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.

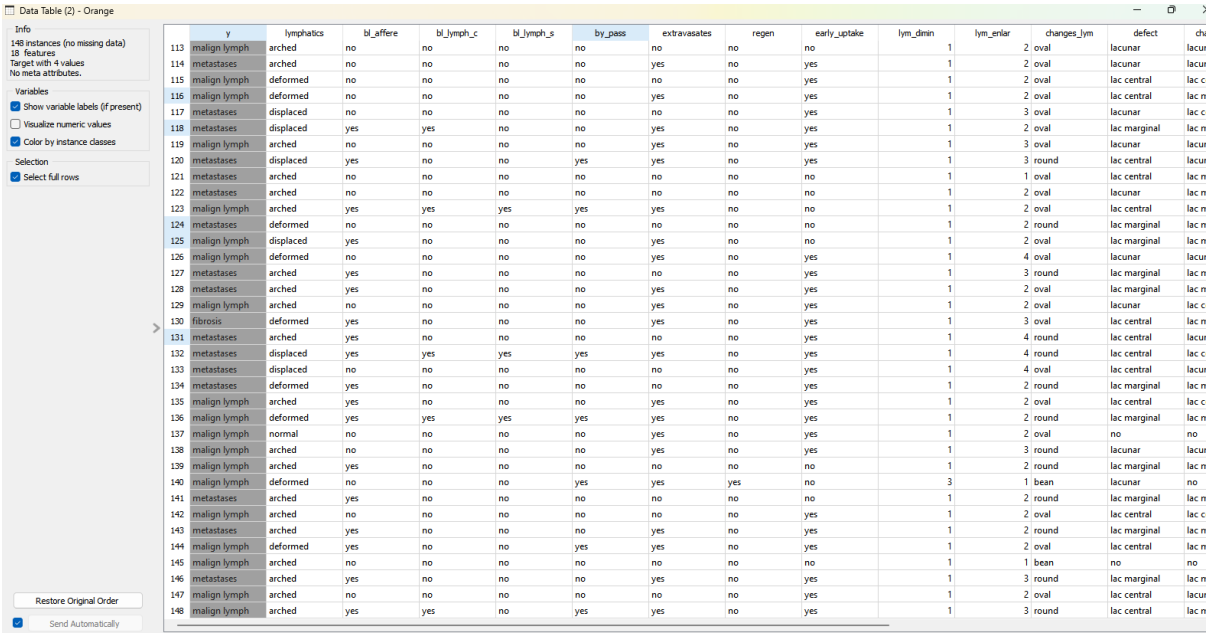
The **Lymphography** dataset comprises instances characterized by various features, including lymphatic state, blockage indicators, and lymph node properties. Its primary objective is classification, aiming to predict lymphography outcomes based on these features.

Data Quality:

- Ensuring data quality is crucial for building accurate models.

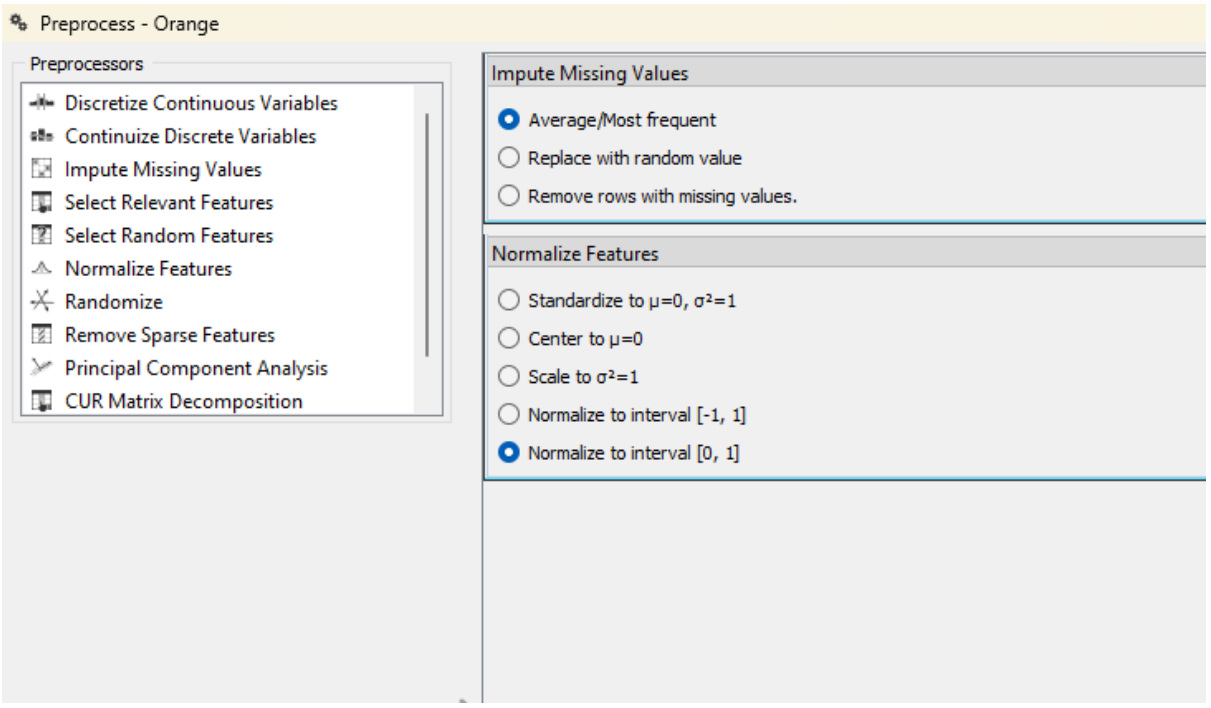
- We should check for missing values, outliers, and inconsistencies.

