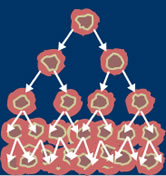
**Case Study using Breast cancer data set**

Breast cancer is one of the most common life-threatening cancers in women of any age group. Proper diagnosis of the disease is important in treating the patient. Healthcare personnel has collected vast amounts of data which could be used to train the machine to identify whether the tumor is cancerous or not. These large databases should be preprocessed, analyzed so that, predicting the future patterns of disease and mortality becomes easier, thereby enhancing our understanding the severity and outcomes.



|  |
| --- |
|  |
| **Details of the IRIS data set** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 570 | **Area:** | Health science |
| **Attribute Characteristics:** | Real | **Number of Attributes:** | 9 | **Associated Tasks:** | Classification |

**Attribute Information:**

1. clump\_thickness
2. uniformity\_of\_cell\_size
3. uniformity\_of\_cell\_shape
4. marginal\_adhesion
5. single\_epithelial\_cell\_size
6. bare\_nuclei
7. bland\_chromatin
8. normal\_nucleoli
9. mitosis

**class: 2**

Load the “breast\_cancer\_dataset.xls” and perform the following questions:

1. Describe the data with its statistical properties and discuss what you infer from it.
2. Find out the distribution of the given data set.
3. Design and implement a model to identify the combination of features which can best describe the binomial class for the given data set.
4. Develop a model to predict whether the given instance is cancerous or not.