Classification**

1) Look at the scatter plot below. It shows the state of a machine given a temperature and pressure reading. Red dots indicate the machine failed, green dots indicate the machine worked properly. Ignore the blue square for a moment.

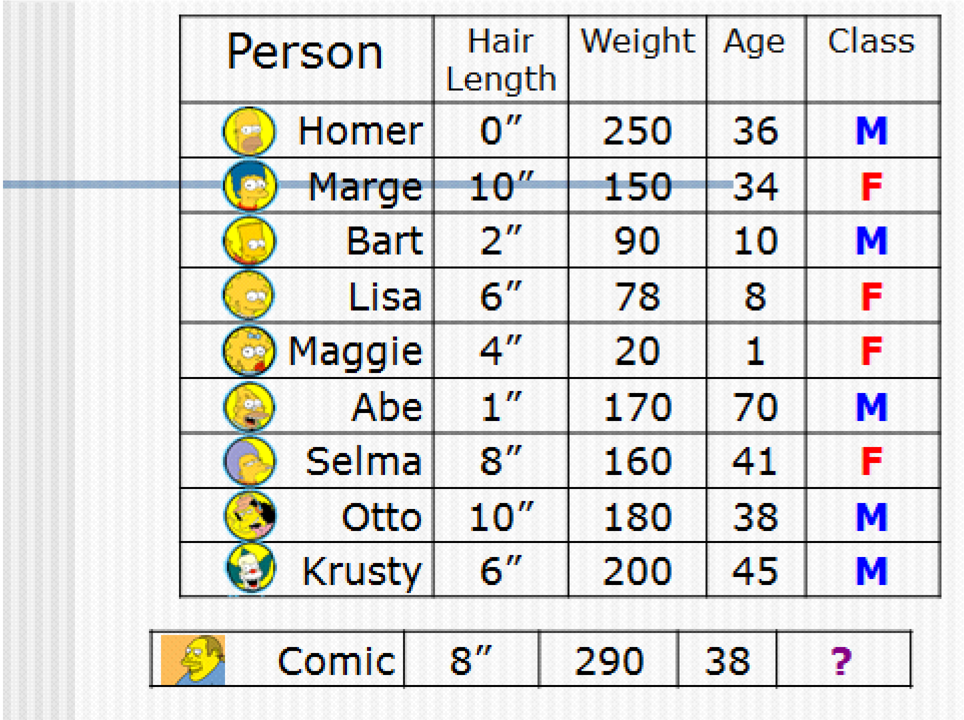


Draw a linear separator across the data that minimises the classification error given pressure and temperature. How many incorrect classifications does this model make on the given data? = 5 errors

Now look at the blue square. Perform a K-Nearest Neighbour classification of the blue square where K=6. What is the most likely class for the machine in this state (working or failed)?

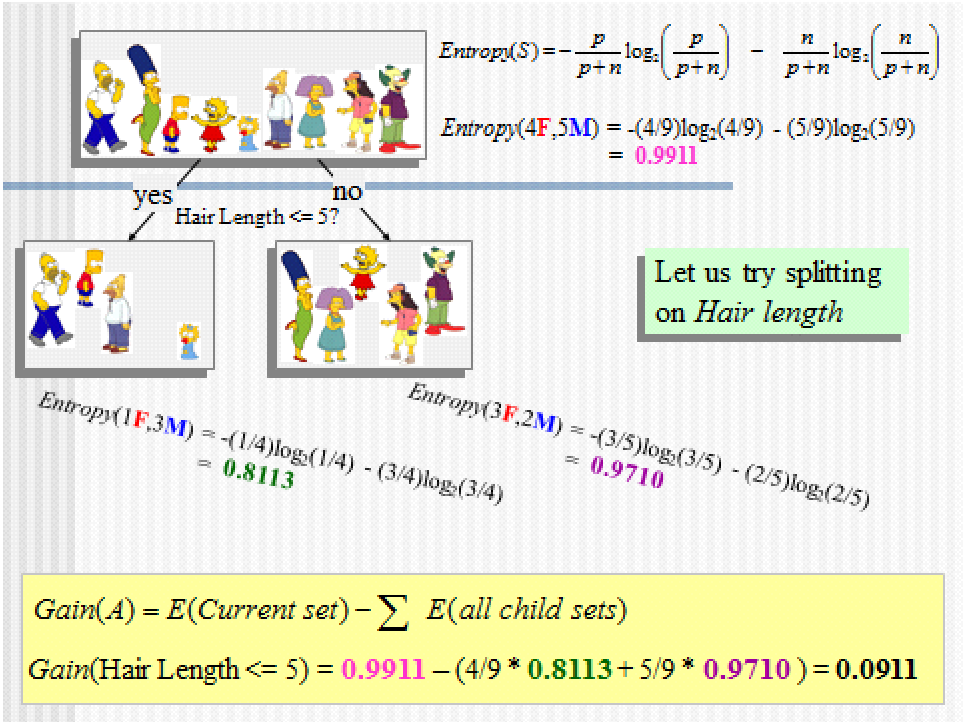
Working = 4
Failed = 2

2)



