

DATA MINING PROJECT:

BATCH NO: B8

ROLLNO: 21481A0575

21481A05B4

22485A0511

21481A0587

DATASET DESCRIPTION: LYMPHOGRAPHY DATASET

The Lymphography dataset is a classification dataset that has been widely studied in machine learning literature.

Objective:

- The primary objective of this dataset is **classification**.
- Specifically, we aim to predict lymphography outcomes (class labels) based on the provided features.

Target Variable (Class Labels):

- The target variable is denoted as **class**.
- It represents different lymphography outcomes.
- The possible class labels are:
 - normal** : Indicates a normal lymphography result.
 - metastases**: Suggests the presence of metastases.
 - malign lymph**: Indicates malignant lymph nodes.
 - fibrosis**: Represents fibrosis.

The **Lymphography** dataset comprises instances characterized by various features, including lymphatic state, blockage indicators, and lymph node properties. Its primary objective is classification, aiming to predict lymphography outcomes based on these features.

Data Quality:

- Ensuring data quality is crucial for building accurate models.
- We should check for missing values, outliers, and inconsistencies.

Dataset:

Data Table (2) - Orange

Info
148 instances (no missing data)
18 features
Target with 4 values
No meta attributes.

Variables
☒ Show variable labels (if present)
☐ Visualize numeric values
☒ Color by instance classes

Select
☒ Select full rows

	y	lymphatics	bl_affe	bl_lymph_c	bl_lymph_s	by_pass	extravasates	regen	early_uptake	lym_dmin	lym_enlar	changes_lym	defect	dx
113	malign lymph	arched	no	no	no	no	no	no	no	1	2	oval	lacunar	lacur
114	metastases	arched	no	no	no	no	yes	no	yes	1	2	oval	lacunar	lacur
115	malign lymph	deformed	no	no	no	no	no	no	yes	1	2	oval	lac central	lac c
116	malign lymph	deformed	no	no	no	no	yes	no	yes	1	2	oval	lac central	lac n
117	metastases	displaced	no	no	no	no	no	no	yes	1	3	oval	lacunar	lac c
118	metastases	displaced	yes	yes	no	no	yes	no	yes	1	2	oval	lac marginal	lac n
119	malign lymph	arched	no	no	no	no	yes	no	yes	1	3	oval	lacunar	lacur
120	metastases	displaced	yes	no	no	yes	yes	no	yes	1	3	round	lac central	lacur
121	metastases	arched	no	no	no	no	no	no	no	1	1	oval	lac central	lac n
122	metastases	arched	no	no	no	no	no	no	no	1	2	oval	lacunar	lac n
123	malign lymph	arched	yes	yes	yes	yes	yes	no	no	1	2	oval	lac central	lac n
124	metastases	deformed	no	no	no	no	no	no	no	1	2	round	lac marginal	lac n
125	malign lymph	displaced	yes	no	no	no	yes	no	no	1	2	oval	lac marginal	lac n
126	malign lymph	deformed	no	no	no	no	yes	no	yes	1	4	oval	lacunar	lacur
127	metastases	arched	yes	no	no	no	no	no	yes	1	3	round	lac marginal	lac n
128	metastases	arched	yes	no	no	no	yes	no	yes	1	2	oval	lac marginal	lac n
129	malign lymph	arched	no	no	no	no	yes	no	yes	1	2	oval	lac central	lac c
130	fibrosis	deformed	yes	no	no	no	yes	no	yes	1	3	oval	lac central	lac n
131	metastases	arched	yes	no	no	no	no	no	yes	1	4	round	lac central	lacur
132	metastases	displaced	yes	yes	yes	yes	yes	no	yes	1	4	round	lac central	lac c
133	metastases	displaced	no	no	no	no	no	no	yes	1	4	oval	lac central	lacur
134	metastases	deformed	yes	no	no	no	no	no	yes	1	2	round	lac marginal	lac n
135	malign lymph	arched	yes	no	no	no	yes	no	yes	1	2	oval	lac central	lac c
136	malign lymph	deformed	yes	yes	yes	yes	yes	no	yes	1	2	round	lac marginal	lac n
137	malign lymph	normal	no	no	no	no	yes	no	yes	1	2	oval	no	no
138	malign lymph	arched	no	no	no	no	yes	no	yes	1	3	round	lacunar	lacur
139	malign lymph	arched	yes	no	no	no	no	no	no	1	2	round	lac marginal	lac n
140	malign lymph	deformed	no	no	no	yes	yes	yes	no	3	1	bean	lacunar	no
141	metastases	arched	yes	no	no	no	no	no	no	1	2	round	lac marginal	lac n
142	malign lymph	arched	no	no	no	no	no	no	yes	1	2	oval	lac central	lac c
143	metastases	arched	yes	no	no	no	yes	no	yes	1	2	round	lac marginal	lac n
144	malign lymph	deformed	yes	no	no	yes	yes	no	yes	1	2	oval	lac central	lac n
145	metastases	arched	no	no	no	no	no	no	no	1	1	bean	no	no
146	metastases	arched	yes	no	no	no	yes	no	yes	1	3	round	lac marginal	lac n
147	malign lymph	arched	no	no	no	no	no	no	yes	1	2	oval	lac central	lacur
148	malign lymph	arched	yes	yes	no	yes	yes	no	yes	1	3	round	lac central	lac n

