# UE17CS303- MACHINE LEARNING ASSIGNMENT

# BREAST CANCER DETECTION

| | | |
|---|---|---|
| 1. | DRASTI N VADHAR | PES1201700686 |
| 2. | G SHREYA | PES1201700084 |
| 3. | ISHWAR CHOUDHARY | PES1201700189 |

## **Problem Statement:**

Build classification models for Breast cancer detection.
We were given a dataset with 699 instances and 10 attributes:

- Clump Thickness
- Uniformity of Cell Size
- Uniformity of Cell Shape
- Marginal Adhesion
- Single Epithelial Cell Size
- Bare Nuclei
- Bland Chromatin
- Normal Nucleoli
- Mitoses

Including a target attribute that could take the values:

- 2 indicating Benign
- 4 indicating Malignant

Also there were some missing values which were filled on the basis of the most frequently occurring value i.e the mode, for that particular attribute .

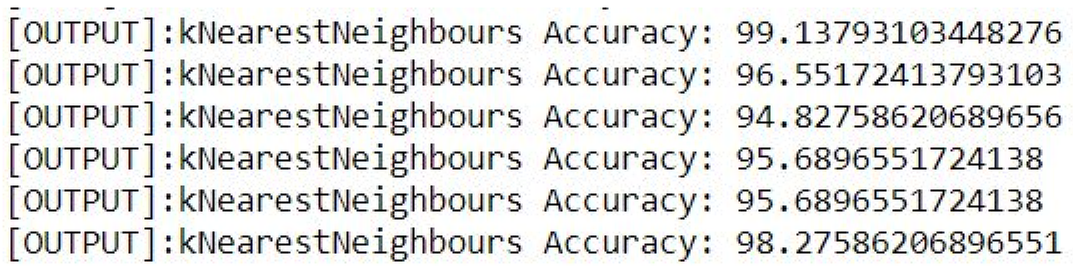
## **ML Techniques Used:**

Two methods were employed

- Supervised Learning-
  1. K Nearest Neighbors
  2. Neural Networks
- Unsupervised Learning- For Visualization and further improvement
  1. K Means Clustering

### *K Nearest Neighbors:*

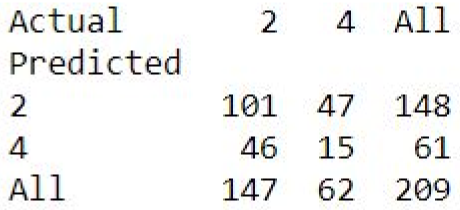
We implemented K Nearest Neighbors using K fold validation. In general K fold validation is employed when the dataset is small to train and test, so we split the dataset into 'K' folds , where we train the dataset for K-1 folds, and test the model on the Kth fold, this is repeated K times. In this way each fold will get the opportunity to be the test set.
So in our model, we chose K to be 6 (as per problem statement), and the following accuracies were obtained:



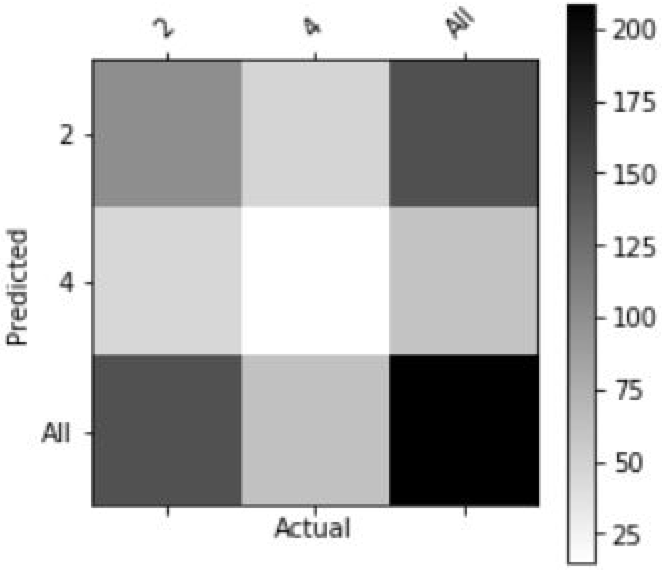
```
[OUTPUT]:kNearestNeighbours Accuracy: 99.13793103448276
[OUTPUT]:kNearestNeighbours Accuracy: 96.55172413793103
[OUTPUT]:kNearestNeighbours Accuracy: 94.82758620689656
[OUTPUT]:kNearestNeighbours Accuracy: 95.6896551724138
[OUTPUT]:kNearestNeighbours Accuracy: 95.6896551724138
[OUTPUT]:kNearestNeighbours Accuracy: 98.27586206896551
```

