



Nhóm 12

Kho và khai phá dữ liệu

Final Report

CÁC KỸ THUẬT KHAI PHÁ DỮ LIỆU

GVHD: TS. PHAN ĐÌNH VẤN

THỰC HIỆN: NHÓM 12

May 28, 2025

NHÓM 12 - THÀNH VIÊN NHÓM

STT	Họ và tên	Phân công công việc	% đóng góp
1	Vương Thị Mỹ Lệ	Tìm dữ liệu, Demo Decision tree và Logistic Regression, Làm slide	20%
2	Phan Thị Thanh Nhân	Tìm dữ liệu, Demo K-mean và Association rule, Làm slide	20%
3	Lâm Hồng Phúc	Tìm dữ liệu, Demo Random forest và Model Evaluation , Làm slide	20%
4	Võ Thị Thu Trang	Tìm dữ liệu, Demo KNN và SVM, Làm slide	20%
5	Nguyễn Ngọc Đan Trâm	Tìm dữ liệu, Demo Hierarchical và Model Evaluation, Làm slide	20%

Content

01

Decision
tree (ID3)

02

Random
forest

03

KNN

04

SVM

05

Logistic
Regression

06

K-mean

07

Hierarchical

08

Association
Rule

09

Model
Evaluation



Tổng quan về dữ liệu

Bộ dữ liệu dự đoán bệnh ung thư phổi dựa trên các yếu tố về nhân khẩu học và hành vi

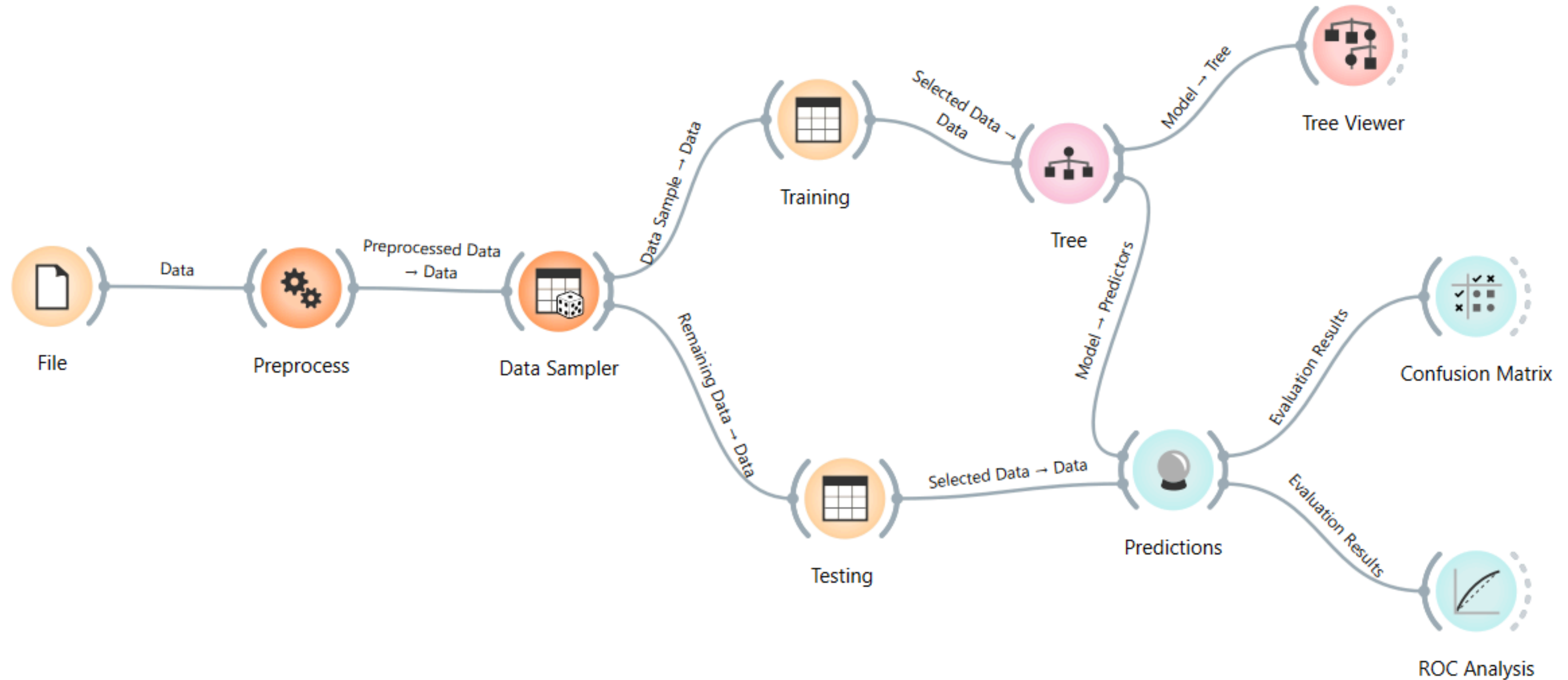


ID	Gender	Age	Marital Status	Children	Smoker	Employed	Years Worked	Income Level	Social Media	Online Gaming	Cancer
25137	Male	23	Married	3	No	No	0	Low	Yes	No	Yes
13063	Male	55	Single	4	Yes	No	0	Low	No	No	Yes
61685	Female	43	Single	3	No	Yes	25	Low	Yes	Yes	Yes
49303	Female	24	Married	0	Yes	Yes	15	High	Yes	Yes	No
84442	Female	54	Widowed	5	Yes	No	0	Low	No	Yes	No
99609	Female	19	Separated	0	No	Yes	2	Low	Yes	Yes	No
55657	Female	67	Separated	0	Yes	No	0	Low	Yes	No	Yes
40949	Male	19	Married	2	No	Yes	34	Low	No	Yes	No
11655	Male	41	Widowed	1	No	No	0	Low	No	Yes	No
77423	Female	87	Married	5	Yes	No	0	Low	No	Yes	Yes
81580	Male	32	Separated	1	Yes	Yes	24	Low	No	No	Yes
99321	Female	33	Married	1	Yes	Yes	2	Low	No	Yes	No
24196	Female	33	Separated	2	Yes	Yes	26	High	No	Yes	No
23230	Male	65	Single	2	No	No	0	Low	No	No	Yes
37932	Male	89	Married	1	Yes	Yes	18	Low	No	No	Yes
97651	Female	35	Separated	4	Yes	No	0	Low	No	No	No
91726	Female	50	Separated	0	No	Yes	25	Medium	Yes	No	No
40281	Female	39	Widowed	4	No	Yes	39	High	No	Yes	Yes
84778	Male	23	Widowed	2	No	Yes	25	High	No	No	Yes
85448	Male	52	Widowed	5	No	Yes	15	High	No	Yes	No
24065	Female	45	Widowed	1	No	Yes	23	Low	No	No	Yes



01

DECISION TREE



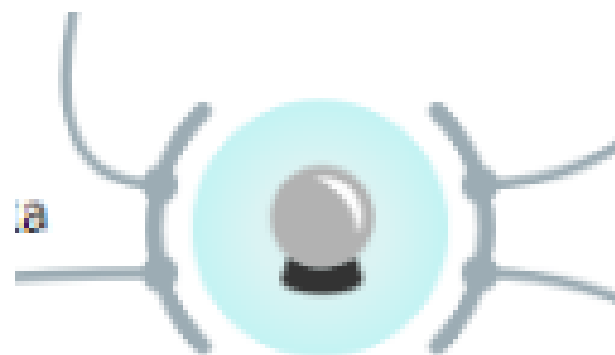
Demo Decision Tree bằng Orange

Vương Thị Mỹ Lệ

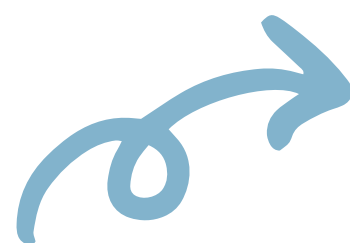


01

DECISION TREE

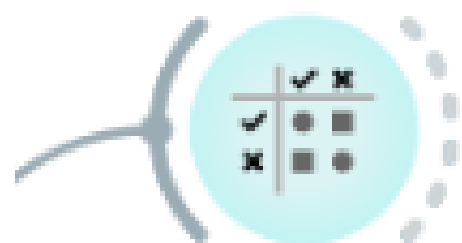


Predictions

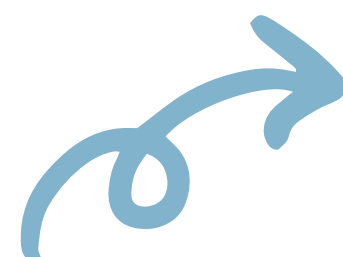


<input checked="" type="checkbox"/> Show performance scores					Target
Model	AUC	CA	F1	Prec	Recall
Tree	0.578	0.718	0.718	0.718	0.718

