**Introduction**

Cancer is a disease characterized by rapid growth of cells in the body, in form of tumor. Tumors can either be malignant (cancerous) or benign (non-cancerous). Malignant tumors destroy healthy body tissues. The breast cancer is a form malignant tumor that developed from cells in the breast.

Breast cancer is the leading cause of death among women between 40 and 55 years of age and is the second overall cause of death among women. It is estimated by world health organization’s International Agency for Research on cancer (IARC) that more than 400,000 women expire each year due to breast cancer.

A key factor in this trend is the early detection and accurate diagnosis of this disease. This is possible by performing various tests like MRI, mammogram, ultrasound and biopsy but these tests lacks high diagnostic accuracy in identifying malignant breast cancer.

The objective of this project is to develop a predictive model with the help of a Classification algorithm, k-Nearest Neighbour algorithm (kNN) that can classify cancerous tumor as benign or malignant with high accuracy and better performance. This will help pathologists to take accurate and timely decisions regarding this disease.

**Literature Review**

The research on medical diagnosis of breast cancer has been conducted several times in past and majority of them reported high classification accuracies.

In 1996, Quinlan reached 94.74% classification accuracy using 10-fold cross-validation with C4.5 decision tree method.

In 2007, Polat and Gunes, an accuracy of 98.53% was obtained through least square SVM.

In 2000, Setiono reported accuracy of 98.10 based on a feed forward neural network rule extraction algorithm.

**Dataset**

In this project, we will be downloading and preprocessing the Wisconsin Breast Cancer Dataset which was collected from University of Wisconsin Hospitals and made available at UCI Machine Learning repository. It comprises 699 patients with a total of 11 different variables. These variables will be used to create model that predict diagnosis as benign or malignant of a particular patient. Each record in the dataset represents one breast cancer tissue sample. Below is a summary of the attributes of data.

1. Sample code number: ID number

2. Clump Thickness

3. Uniformity of Cell Size

4. Uniformity of Cell Shape

5. Marginal Adhesion

6. Single Epithelial Cell Size

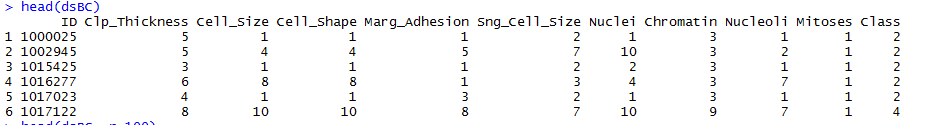
7. Bare Nuclei

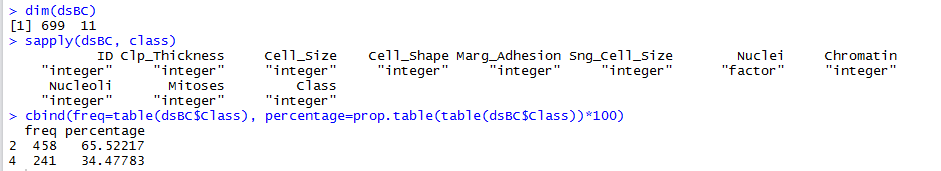
8. Bland Chromatin

9. Normal Nucleoli

10. Mitoses

11. Class: (2 for benign, 4 for malignant).





The dataset consists of nine features (from #2 to #10 above), each of which represented as an integer between 1 and 10. The 458 samples of the dataset belong to benign class (shown as 2) and other 241 are of malignant class (shown as 4). Please note that these features are computed from biopsy results of breast masses that describes the characteristics of the cell nuclei in the image.

