INTRODUCTION

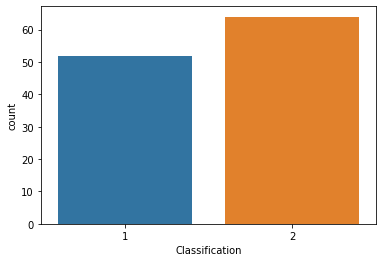
DATASET

The labelled dataset used for this project is obtained from [UCI Machine Learning Repository](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra). The dataset contains nine anthropometric parameters (which can be gathered via routine blood analysis) associated to 116 subjects. Each of these record in turn are classified either as ‘1’ or ‘2’ implying if the record is of a subject with ‘Healthy Controls’ or of a ‘Patient’.

EVALUATION METRIC

DATA ANALYSIS

First we found the distribution of the target class and turned out that of the 116 records 52 were of subjects with ‘Healthy Controls’ and remaining were of subjects who were ‘Patients’.

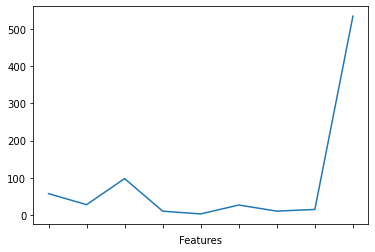


Outlier Identification

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

Since the data seems to have certain outliers we will use Local Outlier Factor as the data is low dimensional

So the distibution of the output data are imbalanced. In the next step we determined the mean of the individual features and it was immedeitly visible that a feature scaling would be required to eliminate tendencies of domination of the features with higher range values during traning of the models.



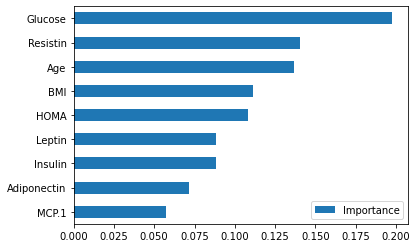
We used **statistical normalization** as our preferred method of feature scaling (Equation – 1).

Equation 1 : Statistical Normalization

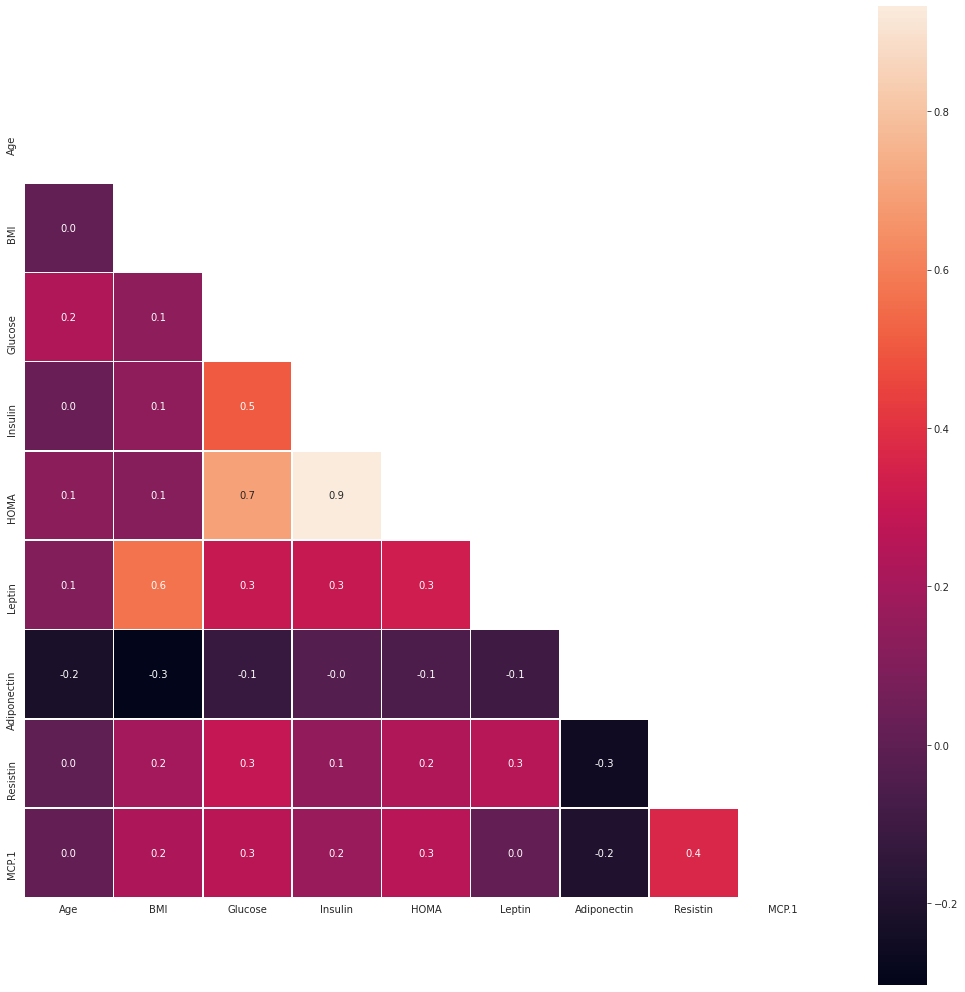
Subsequently determined the importance of the individual features (Table-1) in the eventual determination of the target outcome.

|  |  |
| --- | --- |
| Features | Importance |
| Glucose | .1977570 |
| Resistin | .1404320 |
| Age | .1368260 |
| BMI | .1113990 |
| HOMA | .1082700 |
| Leptin | .0883271 |
| Insulin | .0880923 |
| Adiponectin | .0712979 |
| MCP.1 | .0575987 |

Table : Feature Importance (Descending Order)



Pearson Coefficient Matrix :



MODELS :

**Model Training** : The model training in both cases was done by steadily decreasing the training set concentration (and equally increasing concentration of the test set). Then once the model was trained, its prediction accuracy was measured on the test set. In order to eliminate skewness from the prediction accuracy, median of 100 iterations were measured in each case, instead of one.

**Supervised Models** : The models involved in the analysis are Naïve Bayes (NB), Logistic Regression (LR), K-Nearest Neighbors (KNN), Support Vector Machine (SVM) and Random Forest (RF). The result from that analysis is recorded in **percentage terms** in the table underneath.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Train | Test | NB | LR | KNN | SVM | RF |
| 90% | 10% |  |  |  |  | 79.31 |
| 85% | 15% |  |  |  |  | 79.07 |
| 80% | 20% |  |  |  |  | 75.86 |
| 75% | 25% |  |  |  |  | 76.39 |
| 70% | 30% |  |  |  |  | 75.58 |
| 65% | 35% |  |  |  |  | 69.31 |
| 60% | 40% |  |  |  |  | 73.04 |
| 55% | 45% |  |  |  |  | 70.54 |
| 50% | 50% |  |  |  |  | 72.73 |

Table : Comparative Analysis of Traditional Models

**Deep Learning Model** : I tested the same dataset with a single-layered artificial neural network. The following hyperparameters were used in for all training experiments.

|  |  |
| --- | --- |
| Parameter | Description |
| Kernel Initializer | Uniform |
| Optimizer | SGD |
| Loss | Binary Crossentropy |

Table : List of Hyperparameters

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Train | Test | Median | Minimum | Maximum |
| 90% | 10% | 72.41 |  |  |
| 85% | 15% | 76.74 |  |  |
| 80% | 20% | 67.24 |  |  |
| 75% | 25% | 70.83 |  |  |
| 70% | 30% | 70.93 |  |  |
| 65% | 35% | 62.37 |  |  |
| 60% | 40% | 66.08 |  |  |
| 55% | 45% | 70.54 |  |  |
| 50% | 50% | 68.53 |  |  |

Table : Result of ANN Models over 100 iterations

RESULTS :

CONCLUSION

REFERENCES