≡ ? 📄 | ➡ 149 | 📄 | ➡ 149 | 149 |



Confusion Matrix

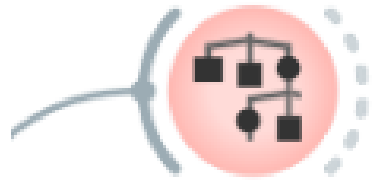


		Predicted		Σ
		No	Yes	
Actual	No	16	21	37
	Yes	21	91	112
Σ		37	112	149

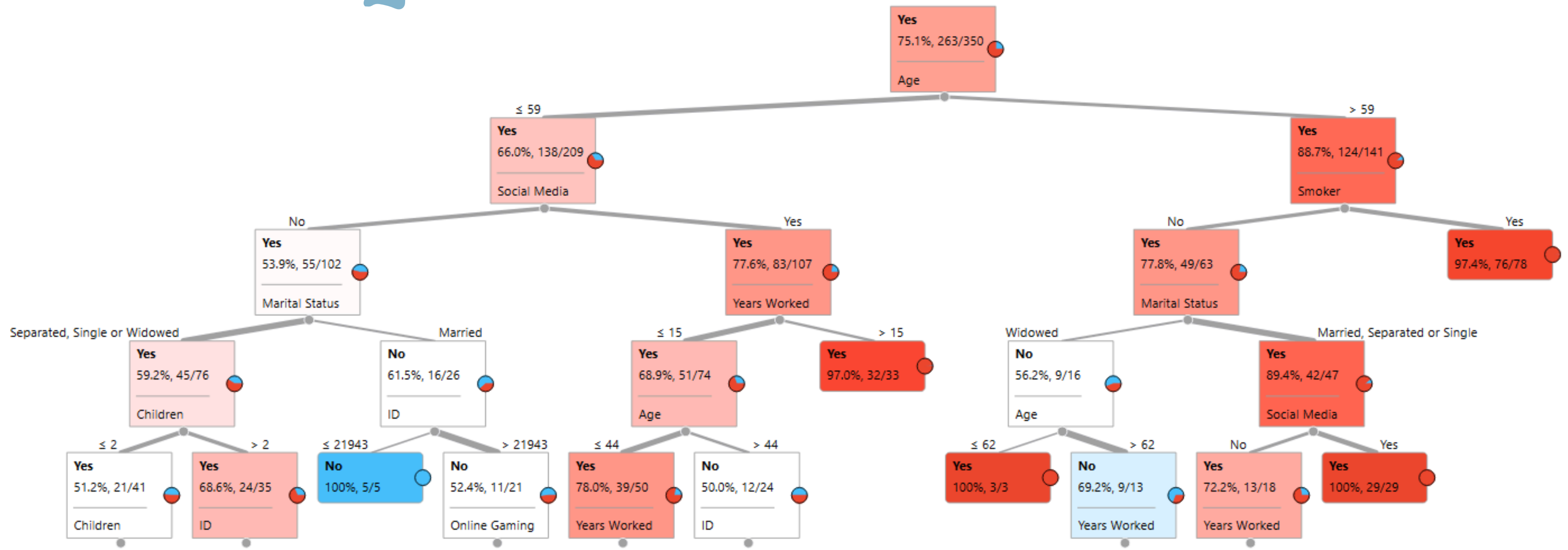
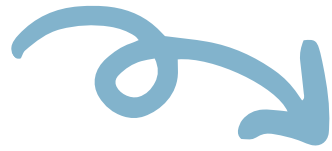


01

DECISION TREE



Tree Viewer

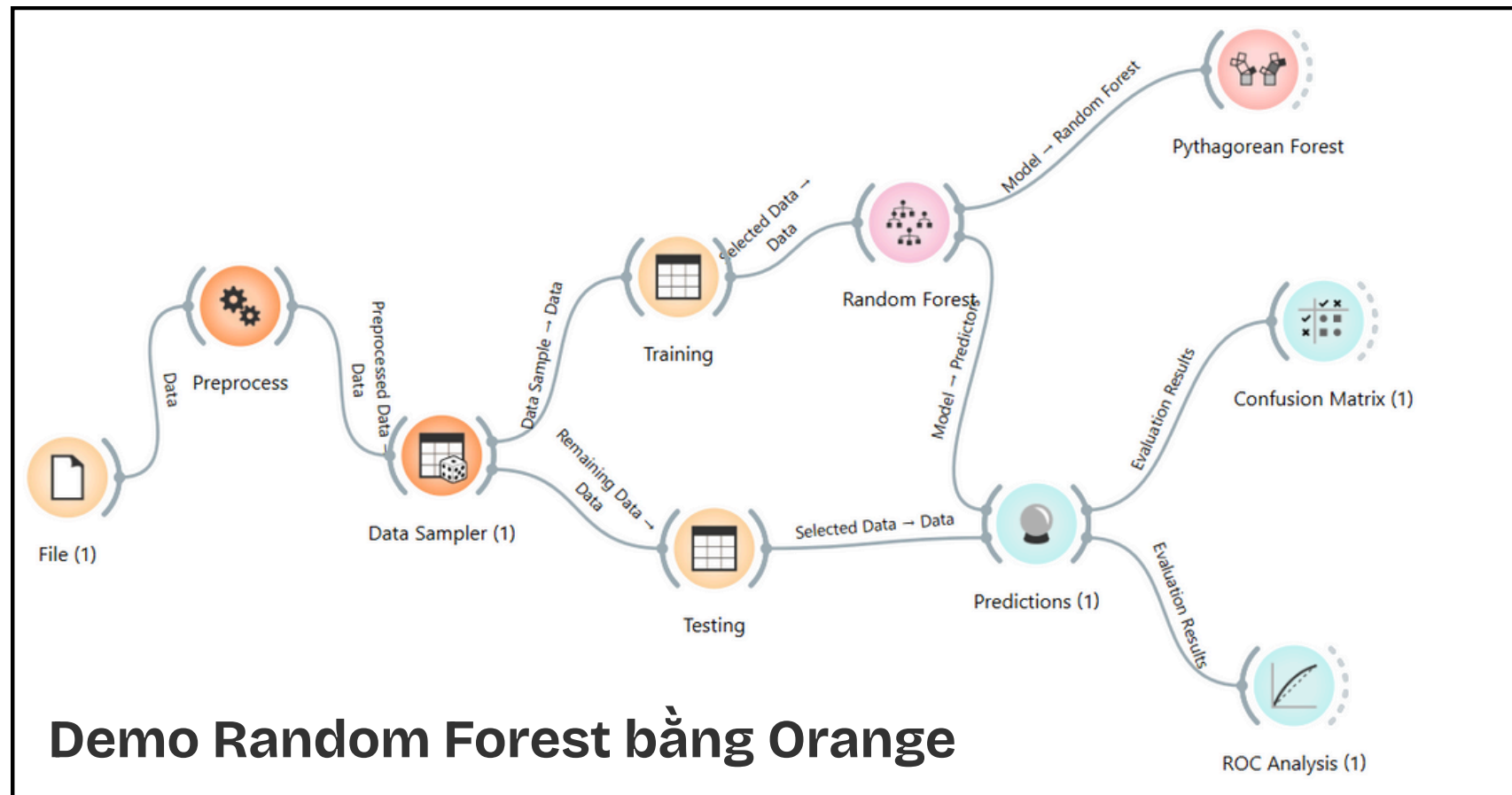


Vương Thị Mỹ Lệ

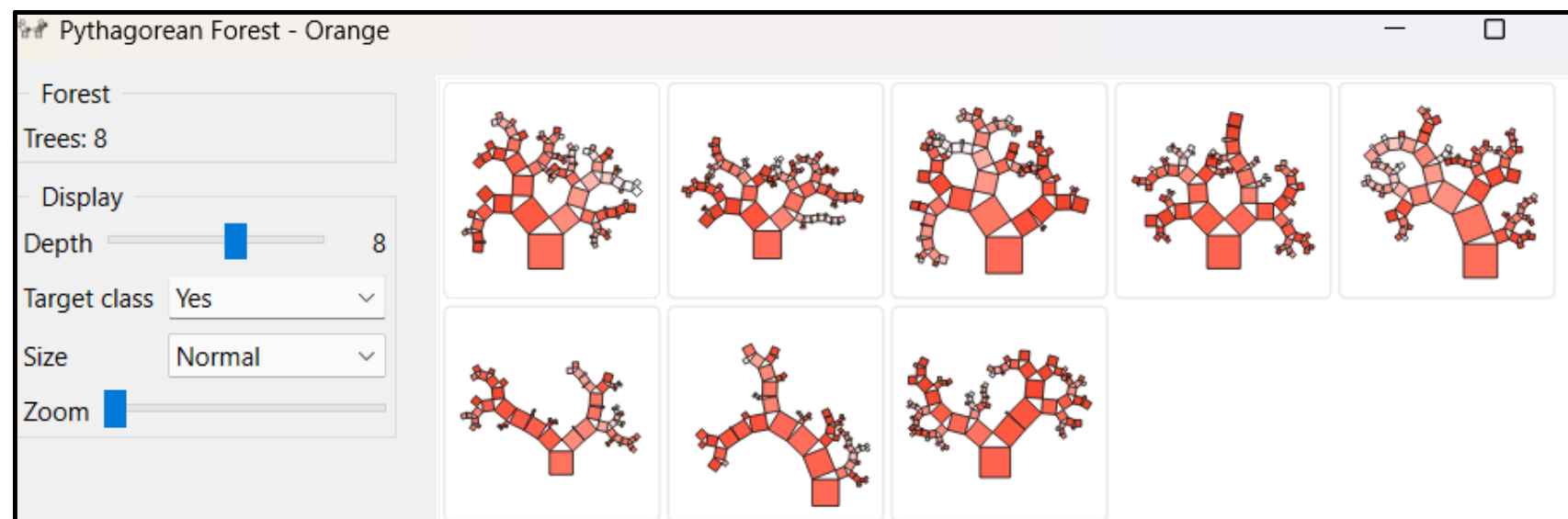


02

RANDOM FOREST



Demo Random Forest bằng Orange



Show performance scores						Target class:
Model	AUC	CA	F1	Prec	Recall	
Random Forest	0.733	0.765	0.738	0.737	0.765	

?

|

→

149

|

→

149

|

149

|

1×149

		Predicted		Σ
		No	Yes	
Actual	No	11	26	37
	Yes	9	103	112
Σ		20	129	149



03 KNN



kNN (k=5) - Or...

