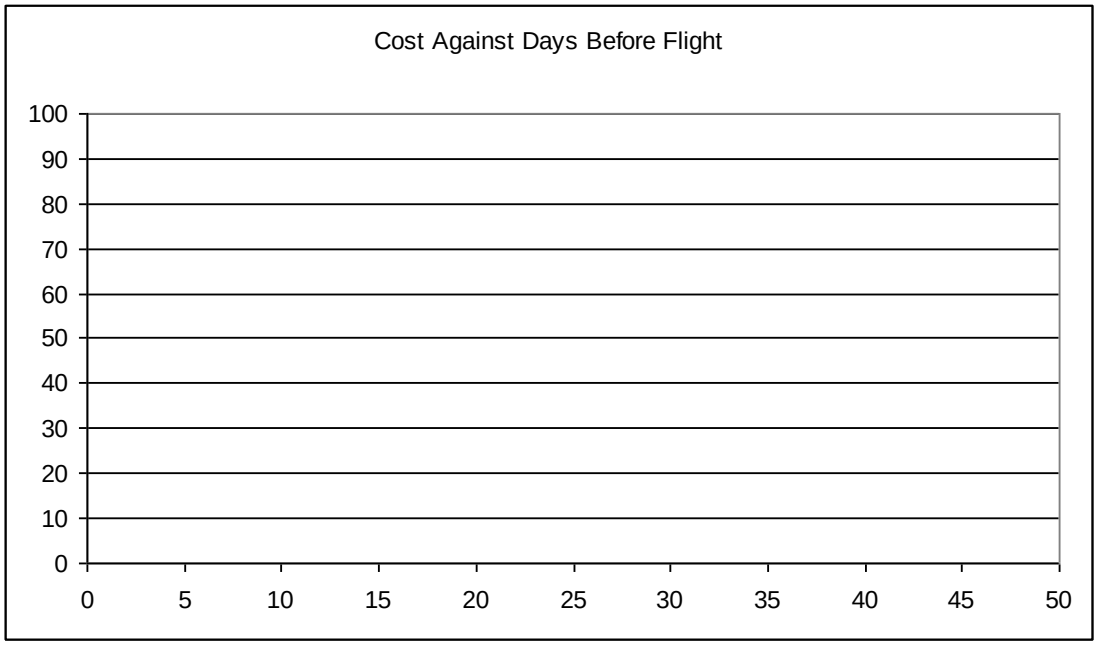
T6 Data Mining Tutorial – Classification & Prediction

1. Prediction

Imagine we want to model the price of an airline ticket given the number of days before the flight. Take the data in the table below and plot each point in the empty graph. Label the two axes correctly.

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



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.

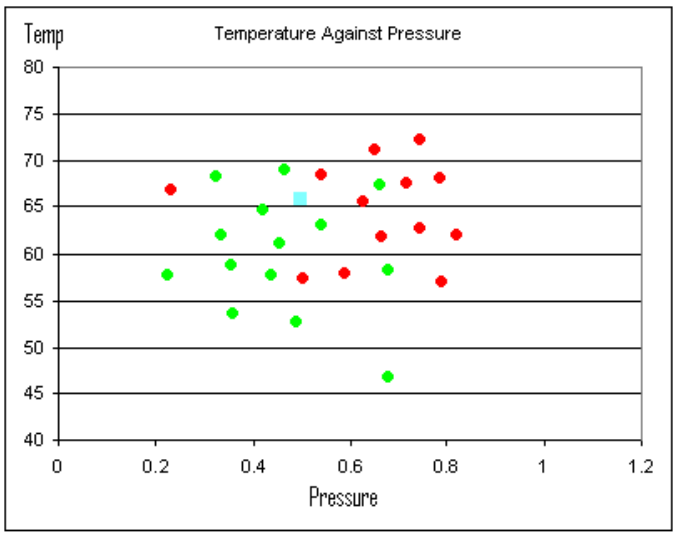
Now complete the table below, using the model you just drew above to make predictions for each day.

Days before	Cost
3	
12	
24	
46	
48	

Suggest an algorithm that you might use to build this function using data with more than 1 input, for example, Days before flight, month, number of passengers booked so far, etc. How would you use your chosen algorithm and what does it do?

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?

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)?

2) Use ID3 algorithm to build a decision tree based on the following training data, then answer what Comic's gender is.

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	**?**