Group Project
Breast Cancer

Alejandra Barrios– 301314502
Paula Carrillo – 301314500
Sebastian Gonzalez – 301336125
Saksham Tiwari- 301334725

Centennial College
Business Analytics Capstone
David Parent
July 15, 2024

Contents

Introduction

The objective of this project is to determine the characteristics of the patients who are more likely to die of breast cancer with a dataset extracted from Kaggle Website[1].

The project has several steps including an initial exploratory analysis to understand a general panorama of the records; cleaning and managing outliers of the dataset; running different models such as decision trees, regressions, and Neural Networks to predict which patients are more likely to die according to different attributes of the dataset; comparing each model considering Average Squared Error (ASE) and ROC index; finally, choosing the best model using ASE as a validation assessment rating and interpret its results to make recommendations to the health industry.

Data Setup and Exploration

The dataset chosen is from patients with breast cancer, it has a total of 4024 records with the following variables:

Name	Description	Role	Level
Age	The age in years of each patient	Input	Interval
Race	The ethnicity or skin color of the patient (White, Black, Other)	Input	Nominal
Marital Status	Status of relationship of the patient (Divorced, Married, Separated, Single, and Widowed)	Input	Nominal
T Stage	Indicates the size of the main tumor. The higher the number after T, the larger the tumor **T1**: Tumor is 2 cm (3/4 of an inch) or less across. **T2**: Tumor is more than 2 cm but not more than 5 cm (2 inches) across. **T3**: Tumor is more than 5 cm across. **T4**: Tumor of any size growing into the chest wall or skin. This includes inflammatory breast cancer.	Input	Nominal

[1] Dataset link: https://www.kaggle.com/datasets/reihanenamdari/breast-cancer

N Stage	Refer to the number and location of lymph nodes that contain cancer. The higher the number after N, the more lymph nodes that contain cancer. **N1**: Cancer has spread to 1 to 3 lymph nodes under the arm with at least one area of cancer spread greater than 2 mm across. **N2**: Cancer has spread to 4 to 9 lymph nodes under the arm, or cancer has enlarged the internal mammary lymph nodes. **N3**: Cancer has spread to 10 or more axillary lymph nodes, with at least one area of cancer spread greater than 2 mm.	Input	Nominal
6th Stage	**IIA:** The tumor is less than 2 centimeters and less than four axillary lymph nodes have cancer cells present. **IIB:** The tumor is between 2 and 5 centimeters and has spread to less than four axillary lymph nodes. **IIIA:** The tumor is larger than the approximate size of a small lime (over 5 centimeters), AND the cancer has spread to 1, 2, or 3 lymph nodes under the arm or near the breastbone. **IIIB:** the tumor has grown into the muscles of the chest wall or skin. **IIIC:** The cancer has spread to 10 or more axillary lymph nodes	Input	Nominal
Differentiate	How the cells look like. Going from well differentiated to undifferentiated.	Rejected	Nominal
Grade	**1**: The cancer cells are well differentiated. They look almost like normal cells. **2**: The cancer cells are moderately differentiated. They are between grades 1 and 3. **3**: The cancer cells are poorly differentiated or undifferentiated. They look less normal, or more abnormal, than healthy cells. **anaplastic; Grade IV:** The cells look undifferentiated or abnormal.	Input	Nominal
A Stage	**Regional:** The cancer has spread outside the breast to nearby lymph nodes. **Distant:** The cancer has spread to distant parts of the body (Such as lungs, liver, bones, etc.)	Input	Nominal
Tumor Size	Size of tumor in mm	Input	Interval

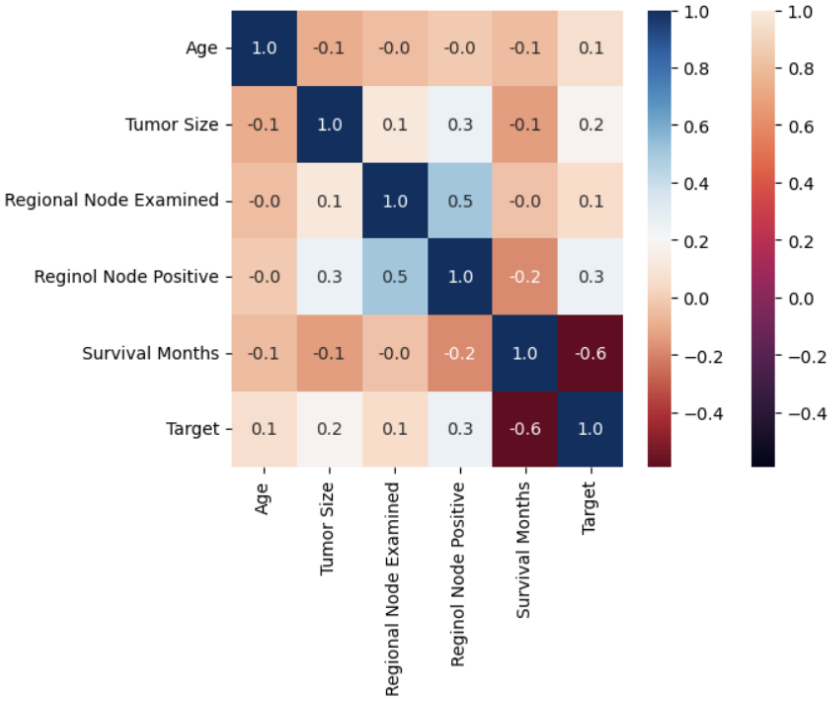
Estrogen Status	**Positive:** Cancer cells have receptors for estrogen. **Negative:** Otherwise	Input	Nominal
Progesterone Status	**Positive:** Cancer cells have receptors for progesterone. **Negative:** Otherwise	Input	Nominal
Regional Node Examined	Total number of regional lymph nodes that were removed and examined by the pathologist.	Input	Interval
Reginol Node Positive	Number of regional lymph nodes examined that were positive for cancer.	Input	Interval
Survival Months	Number of months that the person is being alive after the diagnosis.	Rejected	Interval
Status	If the patient survived (1 dead - 0 alive)	Target	Binary

Variable Analysis

In this section, a Heatmap and Chi-Square Test of independence was created to discover any relationship between the variables for numerical and categorical variables, respectively.

Heatmap

The Heatmap identifies that the variable 'Survival Months' has a strong relationship negative correlation (-0.6) with 'Status' (Target label). This correlation suggests that as the number of survival months increases, the likelihood of Status being 1 (Dead) decreases, which aligns with the expected outcome that longer survival time correlates with being live.



For this reason, Survival Months is rejected in the model to address the curse of dimensionality and ensure that the model remains focused and free from redundancy.

Chi-Square Test of independence

A chi-square test was conducted between the categorical variables 'Differentiate' and 'Grade since it appears to have a relationship. The Contingency table shows the frequency of each category of 'Differentiate' compared across the categories of 'Grade'. The null hypothesis for this test is that there is no relationship between the two variables; whereas the alternative hypothesis is that there is a relationship between the two variables.

Contingency Table:

Grade	anaplastic; Grade IV	1	2	3
differentiate				
Moderately differentiated	0	0	697	0
Poorly differentiated	0	0	0	406
Undifferentiated	10	0	0	0
Well differentiated	0	119	0	0

The Chi-square test, with a significance level of 0.05, shows that the p-value is 0.0, meaning that the null hypothesis is rejected and concluding that there is an association between the two variables. A p-value of 0.0 indicates a extremely strong evidence against the null hypothesis.

```
Chi-Square Statistic: 3696.0
p-value: 0.0
Degrees of Freedom: 9
```

For this reason, the variable differentiate is rejected in the model.

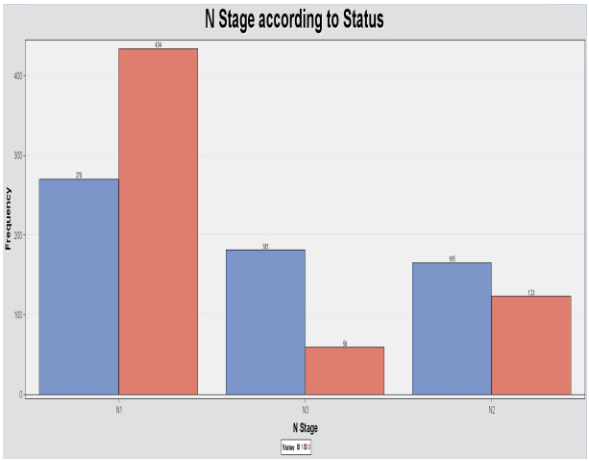
Descriptive Statistics

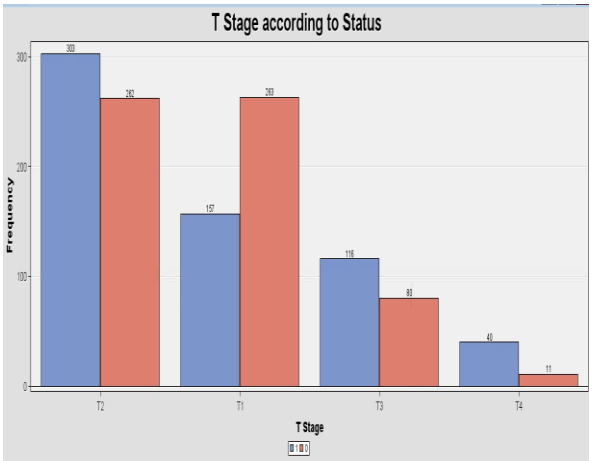
The following table presents the measures of central tendency for each variable of the dataset. The dataset has no missing values for any variable which means it is not necessary to apply techniques to manage these values; also, the numerical variables present a range (minimun and maximun) and mean. For instance, the age of patients is between 30 years and 69 years with a mean of 54.52. The regional Node Positive vary between 1 tumor and 46 tumors and mean of 5.60, which it could indicate that it migh have outliers as well as tumor size with a range between 1 tumor and 140 tumors and mean of 33.17.

As for the categorical varibles, the table shows the mode which indicates the most frequent category of the variable. The most popular characteristics in patients are the marital status married, race white, grade 2 which means the cancer cells are moderately differentiated, N1 stage which means the cancer has spread to 1 to 3 lymph nodes, T2 stage which means Tumor is more than 2 cm but not more than 5 cm.

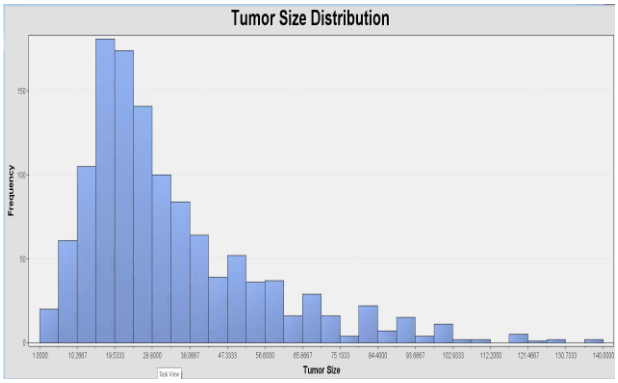
Obs #	Variable Name	Label	Type	Percen...	Minimum	Maximum	Mean	Number of Levels	Mode Percentage	Mode
1	A Stage	A Stage	CLASS	0	.	.	.	2	96.42857	REGIONAL
2	Estrogen Status	Estrogen Status	CLASS	0	.	.	.	2	88.96104	POSITIVE
3	Grade		CLASS	0	.	.	.	4	56.57468	2
4	Marital Status	Marital Status	CLASS	0	.	.	.	5	62.98701	MARRIED
5	N Stage	N Stage	CLASS	0	.	.	.	3	57.14286	N1
6	Progesterone Status	Progesterone Status	CLASS	0	.	.	.	2	76.94805	POSITIVE
7	Race		CLASS	0	.	.	.	3	84.25325	WHITE
8	T Stage	T Stage	CLASS	0	.	.	.	4	45.86039	T2
9	6th Stage	6th Stage	CLASS	0	.	.	.	5	27.51623	IIIA
10	differentiate		CLASS	0	.	.	.	4	56.57468	MODER...
11	Age		VAR	0	30	69	54.51948	.	.	.
12	Reginol Node Positive	Reginol Node Positive	VAR	0	1	46	5.593344	.	.	.
13	Regional Node Examined	Regional Node Examined	VAR	0	1	57	14.56331	.	.	.
14	Status		VAR	0	0	1	0.5	.	.	.
15	Survival Months	Survival Months	VAR	0	2	107	61.44968	.	.	.
16	Tumor Size	Tumor Size	VAR	0	1	140	33.16721	.	.	.

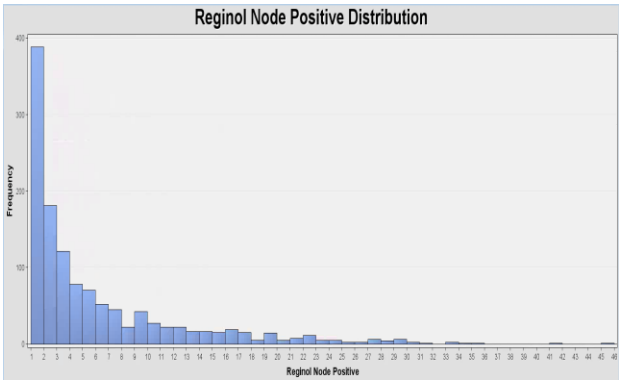
The graphs represent the distribution of N stage and T stage according to the status, showing that the most frequent patients dead are in N1 stage with 270 patients, and in T2 stage with 303 patients who have died.

