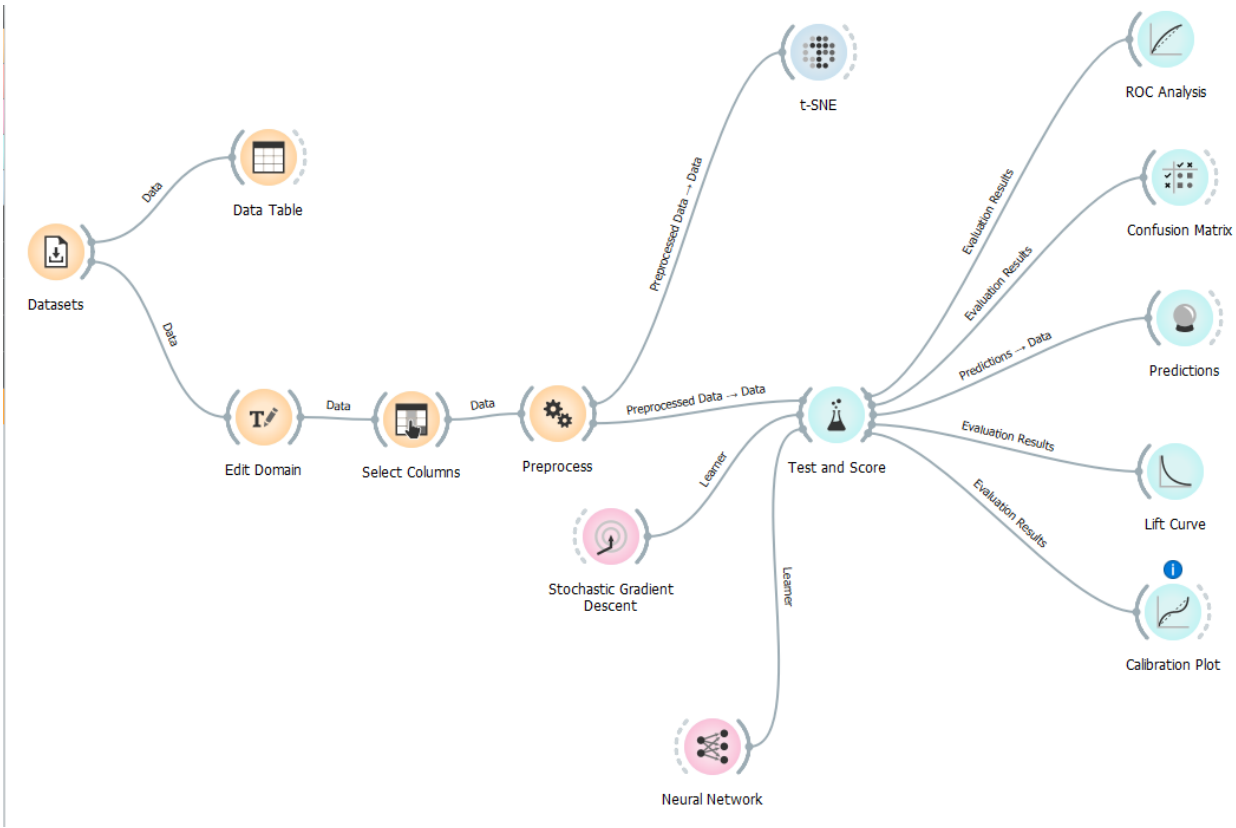
Anastasia Psarou

Lab 2. Loss function, optimization schemes, evaluation

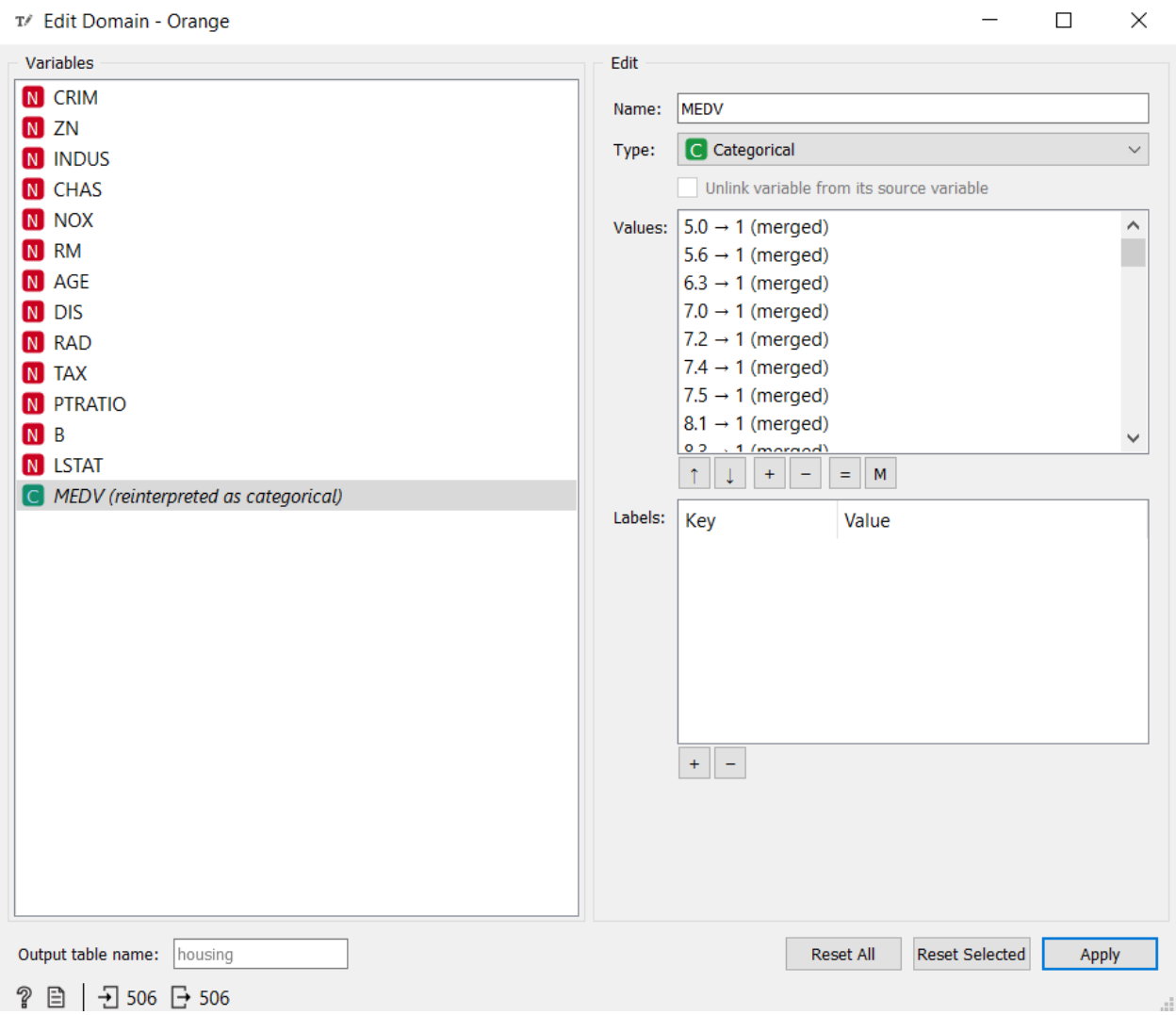
a) Select a dataset with numerical output value (e.g. housing dataset).

b) Read about this dataset: http://lib.stat.cmu.edu/datasets/boston, particularly, learn about the features.

