PGDM / PGDM (IB), 2019-21
Predictive Business Analytics
DM-341 / IB-318
Trimester-III, End-Term Examination: June 2020

The dataset provided to you contains feature values calculated from digitized image of Fine Needle Aspirate (FNA) of 569 samples of breast mass. These features describe the characteristics of the cell nuclei present in the image, and are used to diagnose potential instances of malignant breast cancer. The features are as follows.

1. Field No. 1: ID number
2. Field No. 2: Diagnosis (M = malignant, B = benign)
3. Field Nos. 3-12: Mean value
4. Field Nos. 13-22: Standard Errors
5. Field Nos. 23-32: Worst (most extreme values)

The ten real-valued features which are computed for each cell nucleus are:

a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" – 1)

Use kNN Classification ***OR*** Logit Regression to create a diagnosis (classification) model for such digitized images.

In particular you are required to do the following:

1. Create 70:30 split of the data available such that both training and test data sets have identical proportion of malignant and benign samples.
2. Train the model (using the training data).
3. Apply the model on the test data set and report the accuracy, sensitivity and specificity
4. Do you need to improve / refine your model? Why or why not?
5. If you refine your model how does the performance change for the test data set?

In the Word document that you will submit please paste code snippets, results as well as the methodology you followed along with your explanations. Also please submit your ***.Rnw*** file.