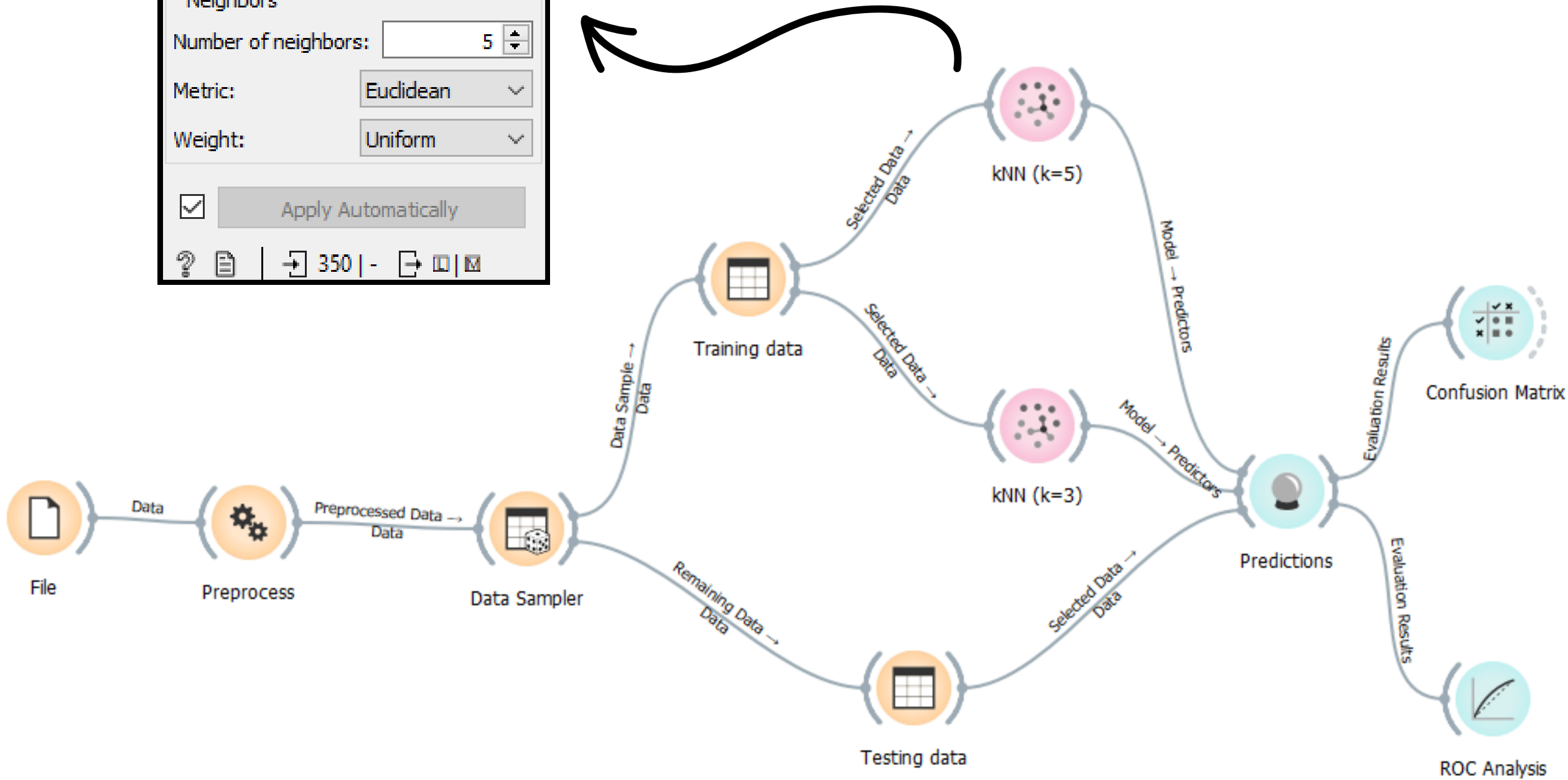
Name: kNN (k=5)

Neighbors: Number of neighbors: 5

Metric: Euclidean

Weight: Uniform

☒ Apply Automatically



K = 5

		Predicted		Σ
		No	Yes	
Actual	No	5	32	37
	Yes	5	107	112
Σ		10	139	149

K = 3

		Predicted		Σ
		No	Yes	
Actual	No	6	31	37
	Yes	10	102	112
Σ		16	133	149

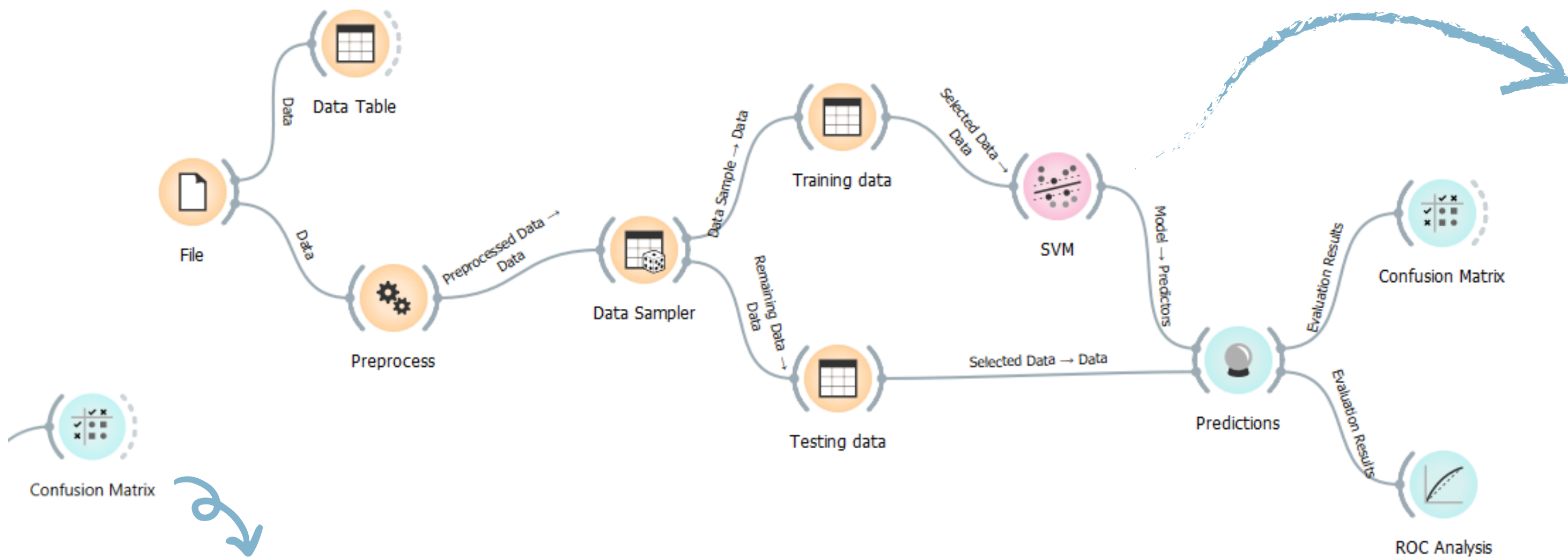
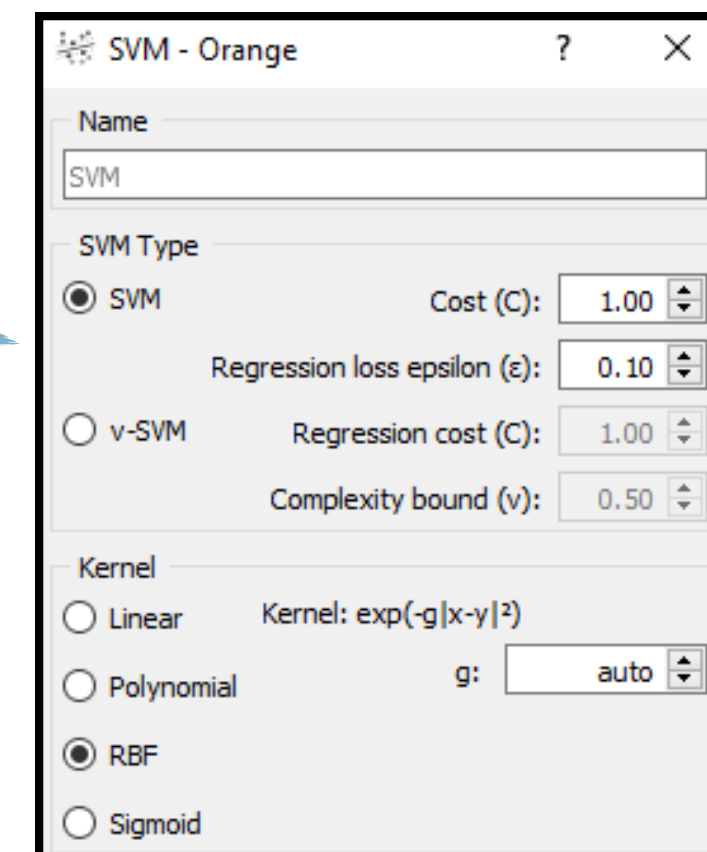


