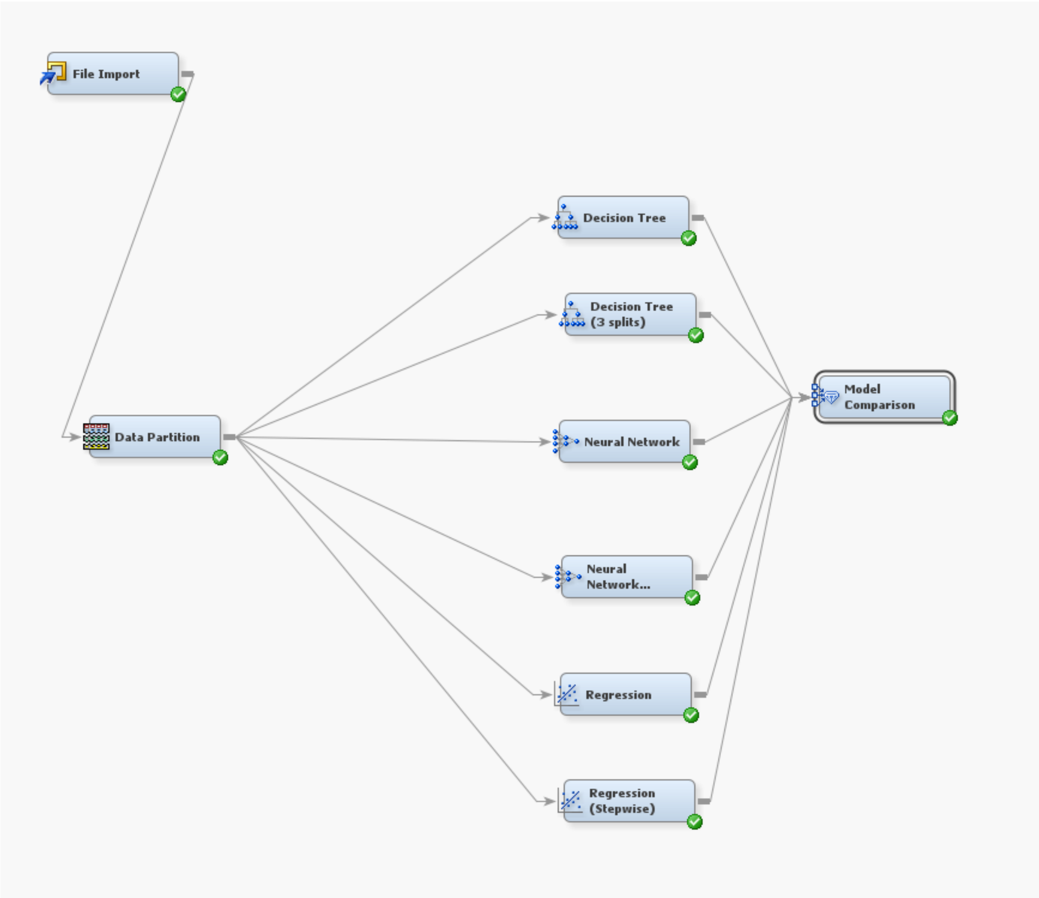
UNIVERSITI MALAYA
Centre for Continuing Education

Master Of Data Science

INDIVIDUAL ASSIGNMENT

COURSE CODE	**: WQD7003**
COURSE TITLE	**: DATA ANALYTICS**
MATRIC NUMBER	**: 22120931**
CLASS	**: 1**
LECTURER	**: Dr. SAW SHIER NEE**

1. SAS Workflow process



The figure above shows the modelling pipeline used in this individual assignment. Firstly, breast.csv dataset is loaded using the “File Import” node. Next, the data is partitioned into training and validation sets with a 50:50 split using the “Data Partition” node to ensure an even distribution for model training and validation. Then, multiple modelling nodes such as Decision Tree, Regression and Neural Network are connected to the “Data Partition”.

For the Decision Tree model, two versions are created to allow comparison between a basic and a more finely tuned Decision Tree model. One model is with the default setting and another model with a specific parameter of a maximum depth of 10, leaf size of 8, 4 surrogate rules and 3 splits.

In the case of the Neural Network model, one model is configured with default settings and another is using the “Back Prop” training technique. Changing the training technique can be done by clicking the “Optimization” of the Property Panel.

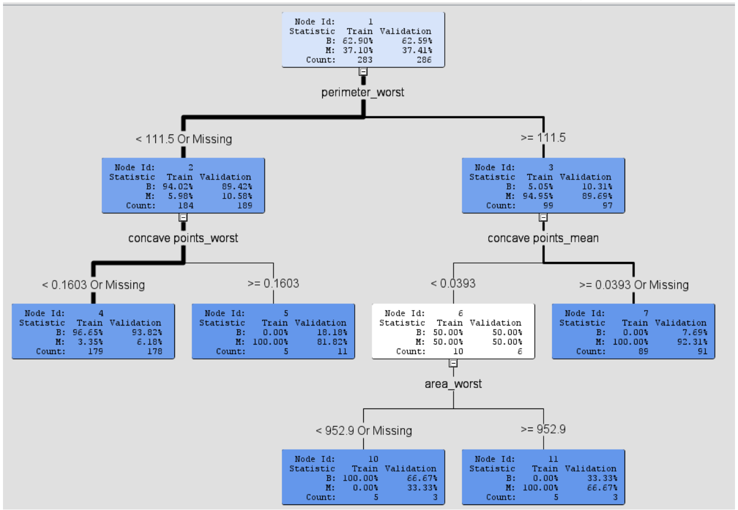
On the other hand, two types of Logistic Regression models are configured. One with default settings and another using the Stepwise regression method. The Stepwise method is selected from the Property Panel.

Lastly, the modelling nodes are connected to the Model Comparison node. This is to evaluate and compare the performance of all the models created. Also, the comparison helps identify the best-performing model, providing a comprehensive view of how different modelling techniques and parameter settings impact overall model performance.

2. Results

i) Decision Tree

Below is the rules created with default settings:



The classification results below show that for the training data show that only 6 counts of malignant cases are misclassified as benign, while benign cases have no misclassifications. For the validation data, the misclassification counts are 12 for malignant cases predicted as benign and 10 for benign cases predicted as malignant.

Classification Table

Data Role=TRAIN Target Variable=diagnosis Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
B	B	96.739	100.000	178	62.8975
M	B	3.261	5.714	6	2.1201
M	M	100.000	94.286	99	34.9823

Data Role=VALIDATE Target Variable=diagnosis Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
B	B	93.3702	94.4134	169	59.0909
M	B	6.6298	11.2150	12	4.1958
B	M	9.5238	5.5866	10	3.4965
M	M	90.4762	88.7850	95	33.2168

Train Confusion Matrix:

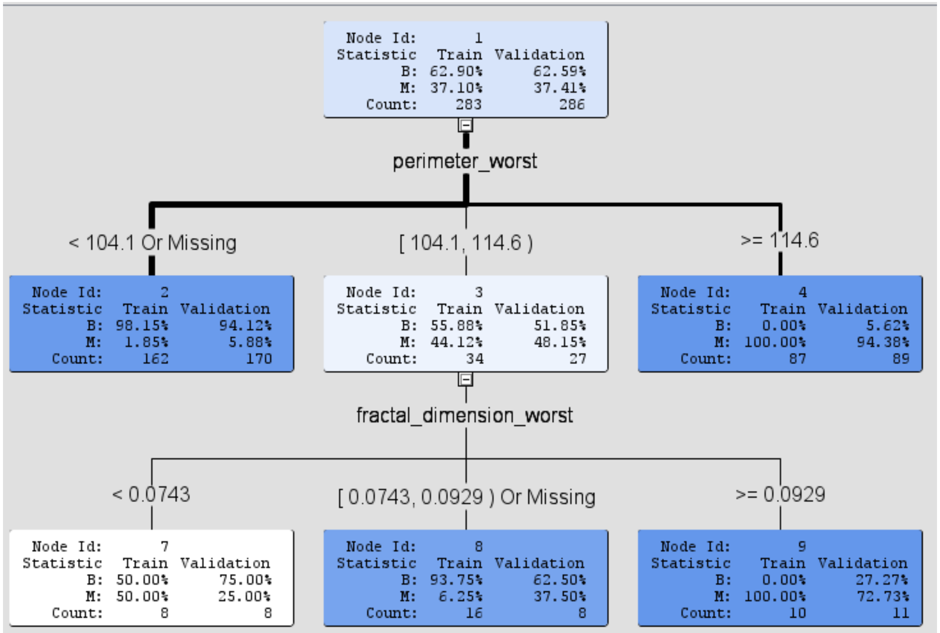
		Target	
		Benign	Malignant
Outcome	Benign	178	6
	Malignant	0	99

