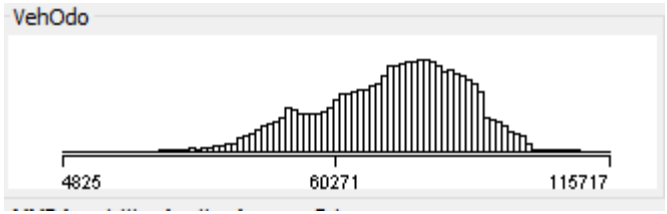
Project#2

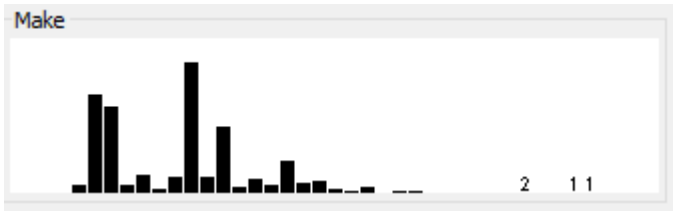
HAMZA ZAFAR
SYED JEHANDAD KAMAL
ABDUL MOIZ

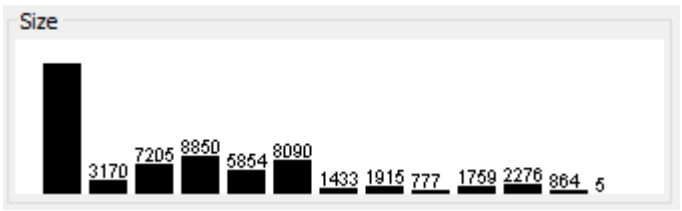
Don't Get Kicked!

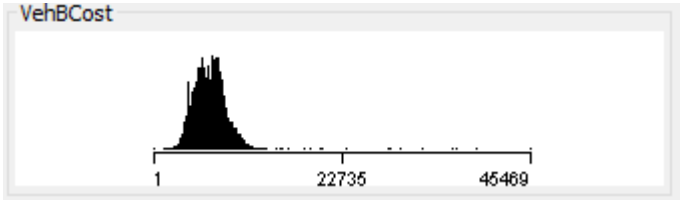
- Predict if a car purchased at auction is a lemon
  - There are columns34
  - The data set is split to 60% (72983)training and 40%(48707) testing.

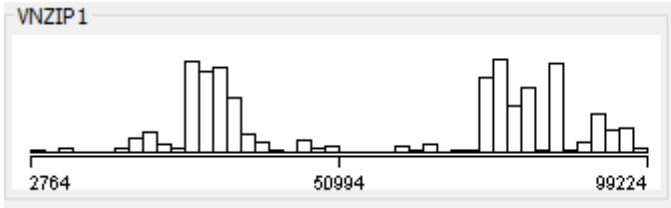
Nature of Data

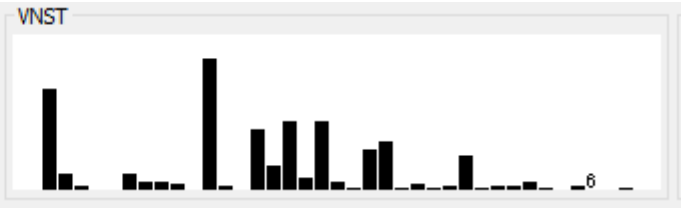












Issues with data

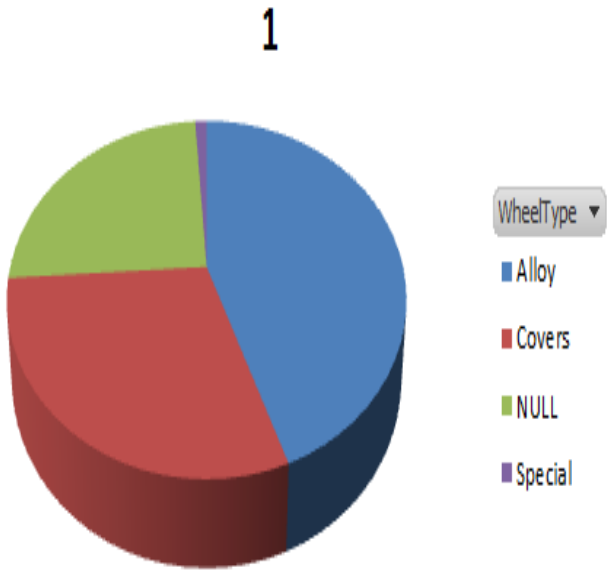
- Date contains many formats.
- Nominal values contain NULL Values
- Skewed data.
- Numeric data may contain outliers.