Model	AUC	CA	F1	Precision	Recall
kNN (k=5)	0.711	0.752	0.694	0.703	0.752
kNN (k=3)	0.665	0.725	0.682	0.670	0.725



04

SVM



Confusion Matrix

		Predicted		
		No	Yes	Σ
Actual	No	6	31	37
	Yes	3	109	112
Σ		9	140	149

Model	AUC	CA	F1	Prec	Recall
SVM	0.775	0.772	0.715	0.751	0.772



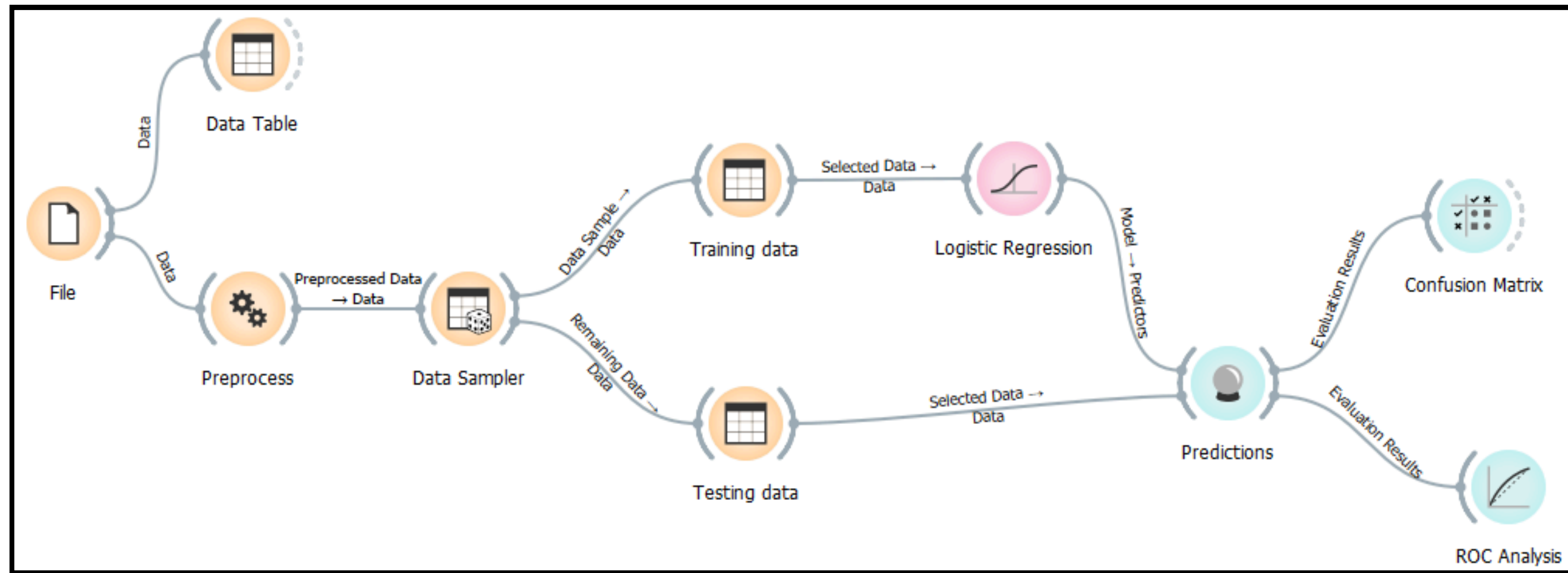
Predictions

Võ Thị Thu Trang



05

LOGISTIC REGRESSION



		Predicted		Σ
		No	Yes	
Actual	No	11	26	37
	Yes	6	106	112
Σ		17	132	149

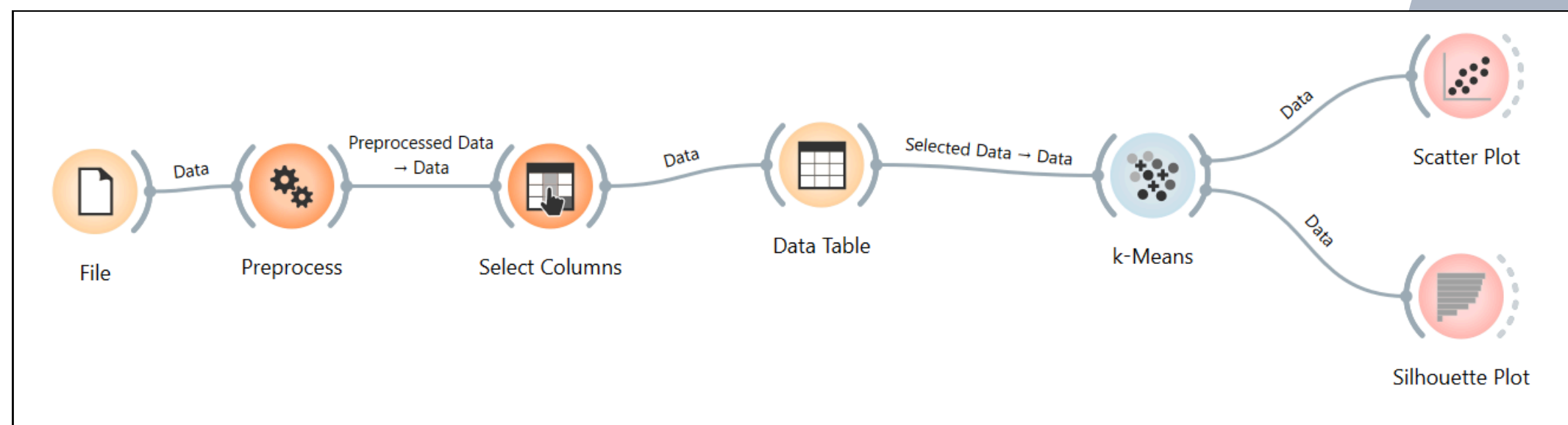


Model	AUC	CA	F1	Precision	Recall
Logistic Regression	0.832	0.785	0.754	0.764	0.785

Phan Thị Thanh Nhân

06

K - MEAN



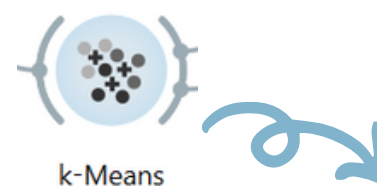
Features (3)

Filter

- ☒ Gender
- ☒ Age
- ☒ Smoker

Target (1)