Validate Confusion Matrix:

		Target	
		Benign	Malignant
Outcome	Benign	169	12
	Malignant	10	95

ii) Decision Tree (3 Splits)

Below is the rules created:



The classification results below show that for the training data show that only 4 counts of malignant cases are misclassified as benign, and 4 counts of benign cases are misclassified as malignant. For the validation data, the misclassification rates are slightly higher compared to the training data, with 13 counts of malignant cases being predicted as benign and 14 counts of benign cases being predicted as malignant.

Classification Table

Data Role=TRAIN Target Variable=diagnosis Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
B	B	97.7528	97.7528	174	61.4841
M	B	2.2472	3.8095	4	1.4134
B	M	3.8095	2.2472	4	1.4134
M	M	96.1905	96.1905	101	35.6890

Data Role=VALIDATE Target Variable=diagnosis Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
B	B	92.6966	92.1788	165	57.6923
M	B	7.3034	12.1495	13	4.5455
B	M	12.9630	7.8212	14	4.8951
M	M	87.0370	87.8505	94	32.8671

Train Confusion Matrix:

		Target	
		Benign	Malignant
Outcome	Benign	174	4
	Malignant	4	101

Validate Confusion Matrix:

		Target	
		Benign	Malignant
Outcome	Benign	165	13
	Malignant	14	94

iii) Neural Network

The model achieved a perfect classification rate on the training data but experienced an 8.39% misclassification rate on the validation data.

In the classification result below, the training set accurately predicted both malignant and benign cases. However, the model showed slight misclassifications in the validation set with 12 counts for both malignant cases being incorrectly predicted as benign and benign cases being incorrectly classified as malignant.

```
Classification Table

Data Role=TRAIN Target Variable=diagnosis Target Label=' '

                       Target      Outcome    Frequency      Total
Target    Outcome    Percentage   Percentage    Count      Percentage

  B          B          100          100         178        62.8975
  M          M          100          100         105        37.1025


Data Role=VALIDATE Target Variable=diagnosis Target Label=' '

                       Target      Outcome    Frequency      Total
Target    Outcome    Percentage   Percentage    Count      Percentage

  B          B        93.2961      93.2961       167        58.3916
  M          B         6.7039      11.2150        12         4.1958
  B          M        11.2150       6.7039        12         4.1958
  M          M        88.7850      88.7850        95        33.2168
```

Train Confusion Matrix:

		Target	
		Benign	Malignant
Outcome	Benign	178	0
	Malignant	0	105

Validate Confusion Matrix:

		Target	
		Benign	Malignant
Outcome	Benign	167	12
	Malignant	12	95

iv) Neural Network (Backdrop)

Similar to Neural Network with default settings, the model achieved a perfect classification rate on the training data but experienced an 8.39% misclassification rate on the validation data.