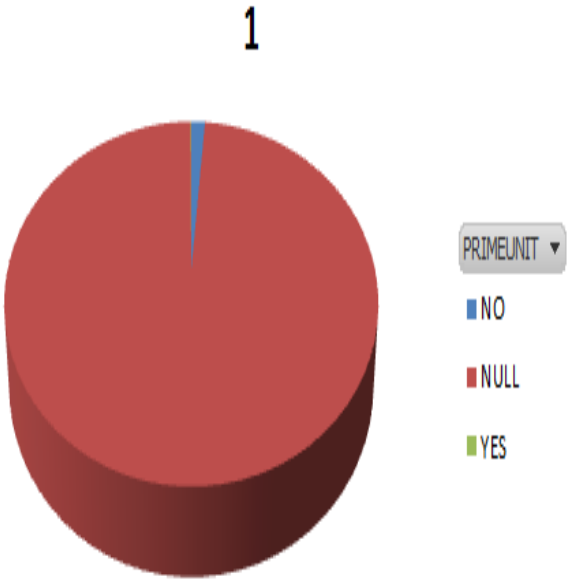
Data Treatment

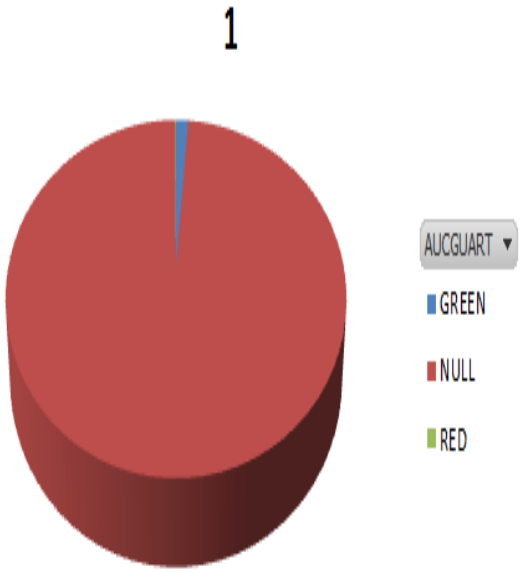
- Missing Values
  - Most Frequently and mean
- DATE:
  - Convert to same format (dd-mmm-yyyy)
  - drill down/ drill up approach is used to see data more deeply i.e. year to month to day.
- NULL:
  - Replace by most frequently values
  - Check the weight of 1 in null values and values having more weight to the null are kept as null.

NULL TREATMENT

Wheel Type	Prime Unit	Acguart

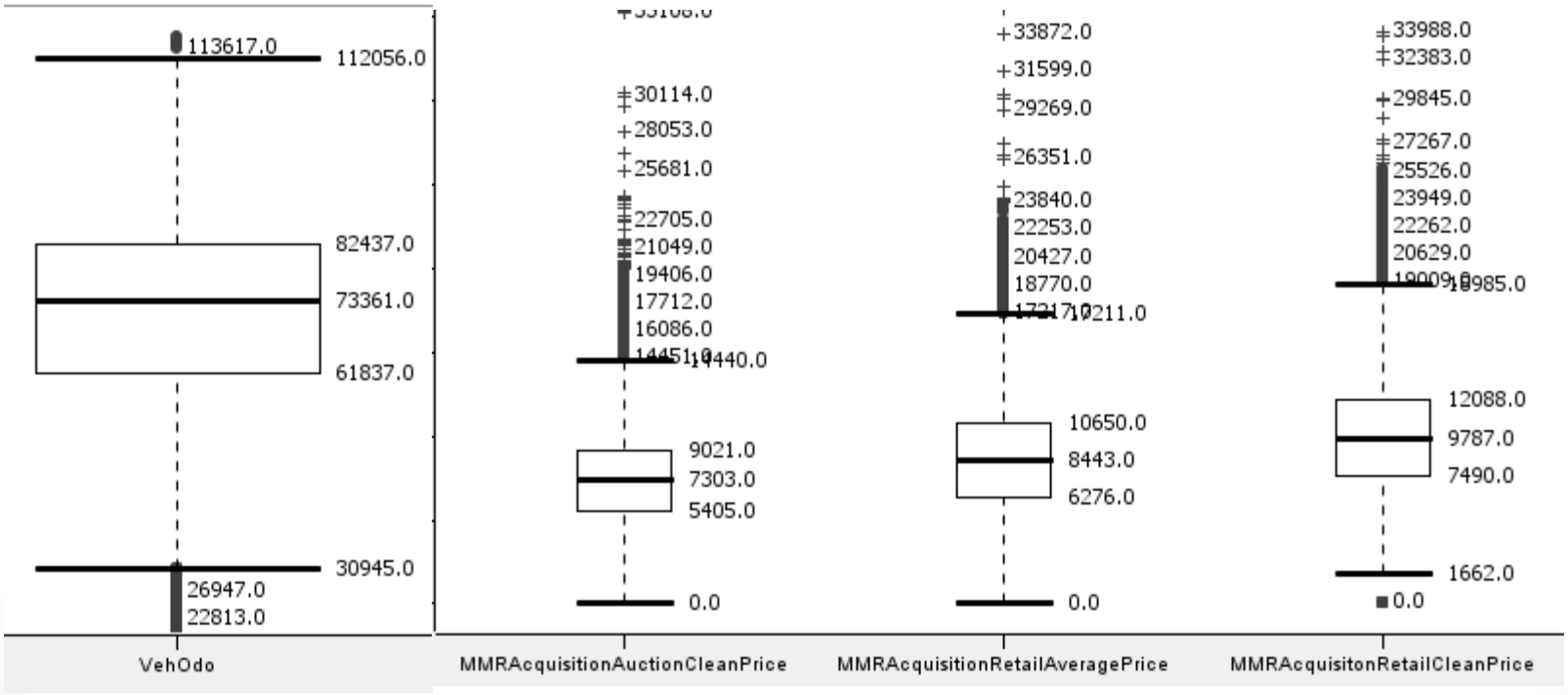






Data Treatment

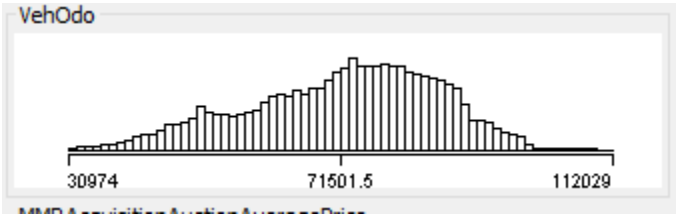
- Outliers:

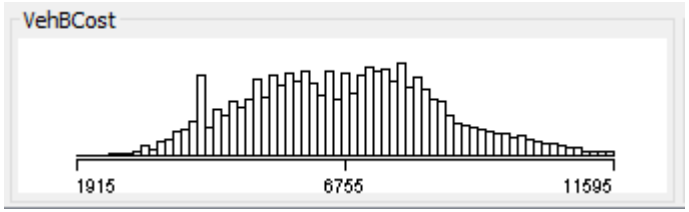


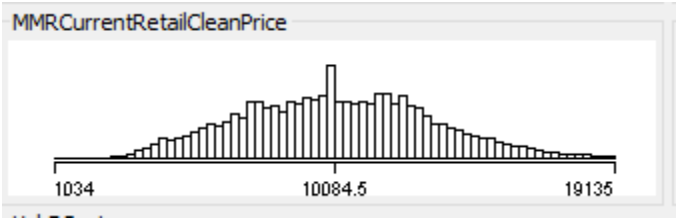
Transformation

COLUMN	VALUE	TRANSFORMED
Transmission	Auto	A
	Manual	M
Wheel type	Alloy	A
	Null	N
	Special	S
	Cover	C
Primeunit	Yes	Y
	No	N
	Null	Nl
Aucguart	Green	G
	Red	R
	Yellow	Y
	Null	N
topThreeName	chrysler	Ch
	Other	O
	Ford	F

After Treatment

- 





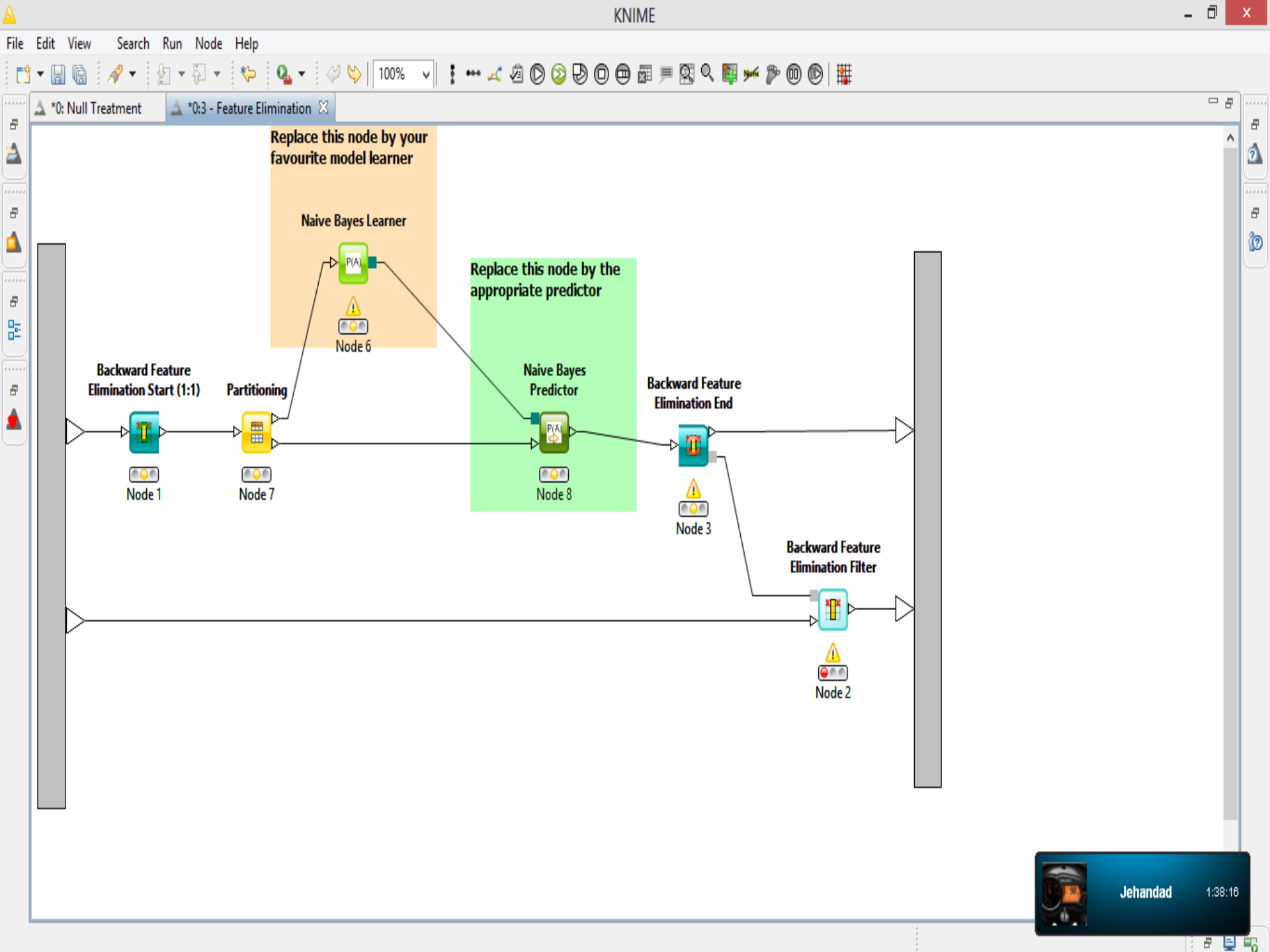
Feature Selection

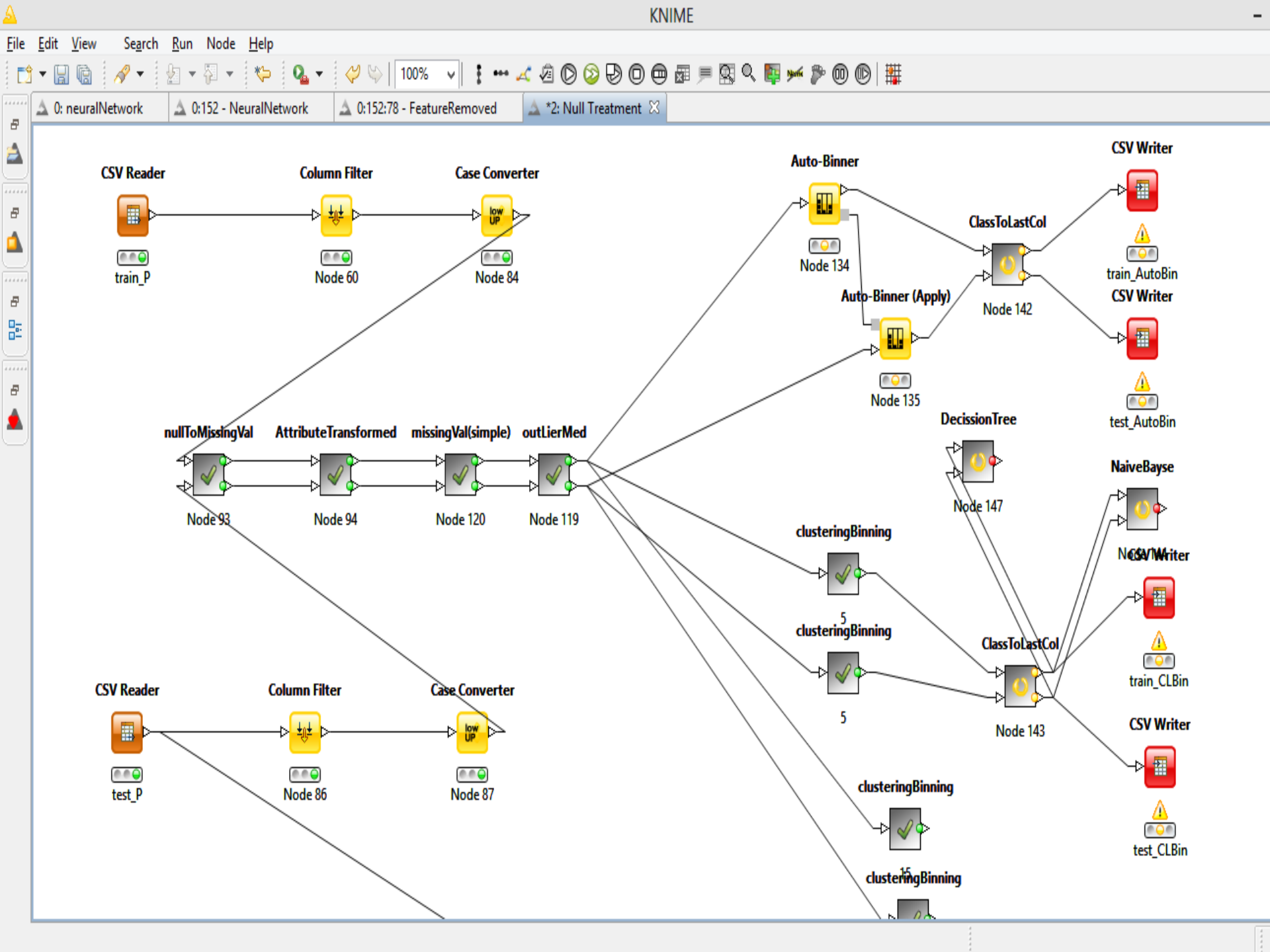
