BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC KINH TẾ TP HỒ CHÍ MINH
KHOA CÔNG NGHỆ THÔNG TIN KINH DOANH

ĐỒ ÁN MÔN HỌC

KHOA HỌC DỮ LIỆU

Đề tài:

Ứng dụng công cụ Orange Data Mining phân tích xu hướng thay đổi công việc Data Scientists từ bộ dữ liệu khảo sát Job Change Of Data Scientists

Thành viên: LÝ TRẦN THIÊN KIM - 31201024493
ĐOÀN ANH THƯ - 31201024519
VÕ NGỌC DUNG - 31201020182

Giảng Viên: Trương Việt Phương

TP Hồ Chí Minh, ngày 8 tháng 10 năm 2022

MỤC LỤC

LỜI NÓI ĐẦU **5**
CHƯƠNG I. TỔNG QUAN ĐỀ TÀI **6**
1. Xác định đề tài **6**
2. Mục đích nghiên cứu **6**
3. Giới hạn phạm vi đề tài **6**
4. Phương pháp nghiên cứu **6**
5. Công cụ sử dụng **6**
CHƯƠNG II. THU THẬP VÀ TIỀN XỬ LÝ DỮ LIỆU NGHIÊN CỨU **7**
1. Tổng quan bộ dữ liệu thu thập **7**
1.1. Giới thiệu bộ dữ liệu 7
1.2. Các thuộc tính chung 7
1.3. Mục tiêu khai thác dữ liệu hướng đến 9
1.4. Phương pháp khai thác dữ liệu sử dụng 9
2. Kiểm tra tình trạng dữ liệu gốc **10**
2.1. Tình trạng bộ dữ liệu gốc 10
2.2. Giải pháp xử lý 10
3. Tiền xử lý dữ liệu **11**
3.1. Lượt bớt dữ liệu 11
3.1.1. Data Sampler *11*
3.1.2. Rank *11*
3.2. Làm sạch dữ liệu 12
3.2.1. Xử lý dữ liệu bị thiếu *12*
CHƯƠNG III. KHAI THÁC DỮ LIỆU NGHIÊN CỨU **13**
1. Phân cụm dữ liệu **13**
1.1. Định nghĩa phân cụm dữ liệu 13
1.2. Mục đích phân cụm cho bộ dữ liệu Job change 13
1.3. Xây dựng mô hình phân cụm 13
1.3.1. Phân cụm phân hoạch *13*
1.3.2. Phân cụm phân cấp *17*
1.4. Đánh giá các phương pháp phân cụm với bộ dữ liệu gốc 20
2. Phân lớp dữ liệu **21**

2.1. Định nghĩa phân lớp dữ liệu 21
2.2. Xây dựng mô hình phân lớp 21
2.2. Sử dụng mô hình phân lớp 24
3. Dự báo từ kết quả phân lớp dữ liệu đạt được. 27
3.1. Trích xuất 10% dữ liệu 27
3.2. Predictions 28
3.3. Các ước lượng và bảng ma trận nhầm lẫn 29
4. Phân tích rút ra từ dự báo mô hình 10% dữ liệu 32
CHƯƠNG V. HƯỚNG PHÁT TRIỂN ĐỀ TÀI 34
MÔ HÌNH DỮ LIỆU ORANGE DATA MINING 36
1. Tiền xử lý dữ liệu 36
2. Phân cụm bộ dữ liệu 36
3. Phân lớp bộ dữ liệu 36
4. Mô hình Orange Data Mining tổng thể 37
TÀI LIỆU THAM KHẢO 39

❖ ***Bảng thành viên nhóm 10***

Thành viên	Vai trò	Phân công
Lý Trần Thiên Kim MSSV: 31201024493	Nhóm trưởng	• Thu thập bộ dữ liệu • Làm Orange • Viết báo cáo Chương 1, Chương 2, Chương 5 • Chỉnh sửa báo cáo toàn bài
Võ Ngọc Dung MSSV: 31201020182	Nhóm viên	• Thu thập bộ dữ liệu • Làm Orange • Viết báo cáo Chương 4, Chương 5
Đoàn Anh Thư MSSV: 31201024519	Nhóm viên	• Thu thập bộ dữ liệu • Làm Orange • Viết báo cáo Chương 2, Chương 3

LỜI NÓI ĐẦU

Khoa học dữ liệu *(Data Science)* là một trong những nhóm ngành lớn và đang phát triển ngày càng mạnh mẽ hiện nay. Khoa học dữ liệu đóng vai trò rất quan trọng, nhất là đối với các doanh nghiệp trong lĩnh vực công nghệ và kinh tế. Nó giúp đưa ra các giải pháp tối ưu cũng như các quyết định chính xác nhờ việc xử lý khối lượng lớn dữ liệu. Thông qua khoa học dữ liệu, ta có thể khám phá được hành vi người tiêu dùng và dự báo những chiều hướng tương lai cũng như giải thích các tác động ảnh hưởng đến chủ thể tham gia nghiên cứu. Qua đó, đem lại lợi ích to lớn cho các lĩnh vực lấy dữ liệu làm nền tảng sử dụng.

Tuy nhiên, để trở thành nhà phân tích dữ liệu chuyên nghiệp, nhóm ngành này lại đòi hỏi những kỹ năng thách thức như: kỹ năng xử lý tốt vấn đề trong mọi tình huống, kỹ năng sử dụng công cụ phân tích thành thạo, kỹ năng thống kê và tính toán, sự phù hợp tính cách người làm nghề, sự ham học hỏi theo độ cấp tiến phát triển của công nghệ... Ta không phủ nhận độ "hot" của ngành Dữ liệu, nhưng khoảng cách giữa việc theo đuổi đam mê và sự phù hợp để gắn bó một hành trình dài với nghề là rất lớn. Vì thế, không thể tránh khỏi các tình huống nhân lực dừng công việc hiện tại để tìm kiếm công việc khác phù hợp hơn. Và điều này sẽ tác động đến các doanh nghiệp nếu như họ chẳng may tuyển dụng một nhân công không lành nghề hay một người có xu hướng đổi việc liên tục theo sở thích cá nhân.

Nhằm hạn chế tác động này, một công ty hoạt động trong lĩnh vực Dữ liệu lớn và Khoa học Dữ liệu đã thực hiện một cuộc khảo sát mang tên Job Change of Data Scientists (*tạm dịch: Xu hướng thay đổi công việc của các nhà khoa học dữ liệu*). Khảo sát nhằm mục đích phân loại các ứng viên nào tham gia thực sự muốn làm việc lâu dài cho công ty và ứng viên nào có xu hướng sẽ tìm việc làm mới. Khảo sát trên đã giúp doanh nghiệp dự đoán xu hướng gắn bó với vị trí tuyển dụng và giải thích được các yếu tố ảnh hưởng đến quyết định làm việc lâu dài, từ đó, doanh nghiệp giảm thiểu được chi phí, thời gian đào tạo nhân lực.

Ở báo cáo này, chúng tôi sẽ thu thập bộ Job Change of Data Scientists để khai thác và tiền xử lý dữ liệu gồm các số liệu ghi nhận của từng ứng viên. Từ đó, kiểm định tính chính xác về xu hướng thay đổi công việc bằng thông qua *phương pháp phân cụm* và *phân lớp* ứng dụng *công cụ Orange Data Mining*. Kết quả thu được sẽ giải thích cho các yếu tố ảnh hưởng đến quyết định gắn bó công việc và dự báo xác suất xu hướng làm việc là dài hạn hay tạm thời của các ứng viên. Đây cũng chính là lý do mà nhóm chúng tôi quan tâm và chọn đề tài "***Ứng dụng công cụ Orange Data Mining phân tích xu hướng thay đổi công việc Data Scientists từ bộ dữ liệu khảo sát Job Change of Data Scientists***" làm đề án cuối kỳ học phần Khoa Học Dữ Liệu khoa Công nghệ thông tin Kinh doanh trường Đại học Kinh tế (UEH).

Do điều kiện về mặt thời gian và nhận thức còn nhiều hạn chế nên đồ án không khỏi có phần sai sót, kính mong quý giảng viên nhiệt tình góp ý để nhóm chúng tôi có những tiếp thu tốt hơn trong môn học này. Qua bài đồ án, nhóm chúng tôi xin cảm ơn giảng viên phụ trách môn học Khoa Học Dữ Liệu, thầy Trương Việt Phương đã truyền đạt những kiến thức quý báu cho học viên bằng cả tấm lòng nhiệt tình và sự tận tâm của thầy.

CHƯƠNG I. TỔNG QUAN ĐỀ TÀI

1. Xác định đề tài

Từ các yếu tố liên quan đến khả năng ảnh hưởng quyết định gắn bó cùng công việc khoa học dữ liệu lâu dài (được xem là thang đo): giới tính, trình độ học vấn, thâm niên làm việc, số lần chuyển công việc,… Chúng tôi sẽ tiến hành khai thác và xử lý dữ liệu thông qua công cụ Orange Data Mining. Kết quả thu được sẽ cho thấy xác suất các ứng viên có nhu cầu làm việc lâu dài và các ứng viên có xu hướng tìm kiếm công việc khác (không có ý định làm việc lâu dài). Việc sử dụng Orange Data Mining sẽ giải thích các yếu tố ảnh hưởng đến quyết định gắn bó công việc, dự báo tiềm năng một ứng viên có thực sự mong muốn sẽ đồng hành bền vững cùng vị trí nhà khoa học dữ liệu công ty đang tuyển dụng hay không.

Khi thực hiện việc xử lý dữ liệu ứng dụng công cụ Orange, chúng ta sẽ thu được các kết quả trực quan giúp ích cho việc tuyển dụng được những nhân lực chất lượng cho doanh nghiệp, từ đó, hạn chế tác động nhảy việc, giảm thiểu nhân lực đột ngột bởi những ứng viên chỉ mong muốn tìm kiếm một công việc tạm thời, ứng viên không chịu được áp lực cao.

Hơn thế nữa, khi thực hiện khai thác bộ dữ liệu này, ta cũng sẽ hiểu rõ hơn nhu cầu làm việc dài hạn của một ứng viên: nếu xác suất cho thấy số lượng người có nhu cầu tìm việc mới cao hơn số lượng người không muốn tìm việc khác, ta cần xem xét lại các hoạt động dành cho vị trí ứng tuyển này, đồng thời đề ra những hướng giải quyết cũng như giải pháp nâng cao chất lượng công việc hiện tại của doanh nghiệp.

2. Mục đích nghiên cứu

- *Dự đoán xác suất các ứng viên tham gia sẽ làm việc cho công ty lâu dài và ứng viên có xu hướng tìm kiếm công việc khác*
- *Phân tích các yếu tố ảnh hưởng đến quyết định làm việc dài hạn của ứng viên khi ứng tuyển vị trí nhà khoa học dữ liệu*

3. Giới hạn phạm vi đề tài

Ứng viên tham gia ứng tuyển vị trí nhà khoa học dữ liệu đến từ nhiều thành phố khác nhau, với đa dạng: độ tuổi, thâm niên làm việc, trình độ,…

4. Phương pháp nghiên cứu

Sử dụng phương pháp phân cụm và phân lớp cho các giá trị thu được từ bộ dữ liệu Job Change Of Data Scientists

5. Công cụ sử dụng

Orange Data Mining là công cụ sử dụng chính trong đề tài này. Cụ thể, Orange Data Ming là một công cụ khai thác dữ liệu đa nền tảng. Công cụ có tính năng lập trình trực quan như là phương tiện trực quan kết hợp phân tích dữ liệu và phương pháp trực quan. Lập trình trực quan cho phép người dùng không phải là lập trình viên quản lý có thể tiền xử lý, khám phá và mô hình hóa dữ liệu đưa vào.

CHƯƠNG II. THU THẬP VÀ TIỀN XỬ LÝ DỮ LIỆU NGHIÊN CỨU

1. Tổng quan bộ dữ liệu thu thập

1.1. Giới thiệu bộ dữ liệu

Bộ dữ liệu **Job Changes Of Data Scientists** được thiết kế để tìm hiểu về các yếu tố dẫn đến ứng viên Data Scientists quyết định rời bỏ công việc hiện tại của mình. Bằng cách sử dụng các thông tin liên quan đến nhân khẩu học, kinh nghiệm làm việc và chức danh hiện tại để dự đoán xác suất các ứng viên sẽ tìm các công việc mới hay tiếp tục làm việc cho công ty hiện tại, đồng thời giải thích các yếu tố ảnh hưởng đến quyết định này.

1.2. Các thuộc tính chung

Tên thuộc tính	Mô tả	Ghi Chú
ID	Mỗi ứng viên khảo sát sẽ được cấp duy nhất một ID.	Từ 3 đến 334000.
City	Mã thành phố đang sinh sống.	City_103, City_21, City_16, City_114, City_160, City_36, City_165, City_116, City_28, City_149, City_27, City_175, City_57, City_74, City_160, City_24, City_138, City_126, City_136, City_89, City_143, City_44, City_173, City_162.
City-Development index	Chỉ số phát triển của thành phố đang sinh sống.	Các chỉ số phát triển: • Từ 0.45 đến 0.95
Gender	Giới tính.	Giới tính thường gặp • Nam (Male): 69% • Nữ (Female): 6% • Khác (Other): 1% • Rỗng (Null): 24%
Relevant-Experience	Các kinh nghiệm liên quan đến vị trí Data Scientist	• Có kinh nghiệm liên quan (Has relevant experience): 72% • Không có kinh nghiệm liên quan (No relevant experience): 28%

Education Level	Trình độ học vấn.	Trình độ học vấn thường gặp • Trung học Phổ thông (High School): 14% • Tốt nghiệp (Graduate) 60% • Thạc Sĩ (Master): 23% • Tiến sĩ (PhD): 3%
Enrolled-University	Khóa học đã tham gia ở Đại học (nếu có)	Các khóa học thường có • Không tham gia (No_enrollment): 71% • Khóa học bán thời gian (Part time course): 21% • Khoá học toàn thời gian (Full time course): 8%
Major-Discipline	Chuyên ngành	Các chuyên ngành thường có: • STEM: 76% • Humanities: 4% • Business Degree: 3% • Khác (Other): 2% • Rỗng (Null): 15%
Experience	Số năm kinh nghiệm làm việc.	Các số năm kinh nghiệm thường có • Từ 1 đến 9 năm: 50% • Từ 10 đến 19 năm: 32% • Hơn 20 năm kinh nghiệm: 18%
Company-Size	Số lượng nhân viên trong công ty hiện tại.	Số lượng nhân viên thường có • Từ 1 đến 49: 24% • từ 50 đến 99: 16% • từ 100-500: 15% • từ 10000 trở lên: 16% • Rỗng (Null): 29%

Company-Type	Loại hình công ty	Các loại hình công ty • Công ty trách nhiệm hữu hạn tư nhân (Pvt Ltd): 54% • Công ty nhà nước (Public sector): 7% • Tổ chức phi chính phủ (NGO): 5% • Công ty vừa khởi nghiệp (Early stage startup): 2% • Công ty khởi nghiệp được tài trợ vốn (Funded startup company): 2% • Rỗng (Null): 30%
Last new job	Sự khác biệt về số năm giữa công việc trước đây và công việc hiện tại.	Số năm kinh nghiệm khác biệt thường có • 1 năm: 42% • 2 năm: 20% • 3 năm: 20% • > 4 năm: 17% • Không có: 1%
Training hours	Số giờ đào tạo đã hoàn thành.	Từ 1 đến 334 giờ
Target	Có thay đổi việc làm hoặc không.	• 1: Có • 0: Không

1.3. Mục tiêu khai thác dữ liệu hướng đến

- *Với sự tham gia khảo sát của các ứng viên thuộc ngành Khoa học dữ liệu và các dữ liệu thu thập được, giúp dự đoán khả năng các ứng viên tham gia sẽ có xu hướng làm việc lâu dài với công ty ứng tuyển hay có xu hướng tìm công việc khác (làm không lâu dài)*
- *So sánh độ chính xác từ mô hình dự báo so với kết quả khảo sát thực tế.*

1.4. Phương pháp khai thác dữ liệu sử dụng

Sử dụng phương pháp phân cụm và phân lớp thông bằng công cụ Orange Data Mining.

