DS 201
Spring 2020
Exam 2

Use your laptop. Take home exam, no time limit. Partial credit may be given for partially correct solutions.

- Create one Jupyter notebook file as DS201_Exam2_LastFirst.ipynb
- Write answers of questions 1-6 as comments (or markdown) on separate block of your notebook file
- Display charts clearly. You don't need to have an exact result as sample.
- Please write comments for your code, brief comments will help make your intention clear in case your code is incorrect.
- You can access any website for reference. No communication allowed. Do not post anything on social media, no email, no chat.
- Submit a DS201_Exam2_LastFirst.ipynb on Canvas

If you have questions, please ask!

Question	Points	Your Score
1	4	
2	4	
3	4	
4	4	
5	4	
6	14	
7	16	
8	20	
9	30	
Total	100	

Part 1: put your answer in a comment in your jupyter notebook file. E.g.,

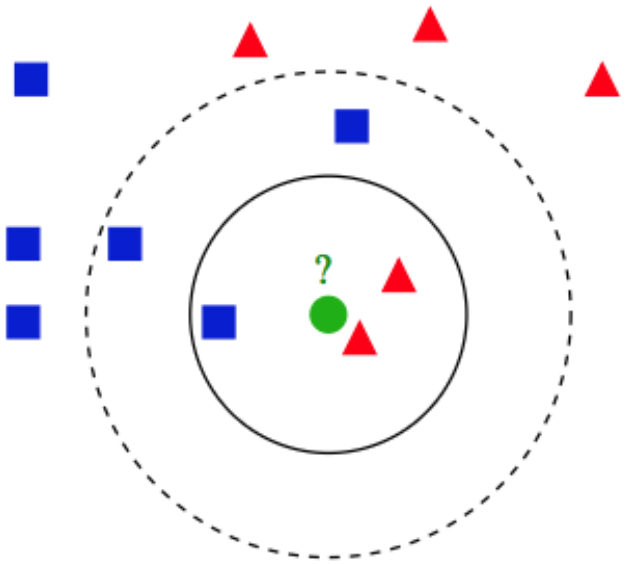
#q1 answer: bla bla bla

#q2 answer: boo boo boo

1. (4 points) According to the following plot image, which Machine Learning model is being used? What is the reasons to using this model?

2. (4 points) What method of machine Learning can “predicts the numeric value for each observation, based on the values of other features column and their labels”?
3. (4 points) From the previous question, what will the model called if the it predicting “categorical values”?

4. (4 points) According to the K-Nearest Neigbor algorithm, if k = 11, in which class would the green observation be classified? Provide the reason.



5. (4 points) If we do not have a label in the dataset, what type of Machine Learning should be used?

Part 2

Breast Cancer Dataset **BreastCancer_data.csv (on Canvas under the Exam#2 module)**

About this Dataset:
Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

Attribute Information:

1) ID number **2) Diagnosis (M = malignant, B = benign)** 3-32)

Ten real-valued features are computed for each cell nucleus:
All feature values are recoded with four significant digits.
Missing attribute values: none
Class distribution: 357 benign, 212 malignant

Use the Breast Cancer Dataset to answer the following question:

6. **(14 points)** Create a simple box plot of radius_mean by diagnosis. Your X axis need to be diagnosis.

7. **(16 points)** Create a pairplot by following columns:
   "radius_mean","texture_mean","perimeter_mean","area_mean","smoothness_mean","compactness_mean".
   Separate colors by “diagnosis”.

8. **(20 points)** Create a bokeh Scatter plot as following:
   - **Saperate** Benign and Malignant data
   - x axis: **area_mean**
   - y axis: **texture_mean**
   - the size of datapoints: radius_mean
   - Benign datapoints: green circle
   - Malignant datapoints: red triangle
   - title: Benign and Malignant **texture_mean(Y)** by **area_mean(X)** with size by **Radius**
   - use the lower alpha value (transparency) for Benign to allow Malignant data to be cleary visible
   - legend location: top_left

9. **(30 points) create a classifier to classify each observation to two class, Benign and Malignant:**
   - **Target:** Benign or Malignant
   - **train/test size: 80:20**
   - random_state=1
   - features columns X :
     "radius_mean","texture_mean","perimeter_mean","area_mean","smoothness_mean","compactness_mean"

   answer the following questions:

   1. print out the shape of X,y,X_train, X_test, Y_train,Y_test
   2. report Logistic Regression accuracy of X,Y
   3. report KNN accuracy of X,Y when k=5
   4. loop k1-k25 and plot KNN accuracy of X,Y

   sample of k1-k25 plot of X (your chart might look different!)