This is my dataflow:



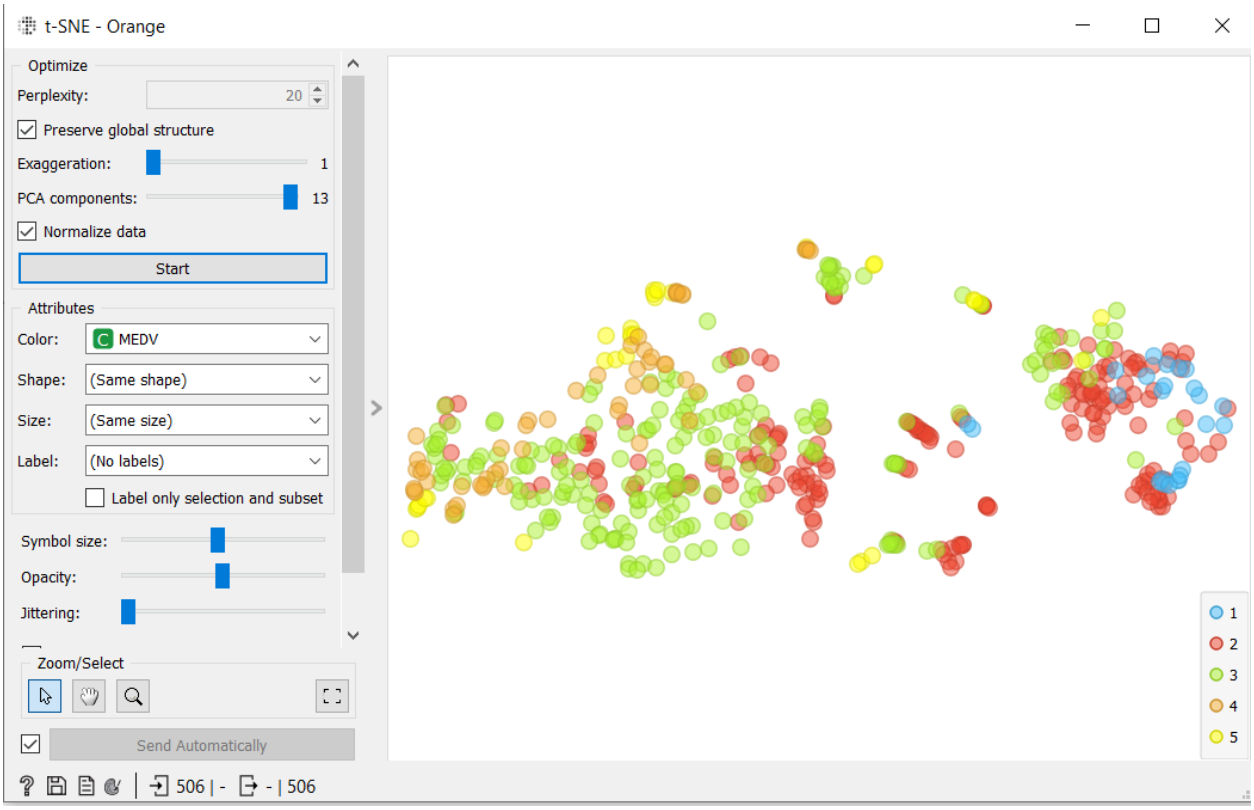
c) Transform the dataset into categorical one e.g. assuming 5 classes. (use widget EditDomain).



In the Edit Domain widget the MEDV is transformed to Categorical and the values are being categorized. In category 1 are the numbers from 0 – 10, in category 2 from 10 – 20, in category 3 from 20 – 30, in category 4 from 30 – 40, in category 5 from 40 – 50.

