Assignment 4

1.1.1 Theoretical Question

Given the data presented in Figure 1, identify the class for the green point using the k-nearest neighbour classifiers with k = 1, 3, and 5. There are only two classes of blue and red, and we want to see which class the green point belongs to. Use major voting with weighting to identify the class. Use the following distance function:

$$\text{Distance} = \min_{i=1}^{2} |x_i - y_i|$$

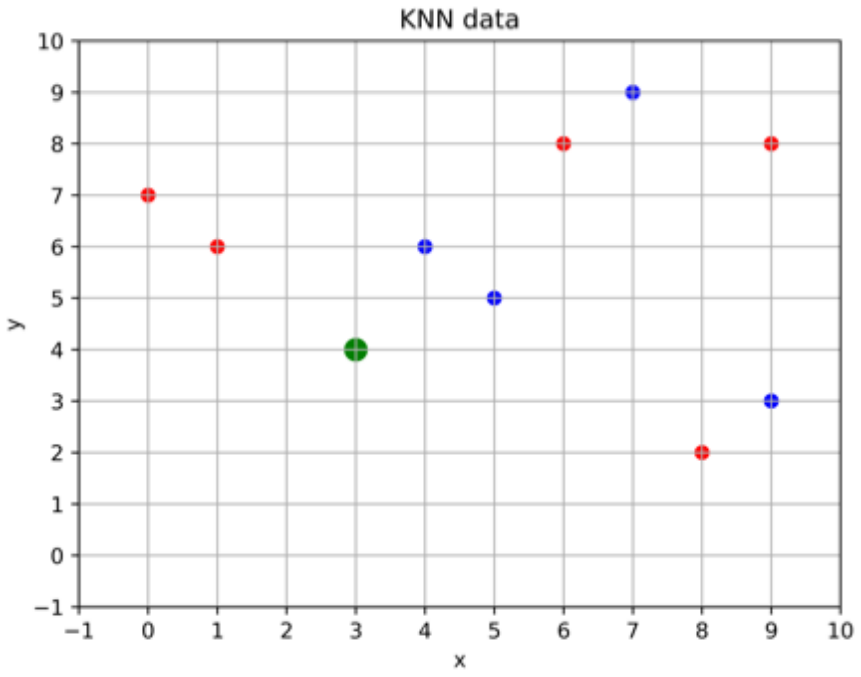

Figure 1

Using the weight function $w = \frac{1}{d^2}$ $where\ d = min(|x_1 - x_2|, |y_1 - y_2|)$

Gives the following results: Using the highest number as the nearest neighbour

	0, 7	1, 6	4, 6	5, 5	6, 8
3, 4	1/9	1/4	1	1	1/9

We can then see that if k = 1, Green belongs to Blue, if k is 3, Green is also Blue, Only at k = 5 does Green become a part of Red.

1.1.5 Conceptual Question

Imagine that you implemented this algorithm for your company, Which relies on your classification method. Your boss is asking you:

- Whether it is a good idea for the company to use accuracy for evaluation. What is your answer? Please justify your answer.

Accuracy alone might not be the best case to go, depending on the balance in the dataset. If the dataset happens to be fairly balanced it can give a nice indication of the evaluation. It also happens to be simple to implement which would have to be taken into consideration.

2.1.1 Conceptual Question

- How is the performance of the Decision Tree compared to the KNN?

  Performance wise a Decision tree is faster than the KNN algorithm, purely looking into runtimes in this Notebook environment.

- Which one has fewer misclassified test items?

  The Decision tree also has fewer misclassified items. With an accuracy of 51.4% on the contrary to the KNN algorithm's ~48.9%. Then having 1303 misclassifications.

- Which one would you select for your company? Justify your choice.

  The choice seems clear on choosing a decision tree rather than the KNN algorithm, for this particular dataset it has a major performance benefit in addition to higher accuracy in its classifications. Given the parameters of the datasets both algorithms could be useful in the classifications, but Decision trees have these two benefits, and the attributes are not expanding, but static, so this is a good choice.