In the classification result below, the training set accurately predicted both malignant and benign cases. However, the model showed slight misclassifications in the validation set with 12 counts for both malignant cases being incorrectly predicted as benign and benign cases being incorrectly classified as malignant. The result here is similar to Neural Network using default settings.

```
Classification Table

Data Role=TRAIN Target Variable=diagnosis Target Label=' '

                        Target       Outcome      Frequency      Total
Target    Outcome     Percentage    Percentage      Count      Percentage

  B          B            100          100           178         62.8975
  M          M            100          100           105         37.1025


Data Role=VALIDATE Target Variable=diagnosis Target Label=' '

                        Target       Outcome      Frequency      Total
Target    Outcome     Percentage    Percentage      Count      Percentage

  B          B          93.2961      93.2961         167         58.3916
  M          B           6.7039      11.2150          12          4.1958
  B          M          11.2150       6.7039          12          4.1958
  M          M          88.7850      88.7850          95         33.2168
```

Train Confusion Matrix:

		Target	
		Benign	Malignant
Outcome	Benign	178	0
	Malignant	0	105

Validate Confusion Matrix:

		Target	
		Benign	Malignant
Outcome	Benign	167	12
	Malignant	12	95

v) Logistic Regression

The model achieved a perfect classification rate on the training data but experienced a 10.49% misclassification rate on the validation data.

In the classification result below, the training set accurately predicted both malignant and benign cases. However, there are some misclassifications in the validation set with 13 counts of malignant cases incorrectly predicted as benign and 17 counts of benign cases incorrectly classified as malignant.

```
Classification Table

Data Role=TRAIN Target Variable=diagnosis Target Label=' '

                        Target       Outcome      Frequency      Total
Target    Outcome     Percentage    Percentage      Count      Percentage

  B          B            100           100          178         62.8975
  M          M            100           100          105         37.1025


Data Role=VALIDATE Target Variable=diagnosis Target Label=' '

                        Target       Outcome      Frequency      Total
Target    Outcome     Percentage    Percentage      Count      Percentage

  B          B          92.5714       90.5028        162         56.6434
  M          B           7.4286       12.1495         13          4.5455
  B          M          15.3153        9.4972         17          5.9441
  M          M          84.6847       87.8505         94         32.8671
```

Train Confusion Matrix:

		Target	
		Benign	Malignant
Outcome	Benign	178	0
	Malignant	0	105

Validate Confusion Matrix:

		Target	
		Benign	Malignant
Outcome	Benign	162	13
	Malignant	17	94

vi) Logistics Regression (StepWise)

The model showed a misclassification rate of 2.12% on the training data and a 5.94% misclassification rate on the validation data.

In the classification result below, there are some misclassifications observed in both sets. In the training set, 4 counts of malignant cases were incorrectly predicted as benign and 2 counts of benign cases were incorrectly classified as malignant. On the other hand, 8 counts of malignant cases were incorrectly predicted as benign and 9 counts of benign cases were incorrectly classified as malignant in the validation set.

```
Classification Table

Data Role=TRAIN Target Variable=diagnosis Target Label=' '

                         Target       Outcome     Frequency      Total
Target    Outcome     Percentage   Percentage      Count      Percentage

  B          B          97.7778      98.8764        176         62.1908
  M          B           2.2222       3.8095          4          1.4134
  B          M           1.9417       1.1236          2          0.7067
  M          M          98.0583      96.1905        101         35.6890


Data Role=VALIDATE Target Variable=diagnosis Target Label=' '

                         Target       Outcome     Frequency      Total
Target    Outcome     Percentage   Percentage      Count      Percentage

  B          B          95.5056      94.9721        170         59.4406
  M          B           4.4944       7.4766          8          2.7972
  B          M           8.3333       5.0279          9          3.1469
  M          M          91.6667      92.5234         99         34.6154
```

Train Confusion Matrix:

		Target	
		Benign	Malignant
Outcome	Benign	176	4
	Malignant	2	101

Validate Confusion Matrix:

		Target	
		Benign	Malignant
Outcome	Benign	170	8
	Malignant	9	99

vii) Model Comparison

Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Reg2	Regression (Stepwise)	0.05944	0.018768	0.021201	0.049497
	Tree	Decision Tree	0.07692	0.020491	0.021201	0.075044
	Neural	Neural Network	0.08392	0.000015	0.000000	0.078533
	Neural2	Neural Network (Backdrop)	0.08392	0.000015	0.000000	0.078533
	Tree2	Decision Tree (3 splits)	0.09441	0.020784	0.028269	0.078126
	Reg	Regression	0.10490	0.000006	0.000000	0.099700

