BAN 210 – FINAL ASSESSMENT

ANALYSIS ON BREAST CANCER DATASET

NAME: DHANANJAY KUMAR
STUDENT ID: 135297208
SUBMITTED TO: PROFESSOR ADEEL JAVED

INTRODUCTION:

For this final assessment, I analysed and predicted the class of the target variable from the breast cancer dataset using predictive modelling. The class variable which we will predicting consist of 85 occurrences of one class and 201 instances of another. Nine attributes, some of which are numeric and others which are nominal, are used to describe the instances in this dataset.

OBJECTIVE:

With the use of this analysis, I'm attempting to provide answers to the two questions below in the context of the results of our research.

- I'm attempting to predict if the Target variable's value represents a "Recurrence Event" or a "Non-Recurrence Event" in this research.
- I will evaluate several models and determine which one is more accurate and performing better overall.

DATASET INFORMATION:

Age: The patient's age. A ged 10 to 99, divided into 10 age groups.

Menopause: Twelve months following the last period for a woman. Divided into three distinct category types: Premeno, lt40, and ge40

Tumor-size: The size of the cancer tumour is represented by the tumor's size. 0 to 59, divided into 12 intervals.

Inv-Nodes: The number of lymph nodes in the armpit that have breast cancer spread apparent on a histological examination, ranging from 0 to 39.

Node caps: Cancer may disclose the risk of lymph node metastases even while the tumor's exterior appears to be confined. (Yes, No)

Degree of malignancy: The grade of cancer that can be seen under a microscope (1, 2, 3)

Breast- Which side of the breast is affected by breast cancer (Left, Right)

Breas Quadrant: Which of the four breast quadrants from the nipple area breast cancer occurred? (left-up, left-low, right-up, right-low, central.)

Irradiation: This medical procedure kills cancer cells. (Yes, No)

Methodology And Inferences:

I used SAS Miner to perform the analysis and used the procedures listed below to forecast the target variable, which is **Class**

File Import

I started by utilising the import node to import the dataset into the system. Using the file Import option in the properties window, the file is uploaded. Using the ellipses button, the dataset is uploaded.

The variables' roles were modified in the following step by selecting "Variable" from the properties menu. I designated the Class variable as "Target" and the remaining features as "Input" as they are independent variables.

Variables - FIMPORT

(none) not Equal to ...

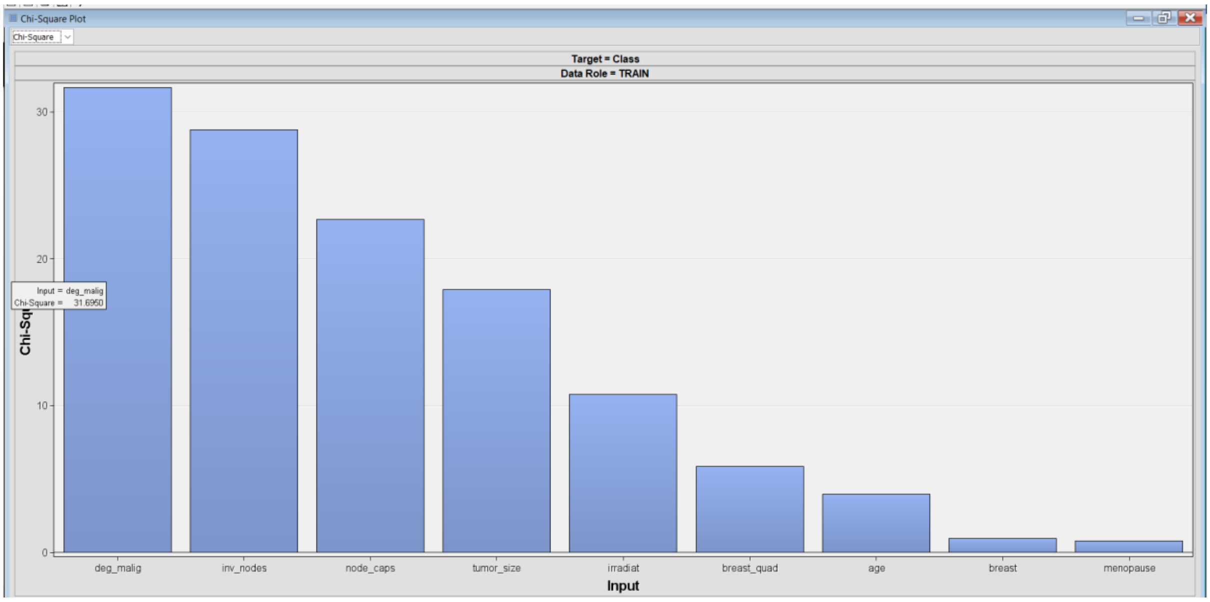
Columns: Label Mining

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Class	Target	Nominal	No		No	.	.
age	Input	Nominal	No		No	.	.
breast	Input	Nominal	No		No	.	.
breast_quad	Input	Nominal	No		No	.	.
deg_malig	Input	Interval	No		No	.	.
inv_nodes	Input	Nominal	No		No	.	.
irradiat	Input	Nominal	No		No	.	.
menopause	Input	Nominal	No		No	.	.
node_caps	Input	Nominal	No		No	.	.
tumor_size	Input	Nominal	No		No	.	.

Stat Explore

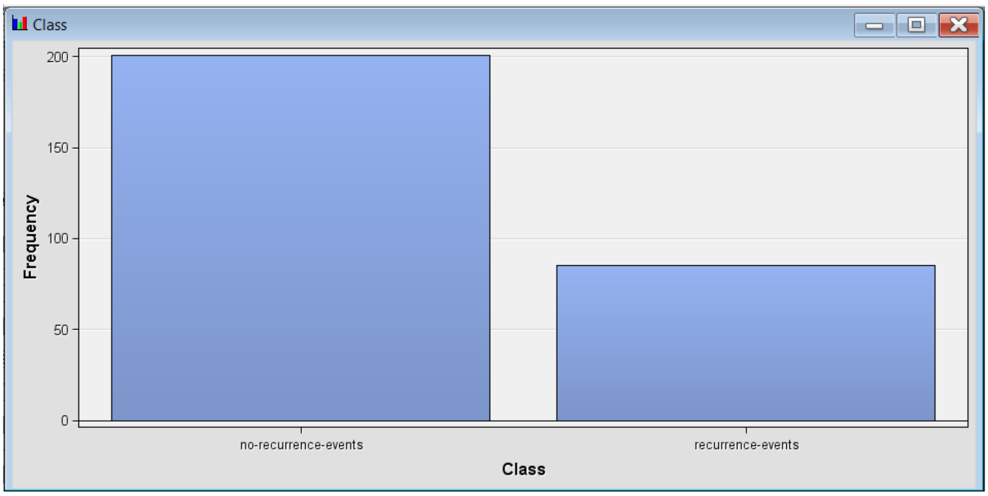
In this step, I added the stat explore node in order to study the class variable. The below screenshot shows the results from Stat Explore node.

In order to identify whether the variables are independent of if there is a relationship between categorical variables. From the results, we can see the degree of malignancy, inv-nodes, and node caps are highly involved in decision making.



Graph Explore

To view the graphical representation of the Target variable, I attached the Graph Explore node to the import node in the following step. The graph below shows distribution of recurrence and non-recurrence events



Data Partition

I divided the data into Train and Validation Datasets in this stage. To prevent underfitting and overfitting, I divided the total amount of data into an 80/20 split between the Train and Validation datasets.

Data Set Allocations	
Training	80.0
Validation	20.0
Test	0.0

The population distribution following data partition is depicted in the image below.