The final accuracy of the model would be the average of these 6 accuracies which is 96.68%

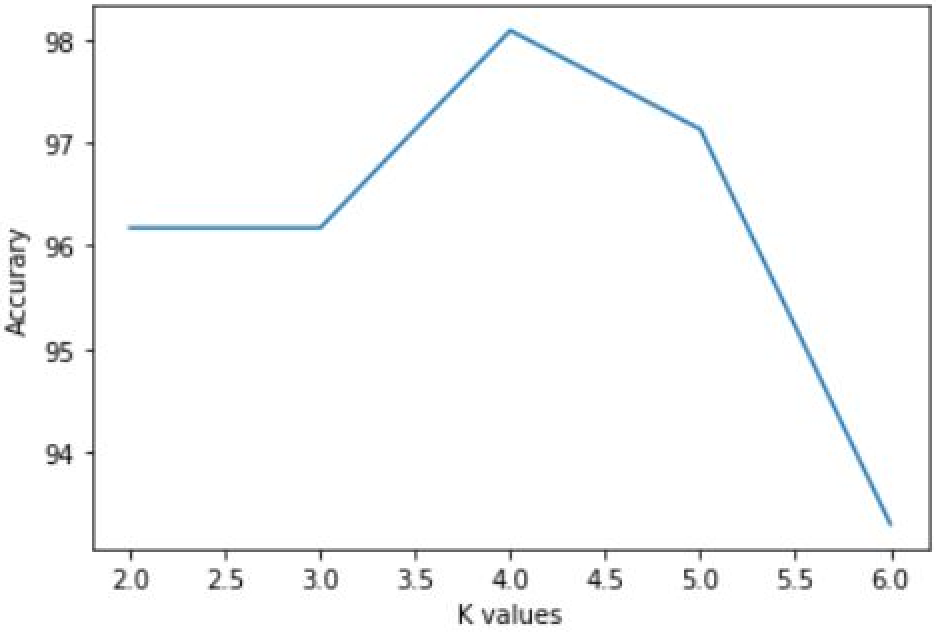
We also implemented this model without K fold validation, where in we used 70% of the dataset for training and 30% for testing and we got an accuracy of 97.6%



```
Actual       2   4  All
Predicted
2          101  47  148
4           46  15   61
All        147  62  209
```



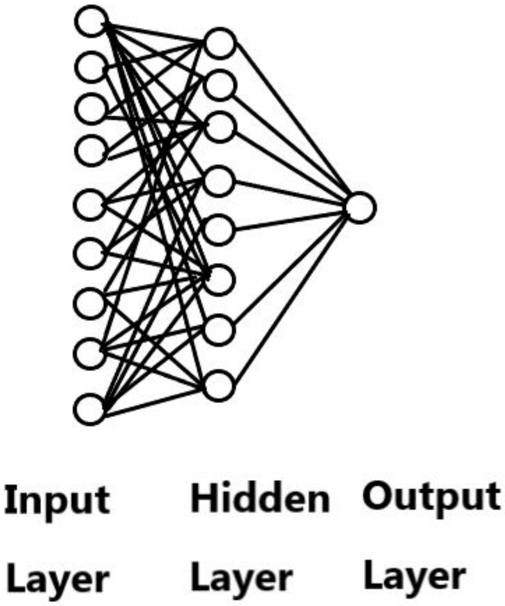
The above figure represents a confusion matrix
The graph below gives us an idea about the optimum k value that can be chosen for the KNN model.