| Person | Hair Length | Weight | Age | Class |
|---|---|---|---|---|
| Homer | 0″ | 250 | 36 | M |
| Marge | 10″ | 150 | 34 | F |
| Bart | 2″ | 90 | 10 | M |
| Lisa | 6″ | 78 | 8 | F |
| Maggie | 4″ | 20 | 1 | F |
| Abe | 1″ | 170 | 70 | M |
| Selma | 8″ | 160 | 41 | F |
| Otto | 10″ | 180 | 38 | M |
| Krusty | 6″ | 200 | 45 | M |

|  |  |  |  |  |
|---|---|---|---|---|
| Comic | 8″ | 290 | 38 | ? |



$$Entropy(S) = -\frac{p}{p+n}\log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n}\log_2\left(\frac{n}{p+n}\right)$$

$$Entropy(4F,5M) = -(4/9)\log_2(4/9) - (5/9)\log_2(5/9) = 0.9911$$

yes

no

Hair Length <= 5?

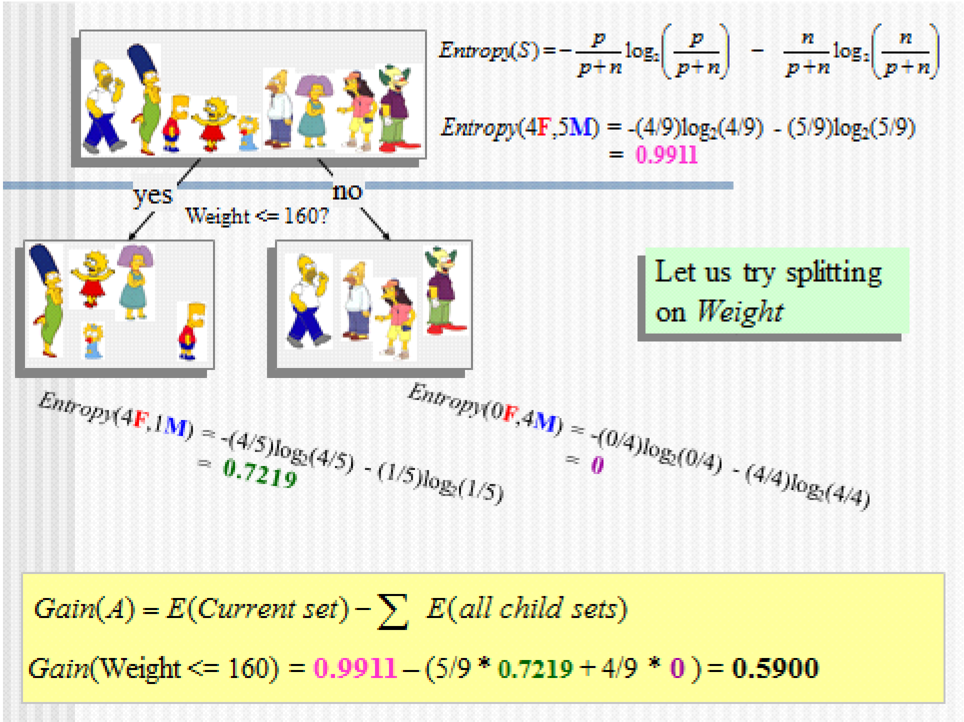
Let us try splitting on *Hair length*

$$Entropy(1F,3M) = -(1/4)\log_2(1/4) - (3/4)\log_2(3/4) = 0.8113$$

$$Entropy(3F,2M) = -(3/5)\log_2(3/5) - (2/5)\log_2(2/5) = 0.9710$$

$$Gain(A) = E(Current\ set) - \sum E(all\ child\ sets)$$

$$Gain(\text{Hair Length} <= 5) = 0.9911 - (4/9 * 0.8113 + 5/9 * 0.9710) = 0.0911$$



$$Entropy(S) = -\frac{p}{p+n}\log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n}\log_2\left(\frac{n}{p+n}\right)$$

