Part 2 – Own Data

Overview

In the second part of this project, we were instructed to choose our data and draw some insights and conclusions based on what we learned this semester. We chose Breast Cancer Wisconsin (Original) data set. This database was obtained from the University of Wisconsin Hospitals, Madison by Dr. William H. Wolberg. The data set is multivariate with 10 attributes and 699 instances. Each instance has a class with results of cancerous cell diagnosis (benign and malignant). As part of Exploratory analysis, we conducted pre-processing on the data and found missing values. The data was then cleaned and prepared for 10-fold cross-validation using some of the techniques we learned so far. After choosing the best model, we further analyzed the data to draw statistical insights and conclusions. The R markdown file is attached separately for reference.

Data and Model

Brief Overview of Data

The data set has a total of 11 columns and 699 rows. The first column indicates a unique identifier for each instance and the last column is the class. The class has two possible values (2=benign, 4=Malignant). Data set was compiled from 8 different groups of samples collected from 1989 to 1991. Each of the attributes from 2 to 10 has values in 1-10 range. For the ease of analysis, we renamed each attribute to a shorter form. Below is a brief description of the attributes in the data set.

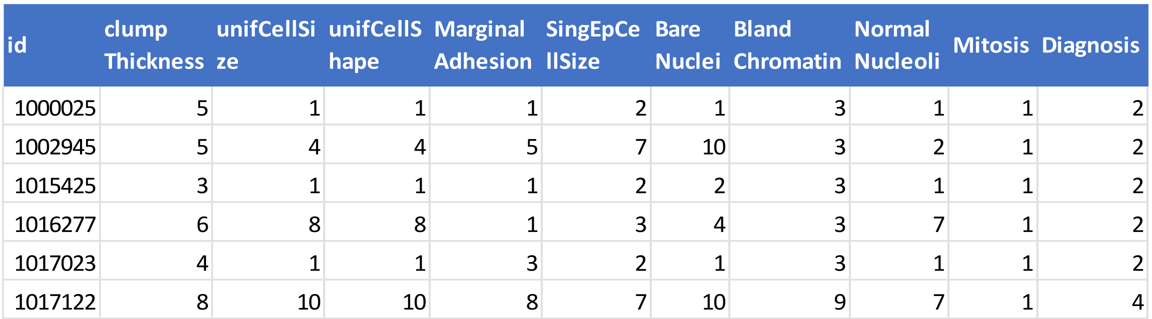
Column No.	Attributes	Domain	Name used in analysis
1	Sample code Number	Id number	id
2	Clump Thickness	1 - 10	clumpThickness
3	Uniformity of Cell Size	1 - 10	unifCellSize
4	Uniformity of Cell Shape	1 - 10	unifCellShape
5	Marginal Adhesion	1 - 10	MarginalAdhesion
6	Single Epithelial Cell Size	1 - 10	SingEpCellSize
7	Bare Nuclei	1 - 10	BareNuclei
8	Bland Chromatin	1 - 10	BlandChromatin
9	Normal Nucleoli	1 - 10	NormalNucleoli
10	Mitosis	1 - 10	Mitosis
11	Class	2 for benign, 4 for malignant	Diagnosis

Brief Description of Pre-processing

As part of the pre-processing, we looked for missing values and noise/outliers in the data. We found 16 missing values for attribute "BareNuclei". We didn't find any noise/outliers. The missing values were estimated based on the modes of the respective classes. Also, before running each model we transformed the data to fit the model requirements.