Restore Original Order
☒ Send Automatically

- **PREPROCESSING:**

Replacing the missing values with most frequent /Average values

Normalize the features

Preprocess - Orange

Preprocessors

- Discretize Continuous Variables
- Continuize Discrete Variables
- Impute Missing Values
- Select Relevant Features
- Select Random Features
- Normalize Features
- Randomize
- Remove Sparse Features
- Principal Component Analysis
- CUR Matrix Decomposition

Impute Missing Values

- ☒ Average/Most frequent
- ☐ Replace with random value
- ☐ Remove rows with missing values.

Normalize Features

- ☐ Standardize to $\mu=0, \sigma^2=1$
- ☐ Center to $\mu=0$
- ☐ Scale to $\sigma^2=1$
- ☐ Normalize to interval $[-1, 1]$
- ☒ Normalize to interval $[0, 1]$

CLASSIFICATION:

1. Decision Tree
2. Logistic Regression
3. Naive Bayes
4. SVM
5. Neural Networks

TEST AND SCORE:

Without preprocessing:

Test and Score (1) - Orange

Cross validation

Number of folds: 2

☒ Stratified

☐ Cross validation by feature

☒ Selected

☐ Random sampling

Repeat train/test: 10

Training set size: 75 %

☒ Stratified

☐ Leave one out

☐ Test on train data

☐ Test on test data

Evaluation results for target (None, show average over classes)

Model	AUC	CA	F1	Prec	Recall	MCC
Tree (1)	0.760	0.736	0.730	0.725	0.736	0.492
Naive Bayes	0.836	0.243	0.354	0.927	0.243	0.258
Logistic Regression	0.943	0.858	0.850	0.849	0.858	0.726
SVM	0.922	0.804	0.785	0.774	0.804	0.618

Compare models by: Area under ROC curve

☐ Negligible diff.: 0.1

	Tree (1)	Naive Bayes	Logistic Regressi...	SVM
Tree (1)		0.246	0.066	0.088
Naive Bayes	0.754		0.114	0.106
Logistic Regression	0.934	0.886		0.843
SVM	0.912	0.894	0.157	

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

With preprocessing:

Test and Score (2) - Orange

Cross validation

Number of folds: 2

☒ Stratified

☐ Cross validation by feature

☐ Random sampling

Repeat train/test: 10

Training set size: 75 %

☒ Stratified

☐ Leave one out

☐ Test on train data

☐ Test on test data

Evaluation results for target (None, show average over classes)

Model	AUC	CA	F1	Prec	Recall	MCC
SVM (1)	0.873	0.804	0.785	0.774	0.804	0.618
Naive Bayes (1)	0.836	0.243	0.354	0.927	0.243	0.258
Tree (2)	0.760	0.736	0.730	0.725	0.736	0.492
Logistic Regression (1)	0.934	0.824	0.807	0.791	0.824	0.657

Compare models by: Area under ROC curve

☐ Negligible diff.: 0.1

	SVM (1)	Naive Bayes (1)	Tree (2)	Logistic Regres...
SVM (1)		0.922	0.803	0.243
Naive Bayes (1)	0.078		0.754	0.112
Tree (2)	0.197	0.246		0.078
Logistic Regression (1)	0.757	0.888	0.922	

