Homework 7&8: More Models+Ensembles

The dataset we will be working with is used to predict whether a job candidate will be invited to a second interview. Each candidate is represented by two features:

- **Feature 1 (x_1)**: Number of coding challenge errors during the technical assessment
- **Feature 2 (x_2)**: Communication skills rating (on a scale from 0 to 3, higher is better)

The label indicates whether the candidate **did receive** (1) or **did not receive** (0) a second interview. Below are the first couple of candidates from the dataset:

Candidate ID	**Coding Errors (x_1)**	**Communication (x_2)**	**Second Interview** (Label)
1	2	1	0
2	1	3	1
3	2	0	0
4	0	2	1
5	3	3	1

We would like to use this dataset to see if a new candidate should be given a second interview. Candidate 6 had 1 error on their technical assessment, and scored a 2 on their communication skills.

1. Using k-nn with k=3, predict whether candidate 6 will receive a second interview.

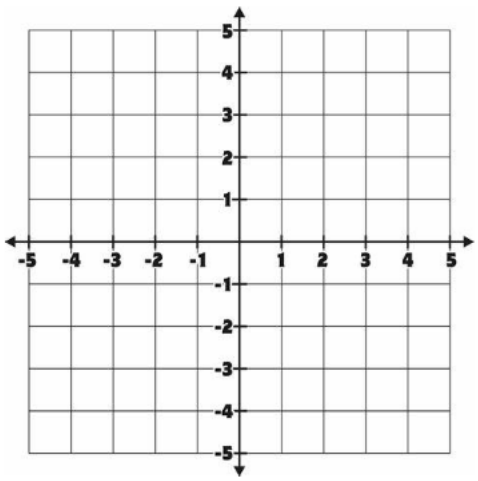
   (a) Calculate the Euclidean distance $d = \sqrt{(x_{1c_i} - x_{1c_6})^2 + (x_{2c_i} - x_{2c_6})^2}$ between candidate 6 (c_6) and all other candidates (c_i) in the dataset.

Candidate ID	**Distance**
1	
2	
3	
4	
5	

   (b) Find the 3 nearest neighbors based on the distances calculated. Which candidates are the nearest?

   (c) Based on the majority class do you predict candidate 6 will receive a second interview? Explain.

2. Using SVM, determine the best hyperplane. You are given two possible hyperplanes that attempt to separate candidates who received a second interview from those who did not. You need to decide which hyperplane does a better job of separating the candidates. Hyperplane 1: $x_1 - 2x_2 = -2$ Hyperplane 2: $x_1 - x_2 = -1$.



   (a) Plot and label the candidate data. Mark the candidates who received a second interview with a star.

   (b) Draw and label both hyperplanes on the plot.

   (c) Decide which hyperplane best separates the two classes. Explain.

   (d) Add candidate 6 to the plot (mark with a big circle). Using the best hyperplane, do you predict candidate 6 will receive a second interview? Explain.

3. Using logistic regression, predict whether candidate 6 will receive a second interview. You are given the following logistic regression equation that was fitted using the existing dataset:

$$\hat{p} = \frac{1}{1 + e^{-(b_0 + b_1 \cdot x_1 + b_2 \cdot x_2)}}$$

where:

- $\hat{p}$ is the probability that a candidate received a second interview.
- $b_0 = -3$ (intercept),
- $b_1 = 1$ (coefficient for Coding Errors, x_1),
- $b_2 = 1.5$ (coefficient for Communication Skills, x_2)

(a) Use the equation to plug in the coefficient values and the information we have on candidate 6 to calculate the probability that candidate 6 will receive a second interview.

(b) Based on that probability, predict whether or not the candidate will receive a second interview (label = 1). Explain.

4. Identifying the best ensemble model. You are now working with a dataset containing information on **250 candidates**. Of these 250 candidates, only 45 candidates (18%) received a second interview (label = 1). The Coding Errors (x_1) ranges from 0 to 5, and the Communication Skills Rating (x_2) ranges from 0 to 3. You suspect there may be some non-linearity in the dataset.

(a) Describe some aspects about our dataset that we must consider when choosing a model. Consider size, number of features, imbalances, linearity, bias-variance tradeoff, and how these aspects can influence over-/under-fitting.

You have run a combination of models, and now need to interpret the results.

Model	**Accuracy**
null model: majority	82%
kNN: k=3	80%
SVM: linear kernel	84%
LogReg	85%
RF: bagging	88%
AdaBoost: boosting LogReg	90%
soft voting: LogReg + SVM + KNN	88%
stacking:LogReg +SVM +KNN→RF	93%

(b) Briefly describe what the null model is doing.

(c) The worst performing model was k-Nearest Neighbors (k=3). Why might this model be doing worse than the null model?

(d) The first two ensemble models used are bagging (using random forest) and boosting (AdaBoost with logistic regression). Describe why boosting may be a better ensemble model for our dataset.

(e) The best performing ensemble model was stacking. Describe why this may be.