The figure above shows that the Logistic Regression model with default setting and both Neural Network models performed very well in the training data with zero misclassification rate. This suggests that these models were able to effectively learn from the training data and accurately classify the instances. On the other hand, the Decision Tree model with 3 splits shows a slightly higher misclassification rate of 2.83% in the training data compared to the Decision Tree model with default settings. This suggests that increasing the complexity of the Decision Tree model by adding more splits can lead to a slightly higher misclassification rate. Additionally, the Decision Tree model with default settings and the Logistic Regression model using the Stepwise method has a similar misclassification rate of 2.12%.

In the validation data, all models exhibit higher misclassification rates compared to the training data. For the validation data, Logistic Regression with default settings showed the highest misclassification rate of 10.49%. The lowest misclassification rate in the validation data is the Logistics Regression model using the Stepwise method with 5.94%.
Both Neural Network model has a similar misclassification rate on the training data and validation data.

In terms of the overfitting model, the Logistic Regression with Default Settings model shows that it has overfitted the training date since it has a zero misclassification rate on the training data but had the highest misclassification rate of 10.49% on the validation data. Furthermore, both Decision Tree models also show overfitting issues due to an overly high misclassification rate of validation data compared to training data.

In conclusion, models that perform well on both training and validation data with low misclassification rates are preferred. In my opinion, the Logistic Regression model using the Stepwise regression method appears to be the best-performing model based on the figure above. It achieved a reasonable misclassification rate in both datasets. However, further improvement and fine-tuning may be necessary to enhance the performance of the models on the validation data.

```
Event Classification Table
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Model                                  Data                    Target    False      True      False      True
Node      Model Description            Role       Target       Label     Negative   Negative  Positive   Positive

Tree2     Decision Tree (3 splits)     TRAIN      diagnosis                 4         174         4        101
Tree2     Decision Tree (3 splits)     VALIDATE   diagnosis                13         165        14         94
Reg       Regression                   TRAIN      diagnosis                 .         178         .        105
Reg       Regression                   VALIDATE   diagnosis                13         162        17         94
Neural    Neural Network               TRAIN      diagnosis                 .         178         .        105
Neural    Neural Network               VALIDATE   diagnosis                12         167        12         95
Tree      Decision Tree                TRAIN      diagnosis                 6         178         .         99
Tree      Decision Tree                VALIDATE   diagnosis                12         169        10         95
Reg2      Regression (Stepwise)        TRAIN      diagnosis                 4         176         2        101
Reg2      Regression (Stepwise)        VALIDATE   diagnosis                 8         170         9         99
Neural2   Neural Network (Backdrop)    TRAIN      diagnosis                 .         178         .        105
Neural2   Neural Network (Backdrop)    VALIDATE   diagnosis                12         167        12         95
```

The figure above is the summarisation of the confusion matrix for all the models.

Train Confusion Matrix:

		Target		
		Benign	Malignant	
Outcome	Benign	178	6	Decision Tree
	Malignant	0	99	
Outcome	Benign	174	4	Decision Tree (3 Splits)
	Malignant	4	101	
Outcome	Benign	178	0	Neural Network
	Malignant	0	105	
Outcome	Benign	178	0	Neural Network (Back Prop)
	Malignant	0	105	
Outcome	Benign	178	0	Logistics Regression
	Malignant	0	105	
Outcome	Benign	176	4	Logistics Regression (StepWise)
	Malignant	2	101	

Validate Confusion Matrix:

		Target		
		Benign	Malignant	
Outcome	Benign	169	12	Decision Tree
	Malignant	10	95	
Outcome	Benign	165	13	Decision Tree (3 Splits)
	Malignant	14	94	
Outcome	Benign	167	12	Neural Network
	Malignant	12	95	
Outcome	Benign	167	12	Neural Network (Back Prop)
	Malignant	12	95	
Outcome	Benign	162	13	Logistics Regression
	Malignant	17	94	
Outcome	Benign	170	8	Logistics Regression (StepWise)
	Malignant	9	99	