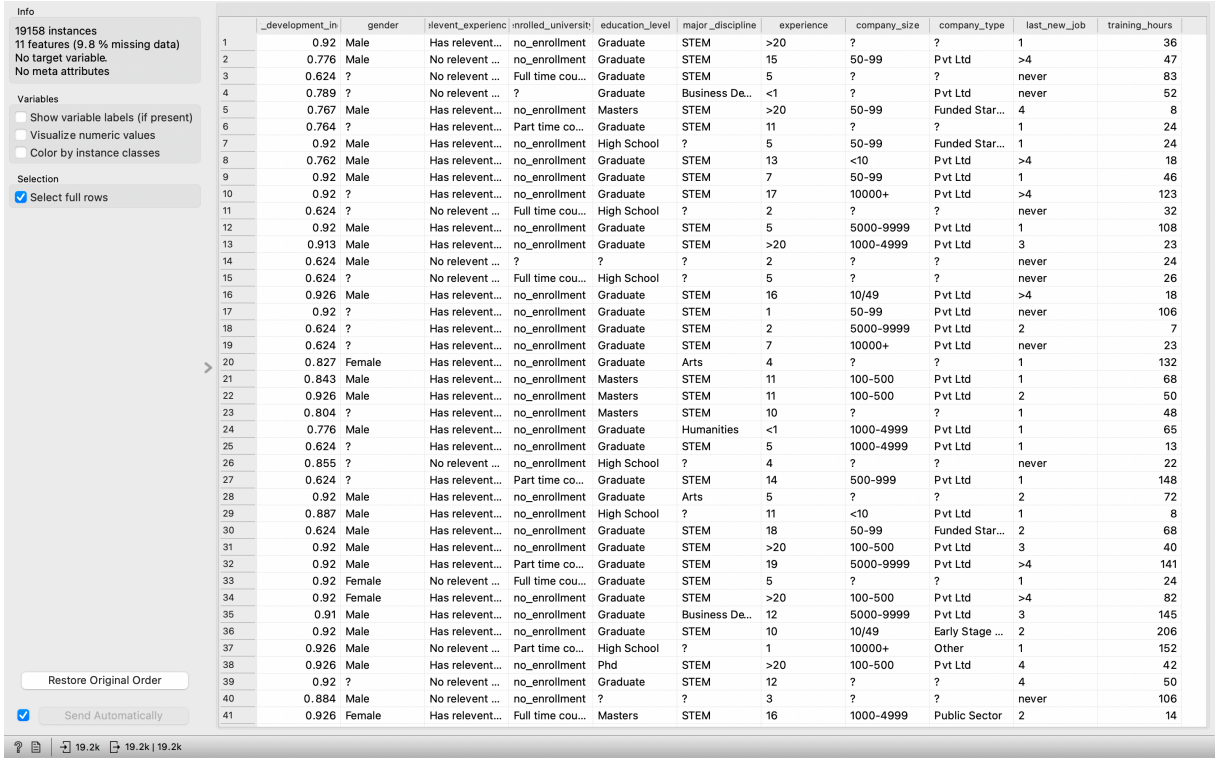
2. Kiểm tra tình trạng dữ liệu gốc

2.1. Tình trạng bộ dữ liệu gốc

Bộ dữ liệu chứa tập hợp dữ liệu lớn (19.158 dòng), khi đưa vào thuật toán K-means sẽ không thể tính toán được điểm Silhouette với những bộ dữ liệu có dữ liệu > 5000 dữ liệu và bên cạnh đó, ta không đủ RAM để tính toán ma trận khoảng cách. Đồng thời tập dữ liệu còn chứa dữ liệu bị thiếu (9.8%). Vấn đề xảy ra missing data khá phổ biến, do đó, ta cũng phải hiểu loại giá trị đang thiếu trong bộ dữ liệu để tìm cách xử lý với từng loại.

❖ ***Có 3 loại missing data ta thường gặp***

- Missing at Random (Dữ liệu khuyết ngẫu nhiên)
- Missing Completely at Random (Dữ liệu thiếu hoàn toàn ngẫu nhiên)
- Missing Not at Random (Dữ liệu khuyết không ngẫu nhiên)



	_development_in	gender	relevent_experienc	enrolled_universit	education_level	major _discipline	experience	company_size	company_type	last_new_job	training_hours
1	0.92	Male	Has relevent...	no_enrollment	Graduate	STEM	>20	?	?	1	36
2	0.776	Male	No relevent ...	no_enrollment	Graduate	STEM	15	50-99	Pvt Ltd	>4	47
3	0.624	?	No relevent ...	Full time cou...	Graduate	STEM	5	?	?	never	83
4	0.789	?	No relevent ...	?	Graduate	Business De...	<1	?	Pvt Ltd	never	52
5	0.767	Male	Has relevent...	no_enrollment	Masters	STEM	>20	50-99	Funded Star...	4	8
6	0.764	?	Has relevent...	Part time co...	Graduate	STEM	11	?	?	1	24
7	0.92	Male	Has relevent...	no_enrollment	High School	?	5	50-99	Funded Star...	1	24
8	0.762	Male	Has relevent...	no_enrollment	Graduate	STEM	13	<10	Pvt Ltd	>4	18
9	0.92	Male	Has relevent...	no_enrollment	Graduate	STEM	7	50-99	Pvt Ltd	1	46
10	0.92	?	Has relevent...	no_enrollment	Graduate	STEM	17	10000+	Pvt Ltd	>4	123
11	0.624	?	No relevent ...	Full time cou...	High School	?	2	?	?	never	32
12	0.92	Male	Has relevent...	no_enrollment	Graduate	STEM	5	5000-9999	Pvt Ltd	1	108
13	0.913	Male	Has relevent...	no_enrollment	Graduate	STEM	>20	1000-4999	Pvt Ltd	3	23
14	0.624	Male	No relevent ...	?	?	?	2	?	?	never	24
15	0.624	?	No relevent ...	Full time cou...	High School	?	5	?	?	never	26
16	0.926	Male	Has relevent...	no_enrollment	Graduate	STEM	16	10/49	Pvt Ltd	>4	18
17	0.92	?	Has relevent...	no_enrollment	Graduate	STEM	1	50-99	Pvt Ltd	never	106
18	0.624	?	Has relevent...	no_enrollment	Graduate	STEM	2	5000-9999	Pvt Ltd	2	7
19	0.624	?	Has relevent...	no_enrollment	Graduate	STEM	7	10000+	Pvt Ltd	never	23
20	0.827	Female	Has relevent...	no_enrollment	Graduate	Arts	4	?	?	1	132
21	0.843	Male	Has relevent...	no_enrollment	Masters	STEM	11	100-500	Pvt Ltd	1	68
22	0.926	Male	Has relevent...	no_enrollment	Masters	STEM	11	100-500	Pvt Ltd	2	50
23	0.804	?	Has relevent...	no_enrollment	Masters	STEM	10	?	?	1	48
24	0.776	Male	Has relevent...	no_enrollment	Graduate	Humanities	<1	1000-4999	Pvt Ltd	1	65
25	0.624	?	Has relevent...	no_enrollment	Graduate	STEM	5	1000-4999	Pvt Ltd	1	13
26	0.855	?	No relevent ...	no_enrollment	High School	?	4	?	?	never	22
27	0.624	?	Has relevent...	Part time co...	Graduate	STEM	14	500-999	Pvt Ltd	1	148
28	0.92	Male	Has relevent...	no_enrollment	Graduate	Arts	5	?	?	2	72
29	0.887	Male	Has relevent...	no_enrollment	High School	?	11	<10	Pvt Ltd	1	8
30	0.624	Male	Has relevent...	no_enrollment	Graduate	STEM	18	50-99	Funded Star...	2	68
31	0.92	Male	Has relevent...	no_enrollment	Graduate	STEM	>20	100-500	Pvt Ltd	3	40
32	0.92	Male	Has relevent...	Part time co...	Graduate	STEM	19	5000-9999	Pvt Ltd	>4	141
33	0.92	Female	No relevent ...	Full time cou...	Graduate	STEM	5	?	?	1	24
34	0.92	Female	Has relevent...	no_enrollment	Graduate	STEM	>20	100-500	Pvt Ltd	>4	82
35	0.91	Male	Has relevent...	no_enrollment	Graduate	Business De...	12	5000-9999	Pvt Ltd	3	145
36	0.92	Male	Has relevent...	no_enrollment	Graduate	STEM	10	10/49	Early Stage ...	2	206
37	0.926	Male	No relevent ...	Part time co...	High School	?	1	10000+	Other	1	152
38	0.926	Male	Has relevent...	no_enrollment	Phd	STEM	>20	100-500	Pvt Ltd	4	42
39	0.92	?	No relevent ...	no_enrollment	Graduate	STEM	12	?	?	4	50
40	0.884	Male	No relevent ...	no_enrollment	?	?	3	?	?	never	106
41	0.926	Female	Has relevent...	Full time cou...	Masters	STEM	16	1000-4999	Public Sector	2	14

2.2. Giải pháp xử lý

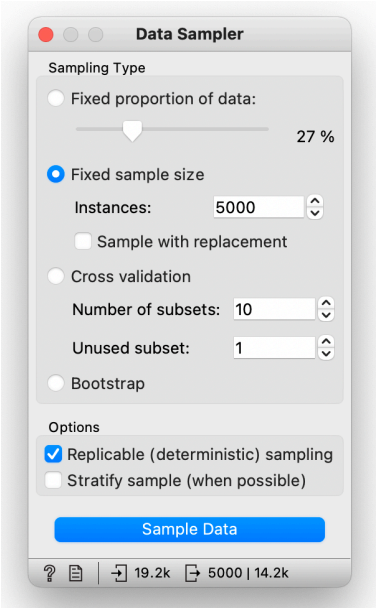
Ở đây missing value có số lượng lớn, chiếm hơn 8% tổng số quan sát trong 1 biến nhất định nên ta thay thế missing value bằng một giá trị khác thay vì loại bỏ hết missing value. Việc loại bỏ missing value sẽ tác động đến các biến phụ thuộc khác và làm thay đổi sự ảnh hưởng của chúng lên nhau ở Ranking phần Gain Ratio.

Đồng thời, để thực hiện tính ma trận khoảng cách ở phương pháp phân cụm phân cấp, cũng như khắc phục tình trạng quá tải dữ liệu khi thực hiện thao tác tính điểm silhouette ta sẽ giảm bớt dữ liệu về 5000 đối tượng

3. Tiền xử lý dữ liệu

3.1. Lượt bớt dữ liệu

3.1.1. Data Sampler



Ta thấy có hơn 19000 trường hợp trong dữ liệu. Như đã đề cập ở *Chương 2, phần 2 mục 2.2. Giải pháp xử lý* chúng ta trích ra 5000 trường hợp mẫu cố định bằng tiện ích **Data Sampler.** Sở dĩ chọn mẫu 5000 vì đây là số dữ liệu tối đa mà ta K-Means có thể xử lý được và thực hiện tính điểm số Silhouette ở thao tác phân cụm.

Bộ dữ liệu **sau khi lấy mẫu** gồm có 5000 dòng với 11 thuộc tính, trong đó có khoảng 9.8% dữ liệu bị thiếu. Việc sử dụng Data Sampler chỉ giải quyết được tình trạng bộ dữ liệu quá lớn, giúp giảm số lượng dữ liệu xuống, ngoài ra không xử lý vấn đề về Missing Values.

3.1.2. Rank

		#	Info.gain	Gain ratio	Gini
1	city_development_index		0.076	0.040	0.042
2	experience	22	0.026	0.006	0.014
3	company_size	8	0.026	0.011	0.014
4	enrolled_university	3	0.013	0.013	0.007
5	relevent_experience	2	0.011	0.013	0.006
6	education_level	5	0.007	0.005	0.003
7	last_new_job	6	0.005	0.002	0.003
8	company_type	6	0.003	0.003	0.002
9	major _discipline	6	0.001	0.002	0.001
10	gender	3	0.000	0.001	0.000
11	training_hours		0.000	0.000	0.000

Ta đưa bộ dữ liệu vào Rank để xếp hạng các biến theo mối tương quan của chúng lên biến mục tiêu dựa trên các bộ ghi điểm áp dụng ở đây là Information Gain, Information Gain Ratio, Gini Decrease. Ta chọn 9 thuộc tính có số điểm tốt nhất (9 biến có ảnh hưởng lên biến mục tiêu). Theo quan sát có biến city_development_index ảnh hưởng lên biến mục tiêu nhiều nhất và 2 biến còn lại gender và training_hourse có điểm bằng 0 nghĩa là không có sự ảnh hưởng lên biến mục tiêu.

3.2. Làm sạch dữ liệu

3.2.1. Xử lý dữ liệu bị thiếu

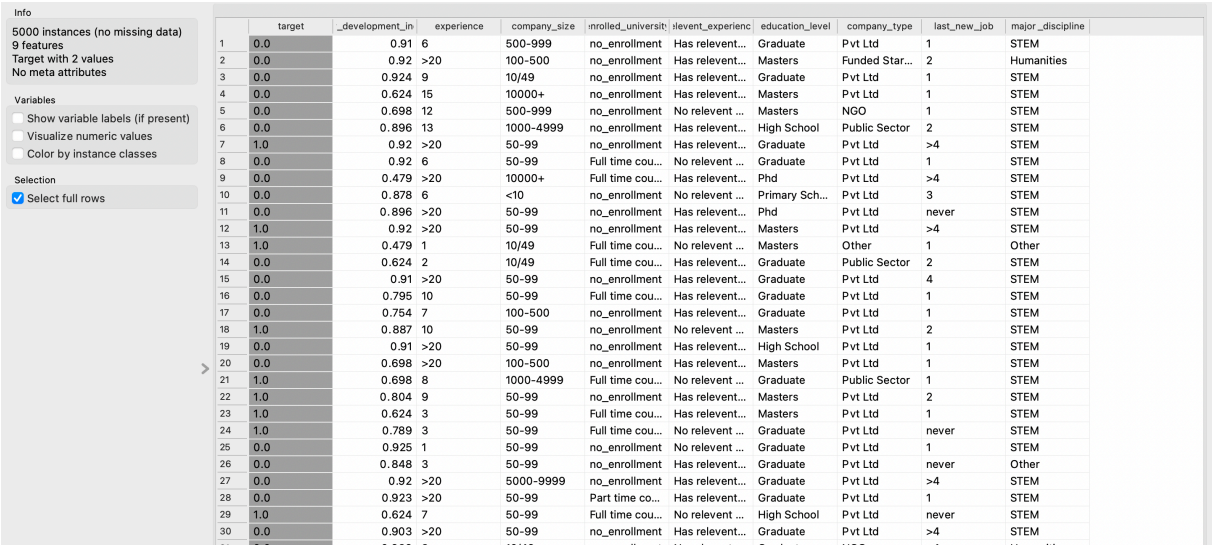
Ở đây biến có missing value là biến categorical: Ta chọn chức năng Impute Missing Values và chỉ định kỹ thuật áp dụng Average/Most frequent (Giá trị trung bình/ thường xuyên nhất) cho tất cả các dữ liệu của thuộc tính bị thiếu. Kết quả sau khi xử lý giá trị bị thiếu là dữ liệu sẽ tự thay vào giá trị xuất hiện thường xuyên nhất.

