# **Project Report- Group 3**

**Breast Cancer Analysis**

**Presented to:**

Prof . Yazan Roumani

School of Business Administration

Oakland University

**Team Members:**

Ashwini Nelagondanahalli Virupakshappa

Aswini Sivakumar

Divya Ramesh

Prachi Dalvi

# **Breast Cancer Analysis**

**Introduction**

Breast cancer is a disease in which the healthy cells of the tissue in the breast are invaded and mutated, which further grow in large numbers to form a malignant tumor. It can most likely occur at any age. However, research has shown that 80% of breast cancer occurs in women older than 50 years. It is one of the most common types of cancer in women, and it can also occur in men, although it is much less common. Also, it is the second most leading cause of cancer deaths in women.

**Data**

The dataset we are using was downloaded from Kaggle site available at:

https://www.kaggle.com/datasets/reihanenamdari/breast-cancer

This dataset of breast cancer patients was obtained from the 2017 November update of the SEER Program of the NCI, which provides information on population-based cancer statistics. The dataset involved female patients with infiltrating duct and lobular carcinoma breast cancer diagnosed in 2006-2010. Patients with unknown tumor size, examined regional LNs, positive regional LNs, and patients whose survival months were less than 1 month were excluded; thus, 4024 patients were ultimately included. The detailed explanation of the variables is as below:

* **Age** - Age of the patient
* **Race** - Race of the patient
* **Marital Status** - Married or Not Married
* **T - Stage** - Refers to the size and/or extent of the main tumor. The higher the number after the T, the larger the tumor or the more it has grown into nearby tissues.
* **N - Stage** - Refers to the number and location of lymph nodes that contain cancer. The higher the number after the N, the more lymph nodes that contain cancer.
* **6th Stage** - The first 2 or 3 letters (II or III) are for the stage of cancer 2nd or 3rd and the last alphabet describes how badly/severely it has spread.
* **Differentiate** - The features that explain the arrangements of the cells in relation to each other.
  + Well-differentiated carcinomas have relatively normal-looking cells that do not appear to be growing rapidly and are arranged in small tubules for ductal cancer and cords for lobular cancer. These cancers tend to grow and spread slowly and have a better prognosis (outlook).
  + Poorly differentiated carcinomas lack normal features, tend to grow and spread faster, and have a worse prognosis.
  + Moderately differentiated carcinomas have features and a prognosis in between these two.
* **Grade** - The higher the grade the more severe the cancer.
* **A -stage** -
  + Regional: The cancer has spread outside the breast to nearby structures or lymph nodes.
  + Distant: The cancer has spread to distant parts of the body such as the lungs, liver, or bones.
* **Tumor Size** - Size of the tumor which further decides the T-stage variable.
* **Estrogen Status** - If breast cancer cells have estrogen receptors, the cancer is called ER-positive breast cancer.
* **Progesterone Status** - If breast cancer cells have Progesterone receptors, the cancer is called PR-positive breast cancer.
* **Regional Node examined**- Records the exact number of regional nodes examined (from around the tumor)
* **Regional Node positive** - Records the exact number of regional nodes examined by the pathologist and found to contain metastases.
* **Survival month** - Number of months the patient has survived since detection.
* **Status** - Whether the patient is Dead or Alive.

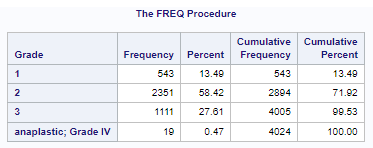
**Problem**

Based on the various variables provided, our goal is to understand the relation between these variables and to analyze and find the best model that will help to predict the status (Alive or Dead) of the new patient. We will also be analyzing which variables are important for both the target variables and how they affect the results in different models.

**Data Cleaning/Validation**

The data available on the Kaggle site was pretty consistent without null and missing fields. However, the following changes were required and made to work using SAS.

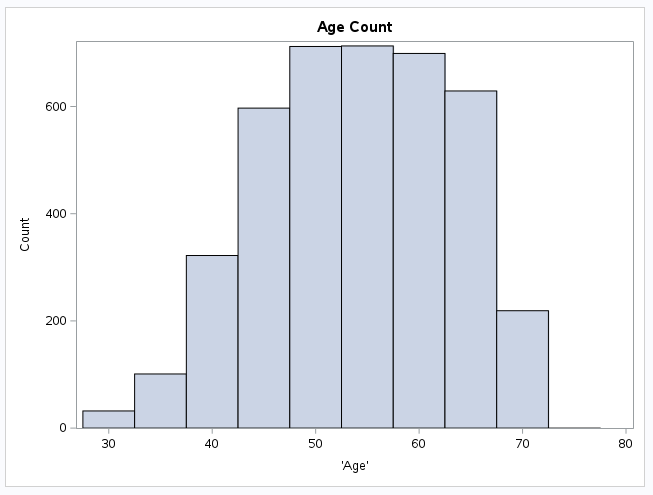
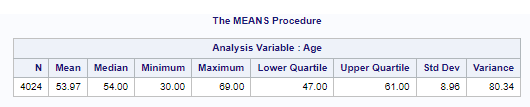
* The variable names in the initial sheet had spaces in them, so to make the variable names as valid SAS names options validvarname=v7; was used.
* For Grade variables the values were 1,2,3 and anaplastic; Grade IV



* + This was giving error when importing the data as 1,2,3 was numeric and ‘anaplastic; Grade IV’ was string. Using guessingrows=max; helped to import the data by determining the appropriate data type and length of variables.
  + Once the data was loaded in the system, anaplastic; Grade IV value was changed to 4 in the same variable.

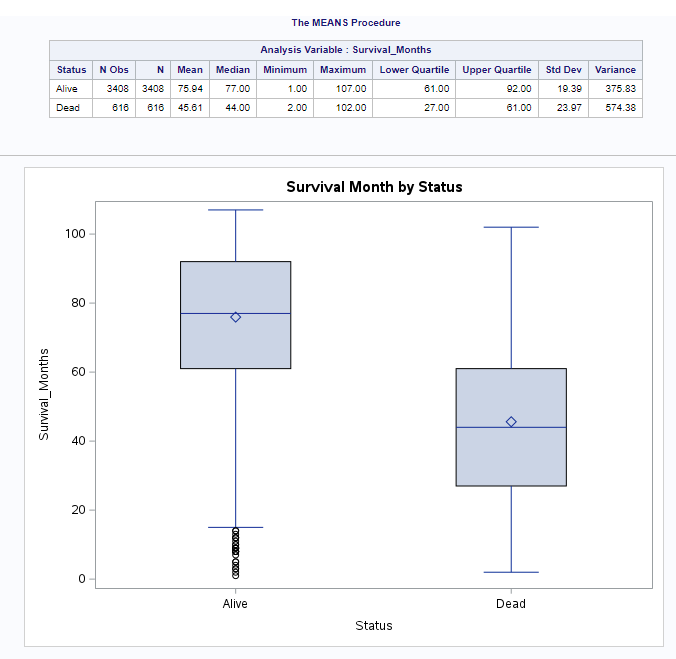
**Exploratory Data Analysis**

**Age**



The range of patients’ age is from 30-69. The Average age of the patient is ~54 and the first quartile and third quartile values are from 47-61, that is also where we see the peak starting and lowering in the histogram.

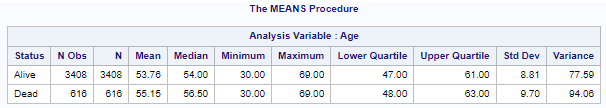
**Survival\_Months by Status**



From the bar plot and analysis variable table we can see that minimum and maximum survival months of alive patients is 1 month and 107 months since detection, we see a lot of lower-level outlier in the alive patients which shows the actual range from around 18 months – 107 months. This might be because the patient was recently detected. The minimum and maximum survival month of the dead patients is 2 months and 102 months respectively. The mean and median of the alive patient is 75.94 months and 77 months and that of the dead patient is 45.61 and 44 months since detection. The boxplot of the survival\_months for alive patients is left skewed as the median is closer to the max value and also because of too many outliers at the lower end.

**Status**

Status By Age



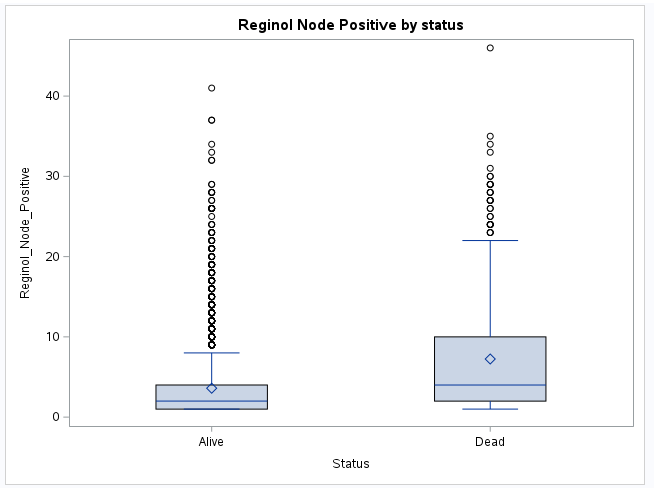
Chart, box and whisker chart

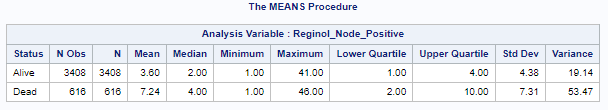
Description automatically generated

From the boxplot and analysis variable table we can see that the minimum age and maximum age of the alive patient as well as the dead patient is 30 and 69 respectively. However, the mean (53.76) and median (54) age of alive patients is slightly lower than the mean (55.15) and median (56.50) of dead patients. This might imply that the chances of older patients surviving are less.

-Status by Regional\_Node\_Positive

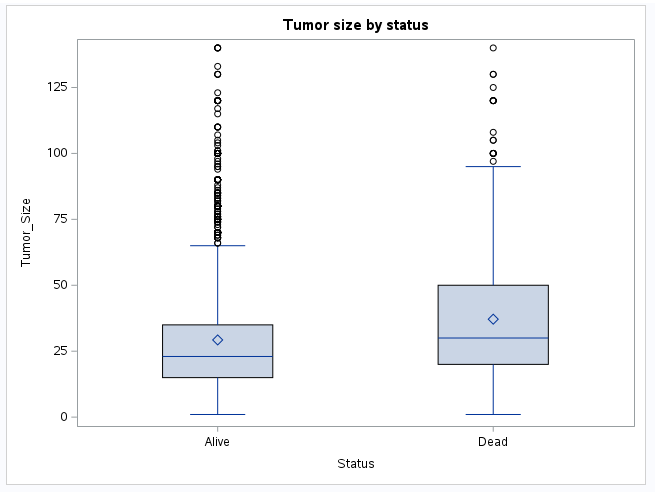
From the box plot and the Analysis Variable table, we can see that the average number of regional nodes that are found to contain metastases by the pathologist is less for alive patients. Similarly, the median number of positive regional nodes is slightly lower in alive patients than in dead patients. The IQR of patients who are alive is smaller compared to that of dead patients. The number of positive regional nodes above 9 in alive patients is considered as outliers and above 22 is considered outliers in dead patients. This could be because there can be other factors like age, stage of the cancer and the spread of the cancer affecting the status of the patient.

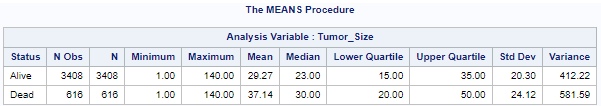




Status by Tumor\_size

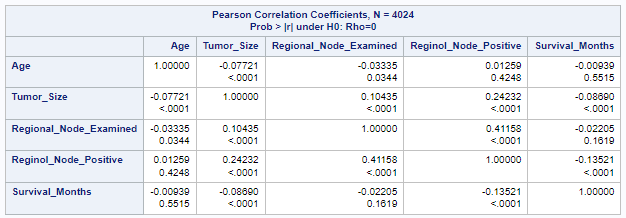
From the box plot and analysis variable table, we can see that the average and median size of the tumor is slightly higher in dead patients than the alive patients. The IQR of the tumor size of alive patients is less than that of the dead patients. The range of the tumor size of both alive and dead patients are 1 - 140. The standard deviation of tumor size among the dead patients is slightly higher than that of the living patients. The living patients with tumor size above 70(approximately) are considered outliers which can be because the patient got diagnosed only at a later stage recently.

****

****

**Correlation matrix**

From the correlation matrix for identifying the correlation between the numerical variables like age, tumor\_size, regional\_node\_examined, reginol\_node\_positive and survival months, we could see that there is no strong relationship between any of the variables.



**Model Selection**

For this project our goal is to predict the status (Alive or Dead) of the new patient. We will also be analyzing which variables are important for the target variable and how they affect the results in different models. For analyzing and predicting the Status (Alive or Dead) of the patient, as our target variable is categorical, we have the following options to try and determine a good fitting model out of them:

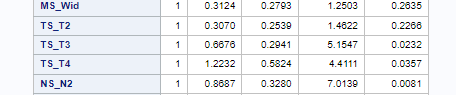
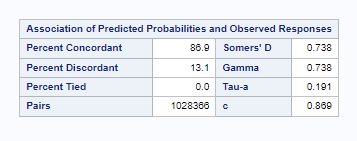
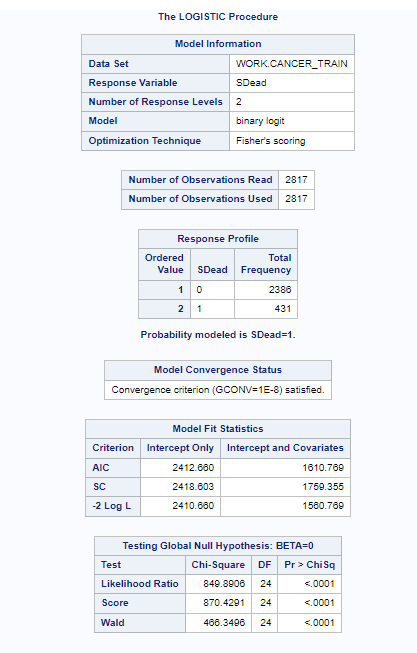
* Logistic Regression
* CART
* Neural Network
* Discriminant Analysis

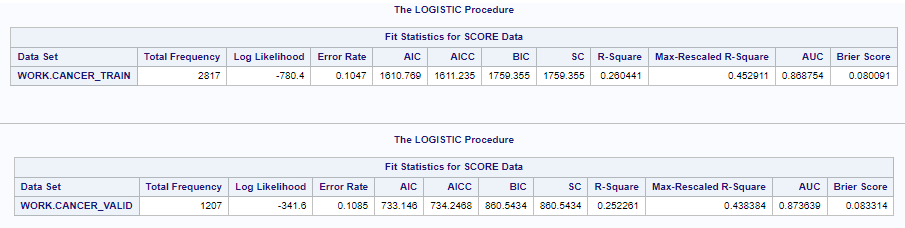
Tumor Size and T\_stage variables are correlated which we are aware of from domain understanding, so we will build all our models with both of them individually in all our models and compare if one has a better fit than another.

**Logistic Regression for Status:**

For the Logistic regression for Status variable, we built the Logistic regression model with all variables with T\_Stage and with Tumor\_size individually.

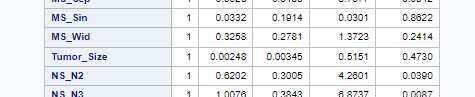
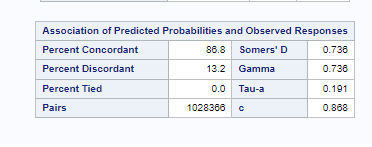
With T\_Stage:

****

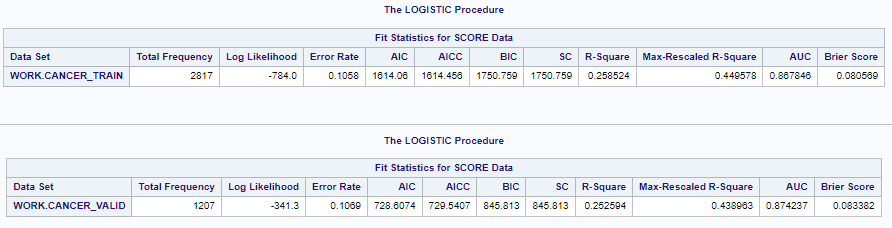
The p-value for the likelihood ratio test is less than alpha so the model is statistically significant. The value of AUC is 86.9% which is quite good. From the Analysis of Maximum LikelihoodEstimates we found that most of the variables were not statistically significant, but for T\_Stage - all variables 2 out of 3 dummy variables are statistically significant as p value is less than 0.05. 

Based on the error metrics of the test and validation set, we see that AUC for validation (87.36%) is better than that of training set (86.87%). Also, the AIC(734), BIC(860) is lower for the validation set than the training set AIC(1610) BIC(1759). There is also not much difference in the error rate of both the sets `0.10, which shows that there is no sign of overfitting using logistic regression with T\_stage.

