**Mild-Term\_Project**

**Wisconsin Breast Cancer Dataset Analysis**

**Annapureddy Priyank**

**Vidit Kalani**

**MSCS 6520 – Business Analytics**

**Spring 2018**

**Prof. RJ Nowling**

***Abstract:***

This report is about the analysis of **Breast Cancer Wisconsin (Diagnostic)** Dataset, obtained from UCI ML repository. The purpose of this project is to build various KNN models, analyze the results and select an effective model.

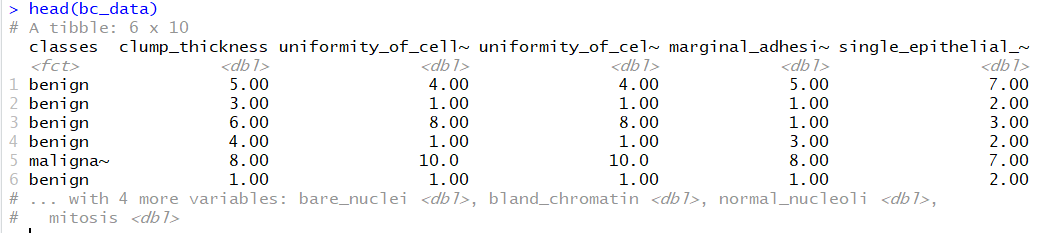
* <<https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29>>.

***About the Dataset:***

This breast cancer database was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. Dataset contains about 698 instances and 10 features. Each instance has one of the two classes: 2 or 4 (2 for benign, 4 for malignant).

There are about 458 instances (65.5%) of Benign class and 241 instances (34.5%) of malignant,

this is a class imbalanced dataset.

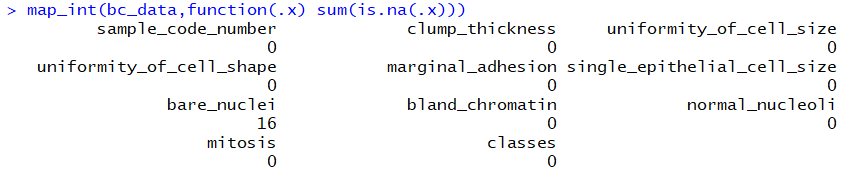


**Preparing the data:**

**Impute missing values:**

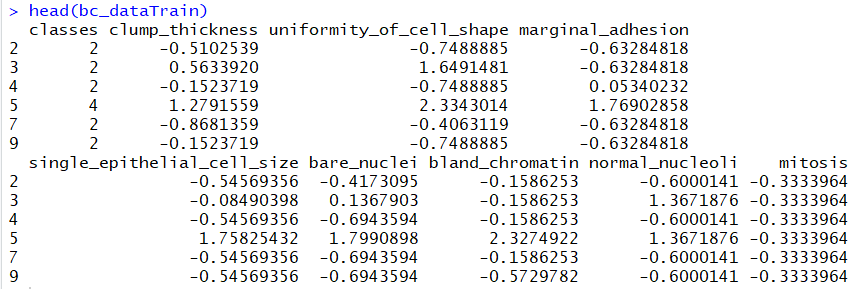
In general, all data sources include errors and missing values. Data cleaning addresses these anomalies. Missing values can be imputed either by dropping the rows or replacing them by a default value. Breast cancer dataset includes 16 instances of missing values denoted by “?”.

Missing values are handled by imputing using the most frequently occurring value 1.



**Normalize the data:**

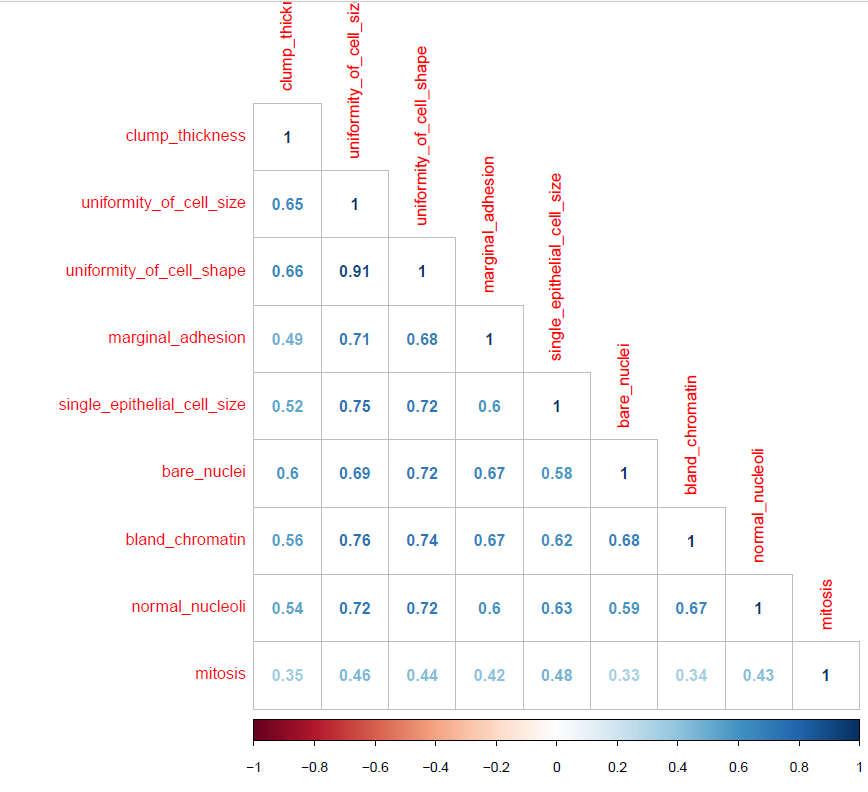
Raw data consists of features with varying scales. For algorithms like KNN where distance between pairs of samples is measured, if distinctive features have different range of values, it can often lead to biasing results or overfitting of the model. In order to avoid this, range of the independent variables need to be standardized. This is done as a part of data pre-processing step. Many machine learning algorithms (such as SVM, K-nearest neighbors, and logistic regression) assume that data is normalized.



**Data Analysis:**

**Correlation Analysis:**

Next step is finding the correlation between the variables. Correlation defines the mutual relationship or association between the variables. It helps in finding the redundancies among the features. Some classifiers assume feature independence. For classifier like kNN, adding more redundant features will artificially inflate their importance. Pearson correlation coefficients are calculated for breast cancer dataset. These coefficients quantify the degree of relationship between the variables. It is observed that uniformity of cell size and shape are highly correlated with a value of 0.91. A threshold value of 0.9 is considered to discard the highly correlated features. Uniformity of cell size is thus eliminated.

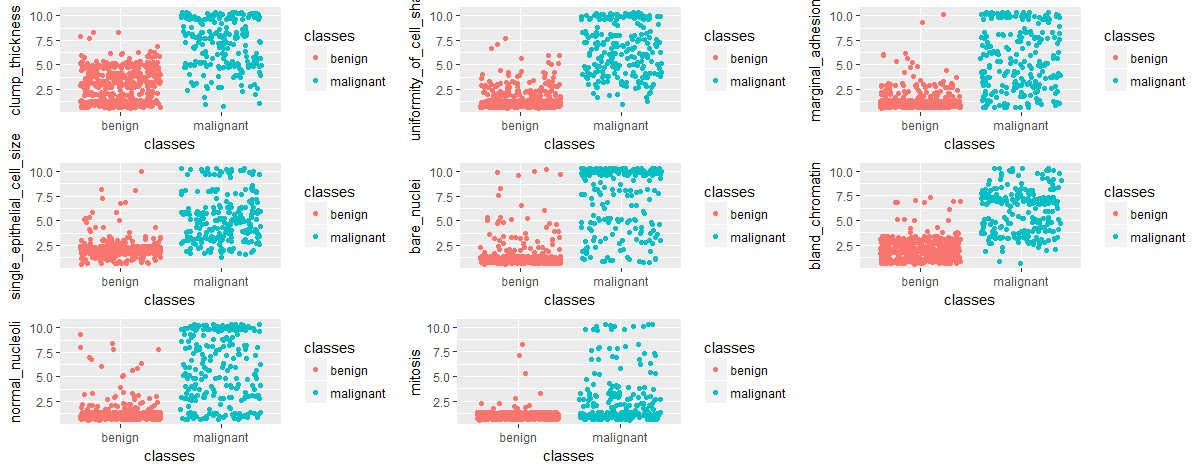


***correlation between different features***

**Jitter plot Analysis:**

In this phase we identify the features that best distinguish the values of outcome variable.