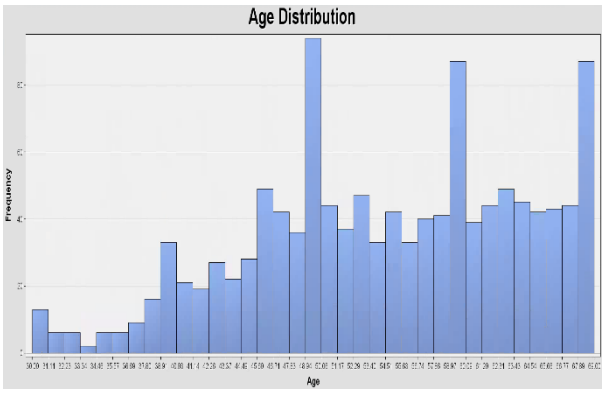


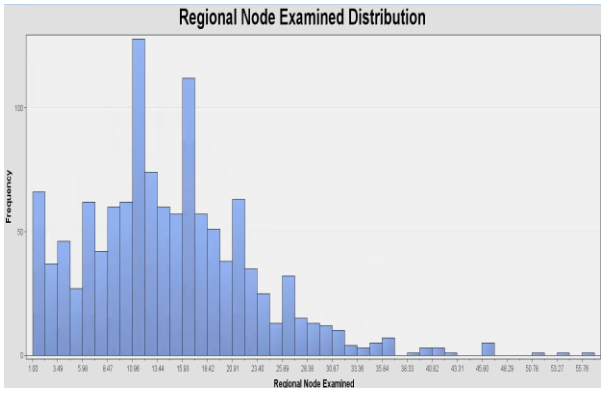


The histograms show the distribution of the variables Tumor size, Reginol Node Positive, Age and Regional Node Examined. The tumor size and Reginol Node positive is skewed in the right tail.



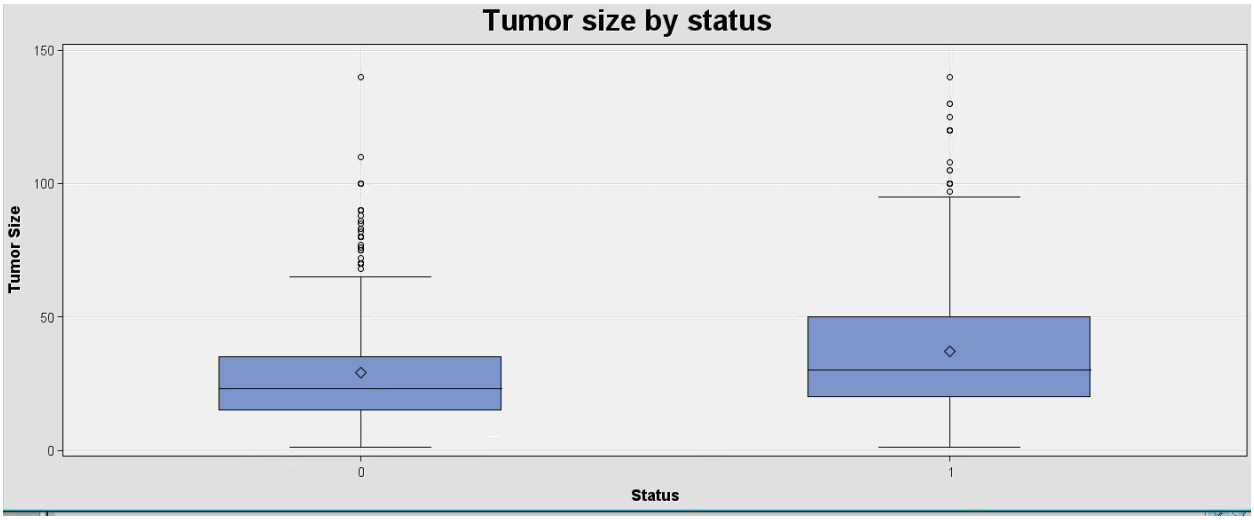






The box plot shows how data are distributed and any outliers. In this case, as mentioned earlier, tumor size has outliers; the median of the tumor size for patients that are alive is less than the median of the patients that died.

The next graph shows the that the tumor size of the patients alive (Status is 0) has more variability than the patients dead (Status is 1) although the patients dead have greater outliers.



Over sampling

The original dataset had 4024 records, 3408 of them had the status of “alive”, and 616 “Dead”. A random sampling was performed to balance the dataset and avoid bias in the model; 616 from the 3408 “alive” records were randomly choosing in Microsoft Excel.

For the target variable **Status**, the original dataset was modified from Nominal to binary, replacing “Dead” to 1 and “Alive” to 0.

Decision Tree Model

Maximal Tree

The Average -Squared error for the Maximal Tree is 0.236363. Figure 1 and Figure 2 in the Appendix section exhibit the statistics results and the diagram for the Maximal Tree.

Misclassification Tree

The Average -Squared error for the Misclassification Tree is 0.229525. Figure 3 and Figure 4 in the Appendix section exhibit the statistics results and the diagram for the Misclassification Tree.

Average Square Error Tree

The Average-Squared error for the Average Square Error Tree is 0.228703. Figure 5 and Figure 6 in the Appendix section exhibit the statistics results and the diagram for the Average Square Error Tree.

Optimal Tree

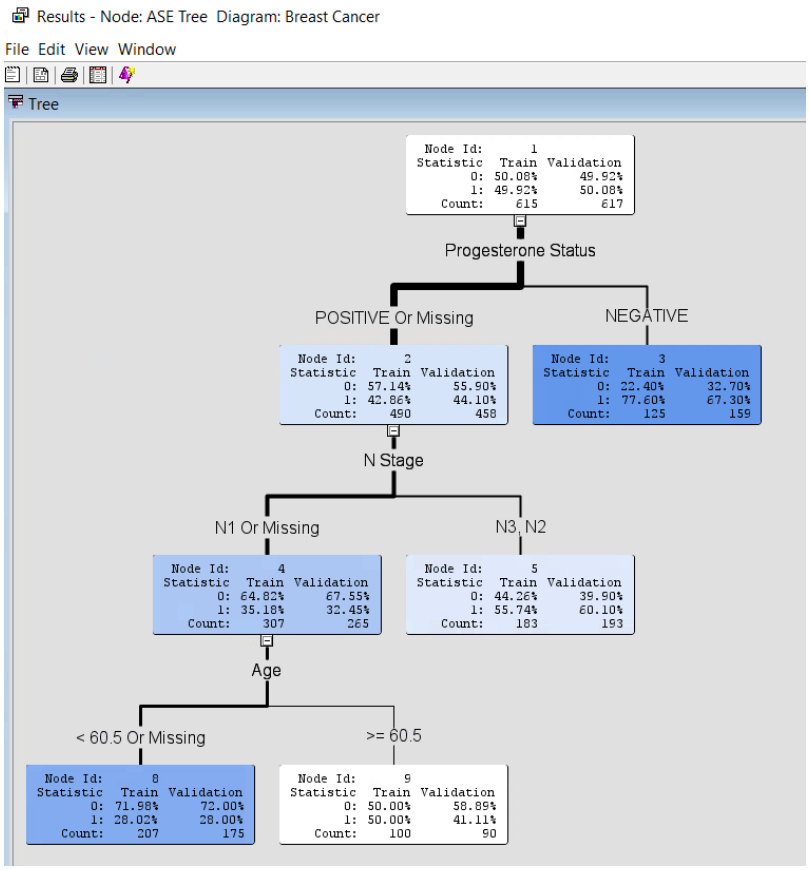
The ASE decision tree with the average square error of 0.228703 is the optimal choice of tree for predicting patients who are more likely to die of breast cancer. Having the lowest average square error indicated the model predictions are closer to the true values on average. The decision tree is better suited for identifying high-risk individuals, contributing to more effective and potentially life-saving interventions.

Based upon the average square error tree, patients with positive or missing progesterone status are less likely to die at 44.10% compared to negative status with 67.3%, indicating **negative progesterone status** might be a positive prognosis indicator.

With **N stages** N3 and N2 the survival rate lowers as potential dead percentage holds to be 60.1%, with N1 stage having 32.45% and this aligns with expected relationship between advance stages and reduced survival.

Under **Age**, younger patients have less death rate of 28% compared to older patients at 41.11%. This suggests age might play a role in deaths within this specific group.

Progesterone status and N stage seems to be the important factors influencing deaths in this model.



Impute

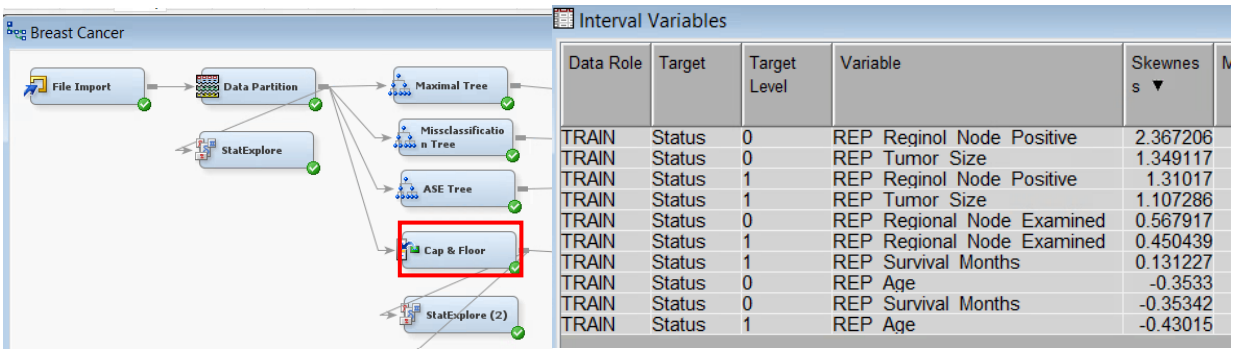
The dataset did not contain any missing values, for that reason the imputation was not needed. The following image shows there were not any missing values for non of the variables.

Interval Variables

Data Role	Target	Target Level	Variable	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness ▼	Kurtosis	Role	Label	Scaled Mean Deviation	Maximum Deviation	Level Id
TRAIN	Status	0	Reginol Node Positive	2	0	308	1	41	4.353896	5.773811	2.881648	10.2049	INPUT	Reginol ...	-0.24975	0.250562	1
TRAIN	Status	1	Reginol Node Positive	2	0	307	1	35	7.257329	7.510487	1.535526	1.828082	INPUT	Reginol ...	0.250562	0.250562	2
TRAIN	Status	0	Tumor Size	25	0	308	1	100	30.12013	20.84756	1.349117	1.356162	INPUT	Tumor S...	-0.11799	0.118379	1
TRAIN	Status	1	Tumor Size	30	0	307	2	125	38.19218	25.12268	1.266795	1.177582	INPUT	Tumor S...	0.118379	0.118379	2
TRAIN	Status	0	Regional Node Examined	14	0	308	1	41	14.70455	8.220518	0.60727	0.516916	INPUT	Regional...	-0.01618	0.01623	1
TRAIN	Status	1	Regional Node Examined	15	0	307	1	47	15.18893	8.007562	0.543017	0.328874	INPUT	Regional...	0.01623	0.01623	2
TRAIN	Status	0	Age	54	0	308	30	69	53.7013	8.95383	-0.3533	-0.36014	INPUT	Age	-0.01393	0.01398	1
TRAIN	Status	1	Age	54	0	307	30	69	55.2215	9.924377	-0.43015	-0.77218	INPUT	Age	0.01398	0.01398	2

Adjust Outliers

In the previous image, there are 2 variables with a skewness higher than 1: Reginol Node Positive and Tumor Size. To address this, capping and flooring was employed to reduce the skewness of those two variables. Although this technique resulted in a reduction in skewness, both variables still presented skewness values exceeding 1.



Interval Variables

Data Role	Target	Target Level	Variable	Skewness ▼
TRAIN	Status	0	REP Reginol Node Positive	2.367206
TRAIN	Status	0	REP Tumor Size	1.349117
TRAIN	Status	1	REP Reginol Node Positive	1.31017
TRAIN	Status	1	REP Tumor Size	1.107286
TRAIN	Status	0	REP Regional Node Examined	0.567917
TRAIN	Status	1	REP Regional Node Examined	0.450439
TRAIN	Status	1	REP Survival Months	0.131227
TRAIN	Status	0	REP Age	-0.3533
TRAIN	Status	0	REP Survival Months	-0.35342
TRAIN	Status	1	REP Age	-0.43015

Skewness

To continue reducing the skewness, a log transformation method was used in both variables mentioned above, to make the data more closely to a normal distribution. The method reduced the skewness in less than 1 for both variables with its respective levels.