❖ ***So sánh bộ dữ liệu trước khi xử lý Missing Values***

- ***Trước khi xử lý***

	target	_development_in	gender	relevent_experienc	nrolled_university	education_level	major_discipline	experience	company_size	company_type	last_new_job	trai
1	1.0	0.92	Male	Has relevent...	no_enrollment	Graduate	STEM	>20	?	?	1	
2	0.0	0.776	Male	No relevent ...	no_enrollment	Graduate	STEM	15	50-99	Pvt Ltd	>4	
3	0.0	0.624	?	No relevent ...	Full time cou...	Graduate	STEM	5	?	?	never	
4	1.0	0.789	?	No relevent ...	?	Graduate	Business De...	<1	?	Pvt Ltd	never	
5	0.0	0.767	Male	Has relevent...	no_enrollment	Masters	STEM	>20	50-99	Funded Star...	4	
6	1.0	0.764	?	Has relevent...	Part time co...	Graduate	STEM	11	?	?	1	
7	0.0	0.92	Male	Has relevent...	no_enrollment	High School	?	5	50-99	Funded Star...	1	
8	1.0	0.762	Male	Has relevent...	no_enrollment	Graduate	STEM	13	<10	Pvt Ltd	>4	
9	1.0	0.92	Male	Has relevent...	no_enrollment	Graduate	STEM	7	50-99	Pvt Ltd	1	
10	0.0	0.92	?	Has relevent...	no_enrollment	Graduate	STEM	17	10000+	Pvt Ltd	>4	
11	1.0	0.624	?	No relevent ...	Full time cou...	High School	?	2	?	?	never	
12	0.0	0.92	Male	Has relevent...	no_enrollment	Graduate	STEM	5	5000-9999	Pvt Ltd	1	
13	0.0	0.913	Male	Has relevent...	no_enrollment	Graduate	STEM	>20	1000-4999	Pvt Ltd	3	
14	0.0	0.624	Male	No relevent ...	?	?	?	2	?	?	never	
15	0.0	0.624	?	No relevent ...	Full time cou...	High School	?	5	?	?	never	
16	0.0	0.926	Male	Has relevent...	no_enrollment	Graduate	STEM	16	10/49	Pvt Ltd	>4	
17	0.0	0.92	?	Has relevent...	no_enrollment	Graduate	STEM	1	50-99	Pvt Ltd	never	
18	0.0	0.624	?	Has relevent...	no_enrollment	Graduate	STEM	2	5000-9999	Pvt Ltd	2	

- ***Sau khi xử lý***



	target	_development_in	experience	company_size	nrolled_university	relevent_experienc	education_level	company_type	last_new_job	major_discipline
1	0.0	0.91	6	500-999	no_enrollment	Has relevent...	Graduate	Pvt Ltd	1	STEM
2	0.0	0.92	>20	100-500	no_enrollment	Has relevent...	Masters	Funded Star...	2	Humanities
3	0.0	0.924	9	10/49	no_enrollment	Has relevent...	Graduate	Pvt Ltd	1	STEM
4	0.0	0.624	15	10000+	no_enrollment	Has relevent...	Masters	Pvt Ltd	1	STEM
5	0.0	0.698	12	500-999	no_enrollment	No relevent ...	Masters	NGO	1	STEM
6	0.0	0.896	13	1000-4999	no_enrollment	Has relevent...	High School	Public Sector	2	STEM
7	1.0	0.92	>20	50-99	no_enrollment	Has relevent...	Graduate	Pvt Ltd	>4	STEM
8	0.0	0.92	6	50-99	Full time cou...	No relevent ...	Graduate	Pvt Ltd	1	STEM
9	0.0	0.479	>20	10000+	Full time cou...	Has relevent...	Phd	Pvt Ltd	>4	STEM
10	0.0	0.878	6	<10	no_enrollment	No relevent ...	Primary Sch...	Pvt Ltd	3	STEM
11	0.0	0.896	>20	50-99	no_enrollment	Has relevent...	Phd	Pvt Ltd	never	STEM
12	1.0	0.92	>20	50-99	no_enrollment	Has relevent...	Masters	Pvt Ltd	>4	STEM
13	1.0	0.479	1	10/49	Full time cou...	No relevent ...	Masters	Other	1	Other
14	0.0	0.624	2	10/49	Full time cou...	Has relevent...	Graduate	Public Sector	2	STEM
15	0.0	0.91	>20	50-99	no_enrollment	Has relevent...	Graduate	Pvt Ltd	4	STEM
16	0.0	0.795	10	50-99	Full time cou...	Has relevent...	Graduate	Pvt Ltd	1	STEM
17	0.0	0.754	7	100-500	no_enrollment	Has relevent...	Graduate	Pvt Ltd	1	STEM
18	1.0	0.887	10	50-99	no_enrollment	No relevent ...	Masters	Pvt Ltd	2	STEM
19	0.0	0.91	>20	50-99	no_enrollment	Has relevent...	High School	Pvt Ltd	1	STEM
20	0.0	0.698	>20	100-500	no_enrollment	Has relevent...	Masters	Pvt Ltd	1	STEM
21	1.0	0.698	8	1000-4999	Full time cou...	No relevent ...	Graduate	Public Sector	1	STEM
22	1.0	0.804	9	50-99	no_enrollment	Has relevent...	Masters	Pvt Ltd	2	STEM
23	1.0	0.624	3	50-99	Full time cou...	Has relevent...	Masters	Pvt Ltd	1	STEM
24	1.0	0.789	3	50-99	Full time cou...	No relevent ...	Graduate	Pvt Ltd	never	STEM
25	0.0	0.925	1	50-99	no_enrollment	No relevent ...	Graduate	Pvt Ltd	1	STEM
26	0.0	0.848	3	50-99	no_enrollment	Has relevent...	Graduate	Pvt Ltd	never	Other
27	0.0	0.92	>20	5000-9999	no_enrollment	Has relevent...	Graduate	Pvt Ltd	>4	STEM
28	0.0	0.923	>20	50-99	Part time co...	Has relevent...	Graduate	Pvt Ltd	1	STEM
29	1.0	0.624	7	50-99	Full time cou...	No relevent ...	High School	Pvt Ltd	never	STEM
30	0.0	0.903	>20	50-99	no_enrollment	Has relevent...	Graduate	Pvt Ltd	>4	STEM

Sau khi tiền xử lý, bộ dữ liệu của ta có 5000 trường hợp dữ liệu (không chứa missing values), 9 biến, 2 mục tiêu phân lớp target (0 và 1), 9 thuộc tính bổ trợ cho target bao gồm: city_development_index, company_size, company_type, education_level, enrolled_university, experience, gender, last_new_job, major_discipline

CHƯƠNG III. KHAI THÁC DỮ LIỆU NGHIÊN CỨU

1. Phân cụm dữ liệu

1.1. Định nghĩa phân cụm dữ liệu

Phân cụm thuộc nhóm phương pháp học không giám sát (unsupervised learning) vì không biết trước được số nhóm (khác với bài toán phân lớp). Quá trình này sẽ gom cụm các đối tượng có đặc điểm tương đồng vào các cụm tương ứng. Thuật toán sẽ tìm cách phân các đối tượng thảnh từng cụm có đặc điểm tương tự, đồng thời độ tương đồng giữa các cụm thấp (khác biệt cao). Hai phương pháp phân cụm sẽ được áp dụng đối với bộ dữ liệu trên bao gồm: phân cụm phân cấp và phân cụm phân hoạch.

Phân cụm dữ liệu nhằm mục đích chính là thành lập các nhóm dữ liệu từ tập dữ liệu lớn, theo đó, cho phép ta khai phá và tìm kiếm thông tin tiềm ẩn, hữu ích phục vụ cho ra quyết định. Đây là một phương pháp xử lý thông tin quan trọng và phổ biến, nó nhằm khám phá mối liên hệ giữa các mẫu dữ liệu bằng cách tổ chức chúng thành các cụm tương tự.

1.2. Mục đích phân cụm cho bộ dữ liệu Job change

Tương ứng với bộ dữ liệu Job Change Database, quá trình phân cụm diễn ra nhằm mục đích phân nhóm các ứng viên ngành Khoa học dữ liệu với 9 thuộc tính (được giữ lại sau quá trình tiền xử lý) vào trong cùng một cụm có những tính chất tương tự nhau, các đối tượng khác nhau sẽ thuộc các cụm khác nhau. Điều này giúp người sử dụng khám phá ra các nhóm thuộc tính nào của ứng viên dẫn đến quyết định rời bỏ hoặc tiếp tục công việc

1.3. Xây dựng mô hình phân cụm

Các kỹ thuật phân cụm có rất nhiều các tiếp cận và ứng dụng trong thực tế, nó hướng tới hai mục tiêu chung đó là chất lượng của các cụm khám phá được và tốc độ thực hiện của thuật toán. Hiện nay, các kỹ thuật phân cụm có thể phân loại theo nhiều cách tiếp cận. Ở đây chúng tôi kết hợp cả hai phương pháp Phân cụm phân hoạch (Partitioning Clusterting) và Phân cụm phân cấp (Hierarchical Clustering) vào bộ dữ liệu Job Change Database. Nghĩa là kết quả của dự liệu thu được sau phương pháp phân cấp sẽ tiếp tục cải tiến thông qua phân cụm phân hoạch

1.3.1. Phân cụm phân hoạch

Loại phân cụm này sẽ phân tập dữ liệu có n phần tử cho trước thành k tập con (k<=n), mỗi tập con biểu diễn một cụm. Số các cụm được thiết lập là các được trưng được lựa chon trước, sao cho: Mỗi đối tượng thuộc duy nhất 1 cụm, các phần tử trong cụm có sự tương tự nhau. Mỗi cụm có ít nhất 1 phần tử. Thuật toán điển hình:

- K-means
- K-mediod,
- Fuzzy C-means

Thuật toán phân cụm dựa trên phương pháp này giúp xây dựng các phân hoạch khác nhau và đánh giá chúng. Sau đó, tìm cách tối thiểu hóa tổng bình phương độ lỗi. Với bộ dữ liệu Job Change Database có dữ liệu lớn thì việc thực hiện phân cụm phân hoạch sẽ hỗ trợ xử lý với tốc độ nhanh chóng.

❖ *Bộ dữ liệu sử dụng phân cụm phân hoạch*

Info
5000 instance(s)
9 feature(s) (no missing values)
Classification; categorical class with 2 values (no missing values)
0 meta attribute(s)

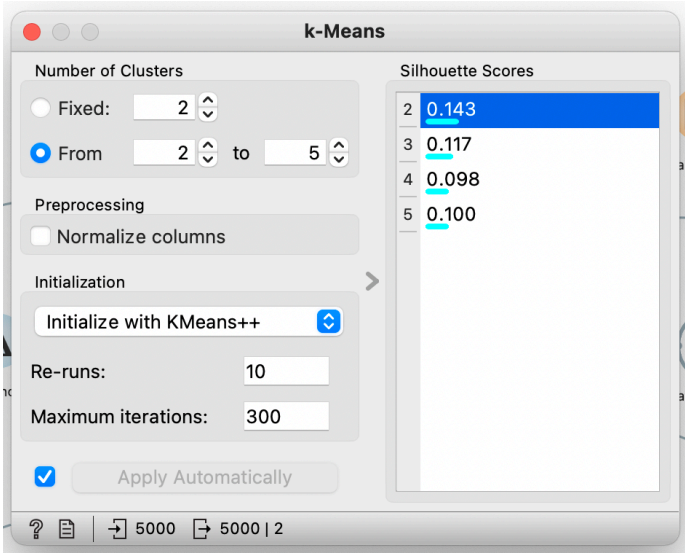
Columns (Double click to edit)

	Name	Type	Role	Values
1	city_develop...	numeric	feature	
2	experience	categorical	feature	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,...
3	company_size	categorical	feature	50-99,100-500,500-999,1000-4999,5000-9999,18172.00,<10,>10000
4	enrolled_uni...	categorical	feature	Full time course,Part time course,no_enrollment
5	relevent_exp...	categorical	feature	Has relevent experience,No relevent experience
6	education_le...	categorical	feature	Graduate,High School,Masters,Phd
7	company_ty...	categorical	feature	Early Stage Startup,Funded Startup,NGO,Other,Public Sector,Pvt Ltd
8	last_new_job	categorical	feature	0,1,2,3,4,>4
9	major_discip...	categorical	feature	Arts,Business Degree,Humanities,No Major,Other,STEM
10	target	categori...	skip	0,1

✳ *K-Means*

Về lý thuyết, tư tưởng của thuật toán là xem mỗi đối tượng trong tập dữ liệu là một điểm trong không gian d chiều (với d là số lượng thuộc tính của đối tượng). Phương pháp này dựa trên độ đo khoảng cách của các đối tượng trong cụm đó. Thuật toán này sẽ dừng cho tới khi tất cả điểm dữ liệu phân về đúng cụm đồng thời số lượt cập nhật tâm đạt tối đa. Ta có thể tóm tắt các bước của thuật toán như sau:

- **Bước 1:** Khởi tạo ngẫu nhiên k tâm cụm
- **Bước 2:** Lặp lại quá trình cập nhật tâm cụm cho tới khi dừng

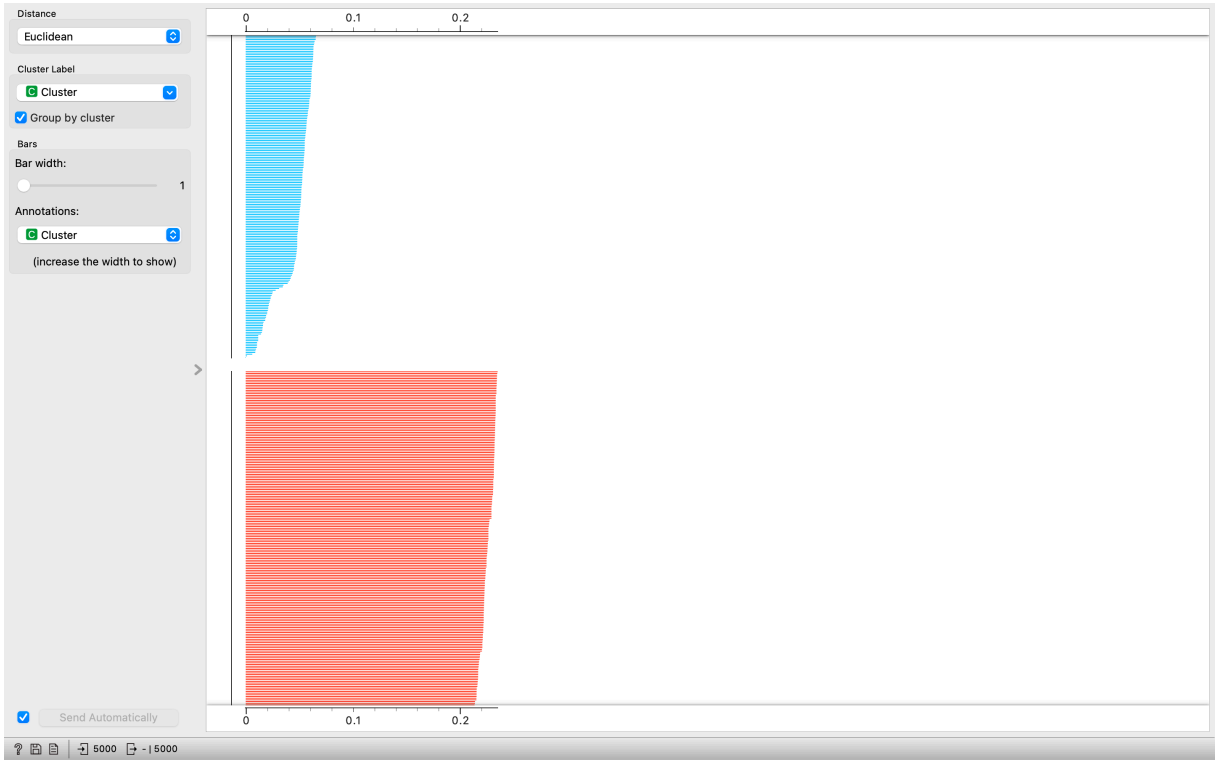


Cụ thể, ta dùng đến tiện ích **k-means** để nhóm dữ liệu thành nhiều nhóm. Vì ta chưa biết trước số cụm là bao nhiêu nên sẽ tiến hành chạy từ 2 đến 5 cụm. Ta chọn cách thuật toán bắt đầu phân cụm bằng phương pháp **Initialize with K-Means** + + để tâm đầu tiên được chọn ngẫu nhiên, các điểm tiếp theo được chọn từ các điểm còn lại với xác suất tương ứng với khoảng cách bình phương từ tâm gần nhất. Tiện ích sẽ hiển thị điểm phân nhóm cho phạm vi cụm đã chọn, bằng thang điểm Silhouette (đối chiếu khoảng cách trung bình đến các phần tử trong cùng 1 cụm với khoảng cách trung bình đến các các phần tử trong cụm khác). Đây là bảng điểm Silhouette Score của bộ dữ liệu trên, ta quan sát thấy với số cụm = 2 có số điểm cao nhất 0.143 (phân nhóm tốt nhất). Vì vậy ta tiến hành **phân thành hai cụm.**