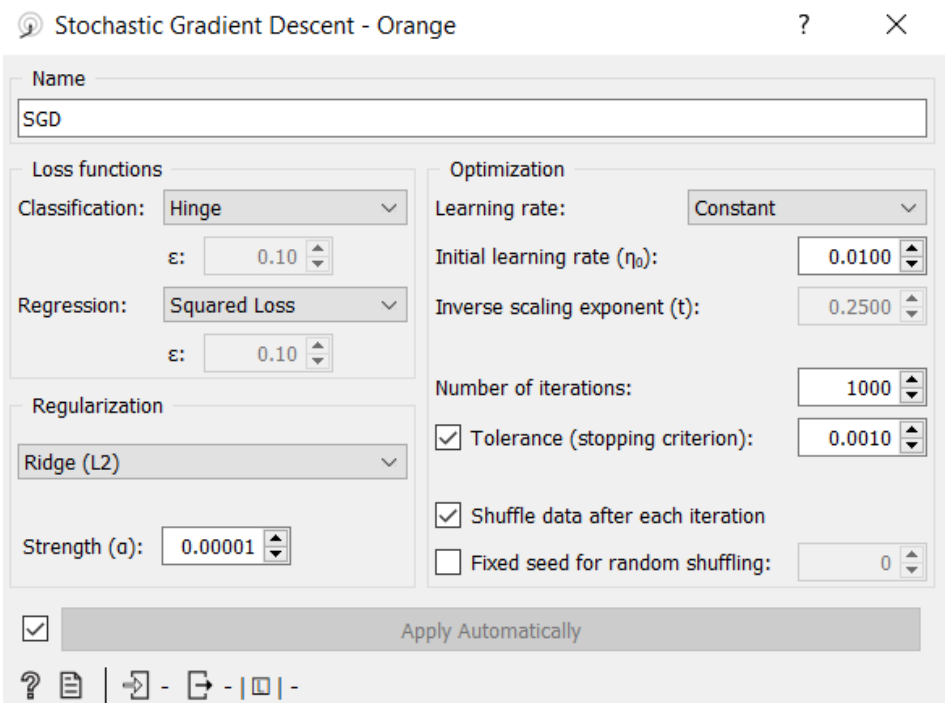
d) Visualize the dataset, and try to divide the dataset on optimal number of classes (i.e. non-overlapping or weak-overlapping classes).

Here we can see the t-SNE graph of this dataflow.



The conclusion is the by overlapping the first 2 categories we have better results.

e) Use SGD classifiers for data classification.



f) Read about the methods you used.

g) Try to combine the best loss function with the best regularizer to obtain the classifier with the highest accuracy.

Some trials were implemented in the Stochastic Gradient Descent in order to determine the best loss function and the best regularizer. In order to find the best ones the Classifier Accuracy is being checked.

Model	Train time [s]	Test time [s]	AUC	CA	F1	Precision	Recall	LogLoss	Specificity
SGD	0.286	0.042	0.897	0.706	0.686	0.697	0.706	0.717	0.828

We conclude that the best loss function is logistic regression and the best regularizer is the elastic net. This result is being produced by the Test and Score widget.

h) How the F1 metrics changes with the number M of cross validation layers (stratified). Compare your result to Leve-one-out scheme (M=number of samples). The best choice of sampling (validation) method is for the lowest F1 value.

At the beginning we observe that for 2 folds the value of F1 is the better than the one for 3 folds. Although, after the 3 folds as we increase the number of folds the value of F1 is increased and has its maximum value for 20 folds.

- 2 folds

Model	Train time [s]	Test time [s]	AUC	CA	F1	Precision	Recall
SGD	0.047	0.010	0.889	0.692	0.665	0.654	0.692

- 3 folds

Model	Train time [s]	Test time [s]	AUC	CA	F1	Precision	Recall
SGD	0.067	0.012	0.891	0.680	0.658	0.656	0.680

- 5 folds

Model	Train time [s]	Test time [s]	AUC	CA	F1	Precision	Recall
SGD	0.113	0.022	0.896	0.690	0.665	0.659	0.690

- 10 folds

Model	Train time [s]	Test time [s]	AUC	CA	F1	Precision	Recall
SGD	0.249	0.041	0.897	0.696	0.676	0.681	0.696

- 20 folds

Model	Train time [s]	Test time [s]	AUC	CA	F1	Precision	Recall
SGD	0.509	0.084	0.898	0.702	0.684	0.692	0.702

The lower value of F1 exists for 3 folds. So, the best choice of sampling is for 3 folds.

i) Compare cross-validation to other sampling methods.

In the Test and Score widget we are going to compare the sampling methods with Cross validation.