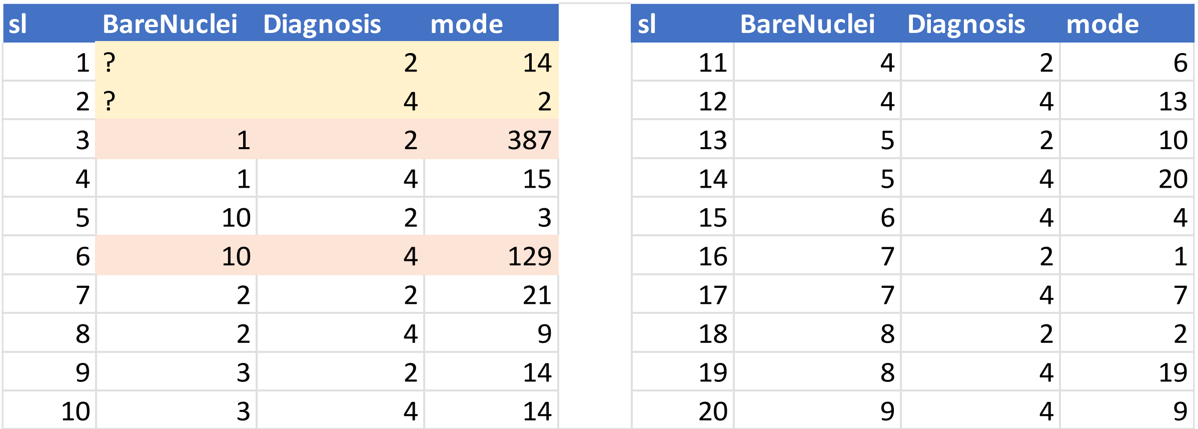
Exploratory Analysis

The data was available as a csv file. Below is a sample of the data after being loaded.



id	clump Thickness	unifCellSize	unifCellShape	Marginal Adhesion	SingEpCellSize	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitosis	Diagnosis
1000025	5	1	1	1	2	1	3	1	1	2
1002945	5	4	4	5	7	10	3	2	1	2
1015425	3	1	1	1	2	2	3	1	1	2
1016277	6	8	8	1	3	4	3	7	1	2
1017023	4	1	1	3	2	1	3	1	1	2
1017122	8	10	10	8	7	10	9	7	1	4

To find out if the data has any noise or missing values, we summarized the data. According to the data set documentations the missing data should be indicated by a "?" mark. After inspecting the summary of the data, we found that there are 16 missing data for attribute "BareNuclei". Of them 14 instances belong to class 2, and 2 instances belong to class 4. We grouped the data on attributes "BareNuclei" and class "Diagnosis" and calculated the modes.



sl	BareNuclei	Diagnosis	mode
1	?	2	14
2	?	4	2
3	1	2	387
4	1	4	15
5	10	2	3
6	10	4	129
7	2	2	21
8	2	4	9
9	3	2	14
10	3	4	14
11	4	2	6
12	4	4	13
13	5	2	10
14	5	4	20
15	6	4	4
16	7	2	1
17	7	4	7
18	8	2	2
19	8	4	19
20	9	4	9

We observed that, for class value 2 (benign), the mode of "BareNuclei" value 1 is highest (387). Similarly, for class value 4 (malignant), the mode of "BareNuclei" value 10 is highest (129). Hence, we replaced the 14 missing values of attribute "BareNuclei" of class 2 (benign) with value 1, and the rest 2 missing values of class 4 (malignant) with value 10.

Apart from missing values the data set don't have any noise/outliers in the data set. The summary of the final data set is given below.

```
##  clumpThickness unifCellSize unifCellShape MarginalAdhesion SingEpCellSize
##  1 :145         1 :384       1 :353        1 :407           1 : 47
##  2 : 50         2 : 45       2 : 59        2 : 58           2 :386
##  3 :108         3 : 52       3 : 56        3 : 58           3 : 72
##  4 : 80         4 : 40       4 : 44        4 : 33           4 : 48
##  5 :130         5 : 30       5 : 34        5 : 23           5 : 39
##  6 : 34         6 : 27       6 : 30        6 : 22           6 : 41
##  7 : 23         7 : 19       7 : 30        7 : 13           7 : 12
##  8 : 46         8 : 29       8 : 28        8 : 25           8 : 21
##  9 : 14         9 :  6       9 :  7        9 :  5           9 :  2
##  10: 69         10: 67       10: 58        10: 55           10: 31
##  BareNuclei BlandChromatin NormalNucleoli Mitosis  Diagnosis
##  1 :416     1 :152         1 :443         1 :579   2:458
##  2 : 30     2 :166         2 : 36         2 : 35   4:241
##  3 : 28     3 :165         3 : 44         3 : 33
##  4 : 19     4 : 40         4 : 18         4 : 12
##  5 : 30     5 : 34         5 : 19         5 :  6
##  6 :  4     6 : 10         6 : 22         6 :  3
##  7 :  8     7 : 73         7 : 16         7 :  9
##  8 : 21     8 : 28         8 : 24         8 :  8
##  9 :  9     9 : 11         9 : 16         10: 14
##  10:134     10: 20         10: 61
```

Models

To find out the best model to we used 10-fold cross validation on the data. The data is sampled into 10 parts randomly and used in the cross validation. Before running each 10-fold cross validation the data is transformed according to the model requirements and attribute "id" was removed. We focused our analysis on the following classifiers:

- Decision Tree (DT)
- Naïve Bayes (NV)
- Artificial Neural Network (ANN)
- Support Vector Machine (SVM)

Model	Accuracy (%)	Precision (%)	Recall (%)
DT	96.137	92.46	96.68
NV	97.42	94.071	98.755
ANN	98.28	95.98	99.17
SVM	98.998	98.75	98.34

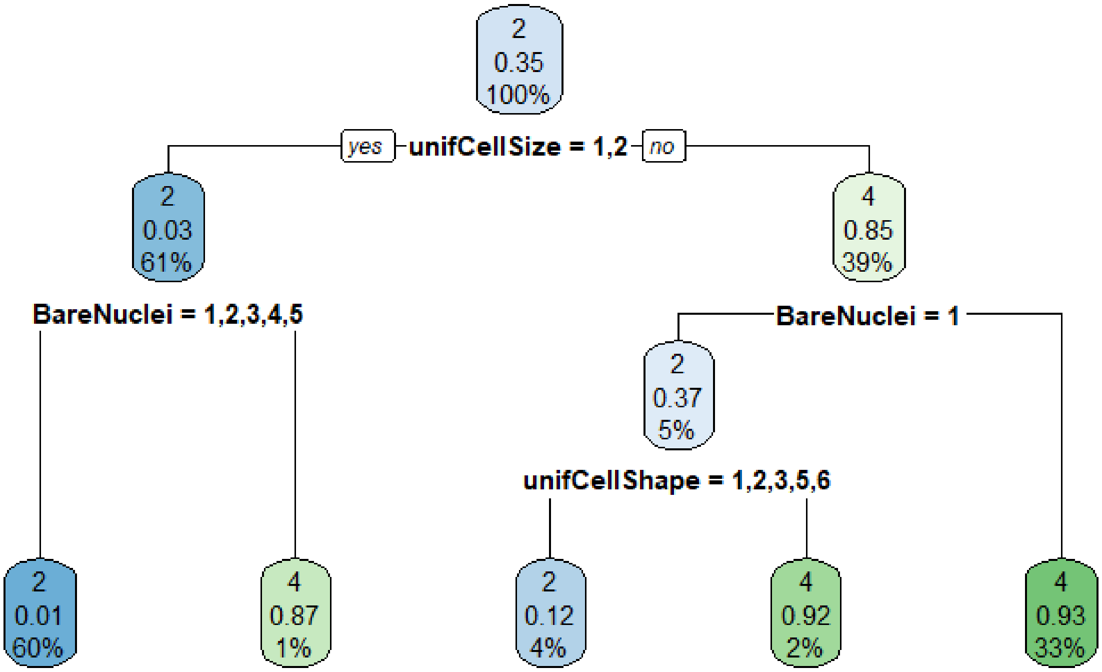
Finding Best Classifier

We inspect the models one by one and analyze the results of the 10-fold cross validation to determine the most suitable classifier for the data set.

Decision Tree

For decision tree the data is transformed accordingly and run 10-fold cross validation. The confusion matrices are accumulated to calculate the final accuracy of the model. The model gives an accuracy of 96.137% and considers the attributes – "unifCellSize", "BareNuclei", and "unifCellShape" as the main determinants.

Confusion Matrix	Actual (0)	Actual (1)
Prediction (0)	439	8
Prediction (1)	19	233



From the Decision tree plot we observe that –

- At the leftmost leaf, 60% of the data is predicted with 99% being 2 when Uniformity of Cell Size is 1 or 2 and Bare Nuclei value is 1, 2, 3, 4, or 5.
- At the rightmost leaf, 33% of the data is predicted with 93% being 4 when Uniformity of Cell Size is from 3 to 10 and Bare Nuclei is any value from 2 to 10.

