

**BAN NBB - 130**

**PROJECT**

BAN 130 NBB



April 7, 2021

:

**SENECA COLLEGE, BLACKBOARD ULTRA**



**DATASET – BREAST CANCER**



Out of the provided options, our group has chosen the dataset of Breast Cancer for the project as per the initial proposal we have submitted in Project Part 1 submission. The dataset contains rich information with several variables that are derived based on the lumps detected from fluid samples taken from breast cancer patients. There are primarily ten features from each one of the cells in the sample, then it calculates the mean value, extreme value and standard error of each feature.

Name: Breast Cancer

Size: 256 kb after loading into SAS

Dimensions: 32 Variables and 569 Observations

File Type: CSV (Comma Separated Values)

The main purpose of our analysis has been to be able to predict the diagnosis of the breast cancer as either Malignant or Benign.

Visualize the frequency of the diagnosis

M= Malignant (indicates presence of cancer cells); B= Benign (indicates absence)

357 observations which account for 62.7% of all observations indicating the absence of cancer cells, 212 which account for 37.3% of all observations shows the presence of cancerous cell.

The percent is unusually large; the dataset does not represents in this case a typical medical analysis distribution. Typically, we will have a considerable large number of cases that represents negative vs. a small number of cases that represents positives (malignant) tumor.

## ANALYSIS:

* **Step – 1 Importing Breast Cancer file and create Data set.**
* **Step – 2 Deciding the Method of preparing Model on the basis of the**

**characteristics of Targeted variable.**

* **Step – 3 Logistic regression with appropriate selection method.**
* **Step – 4 Brief interpretation of significant independent variables.**

/\*CODE-01\*/

/\*We are importing the breast-cancer file in SAS and

creating SAS data set LEARN.BREAST\_CANCER \*/

PROC IMPORT DATAFILE='/folders/myfolders/Lectures/breast-cancer.xlsx'

DBMS=XLSX

OUT=LEARN.BREAST\_CANCER

REPLACE;

GETNAMES=YES;

RUN;

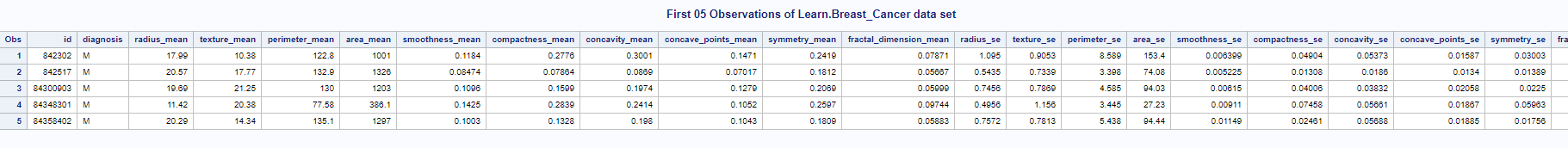
/\*CODE-02\*/

/\*Printing first 05 observations of the data set LEARN.BREAST\_CANCER\*/

TITLE "First 05 Observations of Learn.Breast\_Cancer data set";

PROC PRINT DATA=LEARN.BREAST\_CANCER (OBS=5);

RUN;

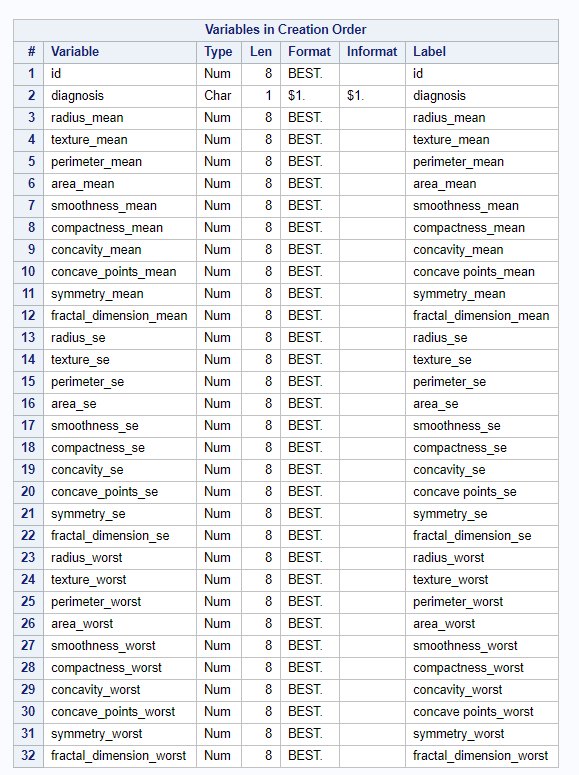
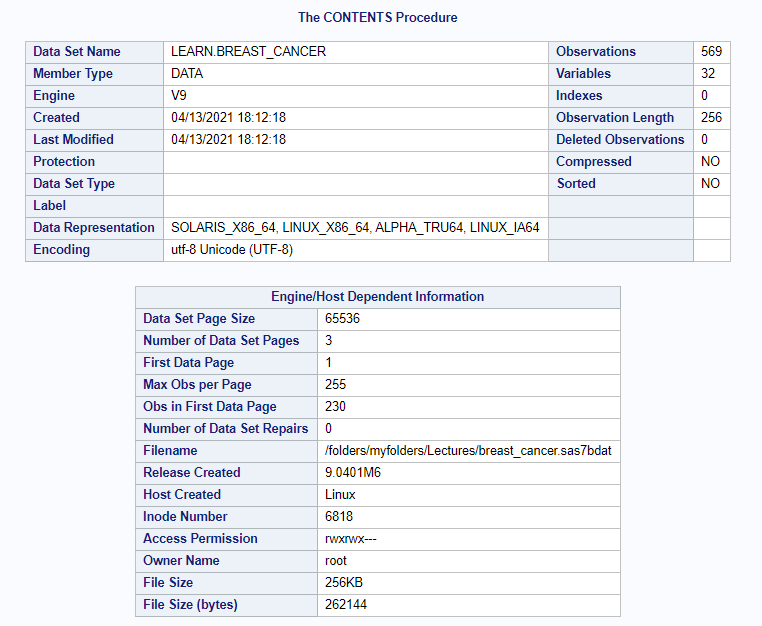


/\*CODE-03\*/

/\*We are creating variable list for Descriptive Statistics See table no.3\*/

PROC CONTENTS DATA=LEARN.BREAST\_CANCER VARNUM;

RUN;



**Deciding the Method of preparing Model on the basis of the characteristics of Targeted variable.**

As Targeted variable Diagnosis is binary M= Malignant and B= Benign. We will go for Logistic Regression.

We need to use selection method to decide most significant independent variables out of 30 independent variables for our model building.

/\*CODE-04 to 07\*/

/\*For doing Logistic regression we need to transfer the character variable to numeric binary class '0' and '1'\*/

DATA LEARN.BREAST\_CANCER1;

SET LEARN.BREAST\_CANCER;

IF diagnosis = 'M' THEN diagnosis = '0';

IF diagnosis = 'B' THEN diagnosis = '1';

RUN;

DATA LEARN.BREAST\_CANCER2;

SET LEARN.BREAST\_CANCER1;

diagnosis\_num = INPUT(diagnosis,3.);

DROP diagnosis;

RUN;

TITLE "FIRST 10 OBSERVATIONS FOR LEARN.BREAST\_CANCER2 DATA SET";

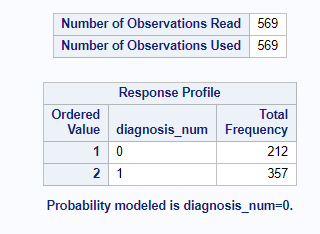
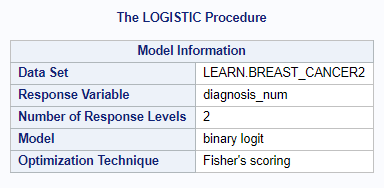
PROC PRINT DATA=LEARN.BREAST\_CANCER2 (OBS=10);

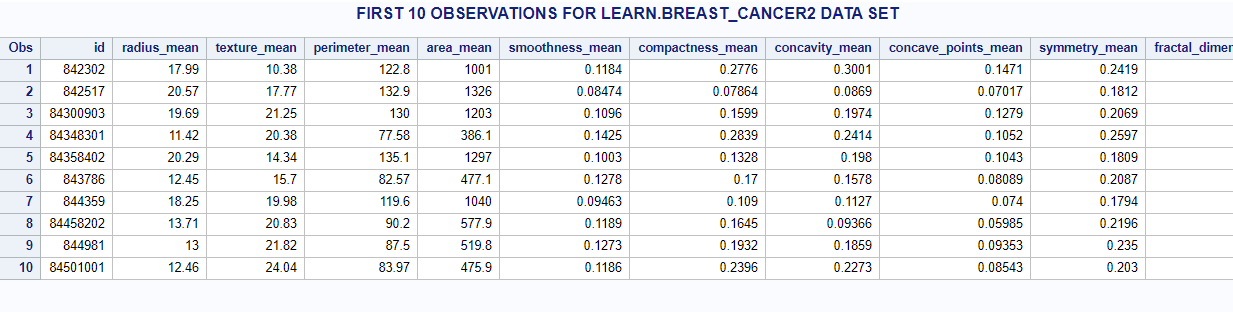
RUN;

TITLE "VARIABLE DESCRIPTION FOR LEARN.BREAST\_CANCER2 DATA SET";

PROC CONTENTS DATA=LEARN.BREAST\_CANCER2 VARNUM;

RUN;





/\*CODE-08 to 12\*/

/\*Now let's first build model with all variables and then select model variables,

by chekcking by all four selection methods FORWARD, BACKWARD, STEPWISE and SCORE\*/

PROC LOGISTIC DATA = BREAST\_CANCER2;

MODEL diagnosis\_num = radius\_mean texture\_mean perimeter\_mean area\_mean smoothness\_mean

compactness\_mean concavity\_mean concave\_points\_mean symmetry\_mean fractal\_dimension\_mean

radius\_se texture\_se perimeter\_se area\_se smoothness\_se compactness\_se concavity\_se

concave\_points\_se symmetry\_se fractal\_dimension\_se radius\_worst texture\_worst perimeter\_worst

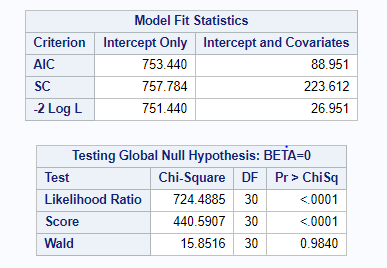
area\_worst smoothness\_worst compactness\_worst concavity\_worst concave\_points\_worst

symmetry\_worst fractal\_dimension\_worst;

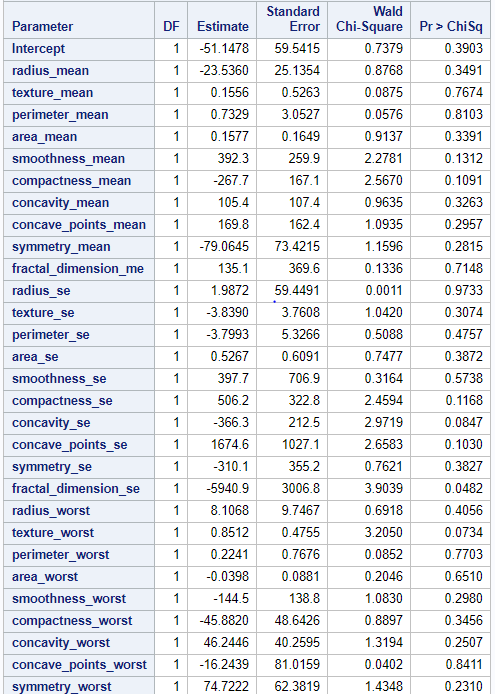
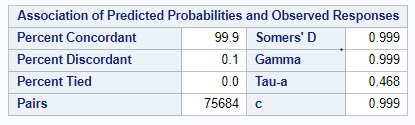
OUTPUT OUT = BREAST\_CANCER3;

RUN;

QUIT;



Analysis for Maximum Likelihood Estimates:

## FEATURE SELECTION:

PROC LOGISTIC DATA =BREAST\_CANCER2;

MODEL diagnosis\_num = radius\_mean texture\_mean perimeter\_mean area\_mean smoothness\_mean

compactness\_mean concavity\_mean concave\_points\_mean symmetry\_mean fractal\_dimension\_mean

radius\_se texture\_se perimeter\_se area\_se smoothness\_se compactness\_se concavity\_se

concave\_points\_se symmetry\_se fractal\_dimension\_se radius\_worst texture\_worst perimeter\_worst

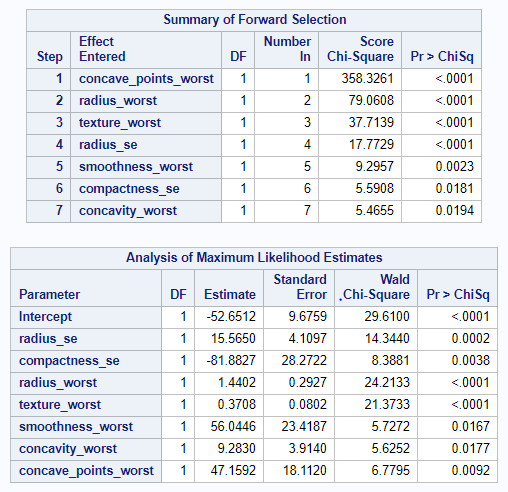
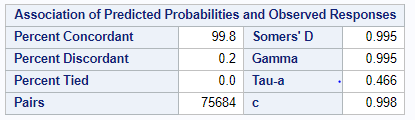
area\_worst smoothness\_worst compactness\_worst concavity\_worst concave\_points\_worst

symmetry\_worst fractal\_dimension\_worst/ SELECTION = FORWARD SLENTRY=.1 SLSTAY=.1;

OUTPUT OUT = BREAST\_CANCER3;

RUN;

QUIT;

PROC LOGISTIC DATA = LEARN.BREAST\_CANCER2;

MODEL diagnosis\_num = radius\_mean texture\_mean perimeter\_mean area\_mean smoothness\_mean

compactness\_mean concavity\_mean concave\_points\_mean symmetry\_mean fractal\_dimension\_mean

radius\_se texture\_se perimeter\_se area\_se smoothness\_se compactness\_se concavity\_se

concave\_points\_se symmetry\_se fractal\_dimension\_se radius\_worst texture\_worst perimeter\_worst

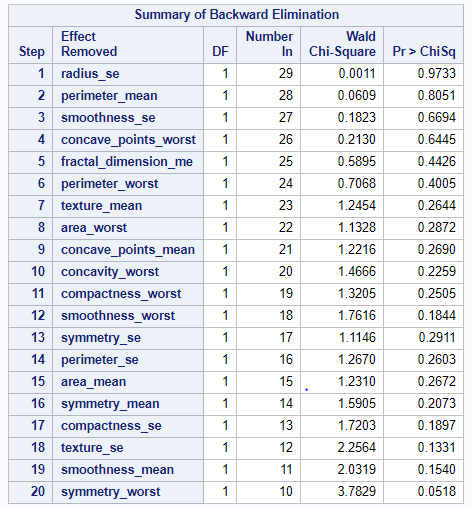
area\_worst smoothness\_worst compactness\_worst concavity\_worst concave\_points\_worst

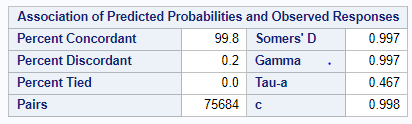
symmetry\_worst fractal\_dimension\_worst/ SELECTION = BACKWARD;

OUTPUT OUT = LEARN.BREAST\_CANCER3;

RUN;

QUIT;





PROC LOGISTIC DATA = LEARN.BREAST\_CANCER2;

MODEL diagnosis\_num = radius\_mean texture\_mean perimeter\_mean area\_mean smoothness\_mean

compactness\_mean concavity\_mean concave\_points\_mean symmetry\_mean fractal\_dimension\_mean

radius\_se texture\_se perimeter\_se area\_se smoothness\_se compactness\_se concavity\_se

concave\_points\_se symmetry\_se fractal\_dimension\_se radius\_worst texture\_worst perimeter\_worst

