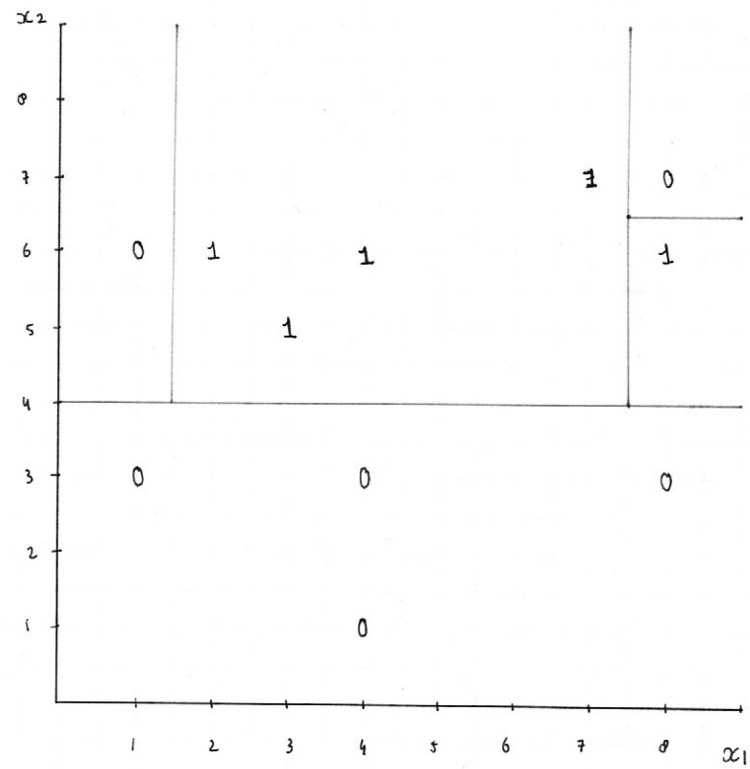
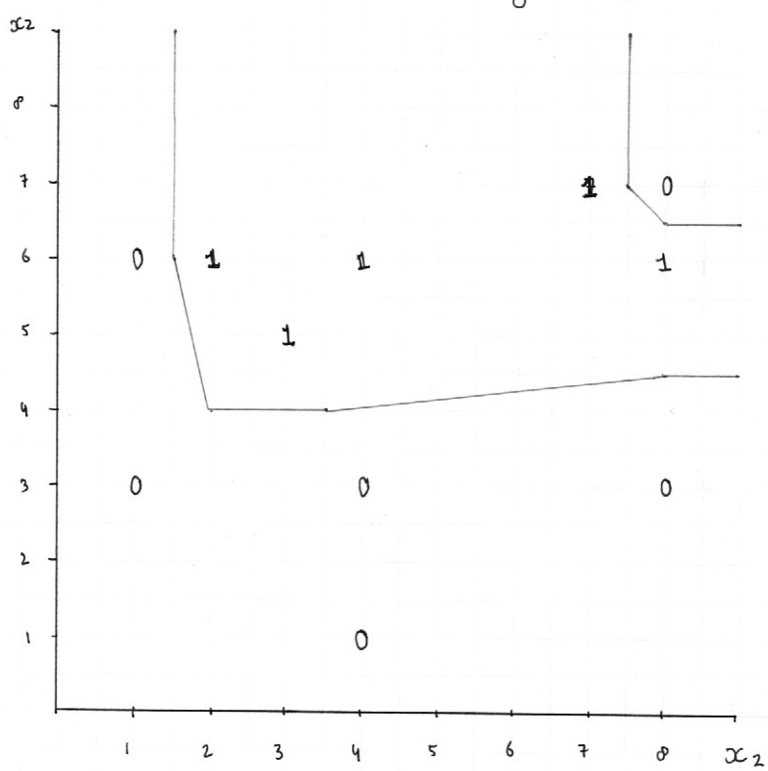
Decision Trees



The dataset

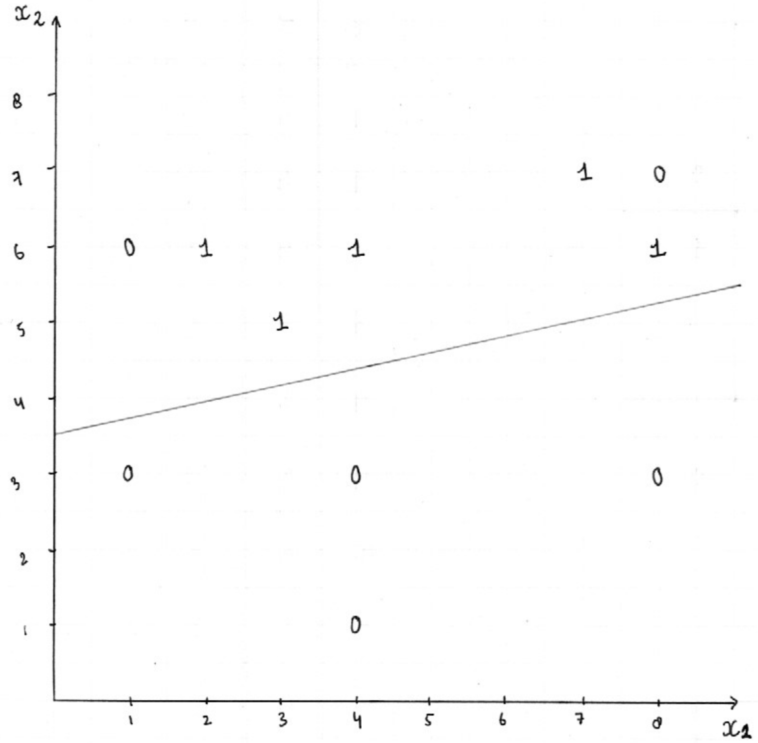
x_1	x_2	y
1	3	0
1	6	0
2	6	1
3	5	1
4	1	0
4	3	0
4	6	1
7	7	1
8	6	1
8	7	0
8	3	0