Naïve Bayes

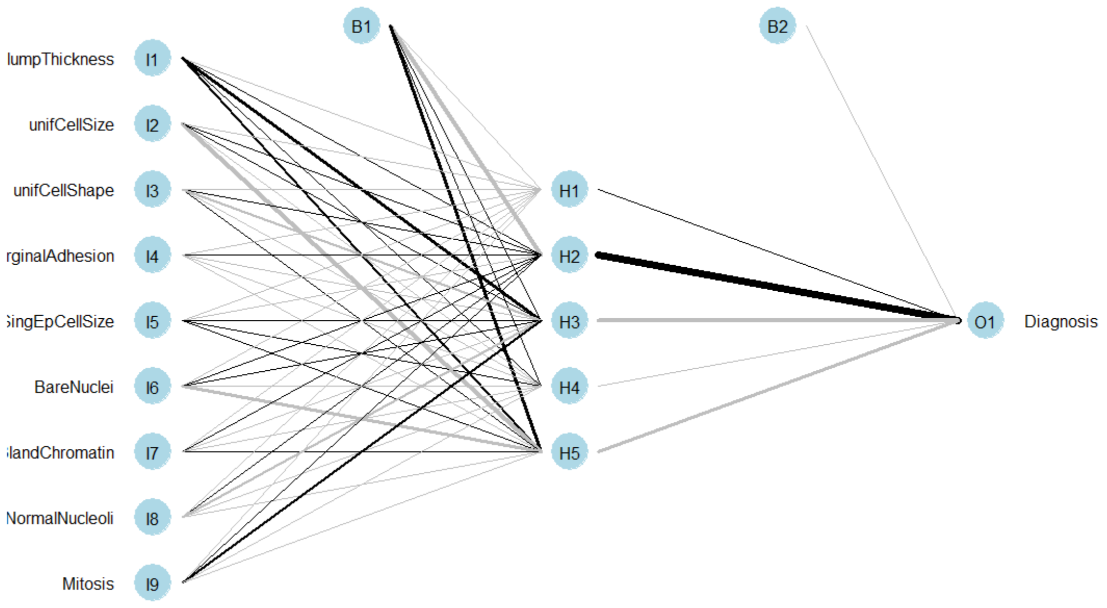
The data is made suitable for Naïve Bayes model and 10-fold cross validation is run on the data. We found that Naïve Bayes performs slightly better than decision tree with an accuracy of 97.42%.

Confusion Matrix	Actual (0)	Actual (1)
Prediction (0)	443	3
Prediction (1)	15	238

Artificial Neural Network

After running 10-fold cross validation we observed that the ANN classifier performs even better than Decision Tree and Naïve Bayes with an accuracy of 98.28%. We used 5 hidden nodes, an initial weight of 0.1 for the classifier.

Confusion Matrix	Actual (0)	Actual (1)
Prediction (0)	448	2
Prediction (1)	10	239



Support Vector Machine

SVM on the data with 10-fold cross-validation produces the highest accuracy. With a linear kernel SVM has an accuracy of 98.99% with 78 support vectors. Since this classifier has the highest accuracy, we consider SVM as the most suitable for the data set.

Confusion Matrix	Actual (0)	Actual (1)
Prediction (0)	455	4
Prediction (1)	3	237

Insights and Charts

From the data we found some interesting insights and patterns. These insights along with supporting statistics, and charts are described below.

1. *Uniformity of Cell Size and Bare Nuclei in determining Malignant cells*

To investigate if Uniformity of Cell Size and Bare Nuclei has any effect on the diagnosis of cancerous cell, we created a count plot (*Figure 1*) with Uniformity of Cell Size along x-axis and Bare Nuclei along y-axis. We used diagnosis result as the color and size as the number of cases.