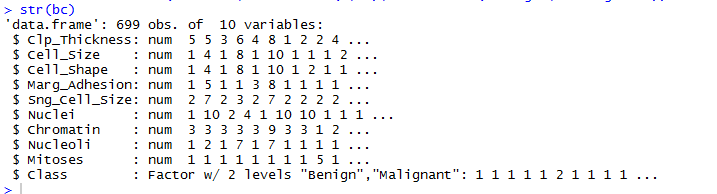
*URL to download data:*

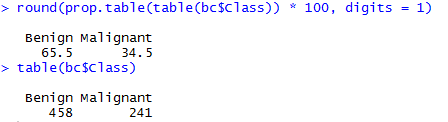
<https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data>

**Approach**

**Data Preprocessing**

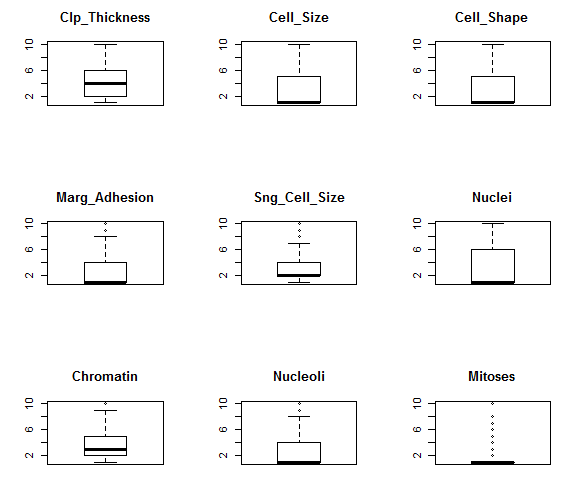
After downloading dataset, headers were added with the help of **R coding** in **RStudio**. Then ID attribute was removed as it was redundant for our experiment. The 16 records that have missing data were also removed. Then all nine features have been converted to numeric data type.

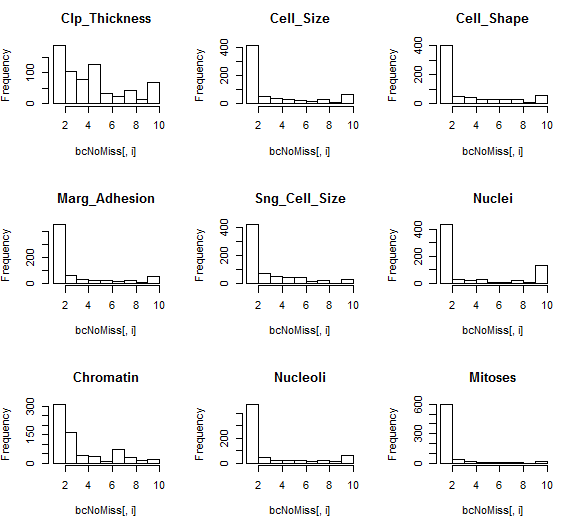




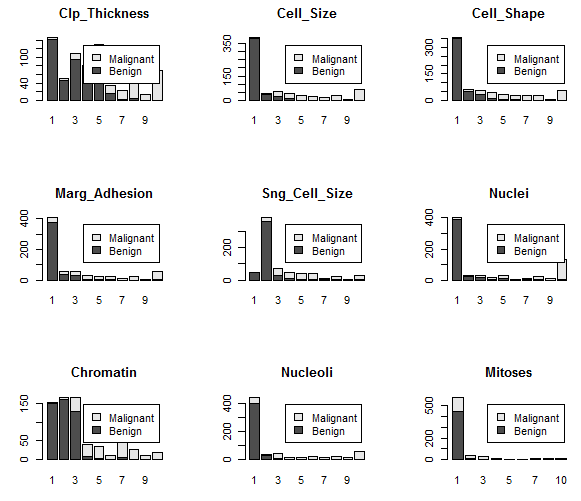
**Data Visualization**

Box and Whisker plots of input attributes shows that data is skewed.





Since data is discrete, bar plots used to observe the interaction of distribution of each attribute and how they breakdown by class value. It shows how the benign values clustered at the left (smaller values) of each distribution and malignant values are all over the place.



**Technique/Algorithm**

**k-nearest neighbors (kNN)** algorithm is the simplest and fastest among all other algorithms. Since we have low dimensionality, it is suitable for this predictive model project. It’s supervised learning as it is provided a labeled training dataset. Depending on the distance metric, kNN can be quite accurate.