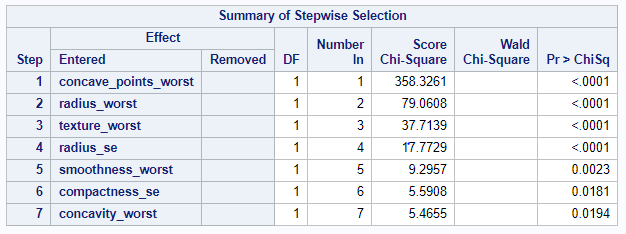
area\_worst smoothness\_worst compactness\_worst concavity\_worst concave\_points\_worst

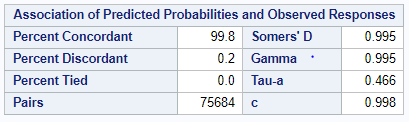
symmetry\_worst fractal\_dimension\_worst/ SELECTION = STEPWISE;

OUTPUT OUT = LEARN.BREAST\_CANCER3;

RUN;

QUIT;





PROC LOGISTIC DATA = LEARN.BREAST\_CANCER2;

MODEL diagnosis\_num = radius\_mean texture\_mean perimeter\_mean area\_mean smoothness\_mean

compactness\_mean concavity\_mean concave\_points\_mean symmetry\_mean fractal\_dimension\_mean

radius\_se texture\_se perimeter\_se area\_se smoothness\_se compactness\_se concavity\_se

concave\_points\_se symmetry\_se fractal\_dimension\_se radius\_worst texture\_worst perimeter\_worst

area\_worst smoothness\_worst compactness\_worst concavity\_worst concave\_points\_worst

symmetry\_worst fractal\_dimension\_worst/ SELECTION = SCORE;

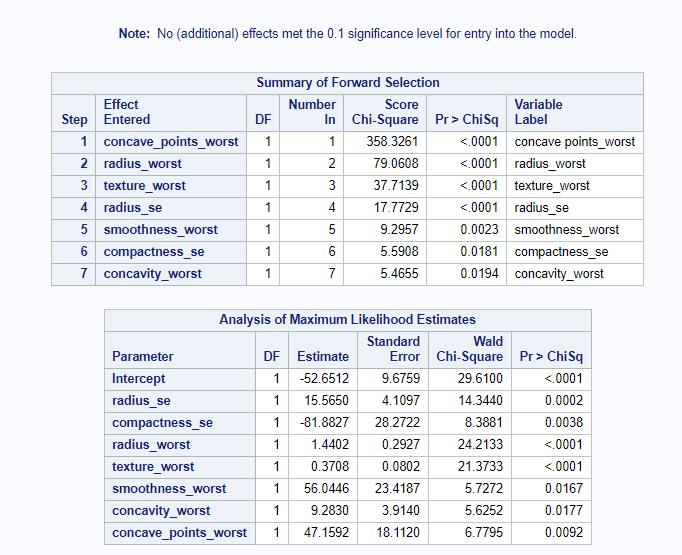
OUTPUT OUT = LEARN.BREAST\_CANCER3;

RUN;

QUIT;

We chose to build model with 7 independent significant variables according to Forward and Stepwise selection method.

* 7 most significant independent variables.
* All pValue <0.05 shows that they are significant but we saw the co-linearity between concavity\_worst & concave\_points\_worst also we saw colinearity between radius\_se & radius\_mean so, we dropped concavity\_worst and radius\_worst from our final model.



Our final model have 4 independent variables, we reduced 2 due to colinearity (concave\_points\_worst & radius\_se) and then forward selection method dropped another one (compactness\_se) as its pValue not remained significant in new combination. So, our final model have radius\_worst, texture\_worst, smoothness\_worst and concave\_points\_worst variables.

