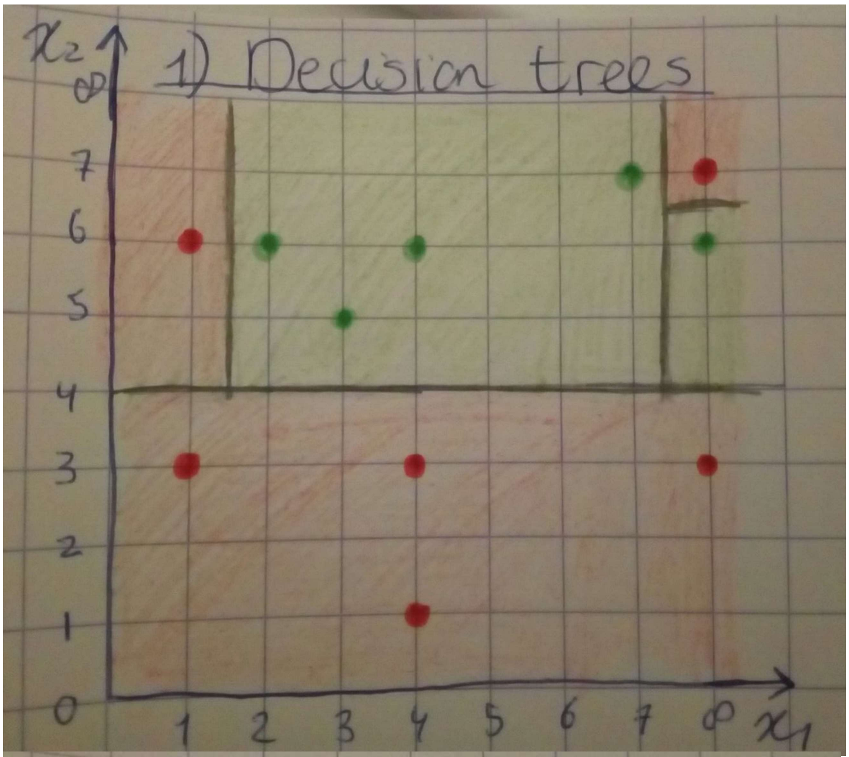
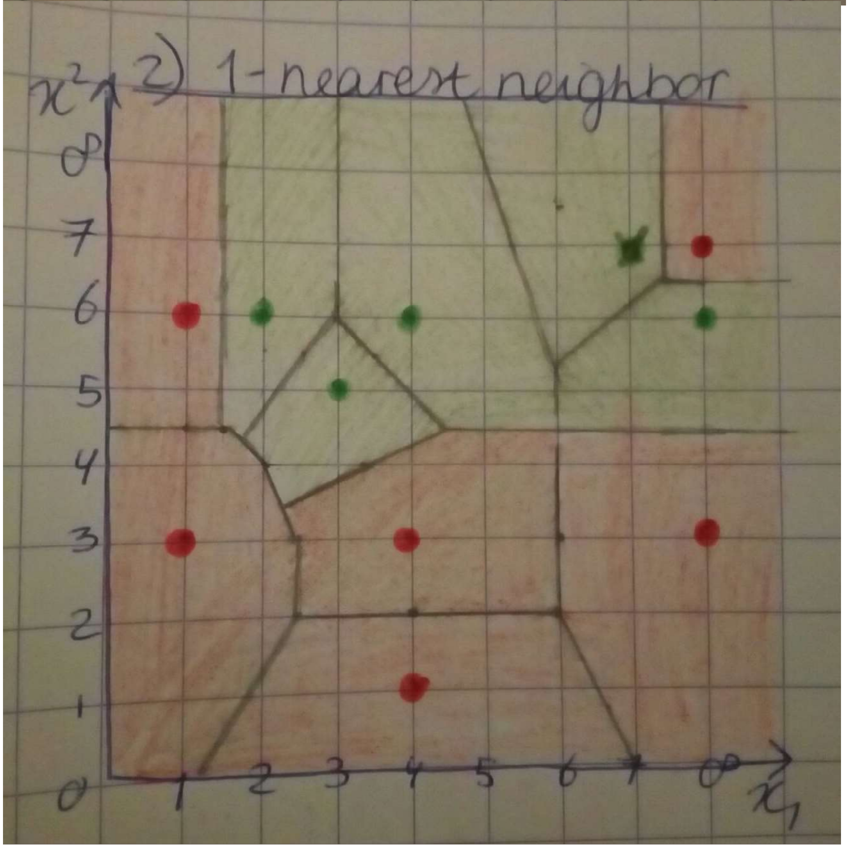
Graded Assignment 2: Written assignment