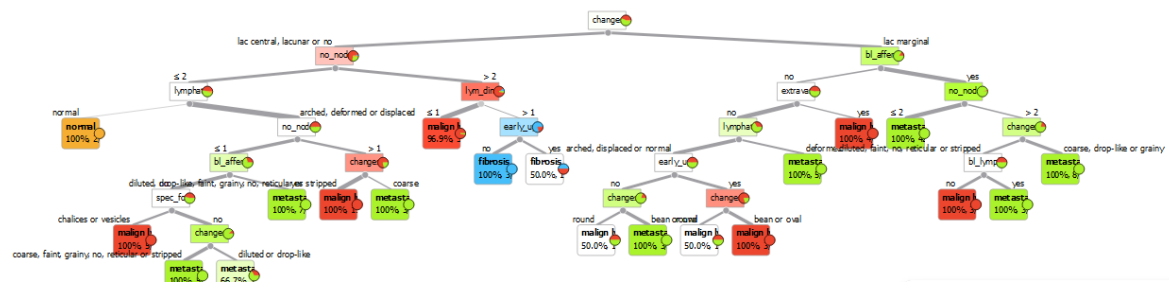
Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

Gain Ratio:

		#	Info. gain	Gain ratio	Gini
1	C changes_node	4	0.402	0.246	0.186
2	N no_nodes		0.264	0.137	0.129
3	N lym_enlar		0.208	0.121	0.087
4	C spec_forms	3	0.184	0.125	0.088
5	C changes_stru	8	0.179	0.071	0.063
6	C bl_affere	2	0.174	0.175	0.101
7	N lym_dimin		0.161	0.565	0.033
8	C lymphatics	4	0.156	0.097	0.026
9	C defect	4	0.148	0.087	0.042
10	C changes_lym	3	0.146	0.122	0.031
11	C regen	2	0.136	0.380	0.025
12	C early_uptake	2	0.134	0.153	0.065
13	C by_pass	2	0.073	0.092	0.009
14	C exclusion	2	0.066	0.089	0.020
15	C dislocation	2	0.064	0.069	0.025
16	C bl_lymph_s	2	0.040	0.145	0.007
17	C bl_lymph_c	2	0.034	0.051	0.012
18	C extravasates	2	0.029	0.029	0.003