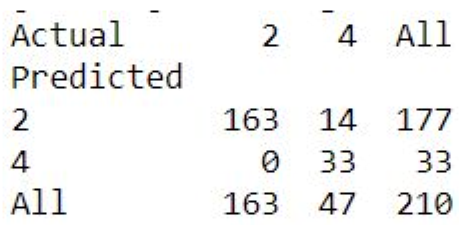
## *Neural Networks:*

We implemented a neural net with the following:

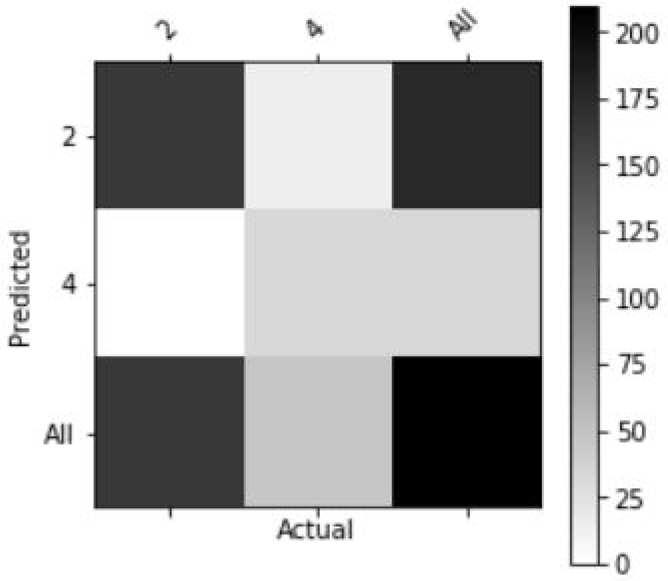
- 9 inputs, 1 hidden layer with 8 neurons and 1 output layer
- Activation unit used - Sigmoid



We used 70% of the dataset for training and 30% for testing and got an accuracy of 88%.
And the confusion matrix for the same is given below-



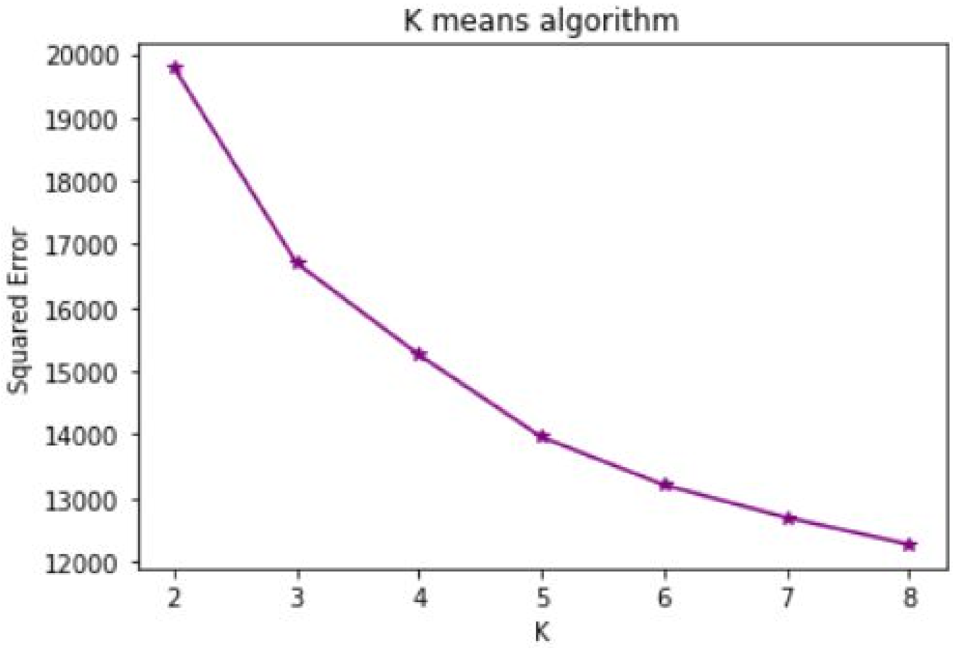
```
Actual       2   4  All
Predicted
2          163  14  177
4            0  33   33
All        163  47  210
```



## *Unsupervised Learning: K Means Clustering*

We ran the K means clustering algorithm for a bunch of k values =[2,3,4,5,6,7,8], and plotted the graph of squared error distance against the k values.
Squared error distance is calculated using the formula:

$$\mathcal{L}(K) = \sum_{j=1}^{m} ||\mu_{C(j)} - \mathbf{x}_j||^2$$



According to this graph the elbow point is found to be at k=4. This is the optimum k value.