✳ Silhouette Plot

Ta chạy k-means với hai cụm và hiển thị trong biểu đồ **Silhouette**. Để ước tính chất lượng cụm của mình, biểu đồ cho ta quan sát được các điểm bất thường và ngoại lai thông qua biểu diễn mức độ gần của mỗi điểm trong một cụm với các điểm trong các cụm lân cận. Có 3 phương pháp được đề xuất:

- **Euclidean:** khoảng cách "đường thẳng" giữa hai điểm
- **Manhattan:** tổng của sự khác biệt tuyệt đối cho tất cả các thuộc tính
- **Cosin:** 1 - cosin của góc giữa hai vectơ

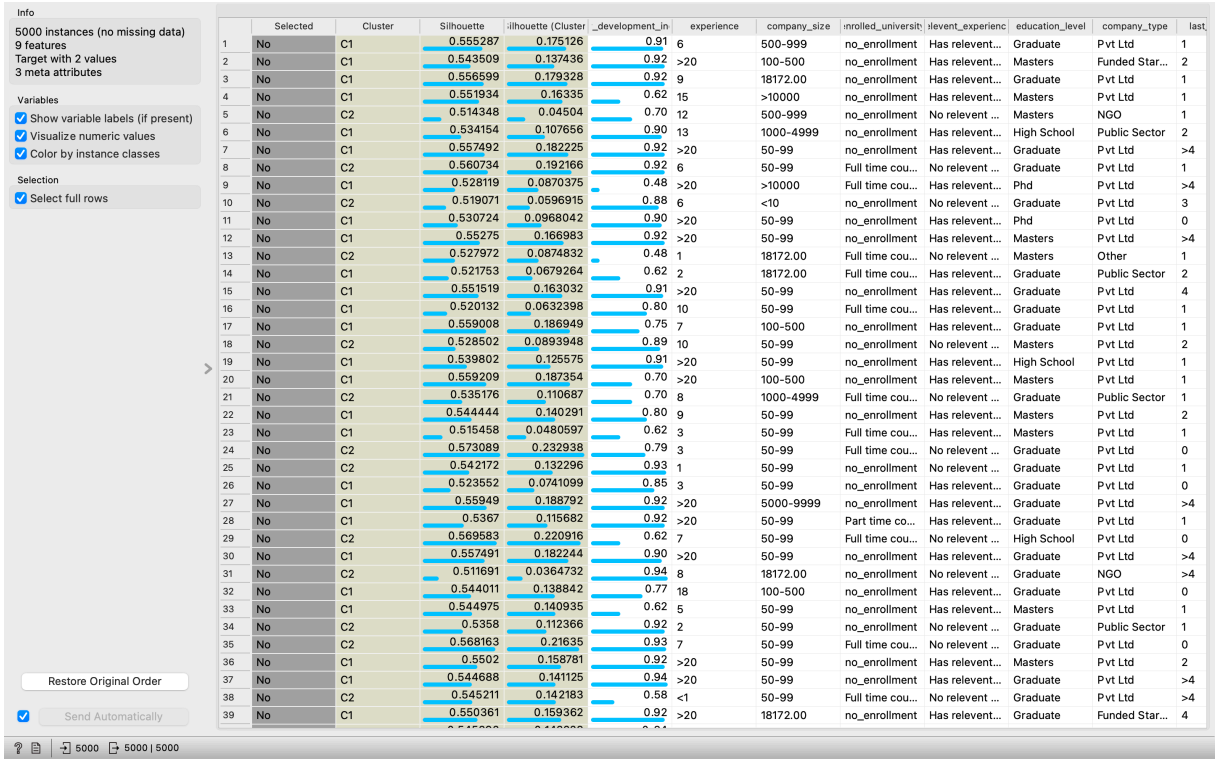


Biểu đồ hiển thị điểm cho các trường hợp dữ liệu riêng lẻ. Với số điểm cao, có giá trị dương đại diện cho các trường hợp có tính đại diện cao cho các cụm, trong khi điểm âm đại diện cho các trường hợp ngoại lệ (không phù hợp với cụm).

Ta thu được biểu đồ Silhouette cho ra 2 cụm cho tổng 5000 đối tượng: Tương ứng với màu xanh biểu diễn cho **3570 đối tượng thuộc C1 (Có muốn thay đổi công việc)**, màu đỏ biểu diễn cho **1430 đối tượng thuộc C2 (Không muốn thay đổi công việc)**. Khi chọn phương pháp tính khoảng cách "đường thẳng" giữa hai điểm thì **không có giá trị nào âm**, nghĩa là không tồn tại đối tượng bị gán sai cụm

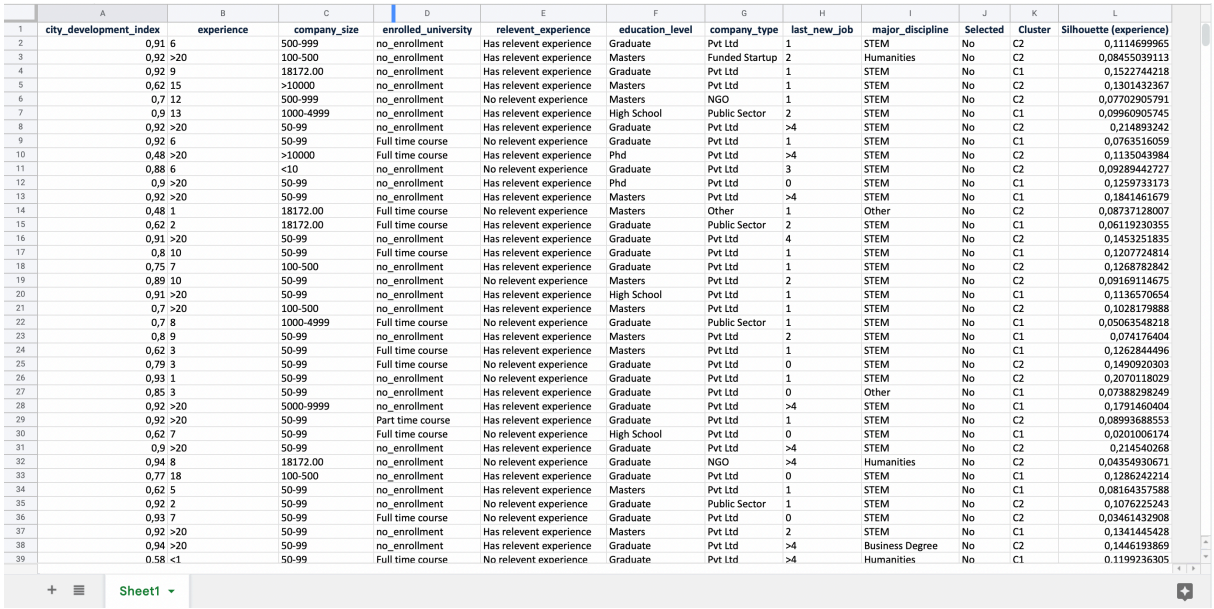
❖ *Dữ liệu sau khi thực hiện phân cụm phân hoạch*

- ***Quan sát từ Orange***



	Selected	Cluster	Silhouette	ilhouette (Cluster	_development_in	experience	company_size	nrolled_universit	elevent_experienc	education_level	company_type	last
1	No	C1	0.555287	0.175126	0.91	6	500-999	no_enrollment	Has relevent...	Graduate	Pvt Ltd	1
2	No	C1	0.543509	0.137436	0.92	>20	100-500	no_enrollment	Has relevent...	Masters	Funded Star...	2
3	No	C1	0.556599	0.179328	0.92	9	18172.00	no_enrollment	Has relevent...	Graduate	Pvt Ltd	1
4	No	C1	0.551934	0.16335	0.62	15	>10000	no_enrollment	Has relevent...	Masters	Pvt Ltd	1
5	No	C2	0.514348	0.04504	0.70	12	500-999	no_enrollment	No relevent ...	Masters	NGO	1
6	No	C1	0.534154	0.107656	0.90	13	1000-4999	no_enrollment	Has relevent...	High School	Public Sector	2
7	No	C1	0.557492	0.182225	0.92	>20	50-99	no_enrollment	Has relevent...	Graduate	Pvt Ltd	>4
8	No	C2	0.560734	0.192166	0.92	6	50-99	Full time cou...	No relevent ...	Graduate	Pvt Ltd	1
9	No	C1	0.528119	0.0870375	0.48	>20	>10000	Full time cou...	Has relevent...	Phd	Pvt Ltd	>4
10	No	C2	0.519071	0.0596915	0.88	6	<10	no_enrollment	No relevent ...	Graduate	Pvt Ltd	3
11	No	C1	0.530724	0.0968042	0.90	>20	50-99	no_enrollment	Has relevent...	Phd	Pvt Ltd	0
12	No	C1	0.55275	0.166983	0.92	>20	50-99	no_enrollment	Has relevent...	Masters	Pvt Ltd	>4
13	No	C2	0.527972	0.0874832	0.48	1	18172.00	Full time cou...	No relevent ...	Masters	Other	1
14	No	C1	0.521753	0.0679264	0.62	2	18172.00	Full time cou...	Has relevent...	Graduate	Public Sector	2
15	No	C1	0.551519	0.163032	0.91	>20	50-99	no_enrollment	Has relevent...	Graduate	Pvt Ltd	4
16	No	C1	0.520132	0.0632398	0.80	10	50-99	Full time cou...	Has relevent...	Graduate	Pvt Ltd	1
17	No	C1	0.559008	0.186949	0.75	7	100-500	no_enrollment	Has relevent...	Graduate	Pvt Ltd	1
18	No	C2	0.528502	0.0893948	0.89	10	50-99	no_enrollment	No relevent ...	Masters	Pvt Ltd	2
19	No	C1	0.539802	0.125575	0.91	>20	50-99	no_enrollment	Has relevent...	High School	Pvt Ltd	1
20	No	C1	0.559209	0.187354	0.70	>20	100-500	no_enrollment	Has relevent...	Masters	Pvt Ltd	1
21	No	C2	0.535176	0.110687	0.70	8	1000-4999	Full time cou...	No relevent ...	Graduate	Public Sector	1
22	No	C1	0.544444	0.140291	0.80	9	50-99	no_enrollment	Has relevent...	Masters	Pvt Ltd	2
23	No	C1	0.515458	0.0480597	0.62	3	50-99	Full time cou...	Has relevent...	Masters	Pvt Ltd	1
24	No	C2	0.573089	0.232938	0.79	3	50-99	Full time cou...	No relevent ...	Graduate	Pvt Ltd	0
25	No	C2	0.542172	0.132296	0.93	1	50-99	no_enrollment	No relevent ...	Graduate	Pvt Ltd	1
26	No	C1	0.523552	0.0741099	0.85	3	50-99	no_enrollment	Has relevent...	Graduate	Pvt Ltd	0
27	No	C1	0.55949	0.188792	0.92	>20	5000-9999	no_enrollment	Has relevent...	Graduate	Pvt Ltd	>4
28	No	C1	0.5367	0.115682	0.92	>20	50-99	Part time co...	Has relevent...	Graduate	Pvt Ltd	1
29	No	C2	0.569583	0.220916	0.62	7	50-99	Full time cou...	No relevent ...	High School	Pvt Ltd	0
30	No	C1	0.557491	0.182244	0.90	>20	50-99	no_enrollment	Has relevent...	Graduate	Pvt Ltd	>4
31	No	C2	0.511691	0.0364732	0.94	8	18172.00	no_enrollment	No relevent ...	Graduate	NGO	>4
32	No	C1	0.544011	0.138842	0.77	18	100-500	no_enrollment	Has relevent...	Graduate	Pvt Ltd	0
33	No	C1	0.544975	0.140935	0.62	5	50-99	no_enrollment	Has relevent...	Masters	Pvt Ltd	1
34	No	C2	0.5358	0.112366	0.92	2	50-99	no_enrollment	No relevent ...	Graduate	Public Sector	1
35	No	C2	0.568163	0.21635	0.93	7	50-99	Full time cou...	No relevent ...	Graduate	Pvt Ltd	0
36	No	C1	0.5502	0.158781	0.92	>20	50-99	no_enrollment	Has relevent...	Masters	Pvt Ltd	2
37	No	C1	0.544688	0.141125	0.94	>20	50-99	no_enrollment	Has relevent...	Graduate	Pvt Ltd	>4
38	No	C2	0.545211	0.142183	0.58	<1	50-99	Full time cou...	No relevent ...	Graduate	Pvt Ltd	>4
39	No	C1	0.550361	0.159362	0.92	>20	18172.00	no_enrollment	Has relevent...	Graduate	Funded Star...	4

