# 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. 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.

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.
* **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** - No of months the patient has survived since detection.
* **Status** - Whether the patient is Dead or Alive.

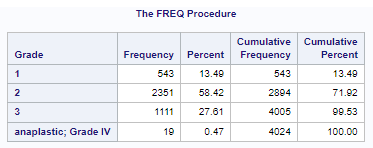
Problem

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 survival month and 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

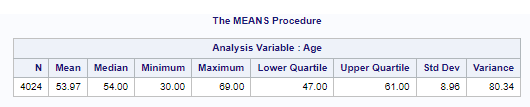
The data available at the Kaggle site was pretty consistent without null and missing fields. However, the following changes were required and made to work using SAS.

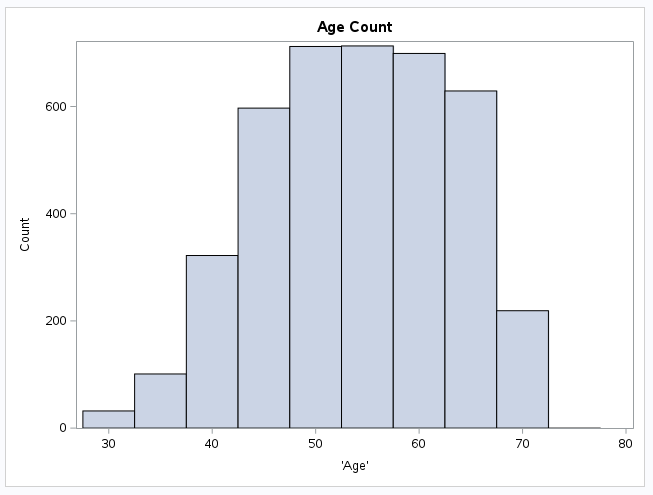
* The variables names in the initial sheet had spaces in them, to make the variables name as valid SAS names options validvarname=v7; was used.
* For Grade variables the values were 1,2,3 and anaplastic; Grade IV
  + This was giving error when importing the data as 1,2,3 were numeric and ‘anaplastic; 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**



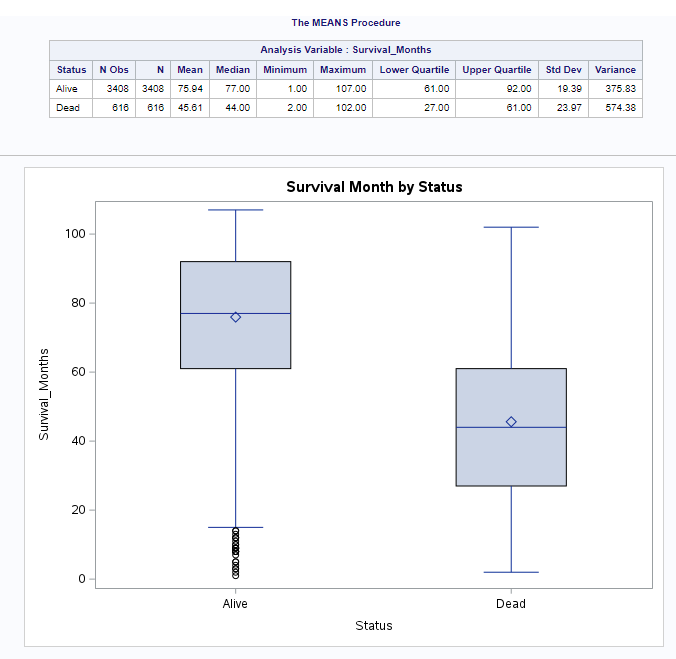


The range of patients age is from 30-69

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**

-Survial\_months by Status



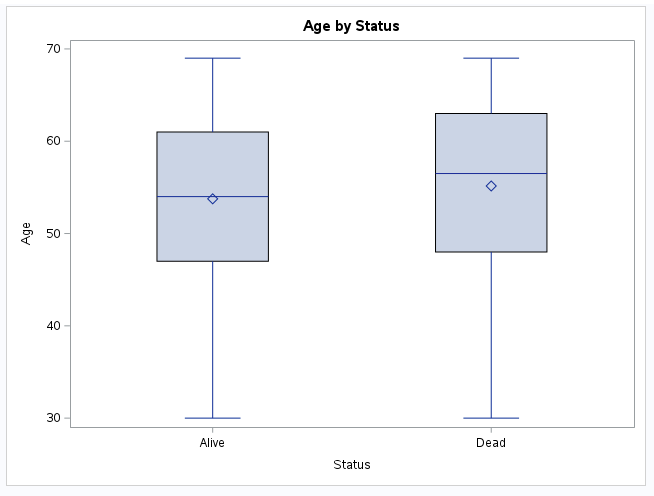
From the bar plot and analysis variable table we can see that minimum and maximum survival months of alive patients is 1 month and 107 months since detection, we see a lot of lower-level outlier in the alive patients which shows the actual range from around 18 months – 107 months. This might be because 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**