Tumor\_size

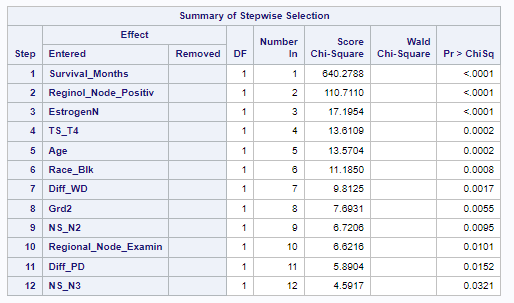
For the Logistic Regression model, with Tumor\_size, the p-value for the likelihood ratio test is less than alpha so the model is statistically significant. The value of AUC is 86.8% which is quite good. And for individual variables again most of the variables have p-value more than alpha including Tumor\_size which shows that the Tumor\_size is not a statistically significant model to be considered for analysis of Status using logistic regression model.



Based on the error metrics of the test and validation set, we see that AUC for validation (87.42%) is better than that of training set (86.78%). Also, the AIC(728), BIC(845) are lower for the validation set than the training set AIC(1614) BIC(1750). There is also not much difference in the error rate of both the sets `0.10, which shows that there is no sign of overfitting using logistic regression with T\_stage.

Next, we built the model with forward, backward and stepwise selection, to find the important variables using T\_stage as Tumor\_size is not statistically significant.

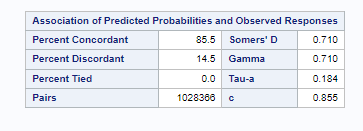
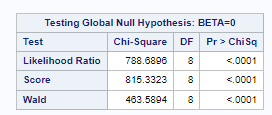
Using stepwise logistic regression, below is the list of important variables:

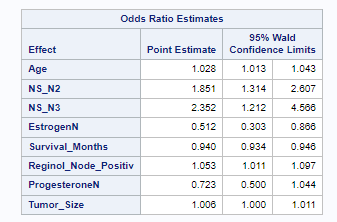
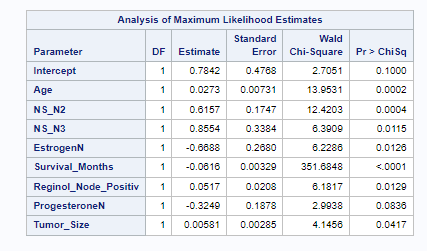


We will be using the below combination of variables to build the Logistic regression and analyze it:

* Age, Race, T\_Stage, N\_Stage, Differentiate, Grade, Estrogen, Survival\_Months, Regional\_Node, Examined Reginol\_Node\_Positive
* Age, Race, N\_Stage\_Differentiate, Grade, Estrogen, Survival\_Months, Regional\_Node, Examined Reginol\_Node\_Positive
* Age, Race, N\_Stage, Differentiate, Estrogen, Survival\_Months, Regional\_Node, Examined Reginol\_Node\_Positive
* Age, Race, N\_Stage\_Differentiate, Grade, Estrogen, Survival\_Months, Regional\_Node, Examined Reginol\_Node\_Positive

Based on above combination of important variables, we got the best AUC (slightly lower but with minimum important variables) with Age, Race, N\_Stage\_Differentiate, Estrogen, Survival\_Months, Regional\_Node, Examined Reginol\_Node\_Positive important variables.

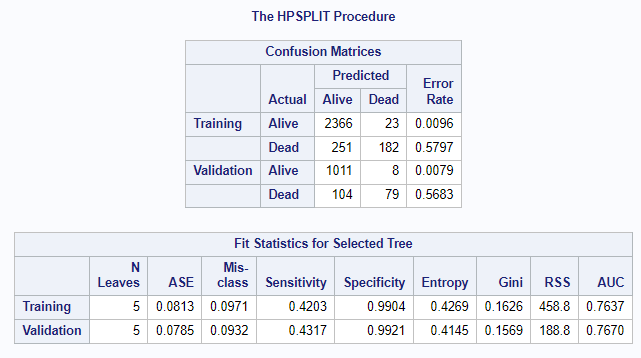
 

When interpreting the variable Age, for every year the patient becomes older, the odds of the status being dead increases by 2.8%. For EstrogenN, if the estrogen status is negative the odds of the patients being dead is reduced by 50.6% compared to that of the patient with estrogen status being positive. For Regional\_Node\_Examined, with each increase in the count of regional node examined the changes of the patient being dead is decreased by 2.3%

**CART**

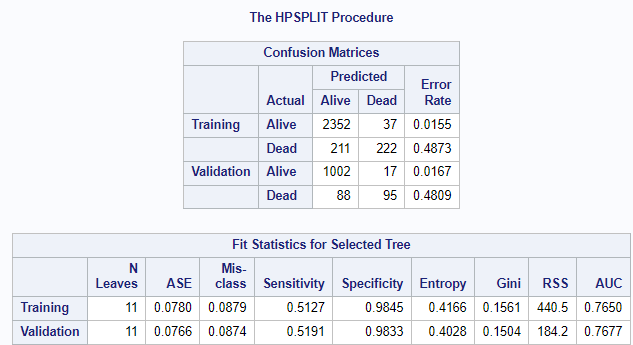
For analyzing status(dead or alive) using the CART model, we used Gini and Entropy measures using all variables and with both Tumor\_size and T\_stage separately.

1. Entropy with T Stage:

****

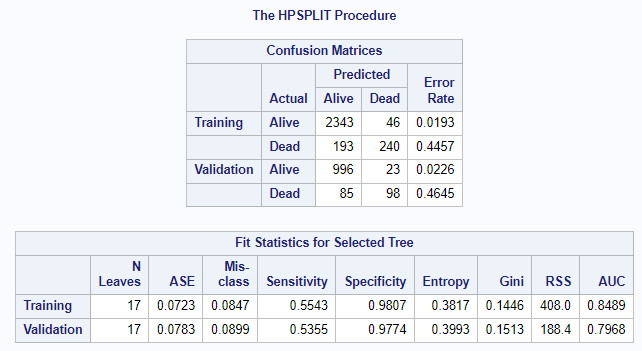
In the Entropy with T stage CART model, the number of leaves after pruning is 5. Evaluating the model performance, we can see from the confusion matrix table that 2366 out of 2389(training) and 1011 out of 1019(validation) observations, the status is classified correctly as Alive which indicates that the specificity is 99.04% and 99.21% on training and validation set respectively, which is very good. Only 182 out of 433(training) and 79 out of 183(validation) observations, the status is classified correctly as Dead which indicates that the sensitivity is 42.03% and 43.17% on training and validation set respectively, which is not good. AUC is 76.37% and 76.70% for training and validation set respectively, which is pretty good and there is no sign of overfitting as there is not much difference between the error metrics of training and validation set.

1. Entropy with Tumor Size:



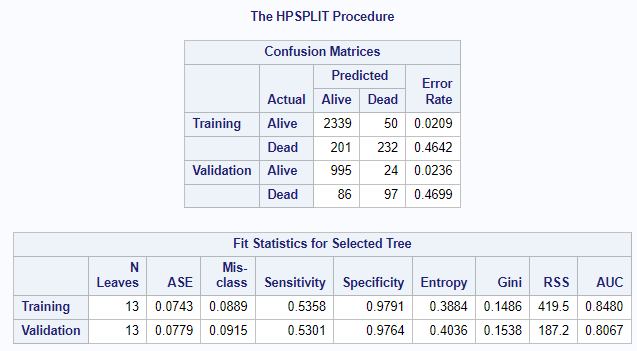
In the Entropy with Tumor size CART model, the number of leaves after pruning is 11 which is approximately double the previous model. We can see from the confusion matrix table that 2352 out of 2389(training) and 1002 out of 1019(validation) observations, the status is classified correctly as Alive which indicates that the specificity is 98.45% and 98.33% on training and validation set respectively, which is very good and is very slightly lower than previous model, but we can say it is very closer. 222 out of 433(training) and 95 out of 183(validation) observations, the status is classified correctly as Dead which indicates that the sensitivity is 51.27% and 51.91% on training and validation set respectively, which is not so good, but we can see there is an improvement compared to the previous model. AUC is 76.50% and 76.77% for training and validation sets respectively, which is pretty good and there is no sign of overfitting as there is not much difference between the error metrics of training and validation set. Compared with the previous model, AUC is very close to the previous model.