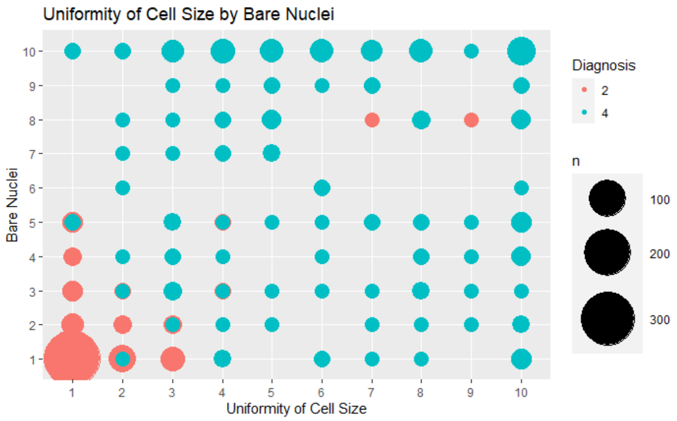


Figure 1: Uniformity of Cell Size by Bare Nuclei

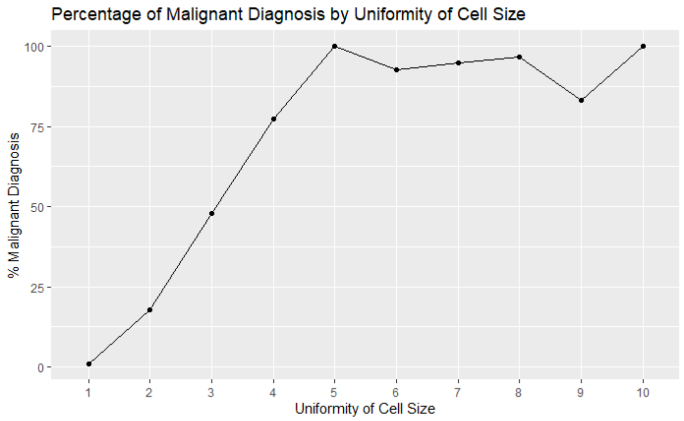


Figure 2: Percentage of Malignant Diagnosis by Uniformity of Cell Size

From *Figure 1* we observe that there is a clear relationship in the diagnosis of a Malignant cell with the value of Uniformity of Cell Size and Bare Nuclei. If the value of Uniformity of cell size is greater than 3 and Bare Nuclei is greater than 5 then there is a very significant chance of the cell to be malignant.

To find out how the number of cases of malignant cell changes with Uniformity of cell size we plotted the line graph (*Figure 2*). This chart has Uniformity of Cell Sizes along the x-axis and percentage of cases with Malignant diagnosis along y-axis.

We observe from the chart that if the value of Uniformity of Cell Size is 4 or greater, then more than 75% of the cells are malignant.

2. *Uniformity of Cell Shape and Bare Nuclei in determining Malignant cells*

Next, we investigate the Uniformity of Cell Shape and Bare Nuclei in diagnosing cancerous cells. We created a mosaic chart (*Figure 3*) to find patterns in the data. Diagnosis (benign, Malignant) is used as the color to clearly define the patterns. We have taken values of Uniformity of Cell Shape as x-axis and Bare Nuclei as y-axis.

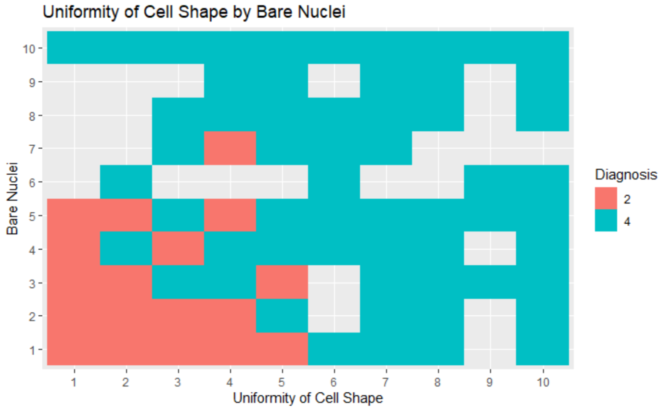


Figure 3: Uniformity of Cell Shape by Bare Nuclei

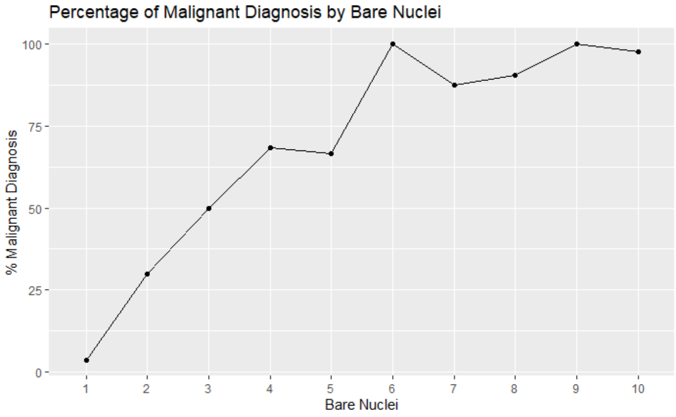


Figure 4: % of Malignant cells by Bare Nuclei

In *Figure 3* we can see a clear pattern and can conclude that a relationship exists between Uniformity of Cell Shape and Bare Nuclei to determine the Malignant diagnosis of a cell. To be more specific, if the value of Uniformity of Cell Shape is greater than 5 and Bare Nuclei is Greater than 5 there is a very significant chance of the cells to be Malignant.