Results - Node: StatExplore (2) Diagram: Breast Cancer

File Edit View Window

Interval Variables

Data Role	Target	Target Level	Variable	Skewness ▼	Median	Missing
TRAIN	Status	0	LOG REP Reginol Node Positive	0.997388	1.098612	0
TRAIN	Status	0	REP Regional Node Examined	0.567917	14	0
TRAIN	Status	1	REP Regional Node Examined	0.450439	15	0
TRAIN	Status	1	LOG REP Reginol Node Positive	0.309194	1.609438	0
TRAIN	Status	1	REP Survival Months	0.131227	47	0
TRAIN	Status	0	LOG REP Tumor Size	-0.21116	3.258097	0
TRAIN	Status	1	LOG REP Tumor Size	-0.25042	3.433987	0
TRAIN	Status	0	REP Age	-0.3533	54	0
TRAIN	Status	0	REP Survival Months	-0.35342	78	0
TRAIN	Status	1	REP Age	-0.43015	57	0

Replace Dummies

For the study case, no categorical variable was replaced; instead, the categorical variables were kept as the original variables without any substitution to not loss any relevant information. The categorical variables in the dataset represent different patterns or conditions associated with breast cancer that are crucial to have a comprehensive understanding within the data.

Logistic Regression Model

Four regression models were applied: Full Regression, Forward Regression, Backward Regression, and Stepwise Regression. Average Square Error was used as the metric to decide which is the optimal regression model.

Full Regression

The Average Square Error for the Full Regression is 0.211012 **(Appendix – Figure 7).**

Forward Regression

The Average Square Error for the Forward Regression is 0.211064 (Appendix – Figure 8).

Backward Regression

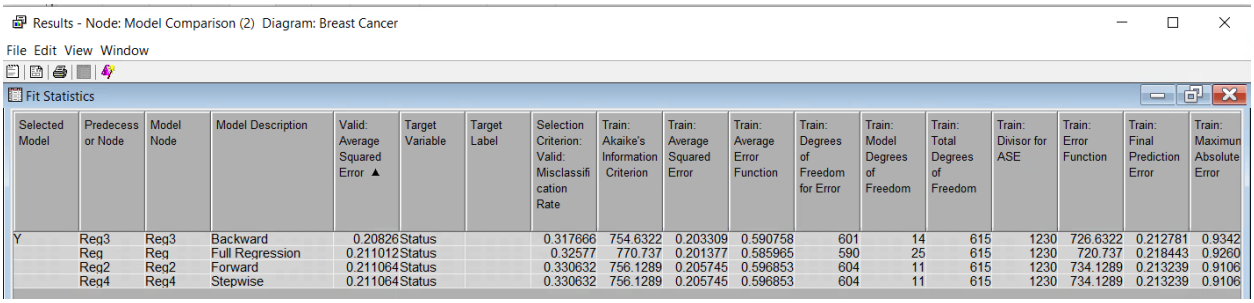
The Average Square Error for the Backward Regression is 0.20826 (Appendix – Figure 9).

Stepwise Regression

The Average Square Error for the Stepwise Regression is 0.211064 (Appendix – Figure 10).

Optimal Regression

After comparing the Average Square Error (ASE) of the four regressions, the optimal regression with the lower ASE is Backward Regression.


Results - Node: Model Comparison (2) Diagram: Breast Cancer

File Edit View Window

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Valid: Average Squared Error ▲	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Akaike's Information Criterion	Train: Average Squared Error	Train: Average Error Function	Train: Degrees of Freedom for Error	Train: Model Degrees of Freedom	Train: Total Degrees of Freedom	Train: Divisor for ASE	Train: Error Function	Train: Final Prediction Error	Train: Maximum Absolute Error
Y	Reg3	Reg3	Backward	0.20826	Status		0.317666	754.6322	0.203309	0.590758	601	14	615	1230	726.6322	0.212781	0.9342
	Reg	Reg	Full Regression	0.211012	Status		0.32577	770.737	0.201377	0.585965	590	25	615	1230	720.737	0.218443	0.9260
	Reg2	Reg2	Forward	0.211064	Status		0.330632	756.1289	0.205745	0.596853	604	11	615	1230	734.1289	0.213239	0.9106
	Reg4	Reg4	Stepwise	0.211064	Status		0.330632	756.1289	0.205745	0.596853	604	11	615	1230	734.1289	0.213239	0.9106

The variables that are included in the final model are:

- Grade.
- LOG_REP_Reginol_Node_Positive.
- Progesterone_Status.
- REP_Age.
- REP_Regional_Node_Examined.
- T_Stage.

Results - Node: Backward Diagram: Breast Cancer

File Edit View Window

Output

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		1	1.1252	22.2256	0.00	0.9596		3.081
Grade	1	1	-2.8161	22.2193	0.02	0.8991		0.060
Grade	2	1	-2.7115	22.2184	0.01	0.9029		0.066
Grade	3	1	-2.1056	22.2185	0.01	0.9245		0.122
LOG_REP_Reginol_Node_Positive		1	0.6266	0.1335	22.03	<.0001	0.2802	1.871
Progesterone_Status	Negative	1	0.6576	0.1231	28.55	<.0001		1.930
REP_Age		1	0.0265	0.00969	7.50	0.0062	0.1386	1.027
REP_Regional_Node_Examined		1	-0.0264	0.0126	4.39	0.0362	-0.1172	0.974
T_Stage	T1	1	-0.5302	0.1860	8.13	0.0044		0.588
T_Stage	T2	1	-0.0997	0.1690	0.35	0.5553		0.905
T_Stage	T3	1	-0.1530	0.2171	0.50	0.4808		0.858

However, the variables with the higher Chi-Square are the most important variables for the model. For this model the variables that are more important are:

- Progesterone_Status with a chi-square of 28.55.
- LOG_REP_Reginol_Node_Positive with a chi-square of 22.03.
- T_Stage T1 with a chi-square of 8.13.
- REP_Age with a chi-square of 7.50.

The next image indicates the odds ratio estimates for each input, following with the interpretation:

Results - Node: Backward Diagram: Breast Cancer

File Edit View Window

Output

Odds Ratio Estimates

Effect		Point Estimate
Grade	1 vs anaplastic; Grade IV	<0.001
Grade	2 vs anaplastic; Grade IV	<0.001
Grade	3 vs anaplastic; Grade IV	<0.001
LOG_REP_Reginol_Node_Positive		1.871
Progesterone_Status	Negative vs Positive	3.725
REP_Age		1.027
REP_Regional_Node_Examined		0.974
T_Stage	T1 vs T4	0.269
T_Stage	T2 vs T4	0.414
T_Stage	T3 vs T4	0.392

NOTE: No (additional) effects met the 0.05 significance level for removal from the model.

- For **Grade**, the odds ratio **(1 vs anaplastic; Grade IV)** estimate equals 0.001. This means that cases with Grade 1 are 99.9% less likely to die of breast cancer than cases with anaplastic Grade 4.

- For **Grade**, the odds ratio **(2 vs anaplastic; Grade IV)** estimate equals 0.001. This means that cases with Grade 2 are 99.9% less likely to die of breast cancer than cases with anaplastic Grade 4.
- For **Grade**, the odds ratio **(3 vs anaplastic; Grade IV)** estimate equals 0.001. This means that cases with Grade 3 are 99.9% less likely to die of breast cancer than cases with anaplastic Grade 4.
- For **LOG_REP_Reginol_Node_Positive**, the odds ratio estimate equals 1.871. This means that every time **LOG_REP_Reginol_Node_Positive** goes up by the factor of 2.74 the probability to die of breast cancer increases by 87.1%.
- For **Progesterone_Status,** the odds ratio estimate equals 3.725. This means that cases with Negative Progesterone are 3.725 times more likely to die of breast cancer that cases with Positive Progesterone.
- For **REP_Age,** the odds ratio estimate equals 1.027. This means that for each additional year, the probability of die of breast cancer increases by 2.7%.
- For **REP_Regional_Node_Examined,** the odds ration estimate equals 0.974. This means that for each additional node examined the probability of die of breast cancer goes down by 97.4%.
- For **T_Stage**, the odds ratio **(T1 vs T4)** estimate equals 0.269. This means that cases with T1 are 73.1% less likely to die of breast cancer than cases with anaplastic T4.
- For **T_Stage**, the odds ratio **(T2 vs T4)** estimate equals 0.414. This means that cases with T2 are 58.6% less likely to die of breast cancer than cases with anaplastic T4.
- For **T_Stage**, the odds ratio **(T3 vs T4)** estimates equals 0.392. This means that cases with T3 are 60.8% less likely to die of breast cancer than cases with anaplastic T4.

Neural Network Model

The Neural Models in the Breast Cancer Model are nine. Seven of these Neural Models are using just the most important variables of the best regression model to mitigate the curse of dimensionality. In this case, the backward regression. The average error was selected as the model selection criterion. Also, the maximum iterations were changed to 100 and the maximum time to 10 minutes.

The other two models were included after the Cap and Floor (NN Cap and Floor) and the Transform Variables (NN Transform Variables) modifications. These models were included to validate if they have lower (ASE) than the Neural Models connected to the backward regression that is the best regression model.

Neural Network: Cap & Floor

The Cap & Floor Neural Network has the model selection criterion Profit/Loss and three hidden units. The Average Square Error is 0.20948 (Appendix – Figure 11).

Neural Network: Transform Variables

The Transform Neural Network has the model selection criterion Profit/Loss and three hidden units. The Average Square Error is 0.211469 (Appendix – Figure 12).

Neural Network: Best Regression (3H)

The Neural Network: Best Regression has the model selection criterion Average Error and three hidden units. The Average Square Error is 0.208826 (Appendix – Figure 13).

Neural Network: 2H

The Neural Network: Best Regression has the model selection criterion Average Error and two hidden units. The Average Square Error is 0.210514 (Appendix – Figure 14).

Neural Network: 4H

The Neural Network: Best Regression has the model selection criterion Average Error and four hidden units. The Average Square Error is 0.210002 (Appendix – Figure 15).

Neural Network: 5H

The Neural Network: Best Regression has the model selection criterion Average Error and five hidden units. The Average Square Error is 0.20793 (Appendix – Figure 16).

Neural Network: 6H

The Neural Network: Best Regression has the model selection criterion Average Error and six hidden units. The Average Square Error is 0.207433 (Appendix – Figure 17).

Neural Network: 7H

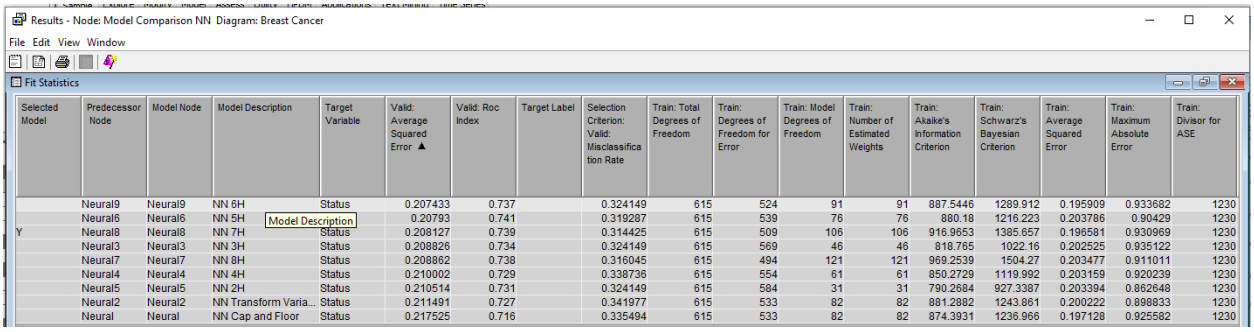
The Neural Network: Best Regression has the model selection criterion Average Error and seven hidden units. The Average Square Error is 0.208127 (Appendix – Figure 18).

Neural Network: 8H

The Neural Network: Best Regression has the model selection criterion Average Error and eight hidden units. The Average Square Error is 0.208862 (Appendix – Figure 19).

