BAN210 – Final Assessment

Using Predictive Modeling on the Breast Cancer Dataset in SAS Enterprise Miner

April Paola Tolosa

Prof. Uzair Ahmad
Predictive Analytics BAN210-ZBB

Date: April 15, 2022

Table of Contents

Academic Integrity Policy

SENECA COLLEGE OF APPLIED ARTS AND TECHNOLOGY
SENECA BUSINESS

Academic Integrity Policy. Seneca upholds a learning community that values academic integrity, honesty, fairness, trust, respect, responsibility and courage. These values enhance Seneca's commitment to students by delivering high-quality education and teaching excellence, while supporting a positive learning environment. The AI policy is always in effect. Note **Sections 2.3 and 2.4:**

*"…**2.3** Should there be a suspected violation of this policy (e.g.…cheating, falsification, impersonation or plagiarism), the academic integrity sanctions will be applied according to the severity of the offence committed. Refer to Appendix B for the academic integrity sanctions. **2.4** Should a suspected violation of this policy be a result of, or in combination with, a suspected violation of Seneca's Student Code of Conduct and/or another non-academic-related Seneca policy, the matter will be investigated and adjudicated through the process found in the Student Code of Conduct."*

TO BE COMPLETED BY STUDENT
SUBJECT SECTION NUMBER (e.g. QNM223 AA): BAN210 ZBB
STUDENT NAME: April Tolosa
STUDENT NUMBER: *131785214*
STUDENT SIGNATURE: April Tolosa

I, **April Tolosa**, declare that the attached assignment is my own work in accordance with the Seneca Academic Policy. I have not copied any part of this assignment, manually or electronically, from any other source including web sites, unless specified as references. I have not distributed my work to other students.

Introduction

In this assessment paper, the breast cancer data set will be analyzed using exploratory and modeling techniques using SAS Enterprise Miner. The goal of this paper is to show the steps and the powerful insights to be gained from the dataset using simple statistical analysis. Steps will be explained as well as the results to show ability to use the tool and to demonstrate data analysis skills.

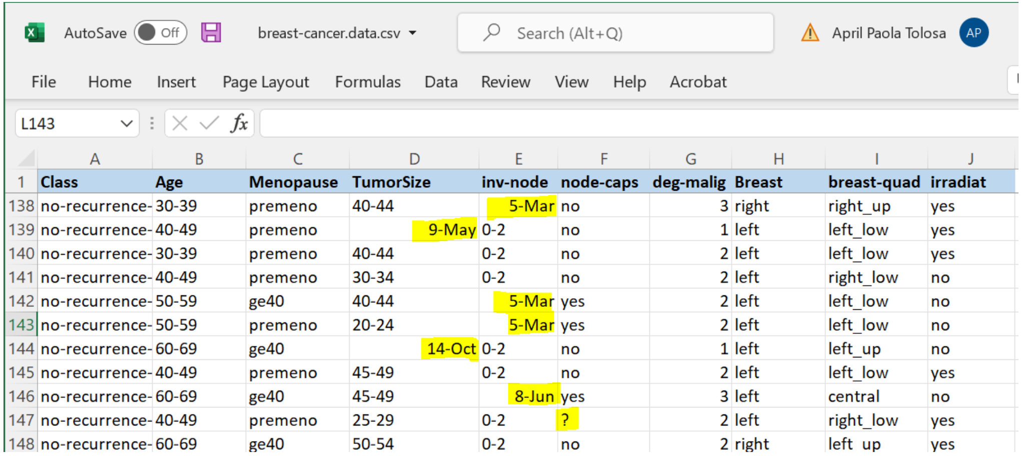
About the Data

The Breast Cancer data set was obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. It includes 201 instances of one class and 85 instances of another class. The instances are described by 10 attributes, some of which are linear, and some are nominal.

Data set Characteristics:	Multivariate	**Number of Instances:**	286	**Area:**	Life
Attribute Characteristics:	Categorical	**Number of Attributes:**	10	**Date Donated**	1988-07-11
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	622435

Attribute Information:
1. Class: no-recurrence-events, recurrence-events
2. age: 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99.
3. menopause: lt40, ge40, premeno.
4. tumor-size: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59.
5. inv-nodes: 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39.
6. node-caps: yes, no.
7. deg-malig: 1, 2, 3.
8. breast: left, right.
9. breast-quad: left-up, left-low, right-up, right-low, central.
10. irradiat: yes, no.

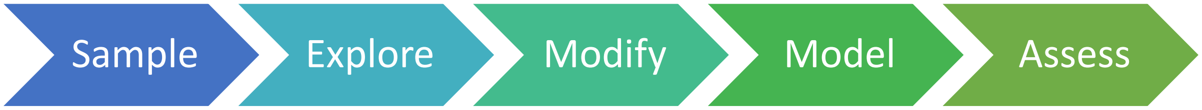
Before analysis, labels were added on the top row and the data cleaned as there were conversion errors from number to dates when the dataset was changed to a .csv file:



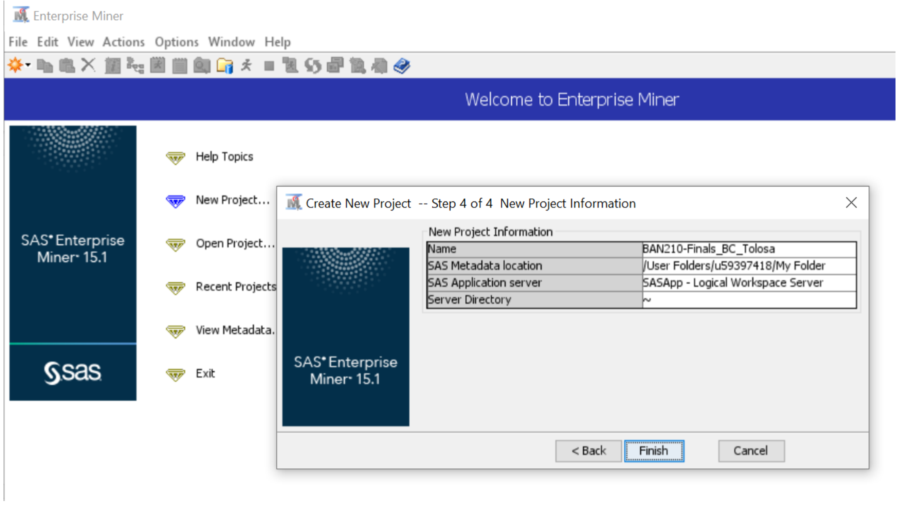
	A	B	C	D	E	F	G	H	I	J
1	**Class**	**Age**	**Menopause**	**TumorSize**	**inv-node**	**node-caps**	**deg-malig**	**Breast**	**breast-quad**	**irradiat**
138	no-recurrence-	30-39	premeno	40-44	5-Mar	no	3	right	right_up	yes
139	no-recurrence-	40-49	premeno	9-May	0-2	no	1	left	left_low	yes
140	no-recurrence-	30-39	premeno	40-44	0-2	no	2	left	left_low	yes
141	no-recurrence-	40-49	premeno	30-34	0-2	no	2	left	right_low	no
142	no-recurrence-	50-59	ge40	40-44	5-Mar	yes	2	left	left_low	no
143	no-recurrence-	50-59	premeno	20-24	5-Mar	yes	2	left	left_low	no
144	no-recurrence-	60-69	ge40	14-Oct	0-2	no	1	left	left_up	no
145	no-recurrence-	40-49	premeno	45-49	0-2	no	2	left	left_low	yes
146	no-recurrence-	60-69	ge40	45-49	8-Jun	yes	3	left	central	no
147	no-recurrence-	40-49	premeno	25-29	0-2	?	2	left	right_low	yes
148	no-recurrence-	60-69	ge40	50-54	0-2	no	2	right	left_up	yes

Analysis

This section will demonstrate how the dataset is analyzed using SAS Miner guided by the data mining process learned from this course



Create a New Project on SAS Enterprise Miner

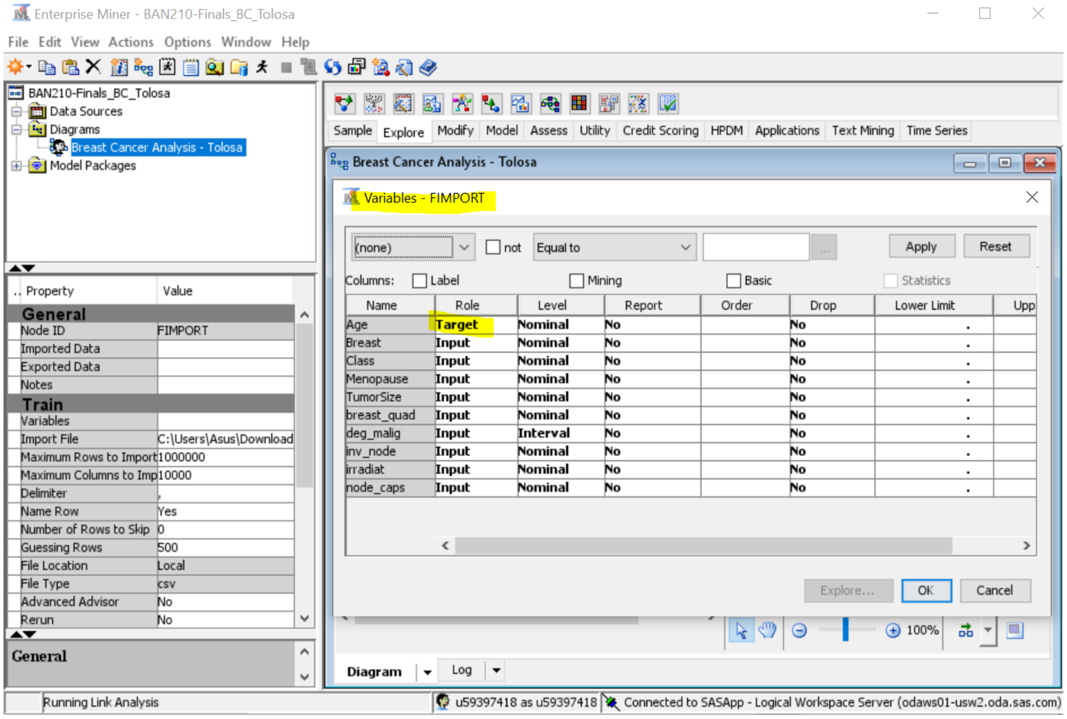


Create New Project -- Step 4 of 4 New Project Information

New Project Information

Name	BAN210-Finals_BC_Tolosa
SAS Metadata location	/User Folders/u59397418/My Folder
SAS Application server	SASApp - Logical Workspace Server
Server Directory	~

Import the file using **File Import** node.
Right click on the File Import node and click Variables to open the Variables window.
Set "Age" as the Target Variable.



Property	Value
General	
Node ID	FIMPORT
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Import File	C:\Users\Asus\Download
Maximum Rows to Import	1000000
Maximum Columns to Imp	10000
Delimiter	,
Name Row	Yes
Number of Rows to Skip	0
Guessing Rows	500
File Location	Local
File Type	csv
Advanced Advisor	No
Rerun	No

(none) □ not Equal to

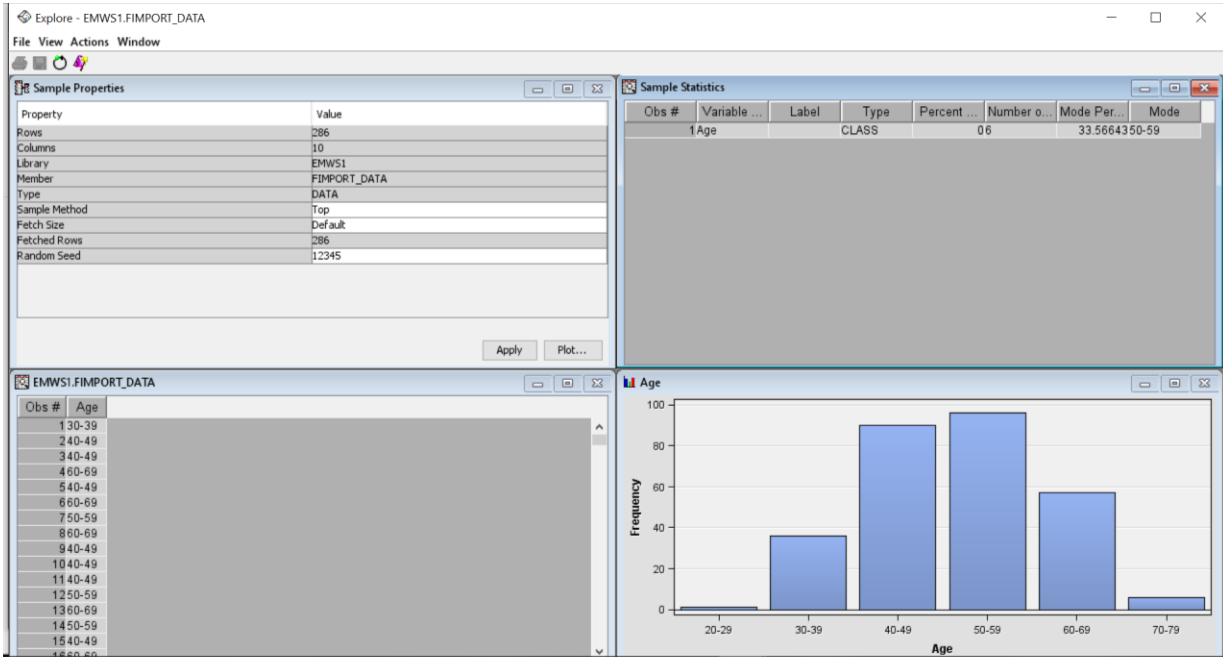
Columns: □ Label □ Mining □ Basic □ Statistics

Name	Role	Level	Report	Order	Drop	Lower Limit	Upp
Age	**Target**	**Nominal**	**No**		**No**	.	
Breast	**Input**	**Nominal**	**No**		**No**	.	
Class	**Input**	**Nominal**	**No**		**No**	.	
Menopause	**Input**	**Nominal**	**No**		**No**	.	
TumorSize	**Input**	**Nominal**	**No**		**No**	.	
breast_quad	**Input**	**Nominal**	**No**		**No**	.	
deg_malig	**Input**	**Interval**	**No**		**No**	.	
inv_node	**Input**	**Nominal**	**No**		**No**	.	
irradiat	**Input**	**Nominal**	**No**		**No**	.	
node_caps	**Input**	**Nominal**	**No**		**No**	.	

Running Link Analysis | u59397418 as u59397418 | Connected to SASApp - Logical Workspace Server (odaws01-usw2.oda.sas.com)

In this window we can click Explore to know more about the data of each variable.

For example, for the Age variable, below is a screenshot of the data:



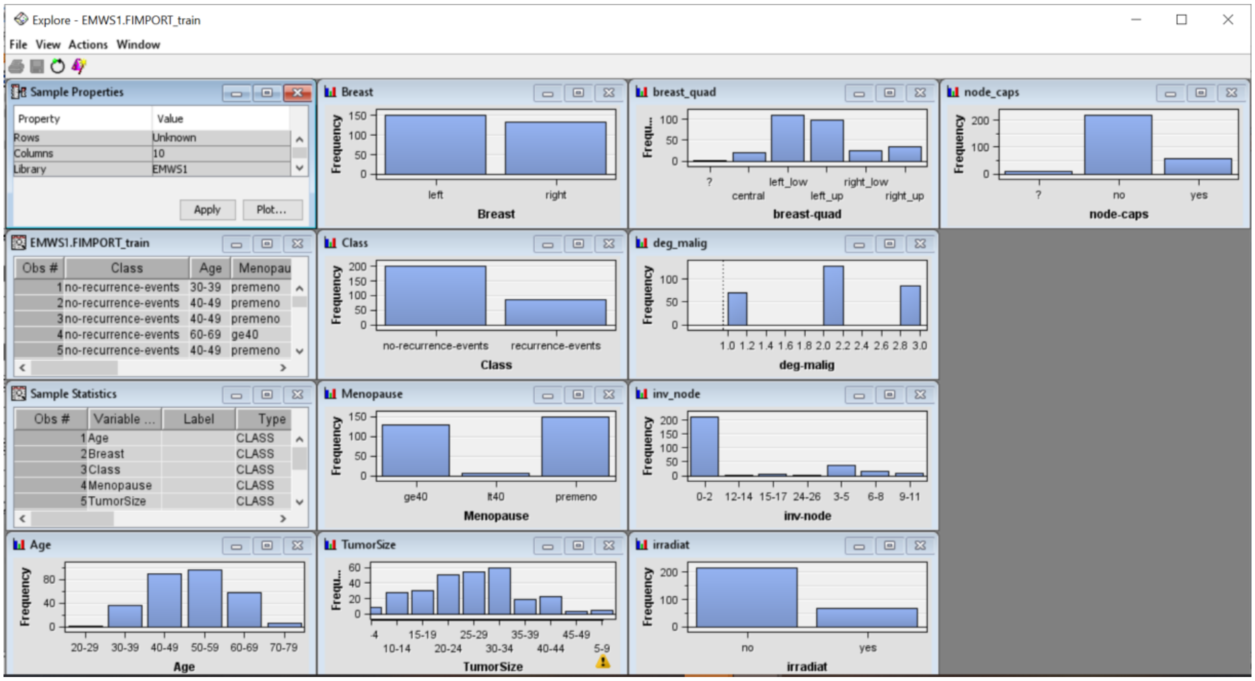
Property	Value
Rows	286
Columns	10
Library	EMWS1
Member	FIMPORT_DATA
Type	DATA
Sample Method	Top
Fetch Size	Default
Fetched Rows	286
Random Seed	12345

Obs #	Variable ...	Label	Type	Percent ...	Number o...	Mode Per...	Mode
1	Age		CLASS	0	6	33.56643	50-59

Obs #	Age
1	30-39
2	40-49
3	40-49
4	60-69
5	40-49
6	60-69
7	50-59
8	60-69
9	40-49
10	40-49
11	40-49
12	50-59
13	60-69
14	50-59
15	40-49

- ❖ The data tells us that there are 6 classifications for age, with the age range of 50-59 being the highest frequency at 33.56%.

Exploring all variables, below are the results derived:



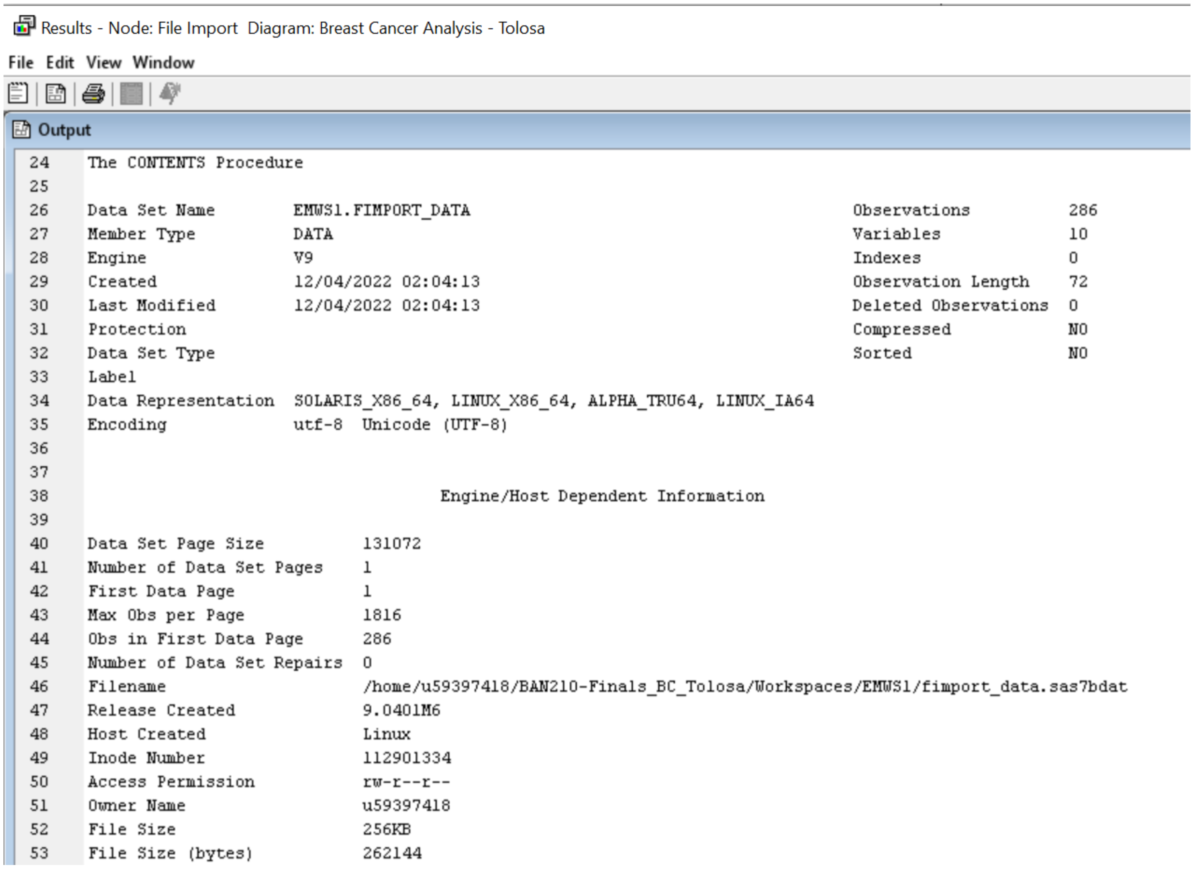
Property	Value
Rows	Unknown
Columns	10
Library	EMWS1

Obs #	Class	Age	Menopau
1	no-recurrence-events	30-39	premeno
2	no-recurrence-events	40-49	premeno
3	no-recurrence-events	40-49	premeno
4	no-recurrence-events	60-69	ge40
5	no-recurrence-events	40-49	premeno

Obs #	Variable ...	Label	Type
1	Age		CLASS
2	Breast		CLASS
3	Class		CLASS
4	Menopause		CLASS
5	TumorSize		CLASS

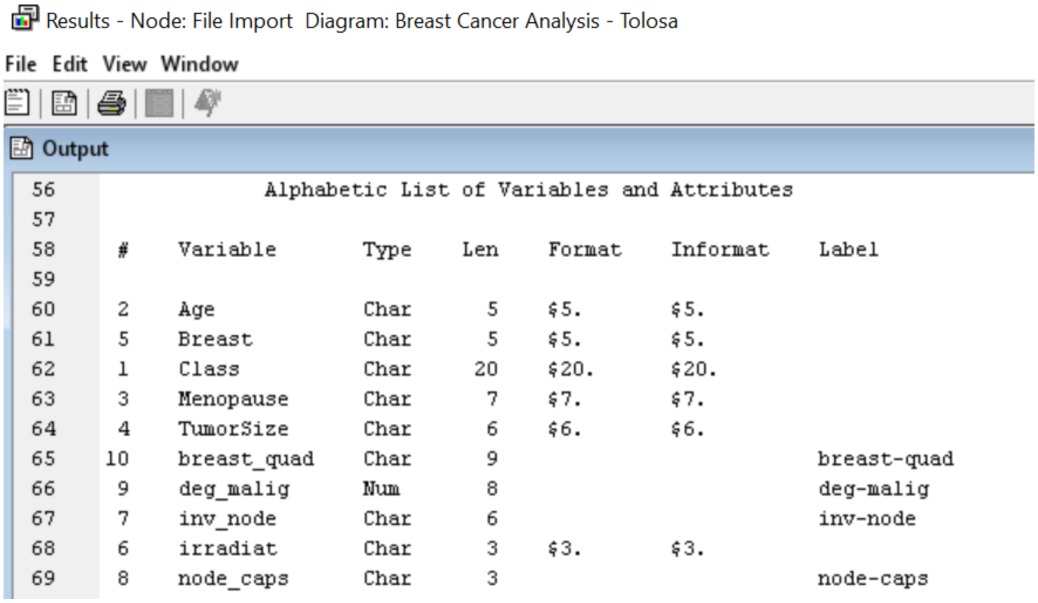
The File Import node is ran to check about the data. Then right click to check the Results.

- We can validate we have 286 observations, and 10 variables:


Results - Node: File Import Diagram: Breast Cancer Analysis - Tolosa
File Edit View Window
Output

```
The CONTENTS Procedure

Data Set Name        EMWS1.FIMPORT_DATA                                          Observations          286
Member Type          DATA                                                        Variables             10
Engine               V9                                                          Indexes               0
Created              12/04/2022 02:04:13                                         Observation Length    72
Last Modified        12/04/2022 02:04:13                                         Deleted Observations  0
Protection                                                                       Compressed            NO
Data Set Type                                                                    Sorted                NO
Label
Data Representation  SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64
Encoding             utf-8  Unicode (UTF-8)


                                    Engine/Host Dependent Information

Data Set Page Size           131072
Number of Data Set Pages     1
First Data Page              1
Max Obs per Page             1816
Obs in First Data Page       286
Number of Data Set Repairs   0
Filename                     /home/u59397418/BAN210-Finals_BC_Tolosa/Workspaces/EMWS1/fimport_data.sas7bdat
Release Created              9.0401M6
Host Created                 Linux
Inode Number                 112901334
Access Permission            rw-r--r--
Owner Name                   u59397418
File Size                    256KB
File Size (bytes)            262144
```

Information about the variables is shown like the type, length, and format:

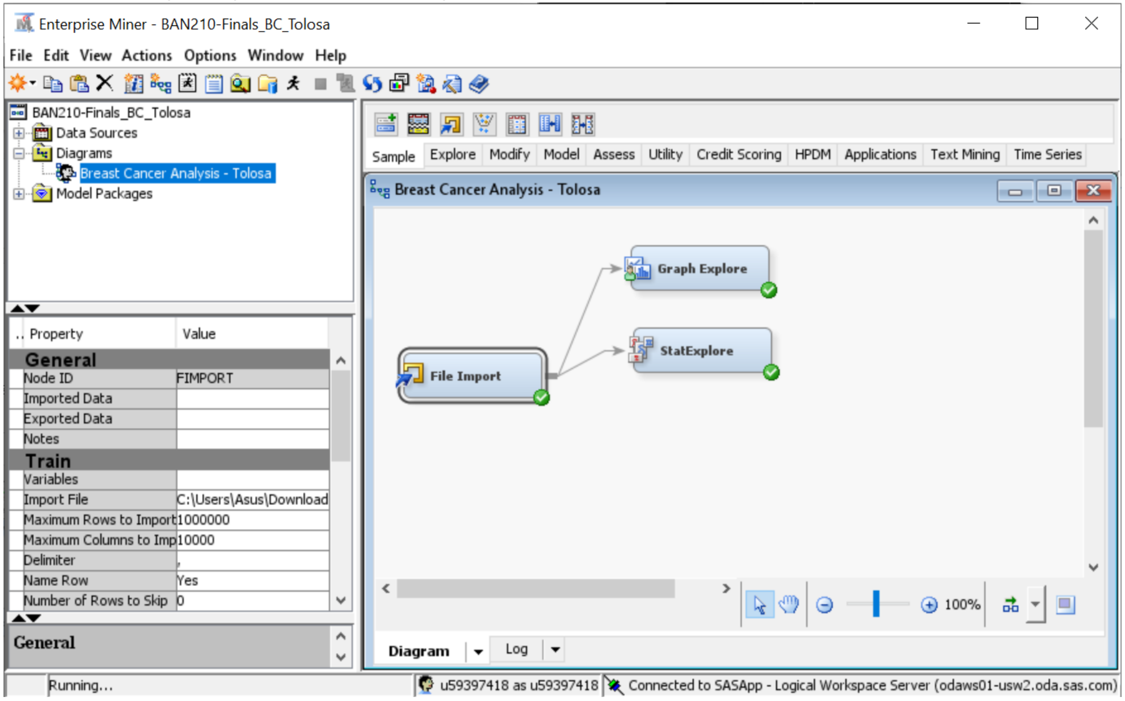

Results - Node: File Import Diagram: Breast Cancer Analysis - Tolosa
File Edit View Window
Output

```
              Alphabetic List of Variables and Attributes

 #    Variable       Type    Len    Format    Informat    Label

 2    Age            Char      5    $5.       $5.
 5    Breast         Char      5    $5.       $5.
 1    Class          Char     20    $20.      $20.
 3    Menopause      Char      7    $7.       $7.
 4    TumorSize      Char      6    $6.       $6.
10    breast_quad    Char      9                          breast-quad
 9    deg_malig      Num       8                          deg-malig
 7    inv_node       Char      6                          inv-node
 6    irradiat       Char      3    $3.       $3.
 8    node_caps      Char      3                          node-caps
```

Exploratory Data Analysis

Next is the Exploratory Data Analysis (EDA) to discover patterns, spot anomalies, test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

Here, the **Graph Explorer** and **StatExplorer** nodes are used.

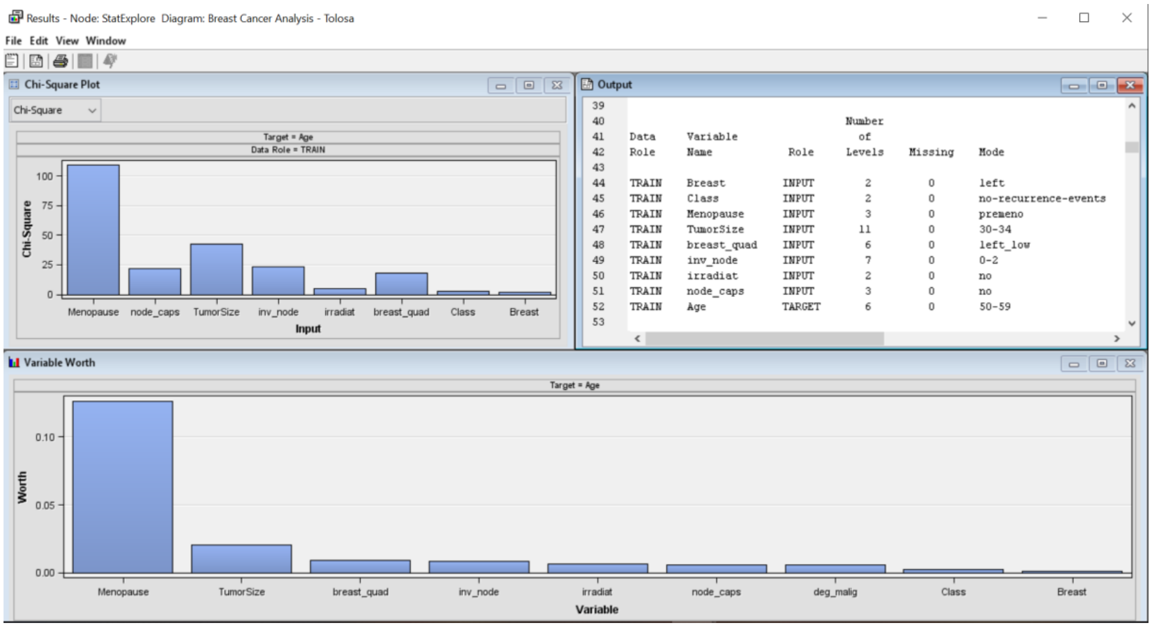

Enterprise Miner - BAN210-Finals_BC_Tolosa
File Edit View Actions Options Window Help
BAN210-Finals_BC_Tolosa
Data Sources
Diagrams
Breast Cancer Analysis - Tolosa
Model Packages
Sample Explore Modify Model Assess Utility Credit Scoring HPDM Applications Text Mining Time Series
Breast Cancer Analysis - Tolosa
Graph Explore
StatExplore
File Import
100%
Diagram
Log
General

Property	Value
General	
Node ID	FIMPORT
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Import File	C:\Users\Asus\Download
Maximum Rows to Import	1000000
Maximum Columns to Imp	10000
Delimiter	,
Name Row	Yes
Number of Rows to Skip	0

Running... u59397418 as u59397418 Connected to SASApp - Logical Workspace Server (odaws01-usw2.oda.sas.com)

StatExplorer Results

The StatExplorer is a versatile tool that gives the variable distributions and statistical information of the dataset. The Chi-Square, Variable Worth, and Output is shown here.

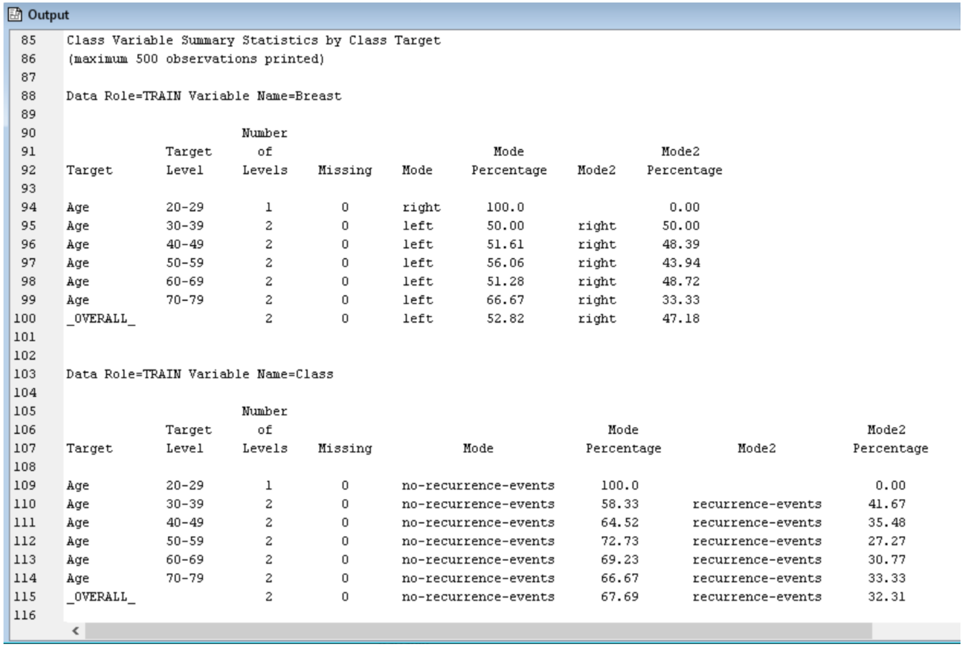


```

                                         Number
Data     Variable                          of
Role     Name              Role          Levels    Missing    Mode

TRAIN    Breast            INPUT            2          0      left
TRAIN    Class             INPUT            2          0      no-recurrence-events
TRAIN    Menopause         INPUT            3          0      premeno
TRAIN    TumorSize         INPUT           11          0      30-34
TRAIN    breast_quad       INPUT            6          0      left_low
TRAIN    inv_node          INPUT            7          0      0-2
TRAIN    irradiat          INPUT            2          0      no
TRAIN    node_caps         INPUT            3          0      no
TRAIN    Age               TARGET           6          0      50-59

```

Concentrating on the Output, more insight is gained on the statistics of how the target Age is compared with other variables.

An example below is the statistics for Age vs. Breast and Age vs. Class.

- ❖ We can say that for Age 40-49, 51.61% developed tumor on the left breast and 48.39% developed tumor on the right breast.

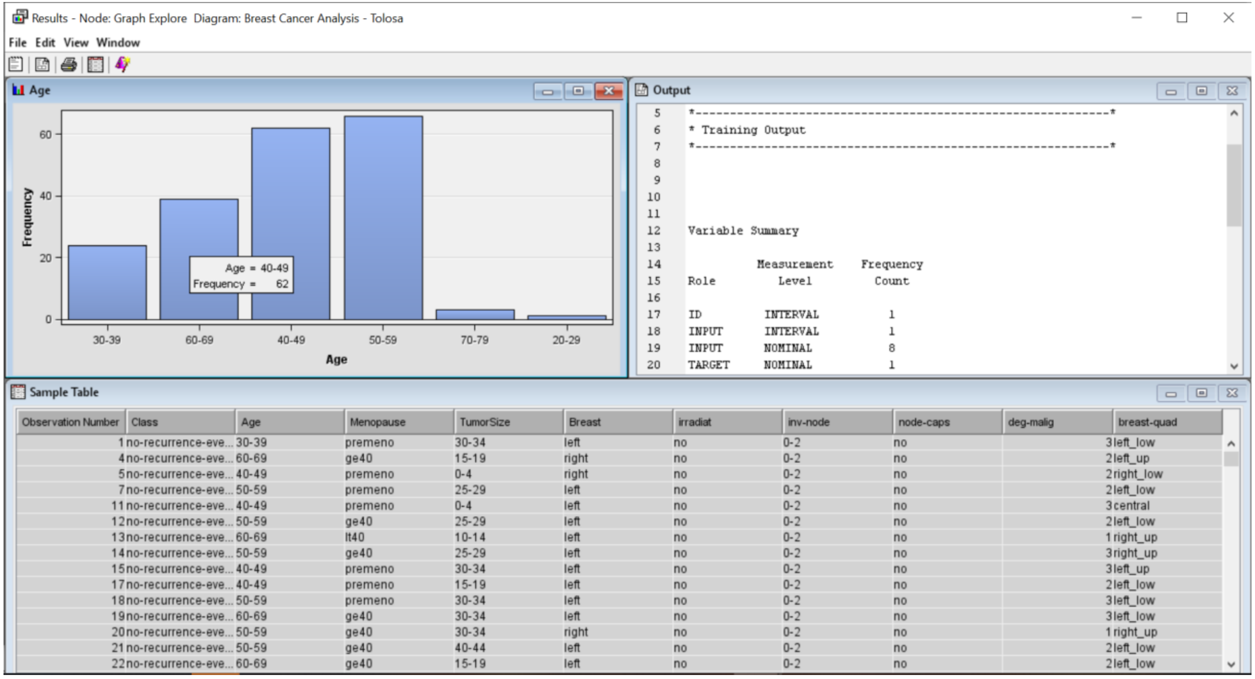


```
Class Variable Summary Statistics by Class Target
(maximum 500 observations printed)

Data Role=TRAIN Variable Name=Breast

                         Number
             Target        of                              Mode               Mode2
Target       Level       Levels    Missing    Mode      Percentage    Mode2    Percentage

Age          20-29          1          0       right       100.0                  0.00
Age          30-39          2          0       left        50.00      right      50.00
Age          40-49          2          0       left        51.61      right      48.39
Age          50-59          2          0       left        56.06      right      43.94
Age          60-69          2          0       left        51.28      right      48.72
Age          70-79          2          0       left        66.67      right      33.33
_OVERALL_                   2          0       left        52.82      right      47.18


Data Role=TRAIN Variable Name=Class

                         Number
             Target        of                                            Mode                                  Mode2
Target       Level       Levels    Missing           Mode             Percentage          Mode2              Percentage

Age          20-29          1          0       no-recurrence-events      100.0                                   0.00
Age          30-39          2          0       no-recurrence-events      58.33       recurrence-events          41.67
Age          40-49          2          0       no-recurrence-events      64.52       recurrence-events          35.48
Age          50-59          2          0       no-recurrence-events      72.73       recurrence-events          27.27
Age          60-69          2          0       no-recurrence-events      69.23       recurrence-events          30.77
Age          70-79          2          0       no-recurrence-events      66.67       recurrence-events          33.33
_OVERALL_                   2          0       no-recurrence-events      67.69       recurrence-events          32.31
```

Graph Explorer Results

The graph explorer shows more statistics about the target variable Age.

- ❖ This confirms that the most frequent data is that from the 50-59 age group and the least for the 20-29 age group.

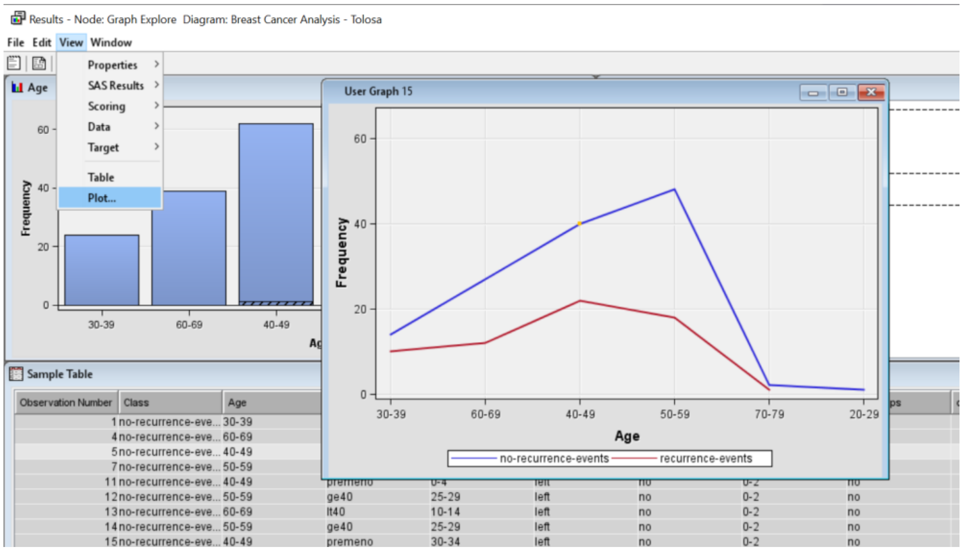


```
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

               Measurement    Frequency
Role              Level         Count

ID              INTERVAL          1
INPUT           INTERVAL          1
INPUT           NOMINAL           8
TARGET          NOMINAL           1
```

Observation Number	Class	Age	Menopause	TumorSize	Breast	irradiat	inv-node	node-caps	deg-malig	breast-quad
1	no-recurrence-eve...	30-39	premeno	30-34	left	no	0-2	no	3	left_low
4	no-recurrence-eve...	60-69	ge40	15-19	right	no	0-2	no	2	left_up
5	no-recurrence-eve...	40-49	premeno	0-4	right	no	0-2	no	2	right_low
7	no-recurrence-eve...	50-59	premeno	25-29	left	no	0-2	no	2	left_low
11	no-recurrence-eve...	40-49	premeno	0-4	left	no	0-2	no	3	central
12	no-recurrence-eve...	50-59	ge40	25-29	left	no	0-2	no	2	left_low
13	no-recurrence-eve...	60-69	lt40	10-14	left	no	0-2	no	1	right_up
14	no-recurrence-eve...	50-59	ge40	25-29	left	no	0-2	no	3	right_up
15	no-recurrence-eve...	40-49	premeno	30-34	left	no	0-2	no	3	left_up
17	no-recurrence-eve...	40-49	premeno	15-19	left	no	0-2	no	2	left_low
18	no-recurrence-eve...	50-59	premeno	30-34	left	no	0-2	no	3	left_low
19	no-recurrence-eve...	60-69	ge40	30-34	left	no	0-2	no	3	left_low
20	no-recurrence-eve...	50-59	ge40	30-34	right	no	0-2	no	1	right_up
21	no-recurrence-eve...	50-59	ge40	40-44	left	no	0-2	no	2	left_low
22	no-recurrence-eve...	60-69	ge40	15-19	left	no	0-2	no	2	left_low

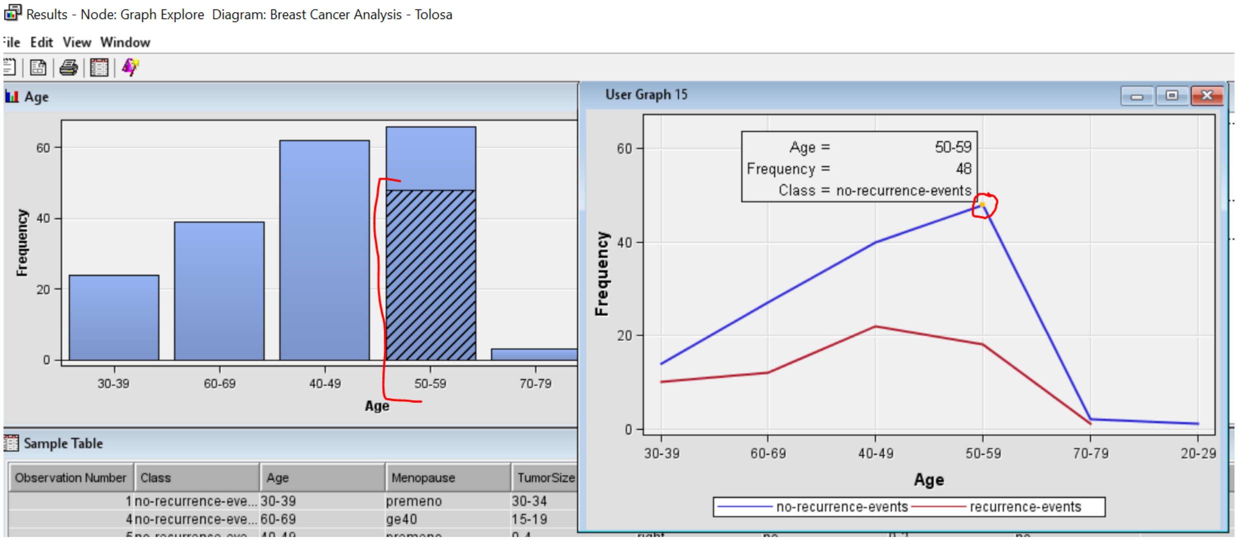
Graphs can also be made by going to View and the Plot. As an example, below is a Line graph with the Age set to Category role and the Class as the Group role

- ❖ The line graph shows us that for all Age range, the tumor being no-recurring (blue line) is greater than recurring tumors.



Observation Number	Class	Age						
1	no-recurrence-eve...	30-39						
4	no-recurrence-eve...	60-69						
5	no-recurrence-eve...	40-49						
7	no-recurrence-eve...	50-59						
11	no-recurrence-eve...	40-49	premeno	0-4	left	no	0-2	no
12	no-recurrence-eve...	50-59	ge40	25-29	left	no	0-2	no
13	no-recurrence-eve...	60-69	lt40	10-14	left	no	0-2	no
14	no-recurrence-eve...	50-59	ge40	25-29	left	no	0-2	no
15	no-recurrence-eve...	40-49	premeno	30-34	left	no	0-2	no

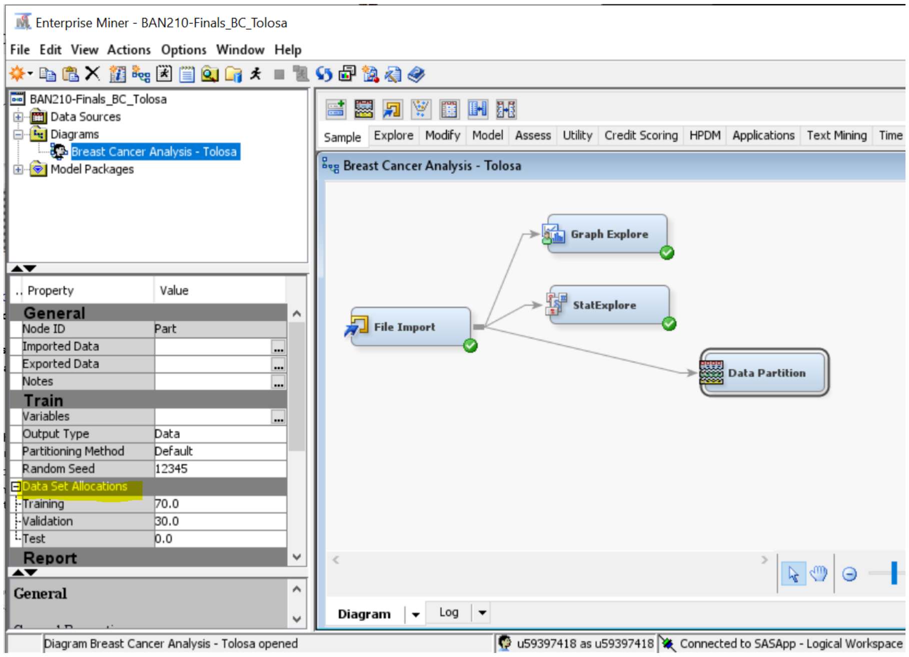
- ❖ There are 48 observations of no-reccurence at the age 50-59 which is the highest frequency. As this is pointed, the frequency graph on the left also shades the corresponding bar that represents this figure.



Observation Number	Class	Age	Menopause	TumorSize
1	no-recurrence-eve...	30-39	premeno	30-34
4	no-recurrence-eve...	60-69	ge40	15-19

Predictive Models

Data Partition node was added, and the Data Set Allocation was set to 70 for training and 30 for validation.


Enterprise Miner - BAN210-Finals_BC_Tolosa
File Edit View Actions Options Window Help
BAN210-Finals_BC_Tolosa
Data Sources
Diagrams
Breast Cancer Analysis - Tolosa
Model Packages
Sample Explore Modify Model Assess Utility Credit Scoring HPDM Applications Text Mining Time
Breast Cancer Analysis - Tolosa
Graph Explore
StatExplore
File Import
Data Partition
Diagram
Log
u59397418 as u59397418
Connected to SASApp - Logical Workspace

Property	Value
General	
Node ID	Part
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	70.0
Validation	30.0
Test	0.0
Report	

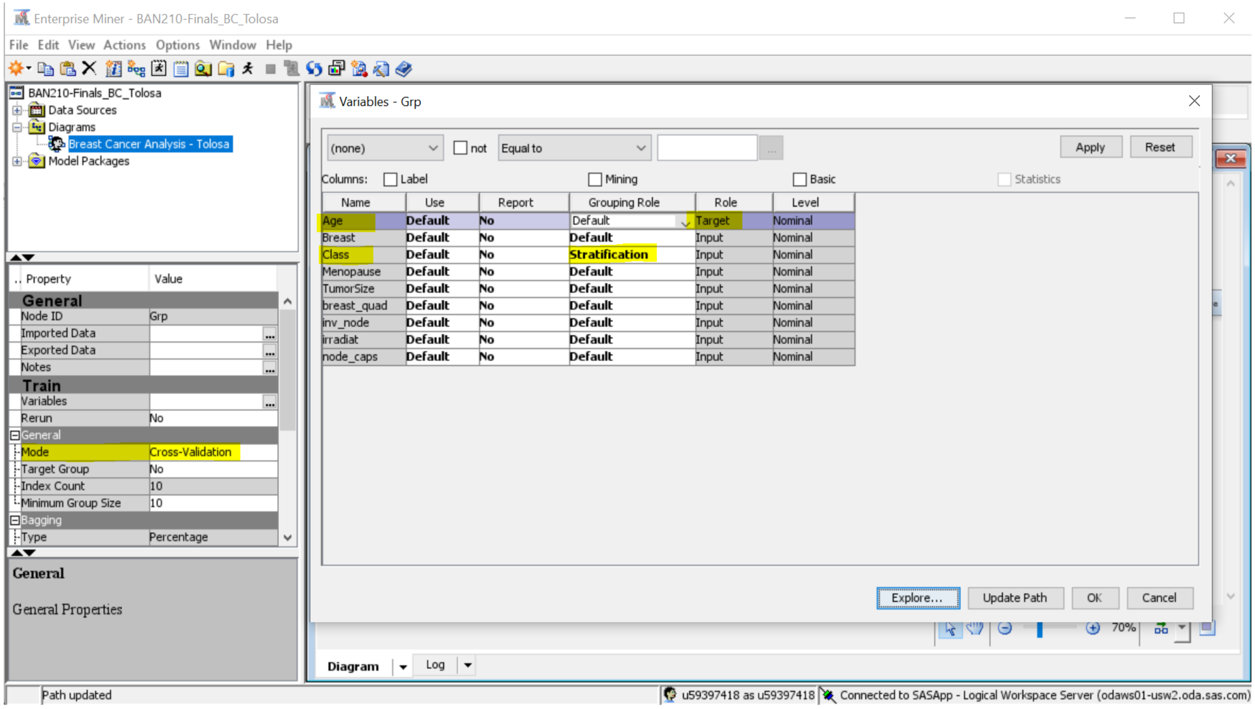
General

Diagram Breast Cancer Analysis - Tolosa opened

K-Fold Validation

Start Group node is added as a way to do the **k-fold validation,** to resample the dataset to ensure generalizability of the predictive models**.** It is on **Cross Validation** mode which exports the complement of the groups specified, as opposed to the groups themselves.

Age is set as the target role and **Class** is set to stratification group role.


Enterprise Miner - BAN210-Finals_BC_Tolosa
File Edit View Actions Options Window Help
BAN210-Finals_BC_Tolosa
Data Sources
Diagrams
Breast Cancer Analysis - Tolosa
Model Packages
Variables - Grp
(none)
□ not
Equal to
Apply
Reset
Columns: □ Label □ Mining □ Basic □ Statistics
Explore...
Update Path
OK
Cancel
Diagram
Log
70%
u59397418 as u59397418
Connected to SASApp - Logical Workspace Server (odaws01-usw2.oda.sas.com)

Property	Value
General	
Node ID	Grp
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Rerun	No
General	
Mode	Cross-Validation
Target Group	No
Index Count	10
Minimum Group Size	10
Bagging	
Type	Percentage

General

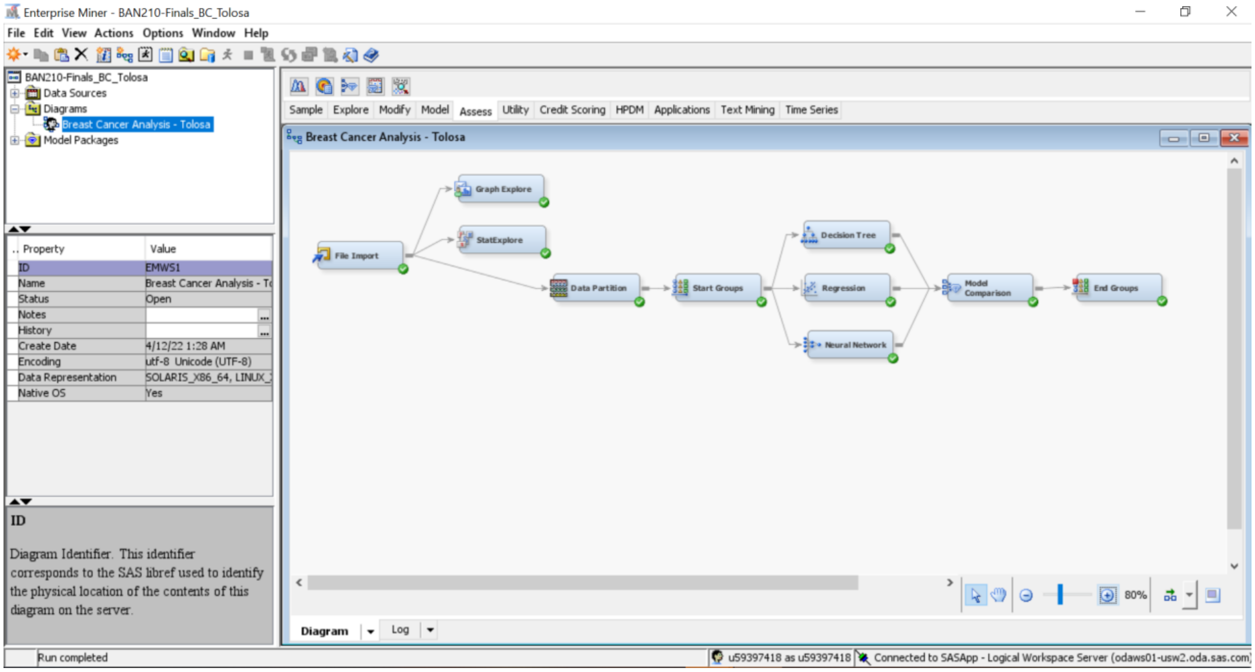
General Properties

Name	Use	Report	Grouping Role	Role	Level
Age	Default	No	Default	Target	Nominal
Breast	Default	No	Default	Input	Nominal
Class	Default	No	Stratification	Input	Nominal
Menopause	Default	No	Default	Input	Nominal
TumorSize	Default	No	Default	Input	Nominal
breast_quad	Default	No	Default	Input	Nominal
inv_node	Default	No	Default	Input	Nominal
irradiat	Default	No	Default	Input	Nominal
node_caps	Default	No	Default	Input	Nominal

Path updated

Next is the completion of the model by connecting the 3 predictive models used namely, the **Decision Tree, Regression, and Neural Network** model nodes.

These three nodes are connected to the **Model Comparison and finally the End Groups** node.


Enterprise Miner - BAN210-Finals_BC_Tolosa
File Edit View Actions Options Window Help
BAN210-Finals_BC_Tolosa
Data Sources
Diagrams
Breast Cancer Analysis - Tolosa
Model Packages
Sample Explore Modify Model Assess Utility Credit Scoring HPDM Applications Text Mining Time Series
Breast Cancer Analysis - Tolosa
Graph Explore
StatExplore
File Import
Data Partition
Start Groups
Decision Tree
Regression
Neural Network
Model Comparison
End Groups
80%
Diagram
Log

Property	Value
ID	EMWS1
Name	Breast Cancer Analysis - To
Status	Open
Notes	
History	
Create Date	4/12/22 1:28 AM
Encoding	utf-8 Unicode (UTF-8)
Data Representation	SOLARIS_X86_64, LINUX_
Native OS	Yes

ID

Diagram Identifier. This identifier corresponds to the SAS libref used to identify the physical location of the contents of this diagram on the server.

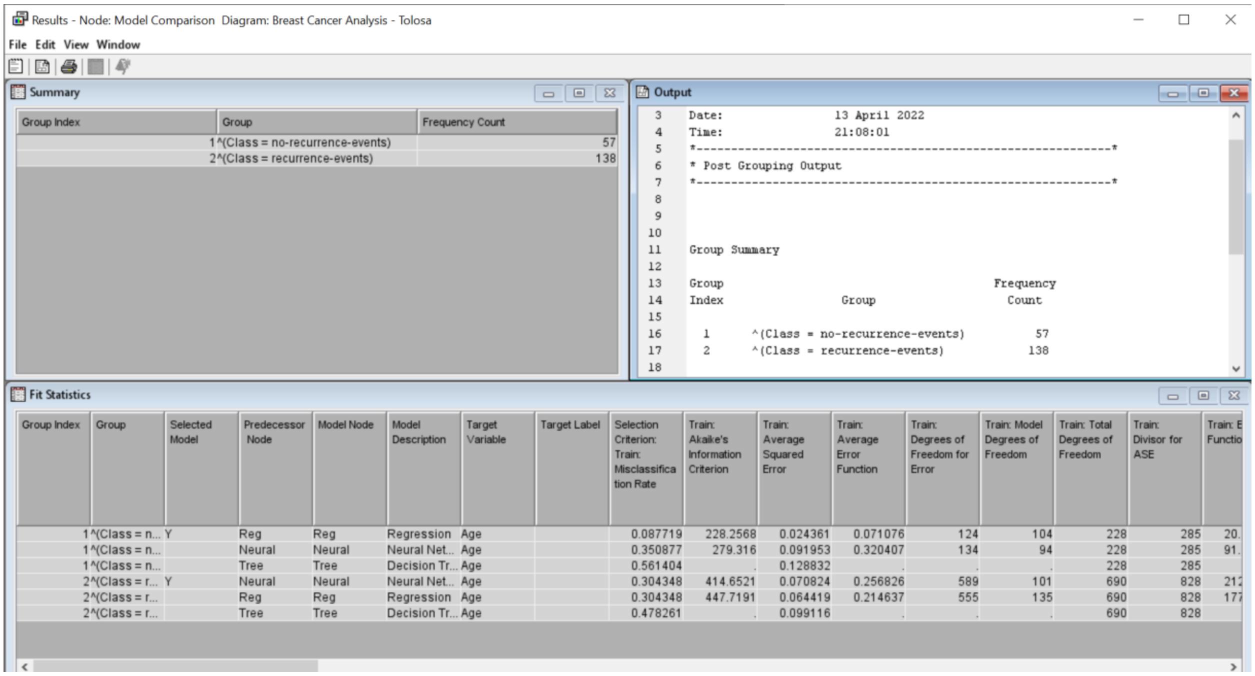
Run completed

u59397418 as u59397418 Connected to SASApp - Logical Workspace Server (odaws01-usw2.oda.sas.com)

Run the End Group Node.

Model Comparison

First, the **Model Comparison Results** is examined— focusing on the Misclassification Rate to see how fit the model is.


Results - Node: Model Comparison Diagram: Breast Cancer Analysis - Tolosa
File Edit View Window
Summary
Output
Fit Statistics

Group Index	Group	Frequency Count
1	^(Class = no-recurrence-events)	57
2	^(Class = recurrence-events)	138

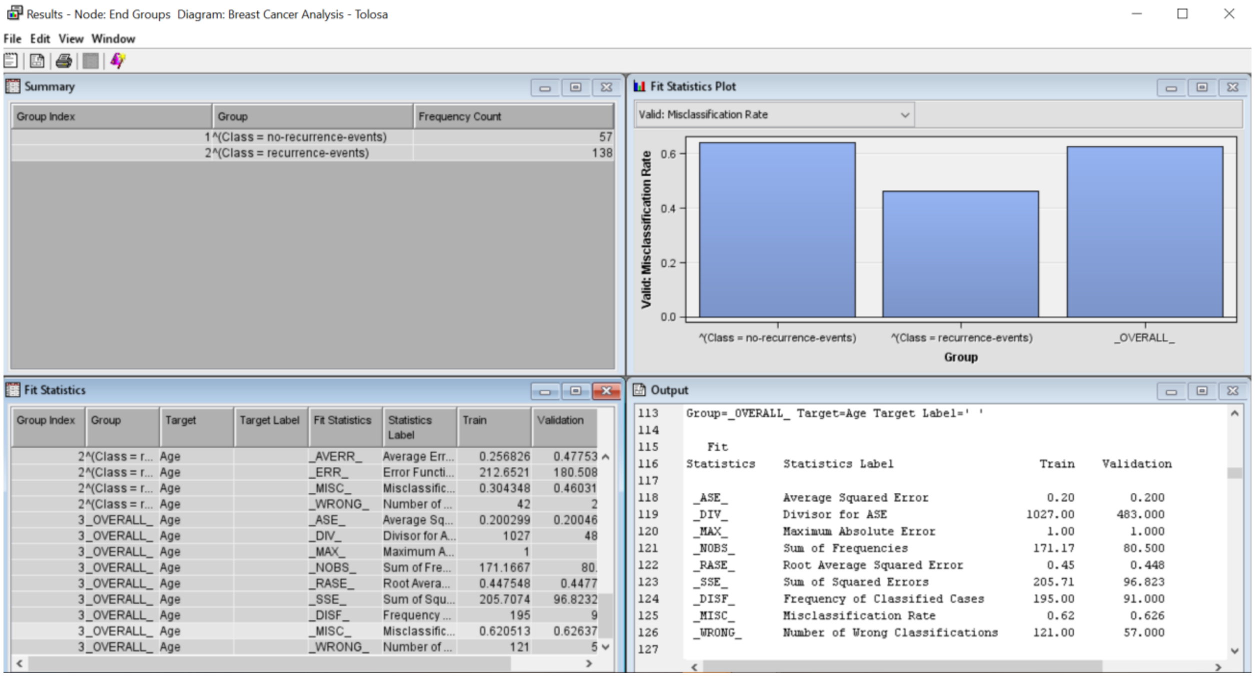
```
Date:                  13 April 2022
Time:                  21:08:01
*------------------------------------------------------------*
* Post Grouping Output
*------------------------------------------------------------*



Group Summary

Group                                              Frequency
Index                 Group                          Count

  1        ^(Class = no-recurrence-events)            57
  2        ^(Class = recurrence-events)              138

