Question 1.

*What is overfitting? If we do not check overfitting when we build a machine learning model, what will happen? To avoid overfitting issues, what approach would you, as data scientists, recommend?*

A model is overfit if it only performs well for the data it was trained on. We would say that the model is not generalizable to non-training data and therefore is not as useful for predictions. If we don’t check for overfitting, the model will perform poorly with any future data sets or points it is given.

To avoid overfitting, we must be sure that the model is just as performant on test data as it is on training data. This can be done with cross validation, and specifically I would recommend k-fold validation to repeat testing on different data partitions. The size of k would be dependent on the size of the data set and available computation resources. Additionally, it is important to understand the types of models that may be prone to overfitting such as unpruned decision trees.

Question 2.

| **X1** | **X2** | **X3** | **Y** | **Euclidian Distance** |
| --- | --- | --- | --- | --- |
| 1 | 0 | 3 | Red | 3.000000 |
| 2 | 2 | 0 | Red | 2.000000 |
| 3 | 0 | 1 | Red | 3.162278 |
| 4 | 0 | 1 | Green | 2.236068 |
| 5 | -1 | 0 | Green | 1.414214 |
| 6 | 1 | 1 | Red | 1.732051 |

*Suppose that we will use the above data for predicting variable Y when X1=0, X2=0, and X3=0 using k-nearest neighbors.*

1. *What is your prediction with k=1? Why? [5 points]*

For k = 1 I would predict Green as it contains the shortest Euclidean distance and is the single nearest neighbor.

1. *What is your prediction with k=3? Why? [5 points]*

For k = 3, I would predict Red as 2/3 nearest neighbors by Euclidean distance are Red.