Breast Cancer Wisconsin

Data Set Information:

Samples arrive periodically as Dr. Wolberg reports his clinical cases. The database therefore reflects this chronological grouping of the data. This grouping information appears immediately below, having been removed from the data itself:

Attribute Information:

1. Sample code number: id number
2. Clump Thickness: 1 - 10
3. Uniformity of Cell Size: 1 - 10
4. Uniformity of Cell Shape: 1 - 10
5. Marginal Adhesion: 1 - 10
6. Single Epithelial Cell Size: 1 - 10
7. Bare Nuclei: 1 - 10
8. Bland Chromatin: 1 - 10
9. Normal Nucleoli: 1 - 10
10. Mitoses: 1 - 10
11. Class: (2 for benign, 4 for malignant)

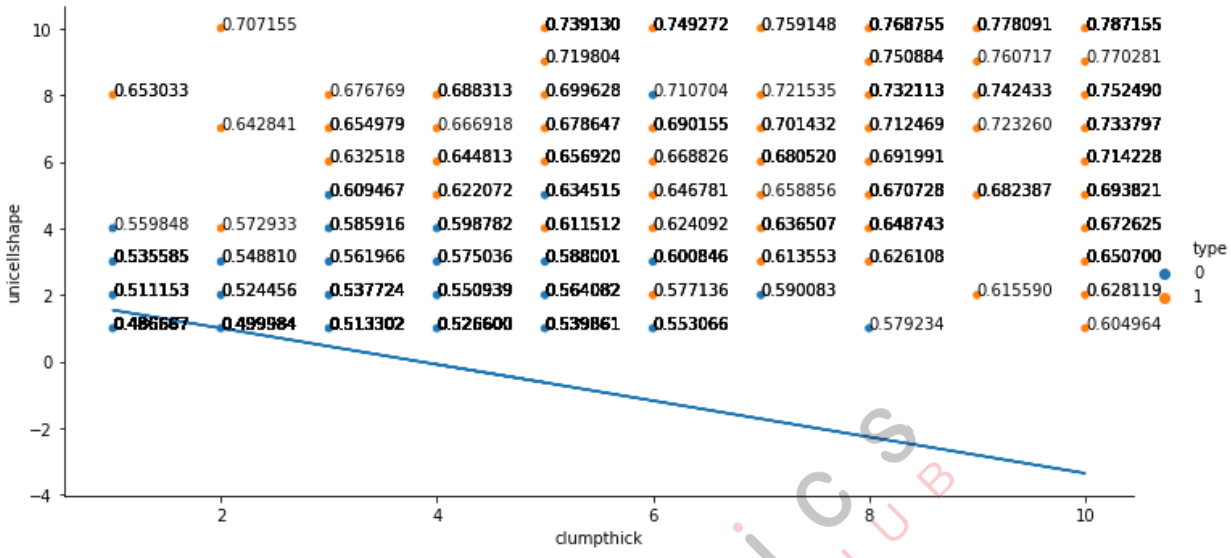
TASK -1

Please kind the following task regarding the dataset:

1. Read the dataset
2. Put the columns names to each column as shown in attribute information
   Use following:
   'idnumber','clumpthick','unicellsize','unicellshape','adhesion','epithelcellsize',
   'barenuclei','chromatin','normnucle','mitoses','class'
3. Identify **categorical** and **numerical** variables in the dataset and write in *markdown.*
4. Remove **idnumber** column
5. Replace **class** columns values '***2***' with ***'0'*** and '***4***' with ***'1'***
6. Consider the following columns
   a. Clump Thickness
   b. Uniformity of Cell Shape
   c. Class
7. Plot '*scatterplot*' between **clump thickness** vs **uniformity of cell shape**
8. **Use all recommended plots will be add-on**
9. Plot distribution of clump thickness and uniformity of cell shape
10. Split the data into training and testing set (80 % training and 20 % training)
11. Build Logistic regression model with training data.
12. Draw a line (logistic regression line) and probability value of each point

LOGISTIC REGRESSION

Eg. Data point with probability values and logistic regression line with p = 0.5



13. Interpret the results
14. Compute **Confusion Matrix** for training set and testing set
15. From confusion matrix compute
    a. Sensitivity, Specificity, Precession and Accuracy for training set
    b. Sensitivity, Specificity, Precession and Accuracy for testing set
16. Also compute **Kappa Score for** training and testing set (optional)
17. Draw ROC curve (Just copy paste the code in today's lecture 22nd March 2019 will explain what ROC and AUC is will explain about)
18. Suggest the approx. threshold value for probability of success from **Scatter plot and logistic regression line**
19. Plot logistics regression for best threshold probability value which you feel
    a. Compute Confusion matrix
        i. Training data
        ii. Testing data
    b. Computer Sensitivity, Specificity, Precession and Accuracy for both training and testing confusion matrix.
20. Any additional recommendation want to made in this model is add-on

Submit the **jupyter notebook** file and also **pdf** of jupyter files by **24 Mar 2019**

TASK -2

1. If I consider all 10 attributes suggest what are the **feature** (columns) are more appropriate. Note: Use either **forward selection** or **backward elimination** method. Submit **jupyter notebook** and **pdf** file by **24 Mar 2019**