There are two sets of data created for training and testing process, one with 80-20 split and other one as 50-50. Then following two models are developed for this experiment.

1. Model 1 - With all nine features (excluding missing values)
2. Model 2 – With eight features (excluding the feature that has all missing values)

*Details of split datasets on observation, variables and Class distribution in each model:*

|  |  |  |  |
| --- | --- | --- | --- |
| Model 1 | | Model 2 | |
| 80-20 split | **50-50 split** | **80-20 split** | **50-50 split** |
| 80: 548 obs. 10 var.  20: 135 obs. 10 var. | **50**: 342 obs. 10 var.  **50:** 341 obs. 10 var. | **80**: 560 obs. 9 var.  **20**: 139 obs. 9 var. | **50**: 350 obs. 9 var.  **50**: 349 obs. 9 var. |
| Ben: 356 Mal: 192  Ben: 88 Mal: 47 | **Ben**: 222 **Mal**: 120  **Ben**: 222 **Mal**: 119 | **Ben**: 367 **Mal**: 193  **Ben**: 91 **Mal**: 48 | **Ben**: 229 **Mal**: 121  **Ben**: 229 **Mal**: 120 |

**Model 1:**

The 10-fold cross-validation applied to the dataset that has all nine features (excluding missing values) which identified the optimum k-value as 7 for 80-20 split dataset. Then knn3() function was implemented with k=7 which produced the accuracy of 98.52%.

For 50-50 split dataset, the 10 -fold CV yields K=9 for train and K=5 for test. The K=5 was selected for knn3() since it provides the better result, accuracy of 95.89%.

In this scenario, 80-20 dataset performed far better than 50-50 one.

*80-20 Split Result with K=7*

|  |  |
| --- | --- |
| 10 Fold CV with Train data | 10 Fold CV with Test data |
|  |  |

|  |  |
| --- | --- |
|  |  |

*50-50 Split Result with K=5*

|  |  |
| --- | --- |
| 10 Fold CV with Train data | 10 Fold CV with Test data |
|  |  |

|  |  |
| --- | --- |
|  |  |

**Model 2:**

The 10-fold cross-validation applied to the dataset that has eight features (excluding one feature that has all missing values) which identified the optimum k-value as 9 for 80-20 split dataset. Then knn3() function was implemented with k=9 which produced 97.84% accuracy.

For 50-50 split dataset, 10-fold CV yields K=5 with the accuracy of 93.98%.

In this scenario, 80-20 dataset performed better than 50-50 one.

*80-20 Split Result with K= 9*

|  |  |
| --- | --- |
| 10 Fold CV with Train data | 10 Fold CV with Test data |
|  |  |

|  |  |
| --- | --- |
|  |  |

*50-50 Split Result with K=5*

|  |  |
| --- | --- |
| 10 Fold CV with Train data | 10 Fold CV with Test data |
|  |  |

|  |  |
| --- | --- |
|  |  |

R Codes are available at below link:

[**https://github.com/farazahma/CKME136-Capstone/tree/master/Final-Results-and-Project-Report**](https://github.com/farazahma/CKME136-Capstone/tree/master/Final-Results-and-Project-Report)

