**ACCURACY ANALYSIS OF VARIOUS CLASSIFICATIONS MODELS APPLIED ON A BREAST CANCER CLASSIFICATION DATASET:**

DATSET NAME-BREAST CANCER.TXT

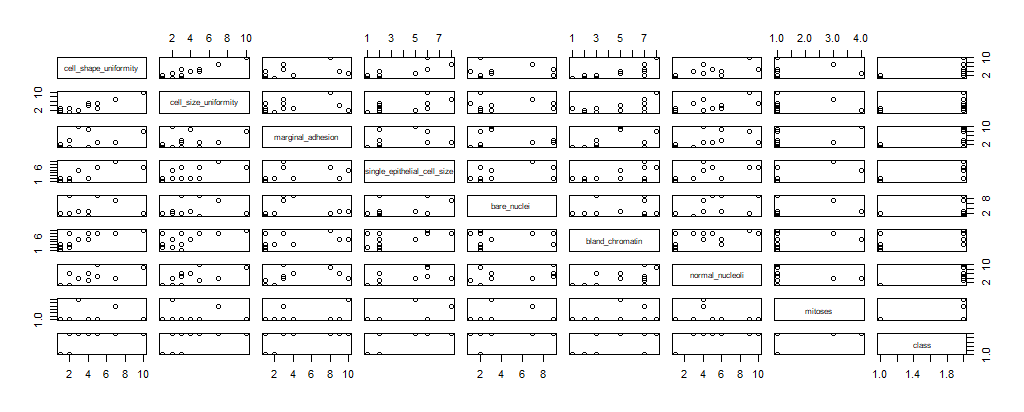
PREDICTION VARIABLES:

* **CLUMP\_THICKNESS**
* **CELL\_SHAPE\_UNIFORMITY**
* **CELL\_SIZE\_UNIFORMITY**
* **MARGINAL\_ADHESION**
* **SINGLE\_EPITHELIAL\_CELL\_SIZE**
* **BARE\_NUCLEI**
* **BLAND\_CHROMATIN**
* **MITOSES**

**DEPENDENT VARIABLE /VARIABLE TO BE CLASSIFIED:**

* **CLASS**

**BASIC VISUAL ANALYSIS OF VARIOUS VARIABLES:**



**IN THIS CASE WE CANNOT JUDGE ANY SIGNS OF MULTICOLLINEARITY BECAUSE DATA IS OF CLASSIFICATION TYPE THERFORE WE HAVE TO USE A CORRELATION MATRIX TO CHECK FOR ANY KIND OF CORRELATION:**

**CORRELATION TABEL FOR BREAST CANCER DATA SET:**

|  |
| --- |
| clump\_thickness cell\_shape\_uniformity cell\_size\_uniformity marginal\_adhesion |
| clump\_thickness 1.0000000 0.6449125 0.6545891 0.4863562 |
| cell\_shape\_uniformity 0.6449125 1.0000000 0.9068819 0.7055818 |
| cell\_size\_uniformity 0.6545891 0.9068819 1.0000000 0.6830792 |
| marginal\_adhesion 0.4863562 0.7055818 0.6830792 1.0000000 |
| single\_epithelial\_cell\_size 0.5218162 0.7517991 0.7196684 0.5995991 |
| bare\_nuclei 0.3630407 0.4403379 0.4410441 0.2999160 |
| bland\_chromatin 0.5584282 0.7557210 0.7359485 0.6667153 |
| normal\_nucleoli 0.5358345 0.7228648 0.7194463 0.6033524 |
| mitoses 0.3500339 0.4586931 0.4389109 0.4176328 |
|  |
|  |
| single\_epithelial\_cell\_size bare\_nuclei bland\_chromatin normal\_nucleoli mitoses |
| clump\_thickness 0.5218162 0.3630407 0.5584282 0.5358345 0.3500339 |
| cell\_shape\_uniformity 0.7517991 0.4403379 0.7557210 0.7228648 0.4586931 |
| cell\_size\_uniformity 0.7196684 0.4410441 0.7359485 0.7194463 0.4389109 |
| marginal\_adhesion 0.5995991 0.2999160 0.6667153 0.6033524 0.4176328 |
| single\_epithelial\_cell\_size 1.0000000 0.3748831 0.6161018 0.6288807 0.4791015 |
| bare\_nuclei 0.3748831 1.0000000 0.3754564 0.4086385 0.2323590 |
| bland\_chromatin 0.6161018 0.3754564 1.0000000 0.6658778 0.3441695 |
| normal\_nucleoli 0.6288807 0.4086385 0.6658778 1.0000000 0.4283357 |
| mitoses 0.4791015 0.2323590 0.3441695 0.4283357 1.0000000 |

