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 Nhàn | 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



# 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

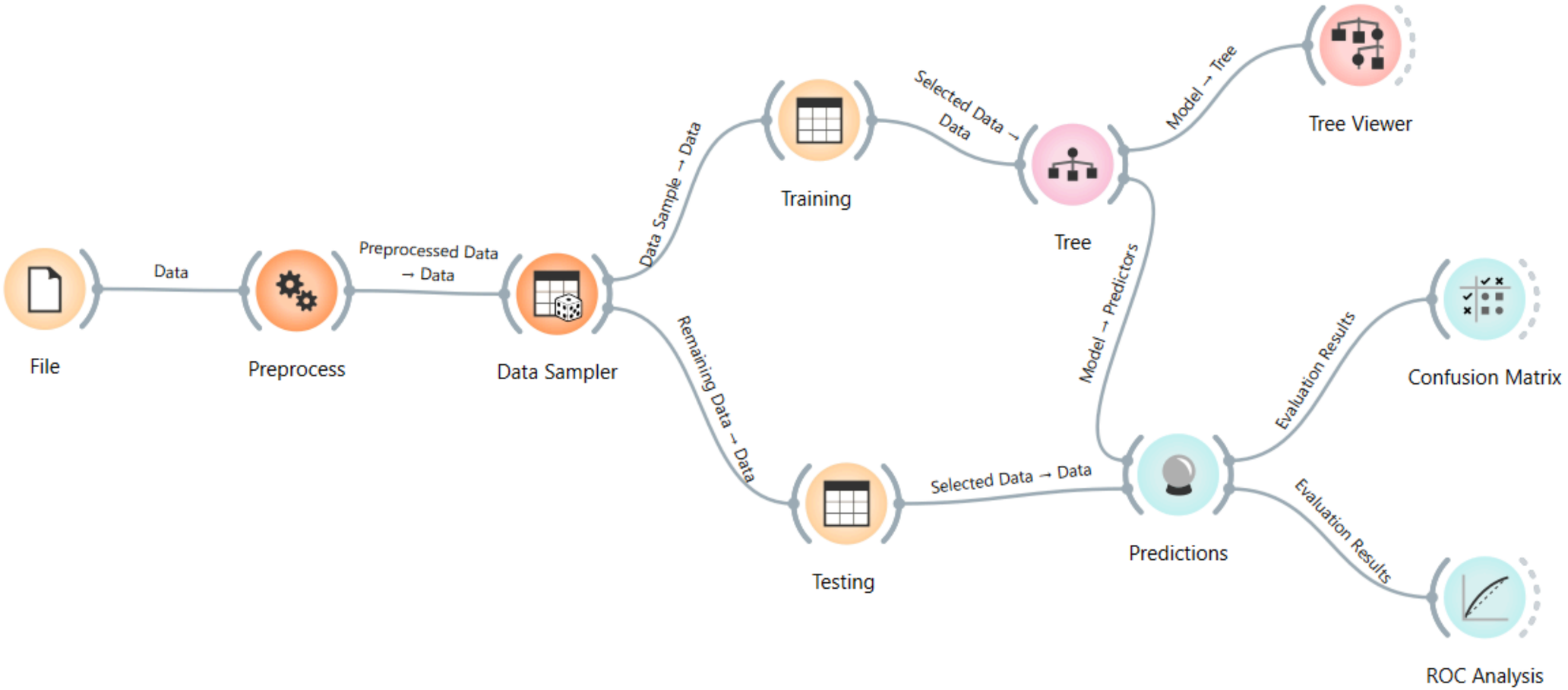
| ID | Gender | Age | Marital Status | Children | Smoker | Employed | Years Worked | Income Level | Social Media | Online Gaming | Cancer |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 25137 | Male | 23 | Married | 3 | No | No | 0 | Low | Yes | No | Yes |
| 13063 | Male | 55 | Single | 4 | Yes | 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 |

12
Cột

499
Dòng

Vương Thị Mỹ Lệ

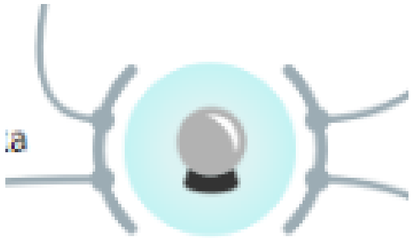
# 01 DECISION TREE




**Demo Decision Tree bằng Orange**

**Vương Thị Mỹ Lệ**

# 01 DECISION TREE

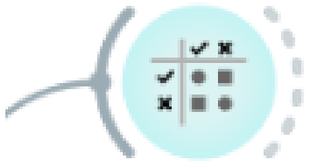


Predictions

Show perfomance scores Target

| Model | AUC | CA | F1 | Prec | Recall |
|---|---|---|---|---|---|
| Tree | 0.578 | 0.718 | 0.718 | 0.718 | 0.718 |

149 | M 149 | 149 |



Confusion Matrix

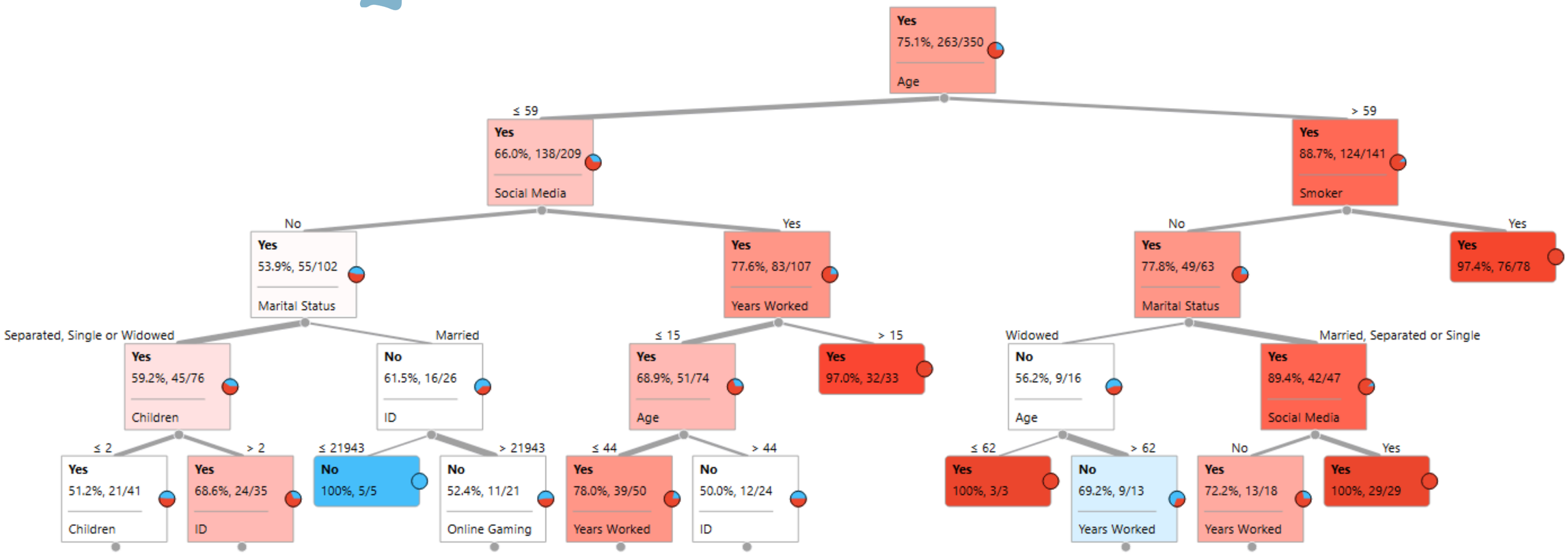
Predicted

| Actual | | No | Yes | Σ |
|---|---|---|---|---|
| | No | 16 | 21 | 37 |
| | Yes | 21 | 91 | 112 |
| | Σ | 37 | 112 | 149 |

**Vương Thị Mỹ Lệ**

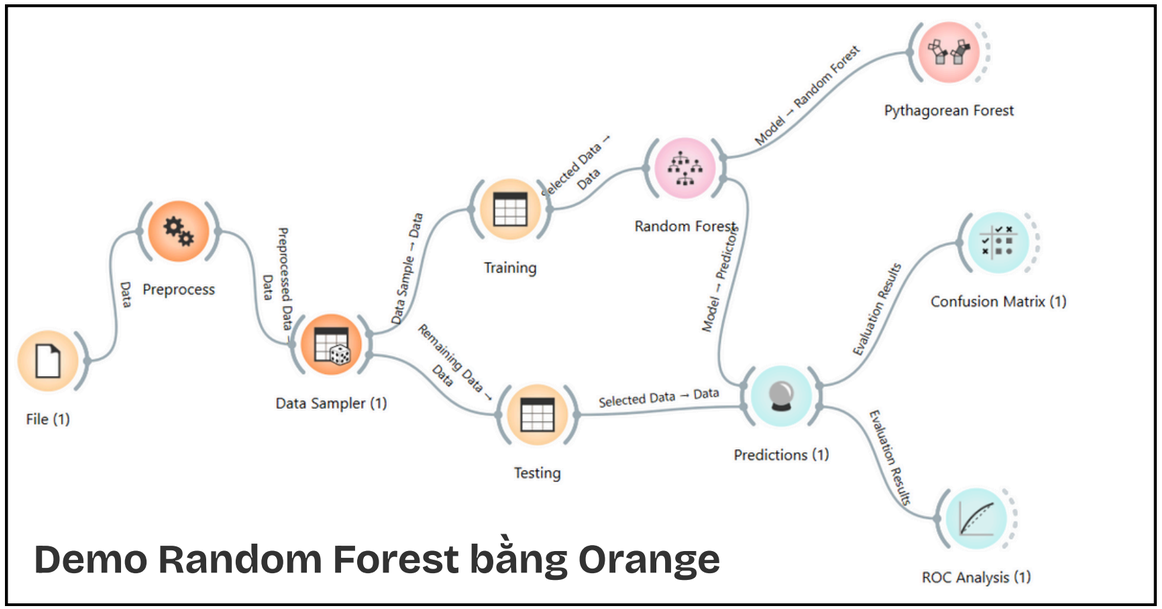
# 01 DECISION TREE

Tree Viewer

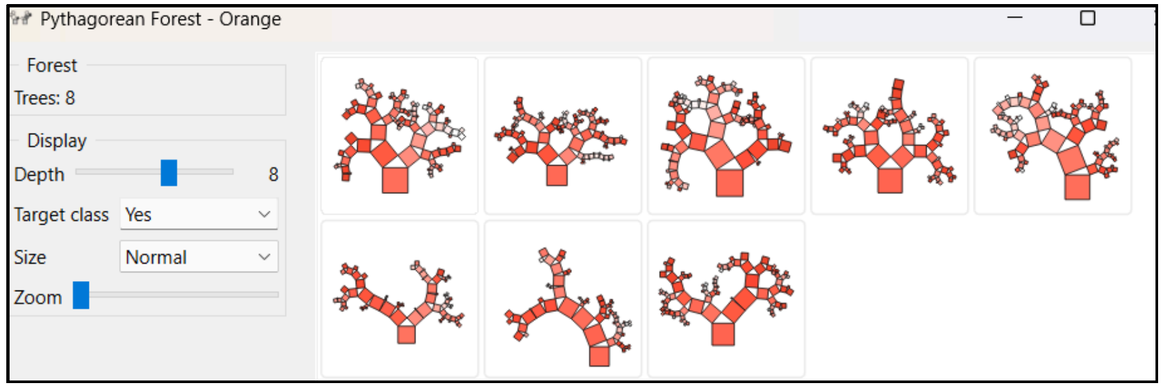



Vương Thị Mỹ Lệ

# 02 RANDOM FOREST




**Demo Random Forest bằng Orange**




Show perfomance scores — Target class:

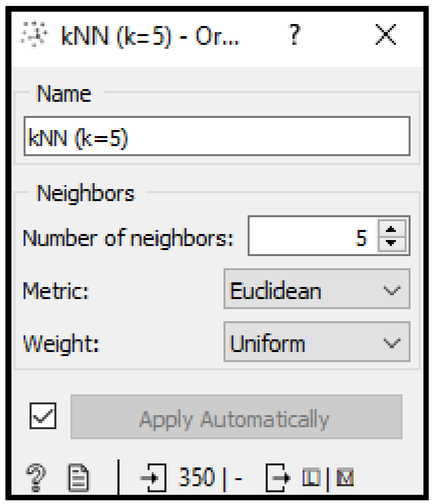
| Model | AUC | CA | F1 | Prec | Recall |
|---|---|---|---|---|---|
| Random Forest | 0.733 | 0.765 | 0.738 | 0.737 | 0.765 |

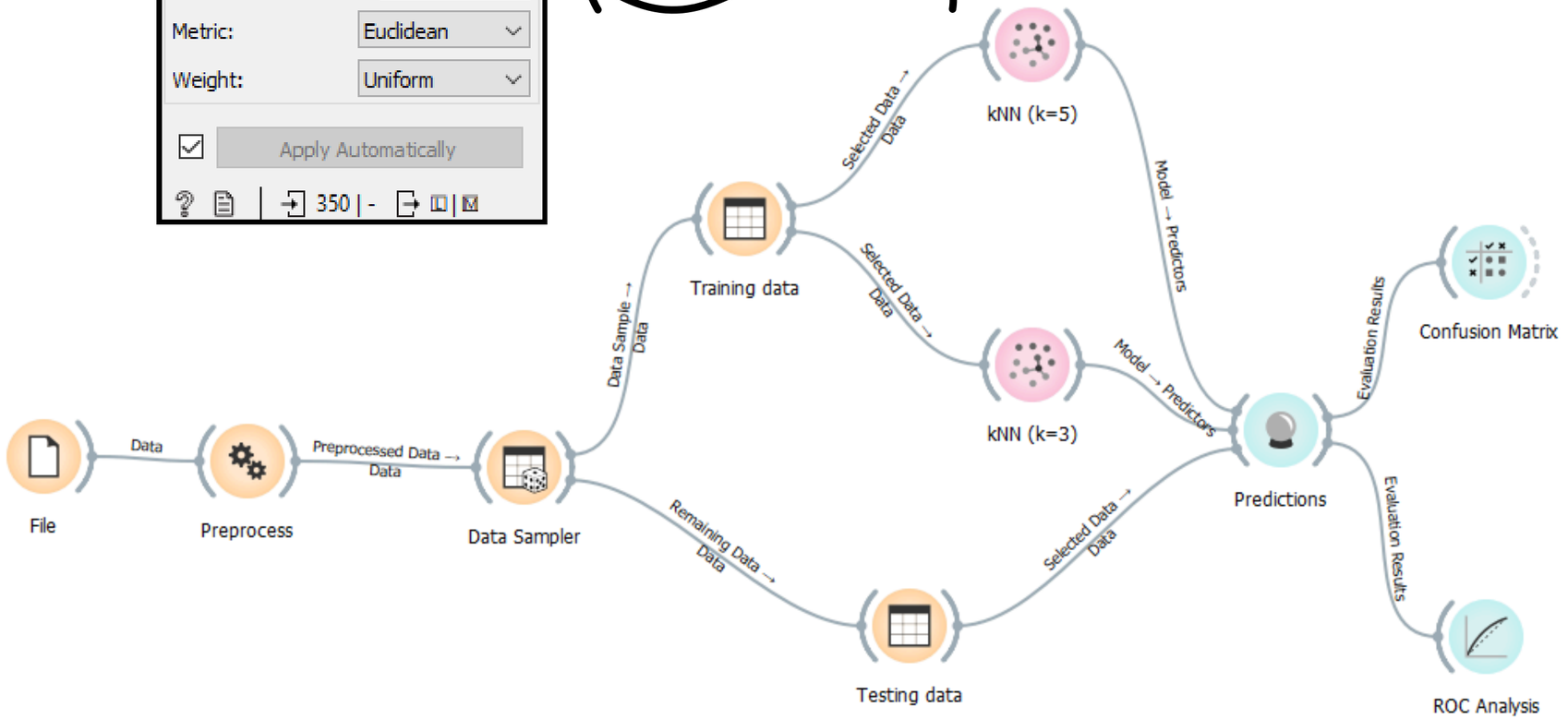
149 | M — 149 | 149 | 1×149

| Actual \ Predicted | No | Yes | ∑ |
|---|---|---|---|
| No | 11 | 26 | 37 |
| Yes | 9 | 103 | 112 |
| ∑ | 20 | 129 | 149 |

**Vương Thị Mỹ Lệ**

# 03 KNN




K = 5

Predicted

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

Predicted

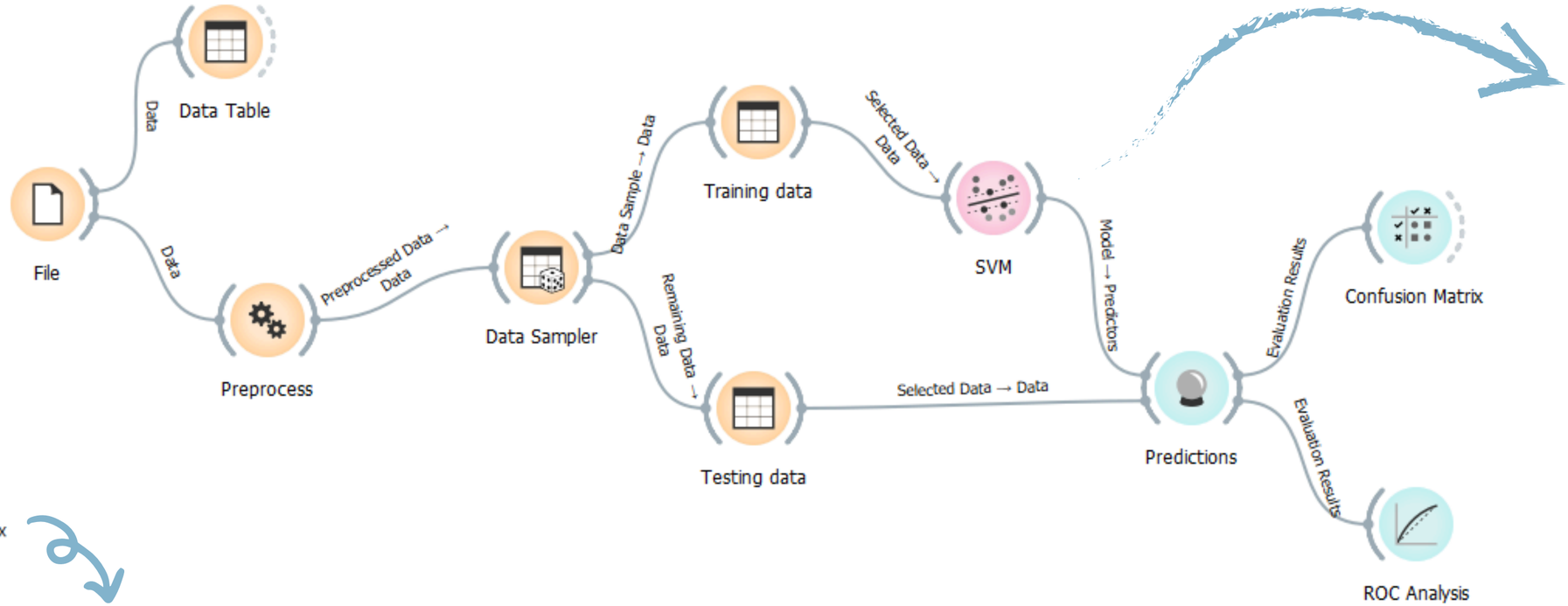
| Actual | No | Yes | Σ |
|---|---|---|---|
| No | 6 | 31 | 37 |
| Yes | 10 | 102 | 112 |
| Σ | 16 | 133 | 149 |

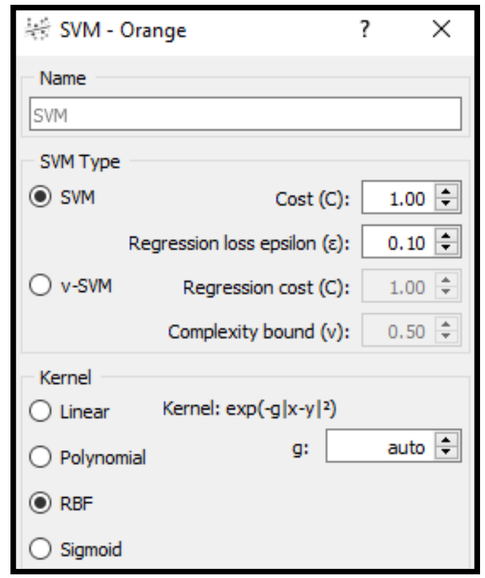
K = 3


| 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 |

Võ Thị Thu Trang

# 04 SVM





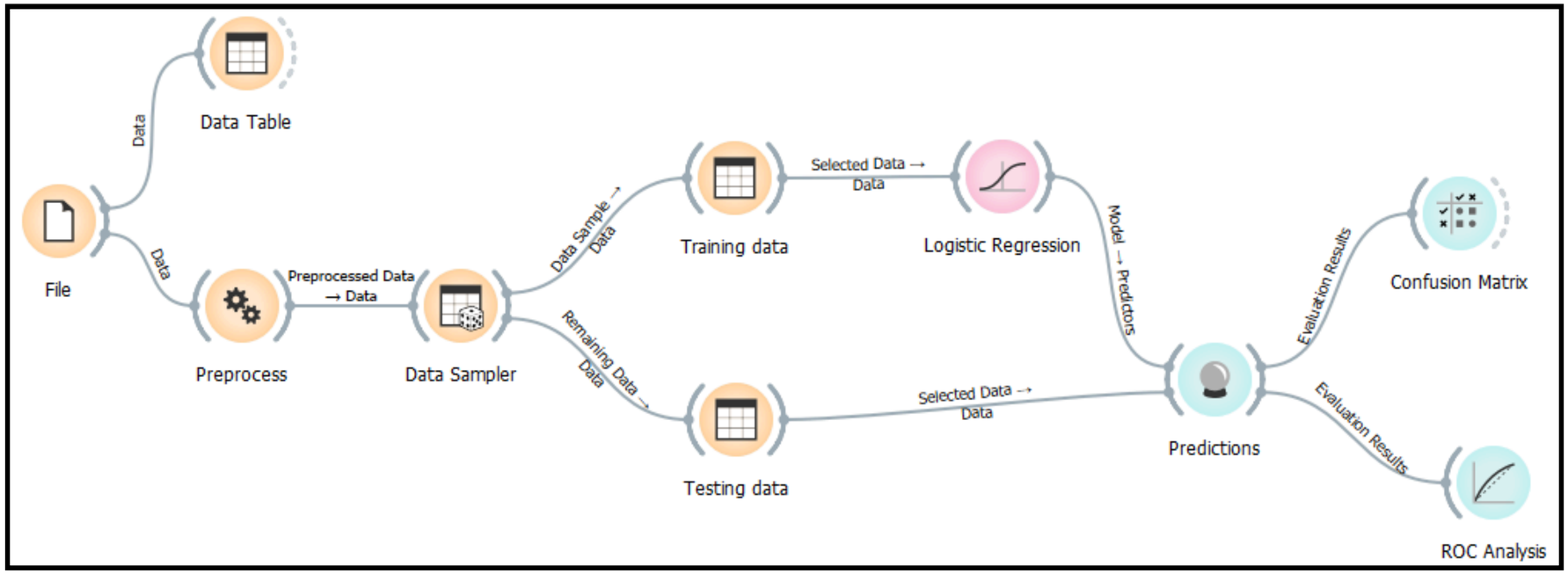
Predicted

| Actual | | No | Yes | ∑ |
|---|---|---|---|---|
| | No | 6 | 31 | 37 |
| | Yes | 3 | 109 | 112 |
| | ∑ | 9 | 140 | 149 |

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

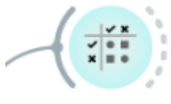

**Võ Thị Thu Trang**

# 05 LOGISTIC REGRESSION





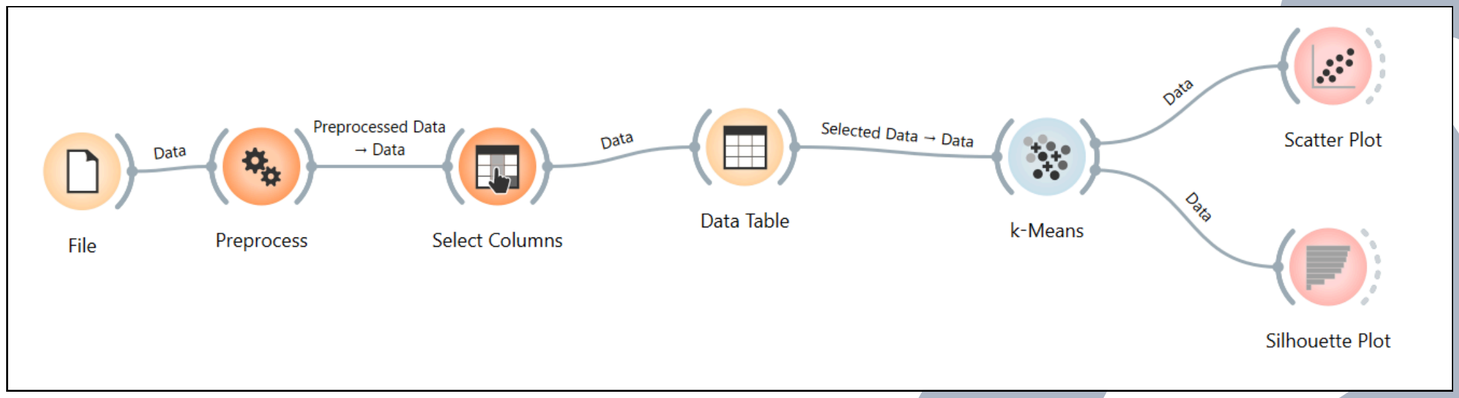
| Actual \ Predicted | No | Yes | Σ |
|---|---|---|---|
| No | 11 | 26 | 37 |
| Yes | 6 | 106 | 112 |
| Σ | 17 | 132 | 149 |



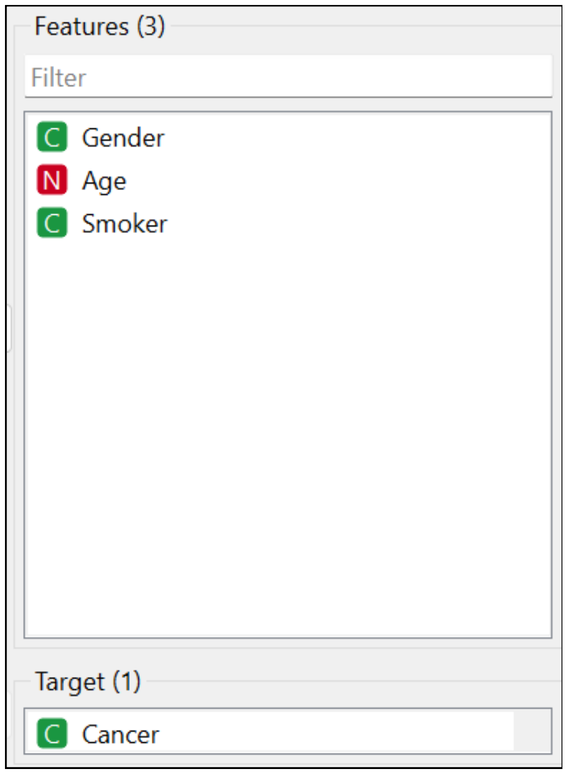

| Model | AUC | CA | F1 | Precision | Recall |
|---|---|---|---|---|---|
| Logistic Regression | 0.832 | 0.785 | 0.754 | 0.764 | 0.785 |

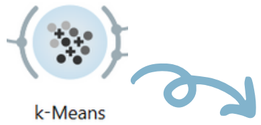
**Phan Thị Thanh Nhàn**

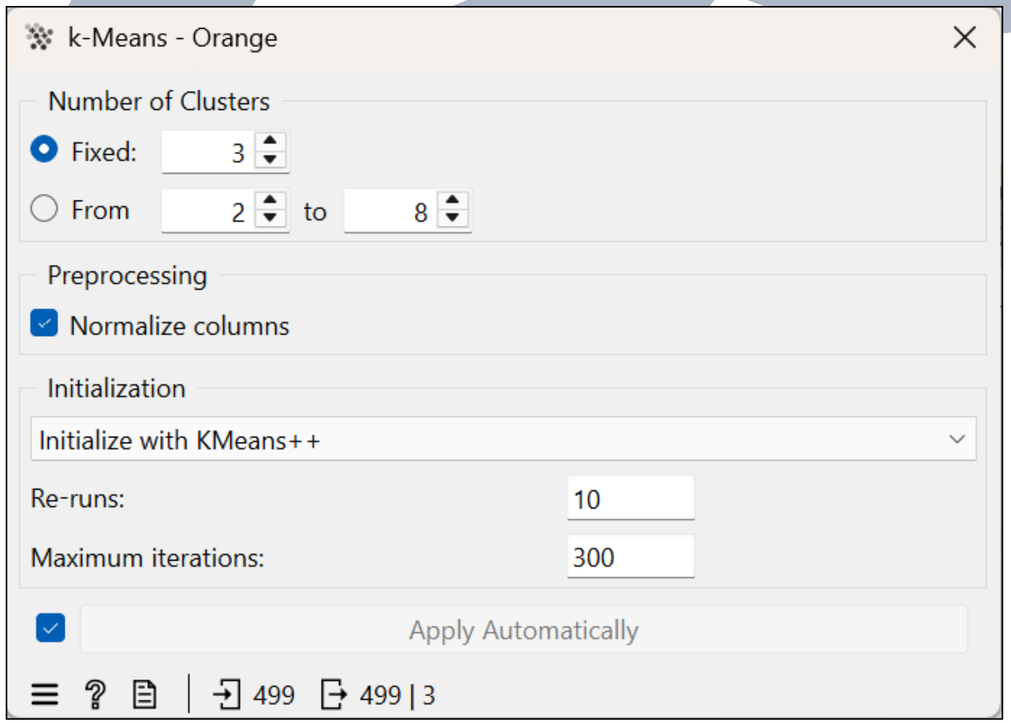
# 06 K - MEAN







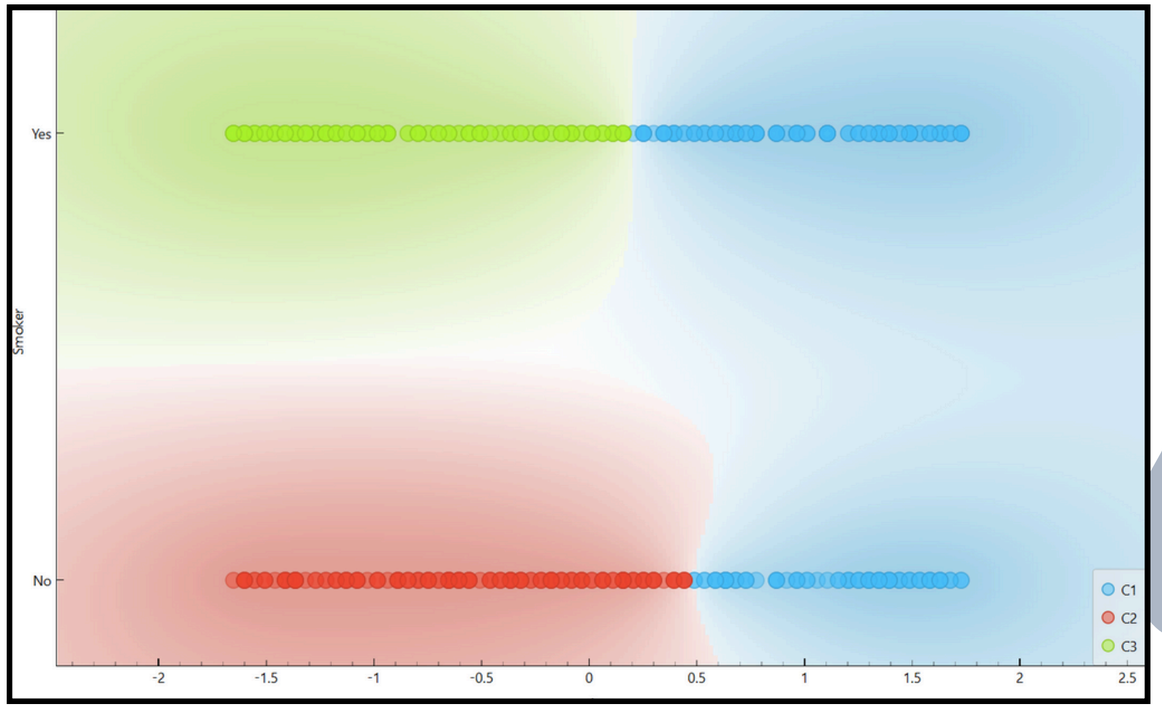



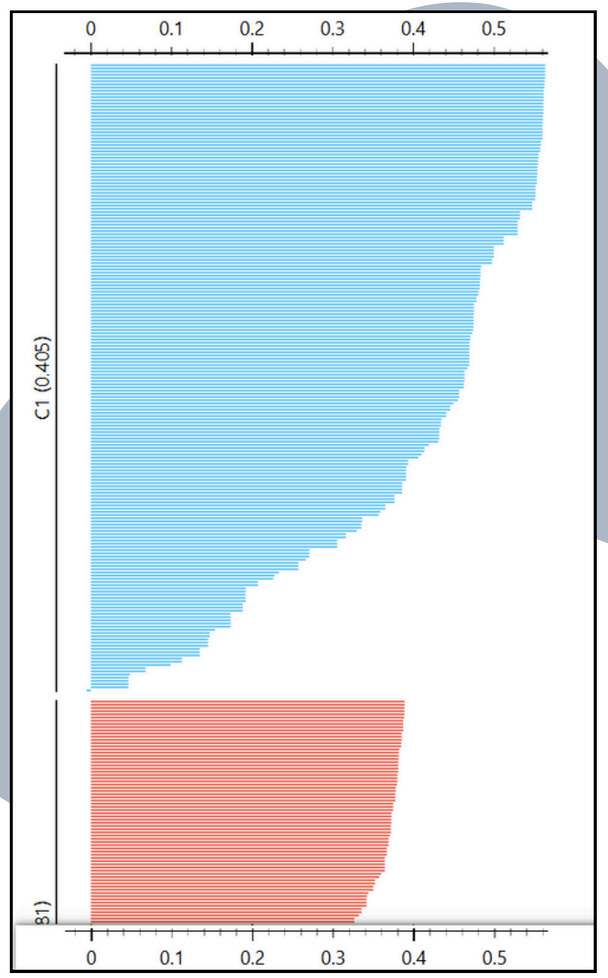

Phan Thị Thanh Nhàn

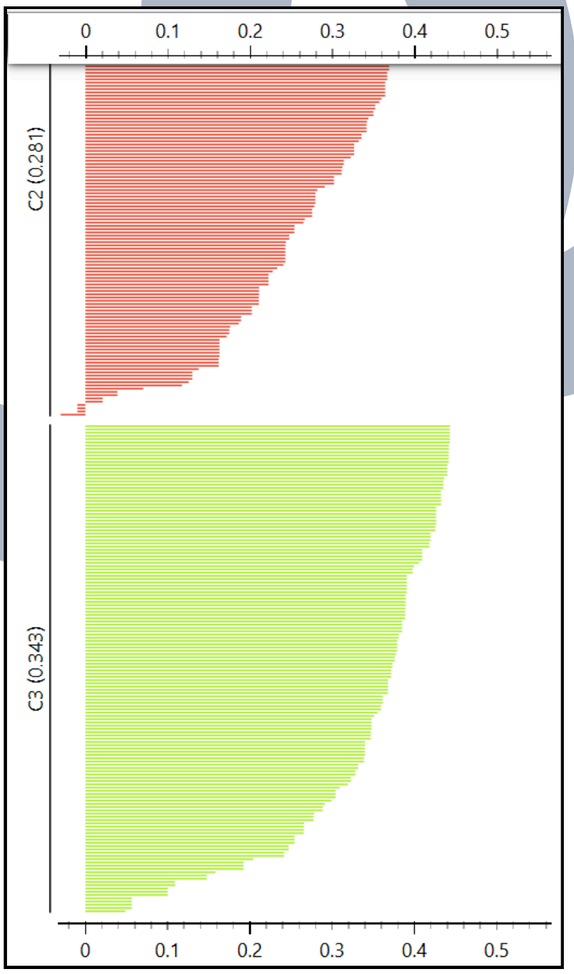
# 06 K - MEAN

Scatter Plot



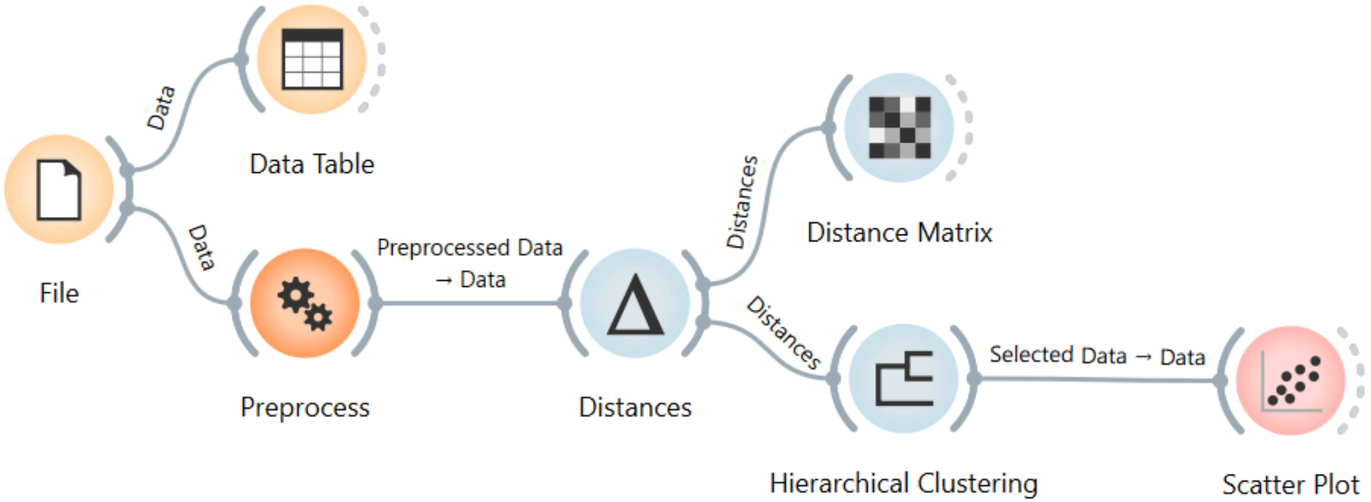

Silhouette Plot



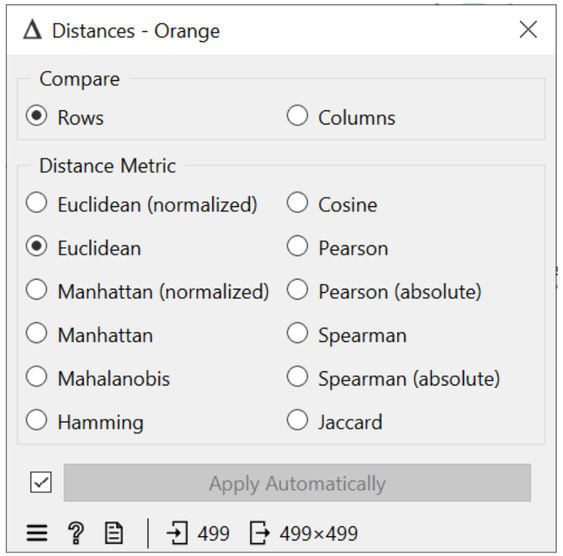




**Phan Thị Thanh Nhàn**

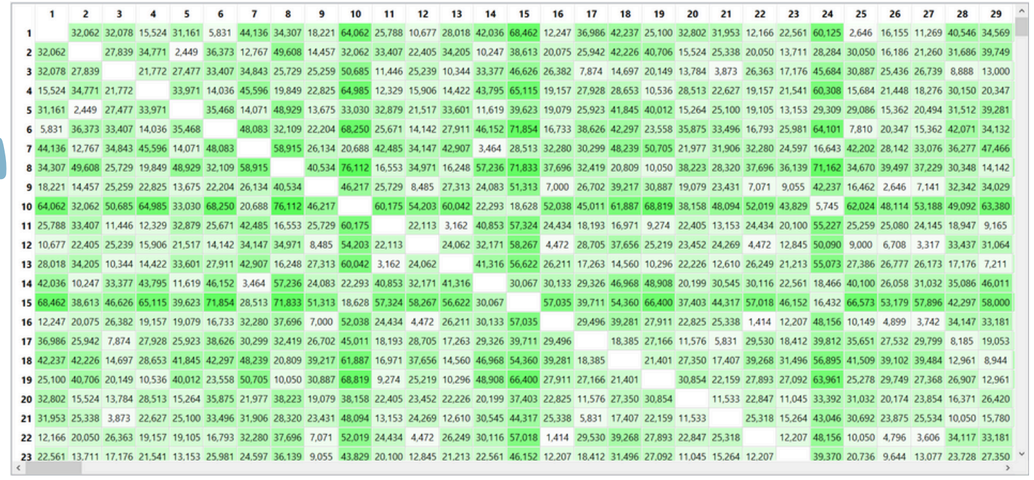
# 07 HIERARCHICAL





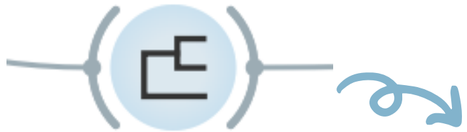




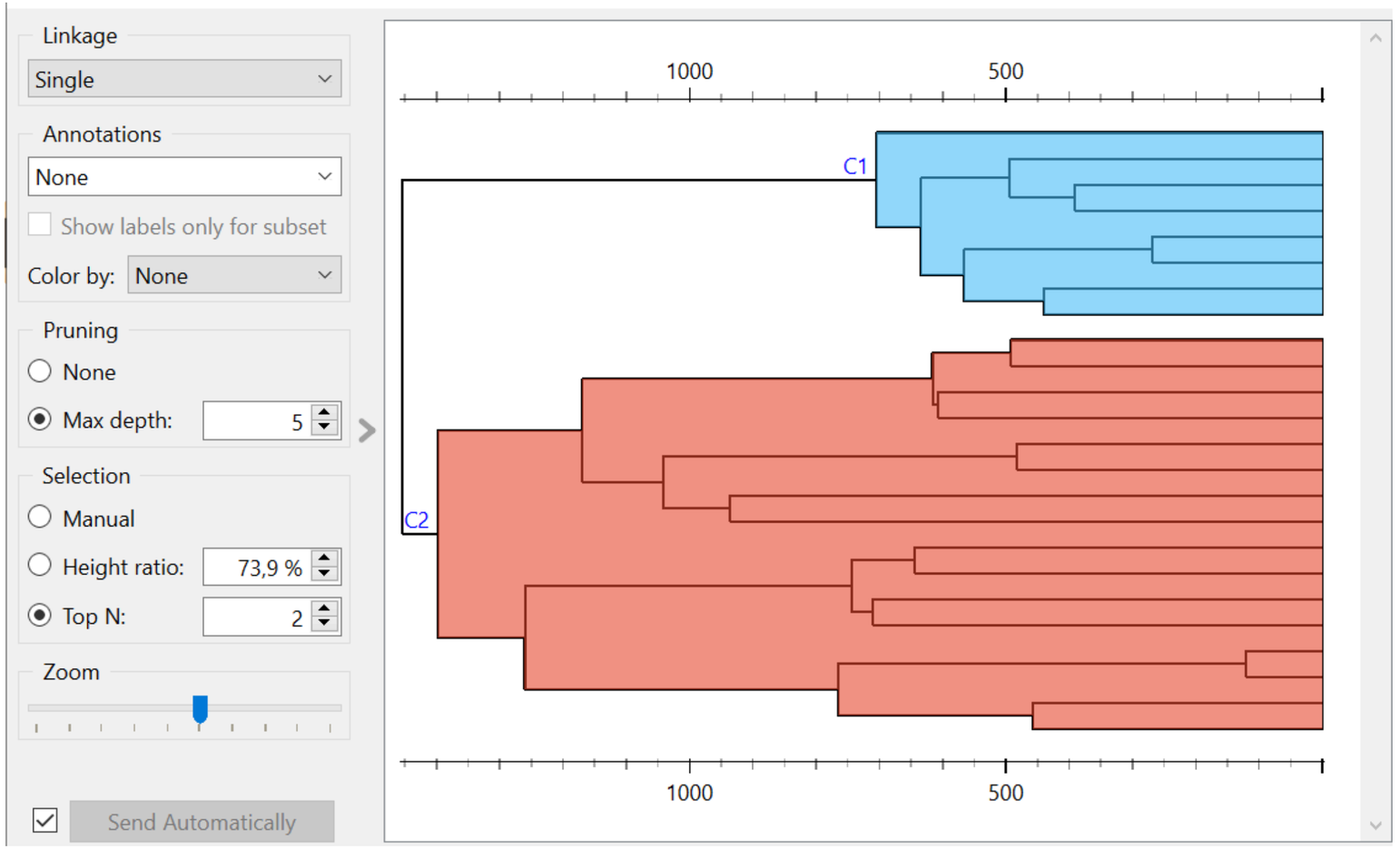


**Nguyễn Ngọc Đan Trâm**

# 07 HIERARCHICAL

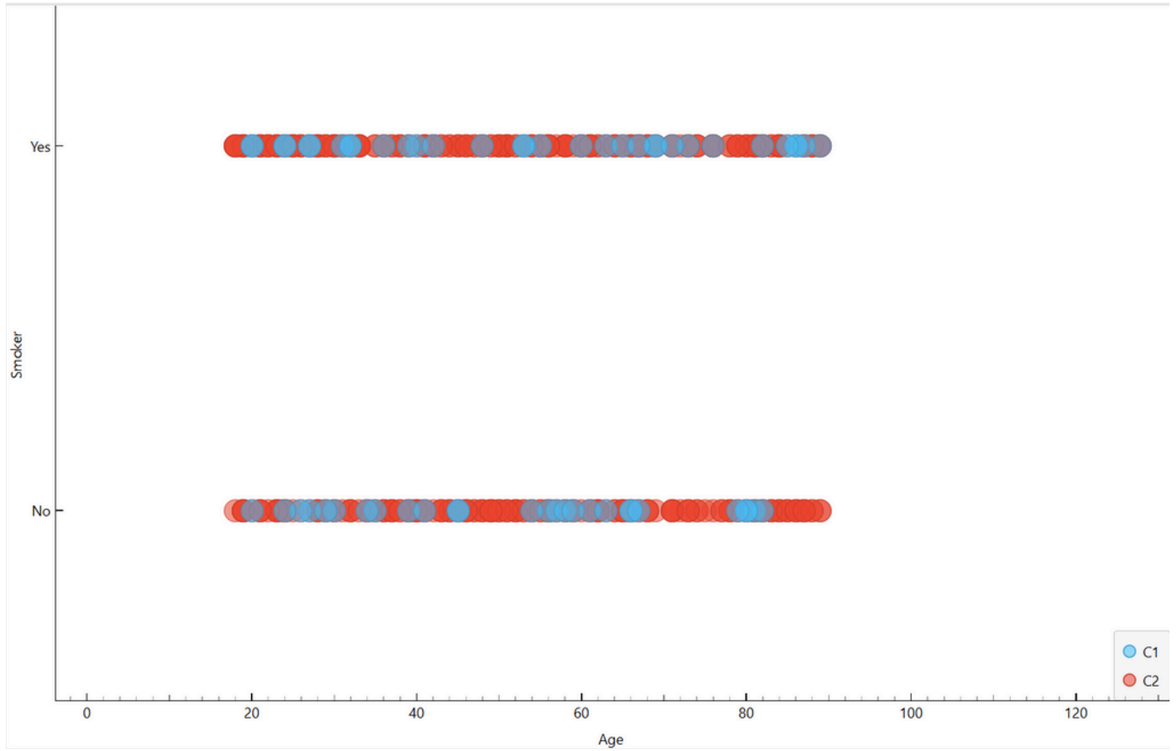


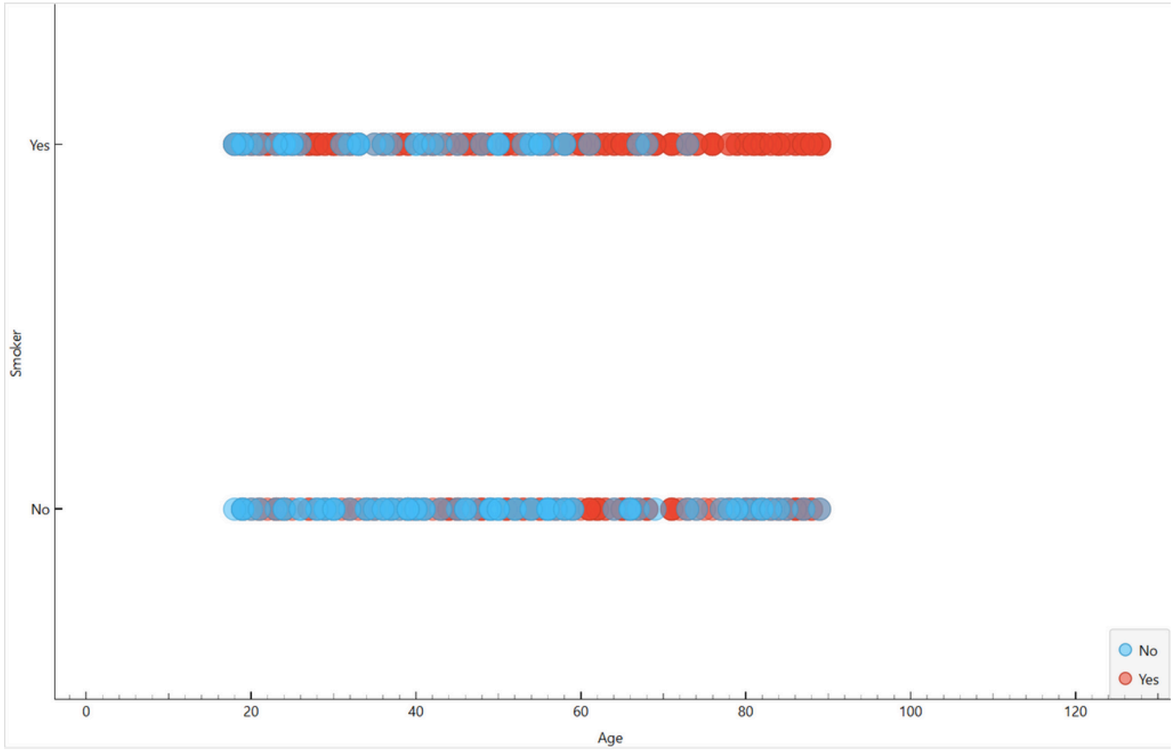
Hierarchical Clustering




Nguyễn Ngọc Đan Trâm

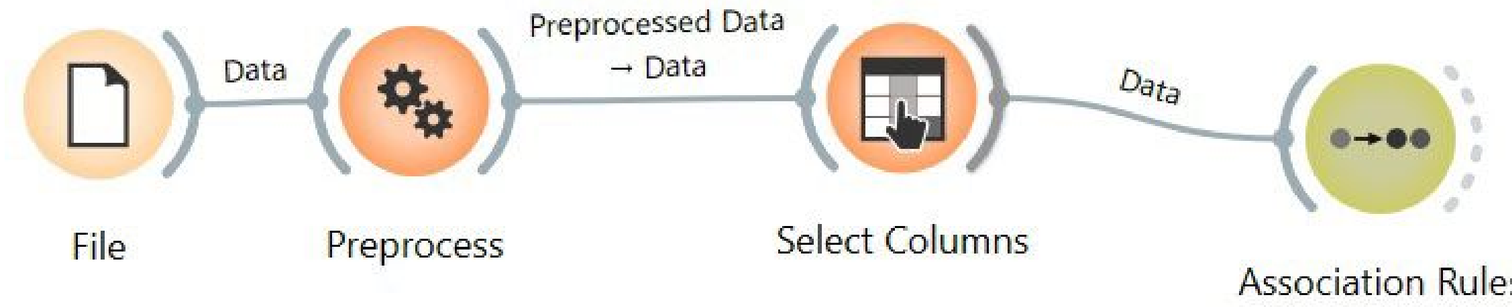
# 07 HIERARCHICAL




Nguyễn Ngọc Đan Trâm

# 08 ASOCIATION RULE




Find association rules

Min. supp.: 30 %

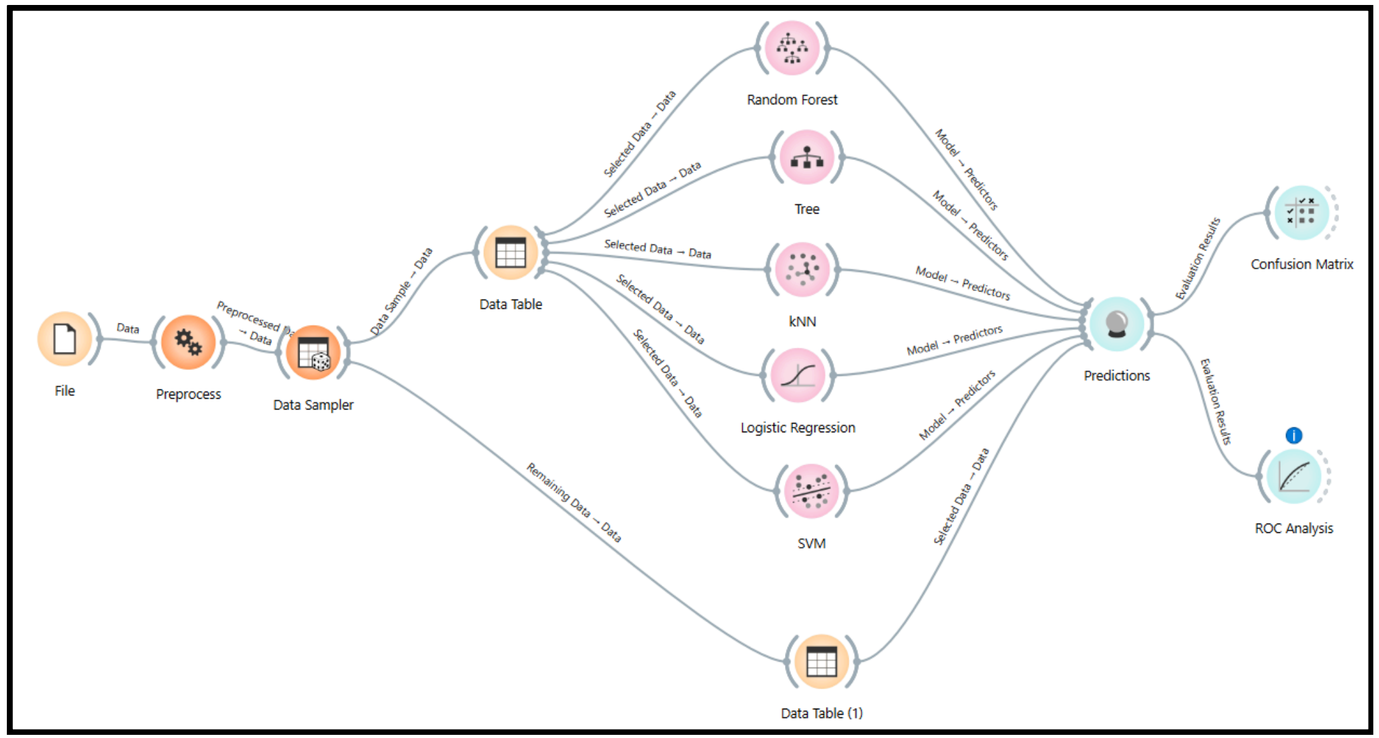
Min. conf.: 60 %

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 |

**Nguyễn Ngọc Đan Trâm**

# 09 MODEL EVALUATION

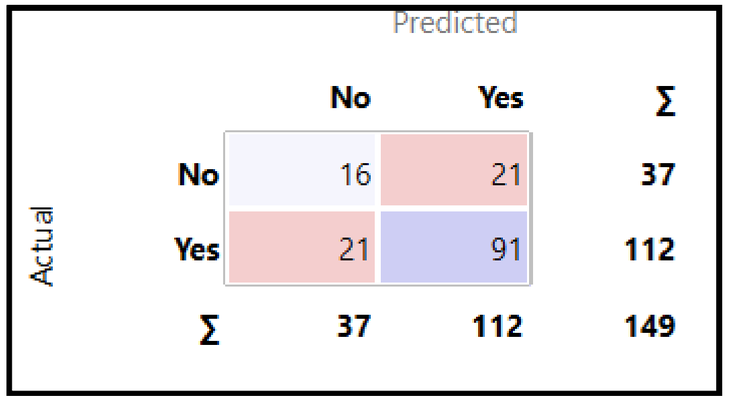



| Model | AUC | CA | F1 | Prec | Recall | MCC |
|---|---|---|---|---|---|---|
| SVM | 0.760 | 0.779 | 0.727 | 0.763 | 0.779 | 0.280 |
| Logistic Regression | 0.803 | 0.799 | 0.770 | 0.785 | 0.799 | 0.380 |
| kNN | 0.680 | 0.725 | 0.703 | 0.693 | 0.725 | 0.171 |
| Tree | 0.532 | 0.705 | 0.705 | 0.705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

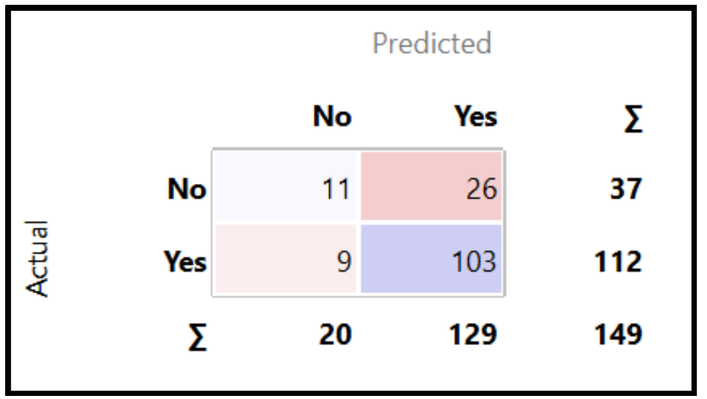
**Lâm Hồng Phúc**

# 09 MODEL EVALUATION

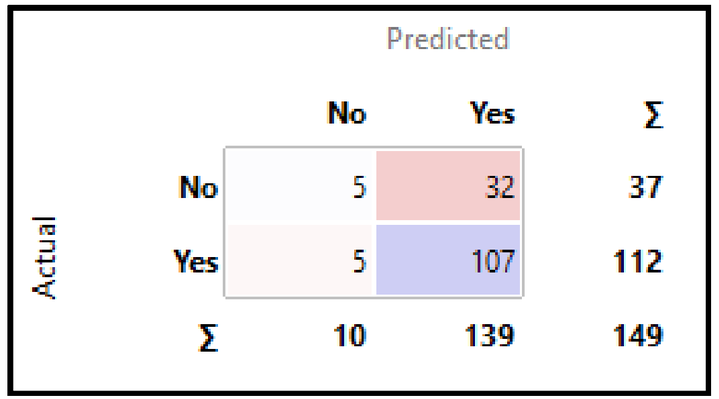
## Decision tree



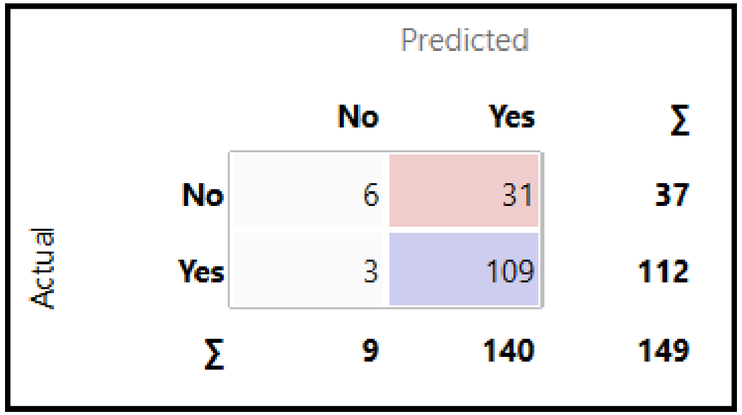

## Random forest



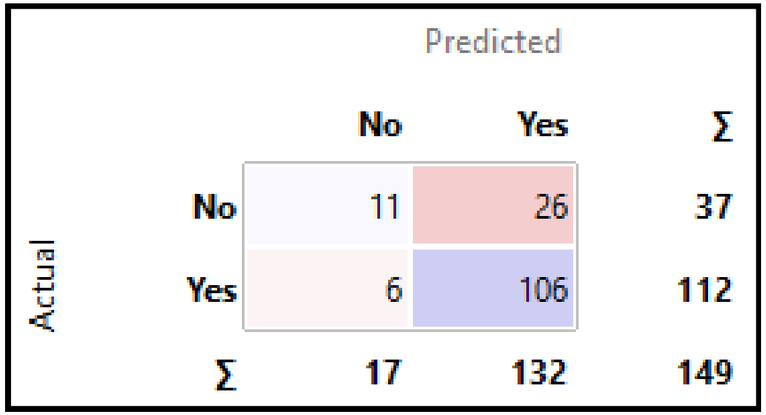

## KNN




## SVM




## Logistic Regression




**Lâm Hồng Phúc**

Nhóm 12

Kho và khai phá dữ liệu

# Thank You

May 28, 2025