**Breast Cancer Dataset**

**Data description:**

All the data points are reported by Dr. WIlliam H. Wolberg (Physician, University of Wisconsin Hospitals Madison, Wisconsin, USA).

**Site:** http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)

**Number of data points:** *N* = 699

**Number of features:** *n* = 9

**Feature Description:** Each feature is evaluated on scale of 1 to 10, with 1 being closest to benign and 10 closest to malignant.

*x*1 : Clump Thickness

*x*2 : Uniformity of Cell Size

*x*3 : Uniformity of Cell Shape

*x*4 : Marginal Adhesion

*x*5 : Single Epithelial Cell Size

*x*6 : Bare Nuclei

*x*7 : Bland Chromatin

*x*8 : Normal Nucleoli

*x*9 : Mitoses

**Output:**

*y* : output class (0 for benign and 1 for malignant)

**Class Distribution:** Benign –458 (65.5%), Malignant - 241(34.5%)

**Missing Values:**

The dataset has total 16 missing values in the feature *x*6. All the missing values have been represented as -1.

**References**:

1. M. F. Akay, “Support vector machines combined with feature selection for breast cancer diagnosis”, *Expert Systems With Applications*, 36(2; 2): 3240-3247, 2009.
2. O.L. Mangasarian, W. N. Street, W.H. Wolberg, “Breast Cancer Diagnosis and Prognosis via Linear Programming”, *Operations Research*, 43(4): 570-577, 1995.
3. G. I. Salama, M.B. Abdelhalim, M. A. Zeid, “Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers”, *International Journal of Computer and Information Technology*, Vol. 1, Issue 1, pp 36-43, 2012.
4. B. L. Rodrigues, “Analysis of the Wisconsin Breast Cancer Dataset and machine Learning for Breast Cancer Detection”, *Proceedings of XI workshop de Visao Computational*, October 2015.
5. D. Bazazeh and R. Shubair, “Comparative study of machine learning algorithms for breast cancer detection and diagnosis,” 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA), Ras Al Khaimah, 2016, pp. 1-4.

**Guidelines for Exercise:**

1. Apply all classification techniques discussed in the classroom.
2. Apply PCA (Principal Component Analysis) to reduce the number of features and then, apply classification techniques. Compare the results with the ones obtained in part 1.