BAN 210 Final Project

BREAST CANCER DATA EXPLORATORY ANALYSIS

Submitted By:
Mitali Maheshwari

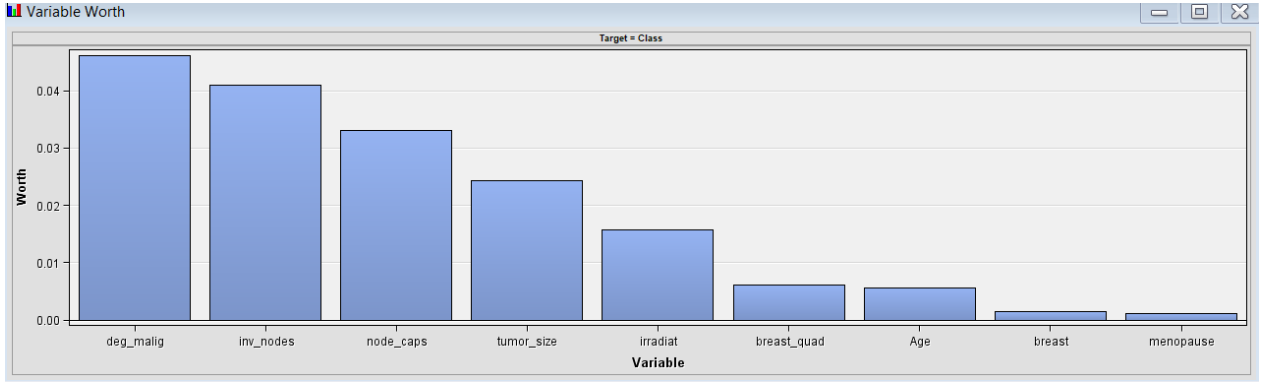
Objective

To determine whether the class of the data point represents a recurrence event or a non-recurrence of the Target variable.

To identify the best performing model for decision making

Dataset

There are nine qualities – some linear and some nominal – are used to characterize the instances.



Variables - FIMPORT

(none) not Equal to ... Apply Reset

Columns: Label Mining Basic Statistics

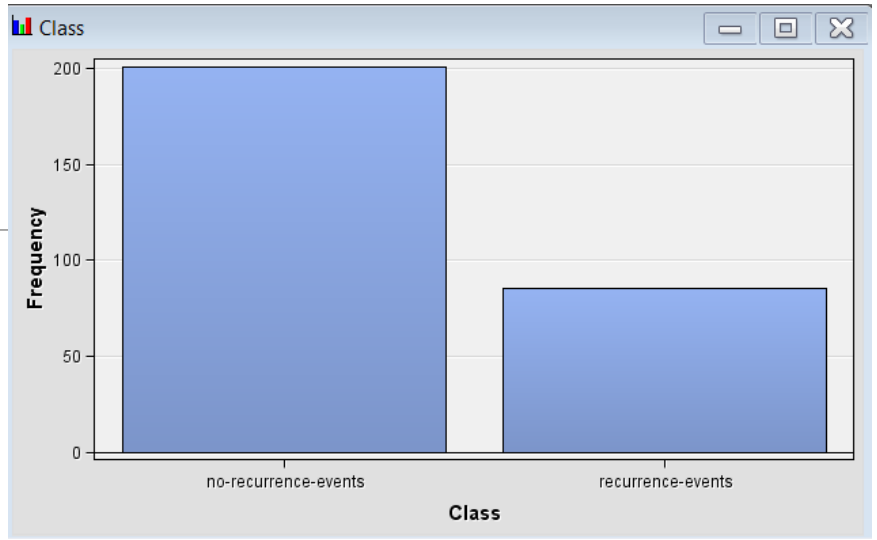
Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Age	Input	Nominal	No		No	.	.
Class	Target	Nominal	No		No	.	.
breast	Input	Nominal	No		No	.	.
breast_quad	Input	Nominal	No		No	.	.
deg_malig	Input	Nominal	No		No	.	.
inv_nodes	Input	Nominal	No		No	.	.
irradiat	Input	Nominal	No		No	.	.
menopause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