Summary Statistics for Class Targets

Data=DATA

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Class	.	no-recurrence-events	201	70.2797	Class
Class	.	recurrence-events	85	29.7203	Class

Data=TRAIN

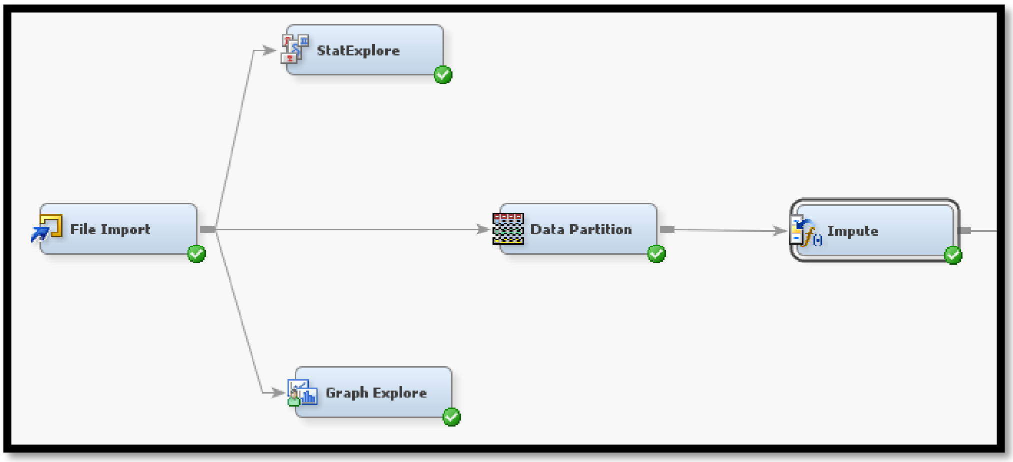
Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Class	.	no-recurrence-events	160	70.4846	Class
Class	.	recurrence-events	67	29.5154	Class

Data=VALIDATE

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Class	.	no-recurrence-events	41	69.4915	Class
Class	.	recurrence-events	18	30.5085	Class

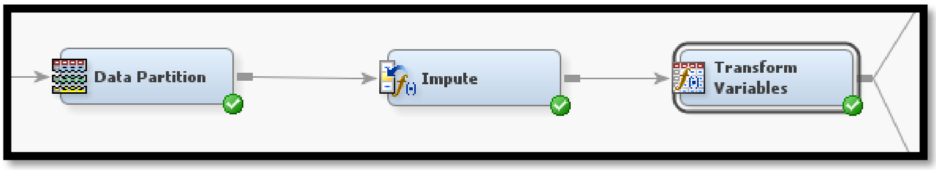
Impute Node

To reduce the possibility of models ignoring observation of missing values, I imputed the missing values in the dataset before training the model. In the impute node, new variables are made to replace any missing data.



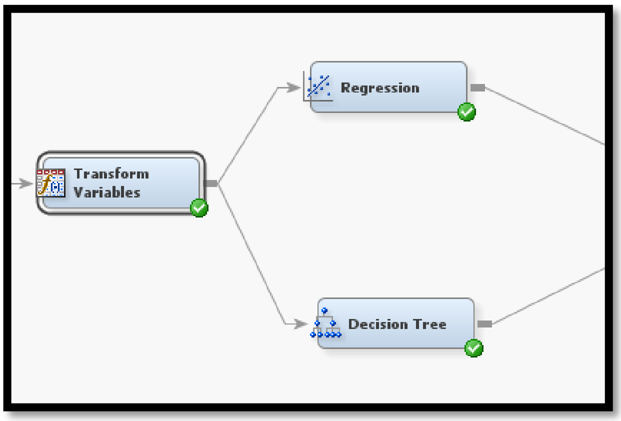
Data Transformation:

The variables for the model were transformed in this stage using the data transformation node. This phase assisted me in reducing variance, eliminating nonlinearity, enhancing additivity, and eliminating non-normality.



PREDICTIVE MODELS (LOGISTIC REGRESSION VS DECISION TREE):

As the target variable Class is binary, I made the decision to create a logistic regression and decision tree model.