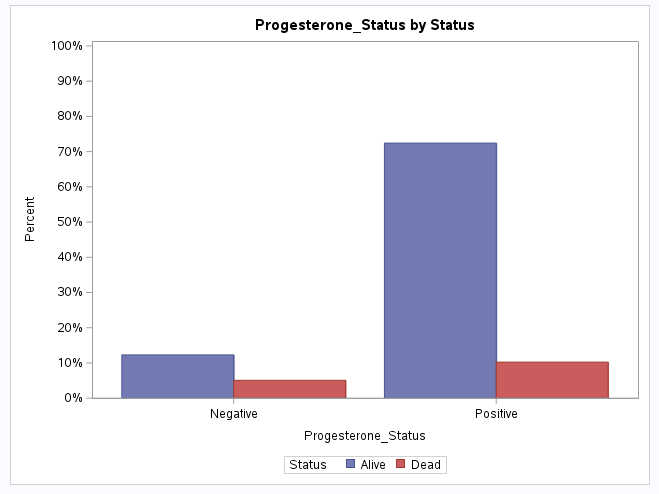
From the box plot and analysis variable table we can see that the minimum age and maximum age of the alive patient as well as the dead patient is 30 and 69 respectively. However, the mean (53.76) and median (54) age of alive patients is slightly lower than the mean (55.15) and median (56.50) of dead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

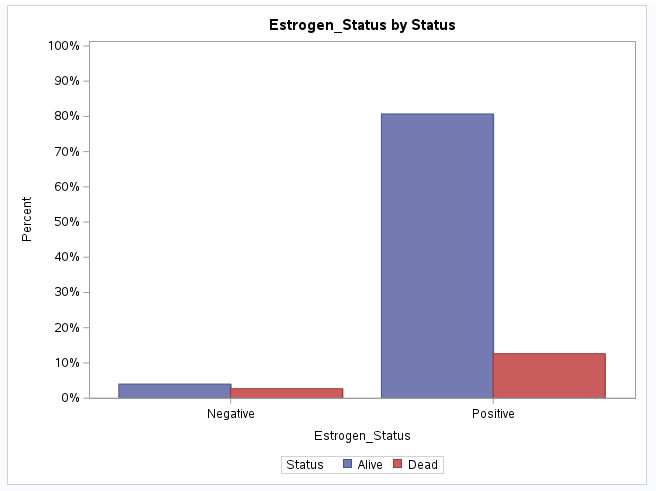
Table

Description automatically generated



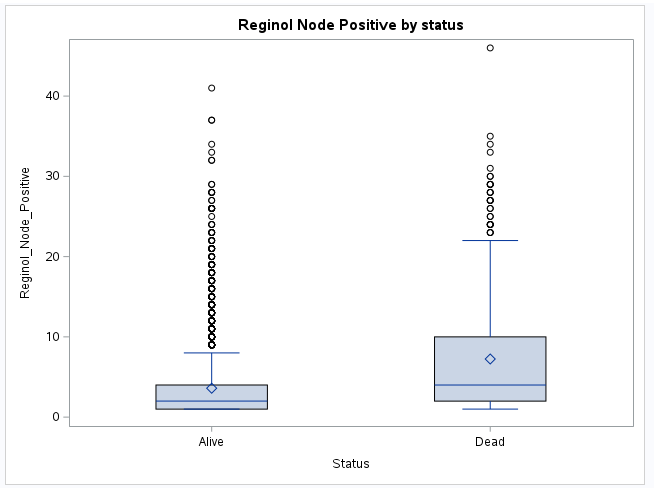
Both the “Progesterone\_Status by Status” and “Estrogen\_Status by Status” bar plots show that we have more patients with Progesterone and Estrogen positive than negative and there are percent of Alive patients (both positive and negative) than Dead. This might be because overall we have more Alive patients than dead, and it's difficult to say with this visualization if there is any relation of having Progesterone and Estrogen positive or negative. The models that we will be building and analyzing will help us to understand it better.

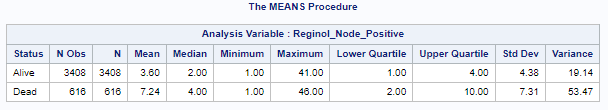




**Regional node positive**

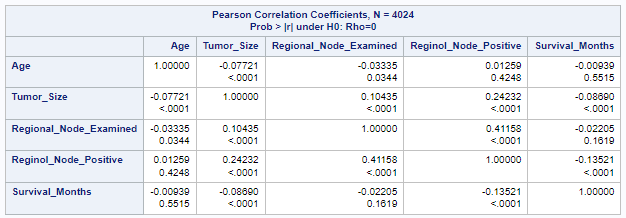
From the box plot and the Analysis Variable table, we can see that the average number of regional nodes that are found to contain metastases by the pathologist is less for living patients. Similarly, the median number of positive regional nodes is slightly lower in alive patients than in dead patients. 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.





**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 survival months and 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. We first start with model analysis of Surival\_Month, we have the following options to try and determine a good fitting model out of them:

* Multiple Linear Regression (MLR)
* CART
* Neural Network

In MLR, we will be using Proc Reg for initial modeling and finding the important variables, we will then use these variables to build the MLR model using Proc Hpreg. The important variables identified from the MLR will also be used to analyze the performance of the CART and Neural Network models. We will also be the different combinations of variables for CART and Neural Network models.