- ***Quan sát từ Excel***



	A	B	C	D	E	F	G	H	I	J	K	L
1	city_development_index	experience	company_size	enrolled_university	relevent_experience	education_level	company_type	last_new_job	major_discipline	Selected	Cluster	Silhouette (experience)
2	0,91	6	500-999	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	1	STEM	No	C2	0,1114699965
3	0,92	>20	100-500	no_enrollment	Has relevent experience	Masters	Funded Startup	2	Humanities	No	C2	0,08455039113
4	0,92	9	18172.00	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	1	STEM	No	C1	0,1522744218
5	0,62	15	>10000	no_enrollment	Has relevent experience	Masters	Pvt Ltd	1	STEM	No	C2	0,1301432367
6	0,7	12	500-999	no_enrollment	No relevent experience	Masters	NGO	1	STEM	No	C2	0,07702905791
7	0,9	13	1000-4999	no_enrollment	Has relevent experience	High School	Public Sector	2	STEM	No	C1	0,09960905745
8	0,92	>20	50-99	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	>4	STEM	No	C2	0,214893242
9	0,92	6	50-99	Full time course	No relevent experience	Graduate	Pvt Ltd	1	STEM	No	C1	0,0763516059
10	0,48	>20	>10000	Full time course	Has relevent experience	Phd	Pvt Ltd	>4	STEM	No	C2	0,1135043984
11	0,88	6	<10	no_enrollment	No relevent experience	Graduate	Pvt Ltd	3	STEM	No	C2	0,09289442727
12	0,9	>20	50-99	no_enrollment	Has relevent experience	Phd	Pvt Ltd	0	STEM	No	C1	0,1259733173
13	0,92	>20	50-99	no_enrollment	Has relevent experience	Masters	Pvt Ltd	>4	STEM	No	C1	0,1841461679
14	0,48	1	18172.00	Full time course	No relevent experience	Masters	Other	1	Other	No	C2	0,08737128007
15	0,62	2	18172.00	Full time course	Has relevent experience	Graduate	Public Sector	2	STEM	No	C1	0,06119230355
16	0,91	>20	50-99	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	4	STEM	No	C2	0,1453251835
17	0,8	10	50-99	Full time course	Has relevent experience	Graduate	Pvt Ltd	1	STEM	No	C1	0,1207724814
18	0,75	7	100-500	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	1	STEM	No	C2	0,1268782842
19	0,89	10	50-99	no_enrollment	No relevent experience	Masters	Pvt Ltd	2	STEM	No	C2	0,09169114675
20	0,91	>20	50-99	no_enrollment	Has relevent experience	High School	Pvt Ltd	1	STEM	No	C1	0,1136570654
21	0,7	>20	100-500	no_enrollment	Has relevent experience	Masters	Pvt Ltd	1	STEM	No	C2	0,1028179888
22	0,7	8	1000-4999	Full time course	No relevent experience	Graduate	Public Sector	1	STEM	No	C1	0,05063548218
23	0,8	9	50-99	no_enrollment	Has relevent experience	Masters	Pvt Ltd	2	STEM	No	C1	0,074176404
24	0,62	3	50-99	Full time course	Has relevent experience	Masters	Pvt Ltd	1	STEM	No	C1	0,1262844496
25	0,79	3	50-99	Full time course	No relevent experience	Graduate	Pvt Ltd	0	STEM	No	C2	0,1490920303
26	0,93	1	50-99	no_enrollment	No relevent experience	Graduate	Pvt Ltd	1	STEM	No	C2	0,2070118029
27	0,85	3	50-99	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	0	Other	No	C1	0,07388298249
28	0,92	>20	5000-9999	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	>4	STEM	No	C1	0,1791460404
29	0,92	>20	50-99	Part time course	Has relevent experience	Graduate	Pvt Ltd	1	STEM	No	C2	0,08993688553
30	0,62	7	50-99	Full time course	No relevent experience	High School	Pvt Ltd	0	STEM	No	C1	0,0201006174
31	0,9	>20	50-99	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	>4	STEM	No	C2	0,214540268
32	0,94	8	18172.00	no_enrollment	No relevent experience	Graduate	NGO	>4	Humanities	No	C2	0,04354930671
33	0,77	18	100-500	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	0	STEM	No	C1	0,1286242214
34	0,62	5	50-99	no_enrollment	Has relevent experience	Masters	Pvt Ltd	1	STEM	No	C1	0,08164357588
35	0,92	2	50-99	no_enrollment	No relevent experience	Graduate	Public Sector	1	STEM	No	C2	0,1076225243
36	0,93	7	50-99	Full time course	No relevent experience	Graduate	Pvt Ltd	0	STEM	No	C2	0,03461432908
37	0,92	>20	50-99	no_enrollment	Has relevent experience	Masters	Pvt Ltd	2	STEM	No	C1	0,1341445428
38	0,94	>20	50-99	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	>4	Business Degree	No	C2	0,1446193869
39	0,58	<1	50-99	Full time course	No relevent experience	Graduate	Pvt Ltd	>4	Humanities	No	C1	0,1199236305

Sheet1

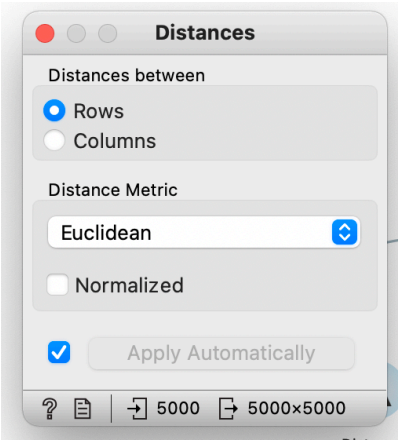
1.3.2. Phân cụm phân cấp

Chiến lược phân cấp chia thành hai mô hình cơ bản: hợp nhất (agglomerative) và phân chia (division). Ta ứng dụng phân cụm phân cấp vào bộ dữ liệu Job Change Database và sử dụng chiến lược phân chia. Xây dựng một cây phân cấp cho dữ liệu cần gom cụm dựa trên: Ma trận khoảng cách giữa các phần tử; Độ đo khoảng cách giữa các cụm. Phương pháp này không cần xác định trước số cụm nhưng cần xác định điều kiện dừng. Có hai chiến lược phân chia chính phụ thuộc vào chiều di chuyển trên biểu đồ dendrogram, ở đây ta thực hiện chiến lược hợp nhất

Mục đích của phân cụm phân cấp theo chiến lược hợp nhất này là xây dựng các bước để hợp nhất hai nhóm có khoảng cách gần nhất với nhau và đồng thời giảm số lượng cụm trước khi chúng đạt được tới node gốc gồm toàn bộ các quan sát. Qua đó, bộ dữ liệu Job Change Database sẽ được xác định điểm đại diện cho từng cụm, khoảng cách giữa 2 cụm để thực hiện phân chia các đối tượng trong tập dữ liệu thu được.

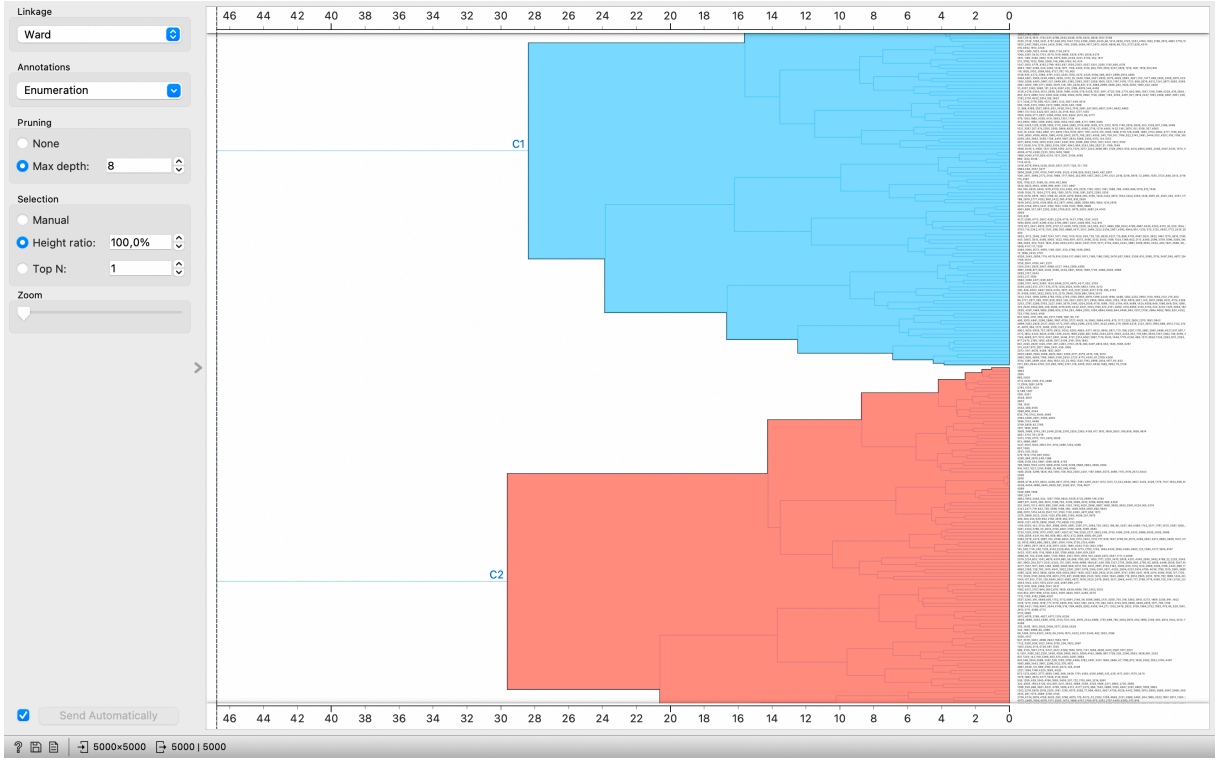
✳ ***Distances***

Độ đo phân cụm được sử dụng tính khoảng cách giữa các hàng hoặc cột trong tập dữ liệu nhằm phục vụ cho quá trình gom cụm. Ta tính toán khoảng cách giữa các cá thể dữ liệu hàng **Rows** và dùng thước đo khoảng cách **Euclidean**. Sau đó chuyển kết quả đến phân cụm phân cấp để tìm các nhóm cá thể dữ liệu.

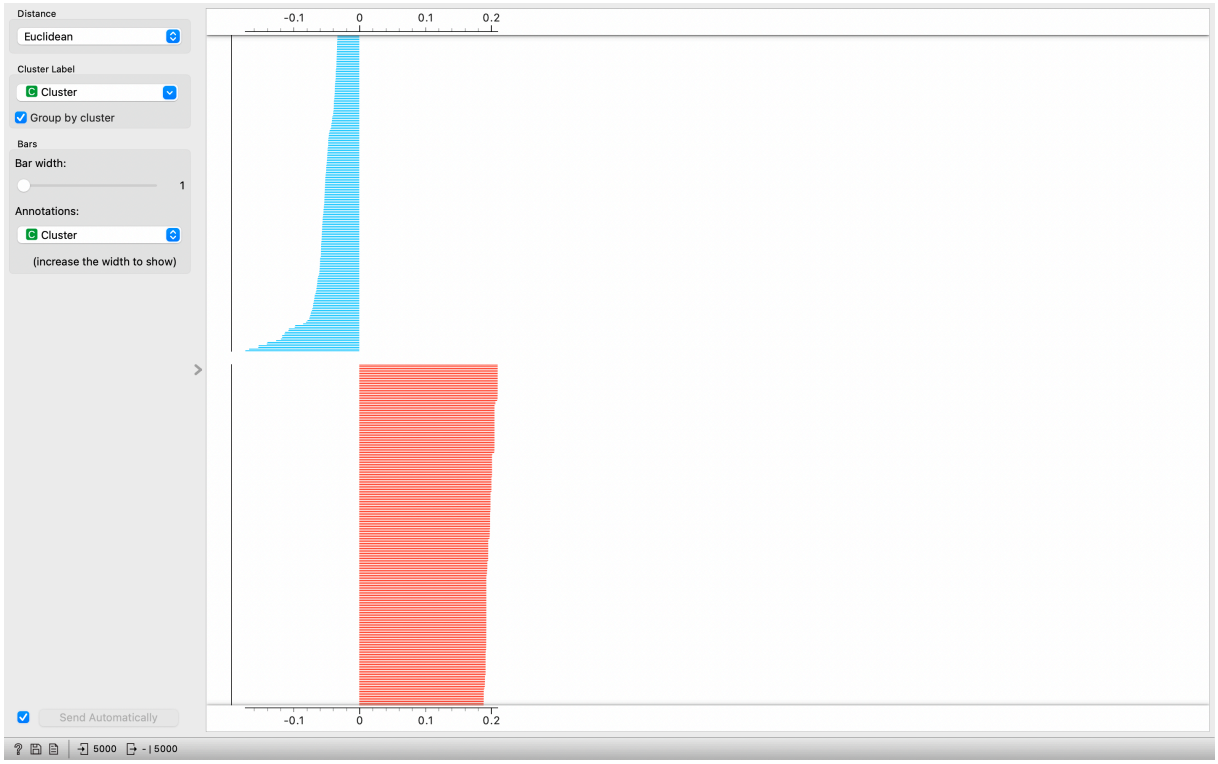


✳ ***Hierarchical Clustering***

Sau khi dữ liệu được chuyển đến phân cụm phân cấp, Hierarchical Clustering sẽ hiển thị một biểu đồ Dendrogram, một cấu trúc phân nhóm giống như cây. Mỗi nút đại diện cho một thể hiện trong tập dữ liệu. Để đào tạo các cụm, chúng ta nhấp vào thước ở ngưỡng mong muốn. Trong trường hợp này, chúng ta đã chọn 2 cụm dựa vào phương pháp Ward linkage tính toán sự gia tăng của tổng sai số của các bình phương. Nói cách khác, tiêu chí phương sai tối thiểu của Ward giảm thiểu tổng phương sai trong cụm. Sau đó chuyển các cụm đó sang MDS, nó hiển thị hình chiếu 2D của các cá thể dữ liệu, được tô màu theo nhãn cụm. Ta quan sát được 2 cụm với cụm C1 có nhãn màu xanh, cụm C2 có nhãn màu đỏ. Tương ứng nhãn của nút trong biểu đồ ta chọn tùy ý, ở đây **Annotations** là experience cho ta quan sát số năm kinh nghiệm của 5000 đối tượng ứng với từng cụm:



✳ **Silhouette Plot**



Tiếp đó chọn cụm trong Dendrogram và chuyển toàn bộ dữ liệu vào Silhouette Plot. Cũng giống như phân cụm phân hoạch, biểu đồ đánh giá trực quan chất lượng phân nhóm với các điểm trong các cụm lân cận.Ta thu được biểu đồ Silhouette cho ra 2 cụm đối tượng thuộc C1 và C2. Ta thu được biểu đồ Silhouette cho ra 2 cụm cho tổng 5000 đối tượng: Tương ứng với màu xanh biểu diễn cho 1907 đối tượng thuộc C1 (Có muốn thay đổi công việc), màu đỏ biểu diễn cho 3093 đối tượng thuộc C2 (Không muốn thay đổi công việc). Khi chọn phương

pháp tính khoảng cách Euclidean thì xuất hiện giá trị âm, nghĩa là đang tồn tại đối tượng bị gán sai cụm.