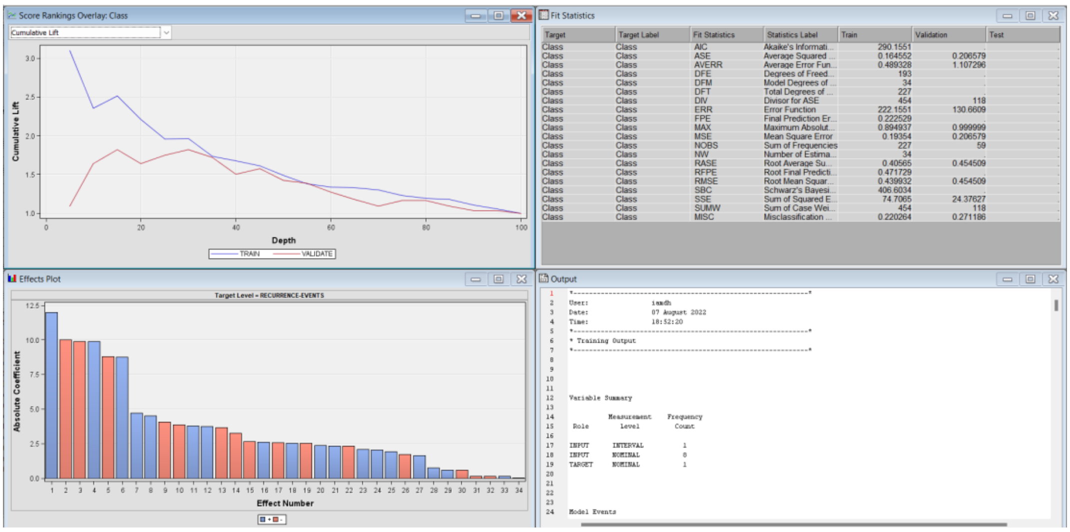


LOGISTIC REGRESSION:

I'm training prediction models in this step. Because we are doing the prediction on a classification variable, I am using logistic regression. The Logistic Regression node and the Data Transformation node are connected, and I have chosen Logit in the properties panel.

Class Targets	
Regression Type	Logistic Regression
Link Function	Logit

The outcome from the regression model is displayed in the screenshot below.



Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Class	Class	AIC	Akaike's Information Criterion	290.1551	.	.
Class	Class	ASE	Average Squared Error	0.164552	0.206579	.
Class	Class	AVERR	Average Error Function	0.489328	1.107296	.
Class	Class	DFE	Degrees of Freedom for Error	193	.	.
Class	Class	DFM	Model Degrees of Freedom	34	.	.
Class	Class	DFT	Total Degrees of Freedom	227	.	.
Class	Class	DIV	Divisor for ASE	454	118	.
Class	Class	ERR	Error Function	222.1551	130.6609	.
Class	Class	FPE	Final Prediction Error	0.222529	.	.
Class	Class	MAX	Maximum Absolute Error	0.894937	0.999999	.
Class	Class	MSE	Mean Square Error	0.19354	0.206579	.
Class	Class	NOBS	Sum of Frequencies	227	59	.
Class	Class	NW	Number of Estimate Weights	34	.	.
Class	Class	RASE	Root Average Sum of Squares	0.40565	0.454509	.
Class	Class	RFPE	Root Final Prediction Error	0.471729	.	.
Class	Class	RMSE	Root Mean Squared Error	0.439932	0.454509	.
Class	Class	SBC	Schwarz's Bayesian Criterion	406.6034	.	.
Class	Class	SSE	Sum of Squared Errors	74.7065	24.37627	.
Class	Class	SUMW	Sum of Case Weights Times Freq	454	118	.
Class	Class	MISC	Misclassification Rate	0.220264	0.271186	.

```
Event Classification Table

Data Role=TRAIN Target=Class Target Label=Class

  False      True      False      True
Negative   Negative   Positive   Positive

   41        151         9          26


Data Role=VALIDATE Target=Class Target Label=Class

  False      True      False      True
Negative   Negative   Positive   Positive

   13         38         3          5
```

Misclassification Tree			
	Detected as 0 (outcome= 0)	Detected as 1 (outcome = 1)	Total
Truly 0 (target = 0)	TN= 38	FP= 3	FP+TN = 41
Truly 1 (target = 1)	FN= 13	TP= 5	TP+FN = 18
Total	TN+FN= 51	TP+FP= 8	