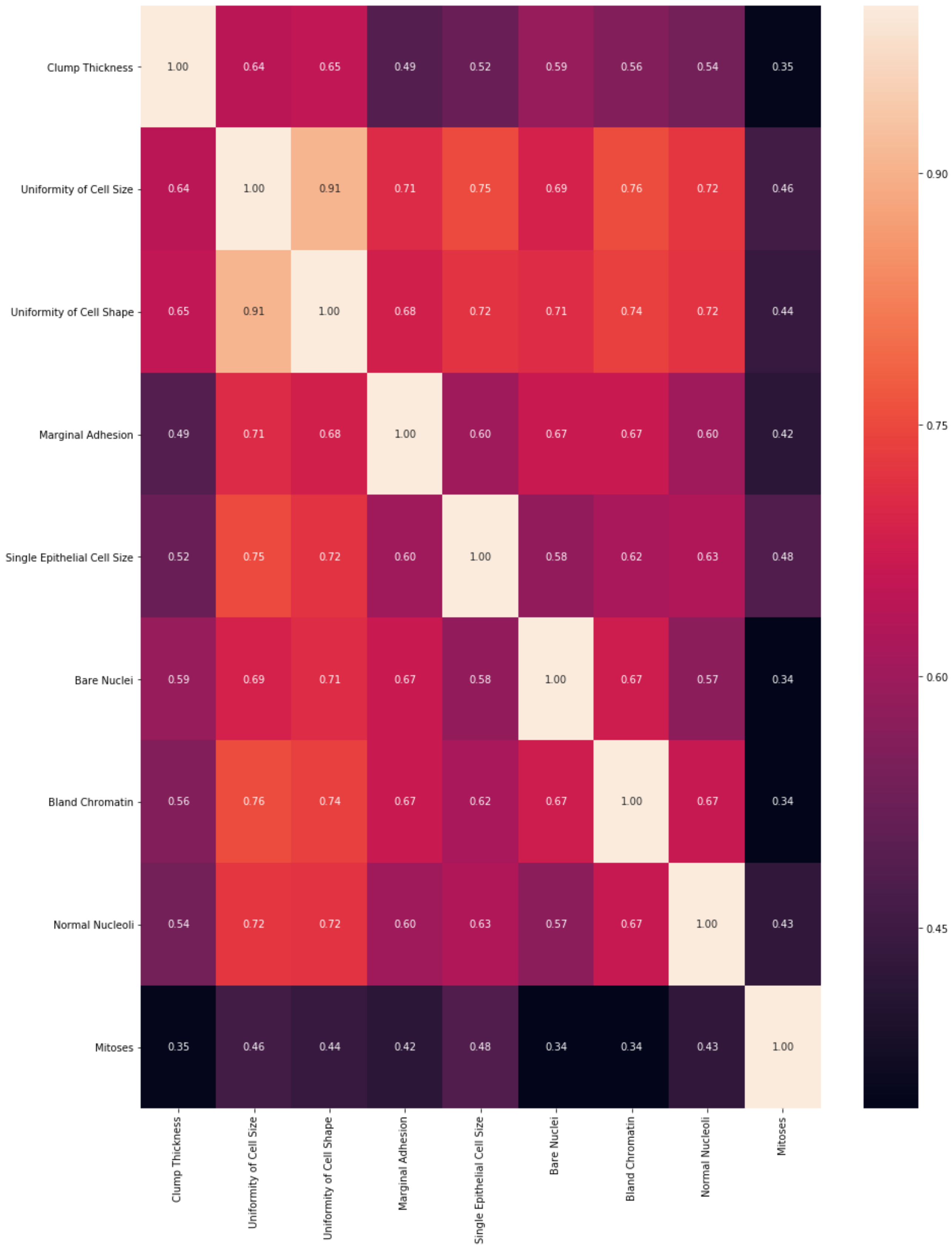
Using this:

- We could train an SVM on the resulting clusters. Then use SVM for classification.
- We could use a 1NN classifier, "trained" on the cluster centroids only.