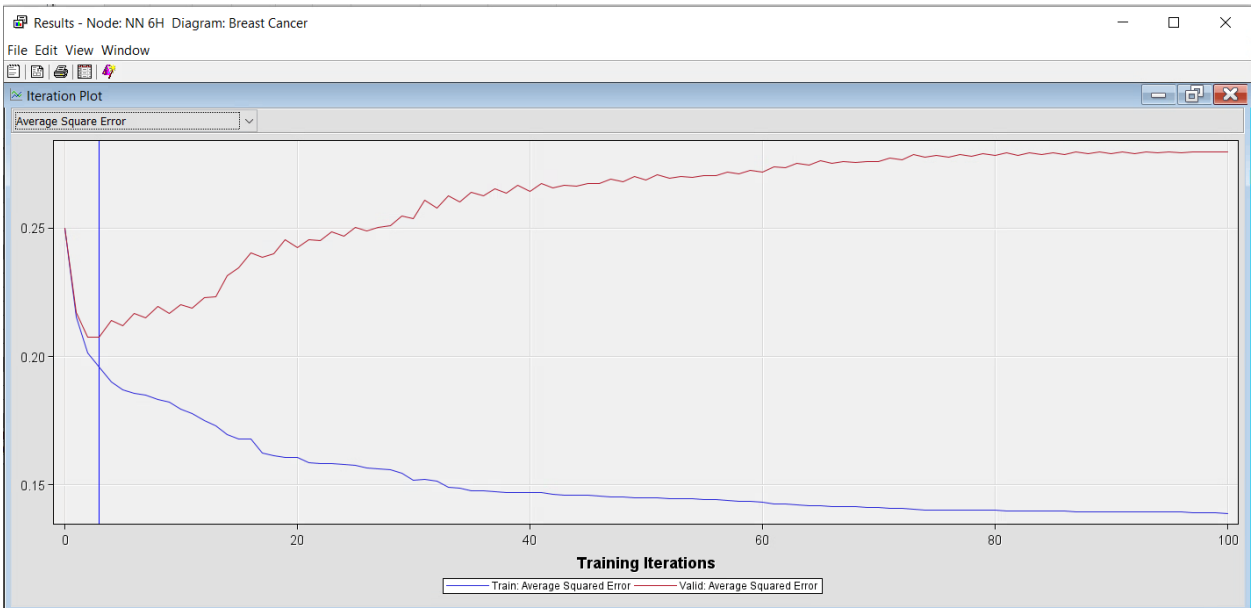
Optimal Neural Network

After running all the Neural Networks models, the best model is the Neural Network using 6 hidden units. The ASE is 0.207433 which is the lowest comparing the other models. Also, this model has a ROC index of 0.737 which is the highest as well. These two parameters indicate that the 6 hidden units neural network model is the best among the neural networks.



Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Valid: Average Squared Error ▲	Valid: Roc Index	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Total Degrees of Freedom	Train: Degrees of Freedom for Error	Train: Model Degrees of Freedom	Train: Number of Estimated Weights	Train: Akaike's Information Criterion	Train: Schwarz's Bayesian Criterion	Train: Average Squared Error	Train: Maximum Absolute Error	Train: Divisor for ASE
	Neural9	Neural9	NN 6H	Status	0.207433	0.737		0.324149	615	524	91	91	887.5446	1289.912	0.195909	0.933682	1230
	Neural6	Neural6	NN 5H	Status	0.20793	0.741		0.319287	615	539	76	76	880.18	1216.223	0.203786	0.90429	1230
Y	Neural8	Neural8	NN 7H	Status	0.208127	0.739		0.314425	615	509	106	106	916.9653	1385.657	0.196581	0.930969	1230
	Neural3	Neural3	NN 3H	Status	0.208826	0.734		0.324149	615	569	46	46	818.765	1022.16	0.202525	0.935122	1230
	Neural7	Neural7	NN 8H	Status	0.208862	0.738		0.316045	615	494	121	121	969.2539	1504.27	0.203477	0.911011	1230
	Neural4	Neural4	NN 4H	Status	0.210002	0.729		0.338736	615	554	61	61	850.2729	1119.992	0.203159	0.920239	1230
	Neural5	Neural5	NN 2H	Status	0.210514	0.731		0.324149	615	584	31	31	790.2684	927.3387	0.203394	0.862648	1230
	Neural2	Neural2	NN Transform Varia...	Status	0.211491	0.727		0.341977	615	533	82	82	881.2882	1243.861	0.200222	0.898833	1230
	Neural	Neural	NN Cap and Floor	Status	0.217525	0.716		0.335494	615	533	82	82	874.3931	1236.966	0.197128	0.925582	1230

The following graph shows that the iteration where with the lowest ASE for the Valid dataset is in the third iteration, that means that the model has converged.



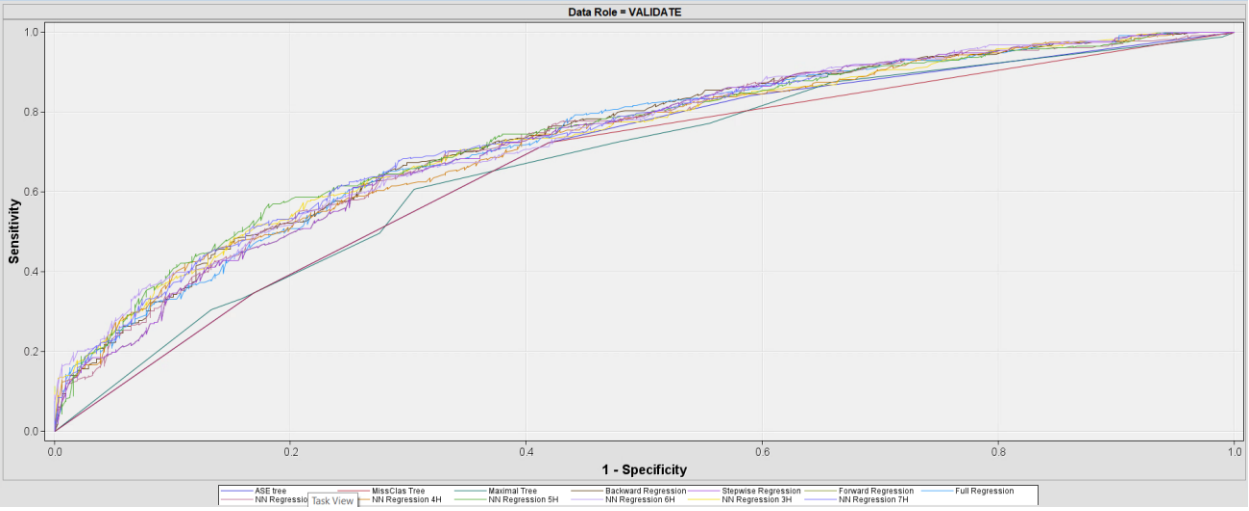
Assessment

The results shows that the best model is the Neural Network after regression with 6 hidden units with an ASE of 0.207433.

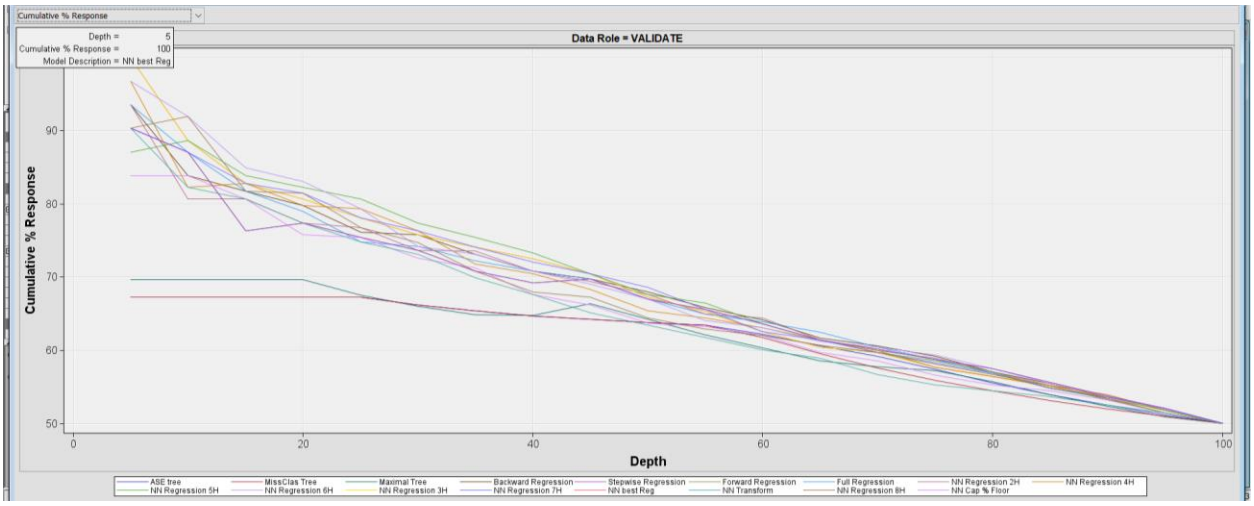
Using ROC as a validation assessment rating the best model is the Neural Network with 5 hidden units with a ROC of 0.741

Model Node	Model Description	Valid: Average Squared Error ▲
Neural6	NN Regression 6H	0.207433
Neural7	NN Regression 5H	0.20793
Neural4	NN Regression 7H	0.208127
Reg4	Backward Regression	0.20826
Neural3	NN best Reg	0.208826
Neural5	NN Regression 3H	0.208826
Neural8	NN Regression 4H	0.210002
Neural9	NN Regression 2H	0.210514
Reg	Full Regression	0.211012
Reg2	Forward Regression	0.211064
Reg3	Stepwise Regression	0.211064
Neural10	NN Regression 8H	0.217083
Neural	NN Cap % Floor	0.224633
Tree3	ASE tree	0.228703
Tree2	MissClas Tree	0.229525
Tree	Maximal Tree	0.236363
Neural2	NN Transform	0.245027

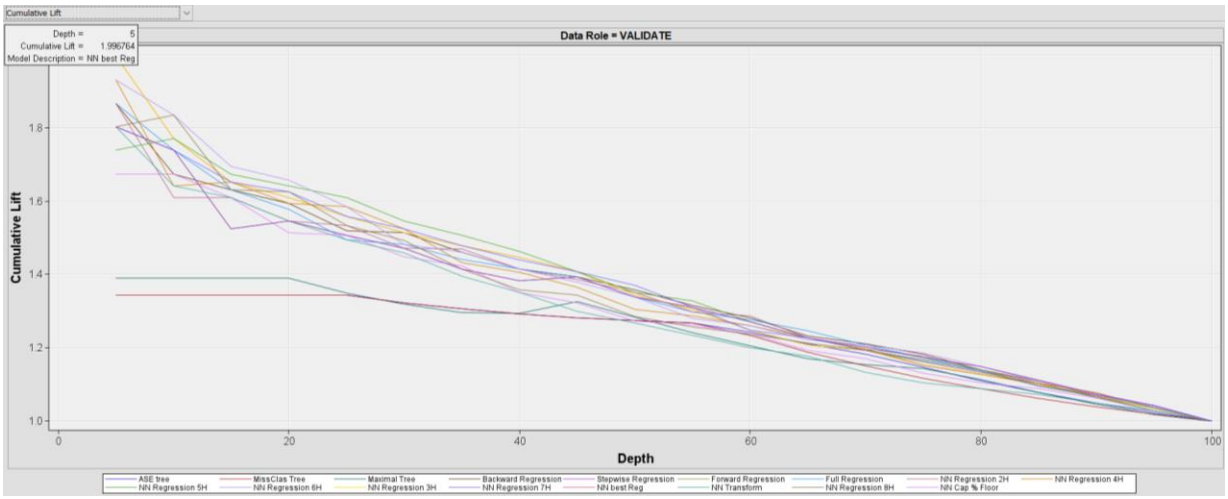
Model Node	Model Description	Valid: Roc Index ▼
Neural7	NN Regression 5H	0.741
Neural4	NN Regression 7H	0.739
Reg4	Backward Regression	0.737
Neural6	NN Regression 6H	0.737
Neural3	NN best Reg	0.734
Neural5	NN Regression 3H	0.734
Neural9	NN Regression 2H	0.731
Reg	Full Regression	0.729
Reg2	Forward Regression	0.729
Reg3	Stepwise Regression	0.729
Neural8	NN Regression 4H	0.729
Neural10	NN Regression 8H	0.719
Neural	NN Cap % Floor	0.708
Tree3	ASE tree	0.674
Tree	Maximal Tree	0.669
Neural2	NN Transform	0.665
Tree2	MissClas Tree	0.663



The cumulative response chart show that the Neural Network with 3 hidden units is the most effective model based on the response rate; choosing the best 5% of the records the response rate of this Neural Network is 100%.



As for the lift, choosing the best 5% of the records the best lift is from the Neural Network with 3 hidden units with a lift value of 1.996764



Conclusion and Recommendations

This project identifies the different characteristics of patients who are more likely to die from breast cancer; to have a consist and reliable model some changes had to be done to the dataset; first, balance the records between dead and alive patients to avoid bias; second, using cap and floor technique to manage the outliers; Finally, transforming the variables to manage properly the skewness of the variables.

After running the different predictive models including decision tree models, logistic regression models, and neural network models,

the backward regression model stands out as the most effective model to identifying individuals that have a higher risk of dying of breast cancer. Although the best model with the ASE criteria is the Neural Network with 6 hidden units, the backward regression was chosen as the best model due to Neural Network cannot be interpretated.

For the backward regression, the variables that are included in the model and that have a notable association with breast cancer mortality are: Grade, LOG_REP_Reginol_Node_Positive, Progesterone_Status, REP_Age, REP_Regional_Node_Examined, and T_Stage.

For the logistic regression, the odds ratio estimates provide a crucial information about the impact of individual variables on the likelihood of death from breast cancer. For the backward regression of the breast cancer model, the odds ratio estimates show:

- For the variable **Grade**, cases with Grades 1,2, and 3 are shown to be 99.9% less likely to die compared to cases with anaplastic Grade4.
- For **LOG_REP_Reginol_Node_Positive**, every time this variable goes up by the factor of 2.74 the probability to die of breast cancer increases by 87.1%.
- For **Progesterone_Status,** cases with Negative Progesterone are 3.725 times more likely to die of breast cancer than cases with Positive Progesterone.
- For **REP_Age,** for each additional year the probability of die of breast cancer increases by 2.7%.
- For **REP_Regional_Node_Examined,** for each additional node examined the probability of die of breast cancer goes down by 97.4%.
- For **T_Stage**, cases with T1 are 73.1% less likely to die of breast cancer than cases with anaplastic T4.
- For **T_Stage**, cases with T2 are 58.6% less likely to die of breast cancer than cases with anaplastic T4.
- For **T_Stage**, cases with T3 are 60.8% less likely to die of breast cancer than cases with anaplastic T4.

Some recommendations for the health industry include focusing on those patients who their progesterone is negative since they have a 272,5% more chance to die than patients with positive progesterone. Also, the number of positive nodes has a major impact on critical patients; the more positive node a patient has, the probability of dead increase in 87,1%.

The cases in anaplastic T4 stage are other important characteristic to consider since these patients have a 26,9% more probability to die. Finally, the age has also an impact with a 2,7% of more chances to die when the age increase.

The study also showed that anaplastic grade 4 has a greater impact in the target comparing with the other categories 1,2 and 3, with 0,1% chance not to die. For further research, it is recommended the use of more techniques for a deeper understanding of this categorical variable.

Appendix

Results - Node: Maximal Tree Diagram: Breast Cancer

File Edit View Window

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
Status		NOBS	Sum of Frequencies	615	617
Status		MISC	Misclassification Rate	0.313821	0.376013
Status		MAX	Maximum Absolute Error	0.875	0.875
Status		SSE	Sum of Squared Errors	246.0806	291.6717
Status		ASE	Average Squared Error	0.200066	0.236363
Status		RASE	Root Average Squared Error	0.447287	0.486172
Status		DIV	Divisor for ASE	1230	1234
Status		DFT	Total Degrees of Freedom	615	

Figure 1. Maximal Tree Results

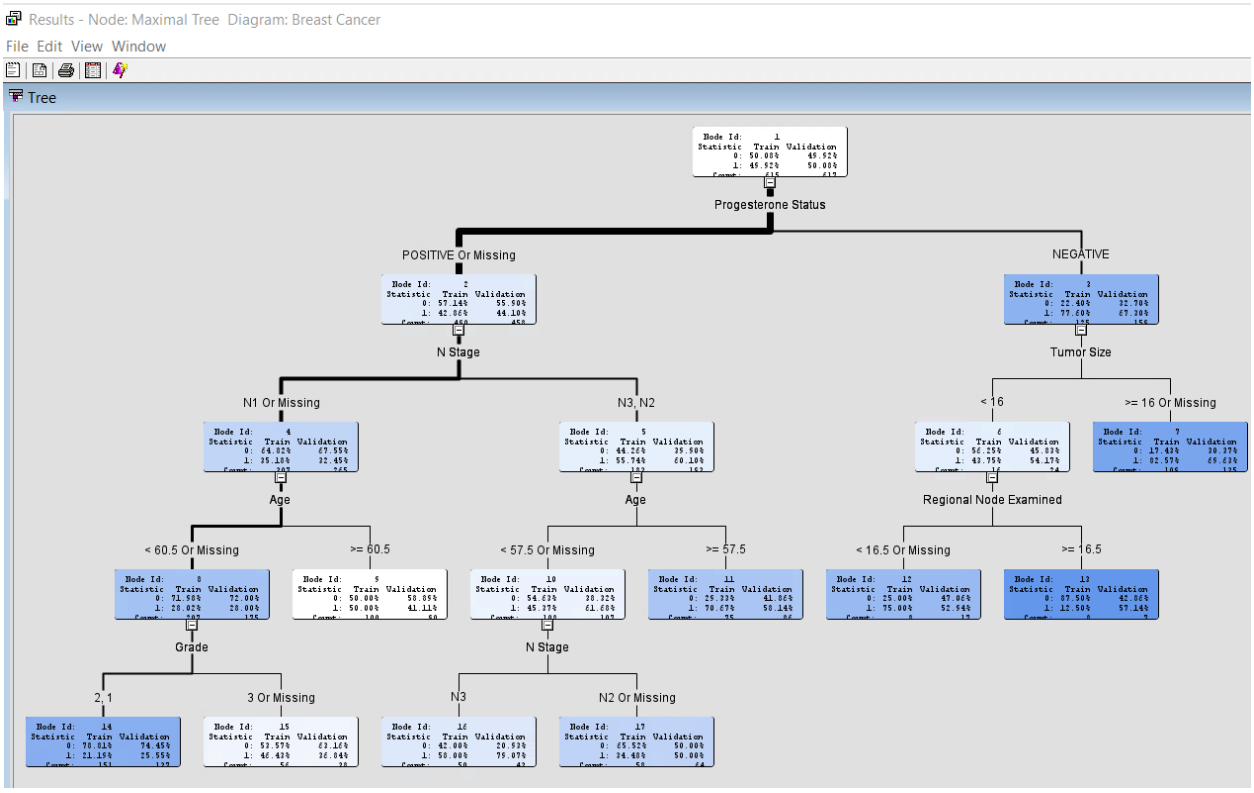


Figure 2. Maximal Tree Diagram

Results - Node: Missclassification Tree Diagram: Breast Cancer

File Edit View Window

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
Status		NOBS	Sum of Frequencies	615	617
Status		MISC	Misclassification Rate	0.352846	0.34846
Status		MAX	Maximum Absolute Error	0.776	0.776
Status		SSE	Sum of Squared Errors	273.7841	283.2344
Status		ASE	Average Squared Error	0.222572	0.229525
Status		RASE	Root Average Squared Error	0.471776	0.479088
Status		DIV	Divisor for ASE	1230	1234
Status		DFT	Total Degrees of Freedom	615	

Figure 3. Misclassification Tree Results

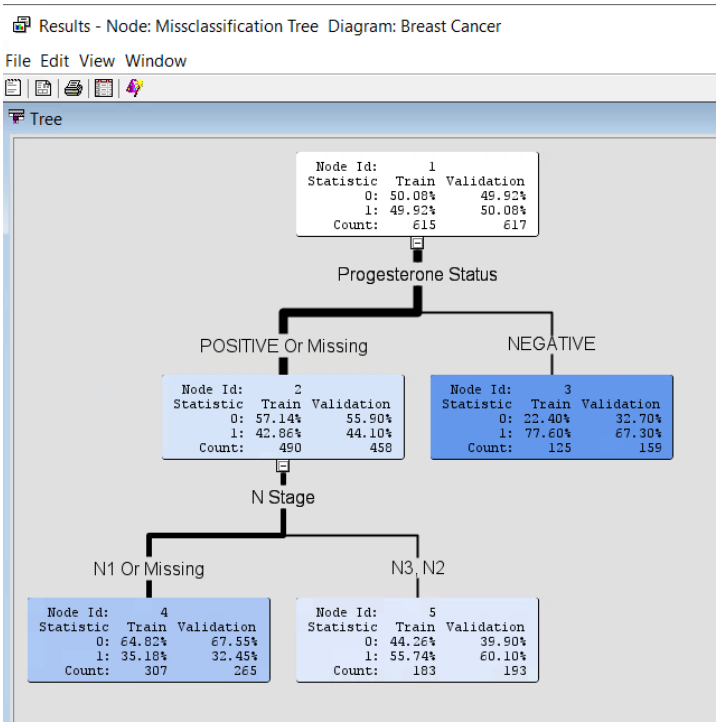
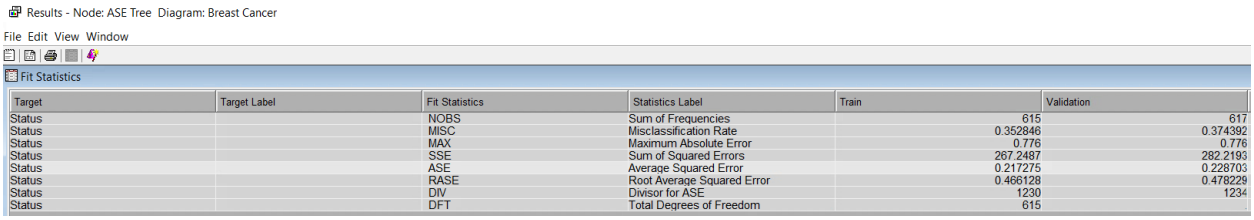


Figure 4. Misclassification Tree Diagram



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
Status		NOBS	Sum of Frequencies	615	617
Status		MISC	Misclassification Rate	0.352846	0.374392
Status		MAX	Maximum Absolute Error	0.776	0.776
Status		SSE	Sum of Squared Errors	267.2487	282.2193
Status		ASE	Average Squared Error	0.217275	0.228703
Status		RASE	Root Average Squared Error	0.466128	0.478229
Status		DIV	Divisor for ASE	1230	1234
Status		DFT	Total Degrees of Freedom	615	

Figure 5. Average Squared Error Tree Results

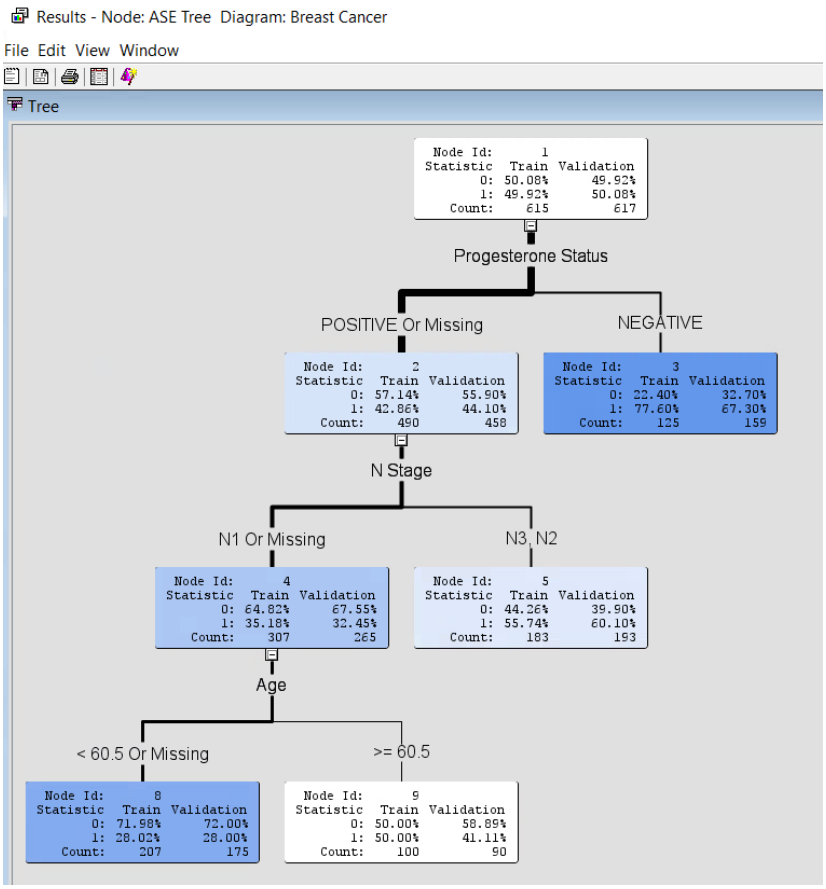
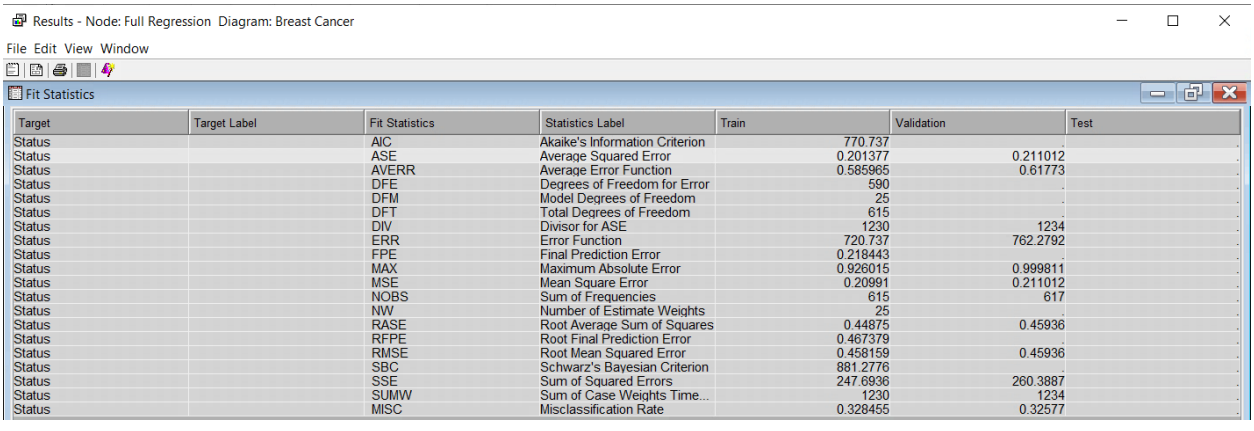
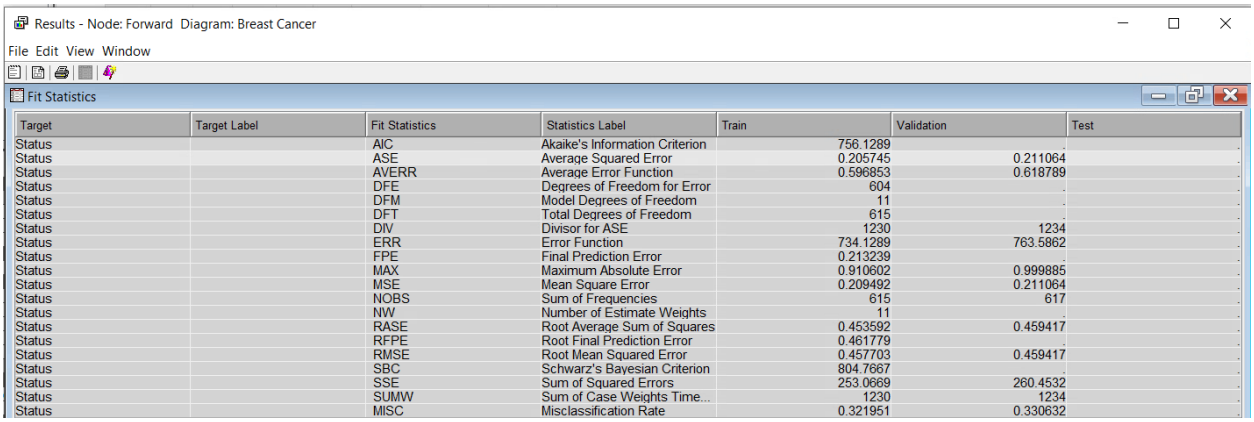