Dataset:



	y	lymphatics	bl_affere	bl_lymph_c	bl_lymph_s	by_pass	extravasates	regen	early_uptake	lym_dimin	lym_enlar	changes_lym	defect	ch
113	malign lymph	arched	no	no	no	no	no	no	no	1	2	oval	lacunar	lacu
114	metastases	arched	no	no	no	no	yes	no	yes	1	2	oval	lacunar	lacu
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	lacu
120	metastases	displaced	yes	no	no	yes	yes	no	yes	1	3	round	lac central	lacu
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	lacu
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	lacunar	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	lacu
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	lacu
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	lacu
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	malign lymph	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	lacu
148	malign lymph	arched	yes	yes	no	yes	yes	no	yes	1	3	round	lac central	lac n

- **<u>PREPROCESSING:</u>**
  Replacing the missing values with most frequent /Average values
  Normalize the features

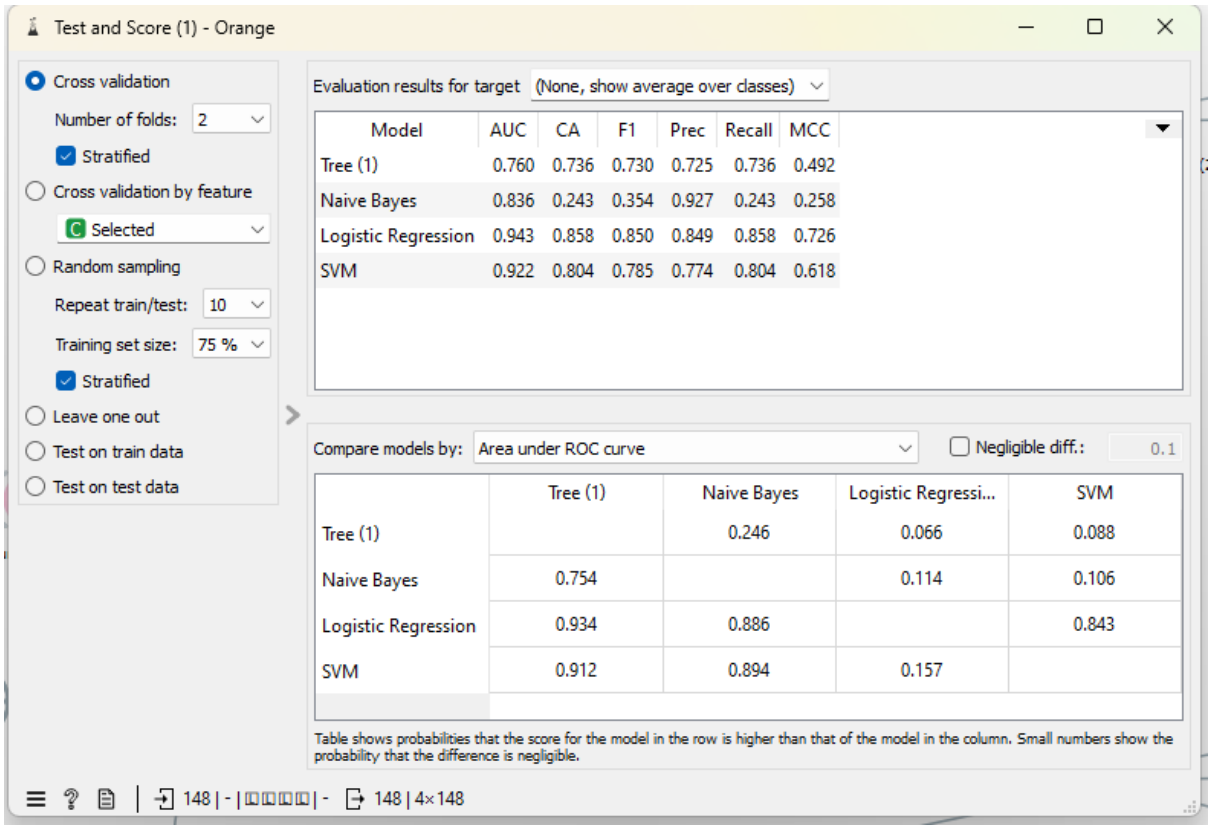


<u>CLASSIFICATION:</u>

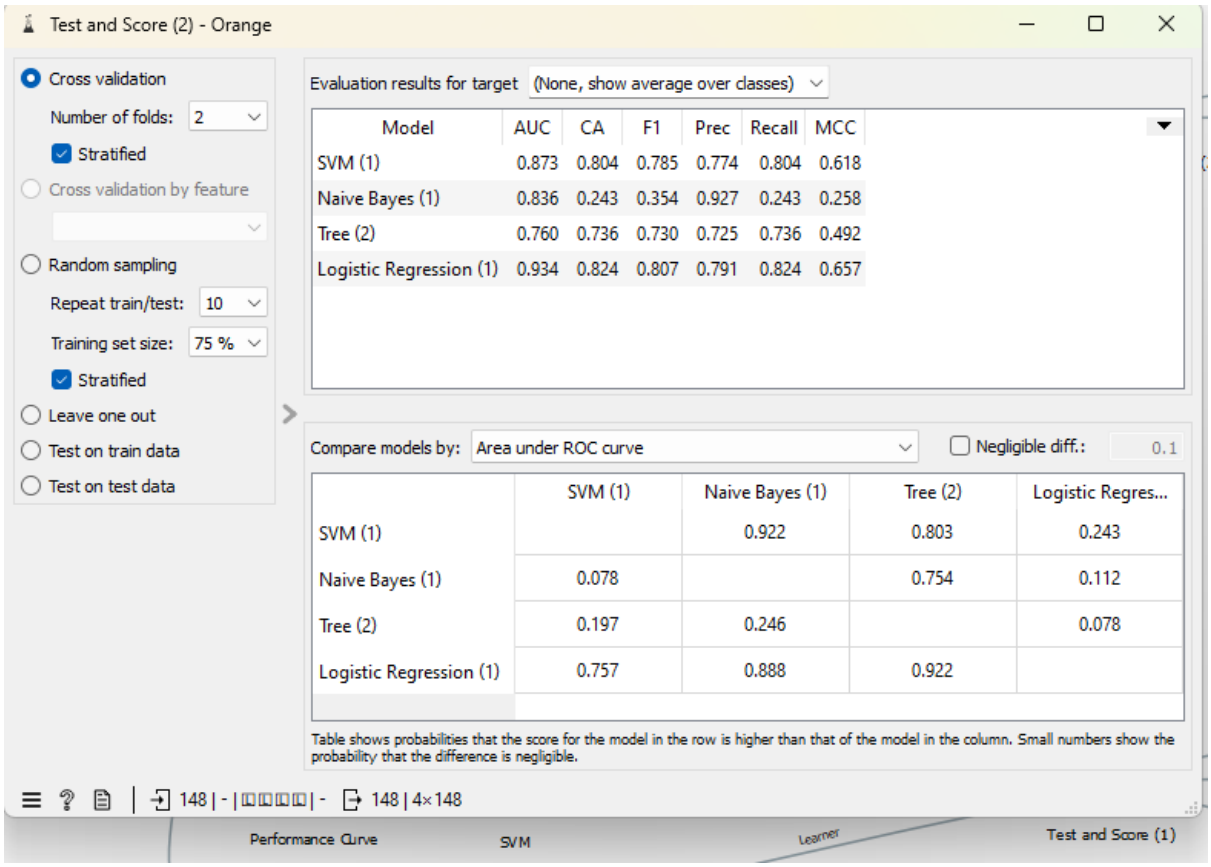
1.Decision Tree
2.Logistic Regression
3.Naive Bayes
4.SVM
5.Neural Networks

TEST AND SCORE:

Without preprocessing:



With preprocessing:

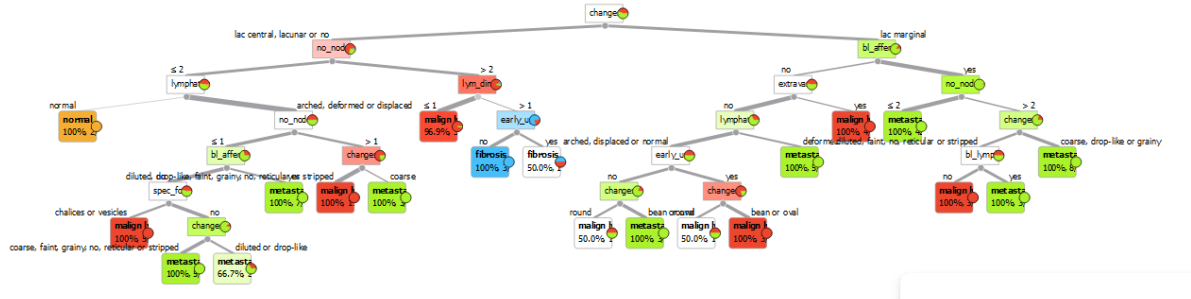


Gain Ratio:

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

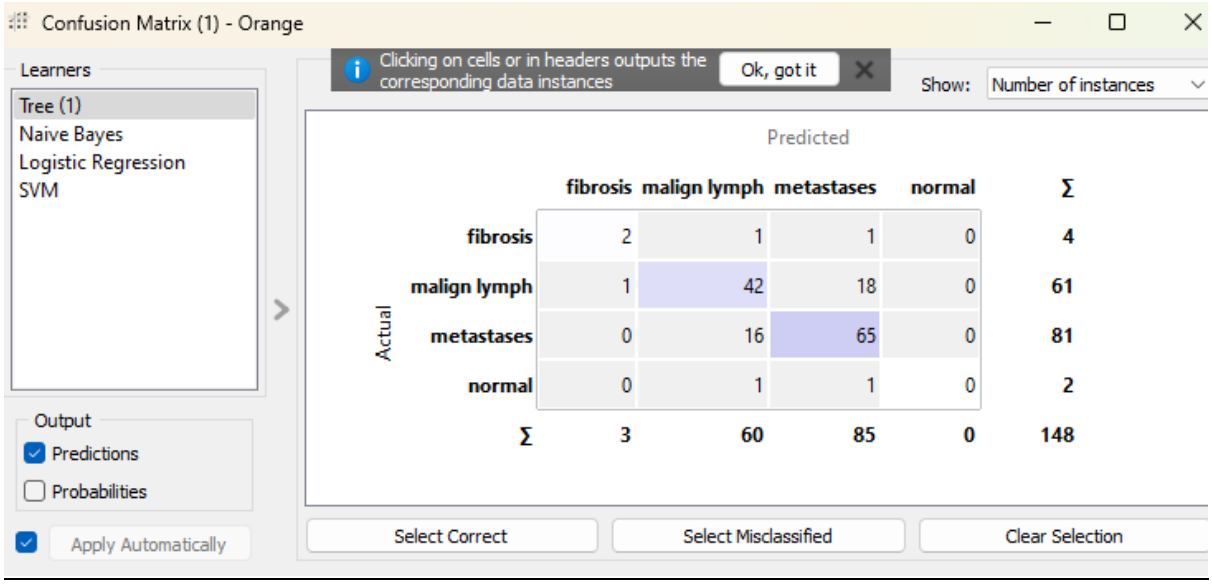
Hence change_node is the root node

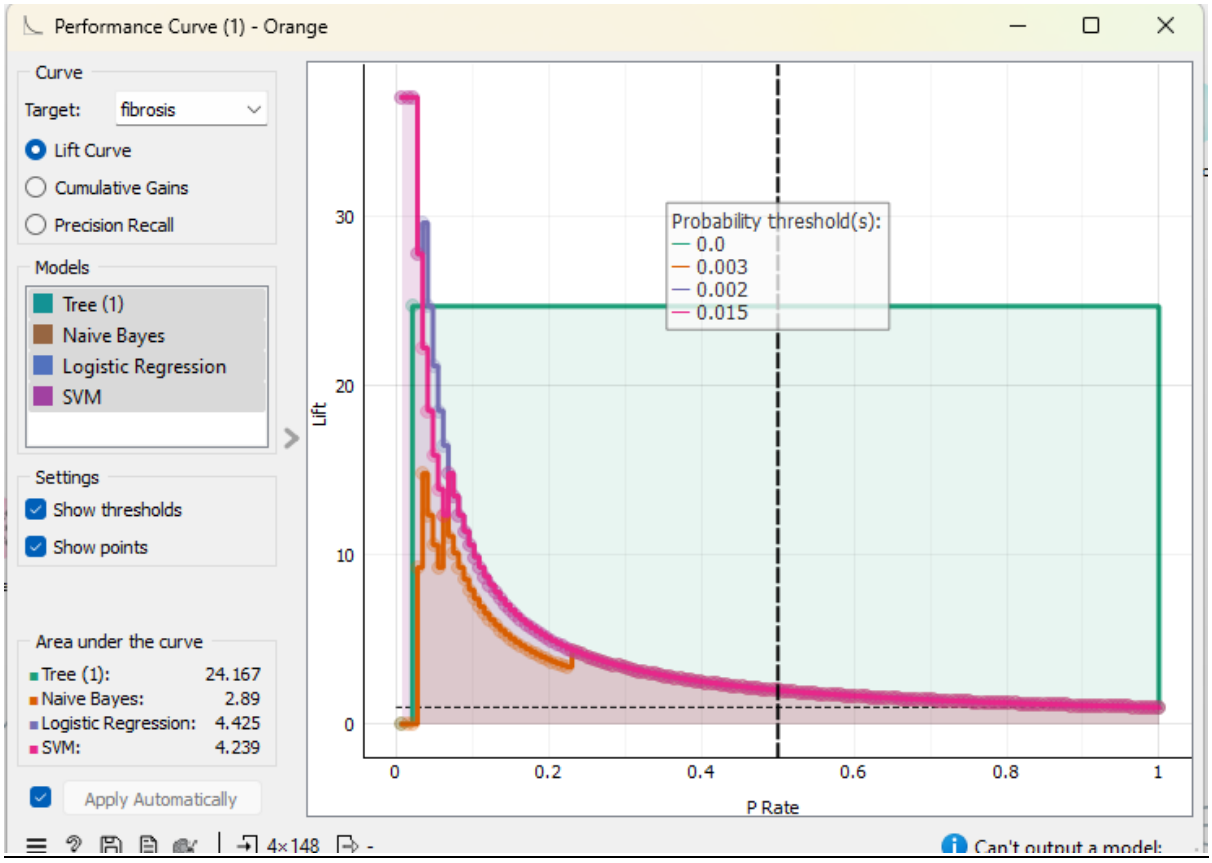
Decision Tree:

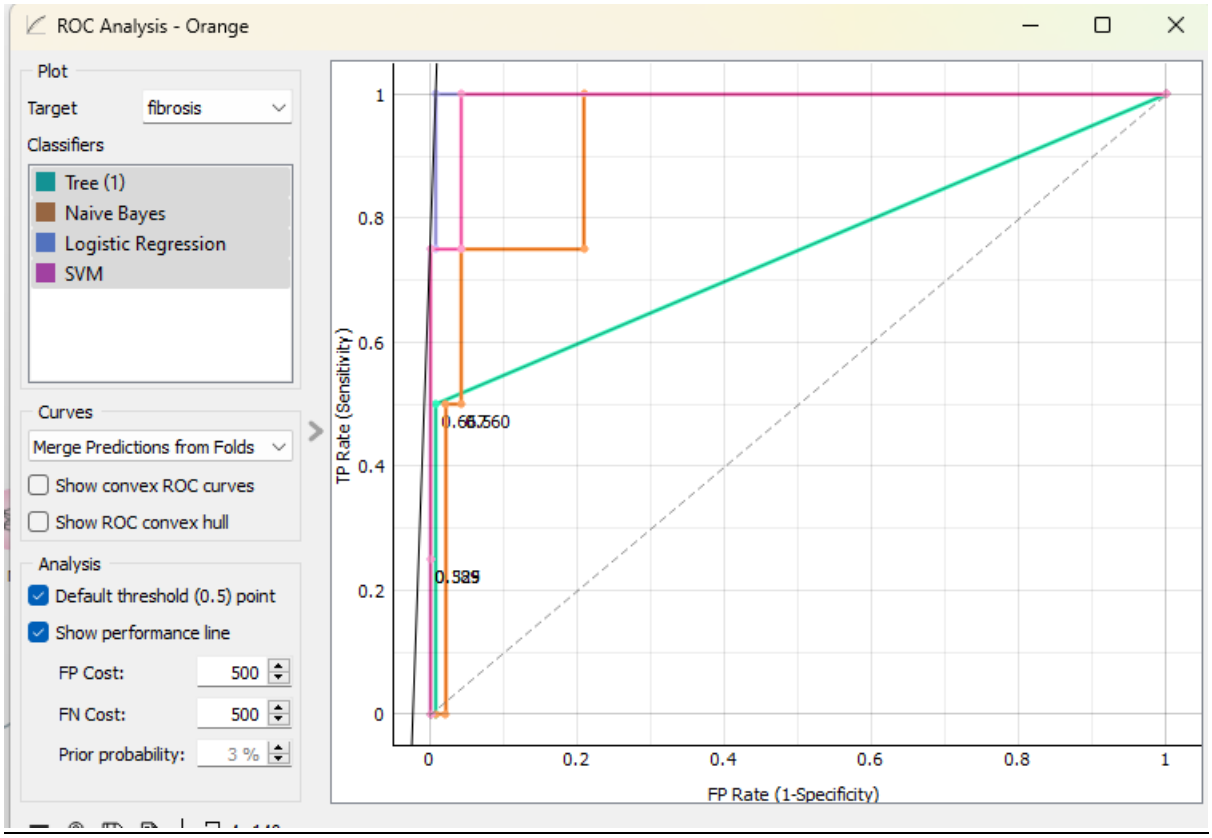


Confusion Matrix, Performance curve and Roc Analysis:

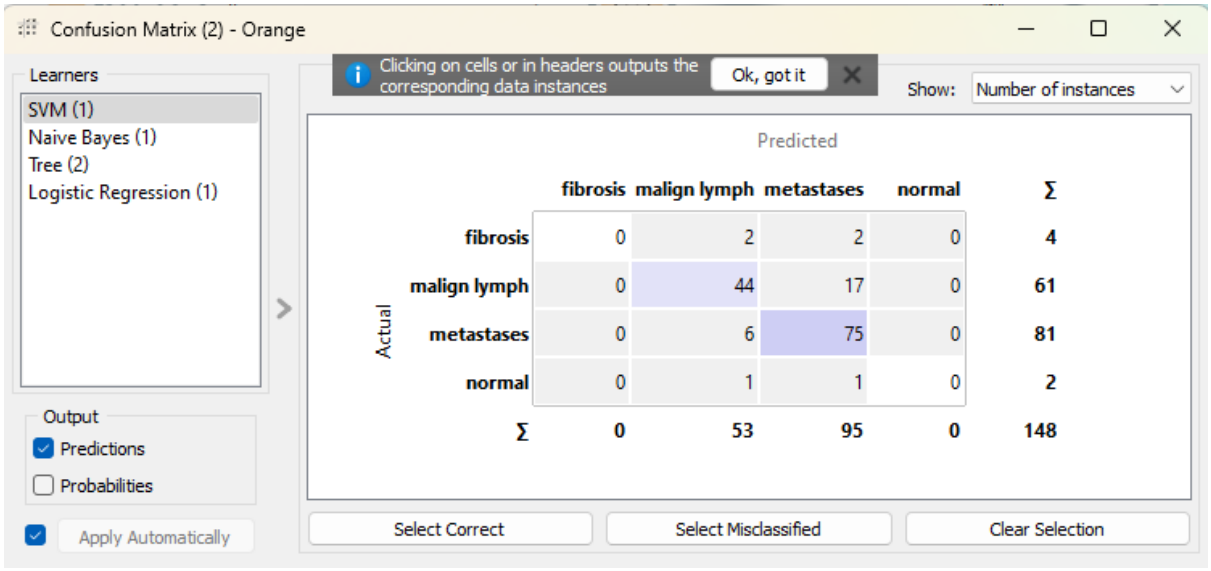
Before preprocessing:

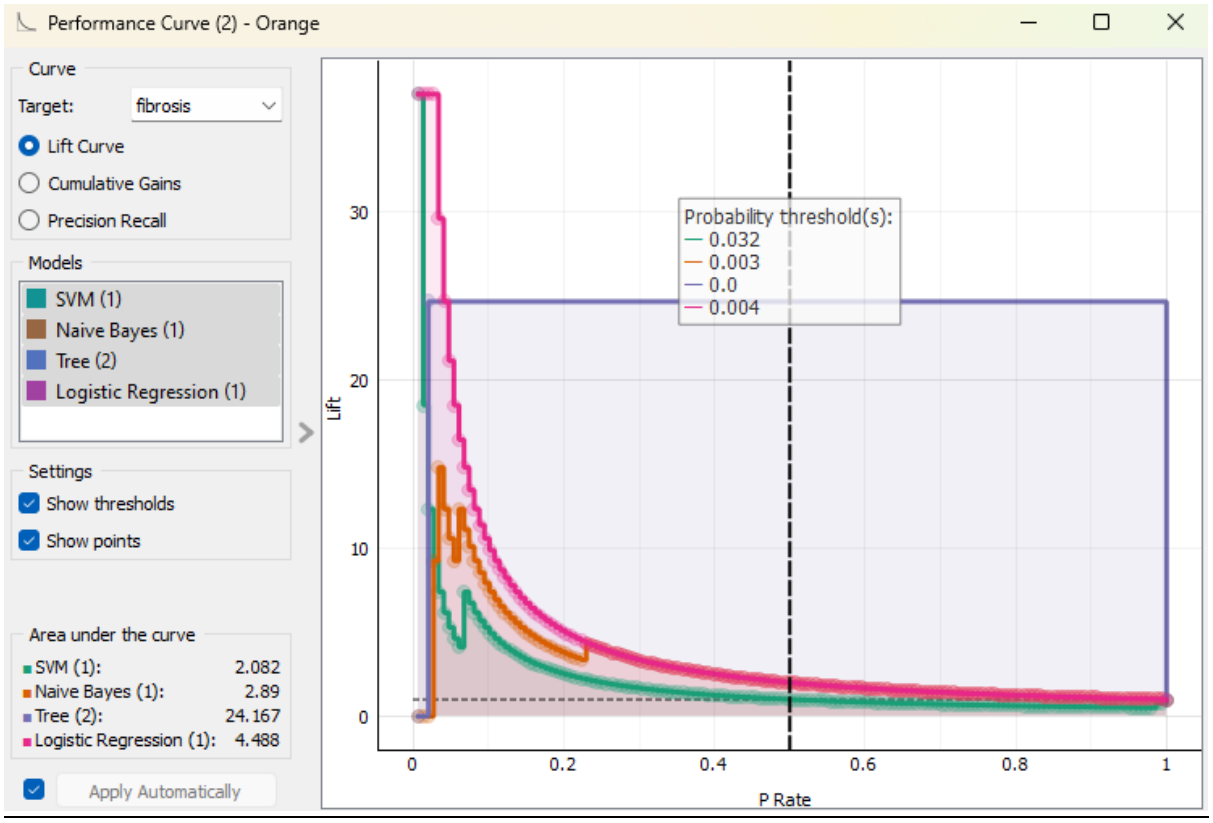


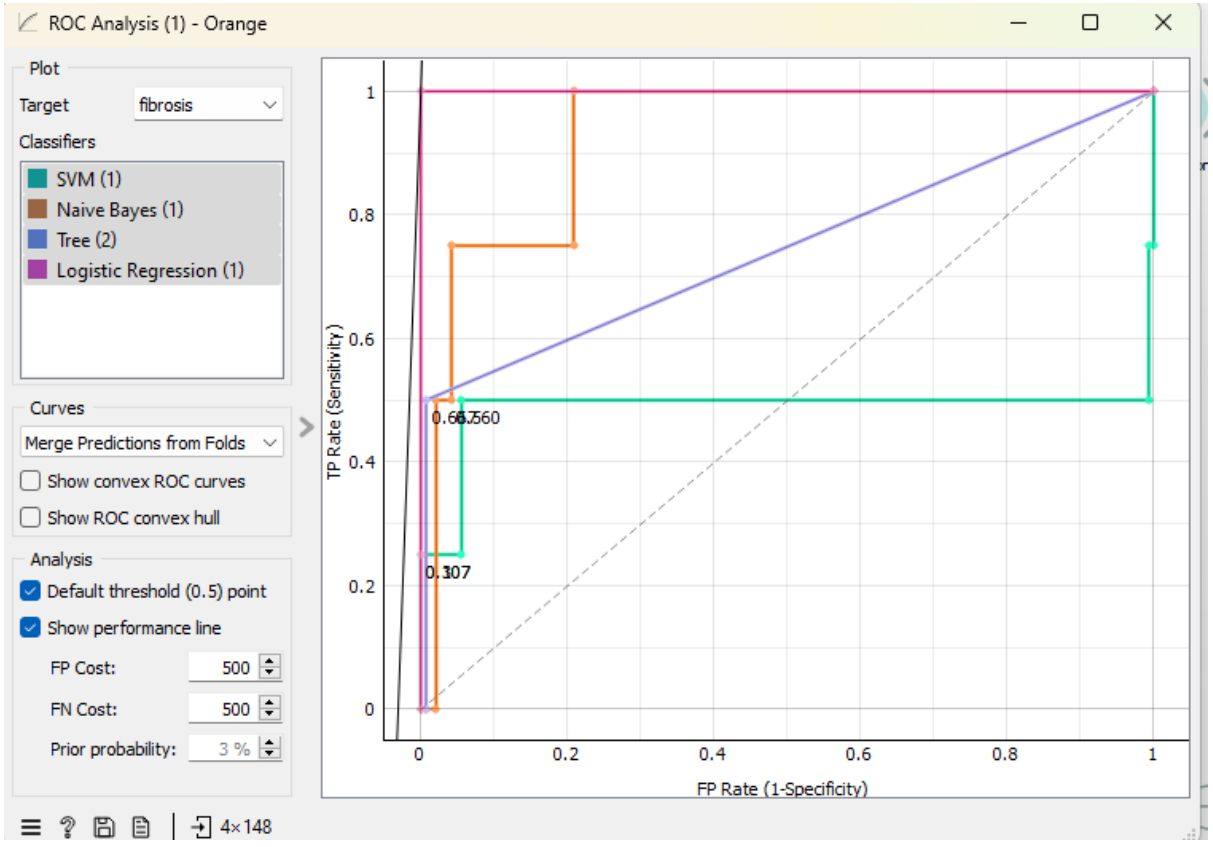




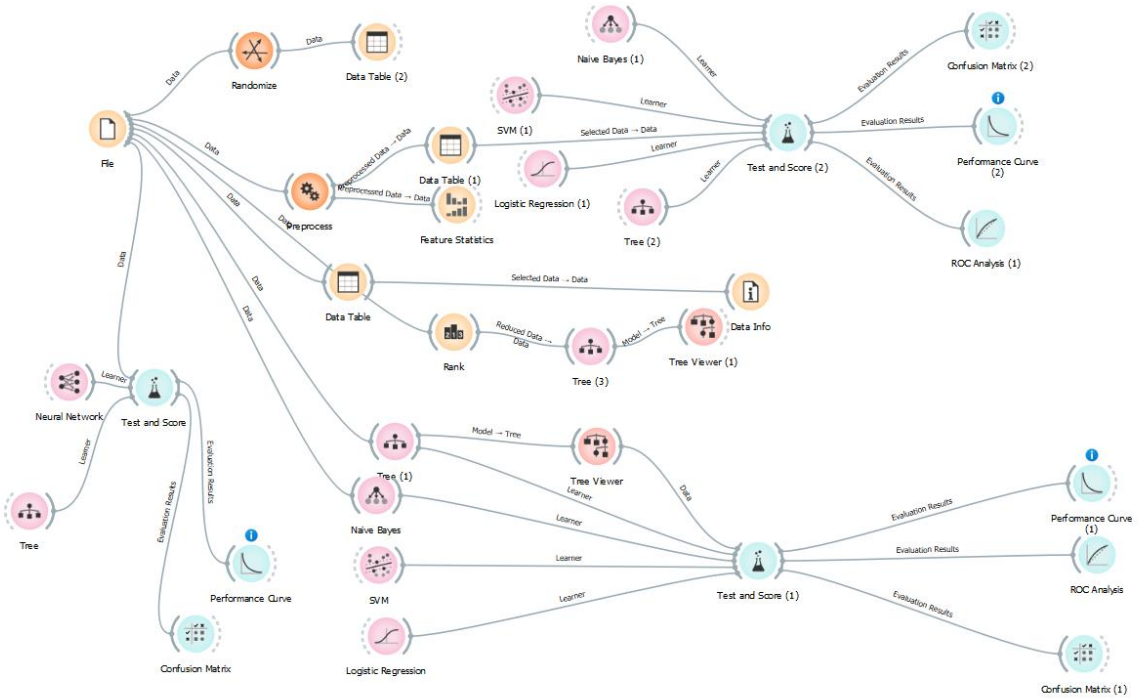
After Preprocessing:




Performance Curve (2) - Orange
Curve
Target: fibrosis
Lift Curve
Cumulative Gains
Precision Recall
Models
SVM (1)
Naive Bayes (1)
Tree (2)
Logistic Regression (1)
Settings
Show thresholds
Show points
Area under the curve
SVM (1): 2.082
Naive Bayes (1): 2.89
Tree (2): 24.167
Logistic Regression (1): 4.488
Apply Automatically
Probability threshold(s):
0.032
0.003
0.0
0.004
Lift
P Rate


ROC Analysis (1) - Orange
Plot
Target fibrosis
Classifiers
SVM (1)
Naive Bayes (1)
Tree (2)
Logistic Regression (1)
Curves
Merge Predictions from Folds
Show convex ROC curves
Show ROC convex hull
Analysis
Default threshold (0.5) point
Show performance line
FP Cost: 500
FN Cost: 500
Prior probability: 3 %
4×148
TP Rate (Sensitivity)
FP Rate (1-Specificity)

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.