1. Gini with T Stage:



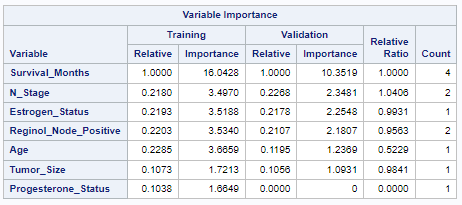
In the Gini with T Stage CART model, the number of leaves after pruning is 17 which is more than the previous 2 models. We can see from the confusion matrix table that 2343 out of 2389(training) and 996 out of 1019(validation) observations, the status is classified correctly as Alive which indicates that the specificity is 98.07% and 97.74% on training and validation set respectively, which is very good and is very slightly lower than previous models, but we can say it is very closer. 240 out of 433(training) and 98 out of 183(validation) observations, the status is classified correctly as Dead which indicates that the sensitivity is 55.43% and 53.55% on training and validation set respectively, which is not so good, but we can again see there is a improvement compared to the previous model. AUC is 84.89% and 79.68% for training and validation set respectively, which is pretty good and there is no sign of overfitting. Compared with the previous 2 models, AUC has improved in both the sets. So, Gini with T Stage CART model(model 3) is better compared to model 1 and 2 as there is improvement in sensitivity and AUC even though the number of leaves is 17 after pruning.

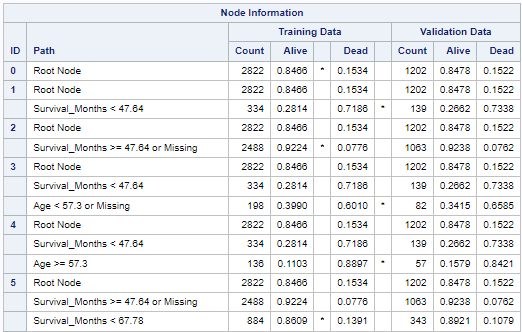
1. Gini with Tumor Size:



In the Gini with Tumor size CART model, the number of leaves after pruning is 13 which is more than the model 1 and 2 but lower than the model 3. We can see from the confusion matrix table that 2339 out of 2389(training) and 995 out of 1019(validation) observations, the status is classified correctly as Alive which indicates that the specificity is 97.91% and 97.64% on training and validation set respectively, which is very good, and it is very closer to model 3. 232 out of 433(training) and 97 out of 183(validation) observations, the status is classified correctly as Dead which indicates that the sensitivity is 55.58% and 53.01% on training and validation set respectively, which is not so good, and values are very closer to previous model( Model 3). AUC is 84.80% and 80.67% for training and validation set respectively, which is pretty good and there is no sign of overfitting. Compared with the model 3 which was better among the first three models, AUC has slightly improved in the validation set and the number of leaves after pruning is lesser, so we can say Gini with Tumor size CART model has performed well compared to other 3 models.

Below is the list of important variables from the CART mode(Gini with Tumor size):





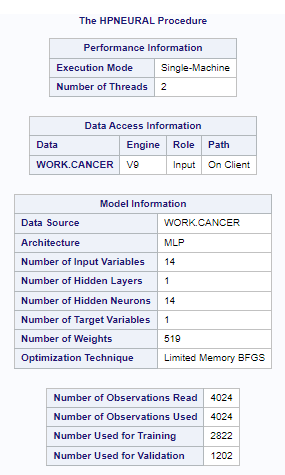
From the node information, Node 5 is interpreted as if Survival\_Months >= 47.64 or Missing and Survival\_Months < 67.78, then there is an 86.09% chance that the patient is classified as Alive.

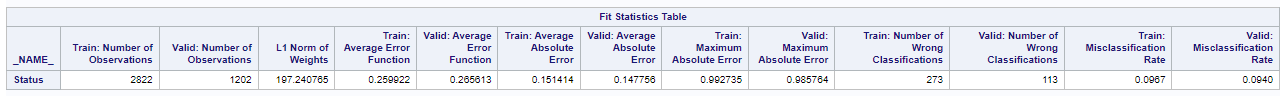
Node 4 is interpreted as if Survival\_Months < 47.64 and Age >= 57.3, then there is an 88.97% chance that the patient is classified as Dead.

**Neural Networks:**

As we did with the Logistic regression and CART model, we built the NN model with T\_Stage and Tumor\_size.

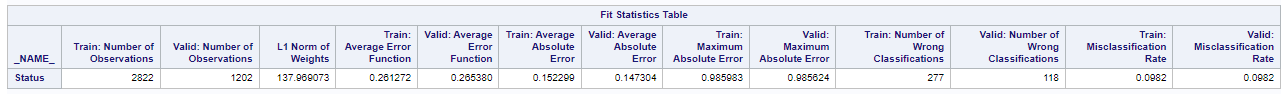
With T\_Stage:





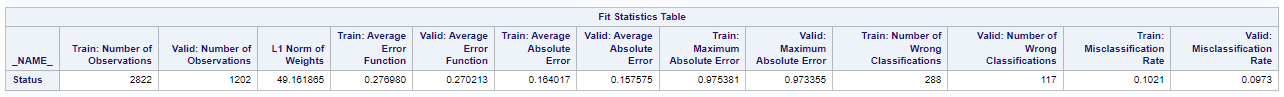
For NN with T\_stage, the hidden layer 1 and with 14 neurons, from the Fit Statistics table, we can see that the average absolute error for the validation set (0.147) is lower than the training set (0.151) and so is the misclassification rate validation – 9.4%, training – 9.67%. As there is not much difference in the error metrics of training and validation set there is no sign of overfitting using this model.

With Tumor\_size:



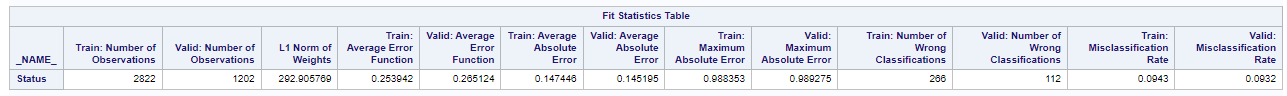
In this model, we only swapped T\_stage with Tumor\_size and build the NN model with 1 hidden layer and 14 neurons, and with tumor\_size model as well their average absolute error for validation (0.147) is lower than the training error (0.152) and the misclassification rate is same for both the training and validation set 9.82%.

As we see by comparing both the models above for NN there is not much difference in the error metrics of the model with T\_Stage and Tumor\_size. Next, we built the NN model with the important variables from the CART model and next tried to optimize the NN model to reduce error metrics but still trying to keep the model simple.

Error metrics for NN with important variables from CART model, hidden layer 1 and 6 neurons:

By using only, the important variables from the CART model the error measure has slightly increased on both the training and validation set.

Optimized NN Model:



For optimization we used all variables with tumor\_size with 1 hidden layer and 10 neurons. This has not only reduced the average absolute error but also the misclassification rate.

**Discriminant Analysis:**

For Discriminant analysis, as we have imbalance data, we tried the priors with proportional option first with the T\_Stage and Tumor\_Size variable individually. However, we also tried the equal priors approach to see how well the equal priors fits the data.

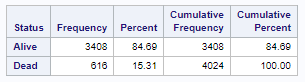
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Priors** |  | **Sensitivity** | **Specificity** | **Accuracy** |
| **With T\_Stage** | **Proportional** | Training | 44.32 | 97.11 | 89.03 |
|  | Validation | 48.65 | 95.99 | 88.73 |
| **Equal** | Training | 73.09 | 81.81 | 80.47 |
|  | Validation | 81.08 | 78.47 | 78.87 |
| **With Tumor\_size** | **Proportional** | Training | 45.24 | 97.11 | 89.18 |
|  | Validation | 51.35 | 95.79 | 88.98 |
| **Equal** | Training | 72.39 | 81.94 | 80.47 |
|  | Validation | 80.54 | 78.47 | 78.79 |

Based on the consolidated metrics of all the models analyzed for discriminant analysis above, we can see that equal priors (with t\_stage and tumor\_size) give us a better sensitivity and specificity than the proportional priors, however, the accuracy is better with the proportional priors model. And also, as our data is imbalanced, we will be considering only proportional priors for our analysis. Using proportional priors there is a very negligible difference between the t\_stage and tumor\_size model and individually in t\_stage and tumor\_size, there is not much difference in the specificity and accuracy of the training and validation set and the sensitivity has improved in the validation set, which shows there is no sign of overfitting in any of the proportional priors models of discriminant analysis.

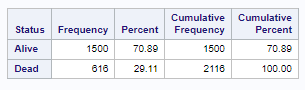
Analysis Using under sampling of majority group:

As our dataset has imbalance data, we used the under sampling of the majority group to analyze if we get better results.