☒ Cancer



k-Means - Orange

Number of Clusters

☒ Fixed: 3

☐ From 2 to 8

Preprocessing

☒ Normalize columns

Initialization

Initialize with KMeans++

Re-runs: 10

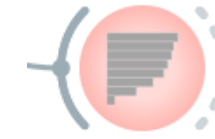
Maximum iterations: 300

☒ Apply Automatically

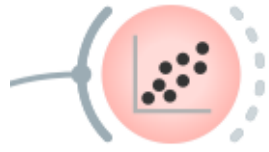
≡ ? | 499 | 499 | 3

06

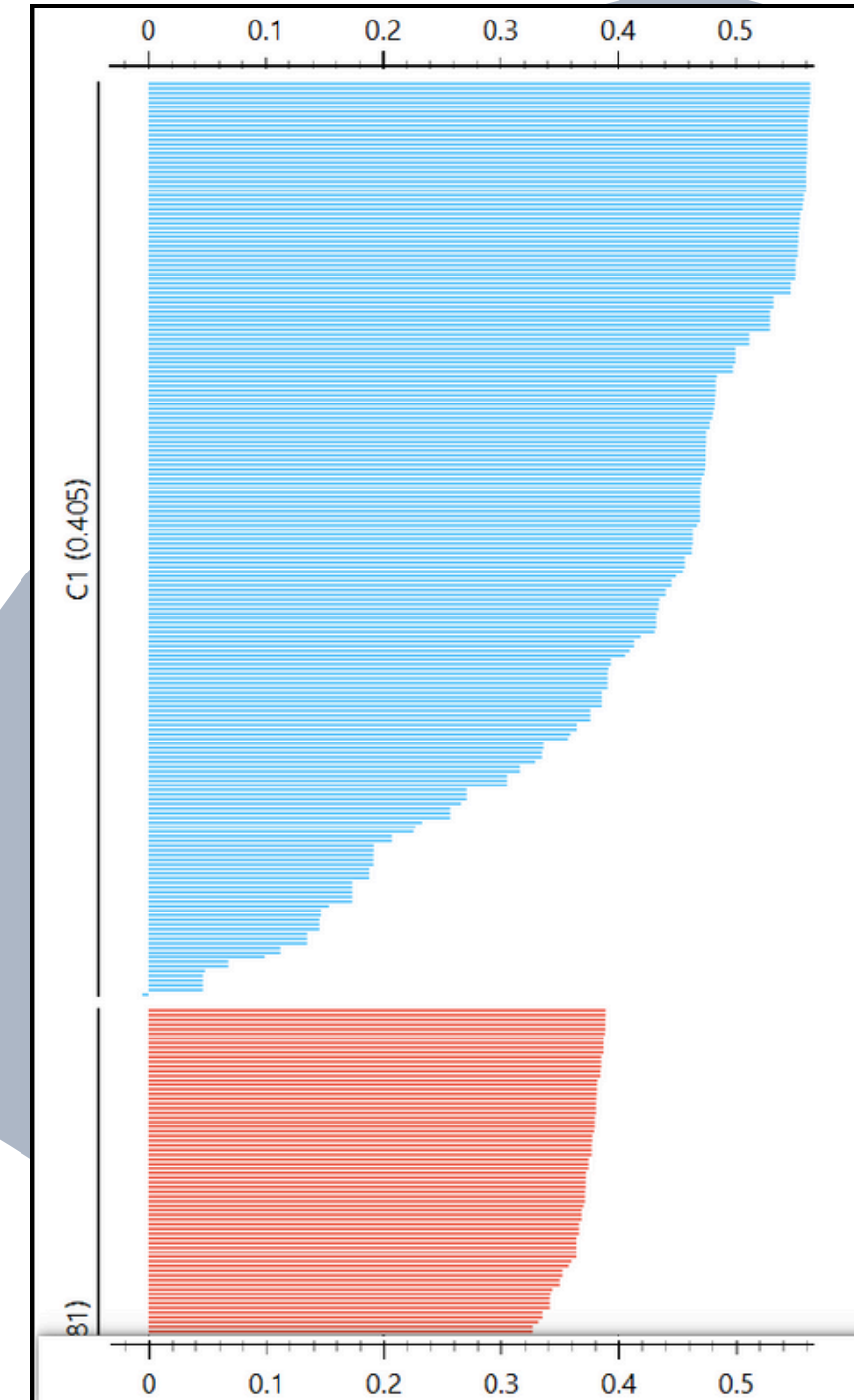
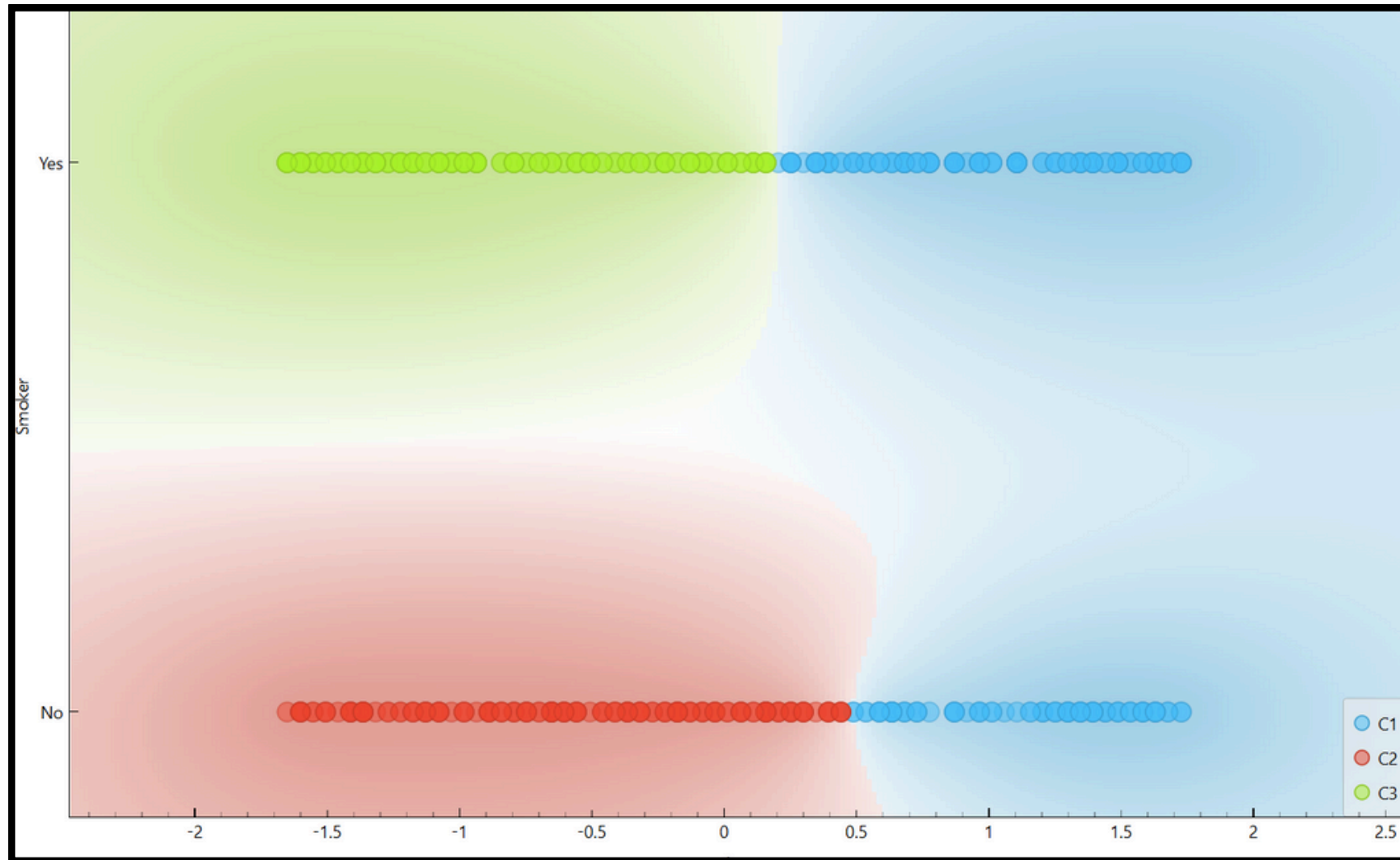
K - MEAN



Silhouette Plot



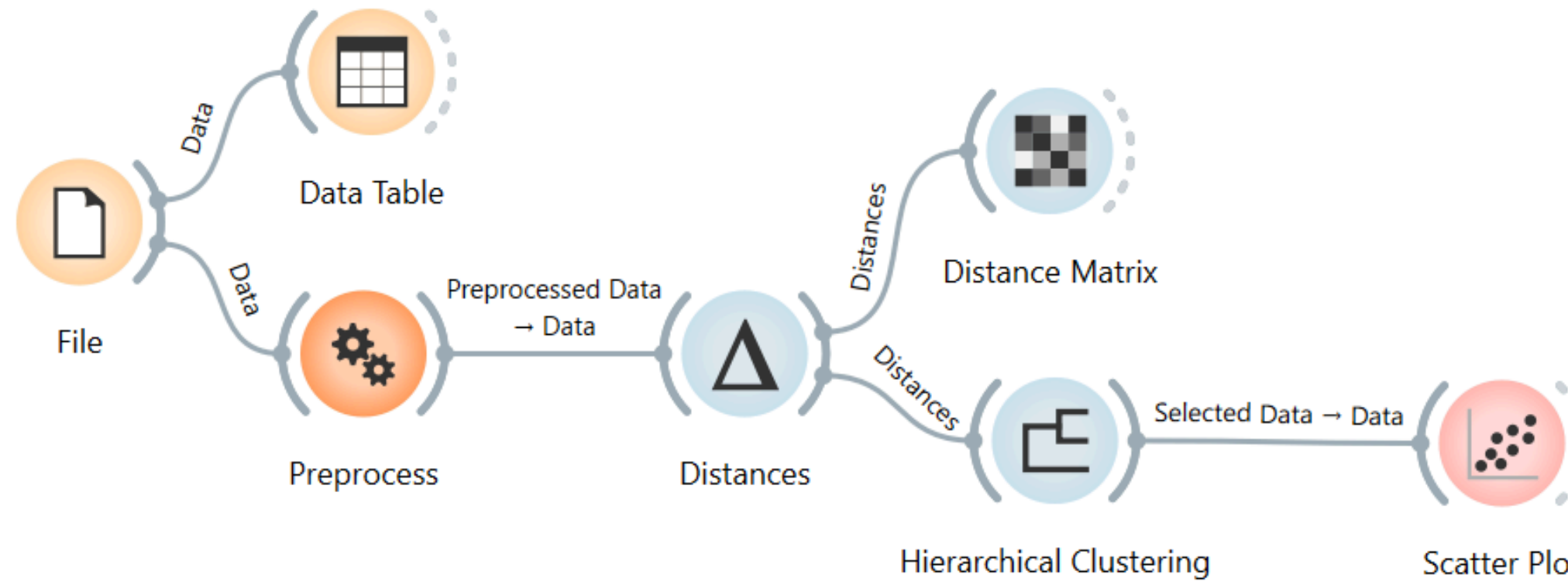
Scatter Plot





07

HIERARCHICAL



Distances

△ Distances - Orange

Compare

☒ Rows

☐ Columns

Distance Metric

☐ Euclidean (normalized)

☐ Cosine

☒ Euclidean

☐ Pearson

☐ Manhattan (normalized)

☐ Pearson (absolute)

☐ Manhattan

☐ Spearman

☐ Mahalanobis

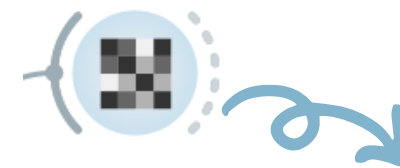
☐ Spearman (absolute)

☐ Hamming

☐ Jaccard

☒ Apply Automatically

≡ ? 📄 | ↗ 499 ↘ 499×499



Distance Matrix

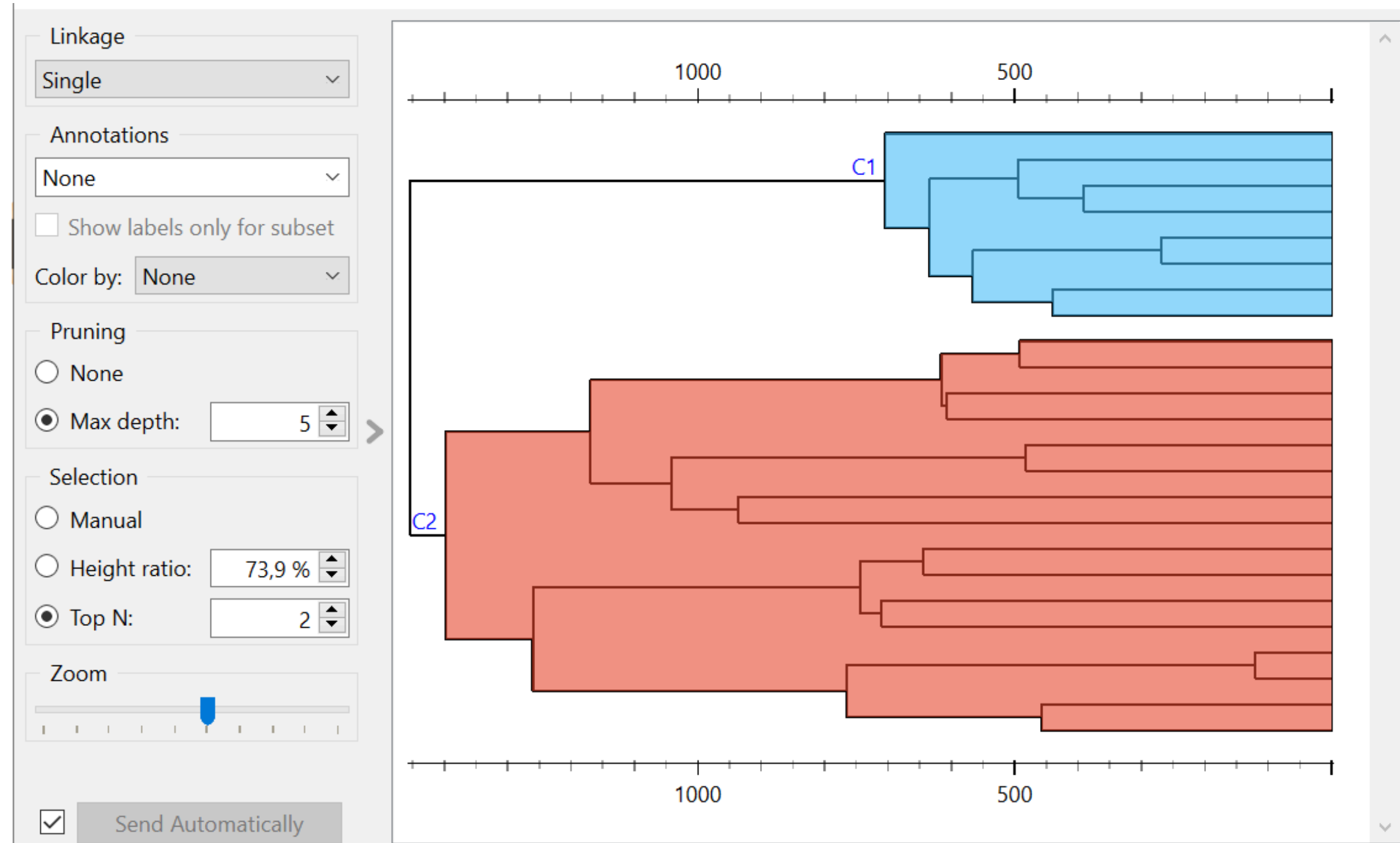
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
1		32.062	32.078	15.524	31.161	5.831	44.136	34.307	18.221	64.062	25.788	10.677	28.018	42.036	68.462	12.247	36.986	42.237	25.100	32.802	31.953	12.166	22.561	60.125	2.646	16.155	11.269	40.546	34.569
2	32.062		27.839	34.771	2.449	36.373	12.767	49.608	14.457	32.062	33.407	22.405	34.205	10.247	38.613	20.075	25.942	42.226	40.706	15.524	25.338	20.050	13.711	28.284	30.050	16.186	21.260	31.686	39.749
3	32.078	27.839		21.772	27.477	33.407	34.843	25.729	25.259	50.685	11.446	25.239	10.344	33.377	46.626	26.382	7.874	14.697	20.149	13.784	3.873	26.363	17.176	45.684	30.887	25.436	26.739	8.888	13.000
4	15.524	34.771	21.772		33.971	14.036	45.596	19.849	22.825	64.985	12.329	15.906	14.422	43.795	65.115	19.157	27.928	28.653	10.536	28.513	22.627	19.157	21.541	60.308	15.684	21.448	18.276	30.150	20.347
5	31.161	2.449	27.477	33.971		35.468	14.071	48.929	13.675	33.030	32.879	21.517	33.601	11.619	39.623	19.079	25.923	41.845	40.012	15.264	25.100	19.105	13.153	29.309	29.086	15.362	20.494	31.512	39.281
6	5.831	36.373	33.407	14.036	35.468		48.083	32.109	22.204	68.250	25.671	14.142	27.911	46.152	71.854	16.733	38.626	42.297	23.558	35.875	33.496	16.793	25.981	64.101	7.810	20.347	15.362	42.071	34.132
7	44.136	12.767	34.843	45.596	14.071	48.083		58.915	26.134	20.688	42.485	34.147	42.907	3.464	28.513	32.280	30.299	48.239	50.705	21.977	31.906	32.280	24.597	16.643	42.202	28.142	33.076	36.277	47.466
8	34.307	49.608	25.729	19.849	48.929	32.109	58.915		40.534	76.112	16.553	34.971	16.248	57.236	71.833	37.696	32.419	20.809	10.050	38.223	28.320	37.696	36.139	71.162	34.670	39.497	37.229	30.348	14.142
9	18.221	14.457	25.259	22.825	13.675	22.204	26.134	40.534		46.217	25.729	8.485	27.313	24.083	51.313	7.000	26.702	39.217	30.887	19.079	23.431	7.071	9.055	42.237	16.462	2.646	7.141	32.342	34.029
10	64.062	32.062	50.685	64.985	33.030	68.250	20.688	76.112	46.217		60.175	54.203	60.042	22.293	18.628	52.038	45.011	61.887	68.819	38.158	48.094	52.019	43.829	5.745	62.024	48.114	53.188	49.092	63.380
11	25.788	33.407	11.446	12.329	32.879	25.671	42.485	16.553	25.729	60.175		22.113	3.162	40.853	57.324	24.434	18.193	16.971	9.274	22.405	13.153	24.434	20.100	55.227	25.259	25.080	24.145	18.947	9.165
12	10.677	22.405	25.239	15.906	21.517	14.142	34.147	34.971	8.485	54.203	22.113		24.062	32.171	58.267	4.472	28.705	37.656	25.219	23.452	24.269	4.472	12.845	50.090	9.000	6.708	3.317	33.437	31.064
13	28.018	34.205	10.344	14.422	33.601	27.911	42.907	16.248	27.313	60.042	3.162	24.062		41.316	56.622	26.211	17.263	14.560	10.296	22.226	12.610	26.249	21.213	55.073	27.386	26.777	26.173	17.176	7.211
14	42.036	10.247	33.377	43.795	11.619	46.152	3.464	57.236	24.083	22.293	40.853	32.171	41.316		30.067	30.133	29.326	46.968	48.908	20.199	30.545	30.116	22.561	18.466	40.100	26.058	31.032	35.086	46.011
15	68.462	38.613	46.626	65.115	39.623	71.854	28.513	71.833	51.313	18.628	57.324	58.267	56.622	30.067		57.035	39.711	54.360	66.400	37.403	44.317	57.018	46.152	16.432	66.573	53.179	57.896	42.297	58.000
16	12.247	20.075	26.382	19.157	19.079	16.733	32.280	37.696	7.000	52.038	24.434	4.472	26.211	30.133	57.035		29.496	39.281	27.911	22.825	25.338	1.414	12.207	48.156	10.149	4.899	3.742	34.147	33.181
17	36.986	25.942	7.874	27.928	25.923	38.626	30.299	32.419	26.702	45.011	18.193	28.705	17.263	29.326	39.711	29.496		18.385	27.166	11.576	5.831	29.530	18.412	39.812	35.651	27.532	29.799	8.185	19.053
18	42.237	42.226	14.697	28.653	41.845	42.297	48.239	20.809	39.217	61.887	16.971	37.656	14.560	46.968	54.360	39.281	18.385		21.401	27.350	17.407	39.268	31.496	56.895	41.509	39.102	39.484	12.961	8.944
19	25.100	40.706	20.149	10.536	40.012	23.558	50.705	10.050	30.887	68.819	9.274	25.219	10.296	48.908	66.400	27.911	27.166	21.401		30.854	22.159	27.893	27.092	63.961	25.278	29.749	27.368	26.907	12.961
20	32.802	15.524	13.784	28.513	15.264	35.875	21.977	38.223	19.079	38.158	22.405	23.452	22.226	20.199	37.403	22.825	11.576	27.350	30.854		11.533	22.847	11.045	33.392	31.032	20.174	23.854	16.371	26.420
21	31.953	25.338	3.873	22.627	25.100	33.496	31.906	28.320	23.431	48.094	13.153	24.269	12.610	30.545	44.317	25.338	5.831	17.407	22.159	11.533		25.318	15.264	43.046	30.692	23.875	25.534	10.050	15.780
22	12.166	20.050	26.363	19.157	19.105	16.793	32.280	37.696	7.071	52.019	24.434	4.472	26.249	30.116	57.018	1.414	29.530	39.268	27.893	22.847	25.318		12.207	48.156	10.050	4.796	3.606	34.117	33.181
23	22.561	13.711	17.176	21.541	13.153	25.981	24.597	36.139	9.055	43.829	20.100	12.845	21.213	22.561	46.152	12.207	18.412	31.496	27.092	11.045	15.264	12.207		39.370	20.736	9.644	13.077	23.728	27.350

Nguyễn Ngọc Đan Trâm



07

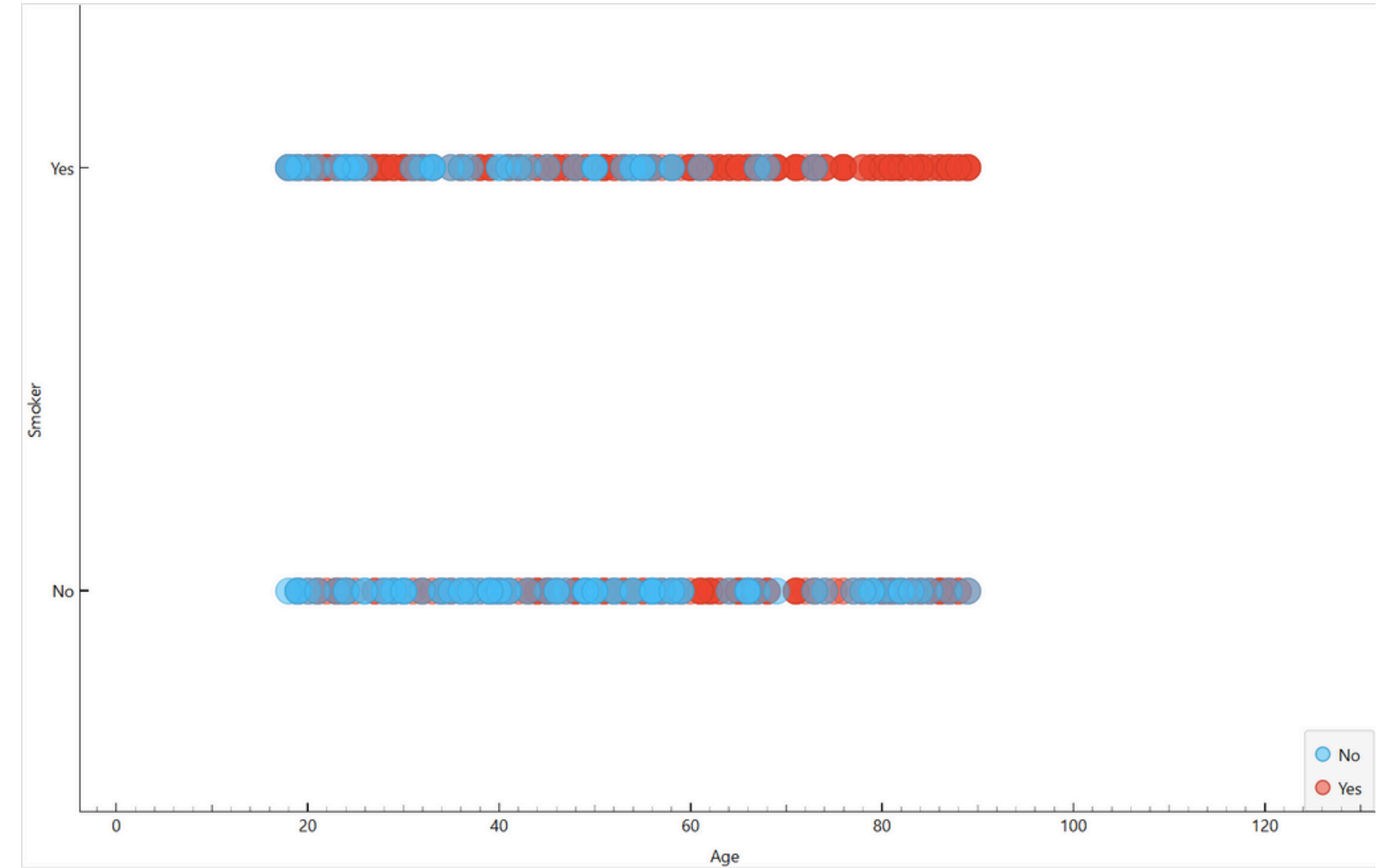
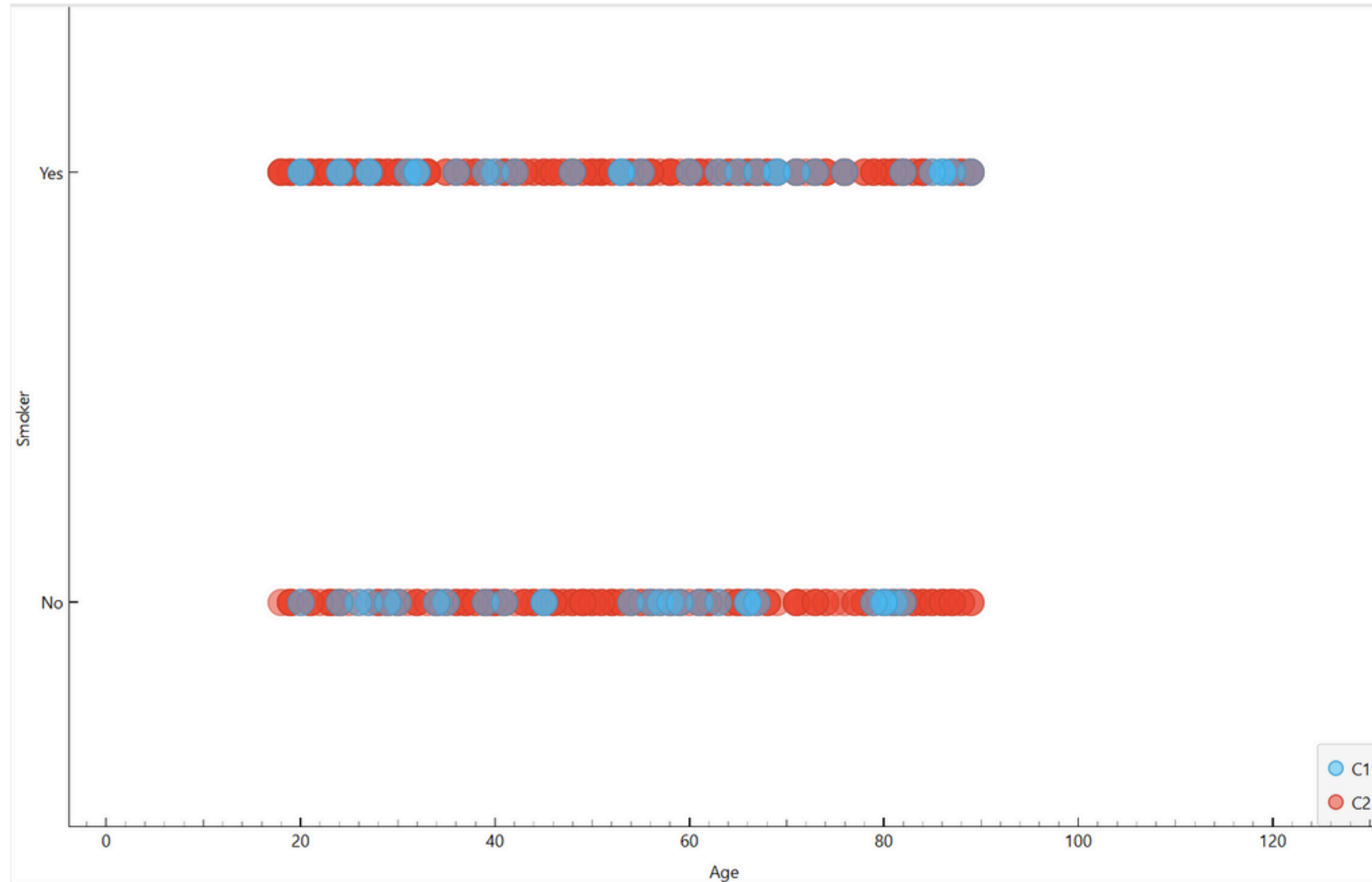
HIERARCHICAL





07

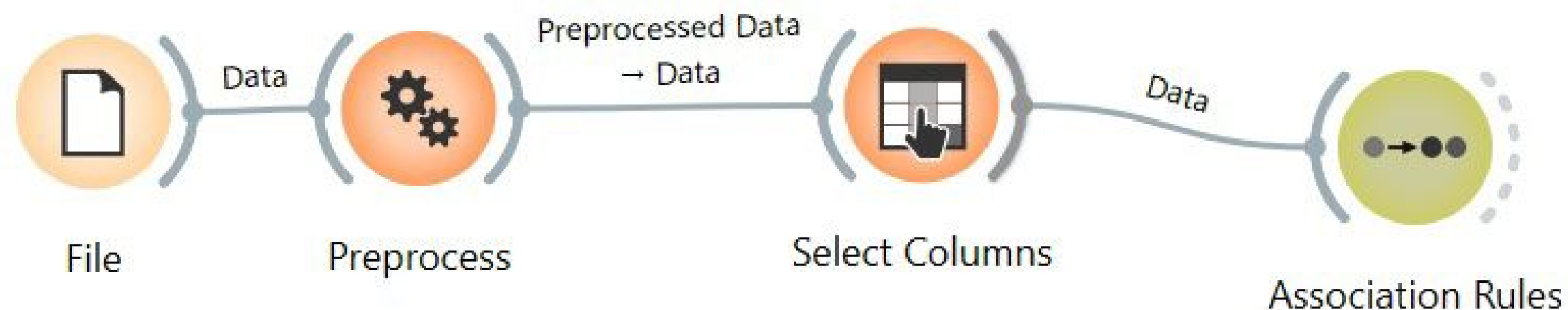
HIERARCHICAL





08

ASOCIATION RULE



Find association rules

Min. supp.: 30 %

Min. conf.: 60 %

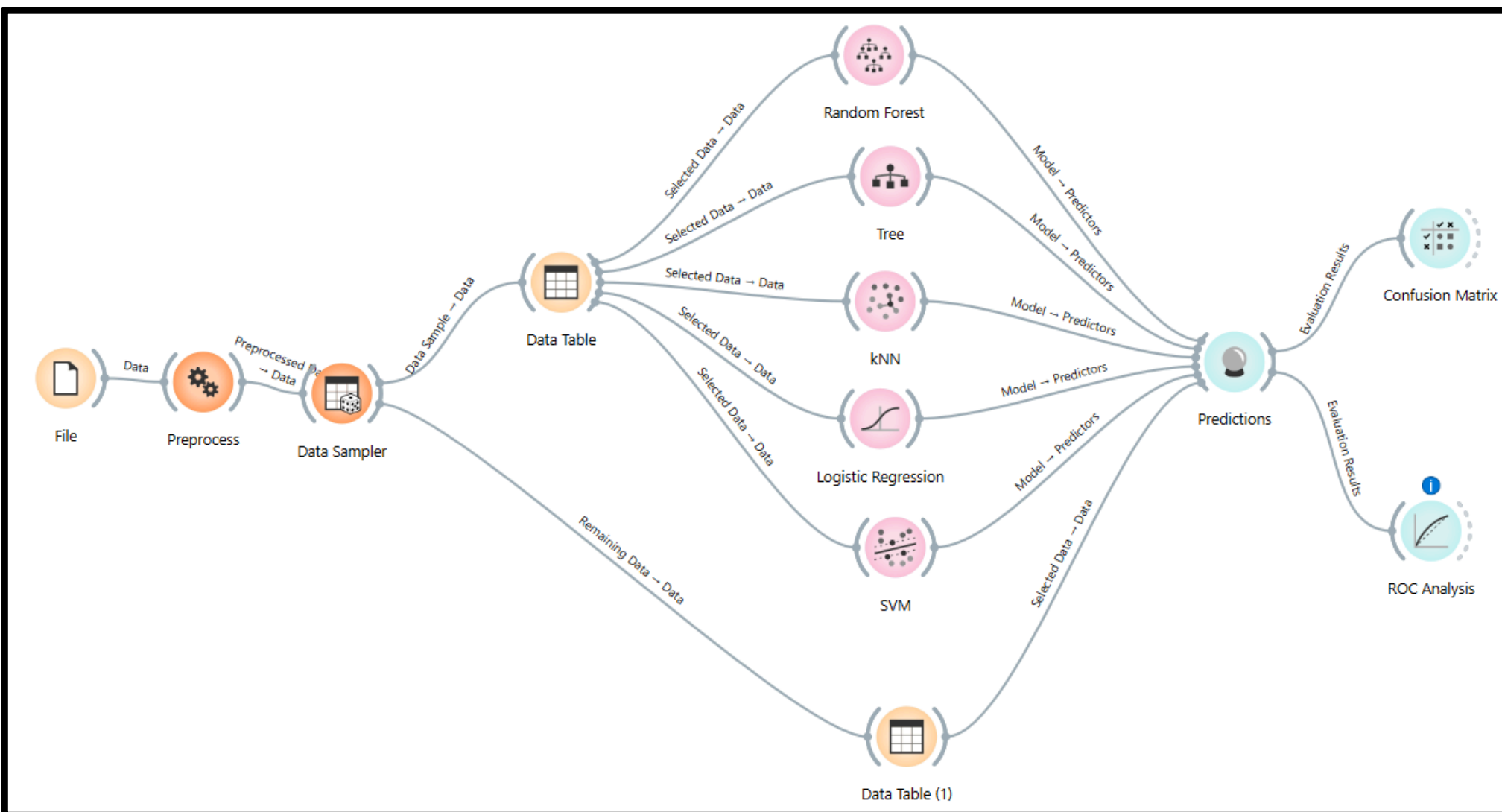
Max. rules: 10k

Supp	Conf	Covr	Strg	Lift	Levr	Antecedent	→	
0.301	0.649	0.463	1.623	0.864	-0.047	Smoker=No	→	Cancer=Yes
0.451	0.600	0.752	0.715	1.117	0.047	Cancer=Yes	→	Smoker=Yes
0.451	0.840	0.537	1.399	1.117	0.047	Smoker=Yes	→	Cancer=Yes
0.397	0.822	0.483	1.556	1.093	0.034	Online Gaming=No	→	Cancer=Yes
0.355	0.686	0.517	1.453	0.913	-0.034	Online Gaming=Yes	→	Cancer=Yes
0.315	0.668	0.471	1.596	0.889	-0.039	Social Media=No	→	Cancer=Yes
0.437	0.826	0.529	1.420	1.099	0.039	Social Media=Yes	→	Cancer=Yes



09

MODEL EVALUATION



Model	AUC	CA	F1	Prec	Recall	MCC
SVM	0.760	0.779	0.727	0.763	0.779	0.280
Logistic Regression	0.803	0.799	0.770	0.785	0.799	0.380
kNN	0.680	0.725	0.703	0.693	0.725	0.171
Tree	0.532	0.705	0.705	0.705	0.705	0.209
Random Forest	0.739	0.772	0.752	0.749	0.772	0.313

Lâm Hồng Phúc



09

MODEL EVALUATION

Decision tree

		Predicted		Σ
		No	Yes	
Actual	No	16	21	37
	Yes	21	91	112
Σ		37	112	149

Random forest

		Predicted		Σ
		No	Yes	
Actual	No	11	26	37
	Yes	9	103	112
Σ		20	129	149

KNN

		Predicted		Σ
		No	Yes	
Actual	No	5	32	37
	Yes	5	107	112
Σ		10	139	149

		Predicted		Σ
		No	Yes	
Actual	No	6	31	37
	Yes	3	109	112
Σ		9	140	149

SVM

		Predicted		Σ
		No	Yes	
Actual	No	11	26	37
	Yes	6	106	112
Σ		17	132	149

Logistic Regression

Lâm Hồng Phúc



Nhóm 12

Kho và khai phá dữ liệu

**Thank
You**



May 28, 2025