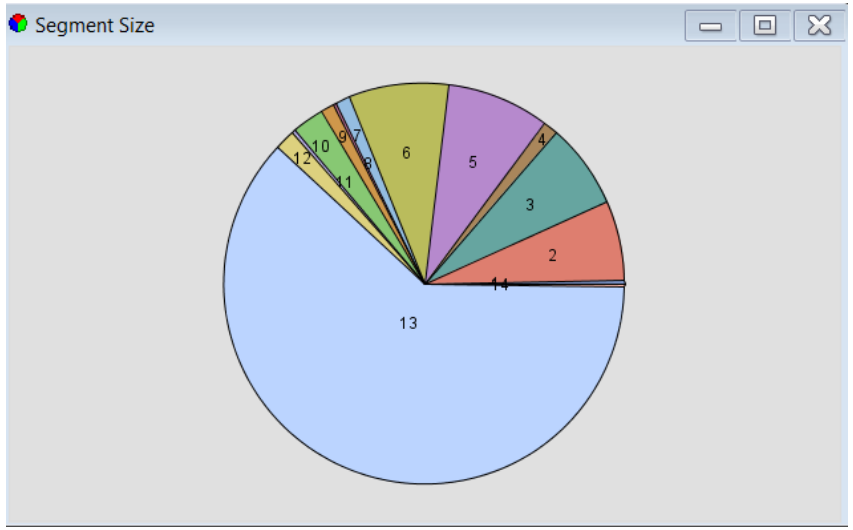
Explore... OK Cancel

Target Variable

There are 85 re-occurrences events and 201 instances of no recurrence events in the target variable.

In the Cluster node, we can see that the variables are well distributed.





Data Partition and Transformation

The data was divided into

60% for training

30% for validation

10% for testing

Partition Summary

Type	Data Set	Number of Observations
DATA	EMWS2.FIMPORT_train	286
TRAIN	EMWS2.Part_TRAIN	170
VALIDATE	EMWS2.Part_VALIDATE	86
TEST	EMWS2.Part_TEST	30

Model Building

Steps:

File Import

Stat Explore

Graph Explore

Multiplot

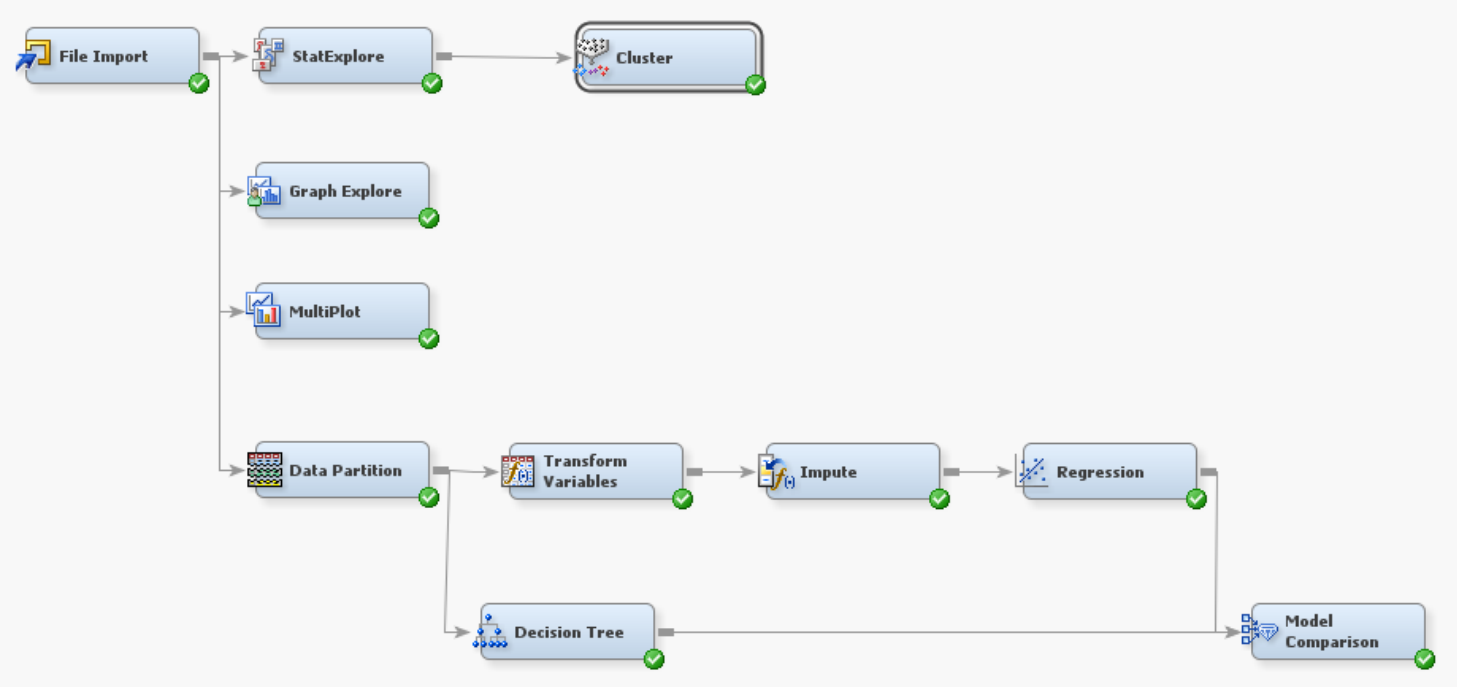
Data Partition

Data Transformation

Logistic Regression

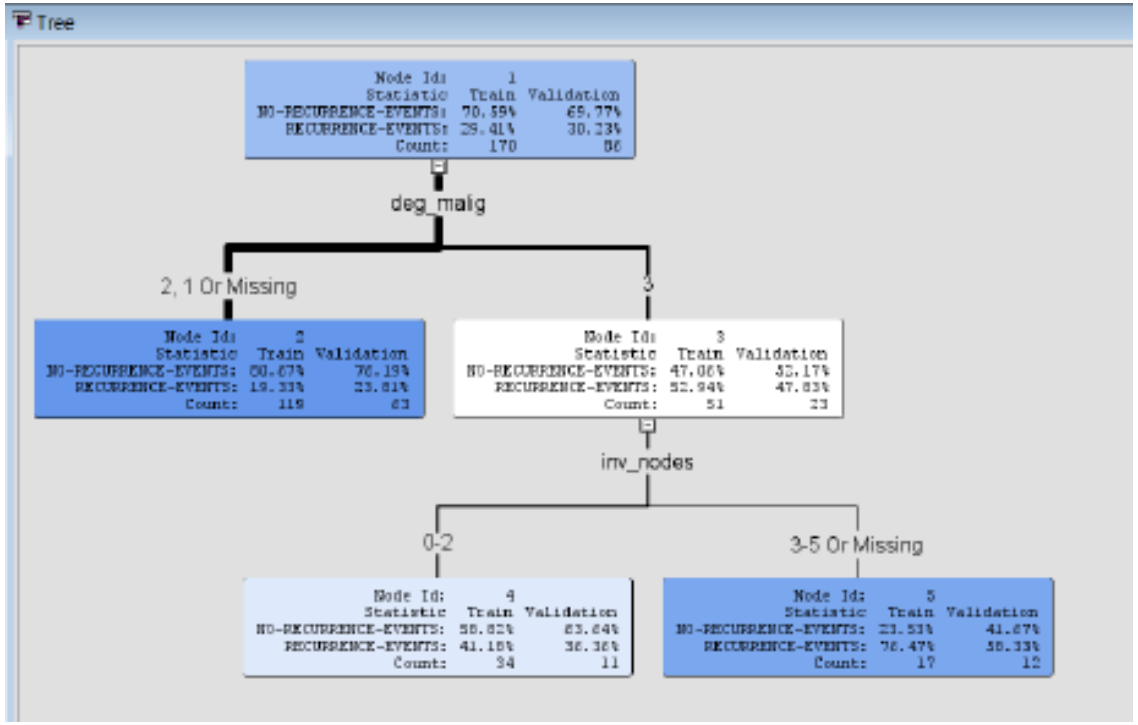
Decision Tree

Model Comparison

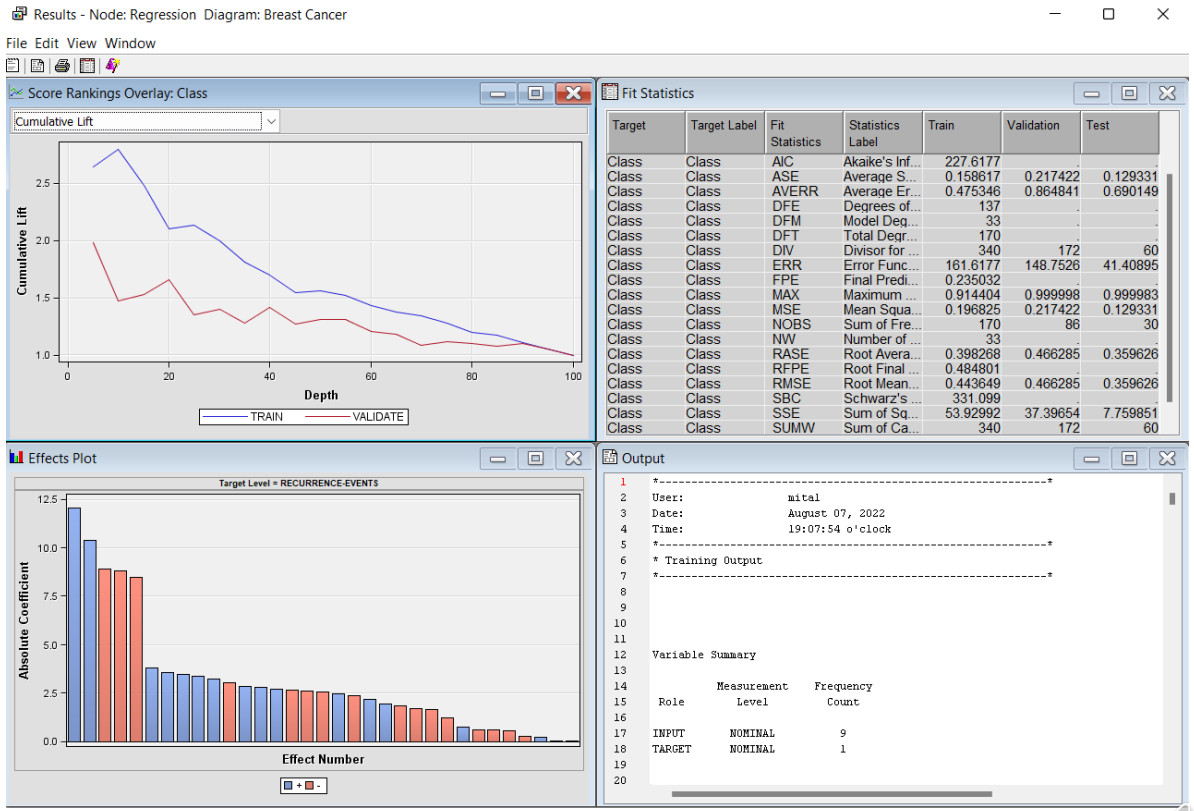


Model Output

Decision Tree



Regression



Model Comparison

The **Decision Tree model** has been selected for the fact that the decision parameters are slightly better than the regression model.

A model is better when the average squared error is low.

Variables Breast and Menopause are of no significance.

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	0.27907	0.17558	0.24118	0.20276
	Reg	Regression	0.31395	0.15862	0.22941	0.21742

Conclusion

The seriousness of the cancer can be determined with the variables – deg-malig and inv-nodes.

The recurrence of cancer is higher of Type 3 category of malig degree and is much higher with 3-5 inv nodes.

Thank You!