For analyzing and predicting the Status (Alive or Dead) of the patient we have the following options to try and determine a good fitting model out of them:

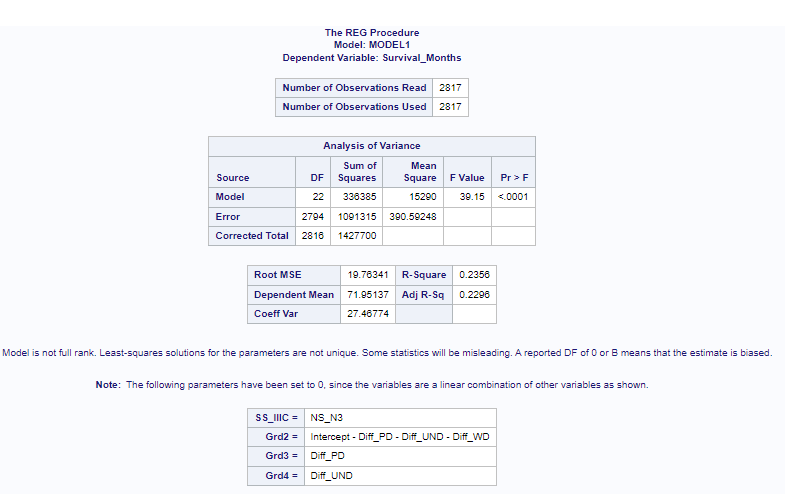
* Logistic Regression
* CART
* Neural Network

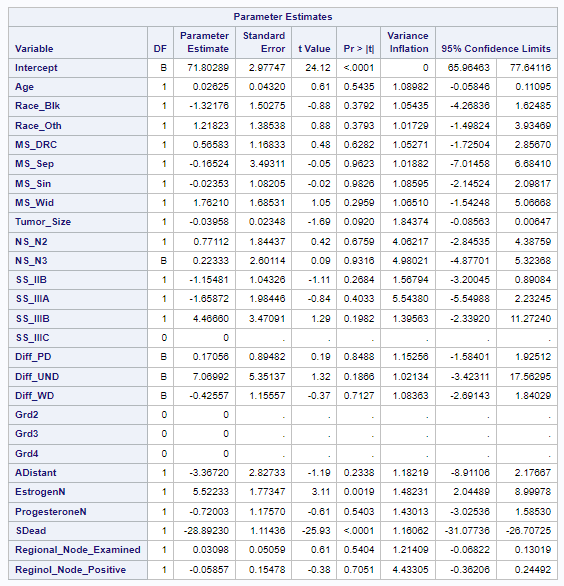
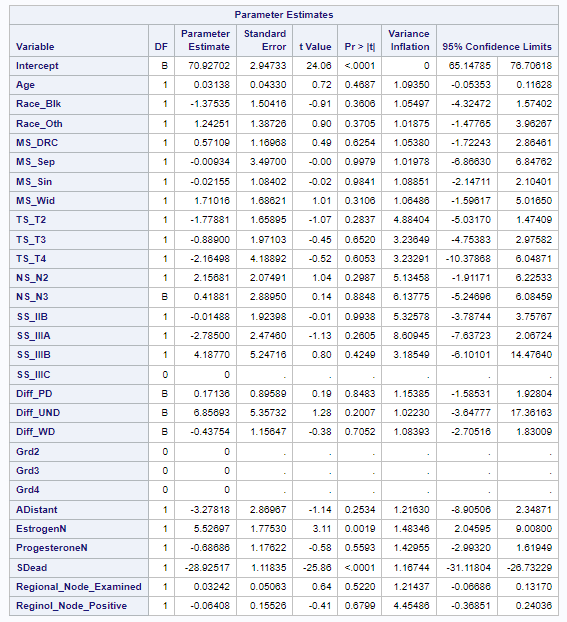
As we did for Survival month, we will use the same approach for finding important variables using the Logistic Regression model (Proc Logistics) and then use them as well as other combinations of variables to build, analyze and compare the models with each other.

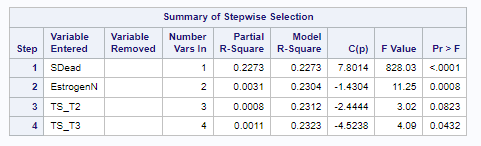
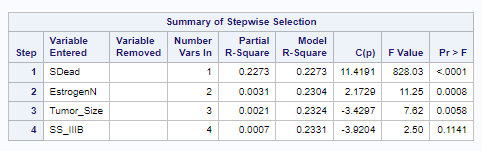
Tumor Size and T\_stage variables are correlated which we are aware of from domain understanding, so we will try with both of them individually in all our models.

**Survival Month Analysis:**

**Multiple Linear Regression –**

With T\_Stage With Tumor\_Size



Using Proc Reg and all variables (one with Tumor\_Size and other with T\_Stage), we see that there is no difference between both the models, based on the tumor variable. For Model evaluation, we can see that both the models are significant from the p-value in the analysis of the variance table, as the p-value is less than alpha (0.05). The adjusted R-Square of both the models is 22.86% and 22.96% which shows that only 23% of variation in the survival month is explained by the model, which is not a good indicator. Also, about the p-value of the individual variables, we can see that only EstrogenN and SDead are statistically significant variables as they have p-values less than alpha and zero is not a possible value in the lower and upper confidence interval. We will next try a forward, backward, and stepwise selection method to find the important variables. 