Figure 6. Average Squared Error Tree Diagram



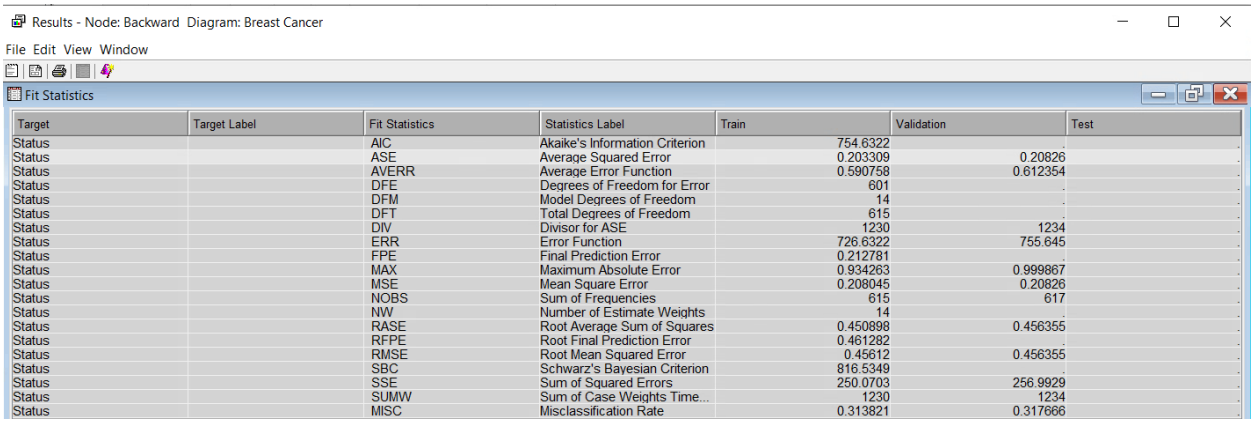
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Status		AIC	Akaike's Information Criterion	770.737		
Status		ASE	Average Squared Error	0.201377	0.211012	
Status		AVERR	Average Error Function	0.585965	0.61773	
Status		DFE	Degrees of Freedom for Error	590		
Status		DFM	Model Degrees of Freedom	25		
Status		DFT	Total Degrees of Freedom	615		
Status		DIV	Divisor for ASE	1230	1234	
Status		ERR	Error Function	720.737	762.2792	
Status		FPE	Final Prediction Error	0.218443		
Status		MAX	Maximum Absolute Error	0.926015	0.999811	
Status		MSE	Mean Square Error	0.20991	0.211012	
Status		NOBS	Sum of Frequencies	615	617	
Status		NW	Number of Estimate Weights	25		
Status		RASE	Root Average Sum of Squares	0.44875	0.45936	
Status		RFPE	Root Final Prediction Error	0.467379		
Status		RMSE	Root Mean Squared Error	0.458159	0.45936	
Status		SBC	Schwarz's Bayesian Criterion	881.2776		
Status		SSE	Sum of Squared Errors	247.6936	260.3887	
Status		SUMW	Sum of Case Weights Time...	1230	1234	
Status		MISC	Misclassification Rate	0.328455	0.32577	

Figure 7. Full Regression Results



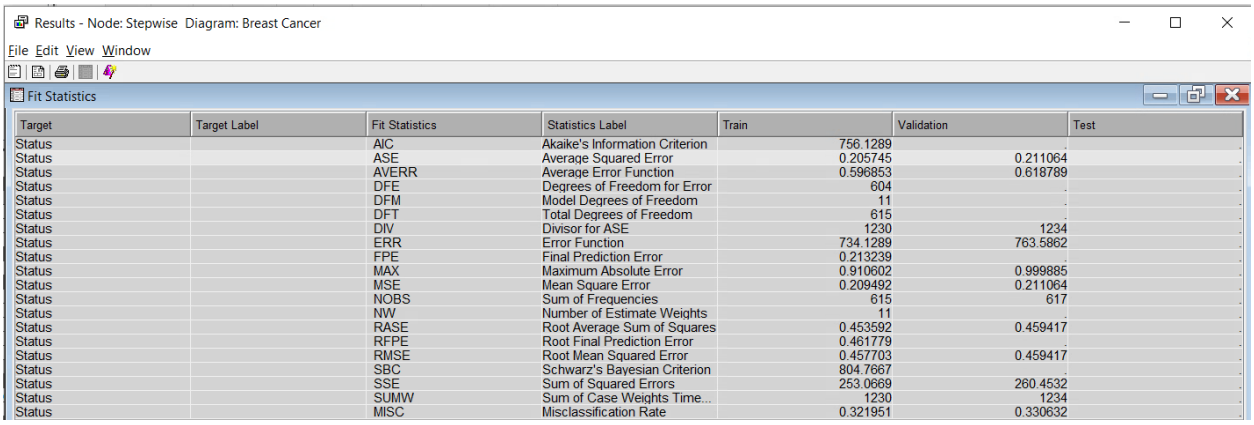
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Status		AIC	Akaike's Information Criterion	756.1289		
Status		ASE	Average Squared Error	0.205745	0.211064	
Status		AVERR	Average Error Function	0.596853	0.618789	
Status		DFE	Degrees of Freedom for Error	604		
Status		DFM	Model Degrees of Freedom	11		
Status		DFT	Total Degrees of Freedom	615		
Status		DIV	Divisor for ASE	1230	1234	
Status		ERR	Error Function	734.1289	763.5862	
Status		FPE	Final Prediction Error	0.213239		
Status		MAX	Maximum Absolute Error	0.910602	0.999885	
Status		MSE	Mean Square Error	0.209492	0.211064	
Status		NOBS	Sum of Frequencies	615	617	
Status		NW	Number of Estimate Weights	11		
Status		RASE	Root Average Sum of Squares	0.453592	0.459417	
Status		RFPE	Root Final Prediction Error	0.461779		
Status		RMSE	Root Mean Squared Error	0.457703	0.459417	
Status		SBC	Schwarz's Bayesian Criterion	804.7667		
Status		SSE	Sum of Squared Errors	253.0669	260.4532	
Status		SUMW	Sum of Case Weights Time...	1230	1234	
Status		MISC	Misclassification Rate	0.321951	0.330632	

Figure 8. Forward Regression Results



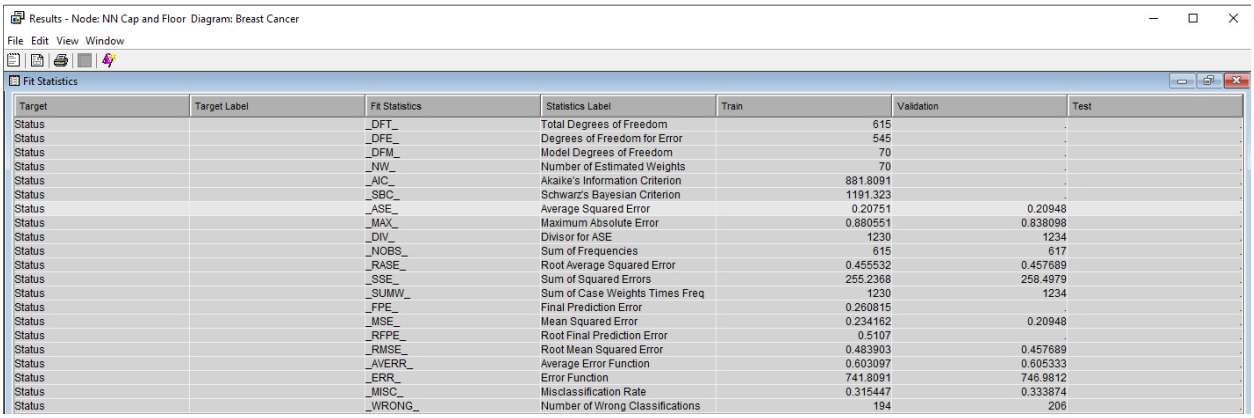
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Status		AIC	Akaike's Information Criterion	754.6322		
Status		ASE	Average Squared Error	0.203309	0.20826	
Status		AVERR	Average Error Function	0.590758	0.612354	
Status		DFE	Degrees of Freedom for Error	601		
Status		DFM	Model Degrees of Freedom	14		
Status		DFT	Total Degrees of Freedom	615		
Status		DIV	Divisor for ASE	1230	1234	
Status		ERR	Error Function	726.6322	755.645	
Status		FPE	Final Prediction Error	0.212781		
Status		MAX	Maximum Absolute Error	0.934263	0.999867	
Status		MSE	Mean Square Error	0.208045	0.20826	
Status		NOBS	Sum of Frequencies	615	617	
Status		NW	Number of Estimate Weights	14		
Status		RASE	Root Average Sum of Squares	0.450898	0.456355	
Status		RFPE	Root Final Prediction Error	0.461282		
Status		RMSE	Root Mean Squared Error	0.45612	0.456355	
Status		SBC	Schwarz's Bayesian Criterion	816.5349		
Status		SSE	Sum of Squared Errors	250.0703	256.9929	
Status		SUMW	Sum of Case Weights Time...	1230	1234	
Status		MISC	Misclassification Rate	0.313821	0.317666	

Figure 9. Backward Regression Results



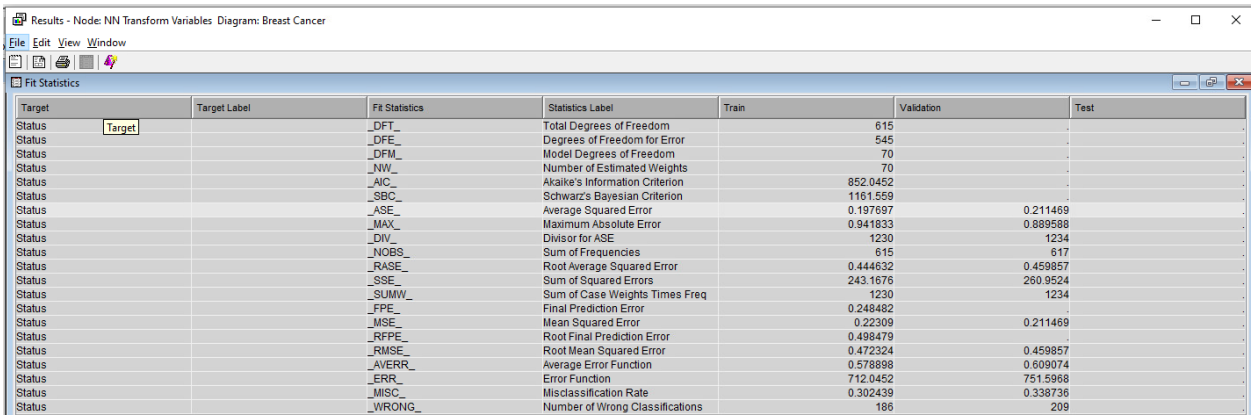
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Status		AIC	Akaike's Information Criterion	756.1289		
Status		ASE	Average Squared Error	0.205745	0.211064	
Status		AVERR	Average Error Function	0.596853	0.618789	
Status		DFE	Degrees of Freedom for Error	604		
Status		DFM	Model Degrees of Freedom	11		
Status		DFT	Total Degrees of Freedom	615		
Status		DIV	Divisor for ASE	1230	1234	
Status		ERR	Error Function	734.1289	763.5862	
Status		FPE	Final Prediction Error	0.213239		
Status		MAX	Maximum Absolute Error	0.910602	0.999885	
Status		MSE	Mean Square Error	0.209492	0.211064	
Status		NOBS	Sum of Frequencies	615	617	
Status		NW	Number of Estimate Weights	11		
Status		RASE	Root Average Sum of Squares	0.453592	0.459417	
Status		RFPE	Root Final Prediction Error	0.461779		
Status		RMSE	Root Mean Squared Error	0.457703	0.459417	
Status		SBC	Schwarz's Bayesian Criterion	804.7667		
Status		SSE	Sum of Squared Errors	253.0669	260.4532	
Status		SUMW	Sum of Case Weights Time...	1230	1234	
Status		MISC	Misclassification Rate	0.321951	0.330632	