1- nearest neighbor

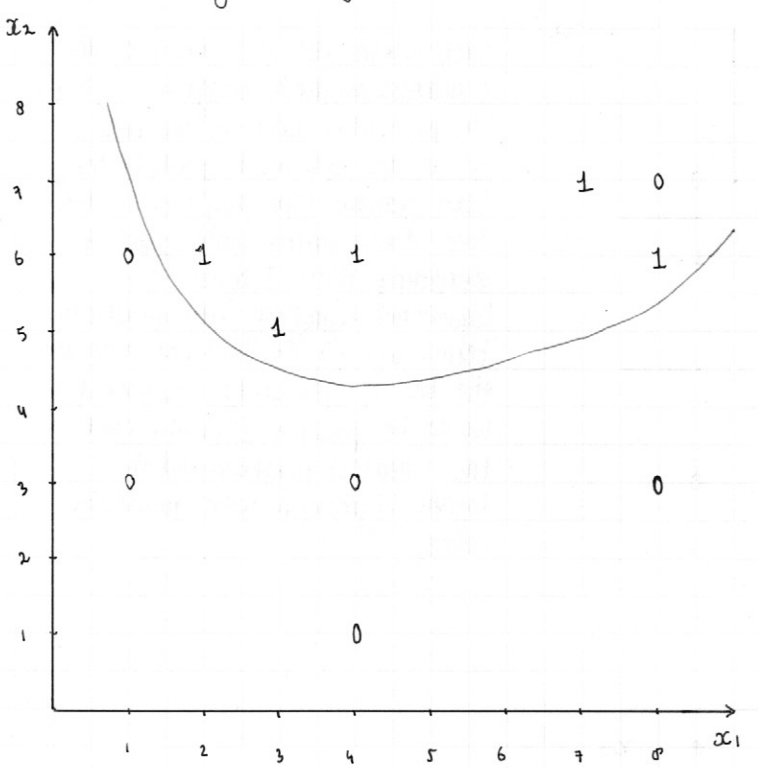


Even tough the decision tree and 1-nearest neighbor alrogrithm will fit this particular data set the best out of all the options, they will be the least general. Especially the decision tree demonstrates some signs of overfitting. Hence I believe that logarithmic regression with quadratic terms will fit the data best without the loss of generality. However if it would be possible, I would combine the 1-nearest neighbor and the logistic regression with quadratic terms.

Plain Logistic Regression



Logistic Regression with quadratic terms



Data	Cluster number	Mean	Up-dated mean
1	1	1	1.00
2	1	1	1.50
3	2	3	3.00
3	2	3	3.00
4	2	3	3.33
5	2	3	3.75
5	2	3	4.00
7	3	8	7.00
10	3	8	8.50
11	3	8	9.33
13	3	8	10.25
14	3	8	11.00
15	3	8	11.67
17	3	8	12.43
20	3	8	13.38
21	3	8	14.22

Step. 1. for each data point we evaluate to which cluster point the data point is closest and we assign it to that cluster.

step. 2. The cost function is given by $J(c^{(1)}, \dots, c^{(m)}, \mu_1, \dots, \mu_K) = \frac{1}{m} \sum_{i=1}^{m} \|x^{(i)} - \mu_{c^{(i)}}\|^2$ since we have 16 data points, $m = 16$, hence the cost before the algoritm is given by:

$$J = \frac{1}{16} \cdot \left[|1-1|^2 + |2-1|^2 + |3-2|^2 + |3-2|^2 + |4-2|^2 + |5-2|^2 + |5-2|^2 + |7-3|^2 + |10-3|^2 + \dots \right.$$
$$\left. + |11-3|^2 + |13-3|^2 + |14-3|^2 + |15-3|^2 + |17-3|^2 + |20-3|^2 + |21-3|^2 \right]$$

$$J = 33$$

step. 3. Now we apply the K- clustering algorithm, for every new data point added to a cluster we will re-calculate the clusters mean

step. 4. Given these new up-dated means for the three clusters, $\mu_1 = 1.50, \mu_2 = 4.00, \mu_3 = 14.22$, we calculate the new value for the cost function.

$$J = \frac{1}{16} \cdot \left[|1-1.50|^2 + |2-1.50|^2 + |3-4|^2 + |3-4|^2 + |4-4|^2 + |5-4|^2 + |5-4|^2 + |7-14.22|^2 + \dots \right.$$
$$+ |10-14.22|^2 + |11-14.22|^2 + |13-14.22|^2 + |14.-14.22|^2 + |15-14.22|^2 + |17-14.22|^2 + |20-14.22|^2 + \dots$$
$$\left. + |21-14.22|^2 \right]$$

$$J \approx 10.88$$