- Cross validation

Model	Train time [s]	Test time [s]	AUC	CA	F1	Precision	Recall
SGD	0.244	0.042	0.899	0.704	0.683	0.691	0.704

- Random sampling

Model	Train time [s]	Test time [s]	AUC	CA	F1	Precision	Recall
SGD	0.215	0.044	0.892	0.686	0.664	0.667	0.686

- Leave one out

Model	Train time [s]	Test time [s]	AUC	CA	F1	Precision	Recall
SGD	13.517	2.733	0.898	0.700	0.681	0.687	0.700

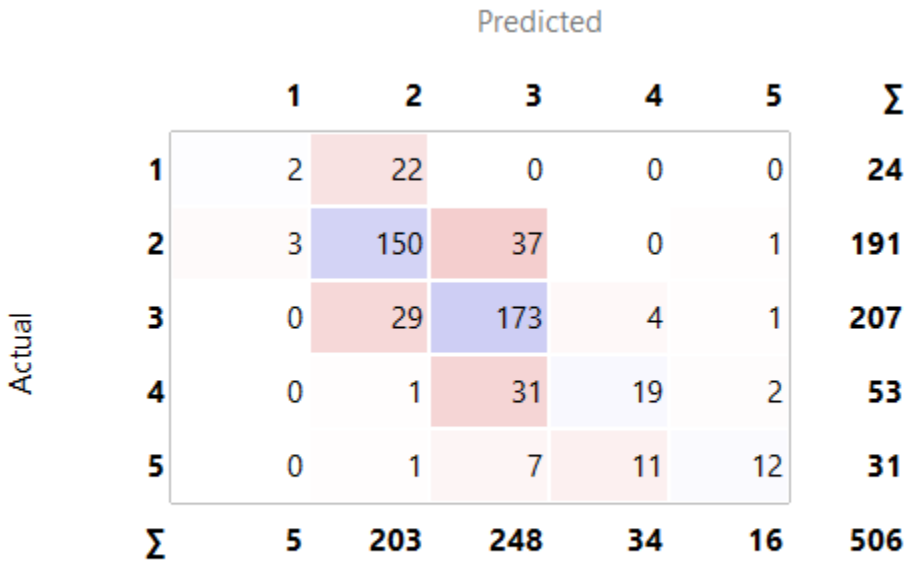
- Test on train data

Model	Train time [s]	Test time [s]	AUC	CA	F1	Precision	Recall
SGD	0.032	0.005	0.918	0.723	0.704	0.721	0.723

As far as F1 is concerned only Test on train data has higher this value in comparison to the other sampling methods that have lower than Cross validation sampling.

As far as CA is concerned both Test on train data and Leave on out have higher this value in comparison to Random sampling that has lower than Cross validation sampling.

j) Based on confusion matrix decide, which classes are the closest ones.