Recall (R) = TP / (TP + FN) = 5/18 = 0.277

Precision (P) = TP / (TP + FP) = 5/8 = 0.625

F_1= 2P.R / (P + R) = 2(0.625) *(0. 277)/ (0.625+0.277)

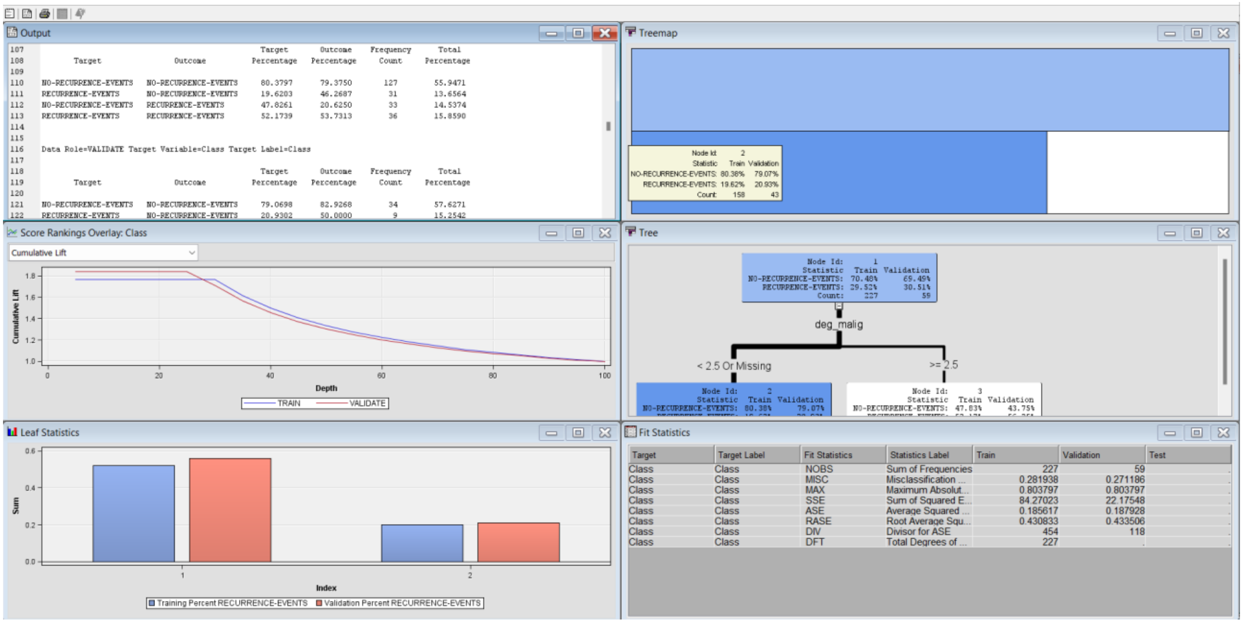
F_1 = 0.346/ 0.902

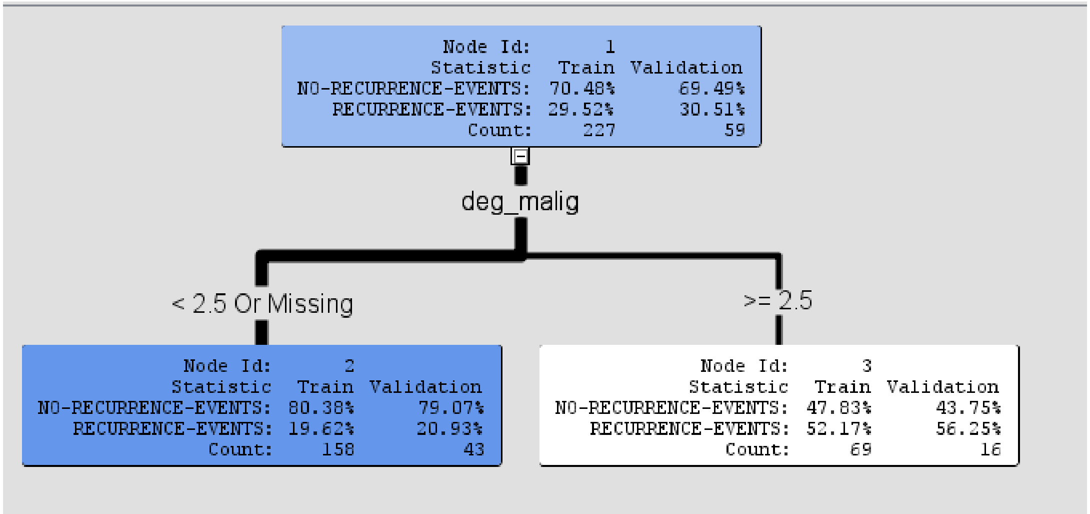
F_1 = *0.383*

Decision Tree

The second model in my analysis is a decision tree, which I have chosen. I linked the Data Transformation node to the Decision tree.

The decision tree's results are displayed in the screens below:





Event Classification Table

Data Role=TRAIN Target=Class Target Label=Class

False Negative	True Negative	False Positive	True Positive