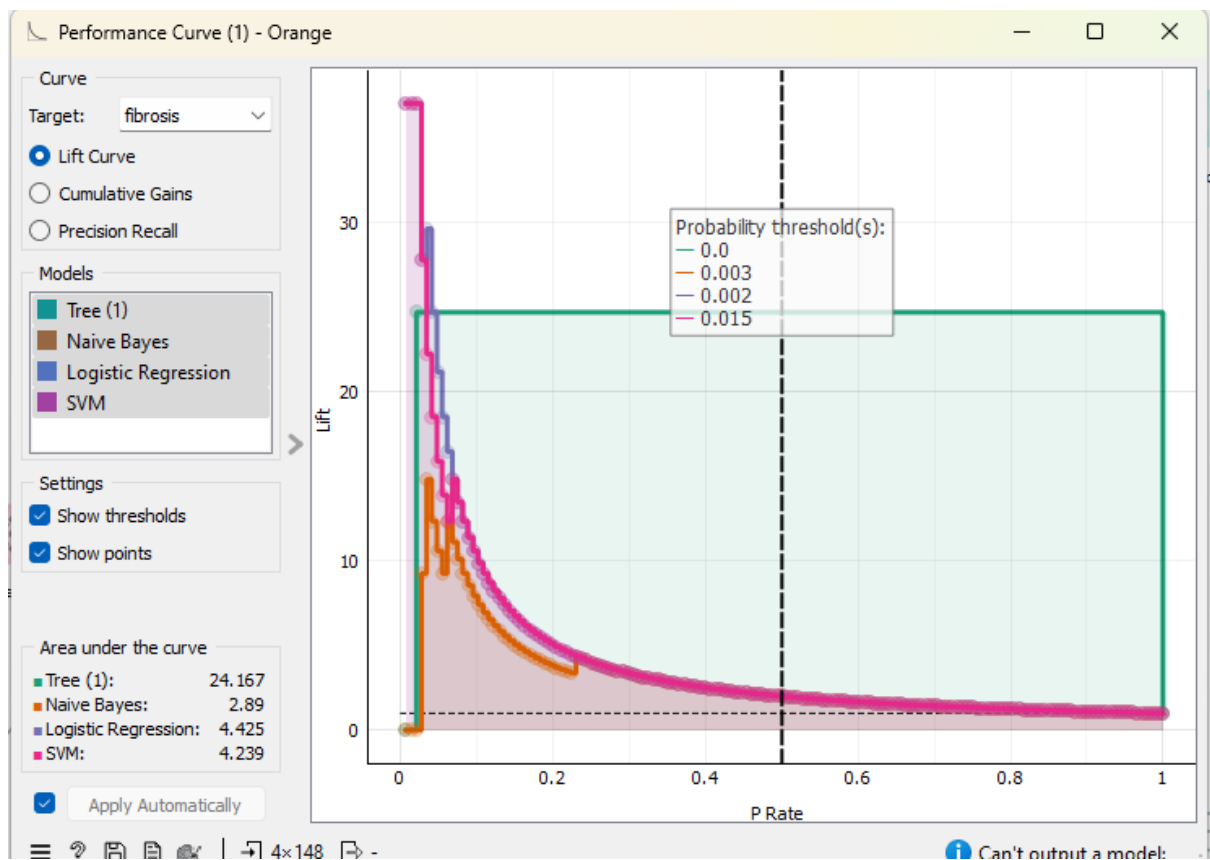
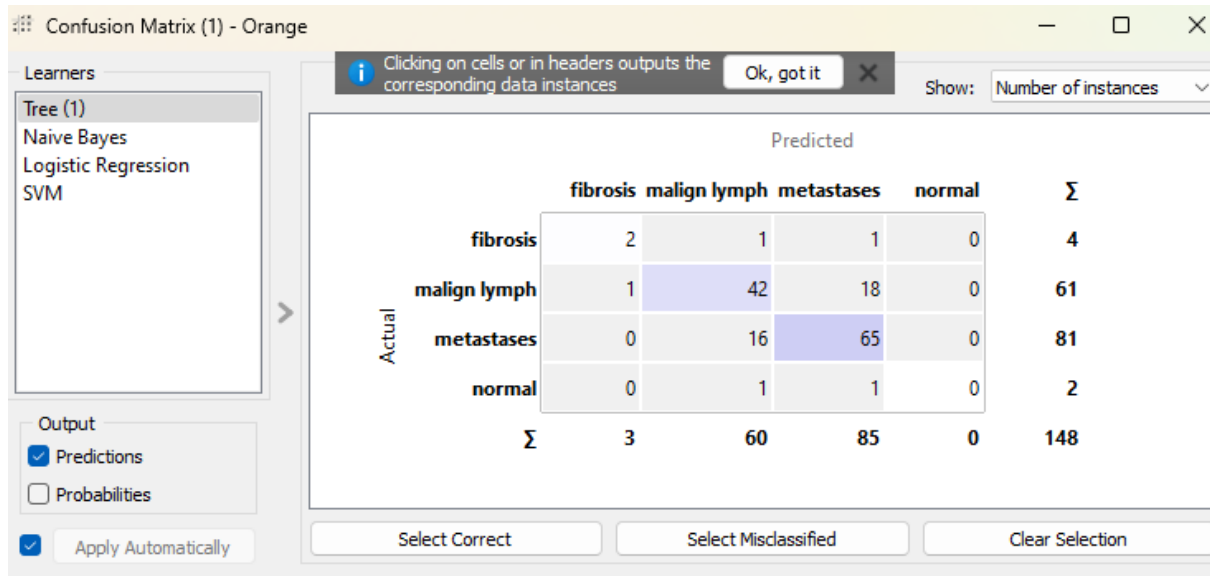
Hence `change_node` is the root node