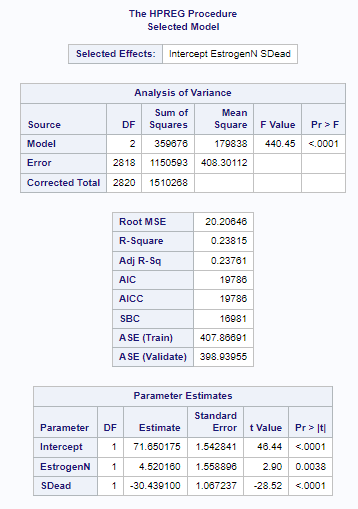
Using stepwise selection method, with Tumor\_Size, the important variables are SDead, EstrogenN, Tumor\_Size and SS\_IIB (Binary variable for 6th Stage), but SS\_IIB has p-value more than alpha, which means it is not statistically significant. With T\_stage the important variables are SDead, EstrogenN, TS\_T2 and TS\_T3 2 out of 3 binary variables of t\_Stage.

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

* SDead, EstrogenN, Tumor\_Size and SS\_IIB, SS\_IIIA, SS\_IIIB, SS\_IIIC (as we can either include and exclude all the binary variables)
* SDead, EstrogenN, Tumor\_Size
* SDead, EstrogenN, TS\_T2, TS\_T3, TS\_T4
* SDead, EstrogenN

After building the models the best variables combination with minimum variables and better ASE for validation set is with SDead and EstrogenN. However, the adjusted R-sq with MLR with these variables is still 24% which is not good.

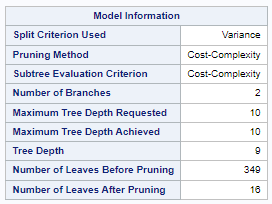
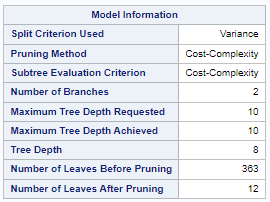
**With SDead, EstrogenN**

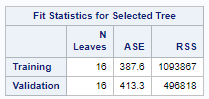
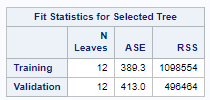


**CART**

For the CART model to predict Survival month, we initially started with all the variables and built 2 models one with Tumor\_size and then with T\_Stage.

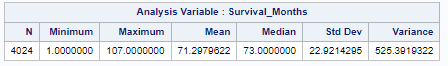
With Tumor\_size With T-Stage



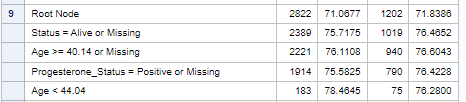


It is observed that the CART model with Tumor\_size has 12 leaves after pruning which is less compared to the model with T-stage. Comparing the model performance on training set and validation set between the 2 models, we see that ASE(Training) in the CART model with Tumor\_size is close to that in the other model. The ASE(validation) in the CART model with Tumor\_size is quite less compared to the other model. Also, there is no sign of overfitting in either of the models.

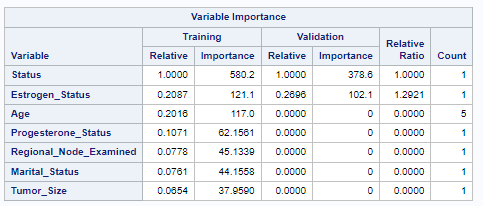
Considering the principle of parsimonious model and aiming to lower the error measure such as ASE, we chose the CART model with Tumor\_size to predict the survival months. Also, by taking the square root value of ASE in validation set(413), we get approximately 20 months which is closer to the minimum value of the survival months. This indicates that this model performed better in predicting the survival months.



From the criteria mentioned in Node 9, we could summarize that if the patient status is Alive and the patient’s age is between 40.14 and 44.04 and the breast cancer cells have progesterone receptors, then the predicted value for survival months is 76.28 months.

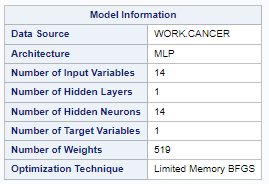


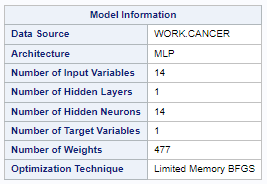
Below are the important variables that are identified by the CART model. This can further be applied to the Neural Network model and check if this improves the model’s performance.



**Neural Networks**

For NNs to predict Survival months we build 2 models using all variables - one with Tumor\_size and other with T\_stage.

**With Tumor size With T-stage**

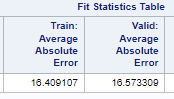
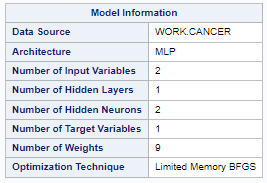


Graphical user interface, text, application, table

Description automatically generated

On the training set, the value of Average Absolute Error is less in the NN model with Tumor\_size variable than the other model. On the validation set, the value of Average Absolute Error is slightly less in the NN model with the T-stage. There is no sign of overfitting in either of the models. This shows that the NN model with T stage is better among the 2 models.

This model uses 14 hidden layers and 14 predictors in total. In order to improve the model performance, we applied only the important variables - Status and Estrogen\_status, identified from the CART model to the NN model.



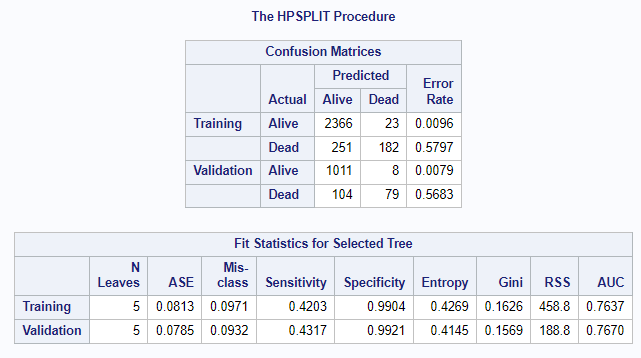
The values of MAE in the training and validation set are close to that of the NN model with T-stage. Also, the value of MAE is close to the minimum value of survival\_months. Hence, the NN model with 2 variables (status, estrogen\_status) did a better job in predicting survival\_months.