$Entropy$(4F,5M) = -(4/9)$\log_2$(4/9) - (5/9)$\log_2$(5/9)
= 0.9911

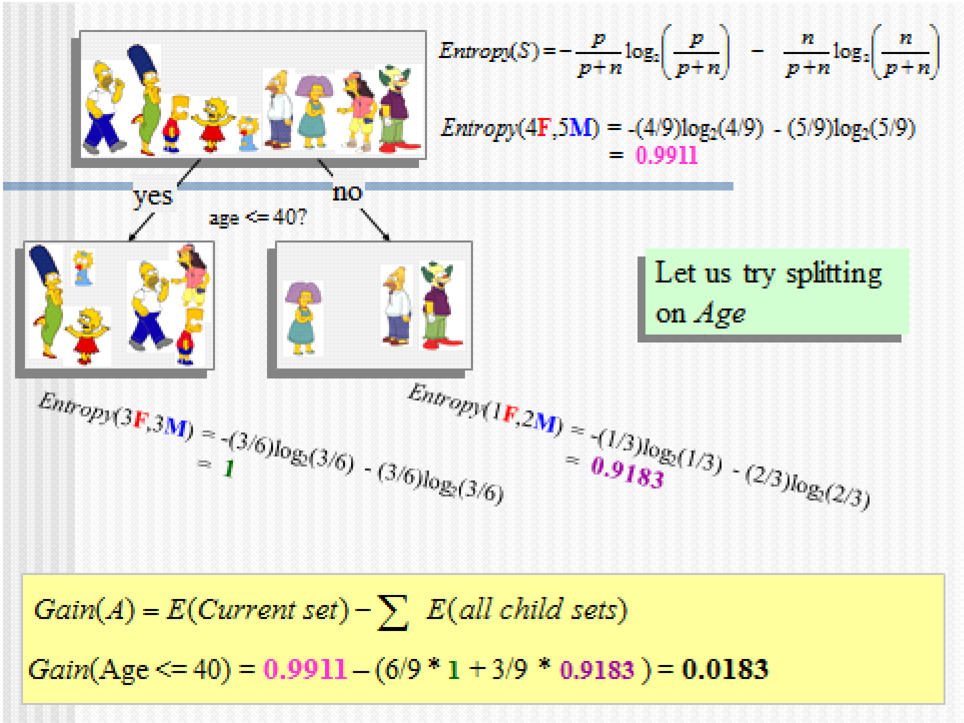
Let us try splitting on *Weight*

$Entropy$(4F,1M) = -(4/5)$\log_2$(4/5) - (1/5)$\log_2$(1/5)
= 0.7219

$Entropy$(0F,4M) = -(0/4)$\log_2$(0/4) - (4/4)$\log_2$(4/4)
= 0

$$Gain(A) = E(Current\ set) - \sum E(all\ child\ sets)$$

$Gain$(Weight <= 160) = 0.9911 – (5/9 * 0.7219 + 4/9 * 0 ) = **0.5900**



$$Entropy(S) = -\frac{p}{p+n}\log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n}\log_2\left(\frac{n}{p+n}\right)$$

$Entropy$(4F,5M) = -(4/9)$\log_2$(4/9) - (5/9)$\log_2$(5/9)
= 0.9911

Let us try splitting on *Age*

$Entropy$(3F,3M) = -(3/6)$\log_2$(3/6) - (3/6)$\log_2$(3/6)
= 1

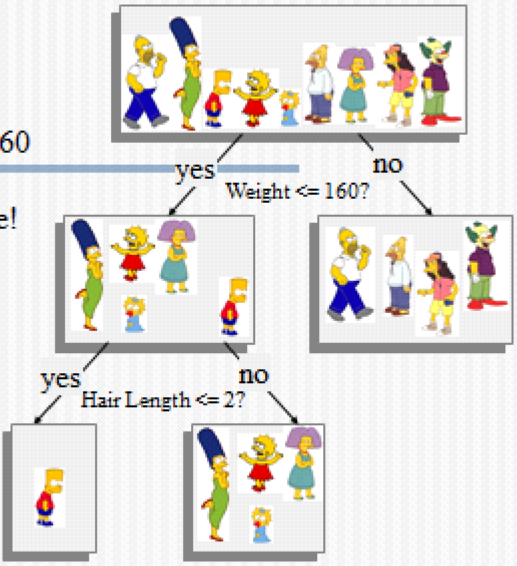
$Entropy$(1F,2M) = -(1/3)$\log_2$(1/3) - (2/3)$\log_2$(2/3)
= 0.9183

$$Gain(A) = E(Current\ set) - \sum E(all\ child\ sets)$$

$Gain$(Age <= 40) = 0.9911 – (6/9 * 1 + 3/9 * 0.9183 ) = **0.0183**

Of the 3 features we had, *Weight* was best. But while people who weigh over 160 are perfectly classified (as males), the under 160 people are not perfectly classified… So we simply recurse!