31	127	33	36

Data Role=VALIDATE Target=Class Target Label=Class

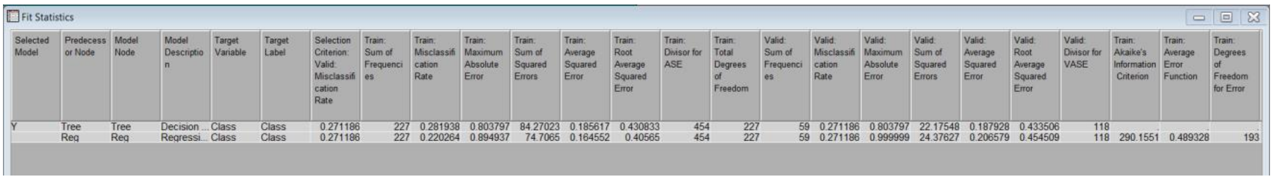
False Negative	True Negative	False Positive	True Positive
9	34	7	9

Misclassification Tree			
	Detected as 0 (outcome= 0)	Detected as 1 (outcome = 1)	Total
Truly 0 (target = 0)	TN= 34	FP= 7	FP+TN = 41
Truly 1 (target = 1)	FN= 9	TP= 9	TP+FN = 18
Total	TN+FN= 43	TP+FP= 16	

Recall (R) = TP / (TP + FN) = 9/18 = 0.5
Precision (P) = TP / (TP + FP) = 9/16 = 0.5625
F_1 = 2P.R / (P + R) = 2(0.5625) * (0.5)/ (0.5625+0.5)
F_1 = 0.5625 / 1.0625
F_1 = *0.5294*