PROC LOGISTIC DATA = LEARN.BREAST\_CANCER2;

MODEL diagnosis\_num = compactness\_se texture\_worst radius\_se

smoothness\_worst concavity\_worst / SELECTION = FORWARD SLENTRY=.1 SLSTAY=.1;

OUTPUT OUT = LEARN.BREAST\_CANCER3;

RUN;

QUIT;

PROC LOGISTIC DATA = LEARN.BREAST\_CANCER2;

MODEL diagnosis\_num = compactness\_se radius\_worst texture\_worst

smoothness\_worst concave\_points\_worst/ SELECTION = FORWARD SLENTRY=.1 SLSTAY=.1;

OUTPUT OUT = LEARN.BREAST\_CANCER3;

RUN;

QUIT;

**/\*Final model with 4 independent variables\*/**

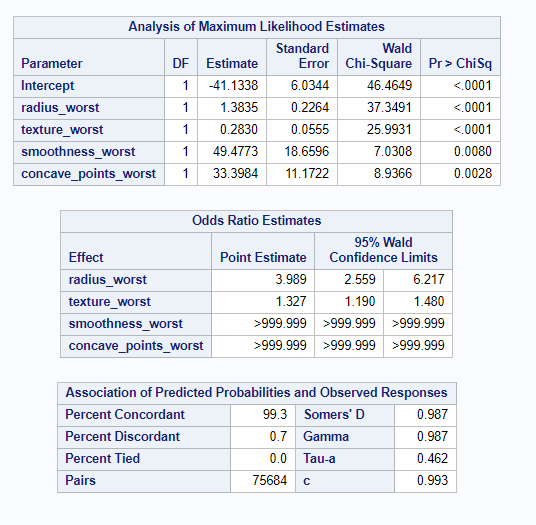
PROC LOGISTIC DATA = LEARN.BREAST\_CANCER2;

MODEL diagnosis\_num = radius\_worst texture\_worst smoothness\_worst

concave\_points\_worst/ SELECTION = FORWARD SLENTRY=.1 SLSTAY=.1;

OUTPUT OUT = LEARN.BREAST\_CANCER3;

RUN;



**FINAL MODEL VARIABLES:**

* Our final binary logit model have 4 signififant independent variables.
* radius\_worst – In tumor distance of edge from centre.
* Texture\_worst – Grey scale value used to find tumor location.
* Smoothness\_worst-Shows variation in radious range and
* Concave\_point worst – it is highest mean value of the concabe portions of the contour.

**CONCLUSION OF ANALYSIS:**

So, we can conclude our analysis that model with 4 independent variables (radius\_worst, texture\_worst, smoothness\_worst and concave\_points\_worst) is highly efficient and we should have combination of regular memmography, doctor’s diagnosis and prediction of our analytic model with this 3 filters we can catch the breast cancer at early stage and over come to it without any fatality.

## INTERPRETATION OF VARIABLES:

**Explanation of 4 independent variable affecting the diagnosis.**

* **Radius\_worst -** It has the highest value of the center for the estimated range. It is essential to know the distance between the center and the point because surgery depends on the size. There is no chance to do surgery with big tumors
* **Texture\_worst** - The highest mean value of standard deviation for gray-scale values is shown as texture worst. Gray-scale is commonly used to find the tumor location, and the standard deviation is essential to find the variation of the data and to explain how to spread out the numbers.
* **Smoothness\_Worst** – It is the largest mean value for regional variation in radius range. Very important statistic to evaluate the growth of the tumor.
* **Concave\_Point\_Worst** – It is the highest mean value of the concave portions of the contour, Contour depends on the volume of the breast and may be consider 2mm,3mm or 5mm deep to see the concavity in the tumor.

Values of above **4 independent variables** are accurate in diagnosing Malignant(M) or Benign(B) with about **99.3% c-score.**

## RECOMMENDATIONS:

1. Regular mammograms can help find **breast cancer** at an **early** stage, when treatment is most successful. **Women at high risk** should have yearly mammograms along with an MRI starting at age 30.
2. Risk factors to consider are – Age, Personal or Family history, Genetic factors, childbearing or menstrual history.
3. There is a need to raise awareness on the importance of early screenings by working with all possible social, health and media organizations
4. Integrate technology, biology and risk models to improve existing infrastructure to better detect, treat and reduce possibility of recurrence.