- Wheelid,Refid,buyerNo,Vnzip removed..

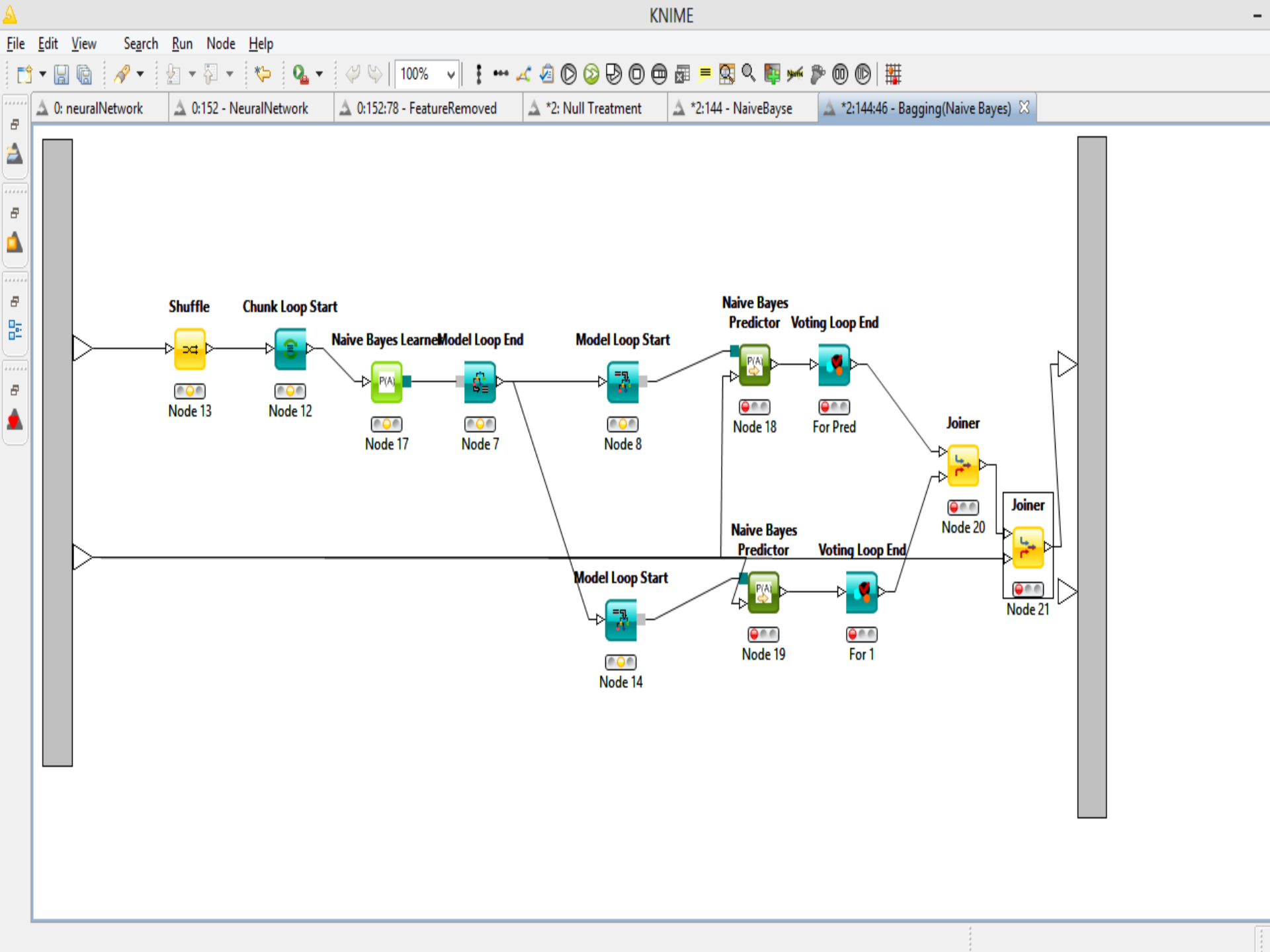
- Backward Feature Elemination
- InfoGain
- Chisquare

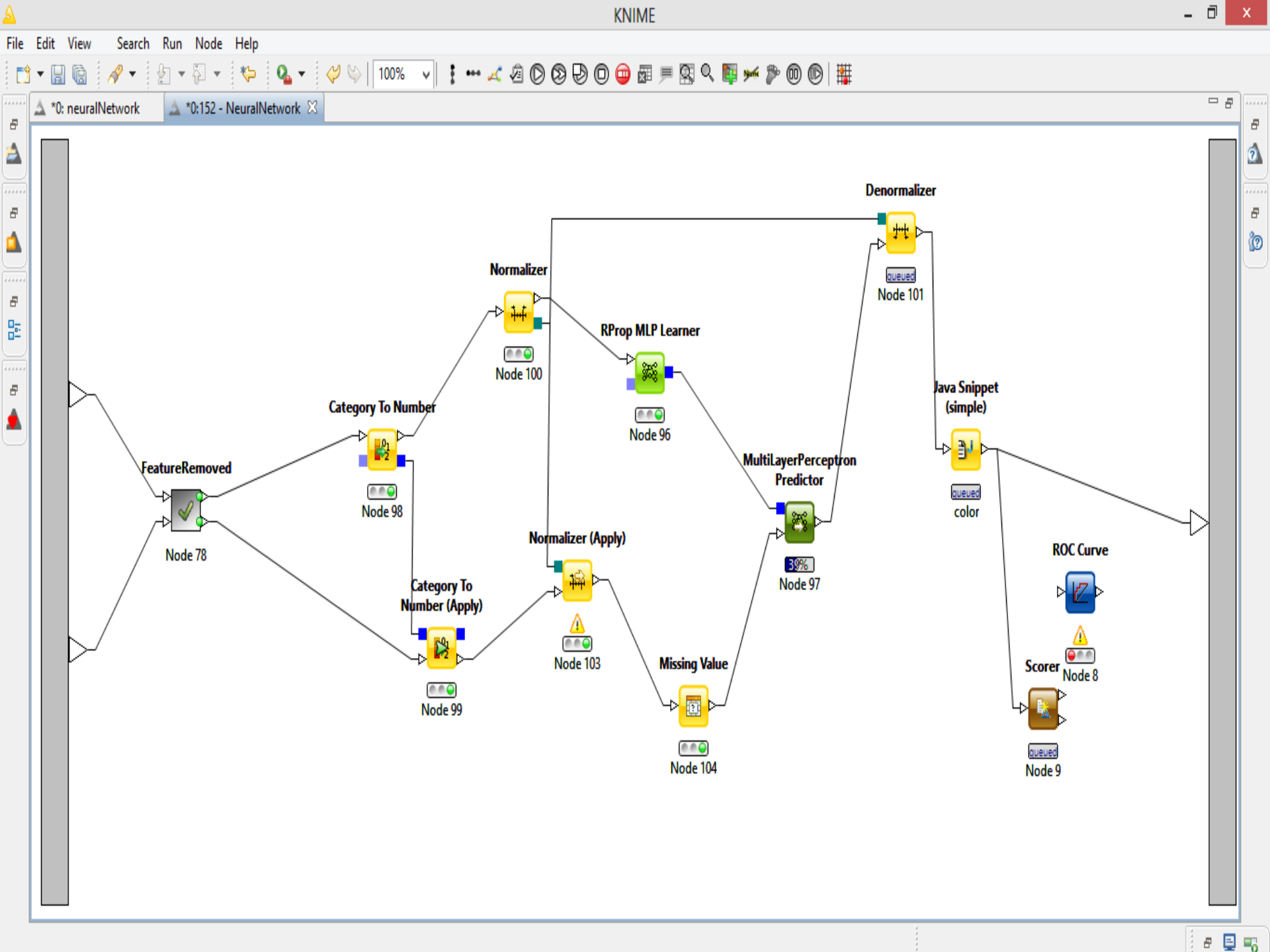
Binning

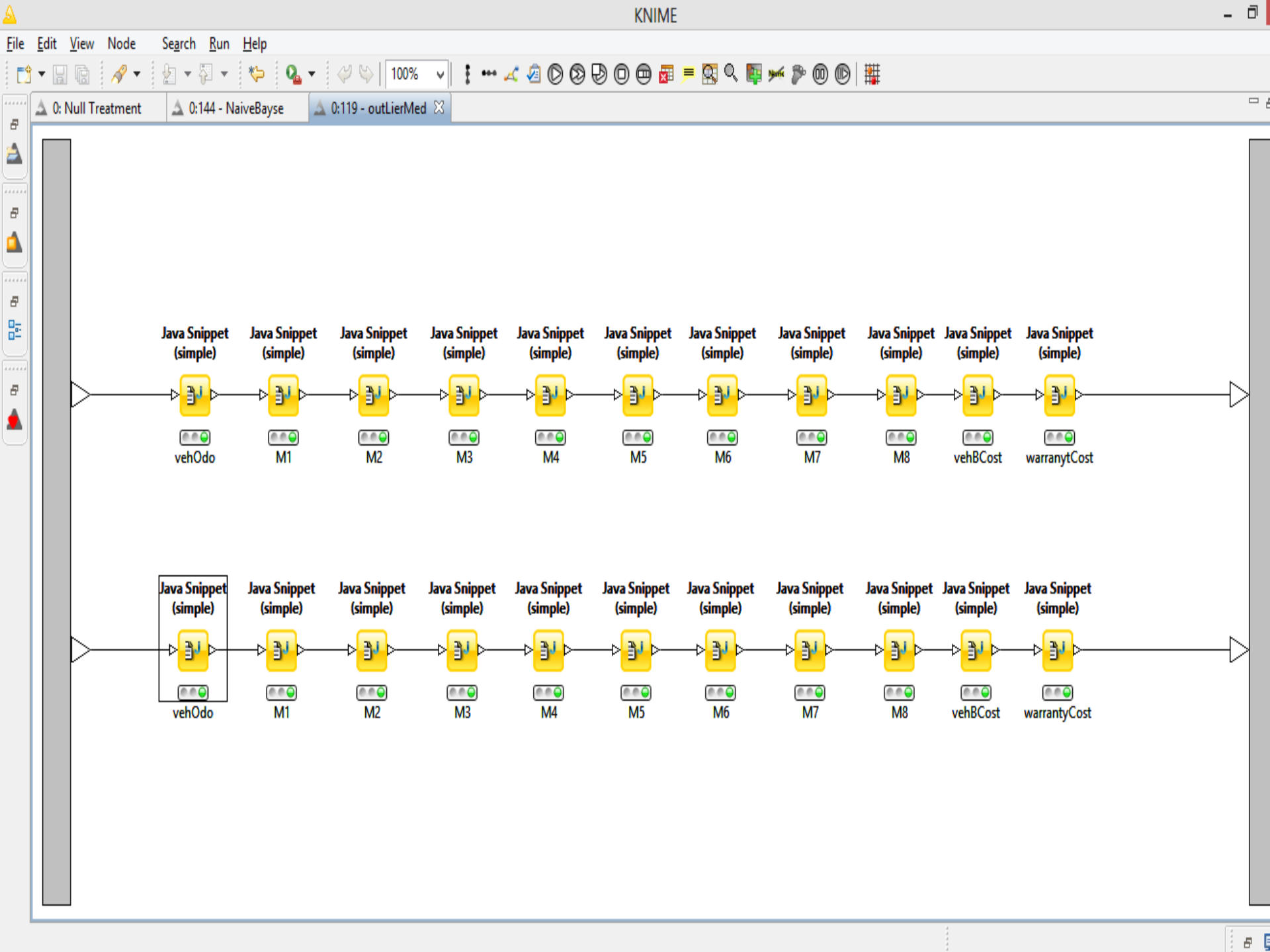
- fazzy c cluster is used
- Auto Binning


KNIME
File Edit View Search Run Node Help
100%
*0: Null Treatment
*0:3 - Feature Elimination
Replace this node by your favourite model learner
Naive Bayes Learner
Node 6
Replace this node by the appropriate predictor
Backward Feature Elimination Start (1:1)
Node 1
Partitioning
Node 7
Naive Bayes Predictor
Node 8
Backward Feature Elimination End
Node 3
Backward Feature Elimination Filter
Node 2
Jehandad
1:38:16


KNIME
File Edit View Search Run Node Help
100%
0: neuralNetwork
0:152 - NeuralNetwork
0:152:78 - FeatureRemoved
*2: Null Treatment
CSV Reader
Column Filter
Case Converter
train_P
Node 60
Node 84
Auto-Binner
Node 134
Auto-Binner (Apply)
Node 135
ClassToLastCol
Node 142
CSV Writer
train_AutoBin
CSV Writer
test_AutoBin
nullToMissingVal
AttributeTransformed
missingVal(simple)
outLierMed
Node 93
Node 94
Node 120
Node 119
DecissionTree
Node 147
NaiveBayse
clusteringBinning
5
clusteringBinning
5
ClassToLastCol
Node 143
CSV Writer
train_CLBin
CSV Writer
test_CLBin
clusteringBinning
clusteringBinning
CSV Reader
Column Filter
Case Converter
test_P
Node 86
Node 87


KNIME
File Edit View Search Run Node Help
100%
0: neuralNetwork
0:152 - NeuralNetwork
0:152:78 - FeatureRemoved
*2: Null Treatment
*2:144 - NaiveBayse
*2:144:46 - Bagging(Naive Bayes)
Shuffle
Node 13
Chunk Loop Start
Node 12
Naive Bayes Learner
Node 17
Model Loop End
Node 7
Model Loop Start
Node 8
Naive Bayes Predictor
Node 18
Voting Loop End
For Pred
Joiner
Node 20
Joiner
Node 21
Naive Bayes Predictor
Node 19
Voting Loop End
For 1
Model Loop Start
Node 14


KNIME
File Edit View Search Run Node Help
100%
*0: neuralNetwork
*0:152 - NeuralNetwork
Denormalizer
Node 101
Normalizer
Node 100
RProp MLP Learner
Node 96
Java Snippet (simple)
color
Category To Number
Node 98
FeatureRemoved
Node 78
MultiLayerPerceptron Predictor
39%
Node 97
Normalizer (Apply)
Node 103
Category To Number (Apply)
Node 99
Missing Value
Node 104
ROC Curve
Node 8
Scorer
Node 9


KNIME
File Edit View Node Search Run Help
100%
0: Null Treatment
0:144 - NaiveBayse
0:119 - outLierMed
Java Snippet (simple)
vehOdo
M1
M2
M3
M4
M5
M6
M7
M8
vehBCost
warranytCost
Java Snippet (simple)
vehOdo
M1
M2
M3
M4
M5
M6
M7
M8
vehBCost
warrantyCost

RESULT COMPARISION

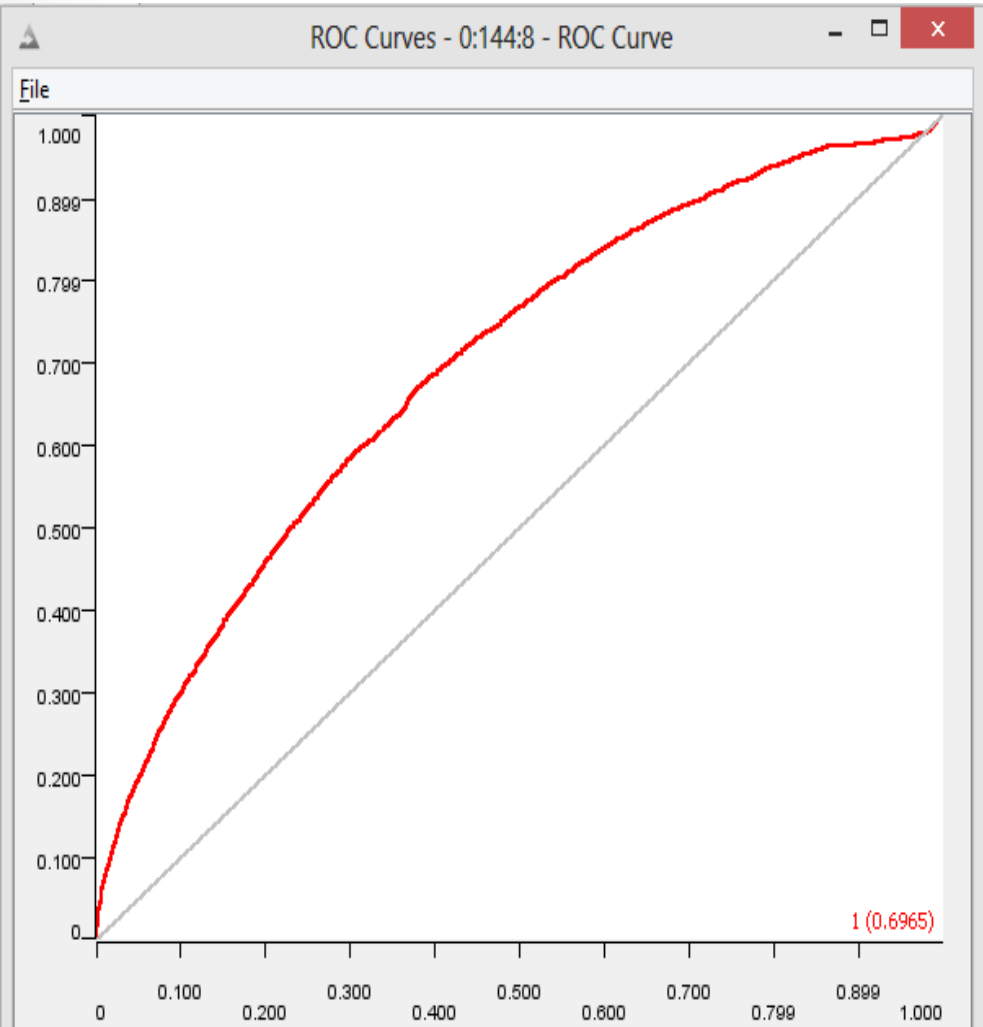
NAÏVE BAYSE	CLUSTER BINNING		AUTO-BINNER	
	ROC	FMEASURE	ROC	FMEASURE
Simple	0.696	0.318	0.693	0.317
Bagging	0.653	0.308	0.66	0.306
Boosting	0.694	0.315	0.693	0.317
DECISSION TREE				
Simple	0.597	0.318	0.603	0.322
Bagging	0.595	0.344	0.601	0.344
Boosting	0.607	0.316	0.624	0.317
Tree Ensembler	0.667	0.318	0.697	0.337

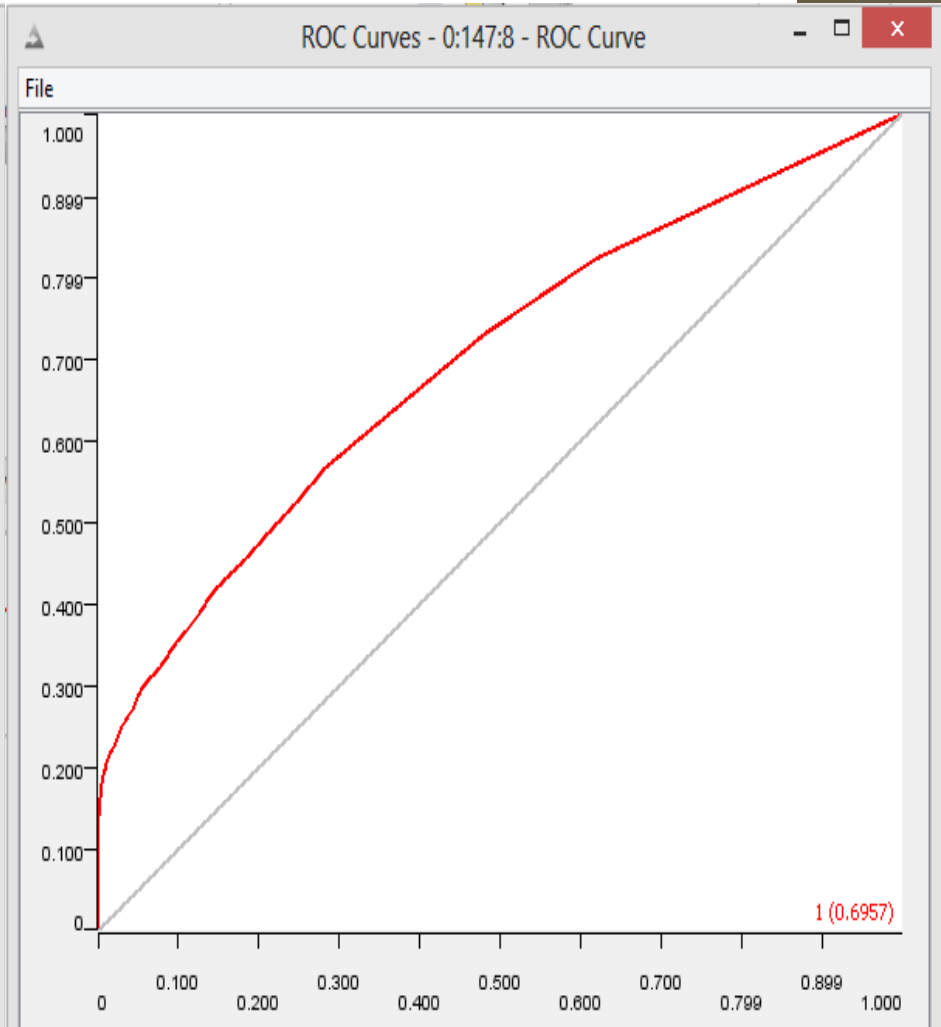
For Evaluation of model train data is spilted in 2/3 as training and 1/3 for testing

ROC

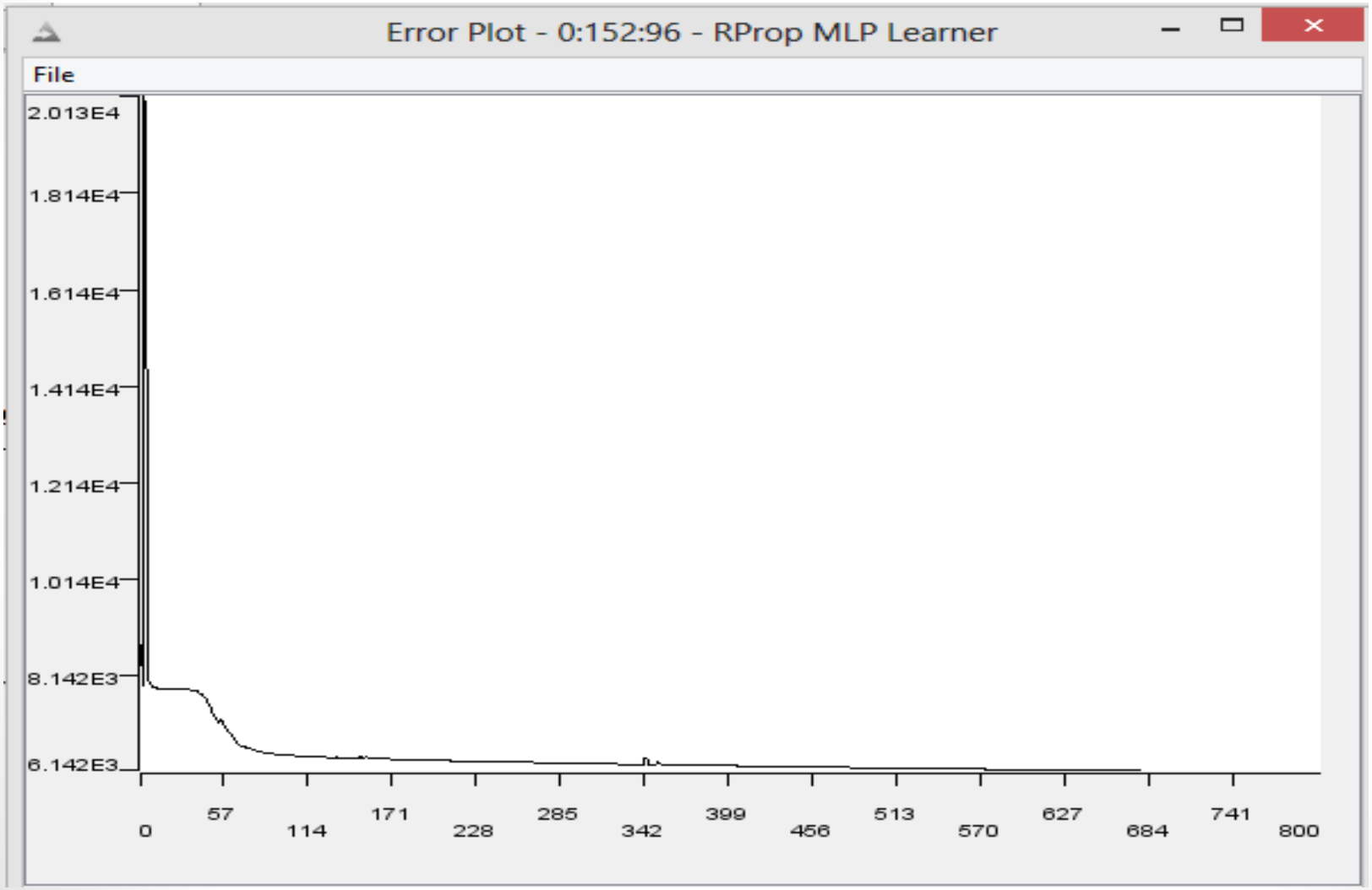
- NAÏVE BAYSE

DECISION TREE





SSE Neural Network



Gini

Learner	Gini
Naïve Bayse (Simple) (5bins)	0.12374(446)
Decission Tree Ensemebler (50 Models,All columns)	**0.09789**
Neural Network (800it,10hidenLayer,25neurons)	0.1053

Conclusion

- The Decision Tree with Ensembling Meta Node gave the best Roc and fmeasure when train data is spited in 2/3 as training and 1/3 as testing
- Naïve Bayes with Simple Learner gave the best Gini i.e. 0.12374(It scored 446 on leaderboard)
- Neural Network gave 0.1053, and may be improved with more numbers of iterations.