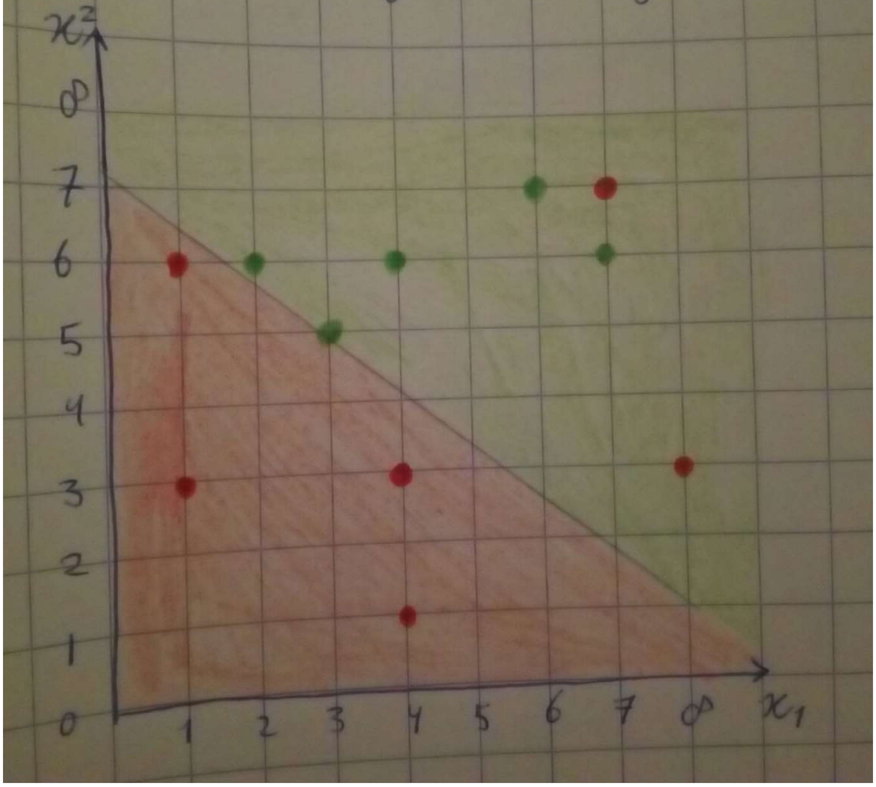
Miriam Riefel 11332549

1) Red signifies a datapoint with y=0, and the area that is classified as 0
   Green signifies a datapoint with y=1, and the area that is classified as 1

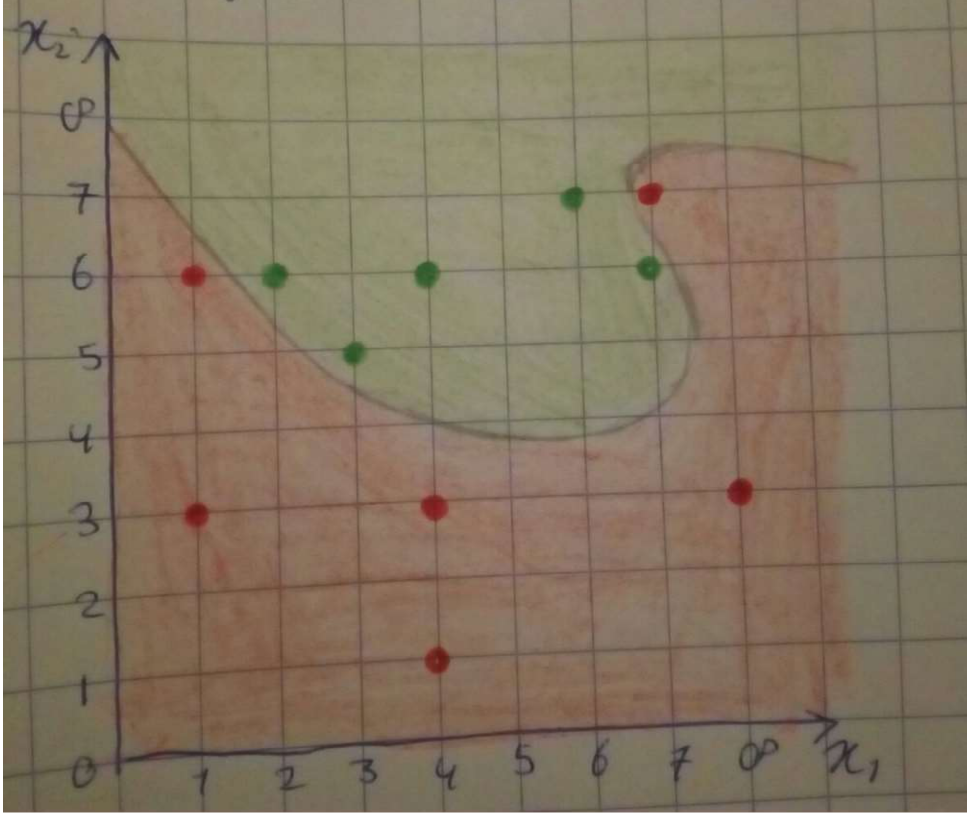




3) plain logistic regression



4) polynomial logistic regression



2) Firstly, a decision boundary made by plain logistic regression boundary, without polynomials, is not suited for this dataset. The linear line underfits the data and creates a high-bias classification.

All other classification methods suffice. However, the polynomial logistic regression would be quite complicated if it had to fit the data and thus be computationally expensive. The 1-nearest neighbour classification can also be computationally expensive, because for each new data-sample, it needs to go through all data instances. Finally, the decision tree is also quite computationally expensive, because it needs for decision boundaries.

When we look at the data points, we see that it is mainly one sample (7,7, y=0), that makes the classification very complicated. If we could take this area out, then we could have a quite simple logistic regression or a two-boundary decision tree. The cell around (7,7, y=0), can be created through two Booleans: x1 > 7.5 and x2 > 7.5. Thus, what we could do is add one if statement before other classification methods. Namely, if x1 > 7.5 and x2 > 7.5, then y = 0. Then we could use a logistic regression function that would work with a function using the polynomials [0.5,1,2]. Alternatively, this nearest neighbour condition could also be combined with a decision trees classification method, that would basically come down to if x2< 4: y =0, if x1 < 1.5: y = 0, else y = 1.