**Status Analysis**

**CART**

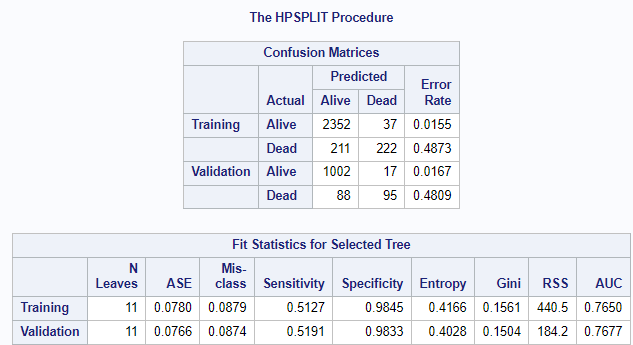
For analyzing status(dead or alive) using the CART model, we used Gini and Entropy measures using all variables and with both Tumor\_size and T\_stage separately.

1. **Entropy with T Stage:**

****

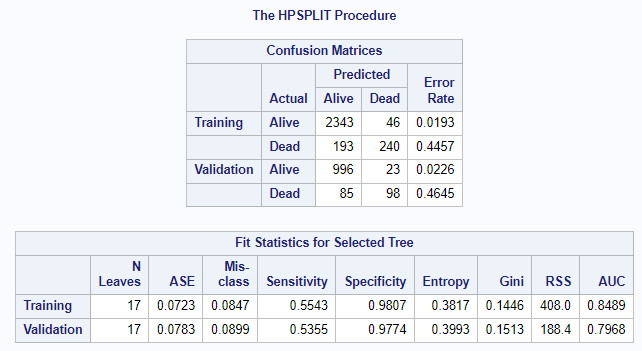
In the Entropy with T stage CART model, the number of leaves after pruning is 5. Evaluating the model performance, we can see from the confusion matrix table that 2366 out of 2389(training) and 1011 out of 1019(validation) observations, the status is classified correctly as Alive which indicates that 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:**



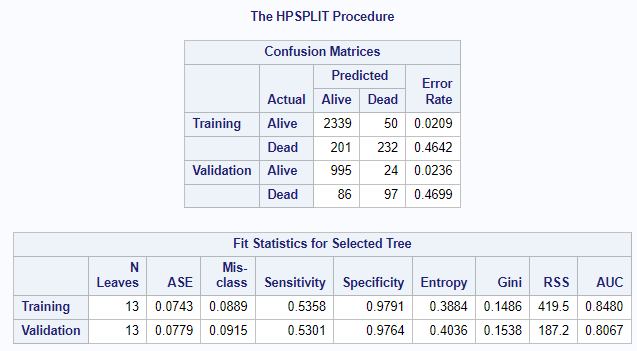
In the Entropy with Tumor size CART model, the number of leaves after pruning is 11 which is approximately double the previous model. We can see from the confusion matrix table that 2352 out of 2389(training) and 1002 out of 1019(validation) observations, the status is classified correctly as Alive which 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 a improvement compared to the previous model. AUC is 76.50% and 76.77% 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. Compared with the previous model, AUC is very close to the previous model.

1. **Gini with T Stage:**



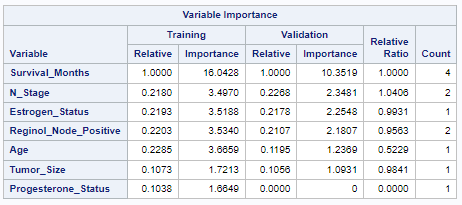
In the Gini with T Stage CART model, the number of leaves after pruning is 17 which is more than the previous 2 models. We can see from the confusion matrix table that 2343 out of 2389(training) and 996 out of 1019(validation) observations, the status is classified correctly as Alive which indicates 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 number of leaves are 17 after pruning.

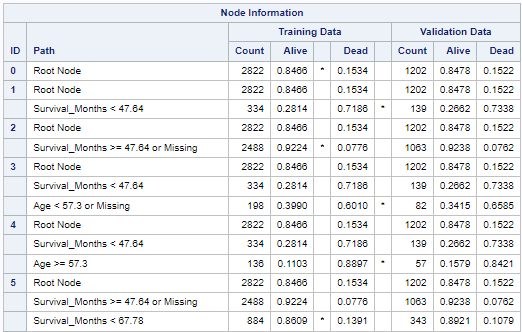
1. **Gini with Tumor Size:**



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 model (Gini with Tumor size):



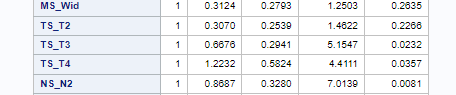
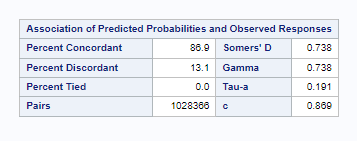
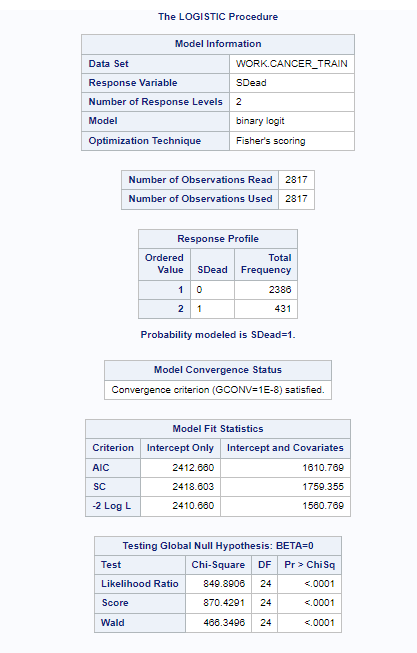


From the node information, Node 5 is interpreted as if Survival\_Months >= 47.64 or Missing and Survival\_Months < 67.78, then there is an 86.09% chance that the patient is classified as Alive.

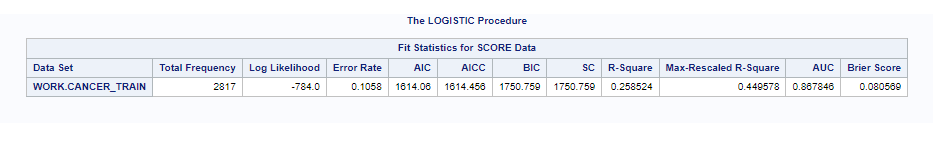
Node 4 is interpreted as if Survival\_Months < 47.64 and Age >= 57.3, then there is an 88.97% chance that the patient is classified as Dead.

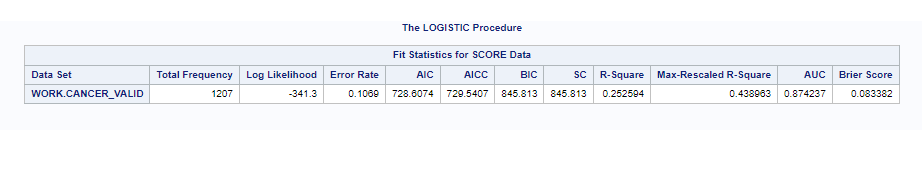
**Logistic Regression for Status:**

**With T\_Stage:**

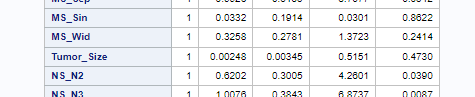
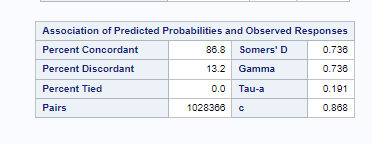
****

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 . T\_Stage - all variables are statistically significant as p value is less than 0.05.





**Tumor\_Size**

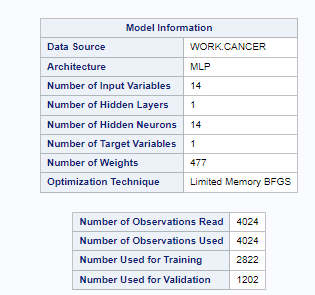
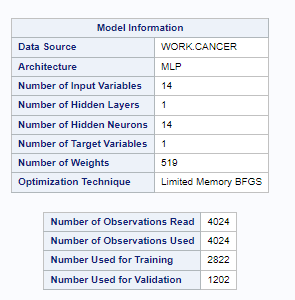


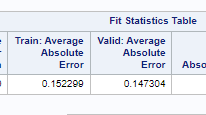
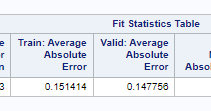
The p-value for the likelihood ratio test is less than alpha so the model is statistically significant. The value of AUC is 86.8% which is quite good. Tumor\_size - variable is statistically not significant as p value is greater than 0.05.

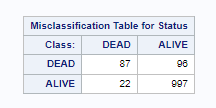
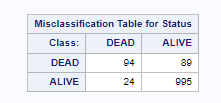
**Neural Networks for Status:**

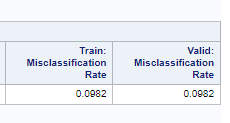
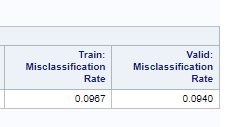
For the Neural Network's to predict status, we built 3 models - first model with Tumor\_Size , second model with T\_stage and third model with the important variables that were obtained from CART model analysis.

With Tumor\_size With T-stage

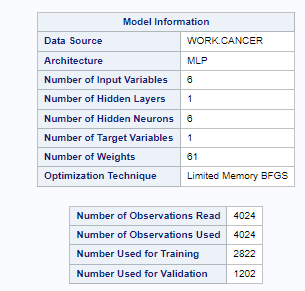
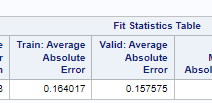
On the training set, the value of Average Absolute Error is almost the same in both the models , but the value is slightly less in the NN model with T\_Stage. On the validation set the average absolute error value is the same.

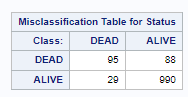
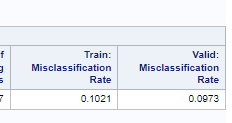
The misclassification error rate on the training set is less in the NN model with T-stage which is around 9.6%, when compared with other NN model with tumor\_size which is 9.8%. On the validation set the misclassification error rate value is less in the NN model with T-stage which is around 9.4%, when compared with other NN model with tumor\_size which is 9.8%.

There is no sign of overfitting in either the models or the NN model with T-Stage is better when compared to the other model.

**NN for status with important variables from CART model:**

Inorder to check if any improvements can be made to improve the model performance and to improve the misclassification error rate, we considered the important variables -age survival\_months Reginol\_Node\_positive tumor\_size N\_stage estrogen\_status , obtained from the CART model . The below values were obtained for the NN model with important variables.

** **

** **

The Average Absolute error on the training set is around 0.164 , which is close and slightly higher and close to the value obtained from the above NN models.On the validation set the value is 0.1575 which is almost same and close to the values obtained from the above models.

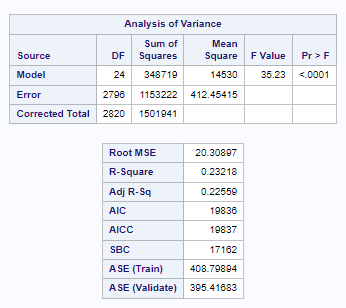
There is no sign of overfitting and if we observe the confusion matrix , it can be seen that the model was able to correctly classify 95 correctly out of 183 as to 94 out of 183 in the previous 2 NN models. If we observe the confusion matrixMisclassification error rate on training set is 10.21% which higher than NN model with T\_stage which is around 9.6% but if we the misclassification error rate on Validation set is 9.7% which is almost near to error rate of NN model with T-stage which is around 9.4%.

Overall, the NN model with T\_Stage has good model performance.

Conclusion:

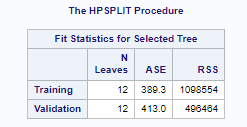
**Survival\_Months model comparison:**

**Error Metrics - MLR**



As seen from the error metrics details above of MLR, the overall model is statistically significant but the adjusted-R-sq is very low of 22%. The ASE value of the training set is 408.8 and that of validation is 395.4 which shows that there is no sign.

**Error Metrics – Cart Model**



The number of leaves using the CART model is 12 and the ASE for the training set is 389 and for the validation set is 413.

/\*SAS Code\*/

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

proc import out=Cancer 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;

table grade;

run;

/\*Data Cleaning\*/

data Cancer;

set Cancer;

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

run;

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

proc freq data=Cancer;

table grade;

run;

/\*Creating Dummy Variable\*/

data Cancer;

set Cancer;

/\*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 Count";

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;

/\*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;

/\*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;/\*stat is to change frequency to %\*/

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

title "Progesterone\_Status by Status";

run;

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

proc sgplot data=Cancer;

vbox reginol\_node\_positive/category=Status;

Title"Reginol Node Positive by status";

run;

/\* Correlation matrix \*/

proc corr data=Cancer;

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

run;

/\*------------------------------------------------------SURVIVAL\_MONTH---------------------------------------------------------------\*/

/\*----------------------------Multiple Linear Regression------------------------------------\*/

/\*Multiple Linear Regression to predict Survival Months using all variables except Tumor\_Size and by Proc Reg\*/

Proc Reg data=Cancer\_train; /\*Reg = Regression\*/

Model Survival\_Months=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 SDead Regional\_Node\_Examined Reginol\_Node\_Positive/clb corrb vif ;

run;

/\*Multiple Linear Regression to predict Survival Months using all variables except T\_Stage and by Proc Reg\*/

Proc Reg data=Cancer\_train; /\*Reg = Regression\*/

Model Survival\_Months=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 SDead Regional\_Node\_Examined Reginol\_Node\_Positive/clb corrb vif ;

run;

/\*Multiple Linear Regression to predict Survival Months using all variables except Tumor\_Size and by Proc HpReg\*/

proc hpreg data=Cancer seed=12345;

partition fraction(validate=0.3);

model Survival\_Months=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 SDead Regional\_Node\_Examined Reginol\_Node\_Positive;

run;

/\*Multiple Linear Regression to predict Survival Months using all variables except T\_Stage and by Proc HpReg\*/

proc hpreg data=Cancer seed=12345;

partition fraction(validate=0.3);

model Survival\_Months=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 SDead Regional\_Node\_Examined Reginol\_Node\_Positive;

selection method=stepwise(choose=validate);

run;

/\*Variables selection\*/

/\*Multiple Linear Regression to predict Survival Months with Forward Selection\*/

Proc Reg data=Cancer\_train; /\*Reg = Regression\*/

Model Survival\_Months=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 SDead Regional\_Node\_Examined Reginol\_Node\_Positive/selection=forward;

run;

/\*Multiple Linear Regression to predict Survival Months with backward Selection\*/

Proc Reg data=Cancer\_train; /\*Reg = Regression\*/

Model Survival\_Months=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 SDead Regional\_Node\_Examined Reginol\_Node\_Positive/selection=backward ;

run;

/\*Multiple Linear Regression to predict Survival Months with Stepwise Selection with T\_Stage\*/

Proc Reg data=Cancer\_train; /\*Reg = Regression\*/

Model Survival\_Months=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 SDead Regional\_Node\_Examined Reginol\_Node\_Positive/selection=Stepwise ;

run;

/\*Multiple Linear Regression to predict Survival Months with Stepwise Selection with Tumor\_Size\*/

Proc Reg data=Cancer\_train; /\*Reg = Regression\*/

Model Survival\_Months=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 SDead Regional\_Node\_Examined Reginol\_Node\_Positive/selection=Stepwise ;

run;

/\*Hpreg with t\_Stage and all variables with stepwise selection\*/

proc hpreg data=Cancer seed=12345;

partition fraction(validate=0.3);

model Survival\_Months=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 SDead Regional\_Node\_Examined Reginol\_Node\_Positive;

selection method=stepwise(choose=validate);

run;

/\*Hpreg with Tumor\_size and all variables with stepwise selection\*/

proc hpreg data=Cancer seed=12345;

partition fraction(validate=0.3);

model Survival\_Months=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 SDead Regional\_Node\_Examined Reginol\_Node\_Positive;

selection method=stepwise(choose=validate);

run;

/\*with SDead, EstrogenN, Tumor\_Size and SS\_IIB, SS\_IIIA, SS\_IIIB, SS\_II\*/

proc hpreg data=Cancer seed=12345;

partition fraction(validate=0.3);

model Survival\_Months=Tumor\_size SS\_IIB SS\_IIIA SS\_IIIB SS\_IIIC EstrogenN SDead;

selection method=stepwise(choose=validate);

run;

/\*with SDead, EstrogenN, Tumor\_Size \*/

proc hpreg data=Cancer seed=12345;

partition fraction(validate=0.3);

model Survival\_Months=Tumor\_size EstrogenN SDead;

selection method=stepwise(choose=validate);

run;

/\*with SDead, EstrogenN, TS\_T2 TS\_T3 TS\_T4 \*/

proc hpreg data=Cancer seed=12345;

partition fraction(validate=0.3);

model Survival\_Months=TS\_T2 TS\_T3 TS\_T4 EstrogenN SDead;

selection method=stepwise(choose=validate);

run;

/\*with SDead, EstrogenN \*/

proc hpreg data=Cancer seed=12345;

partition fraction(validate=0.3);

model Survival\_Months= EstrogenN SDead;

selection method=stepwise(choose=validate);

run;

/\*Final model with SDead, EstrogenN \*/

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

/\* CART for Survival months\*/

/\* 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 survival\_months=age Race Marital\_Status T\_Stage N\_Stage \_6th\_Stage Differentiate A\_Stage Grade Estrogen\_Status Progesterone\_Status

regional\_node\_examined reginol\_node\_positive status;

grow rss;

prune cc;

partition fraction(validate=0.3 seed=12345);

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 survival\_months=age Race Marital\_Status N\_Stage \_6th\_Stage Differentiate A\_Stage Grade Estrogen\_Status Progesterone\_Status

tumor\_size regional\_node\_examined reginol\_node\_positive status;

grow rss;

prune cc;

partition fraction(validate=0.3 seed=12345);

run;

proc means data=cancer n min max mean median stddev var;

var survival\_months;

run;

/\* with only important variables from CART -\*/

proc hpsplit data=cancer nodes=detail;

class Estrogen\_Status status;

model survival\_months=Estrogen\_Status status;

grow rss;

prune cc;

partition fraction(validate=0.3 seed=12345);

run;

/\*----------------------------Neural Netwrok------------------------------------\*/

/\* NN for Survival months \*/

/\* with Tumor\_size\*/

proc hpneural data=cancer;

partition fraction(validate=0.3 seed=12345);

target survival\_months/level=int;

input age tumor\_size regional\_node\_examined reginol\_node\_positive /level=int;

input Race Marital\_Status N\_Stage \_6th\_Stage Differentiate A\_Stage Grade Estrogen\_Status Progesterone\_Status status/level=nom;

hidden 14;

train maxiter=1000 numtries=5;

run;

/\* with T stage \*/

proc hpneural data=cancer;

partition fraction(validate=0.3 seed=12345);

target survival\_months/level=int;

input age regional\_node\_examined reginol\_node\_positive /level=int;

input Race Marital\_Status T\_Stage N\_Stage \_6th\_Stage Differentiate A\_Stage Grade Estrogen\_Status Progesterone\_Status status/level=nom;

hidden 14;

train maxiter=1000 numtries=5;

run;

/\* using important variables from CART tumor\_size\*/

proc hpneural data=cancer;

partition fraction(validate=0.3 seed=12345);

target survival\_months/level=int;

\*input age tumor\_size regional\_node\_examined/level=int;

input Estrogen\_Status status/level=nom;

hidden 2;

train maxiter=1000 numtries=5;

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;

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

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 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\*/

/\*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\*/

/\*Logistic Regression with important variablesfrom 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 exculding T\_Stage \*/

/\*Logistic Regression with important variablesfrom the Stepwise selectionremoving 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;

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;

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;

proc hpsplit data=cancer nodes=detail;

class N\_Stage Estrogen\_Status Progesterone\_Status status;

model status(event="Dead")=age N\_Stage Estrogen\_Status Progesterone\_Status survival\_months tumor\_size reginol\_node\_positive;

partition fraction(validate=0.3 seed=12345);

grow gini;

prune cc;

run;

proc hpsplit data=cancer nodes=detail;

class N\_Stage Race Estrogen\_Status status Differentiate;

model status(event="Dead")=Age Race N\_Stage Differentiate Estrogen\_Status Survival\_Months Regional\_Node\_Examined Reginol\_Node\_Positive;

partition fraction(validate=0.3 seed=12345);

grow gini;

prune cc;

run;