❖ ***Dữ liệu sau khi thực hiện phân cụm phân cấp***

- ***Quan sát từ Orange***

Info
5000 instances (no missing data)
9 features
Target with 2 values
2 meta attributes

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

Selection
Select full rows

Restore Original Order
Send Automatically

5000 | 5000 | 5000

	Selected	Cluster	ilhouette (Cluster	_development_in	experience	company_size	nrolled_university	elevent_experienc	education_level	company_type	last_new_job	major
1	No	C2	0.158408	0.91	6	500-999	no_enrollment	Has relevent...	Graduate	Pvt Ltd	1	STEN
2	No	C2	0.143306	0.92	>20	100-500	no_enrollment	Has relevent...	Masters	Funded Star...	2	Huma
3	No	C2	0.161785	0.92	9	18172.00	no_enrollment	Has relevent...	Graduate	Pvt Ltd	1	STEN
4	No	C2	0.153191	0.62	15	>10000	no_enrollment	Has relevent...	Masters	Pvt Ltd	1	STEN
5	No	C1	0.000461706	0.70	12	500-999	no_enrollment	No relevent ...	Masters	NGO	1	STEN
6	No	C2	0.109785	0.90	13	1000-4999	no_enrollment	Has relevent...	High School	Public Sector	2	STEN
7	No	C2	0.192925	0.92	>20	50-99	no_enrollment	Has relevent...	Graduate	Pvt Ltd	>4	STEN
8	No	C1	0.184995	0.92	6	50-99	Full time cou...	No relevent ...	Graduate	Pvt Ltd	1	STEN
9	No	C2	0.051531	0.48	>20	>10000	Full time cou...	Has relevent...	Phd	Pvt Ltd	>4	STEN
10	No	C1	0.00713434	0.88	6	<10	no_enrollment	No relevent ...	Graduate	Pvt Ltd	3	STEN
11	No	C2	0.1129	0.90	>20	50-99	no_enrollment	Has relevent...	Phd	Pvt Ltd	0	STEN
12	No	C2	0.179745	0.92	>20	50-99	no_enrollment	Has relevent...	Masters	Pvt Ltd	>4	STEN
13	No	C1	0.0830137	0.48	1	18172.00	Full time cou...	No relevent ...	Masters	Other	1	Othe
14	No	C1	-0.0194079	0.62	2	18172.00	Full time cou...	Has relevent...	Graduate	Public Sector	2	STEN
15	No	C2	0.166469	0.91	>20	50-99	no_enrollment	Has relevent...	Graduate	Pvt Ltd	4	STEN
16	No	C1	0.0187396	0.80	10	50-99	Full time cou...	Has relevent...	Graduate	Pvt Ltd	1	STEN
17	No	C2	0.170329	0.75	7	100-500	no_enrollment	Has relevent...	Graduate	Pvt Ltd	1	STEN
18	No	C1	0.0287673	0.89	10	50-99	no_enrollment	No relevent ...	Masters	Pvt Ltd	2	STEN
19	No	C2	0.125134	0.91	>20	50-99	no_enrollment	Has relevent...	High School	Pvt Ltd	1	STEN
20	No	C2	0.182743	0.70	>20	100-500	no_enrollment	Has relevent...	Masters	Pvt Ltd	1	STEN
21	No	C1	0.106757	0.70	8	1000-4999	Full time cou...	No relevent ...	Graduate	Public Sector	1	STEN
22	No	C2	0.137715	0.80	9	50-99	no_enrollment	Has relevent...	Masters	Pvt Ltd	2	STEN
23	No	C1	0.0209239	0.62	3	50-99	Full time cou...	Has relevent...	Masters	Pvt Ltd	1	STEN
24	No	C1	0.205232	0.79	3	50-99	Full time cou...	No relevent ...	Graduate	Pvt Ltd	0	STEN
25	No	C1	0.0697426	0.93	1	50-99	no_enrollment	No relevent ...	Graduate	Pvt Ltd	1	STEN
26	No	C2	0.0797836	0.85	3	50-99	no_enrollment	Has relevent...	Graduate	Pvt Ltd	0	Othe
27	No	C2	0.194463	0.92	>20	5000-9999	no_enrollment	Has relevent...	Graduate	Pvt Ltd	>4	STEN
28	No	C1	-0.0505792	0.92	>20	50-99	Part time co...	Has relevent...	Graduate	Pvt Ltd	1	STEN
29	No	C1	0.192683	0.62	7	50-99	Full time cou...	No relevent ...	High School	Pvt Ltd	0	STEN
30	No	C2	0.19285	0.90	>20	50-99	no_enrollment	Has relevent...	Graduate	Pvt Ltd	>4	STEN
31	No	C1	-0.0116642	0.94	8	18172.00	no_enrollment	No relevent ...	Graduate	NGO	>4	Huma
32	No	C2	0.137961	0.77	18	100-500	no_enrollment	Has relevent...	Graduate	Pvt Ltd	0	STEN
33	No	C2	0.129933	0.62	5	50-99	no_enrollment	Has relevent...	Masters	Pvt Ltd	1	STEN
34	No	C1	0.0542516	0.92	2	50-99	no_enrollment	No relevent ...	Graduate	Public Sector	1	STEN
35	No	C1	0.190835	0.93	7	50-99	Full time cou...	No relevent ...	Graduate	Pvt Ltd	0	STEN
36	No	C2	0.164351	0.92	>20	50-99	no_enrollment	Has relevent...	Masters	Pvt Ltd	2	STEN
37	No	C2	0.151368	0.94	>20	50-99	no_enrollment	Has relevent...	Graduate	Pvt Ltd	>4	Busir
38	No	C1	0.123127	0.58	<1	50-99	Full time cou...	No relevent ...	Graduate	Pvt Ltd	>4	Huma
39	No	C2	0.1594	0.92	>20	18172.00	no_enrollment	Has relevent...	Graduate	Funded Star...	4	STEN
40	No	C1	0.137955	0.84	5	100-500	Full time cou...	No relevent ...	High School	Pvt Ltd	1	STEN
41	No	C1	0.0196	0.91	7	50-99	Full time cou...	Has relevent...	Graduate	Pvt Ltd	1	STEN

- ***Quan sát từ Excel***

	A	B	C	D	E	F	G	H	I	J	K	L
1	city_development_index	experience	company_size	enrolled_university	relevent_experience	education_level	company_type	last_new_job	major_discipline	Selected	Cluster	Silhouette (experience)
2	0,91	6	500-999	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	1	STEM	No	C2	0,1114699965
3	0,92	>20	100-500	no_enrollment	Has relevent experience	Masters	Funded Startup	2	Humanities	No	C2	0,08455039113
4	0,92	9	18172.00	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	1	STEM	No	C1	0,1522744218
5	0,62	15	>10000	no_enrollment	Has relevent experience	Masters	Pvt Ltd	1	STEM	No	C2	0,1301432367
6	0,7	12	500-999	no_enrollment	No relevent experience	Masters	NGO	1	STEM	No	C2	0,07702905791
7	0,9	13	1000-4999	no_enrollment	Has relevent experience	High School	Public Sector	2	STEM	No	C1	0,09960905745
8	0,92	>20	50-99	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	>4	STEM	No	C2	0,214893242
9	0,92	6	50-99	Full time course	No relevent experience	Graduate	Pvt Ltd	1	STEM	No	C1	0,0763516059
10	0,48	>20	>10000	Full time course	Has relevent experience	Phd	Pvt Ltd	>4	STEM	No	C2	0,1135043984
11	0,88	6	<10	no_enrollment	No relevent experience	Graduate	Pvt Ltd	3	STEM	No	C2	0,09289442727
12	0,9	>20	50-99	no_enrollment	Has relevent experience	Phd	Pvt Ltd	0	STEM	No	C1	0,1259733173
13	0,92	>20	50-99	no_enrollment	Has relevent experience	Masters	Pvt Ltd	>4	STEM	No	C1	0,1841461679
14	0,48	1	18172.00	Full time course	No relevent experience	Masters	Other	1	Other	No	C2	0,08737128007
15	0,62	2	18172.00	Full time course	Has relevent experience	Graduate	Public Sector	2	STEM	No	C1	0,06119230355
16	0,91	>20	50-99	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	4	STEM	No	C2	0,1453251835
17	0,8	10	50-99	Full time course	Has relevent experience	Graduate	Pvt Ltd	1	STEM	No	C1	0,1207724814
18	0,75	7	100-500	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	1	STEM	No	C2	0,1268782842
19	0,89	10	50-99	no_enrollment	No relevent experience	Masters	Pvt Ltd	2	STEM	No	C2	0,09169114675
20	0,91	>20	50-99	no_enrollment	Has relevent experience	High School	Pvt Ltd	1	STEM	No	C1	0,1136570654
21	0,7	>20	100-500	no_enrollment	Has relevent experience	Masters	Pvt Ltd	1	STEM	No	C2	0,1028179888
22	0,7	8	1000-4999	Full time course	No relevent experience	Graduate	Public Sector	1	STEM	No	C1	0,05063548218
23	0,8	9	50-99	no_enrollment	Has relevent experience	Masters	Pvt Ltd	2	STEM	No	C1	0,074176404
24	0,62	3	50-99	Full time course	Has relevent experience	Masters	Pvt Ltd	1	STEM	No	C1	0,1262844496
25	0,79	3	50-99	Full time course	No relevent experience	Graduate	Pvt Ltd	0	STEM	No	C2	0,1490920303
26	0,93	1	50-99	no_enrollment	No relevent experience	Graduate	Pvt Ltd	1	STEM	No	C2	0,2070118029
27	0,85	3	50-99	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	0	Other	No	C1	0,07388298249
28	0,92	>20	5000-9999	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	>4	STEM	No	C1	0,1791460404
29	0,92	>20	50-99	Part time course	Has relevent experience	Graduate	Pvt Ltd	1	STEM	No	C2	0,08993688553
30	0,62	7	50-99	Full time course	No relevent experience	High School	Pvt Ltd	0	STEM	No	C1	0,0201006174
31	0,9	>20	50-99	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	>4	STEM	No	C2	0,214540268
32	0,94	8	18172.00	no_enrollment	No relevent experience	Graduate	NGO	>4	Humanities	No	C2	0,04354930671
33	0,77	18	100-500	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	0	STEM	No	C1	0,1286242214
34	0,62	5	50-99	no_enrollment	Has relevent experience	Masters	Pvt Ltd	1	STEM	No	C1	0,08164357588
35	0,92	2	50-99	no_enrollment	No relevent experience	Graduate	Public Sector	1	STEM	No	C2	0,1076225243
36	0,93	7	50-99	Full time course	No relevent experience	Graduate	Pvt Ltd	0	STEM	No	C2	0,03461432908
37	0,92	>20	50-99	no_enrollment	Has relevent experience	Masters	Pvt Ltd	2	STEM	No	C1	0,1341445428
38	0,94	>20	50-99	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	>4	Business Degree	No	C2	0,1446193869
39	0,58	<1	50-99	Full time course	No relevent experience	Graduate	Pvt Ltd	>4	Humanities	No	C1	0,1199236305

Sheet1

1.4. Đánh giá các phương pháp phân cụm với bộ dữ liệu gốc

Sử dụng các phương pháp: **đánh giá ngoài, đánh giá nội bộ, đánh giá tương đối.**

- **Đánh giá ngoài (external validation):** Là đánh giá kết quả phân cụm dựa vào cấu trúc/ xu hướng phân cụm được chỉ định trước cho tập dữ liệu.
- **Đánh giá nội bộ (internal validation):** Là đánh giá kết quả phân cụm mà không có thông tin từ bên ngoài, chủ yếu dựa trên các vector chính của dữ thông qua ma trận xấp xỉ (proximity matrix). Độ đo được ta sử dụng trong 2 phương pháp phân cụm ở trên Silhouette index. Với chỉ số:
  - **>= 0.5:** sát thực tế
  - **0.25 < 0.5:** cần đánh giá lại (kinh nghiệm chuyên gia)
  - **< 0.25:** không tin tưởng vào cluster, tìm phương pháp đánh giá khác
- **Đánh giá tương đối (relative validation)**: :Là đánh giá kết quả gom cụm bằng việc so sánh với kết quả gom cụm ứng với các bộ trị thông số khác nhau. Kết quả gom cụm của các phương pháp khác. Ta đã thực hiện phương pháp đánh giá tương đối bằng việc so sánh kết quả gom cụm của phương pháp phân cụm phân cấp và phương pháp phân cụm phân hoạch. Ta xuất ra bảng excel và thực hiện đánh giá 2 phương pháp :
  - **Cột so sánh K-means** (đối chiếu kết quả phân cụm bằng phương pháp phân hoạch với nhãn ban đầu). Ta thực hiện so sánh và tính được ***tỷ lệ chính xác là 100%***. Đây là kết quả thể hiện độ tương đồng tuyệt đối khi phân cụm bằng phương pháp này.
  - **Cột so sánh Distances** (đối chiếu kết quả phân cụm bằng phương pháp phân cấp với nhãn ban đầu). Ta thực hiện so sánh và tính được ***tỷ lệ chính xác là 12,14%.***