HERE WE SEE STRONG SIGNS OF COLLINEARITY BETWEEN **CELL\_SHAPE\_UNIFORMITY AND CELL\_SIZE\_UNIFORMITY**

**WE CAN AVOID** ANY CORRECTIVE ACTION AS BOTH THESE VARIABLES DOESN’T EFFECT THE CLASSIFICATION MODELS LARGELY WHEN USED TOGETHER.

**CLASSIFICATION MODELS USED:**

* **KNN-K NEAREST NEIGHBOUR**
* **LOGISTICS REGRESSION**
* **LDA-LINEAR DISCRIMINANT ANALYSIS**
* **SVM-SUPPORT VECTOR MACHINES**
* **TREE CLASSIFICATION**
* **NEURAL NETWORKS**

|  |  |  |
| --- | --- | --- |
| **SNO** | **MODEL USED** | **DETAILS AND ACCURACY REPORTED** |
| **1** | **KNN** | **> #without scaling we find highest accuracy of 95.161 for k=4,5**  **> #with scaling we find highest accuracy at k=5,6,7,10 i.e. 94.354% less then without scaled data** |
| **2** | **LOGISTIC REGRESSION** | **REPORTED ACCURACY-92.74194**  **> #we cannot apply the principle of precision and recall here**  **> #as these are not true and false type categorization** |
| **3** | **LDA** | **REPORTED ACCURACY-93.54839**  **> #we cannot apply the principle of precision and recall here**  **> #as these are not true and false type categorization** |
| **4** | **SVM** | **REPORTED ACCURACY-** **93.54839** |
| **5** | **DECISION TREE** | **REPORTED ACCURACY-** **91.93548** |
| **6** | **BAGGING** | **REPORTED ACCURACY-** **93.54839** |
| **7** | **RANDOM FOREST** | **REPORTED ACCURACY-** **95.16129** |
| **8** | **NEURAL NETWORK** | **REPORTED ACCURACY-** **90.32258** |

OBSERVATIONS:

* OUT OF THE AVAILABLE CLASSIFICATION MODELS IT IS FOUND THAT THE MODELS KNN AND RANDOM FOREST HAVE PREDICTED THE CLASS OF BREAST CANCER MORE ACCURATELY THEN ANY OTHE MODEL
* WE HAVE NOT USED THE PRINCIPLES SUCH AS PRECISION,RECALL,FALSE POSITIVE RATE,FALSE NEGATIVE RATES ETC.BECAUSE HERE WE ARE NOT CATEGORIZING DATA IN TRUE AND FALSE VALUE BUT RATHER WE ARE CLASSIFYING TO THE CLASS OF CANCER A PARTICULAR DATA INSTANCE MIGHT BELONG TO
* IN CASE OF KNN THE MODEL USED IS THE ONE WITH K=5,4 AND USING THE SEED VALUE GIVEN IN THE R CODE ATTACHED WITH THE FILE UNDER NAME KNN
* RANDOM FOREST HAS ALSO BEEN ABLE TO CLASSIFY THE DATA IN CORRECT CLASS OF CANCER BETTER THEN OTHER MODELS

RESULT:

* KNN AND RANDOM FOREST SEEMS TO BE THE MOST USEFUL MODELS FOR CALSSIFICATIONS OF THIS DATASET