To find out how Bare Nuclei influences the cells to be malignant we plotted a line chart (*Figure 4*) taking Bare Nuclei along the x-axis and percentage of malignant diagnosis of cells along the y-axis. This show an upward trend of cells to be malignant with the increase of Bare Nuclei.

As an example, if the Bare Nuclei value is 5 or greater, then in more than 68% of cases cells are malignant.

3. Uniformity of Cell Size and Shape in determining Malignant cells

Finally, we investigated the relationship in Uniformity of Cell Size and Shape in determining a cells diagnosis. We created a count chart (*Figure 5*) taking Uniformity of Cell Size along x-axis and Cell Shape along y-axis. Like above charts, diagnosis (benign, malignant) is used to color the points and the number of cases is used as the size of the points.

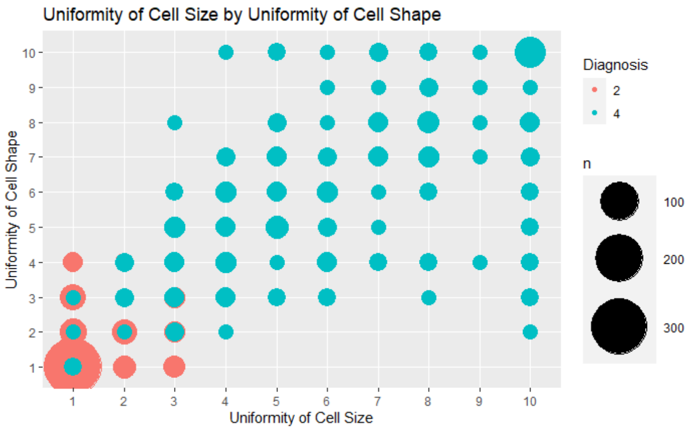


Figure 5: Uniformity of Cell Size by Cell Shape

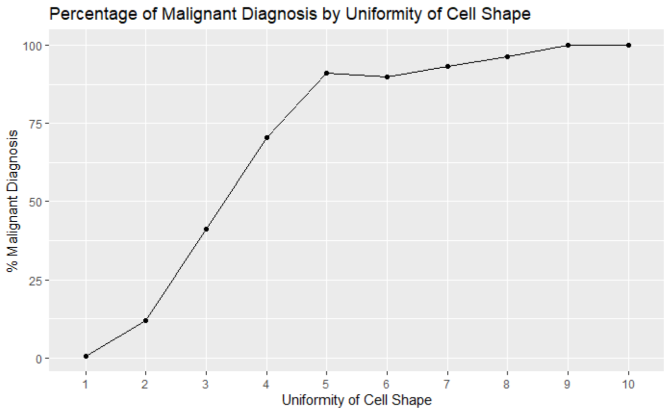


Figure 6: % of Malignant cells by Unif. of Cell Shape

From *Figure 5* we observe a clear patter in the diagnosis of a cell to be malignant with Uniformity of Cell Size and Shape. Specifically, if Cell Size goes above 3 and Cell Shape goes above 4 the likelihood of cells to be cancerous is very high.

To statistically show how Uniformity of Cell Shape influences the cells to be malignant we plotted a line chart (*Figure 6*) taking Uniformity of Cell Shape along the x-axis and percentage of malignant diagnosis of cells along the y-axis. This also show an upward trend of cells to be malignant with the increase of uniformity of cell shape.

As an example, if the Uniformity of Cell Shape is 5 or greater then in more than 88% of cases cells are malignant with at 9 reaching 100%.

Conclusion

For the chosen data we conclude that Support Vector Machine with a linear kernel has the highest accuracy and therefore, given the attributes, is best suited to predict the diagnosis of cancerous cells. The accuracy of Artificial Neural Network is also noteworthy and can be substituted as a classifier for this data set. From the chart and statistics of the data we observe that Uniformity of Cell Size, Uniformity of Cell Shape and Bare Nuclei attributes are the main determinants in predicting malignant cells.