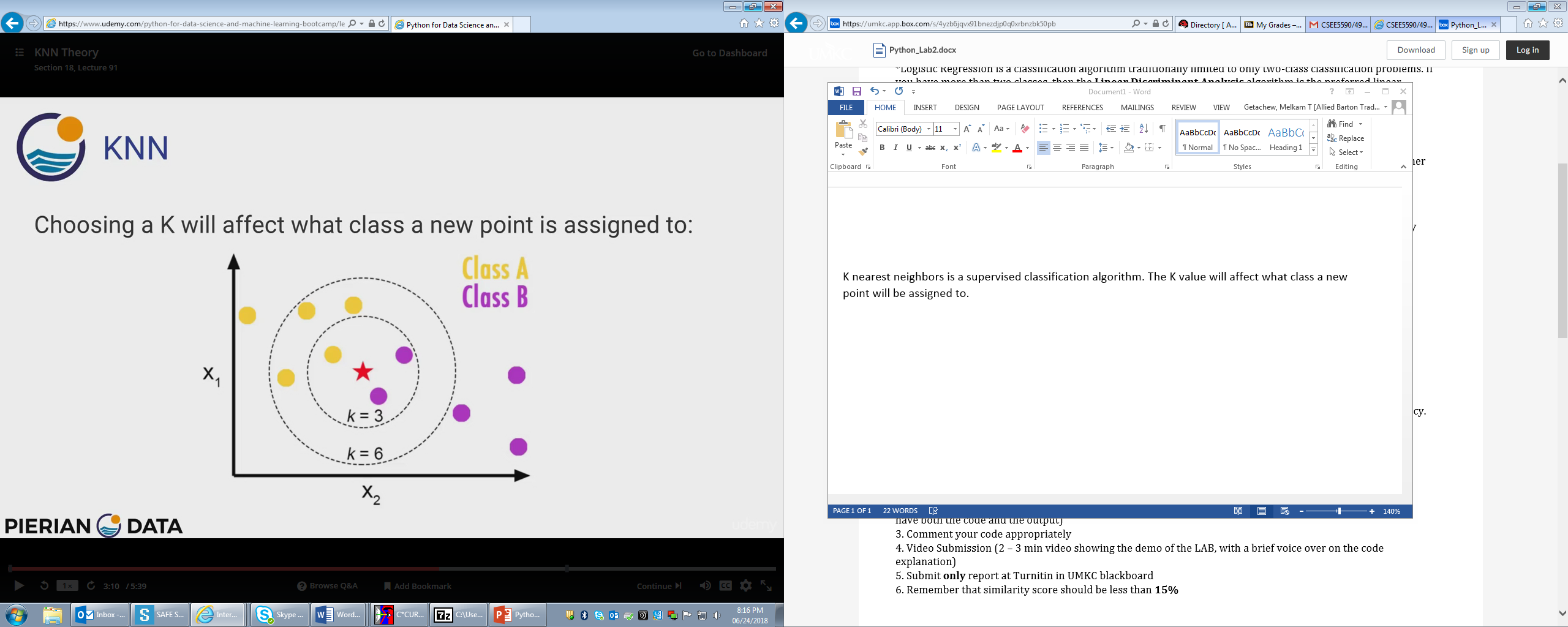
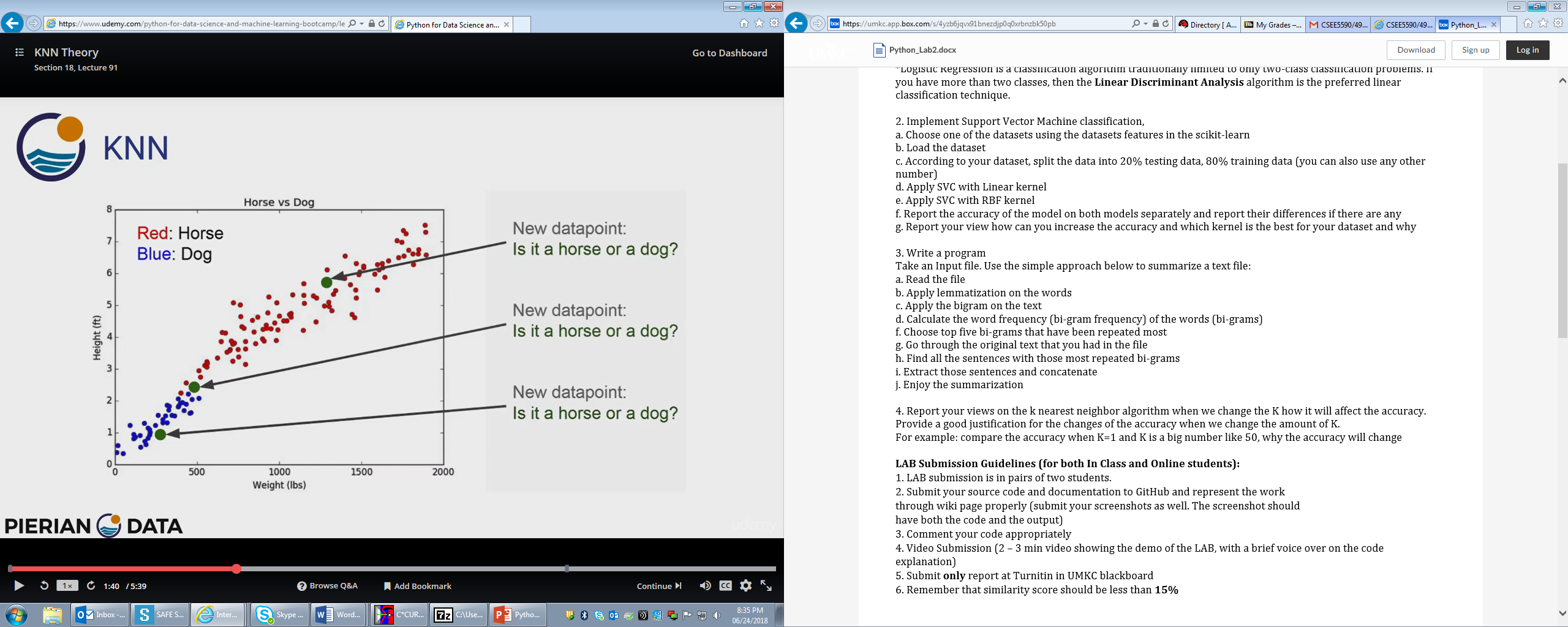
K nearest neighbors is a supervised classification algorithm. The K value will affect what class a new point will be assigned to.