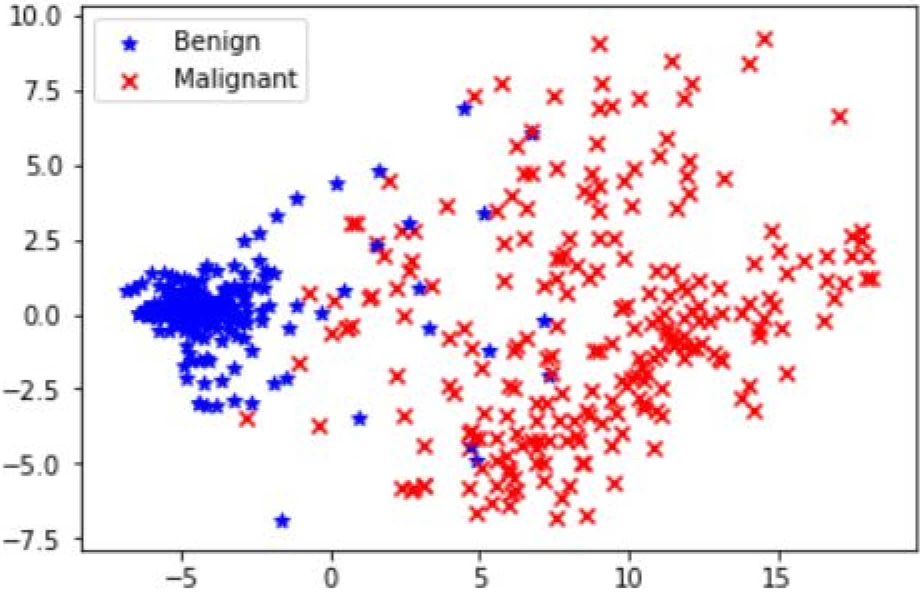
So unsupervised learning could further be used to in addition with other models to get improved results.

## **Graphs/Visualizations:**

- Correlation matrix gives a picture of how related any 2 attributes are
  We can see that the diagonal elements have a correlation of 1 indicating that every attribute is strongly related to itself. The beige/ skin color in the matrix is indicative of the next highest correlation value which is 0.91 that corresponds to the attributes- Uniformity in Cell Shape and Uniformity in Cell Size.


Clump Thickness
1.00 0.64 0.65 0.49 0.52 0.59 0.56 0.54 0.35
Uniformity of Cell Size
0.64 1.00 0.91 0.71 0.75 0.69 0.76 0.72 0.46
Uniformity of Cell Shape
0.65 0.91 1.00 0.68 0.72 0.71 0.74 0.72 0.44
Marginal Adhesion
0.49 0.71 0.68 1.00 0.60 0.67 0.67 0.60 0.42
Single Epithelial Cell Size
0.52 0.75 0.72 0.60 1.00 0.58 0.62 0.63 0.48
Bare Nuclei
0.59 0.69 0.71 0.67 0.58 1.00 0.67 0.57 0.34
Bland Chromatin
0.56 0.76 0.74 0.67 0.62 0.67 1.00 0.67 0.34
Normal Nucleoli
0.54 0.72 0.72 0.60 0.63 0.57 0.67 1.00 0.43
Mitoses
0.35 0.46 0.44 0.42 0.48 0.34 0.34 0.43 1.00
Clump Thickness
Uniformity of Cell Size
Uniformity of Cell Shape
Marginal Adhesion
Single Epithelial Cell Size
Bare Nuclei
Bland Chromatin
Normal Nucleoli
Mitoses
0.90
0.75
0.60
0.45

- PCA Visualization:


★ Benign
× Malignant