Samuele Serri 7069839
Ege Mert Balcik 7071632

a)

Compute the distances between $I_{new} = (168, 80)$ and all the other points in the dataset:

$$d((168, 80), (167, 75)) = \sqrt{(168 - 167)^2 + (80 - 75)^2} = \sqrt{26} \approx 5.0990$$

$$d((168, 80), (183, 62)) = \sqrt{(168 - 183)^2 + (80 - 62)^2} = \sqrt{549} \approx 23.431$$

$$d((168, 80), (175, 64)) = \sqrt{(168 - 175)^2 + (80 - 64)^2} = \sqrt{305} \approx 17.464$$

$$d((168, 80), (170, 85)) = \sqrt{(168 - 170)^2 + (80 - 85)^2} = \sqrt{29} \approx 5.3852$$

The closest point using euclidean distance is $(167, 75)$ so class 0 will be assigned to I_{new} with 1-NN.
With 3-NN classifier we have to consider the three closest neighbor: $(160, 175), (170, 85), (175, 64)$ respectively of class 0,2 and 1. The point I_{new} will take the class of the majority vote but since we have one representative for each class it depends on the weight assigned to each point.

b)

If we have classification problems we only 2 classes then choosing an odd number as k would be better because it avoids ties.

c)

In high-dimensional spaces almost all vectors have the same euclidean distance with I_{new}.

d)

The choice of k is an hyper-parameter and is strongly data-dependent. We can divide data into training, validation and testing sets, choose the value of k with validation set and test it with the test set. Alternatively, if the dataset is not too large, we can use cross-validation which consists in splitting data into folds, using each fold as validation and then averaging the results.

a)

The purpose of regularization is to make the model work better on test data. We add a penalization to avoid over-fitting or under-fitting during the training process.

b)

Elastic net and dropout are examples of regularization techniques.

c)

$\lambda \in [0, +\infty)$

d)

If $\lambda = 0$ then we express no preferences on the complexity of the model and we over-fit.
If $\lambda >> 0$ then the weights are forced to be small and we under-fit.

e)

Elastic net combines L1 and L2 regularization minimizing the loss:

$$L(w) = \|y - Xw\|_2^2 + \lambda_1 \|w\|_1 + \lambda_2 \|w\|_2^2$$