```

Group Index	Group	Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Train: 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: E Functio
1	^(Class = n...	Y	Reg	Reg	Regression	Age		0.087719	228.2568	0.024361	0.071076	124	104	228	285	20.
1	^(Class = n...		Neural	Neural	Neural Net...	Age		0.350877	279.316	0.091953	0.320407	134	94	228	285	91.
1	^(Class = n...		Tree	Tree	Decision Tr...	Age		0.561404	.	0.128832	.	.	.	228	285	
2	^(Class = r...	Y	Neural	Neural	Neural Net...	Age		0.304348	414.6521	0.070824	0.256826	589	101	690	828	212
2	^(Class = r...		Reg	Reg	Regression	Age		0.304348	447.7191	0.064419	0.214637	555	135	690	828	177
2	^(Class = r...		Tree	Tree	Decision Tr...	Age		0.478261	.	0.099116	.	.	.	690	828	

- Based on the Fit Statistics, the **Regression** is the selected model for the no-recurring group and **Neural Network** is the selected model for the recurring group, both with the Age as target variable.

These were picked to be the best fit because they had the lowest misclassification rate and the lower, the better because it means that the forecast is closer to the actual.

The Results of the three models is shown when clicking on the End Group Results. It shows statistics of all three models and their overall results.


Results - Node: End Groups Diagram: Breast Cancer Analysis - Tolosa
File
Edit
View
Window
Summary
Fit Statistics Plot
Valid: Misclassification Rate
Valid: Misclassification Rate
0.0
0.2
0.4
0.6
^(Class = no-recurrence-events)
^(Class = recurrence-events)
OVERALL
Group
Fit Statistics
Output

Summary

Group Index	Group	Frequency Count
1	^(Class = no-recurrence-events)	57
2	^(Class = recurrence-events)	138

Fit Statistics

Group Index	Group	Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
2	^(Class = r...	Age		_AVERR_	Average Err...	0.256826	0.47753
2	^(Class = r...	Age		_ERR_	Error Functi...	212.6521	180.508
2	^(Class = r...	Age		_MISC_	Misclassific...	0.304348	0.46031
2	^(Class = r...	Age		_WRONG_	Number of ...	42	2
3	_OVERALL_	Age		_ASE_	Average Sq...	0.200299	0.20046
3	_OVERALL_	Age		_DIV_	Divisor for A...	1027	48
3	_OVERALL_	Age		_MAX_	Maximum A...	1	
3	_OVERALL_	Age		_NOBS_	Sum of Fre...	171.1667	80.
3	_OVERALL_	Age		_RASE_	Root Avera...	0.447548	0.4477
3	_OVERALL_	Age		_SSE_	Sum of Squ...	205.7074	96.8232
3	_OVERALL_	Age		_DISF_	Frequency ...	195	9
3	_OVERALL_	Age		_MISC_	Misclassific...	0.620513	0.62637
3	_OVERALL_	Age		_WRONG_	Number of ...	121	5

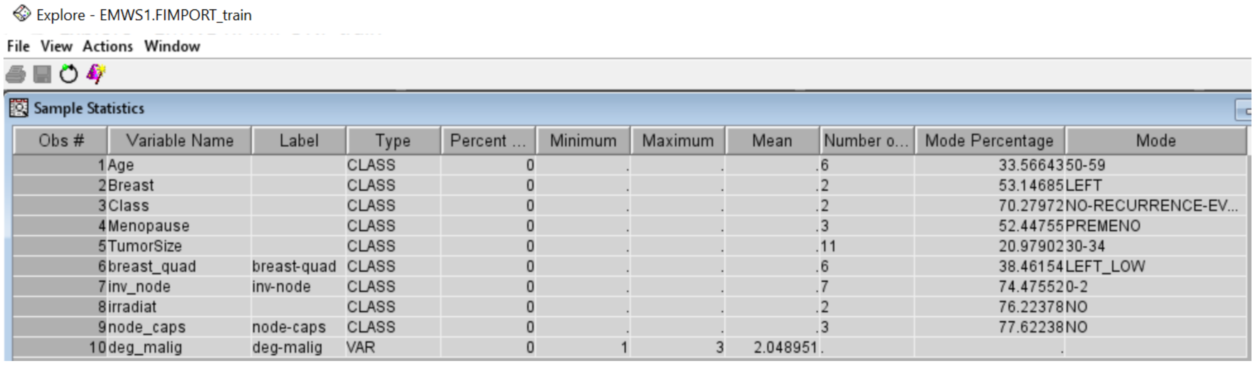
```
Group=_OVERALL_ Target=Age Target Label=' '

   Fit
Statistics    Statistics Label                   Train    Validation

 _ASE_        Average Squared Error               0.20       0.200
 _DIV_        Divisor for ASE                  1027.00     483.000
 _MAX_        Maximum Absolute Error              1.00       1.000
 _NOBS_       Sum of Frequencies                171.17      80.500
 _RASE_       Root Average Squared Error          0.45       0.448
 _SSE_        Sum of Squared Errors             205.71      96.823
 _DISF_       Frequency of Classified Cases     195.00      91.000
 _MISC_       Misclassification Rate              0.62       0.626
 _WRONG_      Number of Wrong Classifications   121.00      57.000

```

The Fit Statistics plot can be navigated to show the bar graph for each fit statistic label.

Conclusion

In this report, a simple exploratory data analysis was run wherein Age vs. Breast was first analyzed and a sample conclusion was mentioned as seen from the result, that is, for Age 40-49, 51.61% developed tumor on the left breast and 48.39% developed tumor on the right breast. This is a statistical analysis which can be presented to the business about the likelihood of the location of the tumor for those age group. With insights like this, looking at the bigger picture, medical diagnosis can be shifted when the patient's age is known as well as testing and prescription that might result to savings in time and money.

Another sample statistic that can be looked at is the mode of each variable, to see their likelihood:



Obs #	Variable Name	Label	Type	Percent ...	Minimum	Maximum	Mean	Number o...	Mode Percentage	Mode
1	Age		CLASS	0	.	.	.	6	33.5664	50-59
2	Breast		CLASS	0	.	.	.	2	53.14685	LEFT
3	Class		CLASS	0	.	.	.	2	70.27972	NO-RECURRENCE-EV...
4	Menopause		CLASS	0	.	.	.	3	52.44755	PREMENO
5	TumorSize		CLASS	0	.	.	.	11	20.97902	30-34
6	breast_quad	breast-quad	CLASS	0	.	.	.	6	38.46154	LEFT_LOW
7	inv_node	inv-node	CLASS	0	.	.	.	7	74.47552	0-2
8	irradiat		CLASS	0	.	.	.	2	76.22378	NO
9	node_caps	node-caps	CLASS	0	.	.	.	3	77.62238	NO
10	deg_malig	deg-malig	VAR	0	1	3	2.048951	.	.	

Analyzing the chart above, it can be said to the business that the likelihood of developing breast cancer is at the age of 50-59 (33.5%) based on the data set. The tumor is also 53% likely to be on the left breast. It is, however 70% not recurring, and 52% in the premeno stage. There is a 20.9% change that the tumor size is 30-34, 38.5% chance it is in the left lower breast quadrant, 74% chance it is in the 0-2 inv_node, and 76% not irradiat and 77% likelihood that it has no node_caps.

When it comes to predictive analysis, three models were ran, namely, the Decision Tree, Logistic Regression, and Neural Network. Age was set as Target variable and Class as the Stratification in the grouping role.

After running the model comparison and checking the Fit Statistics, the **Regression** is the selected model for the no-recurring group and **Neural Network** is the selected model for the recurring group. Knowing which model to use is important because the goal is to minimize the error or difference in the prediction to the actual, in this case, the class.

Work Cited

Zwitter, Matjaz and & Soklic, Milan. "Breast Cancer Data Set." *UCI Machine Learning*. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer.