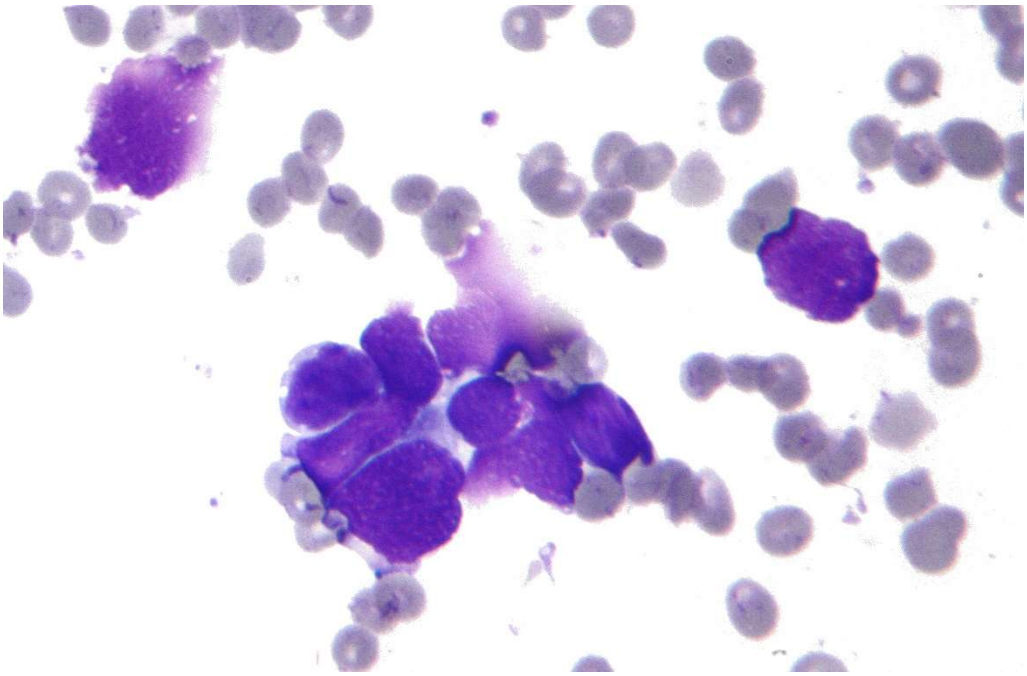


Breast Cancer Diagnosis Classification

Predict whether the cancer is benign or malignant

BY SAMANEH MAROUZI

Overview:

In this project we are going to use 30 different features to predict the Stage of Breast Cancer, "M" (Malignant) and "B" (Bengin). This classification has been done using Basic Machine Learning Algorithms. Our data frame consists of 569 obs. Of 32 variables. Features 3- 32 are divided into three parts: Mean (3-13), SE (13-23) and Worst(23-32) which each contains 10 features. Here "Mean" means the means of the all cells, "SE" means standard error of all cells and Worst means the worst cells.

Attribute information:

ID number

Diagnosis (M = malignant, B = benign)

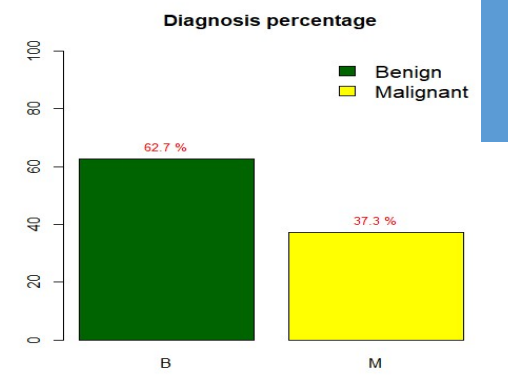
To 32. Ten real-valued features are computed for each cell nucleus: a) radius (mean of distances from center to points on the perimeter) b) texture (standard deviation of gray-scale values)c) perimeter. d) area. e) smoothness (local variation in radius lengths) f) compactness (perimeter^2 / area - 1.0) g) concavity (severity of concave portions of the contours. h) concave points (number of concave portions of the contour) i) symmetry j) fractal dimension

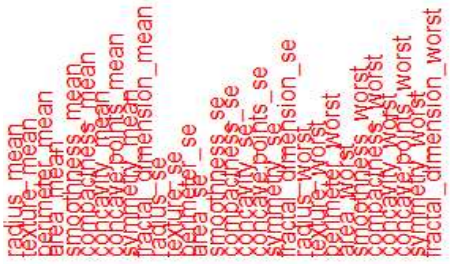
Preprocessing:

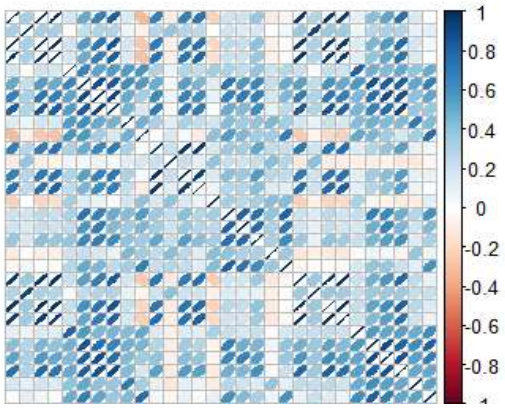
There is a column for ID numbers that we remove it. The range of data in different features is different. From 0 to 4254 which all are positive. So we do "Normalization" on data to make the range of all data from 0 to 1. Also we need to make the diagnosis column as factor. There is no missing data in data set and the data is clean. There is 62.7% of Benign and 37.3% of Malignant type.

Analysis:

We do the correlation for our features and re move the columns which has the correlation over 0.9. Since then we will have 20 features out of 30. (index of highly correlated features : 7 8 23 21 3 24 1 13 14 2)
Then we take a look at histogram of each fea ture which is separated to two type for "M" a nd "B". We see some means of two type histo grams are very close to each other. That mea ns they reveal no significant information for classification.so we remove the ones with closer means of 0.05. the number of features we use for training our machine learning models are 14.

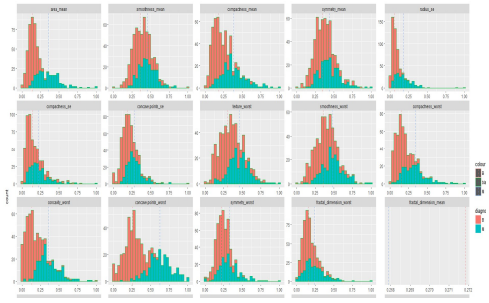






Classification:

The histogram of data we use for classification:



"diagnosis" "area_mean" "smoothness_mean" "compactness_mean" "symmetry_mean" "radius_se" "compactness_se" "concave.points_se" "texture_worst" "smoothness_worst" "compactness_worst" "concavity_worst" "concave.points_worst" "symmetry_worst" "fractal_dimension_worst"

First of all the data should be split up into two sets: train set and test set. Here there are 75% of total data (569) as train set (427) and the rest as test set (142).

Different classification algorithms are used: Knn, GLM (Logestic regression multiclass), RF (random forest), NNET (neural networks) and SVM (support vector machines).

The result of modeling each classification algorithms and testing it over test data is as follows:

	Accuracy	*AUC*
GLM	98.59%	0.989
RF	97.89%	0.989
NNET	98.59%	0.965
KNN	96.48%	0.955
SVM	91.55%	0.935

According to the accuracy GLM has the best result of 98.56% for accuracy and 0.989 for AUC.and SVM has the worst result of 91.55% for accuracy and 0.935 for AUC.

Confusion matrix:

Knn:

datatest$diagnosis	pred_knn B	pred_knn M	Row Total
B	107 1.000 0.982 0.754	0 0.000 0.000 0.000	107 0.754
M	2 0.057 0.018 0.014	33 0.943 1.000 0.232	35 0.246
Column Total	109 0.768	33 0.232	142

GLM:

datatest$diagnosis	pred_glmnet B	pred_glmnet M	Row Total
B	104 0.972 0.981 0.732	3 0.028 0.083 0.021	107 0.754
M	2 0.057 0.019 0.014	33 0.943 0.917 0.232	35 0.246
Column Total	106 0.746	36 0.254	142

RF:

datatest$diagnosis	pred_rf B	pred_rf M	Row Total
B	105 0.981 0.991 0.739	2 0.019 0.056 0.014	107 0.754
M	1 0.029 0.009 0.007	34 0.971 0.944 0.239	35 0.246
Column Total	106 0.746	36 0.254	142

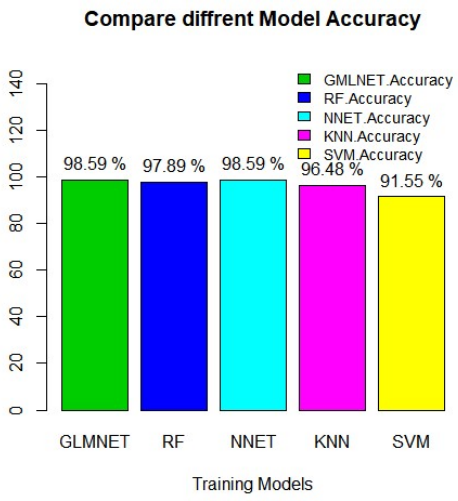
NNET:

datatest$diagnosis	pred_nnet B	pred_nnet M	Row Total
B	106 0.991 0.981 0.746	1 0.009 0.029 0.007	107 0.754
M	2 0.057 0.019 0.014	33 0.943 0.971 0.232	35 0.246
Column Total	108 0.761	34 0.239	142

SVM:

datatest$diagnosis	pred_svmRadial B	pred_svmRadial M	Row Total
B	96 0.897 0.980 0.676	11 0.103 0.250 0.077	107 0.754
M	2 0.057 0.020 0.014	33 0.943 0.750 0.232	35 0.246
Column Total	98 0.690	44 0.310	142

Accuracy bar chart:



ROC AUC: