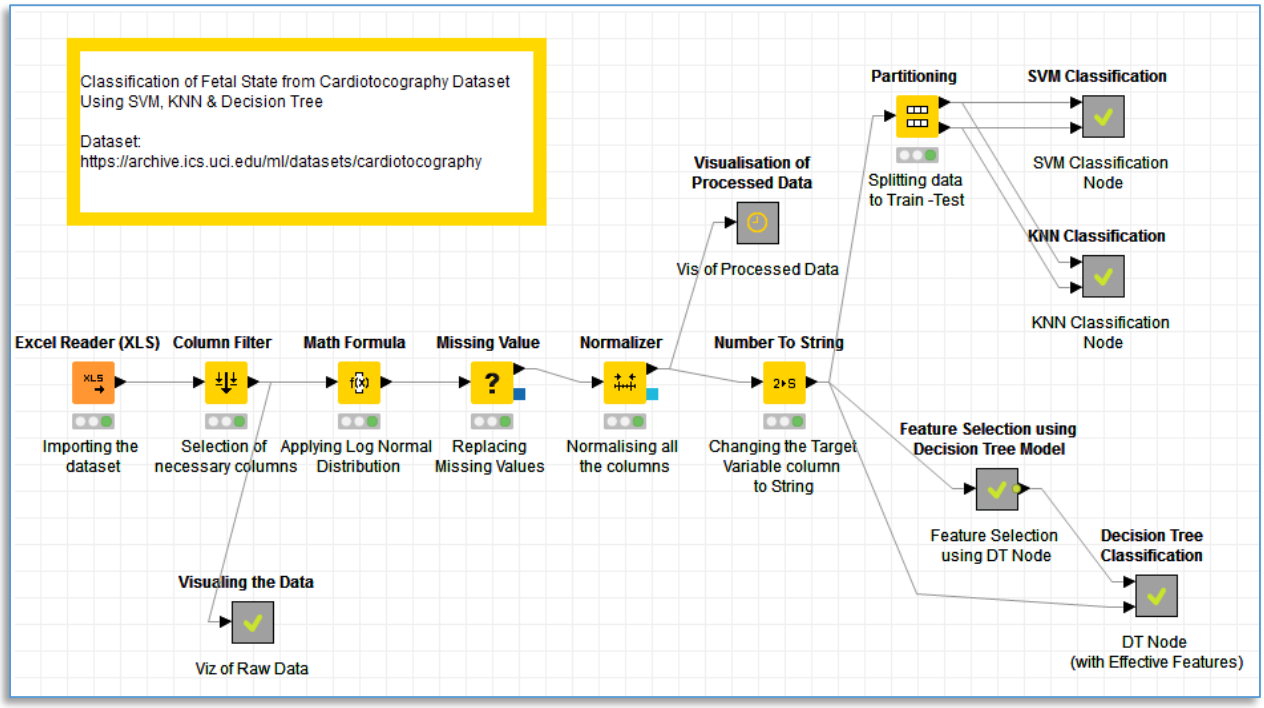
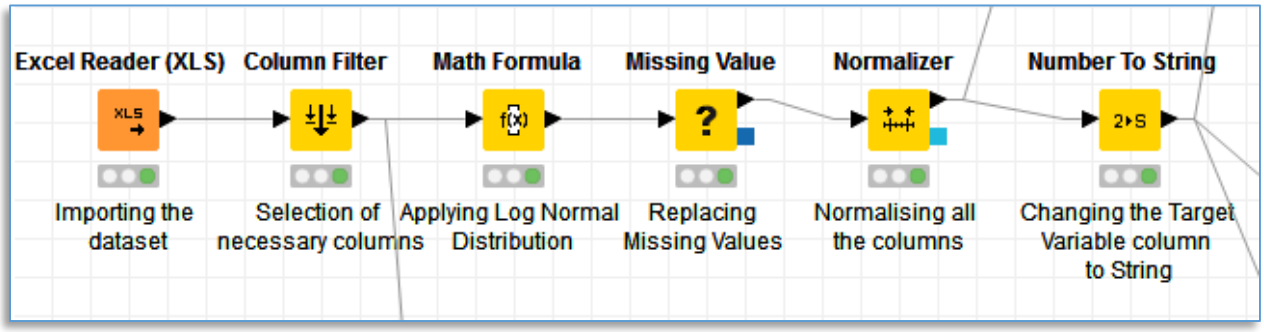
Classification of Fetal State from Cardiotocogram Data with SVM, KNN & Decision Tree using Knime

2018-19 CSD4060 Visual Data Analytics

Coursework 3 - Dr Leishi Zhang



Component(s) for data processing:



The Components used for data processing are explained below:

Excel Reader: The Excel dataset file is imported to Knime through this component. It has 2126 rows and 38 columns.

Table "Cardiotocogram.xlsx [Sheet1]" - Rows: 2126 | Spec - Columns: 38 | Properties

Colum...	Column Type	Column Index	Lower Bound	Upper Bound
Date	Local Date Time	0	1980-01-04T00...	1998-06-06T00...
b	Number (integer)	1	0	3296
e	Number (integer)	2	287	3599
LBE	Number (integer)	3	106	160
LB	Number (integer)	4	106	160
AC	Number (integer)	5	0	26
FM	Number (integer)	6	0	564
UC	Number (integer)	7	0	23
ASTV	Number (integer)	8	12	87
MSTV	Number (double)	9	0.2	7.0

Column Filter:
With this filter we select the necessary columns in the dataset. In this workflow, we have selected only 22 columns out of 38 columns which are present in the raw dataset for our analysis.

Table "default" - Rows: 2126 | Spec - Columns: 22 | Properties | Flow Variables

Columns: 22	Column Type	Lower Bound	Upper Bound	Column Index
LB	Number (integer)	106	160	0
AC	Number (integer)	0	26	1
FM	Number (integer)	0	564	2
UC	Number (integer)	0	23	3
ASTV	Number (integer)	12	87	4
MSTV	Number (double)	0.2	7	5
ALTV	Number (integer)	0	91	6
MLTV	Number (double)	0	50.7	7
DL	Number (integer)	0	16	8
DS	Number (integer)	0	1	9
DP	Number (integer)	0	4	10
Width	Number (integer)	3	180	11
Min	Number (integer)	50	159	12
Max	Number (integer)	122	238	13

Math Formula:

This component applies log normal distribution to the FM attribute column. This appends a new column FM(Log) to the table with the generated log calculation against each row of the table.

I FM	D FM (Log)
0	?
0	?
0	?
57	1.756
147	2.167
489	2.689
273	2.436
290	2.462
251	2.4
317	2.501
557	2.746
304	2.483

Missing Value:

This component replaces the missing values to user defined input.

In this case the log transformation changed the missing values '?' in FM column to 0.001 in FM (log) column.

I FM	D FM (Log)
0	0.001
0	0.001
57	1.756
147	2.167
489	2.689
273	2.436
290	2.462
251	2.4
317	2.501
557	2.746
304	2.483
272	2.435
219	2.34

Normalizer:

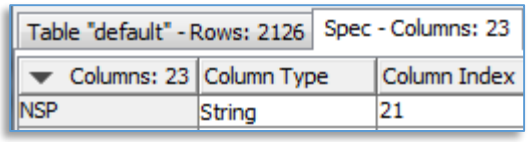
This component normalizes the different range of numerical attributes.

In the Figure, ASTV and ALTV are Integers and have a different scale, Therefore we have normalized these columns along with the MSTV and MLTV columns by rescaling and changing the attribute type to double.

I ASTV	D MSTV	I ALTV	D MLTV
61	0.5	40	6.2
70	0.3	69	5.1
57	1.2	54	12.8
58	1.3	53	13.2
39	0.8	38	5.5
41	0.8	29	6.4

D ASTV	D MSTV	D ALTV	D MLTV
0.653	0.044	0.44	0.122
0.773	0.015	0.758	0.101
0.6	0.147	0.593	0.252
0.613	0.162	0.582	0.26
0.36	0.088	0.418	0.108
0.387	0.088	0.319	0.126

Number to String:
This component changes the Target Variable numeric attribute NSP to string attribute.

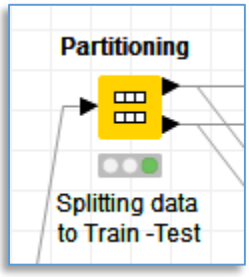


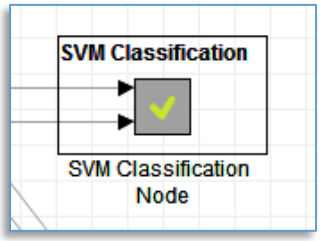
Component(s) for data mining:

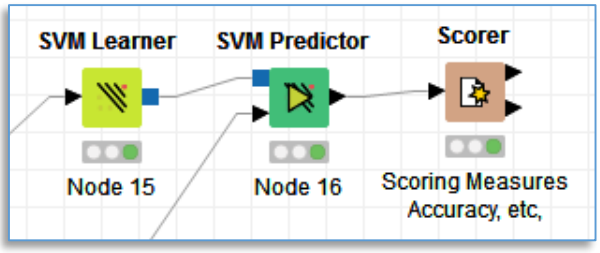
We have applied 3 Classification methods in this analysis.

SVM Classification:

We have used SVM Learner, SVM Predictor and the Scorer Components.
The processed data is splitted to 80% Train and 20% Test through the Partinioning Component. Then we have used the train data with SVM Learner and the test data with SVM Predictor. The output of the SVM Classification results can be viewed through Scorer Component.

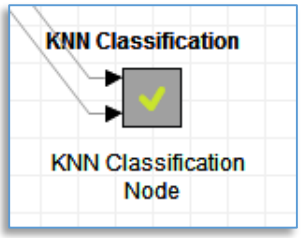


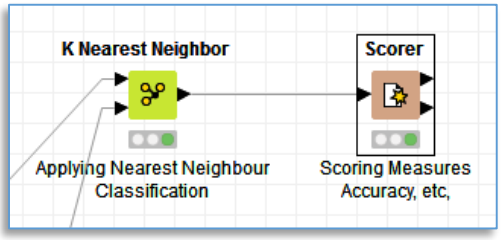




KNN Classification:

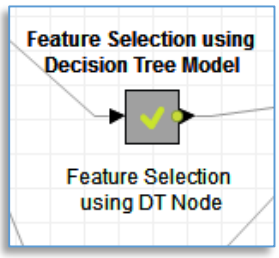
We have used KNN Component and the Scorer.
The processed data is splitted to 80% Train and 20% Test through the Partinioning Component.
Both the Train and Test data are used as input to the KNN Component. The output of the KNN Classification results can be viewed through Scorer Component.





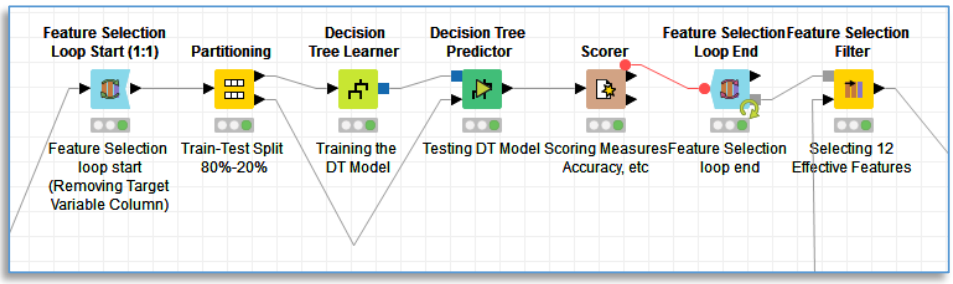
Feature Selection and Decision Tree:

We have used the Feature selection components to select the most effective features for the Decision tree algorithm.



In the Feature Seection Node, Feature Selection Start and End Loop Nodes continuously Partions the data in small chunks and runs the Decision Tree Learner and Predictor Components.

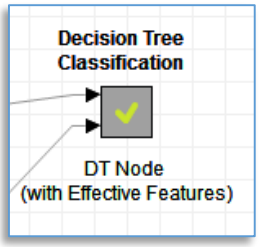
The Scorer passes the score of each loop to the Feature Selection Loop End Component, where it is stored.



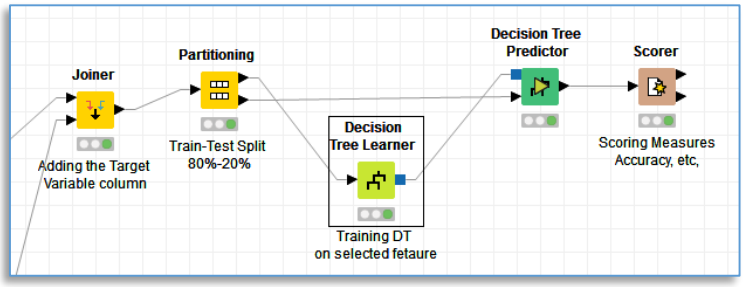
Through the Feature Selection Filter we can select a model with higher accuracy and optimum number of features.

The Decision Tree Classification Components uses the selected features.

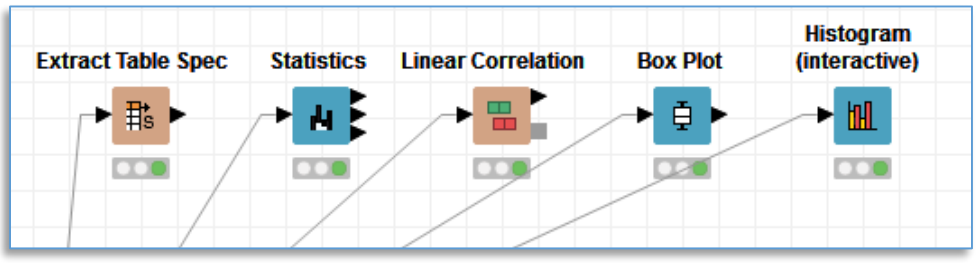
The Joiner which appends the Target Variable NSP with the selected features received as input from Feature Selection Components.



The full dataset is partioned to train and test and finally applied to Decision Tree Learner and Predictor with selected features. The Scorer shows the accuracy and scoring statistics.



Component(s) for visualizing the data and/or data mining results:



The Components used for Data Visualization are the following:

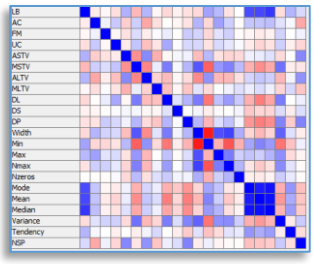
Extract Table Spec:
This Component extracts the table lower and upper bounds of the numerical attributes.

Column...	Lower Bound	Upper Bound
LB	106	160
AC	0	26
FM	0	564

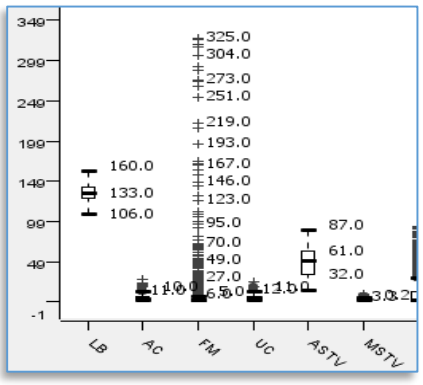
Statistics:
This Component gives the statistical analysis of the each attribute along with a histogram view.

Column	Min	Max	Mean	Std. dev...	Histogram
LB	106	160	133.304	9.841	106 160

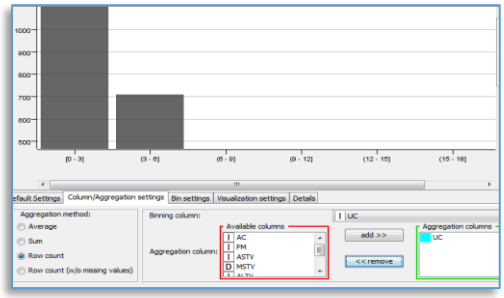
Linear Correlation:
This component helps us to view the correlation between different attributes with the correlation matrix.



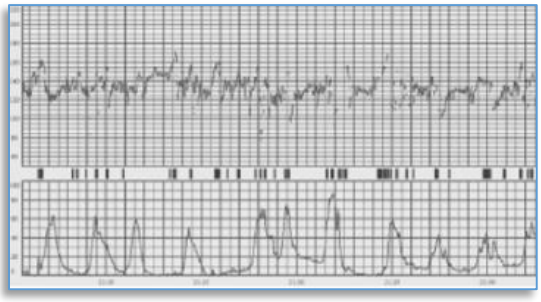
Box Plot:
This Component helps us to view the outliers and the IQR and, median of the attributes.



Histogram with Interactive View:
This is a component with which we can have a histogram view.
The component is interactive, hence we can select and change the attributes and analyse.