The graph below is the jitter plot between classes and other features. From the plot, clump\_thickness, uniformity\_of\_cell\_shape, bland\_chromatin, bare\_nuclei, single\_epithelial\_cell\_size, marginal\_adhesion, normal\_nucleoli are good predictors for our model. Mitosis seems to be not good at classifying the output variable hence eliminating it for building the classifier.

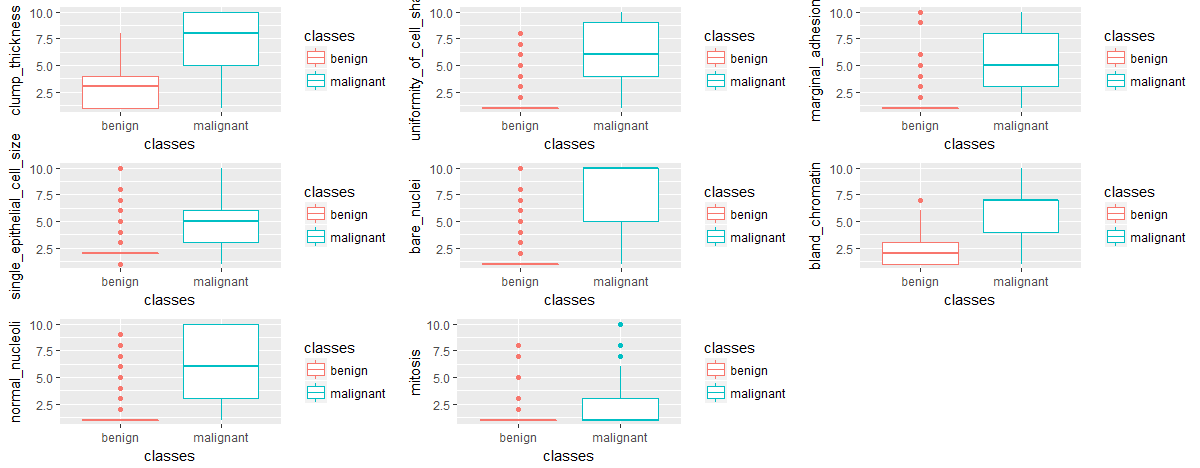


***Jitter plot analysis***

**Box plot analysis:**

Box plots tell more about the data distribution of the variables. Data distribution of bare\_nuclei, mitosis, bland\_chromatin is highly skewed. There are far more outliers for uniformity\_cell\_shape, single\_epithelial\_cell\_size , bare \_nuclei, normal\_nucleoli and mitosis.

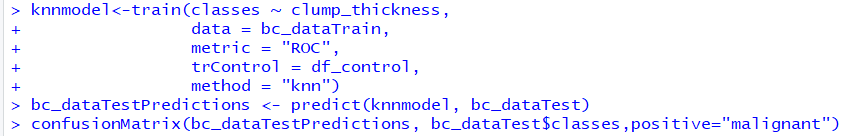
Box plot also shows that clump\_thickness and bland\_chromatin can act as good predictors.



***Boxplot analysis***

**Features & Metric Selection:**

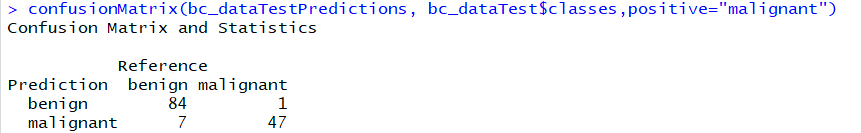
From the above plots clump\_thickness, bland\_chromatin, marginal\_adhesion, single\_epithelial\_cell\_size, bare\_nuclei, normal\_nucleoli are selected for building the first round of our model. Since the classes are imbalanced, model selection from various rounds is based on sensitivity of malignant class rather than accuracy. kNN model is trained using the metric ROC.

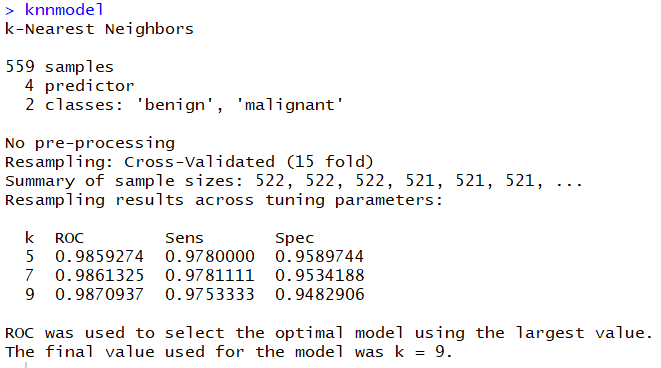


**Results:**

After the recursive forward feature selection, the effective model that is selected has 97.92% sensitivity and 94.24% accuracy. The model has 4 predictors namely Single\_epithelial\_size, bland\_chromatin, clump\_thickness and bare\_nuclei. Initially from the box plots and jitter plots though the uniformity\_of\_cell\_shape and bland\_chromatin seemed to be good predictors, single\_epithelial\_cell\_size, clump\_thickness and bare\_nuclei turned out to be good combination of predictors along with bland\_chromatin w.r.t both accuracy and sensitivity.

Confusion matrix of the final model is as shown below. Model could predict 47 out of 48 malignant classes correctly. The value of K chosen is 9





**Test Results:**

**Round: - 1**

|  |  |  |
| --- | --- | --- |
| Feature | Sensitivity | Accuracy |
| clump\_thickness | 62.5 | 83.45 |
| uniformity\_of\_cell\_shape | 85.42 | 92.09 |
| marginal\_adhesion | 60.42 | 82.01 |
| single\_epithelial\_cell\_size | 95.83 | 92.09 |
| bare\_nuclei | 89.58 | 91.37 |
| bland\_chromatin | 91.67 | 93.53 |
| normal\_nucleoli | 85.42 | 90.65 |

**Round: - 2**

|  |  |  |
| --- | --- | --- |
| Feature | Sensitivity | Accuracy |
| single\_epithelial\_cell\_size +clump\_thickness | 91.67 | 92.81 |
| single\_epithelial\_cell\_size +  uniformity\_of\_cell\_shape | 91.67 | 92.81 |
| single\_epithelial\_cell\_size + marginal\_adhesion | 79.17 | 87.05 |
| Bland\_chromatin+ single\_epithelial\_cell\_size | 97.92 | 93.53 |
| single\_epithelial\_cell\_size +  bare\_nuclei | 93.75 | 92.09 |
| single\_epithelial\_cell\_size + normal\_nucleoli | 89.58 | 90.65 |

**Round: - 3**

|  |  |  |
| --- | --- | --- |
| Feature | Sensitivity | Accuracy |
| Bland\_chromatin+  Single\_epithelial\_cell\_size+  Clump\_thickness | 97.92 | 94.96 |
| Bland\_chromatin+  Single\_epithelial\_cell\_size+  Normal\_nucleoli | 97.92 | 93.53 |
| Bland\_chromatin+  Single\_epithelial\_cell\_size+  Uniformity\_of\_cell\_shape | 93.75 | 94.24 |
| Bland\_chromatin+  Single\_epithelial\_cell\_size+  Marginal\_adhesion | 95.83 | 92.09 |
| Bland\_chromatin+  Single\_epithelial\_cell\_size+  Bare\_nuclei | 97.92 | 93.53 |

**Round: - 4**

|  |  |  |
| --- | --- | --- |
| Feature | Sensitivity | Accuracy |
| Bland\_chromatin+  Single\_epithelial\_cell\_size+  Clump\_thickness+  bare\_nuclei | 97.92 | 94.24 |

**References:**

1. O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear

programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.

1. <http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.names>
2. <https://cran.r-project.org/web/packages/caret/vignettes/caret.pdf>