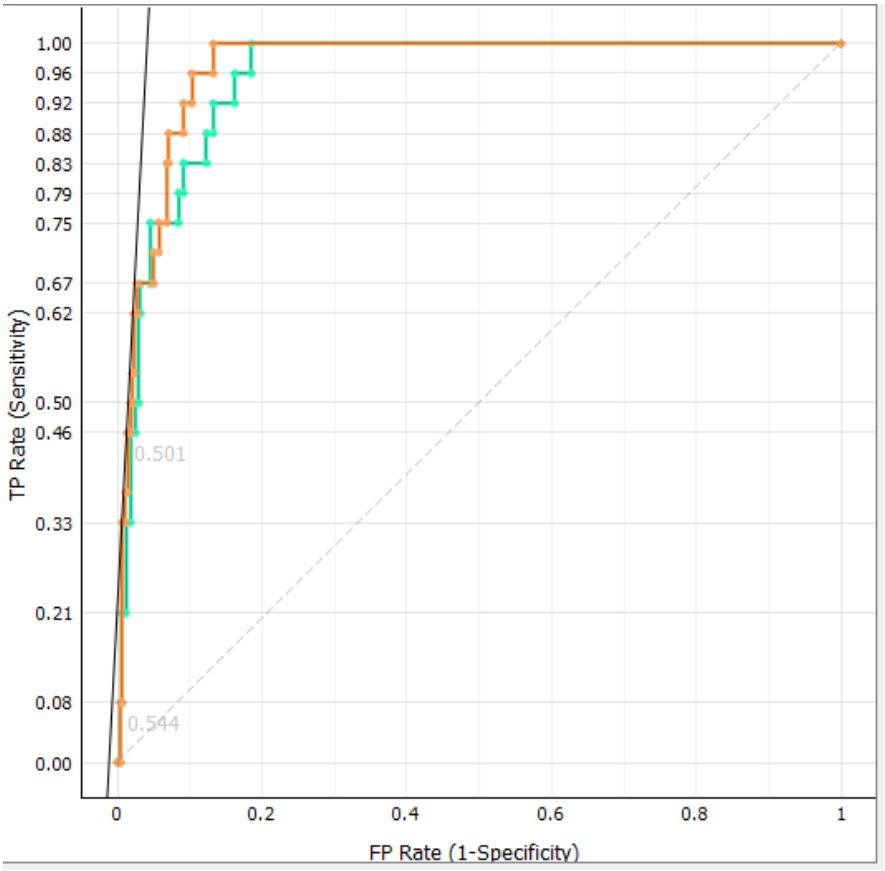


Based on the confusion matrix we conclude that classes 2 and 3 are the closest ones.

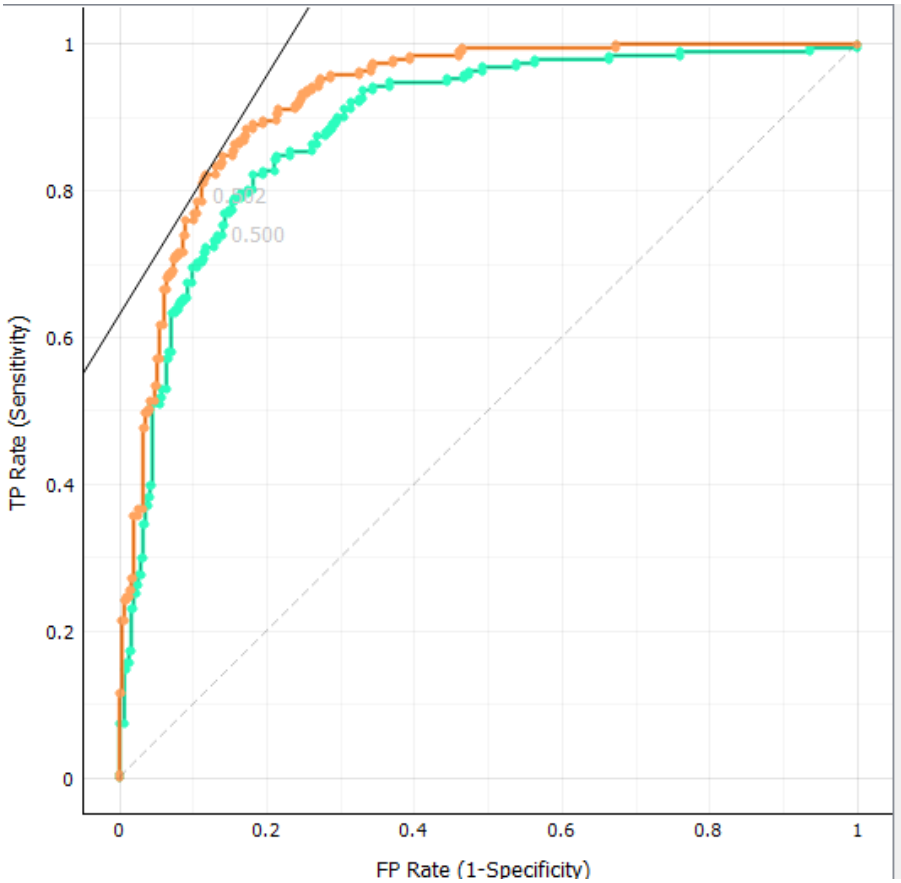
k) Based on ROC curve, decide from which classes the samples are the most difficult (and the easiest) to recognize.