Goal (research question) of the analysis:



Cardiotocography is widely used in prenatal care. FHR has been recognized as an important indicator of fetal status and considered to be sufficiently accurate for analysis.
It is the visual representation of Fetal Heart Rate and uterine contractions through a dopler diagram(Fig)

Our goal is to select important features and classify the data to the following (NSP Column):

1. Normal: All features assures fetal in good health and no ongoing risk factors.
2. Suspect: Correct any underlying causes, Inform an obstetrician of possible risks.
3. Pathologic: Obtain a review by an obstetrician, suspected placental abruption or suspected uterine rupture, consider fetal blood sampling, etc

In this problem, we have used this CTG data with 3 classification methods (Decision Tree, SVM and KNN) and compared them. The metrics Accuracy, Sensitivity, Specificity, F-Score are used to evaluate them.
In order to improve the accuracy, feature selection methods are also applied with Decision Tree and the performance of classifiers is compared.

Based on the guidelines issued by National Institute for Health and Care Excellence 2017. NSP is a criterion classifying fetal state.

Dataset:
The dataset of cardiotocography is downloaded from UCI Machine Learning Repository. It contains 2126 instances and 38 attributes.
https://archive.ics.uci.edu/ml/datasets/cardiotocography

Attribute information of CTG dataset

LB	baseline value
AC	accelerations
FM	foetal movement
UC	uterine contractions
ASTV	percentage of time with abnormal short term variability
mSTV	mean value of short term variability
ALTV	percentage of time with abnormal long term variability
mLTV	mean value of long term variability
DL	light decelerations
DS	severe decelerations
DP	prolongued decelerations
Width	histogram width
Min	low freq. of the histogram
Max	high freq. of the histogram
Nmax	number of histogram peaks
Nzeros	number of histogram zeros
Mode	histogram mode
Mean	histogram mean
Median	histogram median
Variance	histogram variance
Tendency	histogram tendency: -1=left assymetric; 0=symmetric; 1=right assymetric
NSP	Normal=1; Suspect=2; Pathologic=3

Classification Algorithm:

Decision Tree:
Decision trees are powerful and popular tools for classification. It is a popular classifier which is simple and easy to implement. It requires no domain knowledge or parameter setting and can handle high dimensional data. Hence it is more appropriate for exploratory knowledge discovery. The performance of decision trees can be enhanced with suitable attribute selection. We have used Featured Selection components to find out effective features.

K Nearest Neighbours:

KNN is a non-parametric, lazy learning algorithm. Its purpose is to use a database in which the data points are separated into several classes to predict the classification of a new sample point.

KNN is effective for a classification study when there is little or no prior knowledge about the distribution data. It is a simple algorithm providing high accuracies, with no assumption about the data.

SVM:

Support Vector Machine is a discriminative classifier formally defined by a separating hyperplane.

Therefore, for given labeled training data the algorithm outputs an optimal hyperplane which categorizes new examples. SVM is highly effective to separate the data to classes.

Feature Selection:

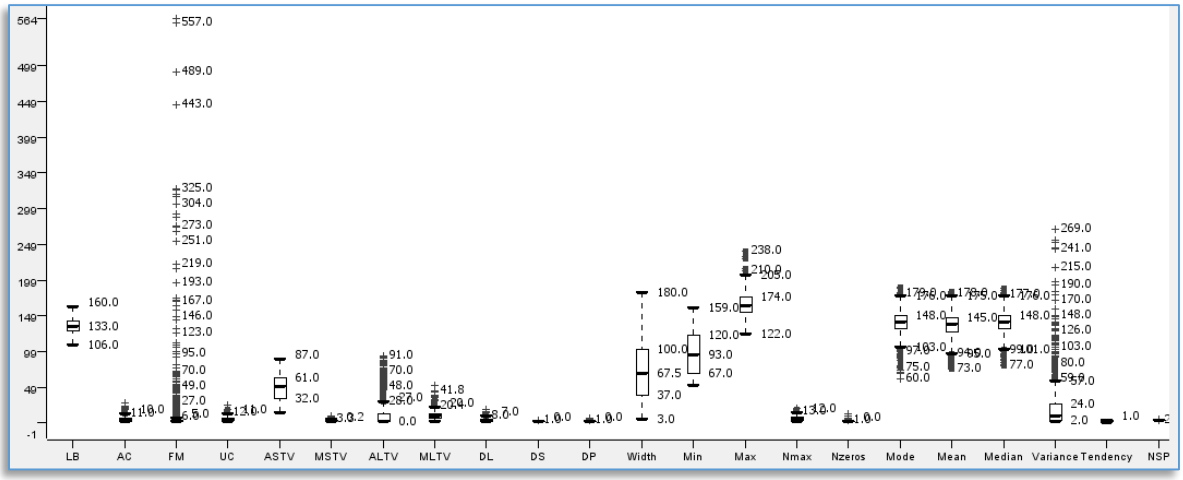
Feature selection is frequently used as a preprocessing step to machine learning. It is a process of choosing a subset of original features so that the feature space is optimally reduced according to a certain evaluation criterion. Feature selection has a main goal of finding a feature subset that produces higher classification accuracy.

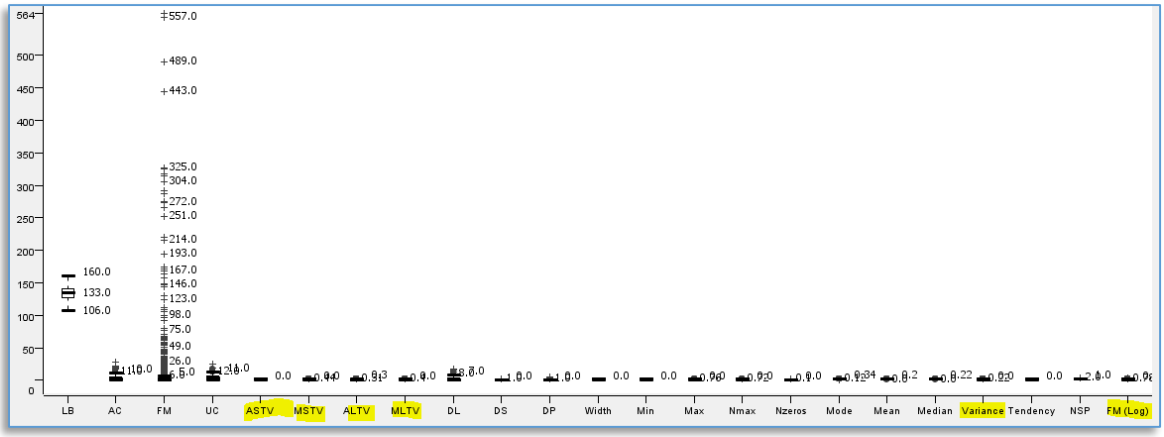
How does your workflow help achieve the goal:

In the Knime workflow the column filter selects the 22 attributes from 38 attributes of the original dataset.
To visualize the data we used Statistics, Histogram and Box plot components.
Both Histogram and Box Plot are interactive components and helped to understand the distribution of data and detecting the outliers.
Using the Math Formula the attributes are rescaled with logarithm normalization.
The Missing Value Component is helpful in replacing the NA Values from the dataset with a user defined input.
Normalizer is used to normalize the numeric attributes having different scales.
We have used statistics and box plot components again to view the processed data.
The Final processed data is Partitioned and passed through the classifying learner and predictor algorithms.
The feature selection loop is helpful to split data, and run a subset of the data with DT algorithm in loop. This results in a feature selection table comparing accuracies with no of features.
The Scorer components are helpful to provide all necessary statistical scorings.

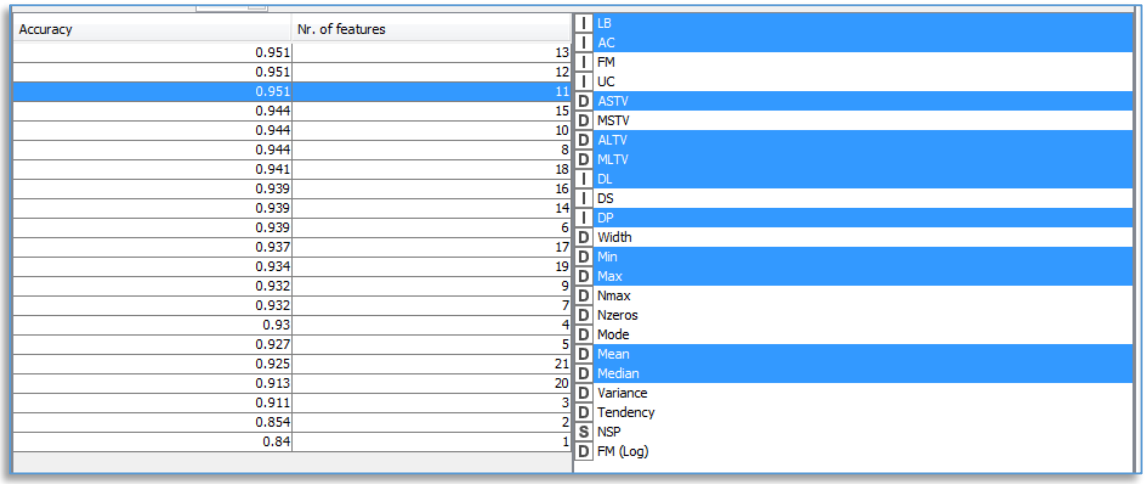
Interesting findings achieved by applying the workflow on the data:

1. Initially the attributes were in different scale and had outliers. For Example, FM & Variance Attributes had outliers.
   ASTV, ALTV were represented in percentage, while ALTV,MLTV were represented in double.
   Post processing with log-norm (FM) and replacing the missing values we rescaled FM.
   Using Normalizer we rescaled ASTV and ALTV attributes and Variance.
   The below Box plot Figs show the data before and after processing.





2. By using the Feature Selection Components with Decision Tree we achieved an excellent result which helps us to select features for the final Decision Tree Algorithm.
   This helps us in reducing Curse of Dimensionality and select minimum and effective features for our DT model.
   In the below figure we can see how feature selection and feature filter components helps out to choose least number of features with highest accuracy for our DT model.
   We have therefore chosen 11 features with Accuracy 0.951 for our DT model.



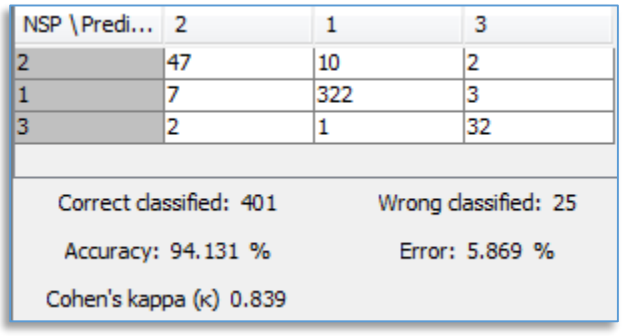
Accuracy	Nr. of features
0.951	13
0.951	12
0.951	11
0.944	15
0.944	10
0.944	8
0.941	18
0.939	16
0.939	14
0.939	6
0.937	17
0.934	19
0.932	9
0.932	7
0.93	4
0.927	5
0.925	21
0.913	20
0.911	3
0.854	2
0.84	1

I	LB
I	AC
I	FM
I	UC
D	ASTV
D	MSTV
D	ALTV
D	MLTV
I	DL
I	DS
I	DP
D	Width
D	Min
D	Max
D	Nmax
D	Nzeros
D	Mode
D	Mean
D	Median
D	Variance
D	Tendency
S	NSP
D	FM (Log)

3. The Scorer helps us identify the Accuracy of each Classification Model and in the below Figs we can compare the scoring.
   The Decision Tree Classifier proves to be the most effective classifier with the help of feature selection used in the workflow.

Decision Tree

The Decision Tree Scores an Accuracy of 94.131% with 25 wrongly classified.



NSP \ Predi...	2	1	3
2	47	10	2
1	7	322	3
3	2	1	32

Correct classified: 401 Wrong classified: 25

Accuracy: 94.131 % Error: 5.869 %

Cohen's kappa (κ) 0.839

SVM
The SVM Scores 90.161%
with 40 wrongly classified.

File Hilite

NSP \ Predi...	2	1	3
2	41	15	1
1	16	313	2
3	4	2	32

Correct classified: 386 Wrong classified: 40

Accuracy: 90.61 % Error: 9.39 %

Cohen's kappa (κ) 0.747

KNN
The KNN scores 87.559%
with 53 wrongly classified

File Hilite

NSP \ Clas...	2	1	3
2	34	22	1
1	20	310	1
3	4	5	29

Correct classified: 373 Wrong classified: 53

Accuracy: 87.559 % Error: 12.441 %

Cohen's kappa (κ) 0.655