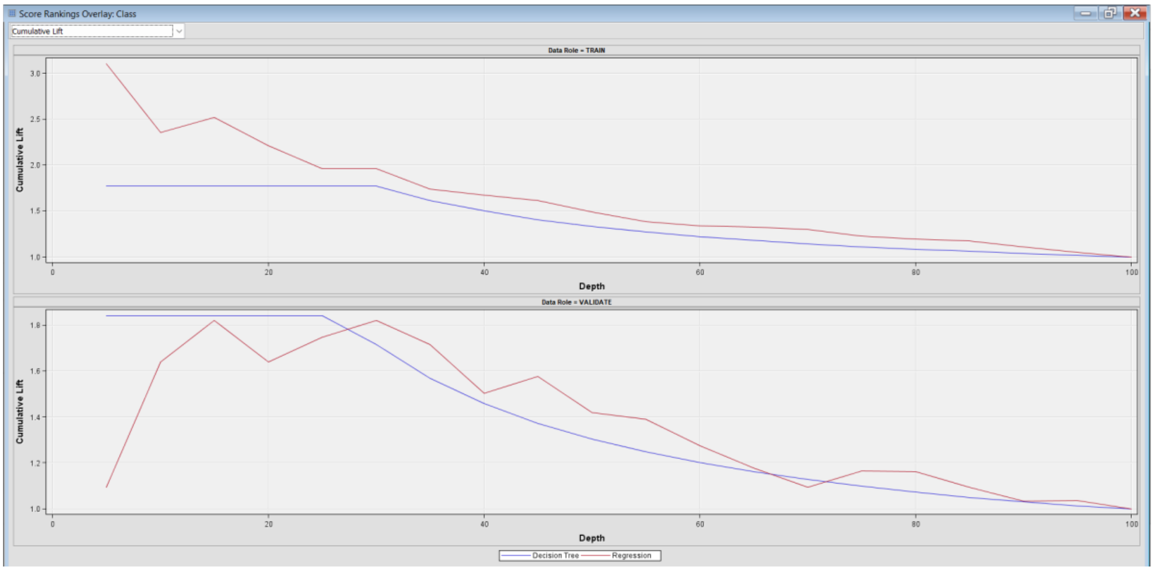
Model Comparison

In order to compare the two models, I connected them with a Model comparison node in this stage. The results of the model comparison are listed below:



Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Sum of Frequencies	Train: Misclassification Rate	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Train: Root Average Squared Error	Train: Divisor for ASE	Train: Total Degrees of Freedom	Valid: Sum of Frequencies	Valid: Misclassification Rate	Valid: Maximum Absolute Error	Valid: Sum of Squared Errors	Valid: Average Squared Error	Valid: Root Average Squared Error	Valid: Divisor for VASE	Train: Akaike's Information Criterion	Train: Average Error Function	Train: Degrees of Freedom for Error
Y	Tree	Tree	Decision ...	Class	Class	0.271186	227	0.281938	0.803797	84.27023	0.185617	0.430833	454	227	59	0.271186	0.803797	22.17548	0.187928	0.433506	118	.	.	.
	Reg	Reg	Regressi...	Class	Class	0.271186	227	0.220264	0.894937	74.7065	0.164552	0.40565	454	227	59	0.271186	0.999999	24.37627	0.206579	0.454509	118	290.1551	0.489328	193

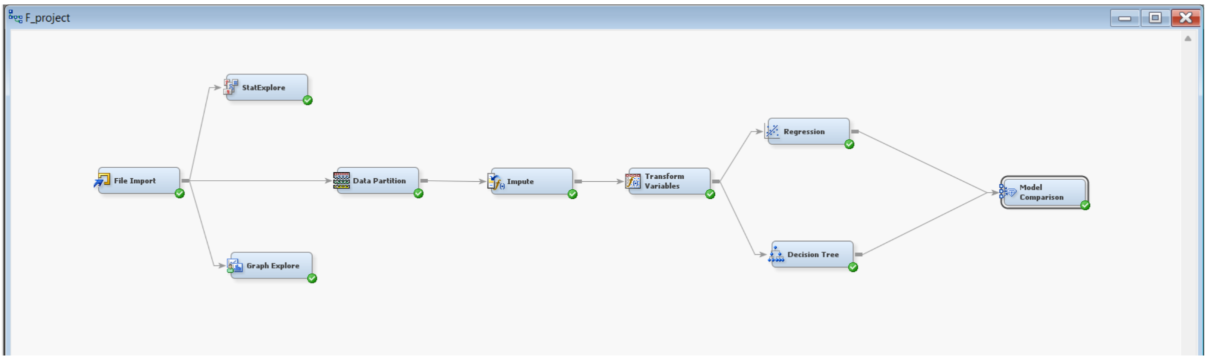


Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Tree	Decision Tree	0.27119	0.18562	0.28194	0.18793
	Reg	Regression	0.27119	0.16455	0.22026	0.20658

Final Diagram

The final diagram that I got is shown below:



Conclusion:

The following metrics are compared between the two models: MSE, Misclassification Rate, Recall, Precision, and F1 Score. The Precision and F1 Score of the Decision Tree model are higher than those of the Logistic Regression model, making it better to the Regression model.

DECLARATION:

I, Dhananjay Kumar, 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.