Figure 10. Stepwise Regression Results



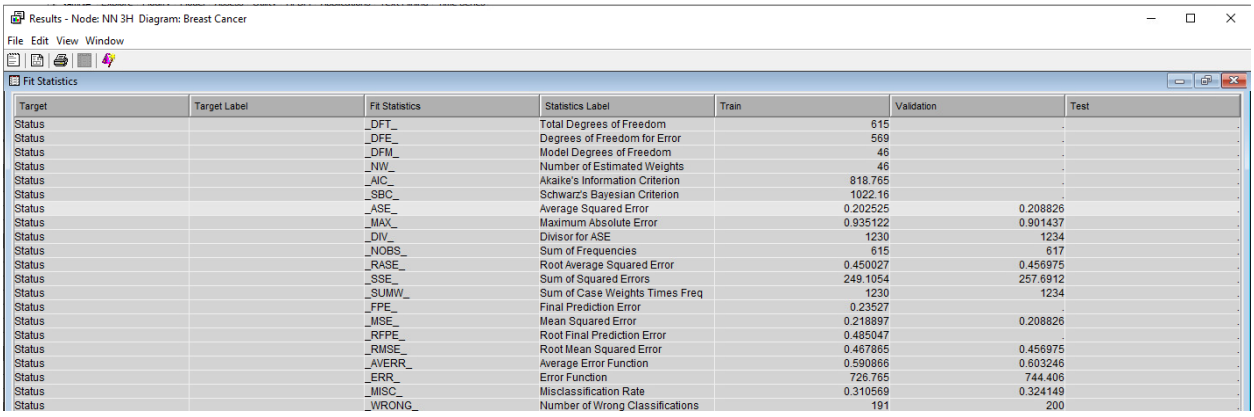
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Status		_DFT_	Total Degrees of Freedom	615	.	.
Status		_DFE_	Degrees of Freedom for Error	545	.	.
Status		_DFM_	Model Degrees of Freedom	70	.	.
Status		_NW_	Number of Estimated Weights	70	.	.
Status		_AIC_	Akaike's Information Criterion	881.8091	.	.
Status		_SBC_	Schwarz's Bayesian Criterion	1191.323	.	.
Status		_ASE_	Average Squared Error	0.20751	0.20948	.
Status		_MAX_	Maximum Absolute Error	0.880551	0.838098	.
Status		_DIV_	Divisor for ASE	1230	1234	.
Status		_NOBS_	Sum of Frequencies	615	617	.
Status		_RASE_	Root Average Squared Error	0.455532	0.457689	.
Status		_SSE_	Sum of Squared Errors	255.2368	258.4979	.
Status		_SUMW_	Sum of Case Weights Times Freq	1230	1234	.
Status		_FPE_	Final Prediction Error	0.260815	.	.
Status		_MSE_	Mean Squared Error	0.234162	0.20948	.
Status		_RFPE_	Root Final Prediction Error	0.5107	.	.
Status		_RMSE_	Root Mean Squared Error	0.483903	0.457689	.
Status		_AVERR_	Average Error Function	0.603097	0.605333	.
Status		_ERR_	Error Function	741.8091	746.9812	.
Status		_MISC_	Misclassification Rate	0.315447	0.333874	.
Status		_WRONG_	Number of Wrong Classifications	194	206	.

Figure 11. Neural Network Cap and Floor Results



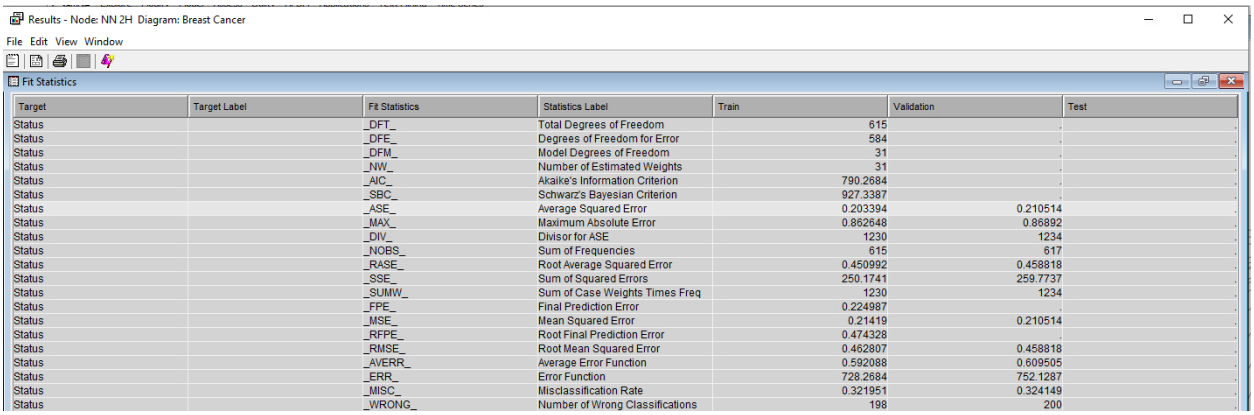
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Status		_DFT_	Total Degrees of Freedom	615	.	.
Status		_DFE_	Degrees of Freedom for Error	545	.	.
Status		_DFM_	Model Degrees of Freedom	70	.	.
Status		_NW_	Number of Estimated Weights	70	.	.
Status		_AIC_	Akaike's Information Criterion	852.0452	.	.
Status		_SBC_	Schwarz's Bayesian Criterion	1161.559	.	.
Status		_ASE_	Average Squared Error	0.197697	0.211469	.
Status		_MAX_	Maximum Absolute Error	0.941833	0.889588	.
Status		_DIV_	Divisor for ASE	1230	1234	.
Status		_NOBS_	Sum of Frequencies	615	617	.
Status		_RASE_	Root Average Squared Error	0.444632	0.459857	.
Status		_SSE_	Sum of Squared Errors	243.1676	260.9524	.
Status		_SUMW_	Sum of Case Weights Times Freq	1230	1234	.
Status		_FPE_	Final Prediction Error	0.248482	.	.
Status		_MSE_	Mean Squared Error	0.22309	0.211469	.
Status		_RFPE_	Root Final Prediction Error	0.498479	.	.
Status		_RMSE_	Root Mean Squared Error	0.472324	0.459857	.
Status		_AVERR_	Average Error Function	0.578898	0.609074	.
Status		_ERR_	Error Function	712.0452	751.5968	.
Status		_MISC_	Misclassification Rate	0.302439	0.338736	.
Status		_WRONG_	Number of Wrong Classifications	186	209	.

Figure 12. Neural Network Transform Variables Results



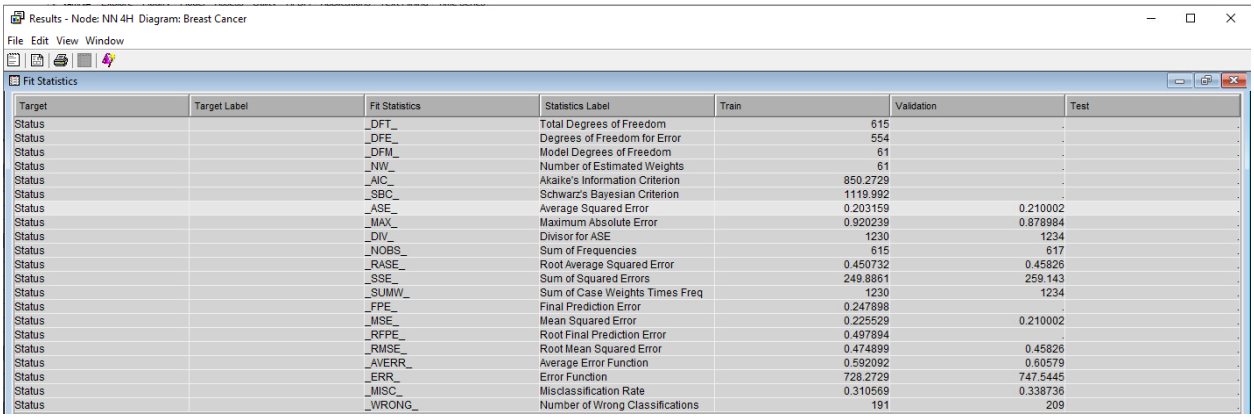
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Status		_DFT_	Total Degrees of Freedom	615	.	.
Status		_DFE_	Degrees of Freedom for Error	569	.	.
Status		_DFM_	Model Degrees of Freedom	46	.	.
Status		_NW_	Number of Estimated Weights	46	.	.
Status		_AIC_	Akaike's Information Criterion	818.765	.	.
Status		_SBC_	Schwarz's Bayesian Criterion	1022.16	.	.
Status		_ASE_	Average Squared Error	0.202525	0.208826	.
Status		_MAX_	Maximum Absolute Error	0.935122	0.901437	.
Status		_DIV_	Divisor for ASE	1230	1234	.
Status		_NOBS_	Sum of Frequencies	615	617	.
Status		_RASE_	Root Average Squared Error	0.450027	0.456975	.
Status		_SSE_	Sum of Squared Errors	249.1054	257.6912	.
Status		_SUMW_	Sum of Case Weights Times Freq	1230	1234	.
Status		_FPE_	Final Prediction Error	0.23527	.	.
Status		_MSE_	Mean Squared Error	0.218897	0.208826	.
Status		_RFPE_	Root Final Prediction Error	0.485047	.	.
Status		_RMSE_	Root Mean Squared Error	0.467865	0.456975	.
Status		_AVERR_	Average Error Function	0.590866	0.603246	.
Status		_ERR_	Error Function	726.765	744.406	.
Status		_MISC_	Misclassification Rate	0.310569	0.324149	.
Status		_WRONG_	Number of Wrong Classifications	191	200	.

Figure 13. Neural Network 3 Hidden Units



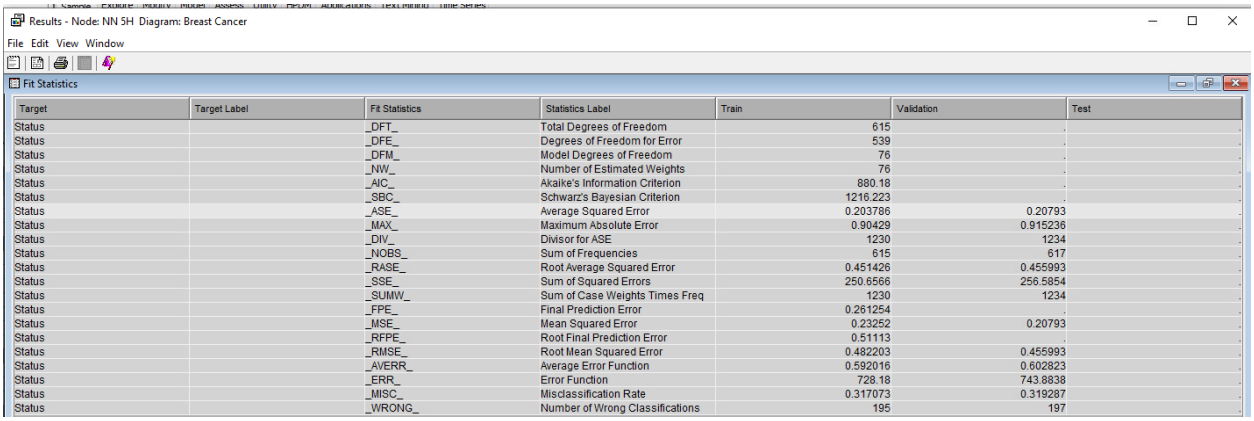
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Status		_DFT_	Total Degrees of Freedom	615		
Status		_DFE_	Degrees of Freedom for Error	584		
Status		_DFM_	Model Degrees of Freedom	31		
Status		_NW_	Number of Estimated Weights	31		
Status		_AIC_	Akaike's Information Criterion	790.2684		
Status		_SBC_	Schwarz's Bayesian Criterion	927.3387		
Status		_ASE_	Average Squared Error	0.203394	0.210514	
Status		_MAX_	Maximum Absolute Error	0.862648	0.86892	
Status		_DIV_	Divisor for ASE	1230	1234	
Status		_NOBS_	Sum of Frequencies	615	617	
Status		_RASE_	Root Average Squared Error	0.450992	0.458818	
Status		_SSE_	Sum of Squared Errors	250.1741	259.7737	
Status		_SUMW_	Sum of Case Weights Times Freq	1230	1234	
Status		_FPE_	Final Prediction Error	0.224987		
Status		_MSE_	Mean Squared Error	0.21419	0.210514	
Status		_RFPE_	Root Final Prediction Error	0.474328		
Status		_RMSE_	Root Mean Squared Error	0.462807	0.458818	
Status		_AVERR_	Average Error Function	0.592088	0.609505	
Status		_ERR_	Error Function	728.2684	752.1287	
Status		_MISC_	Misclassification Rate	0.321951	0.324149	
Status		_WRONG_	Number of Wrong Classifications	198	200	

Figure 14. Neural Network 2 Hidden Units



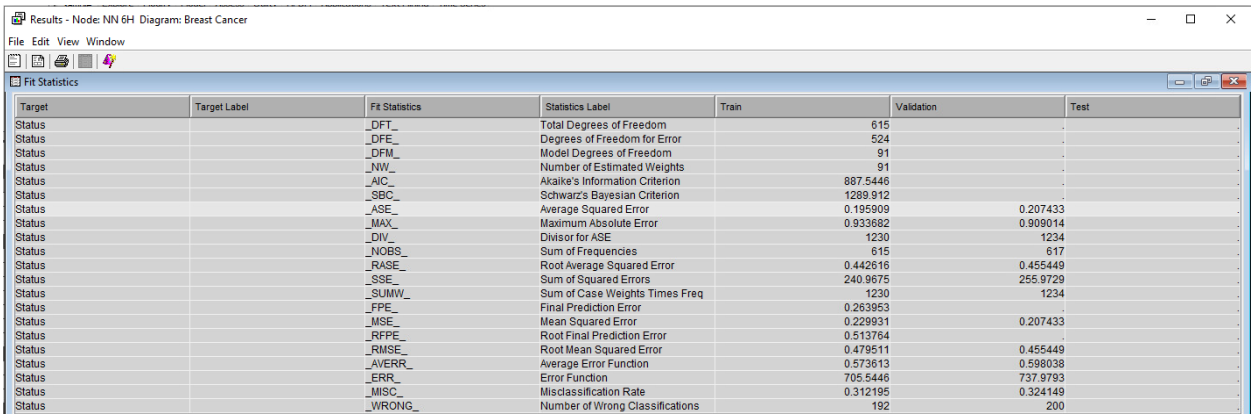
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Status		_DFT_	Total Degrees of Freedom	615		
Status		_DFE_	Degrees of Freedom for Error	554		
Status		_DFM_	Model Degrees of Freedom	61		
Status		_NW_	Number of Estimated Weights	61		
Status		_AIC_	Akaike's Information Criterion	850.2729		
Status		_SBC_	Schwarz's Bayesian Criterion	1119.992		
Status		_ASE_	Average Squared Error	0.203159	0.210002	
Status		_MAX_	Maximum Absolute Error	0.920239	0.878984	
Status		_DIV_	Divisor for ASE	1230	1234	
Status		_NOBS_	Sum of Frequencies	615	617	
Status		_RASE_	Root Average Squared Error	0.450732	0.45826	
Status		_SSE_	Sum of Squared Errors	249.8861	259.143	
Status		_SUMW_	Sum of Case Weights Times Freq	1230	1234	
Status		_FPE_	Final Prediction Error	0.247898		
Status		_MSE_	Mean Squared Error	0.225529	0.210002	
Status		_RFPE_	Root Final Prediction Error	0.497894		
Status		_RMSE_	Root Mean Squared Error	0.474899	0.45826	
Status		_AVERR_	Average Error Function	0.592092	0.60579	
Status		_ERR_	Error Function	728.2729	747.5445	
Status		_MISC_	Misclassification Rate	0.310569	0.338736	
Status		_WRONG_	Number of Wrong Classifications	191	209	

Figure 15. Neural Network 4 Hidden Units



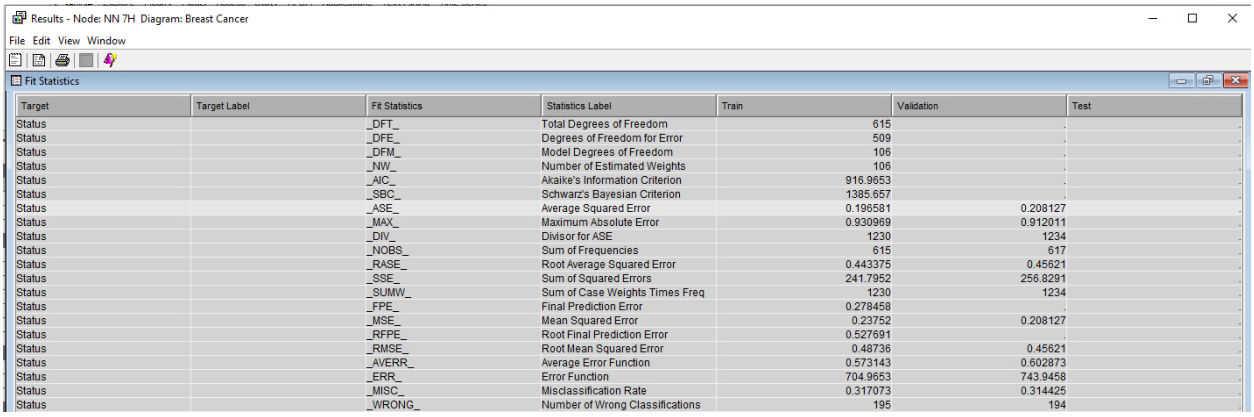
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Status		_DFT_	Total Degrees of Freedom	615		
Status		_DFE_	Degrees of Freedom for Error	539		
Status		_DFM_	Model Degrees of Freedom	76		
Status		_NW_	Number of Estimated Weights	76		
Status		_AIC_	Akaike's Information Criterion	880.18		
Status		_SBC_	Schwarz's Bayesian Criterion	1216.223		
Status		_ASE_	Average Squared Error	0.203786	0.20793	
Status		_MAX_	Maximum Absolute Error	0.90429	0.915236	
Status		_DIV_	Divisor for ASE	1230	1234	
Status		_NOBS_	Sum of Frequencies	615	617	
Status		_RASE_	Root Average Squared Error	0.451426	0.455993	
Status		_SSE_	Sum of Squared Errors	250.6566	256.5854	
Status		_SUMW_	Sum of Case Weights Times Freq	1230	1234	
Status		_FPE_	Final Prediction Error	0.261254		
Status		_MSE_	Mean Squared Error	0.23252	0.20793	
Status		_RFPE_	Root Final Prediction Error	0.51113		
Status		_RMSE_	Root Mean Squared Error	0.482203	0.455993	
Status		_AVERR_	Average Error Function	0.592016	0.602823	
Status		_ERR_	Error Function	728.18	743.8838	
Status		_MISC_	Misclassification Rate	0.317073	0.319287	
Status		_WRONG_	Number of Wrong Classifications	195	197	

Figure 16. Neural Network 5 Hidden Units



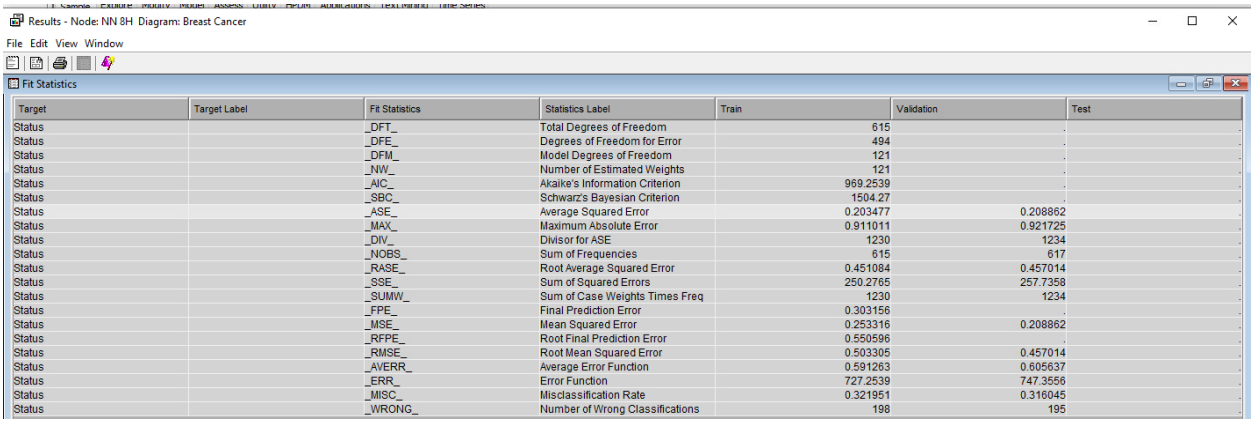
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Status		_DFT_	Total Degrees of Freedom	615	.	.
Status		_DFE_	Degrees of Freedom for Error	524	.	.
Status		_DFM_	Model Degrees of Freedom	91	.	.
Status		_NW_	Number of Estimated Weights	91	.	.
Status		_AIC_	Akaike's Information Criterion	887.5446	.	.
Status		_SBC_	Schwarz's Bayesian Criterion	1289.912	.	.
Status		_ASE_	Average Squared Error	0.195909	0.207433	.
Status		_MAX_	Maximum Absolute Error	0.933682	0.909014	.
Status		_DIV_	Divisor for ASE	1230	1234	.
Status		_NOBS_	Sum of Frequencies	615	617	.
Status		_RASE_	Root Average Squared Error	0.442616	0.455449	.
Status		_SSE_	Sum of Squared Errors	240.9675	255.9729	.
Status		_SUMW_	Sum of Case Weights Times Freq	1230	1234	.
Status		_FPE_	Final Prediction Error	0.263953	.	.
Status		_MSE_	Mean Squared Error	0.229931	0.207433	.
Status		_RFPE_	Root Final Prediction Error	0.513764	.	.
Status		_RMSE_	Root Mean Squared Error	0.479511	0.455449	.
Status		_AVERR_	Average Error Function	0.573613	0.598038	.
Status		_ERR_	Error Function	705.5446	737.9793	.
Status		_MISC_	Misclassification Rate	0.312195	0.324149	.
Status		_WRONG_	Number of Wrong Classifications	192	200	.

Figure 17. Neural Network 6 Hidden Units



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Status		_DFT_	Total Degrees of Freedom	615	.	.
Status		_DFE_	Degrees of Freedom for Error	509	.	.
Status		_DFM_	Model Degrees of Freedom	106	.	.
Status		_NW_	Number of Estimated Weights	106	.	.
Status		_AIC_	Akaike's Information Criterion	916.9653	.	.
Status		_SBC_	Schwarz's Bayesian Criterion	1385.657	.	.
Status		_ASE_	Average Squared Error	0.196581	0.208127	.
Status		_MAX_	Maximum Absolute Error	0.930969	0.912011	.
Status		_DIV_	Divisor for ASE	1230	1234	.
Status		_NOBS_	Sum of Frequencies	615	617	.
Status		_RASE_	Root Average Squared Error	0.443375	0.45621	.
Status		_SSE_	Sum of Squared Errors	241.7952	256.8291	.
Status		_SUMW_	Sum of Case Weights Times Freq	1230	1234	.
Status		_FPE_	Final Prediction Error	0.278458	.	.
Status		_MSE_	Mean Squared Error	0.23752	0.208127	.
Status		_RFPE_	Root Final Prediction Error	0.527691	.	.
Status		_RMSE_	Root Mean Squared Error	0.48736	0.45621	.
Status		_AVERR_	Average Error Function	0.573143	0.602873	.
Status		_ERR_	Error Function	704.9653	743.9458	.
Status		_MISC_	Misclassification Rate	0.317073	0.314425	.
Status		_WRONG_	Number of Wrong Classifications	195	194	.

Figure 18. Neural Network 7 Hidden Units



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Status		_DFT_	Total Degrees of Freedom	615	.	.
Status		_DFE_	Degrees of Freedom for Error	494	.	.
Status		_DFM_	Model Degrees of Freedom	121	.	.
Status		_NW_	Number of Estimated Weights	121	.	.
Status		_AIC_	Akaike's Information Criterion	969.2539	.	.
Status		_SBC_	Schwarz's Bayesian Criterion	1504.27	.	.
Status		_ASE_	Average Squared Error	0.203477	0.208862	.
Status		_MAX_	Maximum Absolute Error	0.911011	0.921725	.
Status		_DIV_	Divisor for ASE	1230	1234	.
Status		_NOBS_	Sum of Frequencies	615	617	.
Status		_RASE_	Root Average Squared Error	0.451084	0.457014	.
Status		_SSE_	Sum of Squared Errors	250.2765	257.7358	.
Status		_SUMW_	Sum of Case Weights Times Freq	1230	1234	.
Status		_FPE_	Final Prediction Error	0.303156	.	.
Status		_MSE_	Mean Squared Error	0.253316	0.208862	.
Status		_RFPE_	Root Final Prediction Error	0.550596	.	.
Status		_RMSE_	Root Mean Squared Error	0.503305	0.457014	.
Status		_AVERR_	Average Error Function	0.591263	0.605637	.
Status		_ERR_	Error Function	727.2539	747.3556	.
Status		_MISC_	Misclassification Rate	0.321951	0.316045	.
Status		_WRONG_	Number of Wrong Classifications	198	195	.

Figure 19. Neural Network 8 Hidden Units