This time we find that we can split on *Hair length,* and we are done!



yes

no

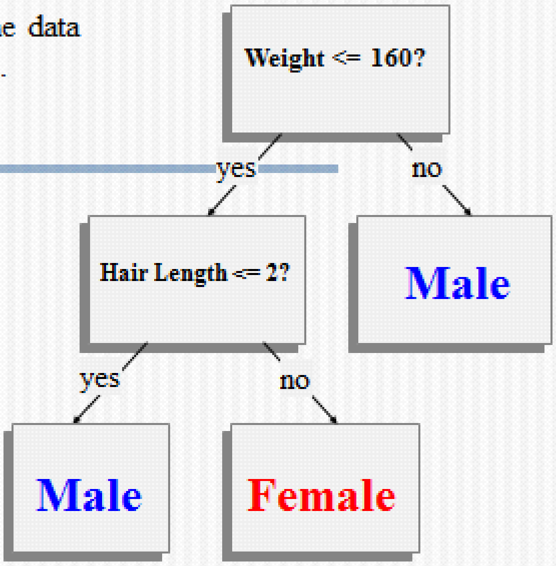
Weight <= 160?

yes

no

Hair Length <= 2?

We need don't need to keep the data around, just the test conditions.



**Weight <= 160?**

yes

no

**Hair Length <= 2?**

Male

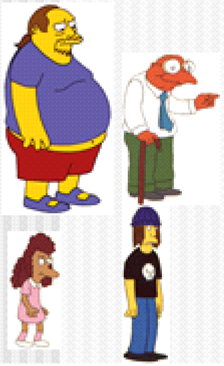
yes

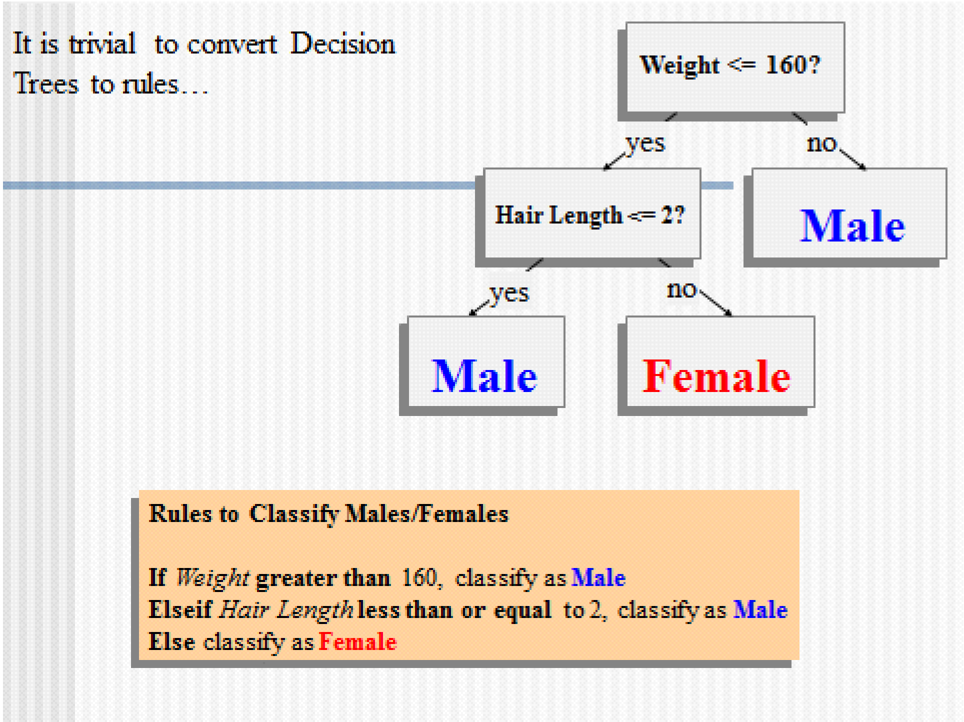
no

Male

Female

How would these people be classified?





It is trivial to convert Decision Trees to rules…

Weight <= 160?

yes

no

Hair Length <= 2?

Male

yes

no

Male

Female

**Rules to Classify Males/Females**

**If** *Weight* **greater than** 160, classify as **Male**
**Elseif** *Hair Length* **less than or equal** to 2, classify as **Male**
**Else** classify as **Female**