Entire data:



Under sampling of majority group data:



|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **With T\_Stage - With all Variables and entire data** | | | | | | | | | | |
| **Model** |  | **Number of Leaves** | **AUC** | **ASE** | **AIC** | **BIC** | **Error Rate / Misclassification error** | **Sensitivity** | **Specificity** | **Accuracy** |
| Logistic Regression |  |  |  |  |  |  |  |  |  |  |
| Training |  | 86.87 |  | 1610 | 1759 | 0.1047 |  |  |  |
| Validation |  | 87.36 |  | 731 | 860 | 0.1085 |  |  |  |
| CART Model (Gini) |  |  |  |  |  |  |  |  |  |  |
| Training | 17 | 84.89 | 0.0723 |  |  | 0.0847 | 55.43 | 98.07 | 91.53 |
| Validation | 17 | 79.63 | 0.0783 |  |  | 0.899 | 53.55 | 97.74 | 91.01 |
| Neural Network |  |  |  |  |  |  |  |  |  |  |
| Training |  |  | 0.1514 |  |  | 0.0967 |  |  |  |
| Validation |  |  | 0.1477 |  |  | 0.094 |  |  |  |

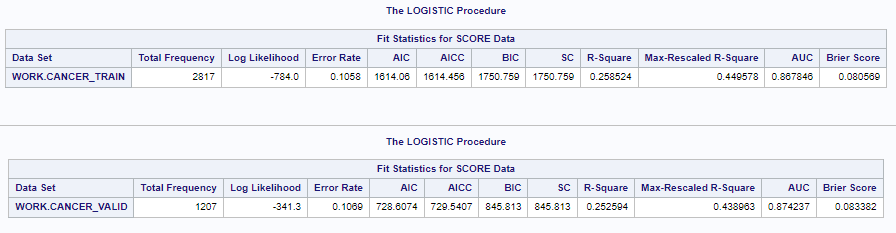
|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **With T\_Stage - With all Variables and under sampling of majority group** | | | | | | | | | | | |
| **Model** |  | **Number of Leaves** | **AUC** | **ASE** | **AIC** | **BIC** | **Error Rate / Misclassification error** | **Sensitivity** | **Specificity** | **Accuracy** |
| Logistic Regression |  |  |  |  |  |  |  |  |  |  |
| Training |  | 86.36 |  | 1803 | 1952 | 0.1132 |  |  |  |
| Validation |  | 87.49 |  | 830 | 958 | 0.1251 |  |  |  |
| CART Model (Gini) |  |  |  |  |  |  |  |  |  |  |
| Training | 8 | 82.66 | 0.1158 |  |  | 0.1448 | 58.39 | 96.76 | 85.52 |
| Validation | 8 | 83.17 | 0.1122 |  |  | 0.1395 | 60.77 | 96.22 | 86.05 |
| Neural Network |  |  |  |  |  |  |  |  |  |  |
| Training |  |  | 0.223 |  |  | 0.1582 |  |  |  |
| Validation |  |  | 0.2196 |  |  | 0.1426 |  |  |  |

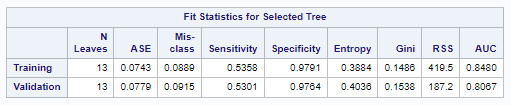
After building the logistic regression, CART and Neural Network model using under sampling of majority group data and comparing the error metrics (above table), we see that there is no positive improvement using the under sampling of majority group data and we are also almost compromising 40-45% of our data. So, we will be using our entire dataset of the final model comparisons and conclusions.

**Model Comparison and Conclusion:**

As there is not much difference between the Tumor\_size and T\_stage model error metrics of all the models, we are using only tumor\_size models for comparing and finding the best model amongst all.

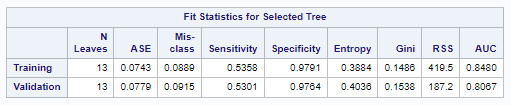
Logistic Regression and CART Model (Gini):

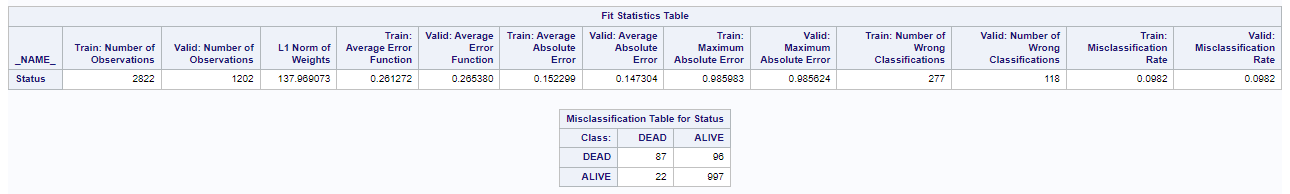




Based on the error metrics of both Logistics (training and validation) and CART’s (training and validation), we can see that the AUC for Logistics model is better for both training and validation than CART’s model, however, the error rate of logistic model high (training – 0.105, validation – 0.106) than that of CART model (training - 0.088, validation – 0.0915). So, out of the two, the CART model is better.

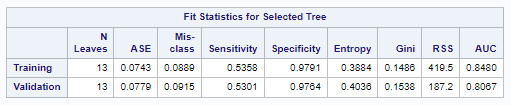
CART Model and Neural Network:

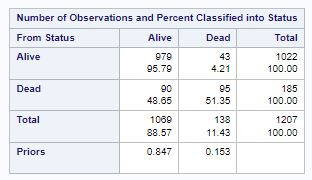
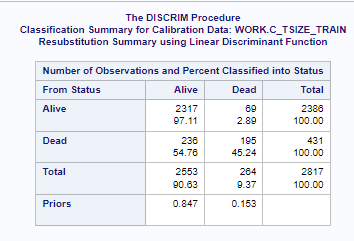




From the error metrics of CART and NN model, the misclassification rate of CART model (training – 0.088, validation – 0.091) is slightly lower than NN (training – 0.098, validation – 0.098) and the ASE of the CART model (training – 0.074, validation – 0.077) is very much low than NN average absolute error (training – 0.152, validation – 0.147). So, out of these two as well, the CART model is better.

CART Model and Discriminant Analysis (Proportional priors):





From the error metrics of CART and confusion matrix of Discriminant analysis, with CART model the sensitivity is training – 53.58%, validation – 53.01%, specificity is training – 97.91%, validation – 97.64% and misclassification rate is training – 0.089, validation – 0.091. Whereas using Discriminant analysis, the sensitivity is training – 45.24, validation – 51.35, specificity is training – 97.11 validation – 95.79 and misclassification rate is training (69+236/2817) – 0.108, validation (43+90/1207) – 0.110. So, the Sensitivity and specificity is better for CART and the misclassification rate is also low for CART. So, in between these two as well, the CART model is the best.

Based on all the above analysis, the CART model using Gini and Tumor\_Size is the best performing model to be used for classifying the patient's Status.

**Improvements:**

The analysis and model prediction we did for Status variables was completely based on the variables and data that we got from the Kaggle. However, based on the domain knowledge and research we think that there are other underlying factors and conditions that need to be considered to help do further analysis and improve the model prediction. If we get more information about the patient's medical history for any other underlines health conditions and also the treatment plan that the patient is undergoing (Surgical extraction, Chemotherapy, radiation, or Mastectomy) as these factors play a very critical and important role in a breast cancer patients survival months and status analysis.

**/\*SAS Code\*/**

options validvarname=v7; /\*change the Variable names of the tables to Valid SAS Name\*/

proc import out=Cancer\_Org datafile="/home/u61150141/sasuser.v94/Data\_Science/Project/Breast\_Cancer.csv"

dbms=csv replace;

guessingrows=max; /\*Column Grade has values 1,2,3 and anaplastic; Grade IV which was giving error in import guessingrows=max will scan entire column before defining the datatype of the column \*/

run;

proc freq data=Cancer\_Org;

table grade;

run;

/\*Data Cleaning\*/

data Cancer\_Org;

set Cancer\_Org;

if Grade='anaplastic; Grade IV' then Grade=4;

run;

/\*Checking count after updating the column \*/

proc freq data=Cancer\_Org;

table grade;

run;

/\*Creating Dummy Variable\*/

data Cancer;

set Cancer\_Org;

/\*Dummy Variable for Race with Race White as base value\*/

if Race='Black' then Race\_Blk=1; else Race\_Blk=0;

if Race='Other' then Race\_Oth=1; else Race\_Oth=0;

/\*Dummy Variable for Marital\_Status with Marital\_Status Married as base value\*/

if Marital\_Status='Divorced' then MS\_DRC=1; else MS\_DRC=0;

if Marital\_Status='Separated' then MS\_Sep=1; else MS\_Sep=0;

if Marital\_Status='Single' then MS\_Sin=1; else MS\_Sin=0;

if Marital\_Status='Widowed' then MS\_Wid=1; else MS\_Wid=0;

/\*Dummy Variable for T\_Stage with T\_Stage T1 as base value\*/

if T\_Stage='T2' then TS\_T2=1; else TS\_T2=0;

if T\_Stage='T3' then TS\_T3=1; else TS\_T3=0;

if T\_Stage='T4' then TS\_T4=1; else TS\_T4=0;