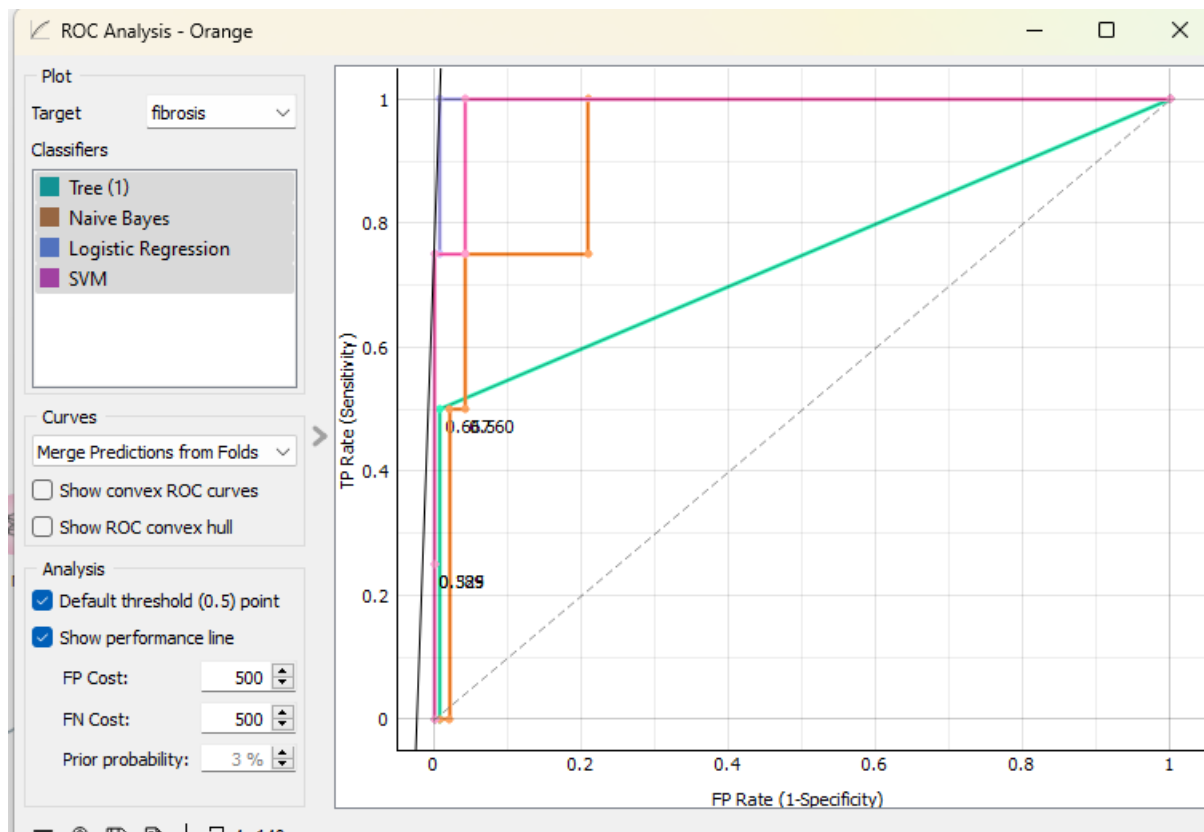
Decision Tree:



Confusion Matrix, Performance curve and Roc Analysis:

Before preprocessing:





After Preprocessing:

Confusion Matrix (2) - Orange

Clicking on cells or in headers outputs the corresponding data instances

Ok, got it

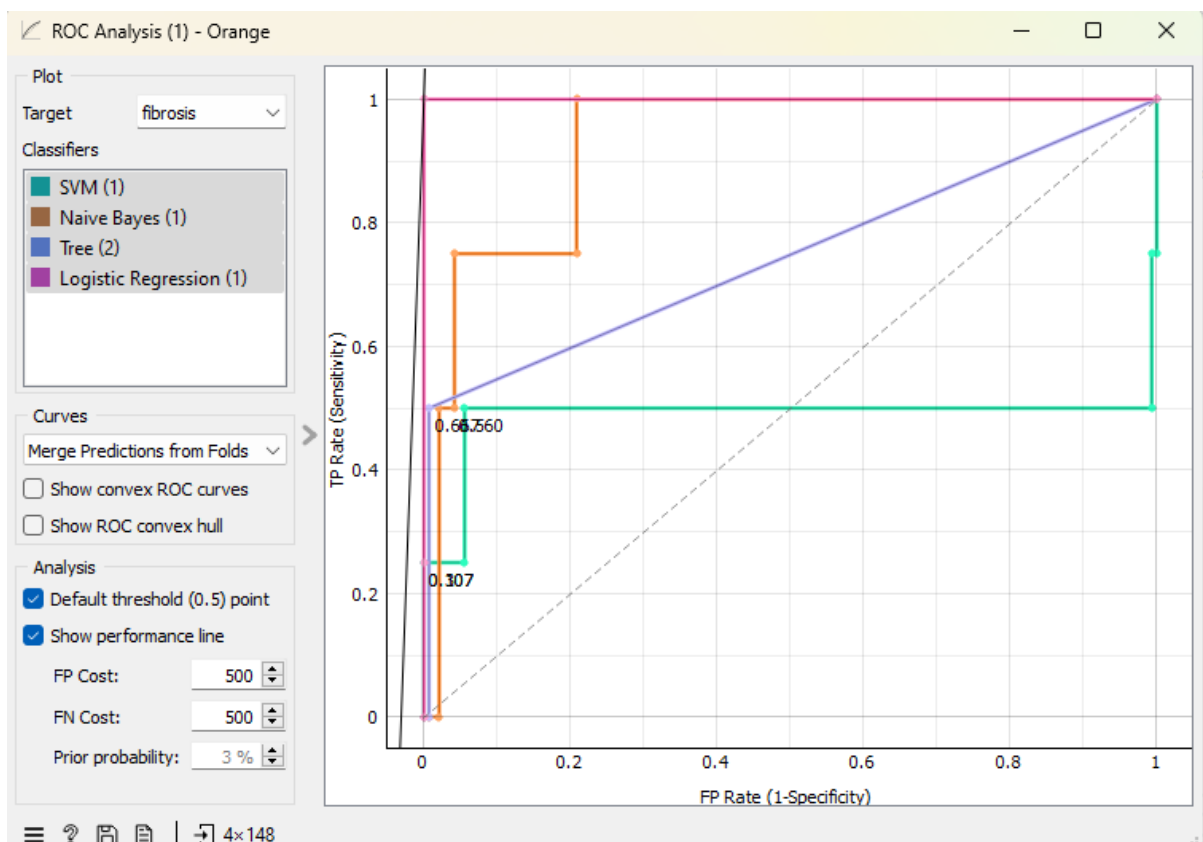
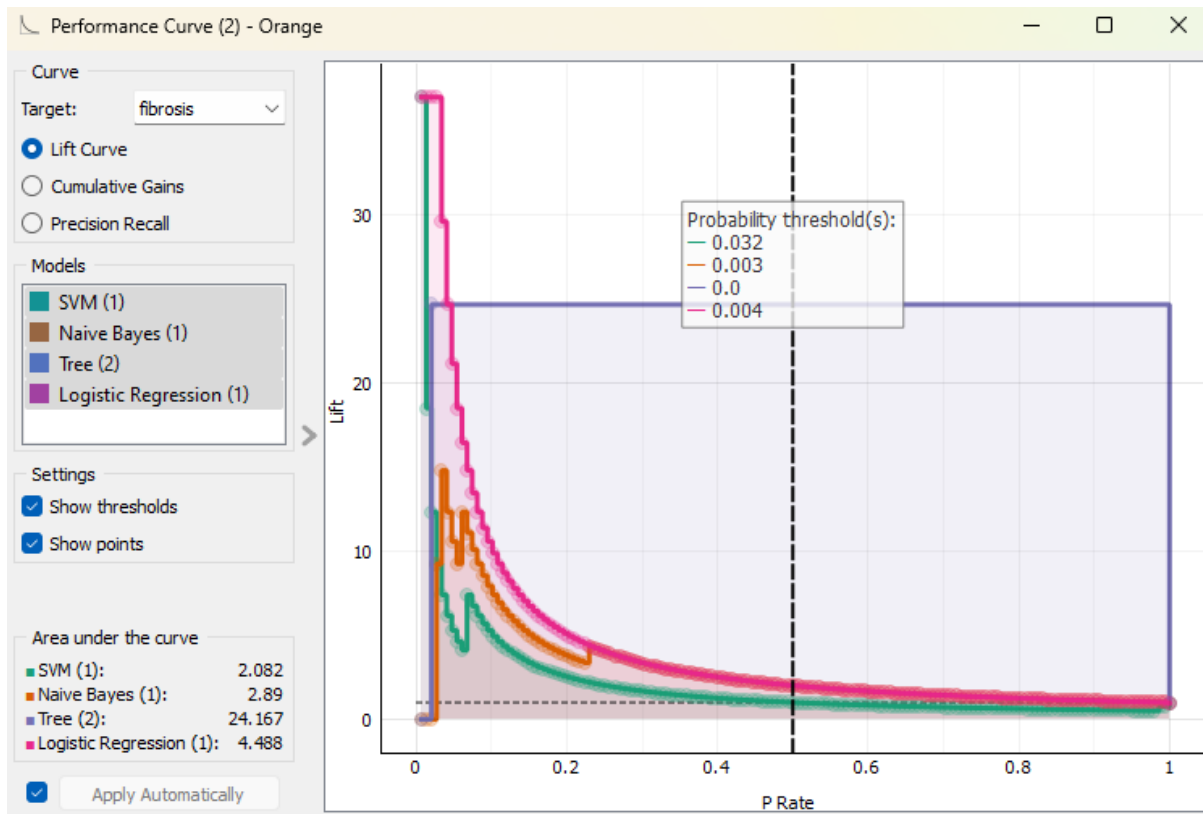
Show: Number of instances