ROC analysis for the different classes:

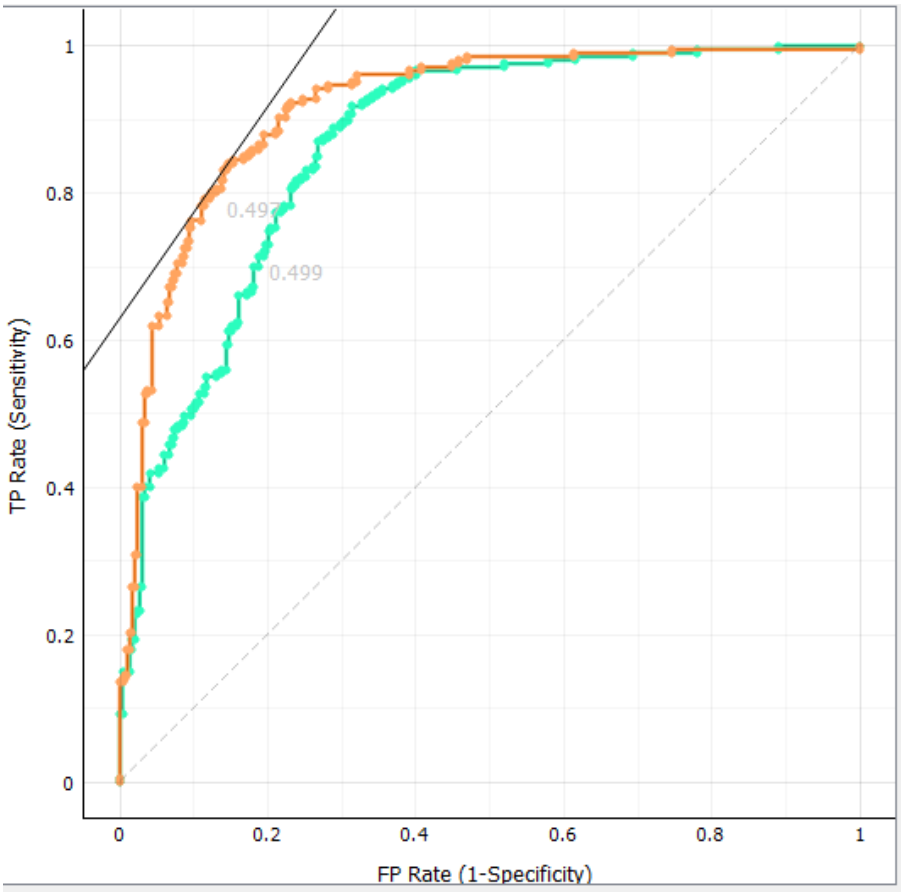
- Class 1



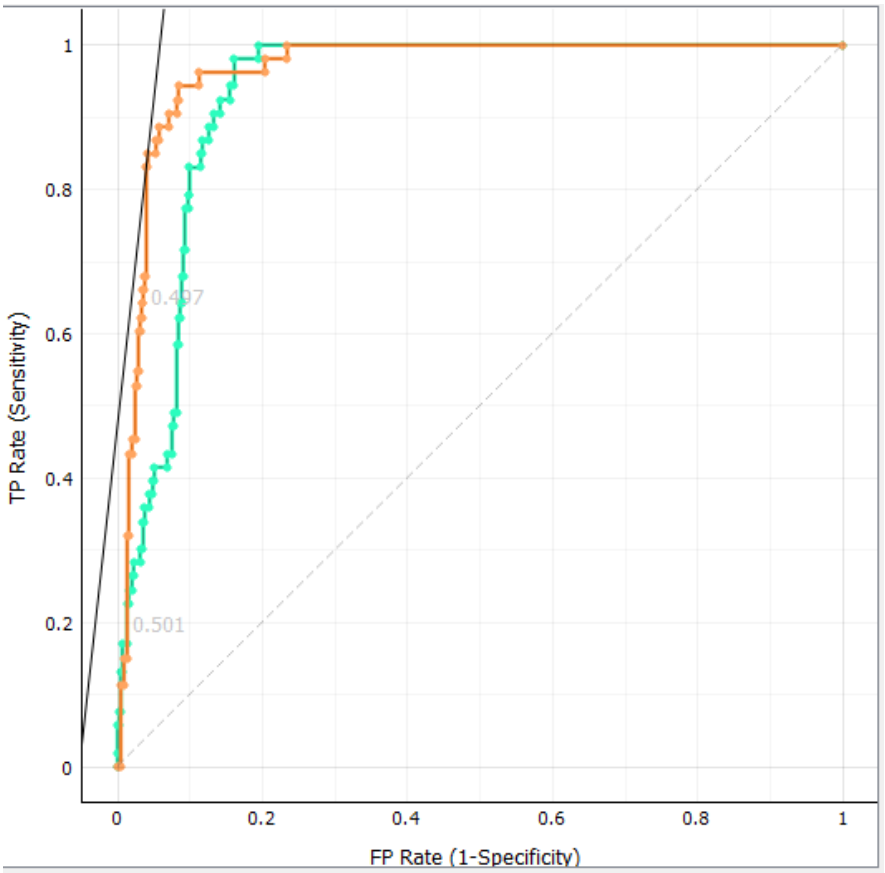
- Class 2


1
0.8
0.6
0.4
0.2
0
TP Rate (Sensitivity)
0.500
0
0.2
0.4
0.6
0.8
1
FP Rate (1-Specificity)

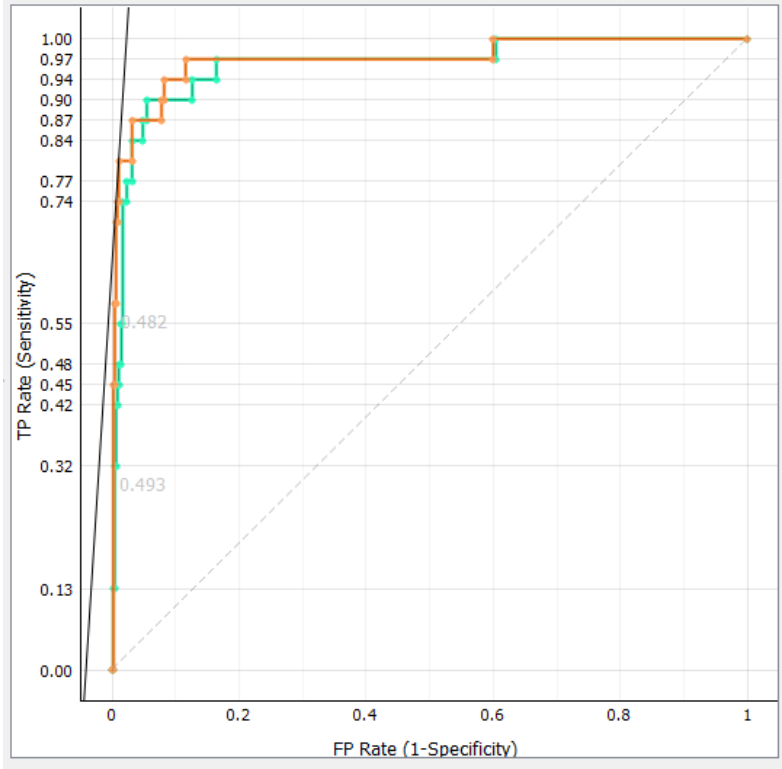
- Class 3


1
0.8
0.6
0.4
0.2
0
TP Rate (Sensitivity)
0.497
0.499
0
0.2
0.4
0.6
0.8
1
FP Rate (1-Specificity)

- Class 4


1
0.8
0.6
0.4
0.2
0
TP Rate (Sensitivity)
0.497
0.501
0
0.2
0.4
0.6
0.8
1
FP Rate (1-Specificity)

- Class 5



From the above diagrams we can see that the samples of Class 3 are the easiest to recognize and the sample of Class 5 are the difficult to recognize.

l) Find the features which decide about samples membership to the most contrasting class.

We can see that the best features are LSTAT RM and INDUS.