	ATTRIBUTES										COMPARE							
city_development_index	experience	company_size	enrolled_university	relevent_experience	education_level	company_type	last_new_job	major_discipline	target	Initial Cluster	K-means Cluster	SO SÁNH K - MEANS	Distances Cluster	SO SÁNH DISTANCES	Silhouette	Silhouette (Cluster)	Selected
1	6	500-999	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	1	STEM	0	C1	C1	TRUE	C2	FALSE	0,555286619	0,175126173	No
1	>20	100-500	no_enrollment	Has relevent experience	Masters	Funded Startup	2	Humanities	0	C1	C1	TRUE	C2	FALSE	0,543508816	0,137436063	No
1	9	18172	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	1	STEM	0	C1	C1	TRUE	C2	FALSE	0,556599394	0,179327838	No
1	15	>10000	no_enrollment	Has relevent experience	Masters	Pvt Ltd	1	STEM	0	C1	C1	TRUE	C2	FALSE	0,551933941	0,163350190	No
1	12	500-999	no_enrollment	No relevent experience	Masters	NGO	1	STEM	0	C2	C2	TRUE	C1	FALSE	0,514348381	0,045040024	No
1	13	1000-4999	no_enrollment	Has relevent experience	High School	Public Sector	2	STEM	0	C1	C1	TRUE	C2	FALSE	0,534153925	0,107656116	No
1	>20	50-99	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	>4	STEM	1	C1	C1	TRUE	C2	FALSE	0,557492353	0,182225477	No
1	6	50-99	Full time course	No relevent experience	Graduate	Pvt Ltd	1	STEM	0	C2	C2	TRUE	C1	FALSE	0,560734438	0,192166010	No
0	>20	>10000	Full time course	Has relevent experience	Phd	Pvt Ltd	>4	STEM	0	C1	C1	TRUE	C2	FALSE	0,528118735	0,087037521	No
1	6	<10	no_enrollment	No relevent experience	Graduate	Pvt Ltd	3	STEM	0	C2	C2	TRUE	C1	FALSE	0,519071122	0,059691513	No
1	>20	50-99	no_enrollment	Has relevent experience	Phd	Pvt Ltd	0	STEM	0	C1	C1	TRUE	C2	FALSE	0,530723699	0,096804230	No
1	>20	50-99	no_enrollment	Has relevent experience	Masters	Pvt Ltd	>4	STEM	1	C1	C1	TRUE	C2	FALSE	0,552749582	0,166983271	No
0	1	18172	Full time course	No relevent experience	Masters	Other	1	Other	1	C2	C2	TRUE	C1	FALSE	0,527972269	0,087483225	No
1	2	18172	Full time course	Has relevent experience	Graduate	Public Sector	2	STEM	0	C1	C1	TRUE	C1	TRUE	0,521752955	0,067926412	No
1	>20	50-99	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	4	STEM	0	C1	C1	TRUE	C2	FALSE	0,551519411	0,163031627	No
1	10	50-99	Full time course	Has relevent experience	Graduate	Pvt Ltd	1	STEM	0	C1	C1	TRUE	C1	TRUE	0,520132102	0,063239842	No
1	7	100-500	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	1	STEM	0	C1	C1	TRUE	C2	FALSE	0,559008371	0,186949427	No
1	10	50-99	no_enrollment	No relevent experience	Masters	Pvt Ltd	2	STEM	1	C2	C2	TRUE	C1	FALSE	0,528502404	0,089394850	No
1	>20	50-99	no_enrollment	Has relevent experience	High School	Pvt Ltd	1	STEM	0	C1	C1	TRUE	C2	FALSE	0,539802320	0,125574708	No
1	>20	100-500	no_enrollment	Has relevent experience	Masters	Pvt Ltd	1	STEM	0	C1	C1	TRUE	C2	FALSE	0,559208711	0,187354073	No
1	8	1000-4999	Full time course	No relevent experience	Graduate	Public Sector	1	STEM	1	C2	C2	TRUE	C1	FALSE	0,535176301	0,110687265	No
1	9	50-99	no_enrollment	Has relevent experience	Masters	Pvt Ltd	2	STEM	1	C1	C1	TRUE	C2	FALSE	0,544443552	0,140290650	No
1	3	50-99	Full time course	Has relevent experience	Masters	Pvt Ltd	1	STEM	1	C1	C1	TRUE	C1	TRUE	0,515457599	0,048059744	No
1	3	50-99	Full time course	No relevent experience	Graduate	Pvt Ltd	0	STEM	1	C2	C2	TRUE	C1	FALSE	0,573088883	0,232938454	No
1	1	50-99	no_enrollment	No relevent experience	Graduate	Pvt Ltd	1	STEM	0	C2	C2	TRUE	C1	FALSE	0,542172333	0,132296266	No
1	3	50-99	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	0	Other	0	C1	C1	TRUE	C2	FALSE	0,523551919	0,074109859	No
1	>20	5000-9999	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	>4	STEM	0	C1	C1	TRUE	C2	FALSE	0,559490217	0,188792387	No
1	>20	50-99	Part time course	Has relevent experience	Graduate	Pvt Ltd	1	STEM	0	C1	C1	TRUE	C1	TRUE	0,536699516	0,115682337	No
1	7	50-99	Full time course	No relevent experience	High School	Pvt Ltd	0	STEM	1	C2	C2	TRUE	C1	FALSE	0,569583189	0,220916149	No
1	>20	50-99	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	>4	STEM	0	C1	C1	TRUE	C2	FALSE	0,557491307	0,182244140	No
1	8	18172	no_enrollment	No relevent experience	Graduate	NGO	>4	Humanities	0	C2	C2	TRUE	C1	FALSE	0,511691342	0,036473234	No
1	18	100-500	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	0	STEM	0	C1	C1	TRUE	C2	FALSE	0,544011240	0,138842336	No
1	5	50-99	no_enrollment	Has relevent experience	Masters	Pvt Ltd	1	STEM	0	C1	C1	TRUE	C2	FALSE	0,544975399	0,140935417	No
1	2	50-99	no_enrollment	No relevent experience	Graduate	Public Sector	1	STEM	0	C2	C2	TRUE	C1	FALSE	0,535800079	0,112365626	No
1	7	50-99	Full time course	No relevent experience	Graduate	Pvt Ltd	0	STEM	0	C2	C2	TRUE	C1	FALSE	0,568163129	0,216350385	No
1	>20	50-99	no_enrollment	Has relevent experience	Masters	Pvt Ltd	2	STEM	0	C1	C1	TRUE	C2	FALSE	0,550199925	0,158780630	No
1	>20	50-99	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	>4	Business Degree	0	C1	C1	TRUE	C2	FALSE	0,544687863	0,141124976	No
1	<1	50-99	Full time course	No relevent experience	Graduate	Pvt Ltd	>4	Humanities	0	C2	C2	TRUE	C1	FALSE	0,545210976	0,142183113	No
1	>20	18172	no_enrollment	Has relevent experience	Graduate	Funded Startup	4	STEM	0	C1	C1	TRUE	C2	FALSE	0,550360658	0,159362239	No
1	5	100-500	Full time course	No relevent experience	High School	Pvt Ltd	1	STEM	0	C2	C2	TRUE	C1	FALSE	0,545391999	0,143227627	No
1	7	50-99	Full time course	Has relevent experience	Graduate	Pvt Ltd	1	STEM	1	C1	C1	TRUE	C1	TRUE	0,519566867	0,061566906	No
1	16	50-99	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	>4	STEM	0	C1	C1	TRUE	C2	FALSE	0,549462366	0,155171708	No
1	9	100-500	no_enrollment	Has relevent experience	Masters	Pvt Ltd	1	STEM	1	C1	C1	TRUE	C2	FALSE	0,554177628	0,171619144	No
1	5	1000-4999	no_enrollment	Has relevent experience	Masters	Pvt Ltd	4	STEM	0	C1	C1	TRUE	C2	FALSE	0,546456301	0,146147174	No
1	2	50-99	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	1	STEM	0	C1	C1	TRUE	C2	FALSE	0,545798227	0,142668042	No
1	2	50-99	no_enrollment	No relevent experience	Graduate	Pvt Ltd	2	STEM	0	C2	C2	TRUE	C1	FALSE	0,538666911	0,121393873	No
1	>20	>10000	no_enrollment	No relevent experience	Masters	Public Sector	>4	STEM	0	C2	C2	TRUE	C2	TRUE	0,509079747	0,028244582	No
1	5	500-999	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	2	STEM	0	C1	C1	TRUE	C2	FALSE	0,552373444	0,165753257	No
1	4	>10000	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	2	STEM	1	C1	C1	TRUE	C2	FALSE	0,552793593	0,166033403	No
1	11	18172	no_enrollment	No relevent experience	Graduate	Pvt Ltd	>4	STEM	0	C2	C2	TRUE	C1	FALSE	0,515942435	0,049767302	No
1	4	100-500	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	1	STEM	0	C1	C1	TRUE	C2	FALSE	0,557670508	0,182760918	No
1	4	100-500	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	1	STEM	0	C1	C1	TRUE	C2	FALSE	0,557670508	0,182760918	No

Sheet1

❖ ***Kết luận***

Sau khi tiến hành so sánh kết quả phân cụm với nhãn ban đầu và tính tỷ lệ chính xác để tìm ra phương pháp tốt hơn. **Kết quả cho thấy phương pháp K-means cho tỷ lệ chính xác cao hơn (tuyệt đối) so với phương pháp Hierarchical Clustering (100% > 12,14%)**

2. Phân lớp dữ liệu

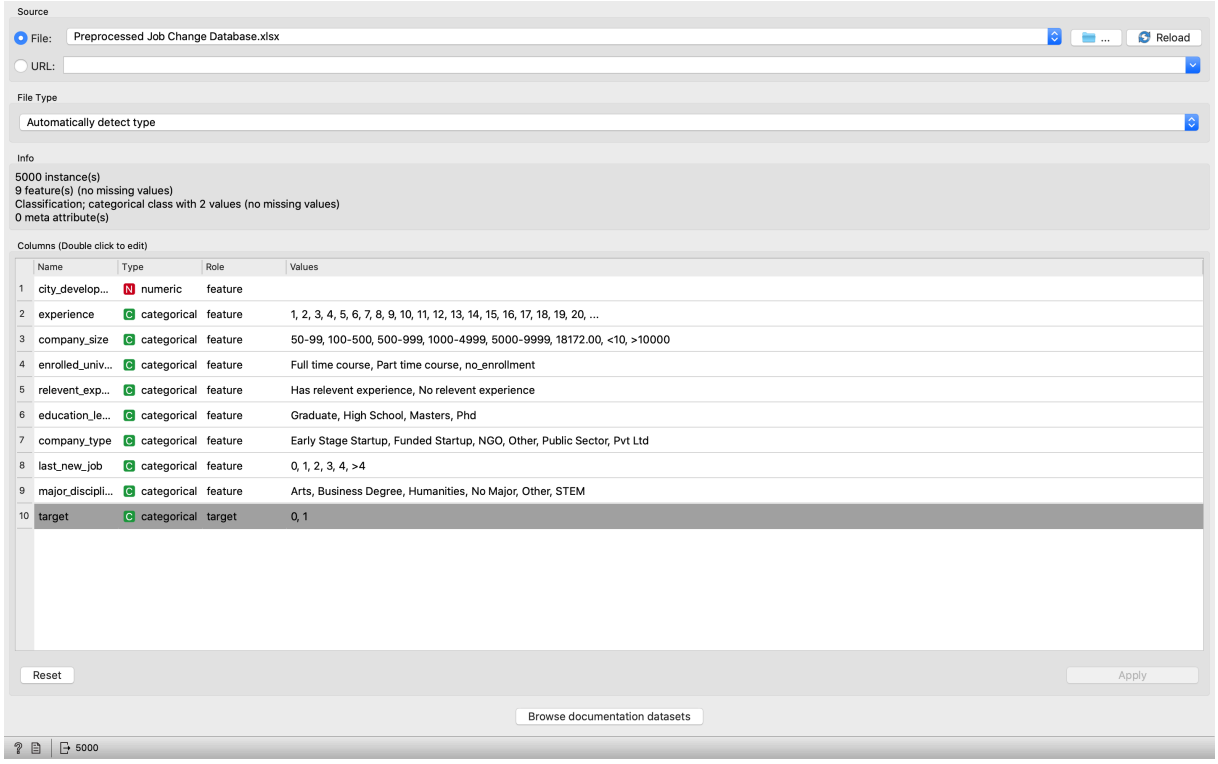
2.1. Định nghĩa phân lớp dữ liệu

Phân lớp dữ liệu là kỹ thuật dựa trên tập huấn luyện và những giá trị hay là nhãn của lớp trong một thuộc tính phân lớp và sử dụng nó trong việc phân lớp dữ liệu mới. Nghĩa là xếp các mẫu dữ liệu hay các đối tượng vào một trong các lớp đã được định nghĩa trước. Kỹ thuật phân lớp được tiến hành bao gồm 2 bước:

- **Bước 1**: Học, mục đích của bước này là xây dựng mô hình xác định tập các lớp dữ liệu.
- **Bước 2** : Kiểm tra và đánh giá, bước này sử dụng mô hình phân lớp đã được xây dựng ở bước 1 vào việc phân lớp.

2.2. Xây dựng mô hình phân lớp

❖ ***Bộ dữ liệu sử dụng phân lớp***

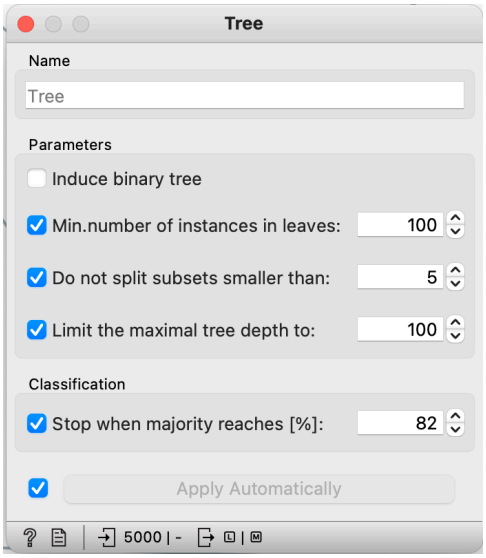


✴ **Cây quyết định**

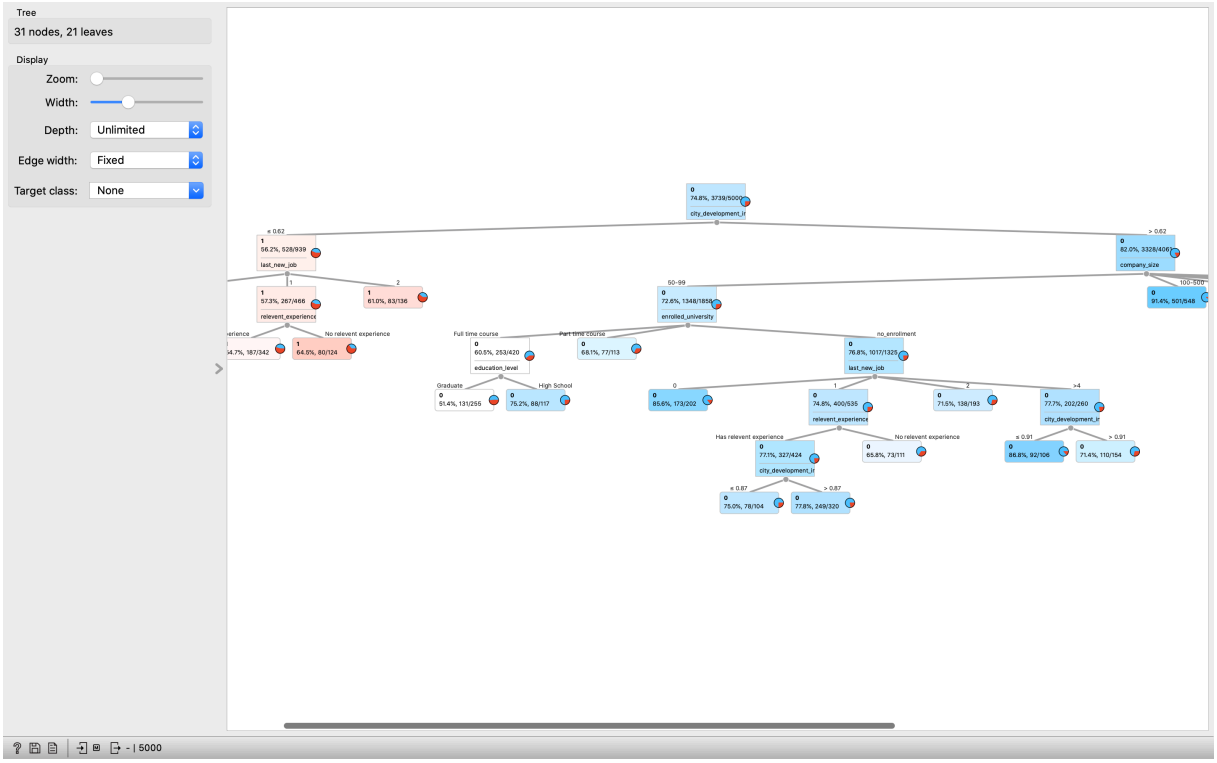
Cây quyết định là một flowchart giống cấu trúc cây, nút bên trong biểu thị một phép kiểm tra trên một thuộc tính, nhánh biểu diễn đầu ra của một phép kiểm tra, nút lá biểu diễn nhãn phân lớp hoặc sự phân chia vào lớp. Xây dựng cây quyết định cho bộ dữ liệu như sau:

- **Induce binary tree**: xây dựng cây nhị phân (chia thành hai nút con). Đối với bộ dữ liệu này, ta không cần xây dựng cây nhị phân nên sẽ không chọn.
- **Min. number of instances in leaves:** Chỉ định (n) số phần tử tối thiểu trong node lá, nếu 1 node lá có số phần tử ít hơn n thì sẽ không tách nữa. Ta chọn số phần tử tối thiểu là 100 trong 1 lá đối với cây quyết định này

- **Do not split subsets smaller than** : không cho phép thuật toán phân chia các nút có ít hơn số lượng ví dụ đã cho. Ở đây ta chọn tập con nhỏ nhất là 5 và không thể phân chia các nút nhỏ hơn số lượng này.
- **Limit the maximal tree depth**: giới hạn độ sâu của cây phân loại ở số cấp nút được chỉ định. Ở đây, ta chọn cây có độ sâu là 100.

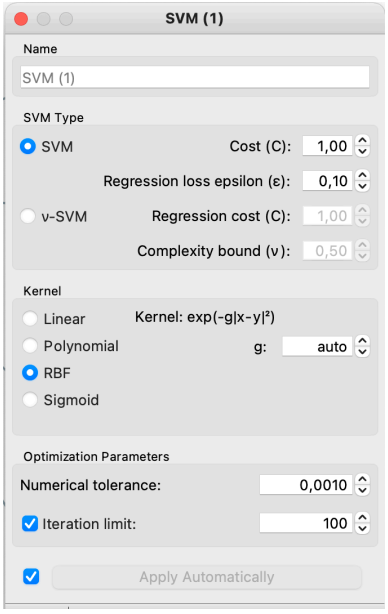


Sử dụng tiện ích **Tree Viewer** để quan sát cây quyết định vừa xây dựng. Với độ sâu không giới hạn, ta có được cây quyết định với tổng cộng 31 nút và 21 lá.