/\*Dummy Variable for N\_Stage with N\_Stage N1 as base value\*/

if N\_Stage='N2' then NS\_N2=1; else NS\_N2=0;

if N\_Stage='N3' then NS\_N3=1; else NS\_N3=0;

/\*Dummy Variable for \_6th\_Stage with \_6th\_Stage IIA as base value\*/

if \_6th\_Stage='IIB' then SS\_IIB=1; else SS\_IIB=0;

if \_6th\_Stage='IIIA' then SS\_IIIA=1; else SS\_IIIA=0;

if \_6th\_Stage='IIIB' then SS\_IIIB=1; else SS\_IIIB=0;

if \_6th\_Stage='IIIC' then SS\_IIIC=1; else SS\_IIIC=0;

/\*Dummy Variable for Differentiate with Differentiate 'Moderately differentiated' as base value\*/

if Differentiate='Poorly differentiated' then Diff\_PD=1; else Diff\_PD=0;

if Differentiate='Undifferentiated' then Diff\_UND=1; else Diff\_UND=0;

if Differentiate='Well differentiated' then Diff\_WD=1; else Diff\_WD=0;

/\*Dummy Variable for Grade with Grade '1' as base value\*/

if Grade=2 then Grd2=1; else Grd2=0;

if Grade=3 then Grd3=1; else Grd3=0;

if Grade=4 then Grd4=1; else Grd4=0;

/\*Dummy Variable for A\_Stage with A\_Stage Regional as base value\*/

if A\_Stage='Distant' then ADistant=1; else ADistant=0;

/\*Dummy Variable for Estrogen\_Status with Estrogen\_Status Positive as base value\*/

if Estrogen\_Status='Positive' then EstrogenN=1; else EstrogenN=0;

/\*Dummy Variable for Progesterone\_Status with Progesterone\_Status Positive as base value\*/

if Progesterone\_Status='Positive' then ProgesteroneN=1; else ProgesteroneN=0;

/\*Dummy Variable for Status with Status Alive as base value\*/

if Status='Dead' then SDead=1; else SDead=0;

run;

proc freq data=Cancer;

table Race Race\_Blk Race\_Oth Marital\_Status MS\_DRC MS\_Sep MS\_Sin MS\_Wid T\_Stage TS\_T2 TS\_T3 TS\_T4 N\_Stage NS\_N2 NS\_N3

\_6th\_Stage SS\_IIB SS\_IIIA SS\_IIIB SS\_IIIC Differentiate Diff\_PD Diff\_UND Diff\_WD Grade Grd2 Grd3 Grd4 A\_Stage ADistant

Estrogen\_Status EstrogenN Progesterone\_Status ProgesteroneN Status SDead;

run;

/\*Splitting the data in 70:30\*/

Proc surveyselect data=Cancer samprate=.7 method=SRS out=Cancer\_Split outall seed=12345; /\*Samprate - size of the trainset\*/

/\*Proc surveyselect is to split the data samprate is for sample rate, dividing the data in 70:30

, method SRS is for simple random sample seed all using the same numbers to split the data \*/

run;

data Cancer\_train Cancer\_Valid;

set Cancer\_Split;

if selected=1 then output Cancer\_train; else output Cancer\_Valid;

run;

/\*Checking for the distribution of Biased variable\*/

proc freq data=Cancer;

table Status;

run;

proc freq data=Cancer\_train;

table Status;

run;

proc freq data=Cancer\_Valid;

table Status;

run;

/\*EDA\*/

proc freq data=Cancer;

table Race Marital\_Status T\_Stage N\_Stage \_6th\_Stage Differentiate A\_Stage Estrogen\_Status Progesterone\_Status;

run;

proc means data=Cancer n mean median min max q1 q3 stddev var maxdec=2;

var Age; /\*class is to add categorical division to the statistics\*/

run;

/\*Age Distribution\*/

Proc sgplot data=Cancer;

histogram Age/nbins=10 scale=count;

title "Age Distribution";

xaxis label="'Age'";

run;

/\*Age by Status\*/

proc means data=Cancer n mean median min max q1 q3 stddev var maxdec=2;

var Age; class Status; /\* class is to add categorical division to the statistics\*/

run;

Proc sgplot data=Cancer;

vbox Age/category=Status;

Title"Age by Status";

run;

/\*Age by Sruvival\_Months\*/

proc means data=Cancer n mean median min max q1 q3 stddev var maxdec=2;

var Survival\_Months; class Status;

run;

Proc sgplot data=Cancer; /\*sgplot is to create scatter plot\*/

scatter x=Age y=Survival\_Months;

title "Scatterplot of Age by Survival Months";

xaxis label="'Age'";

yaxis label="Survival Month";

run;

/\*Survival\_Months by Status\*/

Proc sgplot data=Cancer;

vbox Survival\_Months/category=Status;

Title"Survival Month by Status";

run;

proc means data=Cancer n mean median min max q1 q3 stddev var maxdec=2;

var Survival\_Months; class Estrogen\_Status;

run;

/\*Survival\_Months by Estrogen\_Status\*/

Proc sgplot data=Cancer;

vbox Survival\_Months/category=Estrogen\_Status;

Title"Survival Month by Estrogen\_Status";

run;

/\*Age by T\_Stage\*/

Proc sgplot data=Cancer;

vbox Age/category=T\_Stage;

title "Age by T-Stage";

xaxis label="'Age'";

yaxis label="T-Stage";

run;

/\*Regional\_Node\_Positive by Status\*/

proc sgplot data=Cancer;

vbox reginol\_node\_positive/category=Status;

Title"Reginol Node Positive by status";

run;

/\*Tumor\_Size by Status\*/

proc sgplot data=Cancer;

vbox tumor\_size/category=Status;

Title"Tumor\_Size by status";

run;

/\* Correlation matrix \*/

proc corr data=Cancer;

var age tumor\_size regional\_node\_examined reginol\_node\_positive survival\_months;

run;

/\*Bar Chart --Used for Categorical variables to show percent\*/

Proc sgplot data=Cancer;

vbar Estrogen\_Status/stat=pct group=Status GROUPDISPLAY = CLUSTER;/\*stat is to change frequency to %\*/

yaxis values=(0 to 1 by 0.1) label="Percent";/\*changing scaling for y axis\*/

title "Estrogen\_Status by Status";

run;

Proc sgplot data=Cancer;

vbar Progesterone\_Status/stat=pct group=Status GROUPDISPLAY = CLUSTER;

yaxis values=(0 to 1 by 0.1) label="Percent";

title "Progesterone\_Status by Status";

run;

/\*------------------------------------------------------Status---------------------------------------------------------------\*/

/\*----------------------------Logistic Regression------------------------------------\*/

/\*Logistic Regression with all variables and T\_Stage\*/

proc logistic data=Cancer\_train outmodel=Cancer\_model1; /\*save all the parameter estimates for future analysis \*/

Model SDead(event='1')=Age Race\_Blk Race\_Oth MS\_DRC MS\_Sep MS\_Sin MS\_Wid TS\_T2 TS\_T3 TS\_T4 NS\_N2 NS\_N3

SS\_IIB SS\_IIIA SS\_IIIB SS\_IIIC Diff\_PD Diff\_UND Diff\_WD Grd2 Grd3 Grd4 ADistant

EstrogenN ProgesteroneN Survival\_Months Regional\_Node\_Examined Reginol\_Node\_Positive;

run;

proc logistic inmodel=Cancer\_model1;

score data=Cancer\_train fitstat; /\*Using the statistics from train to evaluate the validation set\*/

run;

proc logistic inmodel=Cancer\_model1;

score data=Cancer\_Valid fitstat; /\*Using the statistics from train to evaluate the validation set\*/

run;

/\*Logistic Regression with all variables and Tumor\_Size\*/

proc logistic data=Cancer\_train outmodel=Cancer\_model2; /\*save all the parameter estimates for future analysis \*/

Model SDead(event='1')=Age Race\_Blk Race\_Oth MS\_DRC MS\_Sep MS\_Sin MS\_Wid Tumor\_Size NS\_N2 NS\_N3

SS\_IIB SS\_IIIA SS\_IIIB SS\_IIIC Diff\_PD Diff\_UND Diff\_WD Grd2 Grd3 Grd4 ADistant

EstrogenN ProgesteroneN Survival\_Months Regional\_Node\_Examined Reginol\_Node\_Positive;

run;

/\*Tumor\_Size is not statistically Significant\*/

proc logistic inmodel=Cancer\_model2;

score data=Cancer\_train fitstat; /\*Using the statistics from train to evaluate the validation set\*/

run;

proc logistic inmodel=Cancer\_model2;

score data=Cancer\_Valid fitstat; /\*Using the statistics from train to evaluate the validation set\*/

run;

/\*Logistic Regression with all variables and T\_Stage with forward selection\*/

proc logistic data=Cancer\_train outmodel=Cancer\_model1; /\*save all the parameter estimates for future analysis \*/

Model SDead(event='1')=Age Race\_Blk Race\_Oth MS\_DRC MS\_Sep MS\_Sin MS\_Wid TS\_T2 TS\_T3 TS\_T4 NS\_N2 NS\_N3

SS\_IIB SS\_IIIA SS\_IIIB SS\_IIIC Diff\_PD Diff\_UND Diff\_WD Grd2 Grd3 Grd4 ADistant

EstrogenN ProgesteroneN Survival\_Months Regional\_Node\_Examined Reginol\_Node\_Positive/selection=forward;

run;

/\*Logistic Regression with all variables and T\_Stage with backward selection\*/

proc logistic data=Cancer\_train outmodel=Cancer\_model1;

Model SDead(event='1')=Age Race\_Blk Race\_Oth MS\_DRC MS\_Sep MS\_Sin MS\_Wid TS\_T2 TS\_T3 TS\_T4 NS\_N2 NS\_N3

SS\_IIB SS\_IIIA SS\_IIIB SS\_IIIC Diff\_PD Diff\_UND Diff\_WD Grd2 Grd3 Grd4 ADistant

EstrogenN ProgesteroneN Survival\_Months Regional\_Node\_Examined Reginol\_Node\_Positive/selection=backward;

run;

/\*Logistic Regression with all variables and T\_Stage with Stepwise selection\*/

proc logistic data=Cancer\_train outmodel=Cancer\_model1;

Model SDead(event='1')=Age Race\_Blk Race\_Oth MS\_DRC MS\_Sep MS\_Sin MS\_Wid TS\_T2 TS\_T3 TS\_T4 NS\_N2 NS\_N3

SS\_IIB SS\_IIIA SS\_IIIB SS\_IIIC Diff\_PD Diff\_UND Diff\_WD Grd2 Grd3 Grd4 ADistant

EstrogenN ProgesteroneN Survival\_Months Regional\_Node\_Examined Reginol\_Node\_Positive/selection=stepwise;

run;

/\*Tryin once again with Tumor\_Size to recheck\*/

/\*Logistic Regression with all variables and Tumor\_Size with Stepwise selection\*/

proc logistic data=Cancer\_train outmodel=Cancer\_model1;

Model SDead(event='0')=Age Race\_Blk Race\_Oth MS\_DRC MS\_Sep MS\_Sin MS\_Wid Tumor\_Size NS\_N2 NS\_N3

SS\_IIB SS\_IIIA SS\_IIIB SS\_IIIC Diff\_PD Diff\_UND Diff\_WD Grd2 Grd3 Grd4 ADistant

EstrogenN ProgesteroneN Survival\_Months Regional\_Node\_Examined Reginol\_Node\_Positive/selection=stepwise;

run;

/\*Removed from stepwise selection\*/

proc logistic inmodel=Cancer\_model1;

score data=Cancer\_train fitstat; /\*Using the statistics from train to evaluate the validation set\*/

run;

proc logistic inmodel=Cancer\_model1;

score data=Cancer\_Valid fitstat; /\*Using the statistics from train to evaluate the validation set\*/

run;

/\*Logistic Regression with important variables from the Stepwise selection\*/

proc logistic data=Cancer\_train outmodel=Cancer\_model1;

Model SDead(event='1')=Age Race\_Blk Race\_Oth TS\_T2 TS\_T3 TS\_T4 NS\_N2 NS\_N3

Diff\_PD Diff\_UND Diff\_WD Grd2 Grd3 Grd4 EstrogenN Survival\_Months Regional\_Node\_Examined Reginol\_Node\_Positive;

run;

/\*Logistic Regression with important variables from the Stepwise selection but removing t\_stage\*/

proc logistic data=Cancer\_train outmodel=Cancer\_model1;

Model SDead(event='1')=Age Race\_Blk Race\_Oth TS\_T2 TS\_T3 TS\_T4 NS\_N2 NS\_N3

Diff\_PD Diff\_UND Diff\_WD Grd2 Grd3 Grd4 EstrogenN Survival\_Months Regional\_Node\_Examined Reginol\_Node\_Positive;

run;

/\*No change in AUC by removing T\_Stage so excluding T\_Stage \*/

/\*Logistic Regression with important variables from the Stepwise selection removing t\_stage and N\_Stage\*/

proc logistic data=Cancer\_train outmodel=Cancer\_model1;

Model SDead(event='1')=Age Race\_Blk Race\_Oth

Diff\_PD Diff\_UND Diff\_WD Grd2 Grd3 Grd4 EstrogenN Survival\_Months Regional\_Node\_Examined Reginol\_Node\_Positive;

run;

/\*Removing N\_stage reduced AUC and reverting to N\_Stage and removing grade\*/

proc logistic data=Cancer\_train outmodel=Cancer\_model1;

Model SDead(event='1')=Age Race\_Blk Race\_Oth NS\_N2 NS\_N3

Diff\_PD Diff\_UND Diff\_WD EstrogenN Survival\_Months Regional\_Node\_Examined Reginol\_Node\_Positive;

run;

/\*Not much difference by removing Grade will test by removing Race \*/

proc logistic data=Cancer\_train outmodel=Cancer\_model1;

Model SDead(event='1')=Age NS\_N2 NS\_N3

Diff\_PD Diff\_UND Diff\_WD EstrogenN Survival\_Months Regional\_Node\_Examined Reginol\_Node\_Positive;

run;

/\*AUC reduced a bit so reverting Race and removing Differentiate\*/

proc logistic data=Cancer\_train outmodel=Cancer\_model1;

Model SDead(event='1')=Age Race\_Blk Race\_Oth NS\_N2 NS\_N3

EstrogenN Survival\_Months Regional\_Node\_Examined Reginol\_Node\_Positive;

run;

/\*AUC reduced further so reverting Differentiate\*/

/\*Final model with best AUC and Parsimony principle\*/

proc logistic data=Cancer\_train outmodel=Cancer\_model1;

Model SDead(event='1')=Age Race\_Blk Race\_Oth NS\_N2 NS\_N3

Diff\_PD Diff\_UND Diff\_WD EstrogenN Survival\_Months Regional\_Node\_Examined Reginol\_Node\_Positive;

run;

/\*Important variables from CART Model\*/

proc logistic data=Cancer\_train outmodel=Cancer\_model1;

Model SDead(event='1')=Age NS\_N2 NS\_N3

EstrogenN Survival\_Months Reginol\_Node\_Positive ProgesteroneN tumor\_size;

run;

/\*----------------------------CART------------------------------------\*/

/\*With T\_Stage\*/

proc hpsplit data=cancer nodes=detail;

class Race Marital\_Status T\_Stage N\_Stage \_6th\_Stage Differentiate A\_Stage Grade Estrogen\_Status Progesterone\_Status status;

model status(event="Dead")=age Race Marital\_Status T\_Stage N\_Stage \_6th\_Stage Differentiate A\_Stage Grade Estrogen\_Status Progesterone\_Status

survival\_months regional\_node\_examined reginol\_node\_positive;

partition fraction(validate=0.3 seed=12345);

grow entropy;

prune cc;

run;

proc hpsplit data=cancer nodes=detail;

class Race Marital\_Status T\_Stage N\_Stage \_6th\_Stage Differentiate A\_Stage Grade Estrogen\_Status Progesterone\_Status status;

model status(event="Dead")=age Race Marital\_Status T\_Stage N\_Stage \_6th\_Stage Differentiate A\_Stage Grade Estrogen\_Status Progesterone\_Status

survival\_months regional\_node\_examined reginol\_node\_positive;

partition fraction(validate=0.3 seed=12345);

grow Gini;

prune cc;

run;

/\* with Tumor\_Size\*/

proc hpsplit data=cancer nodes=detail;

class Race Marital\_Status N\_Stage \_6th\_Stage Differentiate A\_Stage Grade Estrogen\_Status Progesterone\_Status status;

model status(event="Dead")=age Race Marital\_Status N\_Stage \_6th\_Stage Differentiate A\_Stage Grade Estrogen\_Status Progesterone\_Status

survival\_months tumor\_size regional\_node\_examined reginol\_node\_positive;

partition fraction(validate=0.3 seed=12345);

grow entropy;

prune cc;

run;

proc hpsplit data=cancer nodes=detail;

class Race Marital\_Status N\_Stage \_6th\_Stage Differentiate A\_Stage Grade Estrogen\_Status Progesterone\_Status status;

model status(event="Dead")=age Race Marital\_Status N\_Stage \_6th\_Stage Differentiate A\_Stage Grade Estrogen\_Status Progesterone\_Status

survival\_months tumor\_size regional\_node\_examined reginol\_node\_positive;

partition fraction(validate=0.3 seed=12345);

grow Gini;

prune cc;

run;