		Predicted				Σ
		fibrosis	malign lymph	metastases	normal	
Actual	fibrosis	0	2	2	0	4
	malign lymph	0	44	17	0	61
	metastases	0	6	75	0	81
	normal	0	1	1	0	2
Σ		0	53	95	0	148

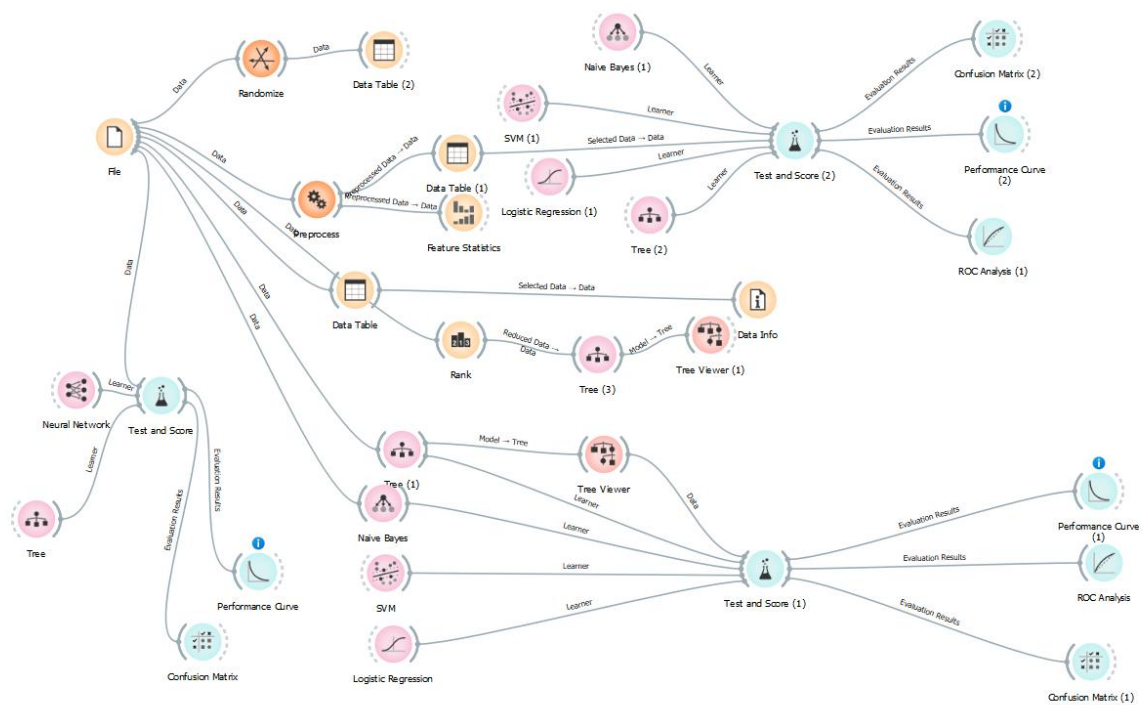
Select Correct

Select Misclassified

Clear Selection



DATA WORKFLOW:



OBSERVATIONS:

- After preprocessing, it's great to see that the accuracy has increased.
- The accuracy of the model has significantly improved after preprocessing steps.
- Removing noise and irrelevant features may have contributed to the accuracy boost.
- After applying preprocessing techniques were effective in enhancing model performance.