**Process Flow**

****

**Results**

Model 1 with 80-20 data split outperformed the Model 2 based on two kNN models created for this binary classification experiment. The accuracy reaches to 98.52% in model 1 as oppose to 97.84% in Model 2. The 10-fold cross-validation process helped to identify the optimum k-value and improved the performance which was 7, in this case.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Model #1** | | | | **Model #2** | | | |
|  | **80-20 Split** | | **50-50 Split** | | **80-20 Split** | | **50-50 Split** | |
| **kNN with 10 Fold Cross Validation** | **Training Accuracy %** | **Test Accuracy %** | **Training Accuracy %** | **Test Accuracy %** | **Training Accuracy %** | **Test Accuracy %** | **Training Accuracy %** | **Test Accuracy %** |
| k=5 | 96.59 | 97.8 | 96.29 | 97.17 | 95.06 | 97.12 | 97.9 | 94.46 |
| k=7 | 96.35 | 97.8 | 96.78 | 96.97 | 95.12 | 97.85 | 97.72 | 93.69 |
| k=9 | 95.99 | 96.85 | 96.78 | 96.77 | 95.54 | 97.85 | 97.62 | 93.5 |
|  |  |  |  |  |  |  |  |  |
|  | **Model #1** | | | | **Model #2** | | | |
|  | **80-20 Split** | | **50-50 Split** | | **80-20 Split** | | **50-50 Split** | |
| **Knn** | **Accuracy %** | | **Accuracy %** | | **Accuracy %** | | **Accuracy %** | |
| k=5 | 97.78 | | 95.89 | | 97.12 | | 93.98 | |
| k=7 | 98.52 | | 95.3 | | 96.4 | | 93.67 | |
| k=9 | 98.52 | | 95.9 | | 97.84 | | 93.4 | |
|  |  |  |  |  |  |  |  |  |
|  | **Model #1** | |  |  | **Model #2** | |  |  |
|  | **80-20 Split (k=7)** | **50-50 Split (k=9)** |  |  | **80-20 Split (k=9)** | **50-50 Split (k=5)** |  |  |
| **Sensitivity** | 0.9886 | 0.9865 |  |  | 1.0000 | 0.9563 |  |  |
| **Specificity** | 0.9787 | 0.9076 |  |  | 0.9375 | 0.9083 |  |  |
| **Pos Pred Value** | 0.9886 | 0.9522 |  |  | 0.9681 | 0.9522 |  |  |
| **Neg Pred Value** | 0.9787 | 0.9730 |  |  | 1.0000 | 0.9160 |  |  |

*80-20 Split Result with K=7*

|  |  |
| --- | --- |
| 10 Fold CV with Train data | 10 Fold CV with Test data |
|  |  |

|  |  |
| --- | --- |
|  |  |

**Conclusion**

In this predictive modeling project, I demonstrate the breast cancer classification process and describe the implementation of k-nearest neighbors (kNN) approach for classifying cancer as benign or malignant. It was observed that kNN technique is more efficient if implement with 10-fold cross-validation (accuracy of 98.52%).

I strongly believe that the proposed model can be beneficial for pathologists to arrive at accurate decision for their patients in timely manner.

**Future Scope**

Further exploration of the data can yield more interesting results. The following tasks will be performed for my future work.

* Use bigger dataset
* Since this is binary classification experiment with two possible output classes, ROC curves can be used in place of 10-fold cross-validation in order to identify best algorithm.
* Further tune with Bagging and boosting process.

**References**

O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.

William H. Wolberg and O.L. Mangasarian: "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December 1990, pp 9193-9196.

National Breast Cancer Foundation Inc., <http://www.nationalbreastcancer.org/about-breast-cancer>.

Xiaowei Songa, Arnold Mitnitskib,c, Jafna Coxb, Kenneth Rockwood - Comparison of Machine Learning Techniques with Classical Statistical Models in Predicting Health Outcomes, MEDINFO 2004 M. Fieschi et al. (Eds) Amsterdam: IOS Press © 2004 IMIA

Abraham Karplus - Machine Learning Algorithms for Cancer Diagnosis, Santa Cruz County Science Fair 2012.

A. Marcano-Cedeño, J. Quintanilla-Domínguez, D. Andina: "WBCD breast cancer database classiﬁcation applying artiﬁcial metaplasticity neural network", <https://www.researchgate.net/publication/222714603_WBCD_breast_cancer_database_classification_applying_artificial_metaplasticity_neural_network>

Kemal Polat, Salih Gunes: "Breast cancer diagnosis using least square support vector machine", <http://www.uz.zgora.pl/~amarcini/wiki/images/Svm.pdf>

Mehmet Fatih Akay: "Support vector machines combined with feature selection for breast cancer diagnosis", <http://download.bioon.com.cn/upload/month_0911/20091113_f233b5811a38be1fc86a5BDAmLqAe5ri.attach.pdf>