For instance in the above figure that represent a certain data set. Let say we introduced a new data that is the star, and we want to determine to which class we should group it to. To which class we group it to depends on the value of k. Let say we chose a k value of 3. K indicates the number of nearest neighbors to the new data. In this case the closest neighbors are two purple and one yellow. Since the number of purple neighbors is greater than the yellow ones we should group it to class B.

Let’s make our k value equal to 6. In this case the closest neighbors are two purples and four yellows. Now the number of green neighbors are greater we should group it to class A.

The above example shows how the value of k affects the classification algorithm. Selecting k value that is too small or too large might not give us the result we need, since it has the effect of biasing the analysis. So which is better? Choosing a small k value like one or choosing a k a large value like 50. Let us see another example.



There are three new data points in this figure, the top, the middle and the bottom. For the top and bottom ones the value of k won’t have that much effect on how to group the data. But for the middle one it makes a huge difference. If we choose k to be one then we have to group it as a horse. If we choose k to be fifty we will group it as a dog.

Choosing a small k value like one will result in a noise that is an error (miss labeling) and choosing a large k value will produce a better accuracy but has a cost of computational expense. Bearing this in mind we have to come up with a formula to compute the k value depending on the desired accuracy and computational expense. But one thing we can be certain of is the formula should depend on the size of the training data.