/\*----------------------------Neural Network------------------------------------\*/

/\*----Neural Network for Status Variable-----\*/

/\*---with T\_stage----\*/

proc hpneural data=cancer;

partition fraction(validate=0.3 seed=12345);

target status/level=nom;

input age Regional\_Node\_Examined Reginol\_Node\_positive survival\_months /level=int;

input race Marital\_Status N\_stage t\_stage \_6th\_stage differentiate grade A\_Stage Estrogen\_Status Progesterone\_Status /level=nom;

hidden 14;

train maxiter=1000 numtries=5;

run;

/\*---with tumor\_size----\*/

proc hpneural data=cancer;

partition fraction(validate=0.3 seed=12345);

target status/level=nom;

input age tumor\_size Regional\_Node\_Examined Reginol\_Node\_positive survival\_months /level=int;

input race Marital\_Status N\_stage \_6th\_stage differentiate grade A\_stage estrogen\_status Progesterone\_Status /level=nom;

hidden 14;

train maxiter=1000 numtries=5;

run;

/\*---- NN model for imp variables obtained from CART Model----\*/

proc hpneural data=cancer;

partition fraction(validate=0.3 seed=12345);

target status/level=nom;

input age survival\_months Reginol\_Node\_positive tumor\_size /level=int;

input N\_stage Estrogen\_Status /level=nom; /\*---Progesterone\_Status---ignoring as importance is 0 and relative importance is zero--\*/

hidden 6;

train maxiter=1000 numtries=5;

run;

proc hpneural data=cancer;

partition fraction(validate=0.3 seed=12345);

target status/level=nom;

input age survival\_months Reginol\_Node\_positive tumor\_size /level=int;

input N\_stage Estrogen\_Status /level=nom; /\*---Progesterone\_Status---ignoring as importance eis 0 and relative importance is zero--\*/

hidden 4;

train maxiter=1000 numtries=5;

run;

proc hpneural data=cancer;

partition fraction(validate=0.3 seed=12345);

target status/level=nom;

input age tumor\_size Regional\_Node\_Examined Reginol\_Node\_positive survival\_months /level=int;

input race Marital\_Status N\_stage \_6th\_stage differentiate grade A\_Stage Estrogen\_Status Progesterone\_Status /level=nom;

hidden 10;

train maxiter=1000 numtries=5;

run;

/\*-------------------------------------Discriminant Analysis------------------------------------\*/

/\*Splitting the original Dataset \*/

Proc surveyselect data=Cancer\_Org samprate=.7 method=SRS out=Cancer\_Org\_Split outall seed=12345;

run;

data C\_Tstage\_train C\_Tstage\_Valid;

set Cancer\_Org\_Split;

if selected=1 then output C\_Tstage\_train; else output C\_Tstage\_Valid;

drop Tumor\_Size Selected;

run;

data C\_Tsize\_train C\_Tsize\_Valid;

set Cancer\_Org\_Split;

if selected=1 then output C\_Tsize\_train; else output C\_Tsize\_Valid;

drop T\_Stage Selected;

run;

/\*Discriminant Analysis with T\_Stage with priors proportional\*/

Proc Discrim data=C\_Tstage\_train testdata=C\_Tstage\_Valid method=normal;

class status;

priors proportional; /\*looking at outcome variable to check for even distribution\*/

run;

/\*Discriminant Analysis with T\_Stage with priors equal\*/

Proc Discrim data=C\_Tstage\_train testdata=C\_Tstage\_Valid method=normal;

class status;

priors equal; /\*looking at outcome variable to check for even distribution\*/

run;

/\*Discriminant Analysis with Tumor\_Size with priors proportional\*/

Proc Discrim data=C\_Tsize\_train testdata=C\_Tsize\_Valid method=normal;

class status;

priors proportional; /\*looking at outcome variable to check for even distribution\*/

run;

/\*Discriminant Analysis with Tumor\_Size with priors equal\*/

Proc Discrim data=C\_Tsize\_train testdata=C\_Tsize\_Valid method=normal;

class status;

priors equal; /\*looking at outcome variable to check for even distribution\*/

run;

/\*-------------------------------------------------Under Sampling Majority Group-------------------------------------------------\*/

/\*Creating new test and train dataset by under sampling the majority group\*/

Proc sort data=Cancer;

key Status/ascending;

run;

Proc Surveyselect data=Cancer out=Cancer\_Bal Seed=12345 method=SRS sampsize=(1500 616);

Strata Status;

run;

Proc surveyselect data=Cancer\_Bal samprate=.8 method=SRS out=Cancer\_Bal\_Split outall seed=12345; /\*Samprate - size of the trainset\*/

run;

proc freq data=Cancer\_Bal;

table Status;

run;

data Cancer\_Bal\_Train Cancer\_Bal\_Valid;

set Cancer\_Bal\_Split;

if selected=1 then output Cancer\_Bal\_Train; else output Cancer\_Bal\_Valid;

drop selected selectionprob samplingweight;

run;

/\*Logistic Regression with all variables and T\_Stage\*/

proc logistic data=Cancer\_Bal\_Train outmodel=Cancer\_model1; /\*save all the parameter estimates for future analysis \*/

Model SDead(event='1')=Age Race\_Blk Race\_Oth MS\_DRC MS\_Sep MS\_Sin MS\_Wid TS\_T2 TS\_T3 TS\_T4 NS\_N2 NS\_N3

SS\_IIB SS\_IIIA SS\_IIIB SS\_IIIC Diff\_PD Diff\_UND Diff\_WD Grd2 Grd3 Grd4 ADistant

EstrogenN ProgesteroneN Survival\_Months Regional\_Node\_Examined Reginol\_Node\_Positive;

run;

/\*Logistic Regression with all variables and T\_Stage\*/

proc logistic data=Cancer\_Bal\_Train outmodel=Cancer\_model1; /\*save all the parameter estimates for future analysis \*/

Model SDead(event='1')=Age Race\_Blk Race\_Oth MS\_DRC MS\_Sep MS\_Sin MS\_Wid Tumor\_Size NS\_N2 NS\_N3

SS\_IIB SS\_IIIA SS\_IIIB SS\_IIIC Diff\_PD Diff\_UND Diff\_WD Grd2 Grd3 Grd4 ADistant

EstrogenN ProgesteroneN Survival\_Months Regional\_Node\_Examined Reginol\_Node\_Positive;

run;

/\*CART Model T\_Stage with Gini\*/

proc hpsplit data=Cancer\_Bal nodes=detail;

class Race Marital\_Status T\_Stage N\_Stage \_6th\_Stage Differentiate A\_Stage Grade Estrogen\_Status Progesterone\_Status status;

model status(event="Dead")=age Race Marital\_Status T\_Stage N\_Stage \_6th\_Stage Differentiate A\_Stage Grade Estrogen\_Status Progesterone\_Status

survival\_months regional\_node\_examined reginol\_node\_positive;

partition fraction(validate=0.3 seed=12345);

grow Gini;

prune cc;

run;

/\*CART Model Tumor\_Size with Gini\*/

proc hpsplit data=Cancer\_Bal nodes=detail;

class Race Marital\_Status N\_Stage \_6th\_Stage Differentiate A\_Stage Grade Estrogen\_Status Progesterone\_Status status;

model status(event="Dead")=age Race Marital\_Status N\_Stage \_6th\_Stage Differentiate A\_Stage Grade Estrogen\_Status Progesterone\_Status

survival\_months tumor\_size regional\_node\_examined reginol\_node\_positive;

partition fraction(validate=0.3 seed=12345);

grow Gini;

prune cc;

run;

/\*Neural Network with T\_Stage\*/

proc hpneural data=Cancer\_Bal;

partition fraction(validate=0.3 seed=12345);

target status/level=nom;

input age tumor\_size Regional\_Node\_Examined Reginol\_Node\_positive survival\_months /level=int;

input race Marital\_Status N\_stage \_6th\_stage differentiate grade A\_Stage Estrogen\_Status Progesterone\_Status /level=nom;

hidden 14;

train maxiter=1000 numtries=5;

run;

/\*Neural Network with Tumor\_Size\*/

proc hpneural data=Cancer\_Bal;

partition fraction(validate=0.3 seed=12345);

target status/level=nom;

input age tumor\_size Regional\_Node\_Examined Reginol\_Node\_positive survival\_months /level=int;

input race Marital\_Status N\_stage \_6th\_stage differentiate grade A\_Stage Estrogen\_Status Progesterone\_Status /level=nom;

hidden 14;

train maxiter=1000 numtries=5;

run;