✳ SVM

SVM là một mô hình phân loại hoạt động bằng việc xây dựng một siêu phẳng trong không gian nhiều chiều làm mặt phân cách các lớp dữ liệu, với các dữ liệu đầu vào sẽ được xem như các vector trong không gian. Đối với bộ dữ liệu trên, ta sẽ thực hiện phân lớp theo kỹ thuật SVM, với các chỉ số như sau:

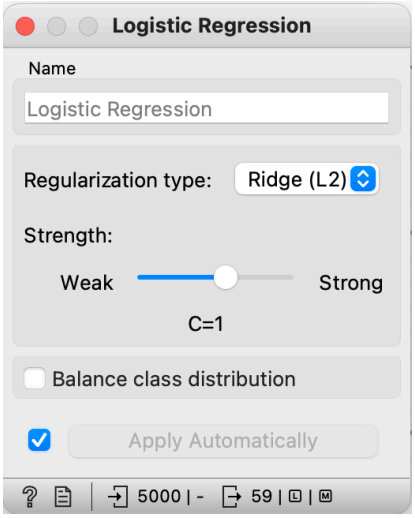


- **Cost:** thời hạn lỗi đối với tổn thất và áp dụng cho các nhiệm vụ phân loại và hồi quy. C càng lớn có nghĩa là SVM càng bị phạt nặng khi thực hiện phân loại sai. Do đó, lề càng hẹp và càng ít vector hỗ trợ được sử dụng. Ở đây chúng ta sẽ chỉ chọn mức phạt là C=1.0 đối với thuật toán SVM này.
- **ε**: một tham số cho mô hình epsilon-SVR, áp dụng cho các tác vụ hồi quy. Xác định khoảng cách từ các giá trị thực mà trong đó không có hình phạt nào được liên kết với các giá trị được dự đoán. Ứng với C = 1.0, ta chọn **ε** = 0.1
- **Kernel** là một hàm biến đổi không gian thuộc tính thành không gian tính năng mới để phù hợp với siêu phẳng có biên độ tối đa, ở đây ta chọn hàm Radial Basis Function kernel (RBF), đây là hàm được sử dụng nhiều trong thực tế.

✳ Logistic Regression

Tải bộ dữ liệu "Preprocessed Job Change Database" và chuyển dữ liệu vào **Logistic Regression**. Đây là một thuật toán phân loại bằng cách gán các đối tượng cho 1 tập hợp giá trị rời rạc (ví dụ như 0,1,2…). Đối với bộ dữ liệu trên, ta thực hiện hồi quy theo các chỉ số sau:

- **Regularization type**: phương pháp hồi quy theo dạng chính quy hóa LASSO (L1) hoặc Ridge (L2)
- **Strength**: cường độ chi phí mặc định C=1



2.2. Sử dụng mô hình phân lớp

Trước khi tiến hành so sánh giữa các chỉ số, ta chọn phương pháp lấy mẫu **Cross Validation**, tức chia dữ liệu thành một số lần nhất định (ở đây ta chọn là 10 lần), thuật toán được kiểm tra bằng cách đưa ra các ví dụ từ một lần trong một lần; mô hình được tạo ra từ các nếp gấp khác và các ví dụ từ nếp gấp được phân loại.

Stratified: quyết định xem có thực hiện Cross-validation phân tầng hay không. Nếu có thực hiện nhưng kích thước của tập dữ liệu không cho phép, nó sẽ sử dụng Cross-validation không phân tầng và thêm cảnh báo danh sách với thông báo cảnh báo.

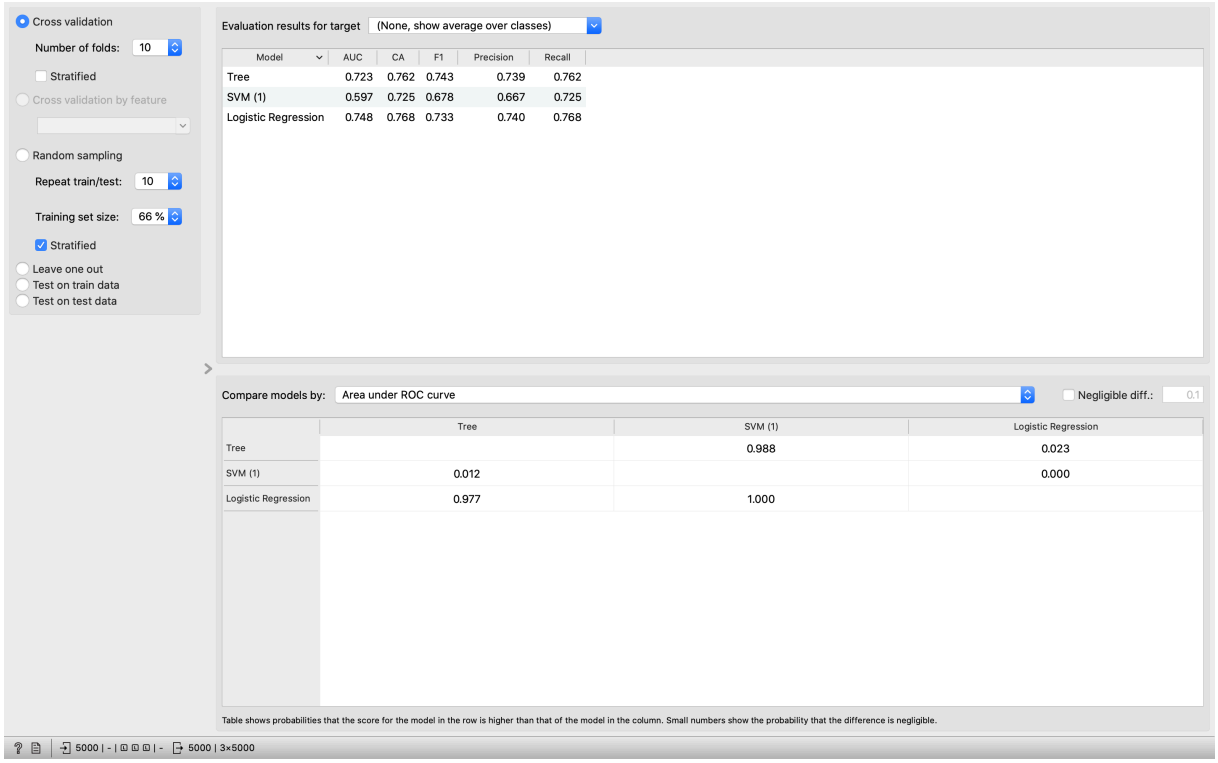
❖ ***Định nghĩa các thang đo:***

- **Accuracy**: tính tỷ lệ giữa số điểm dự đoán đúng và tổng số điểm trong tập dữ liệu kiểm thử, tuy nhiên không chỉ ra được cách mỗi loại được phân loại ra sao, lớp nào có tỷ lệ phân loại đúng nhiều nhất, và dữ liệu thuộc lớp nào thường bị nhầm lẫn vào lớp khác..
- **Precision** (độ chính xác): cho biết trong tổng số các mẫu được phân vào lớp thì có tỷ lệ bao nhiêu mẫu được phân lớp đúng (tránh nhầm lẫn với tính chính xác accuracy)
- **Recall** (độ phủ) còn gọi là độ phủ hay độ nhạy (sensitivity) hay TPR (True Positive Rate)
- **F_1-score**: giá trị trung bình điều hòa (harmonic mean) của hai độ đo Precision và Recall. F_1 có giá trị gần với giá trị nào nhỏ hơn giữa 2 giá trị Precision và Recall. F_1 sẽ có giá trị lớn nếu cả 2 giá trị Precision và Recall đều lớn.

▸ ***Precision và Recall thường dùng cho phân lớp nhị phân, F1-score, Micro F1 score, Macro F1 score dùng cho phân lớp đa lớp hoặc đa lớp đa nhãn***

- **ROC (Receiver Operating Characteristic)**: Là một đồ thị được sử dụng khá phổ biến trong đánh giá các mô hình phân loại nhị phân. Đường cong này được tạo ra bằng cách biểu diễn tỷ lệ dự báo true positive rate (TPR) dựa trên tỷ lệ dự báo false positive rate (FPR) tại các ngưỡng khác nhau. Một mô hình hiệu quả khi có FPR thấp và TPR cao, hay ROC càng tiệm cận với điểm (0;1) trong đồ thị thì mô hình càng hiệu quả
- **AUC (Area Under the Curve)**: Là diện tích nằm dưới đường cong ROC. Giá trị này là một số dương nhỏ hơn hoặc bằng 1. Giá trị này càng lớn thì mô hình càng tốt

- ***Tiến hành sử dụng phương pháp AUC (phần diện tích nằm dưới đường cong ROC) để so sánh các mô hình Cây quyết định, SVM, Logistic Regression***



Model	AUC	CA	F1	Precision	Recall
Tree	0.723	0.762	0.743	0.739	0.762
SVM (1)	0.597	0.725	0.678	0.667	0.725
Logistic Regression	0.748	0.768	0.733	0.740	0.768

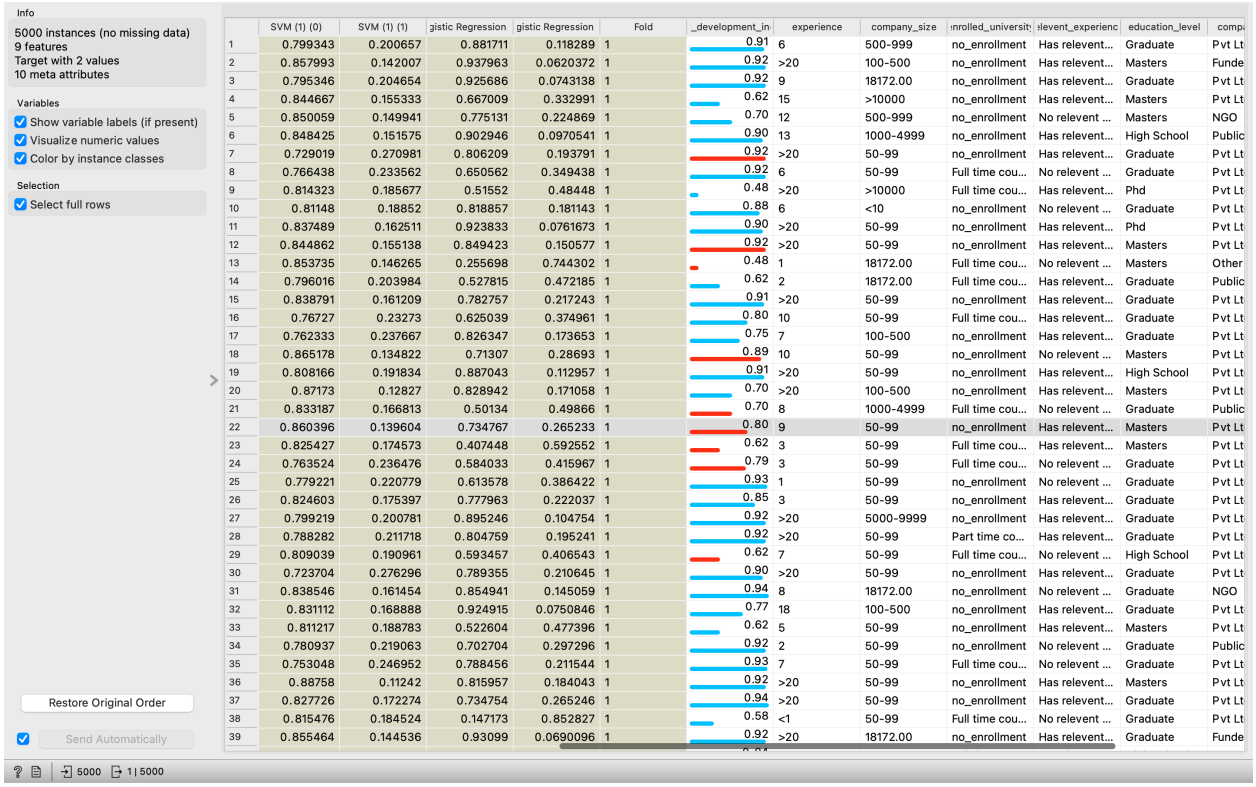
	Tree	SVM (1)	Logistic Regression
Tree		0.988	0.023
SVM (1)	0.012		0.000
Logistic Regression	0.977	1.000	

❖ ***Nhận xét***

Căn cứ vào thang đo AUC (Area under the curve) dùng để biểu diễn mức độ phân loại của mô hình, chỉ số AUC càng cao thì mô hình càng chính xác trong việc phân loại các lớp. Chỉ số AUC càng gần 1 thì mô hình càng phân loại chính xác, càng gần 0.5 thì hiệu suất phân loại càng tệ còn nếu gần 0 thì mô hình sẽ phân loại ngược kết quả (phân loại dương tính thành âm tính và ngược lại). ***Quan sát tiện ích Test and Score, ta nhận thấy chỉ số AUC của phương pháp Logistic Regression là cao nhất với AUC= 0.748, đồng thời gần 1 nhất so với phương pháp Tree và SVM. Vậy ta chọn Logistic Regression là phương pháp phù hợp nhất để dự báo cho bộ dữ liệu trên.***

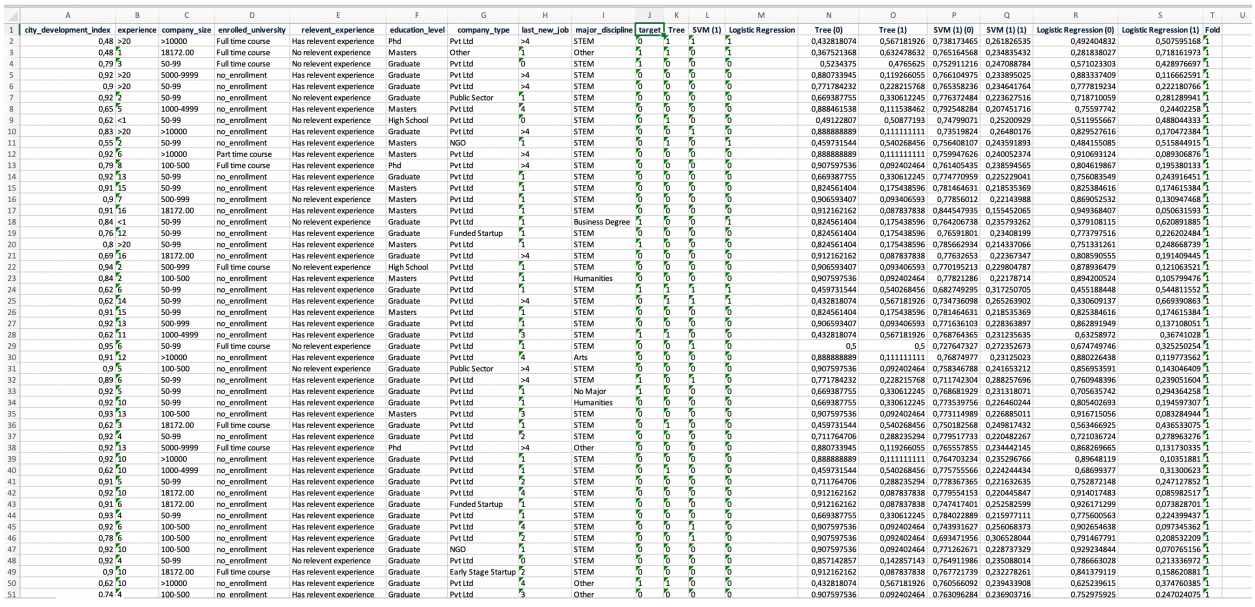
❖ *Bộ dữ liệu sau khi phân lớp*

- ***Quan sát từ Orange***



	SVM (1) (0)	SVM (1) (1)	gistic Regression	gistic Regression	Fold	_development_in	experience	company_size	nrolled_universit	elevent_experienc	education_level	compa
1	0.799343	0.200657	0.881711	0.118289	1	0.91	6	500-999	no_enrollment	Has relevent...	Graduate	Pvt Lt
2	0.857993	0.142007	0.937963	0.0620372	1	0.92	>20	100-500	no_enrollment	Has relevent...	Masters	Funde
3	0.795346	0.204654	0.925686	0.0743138	1	0.92	9	18172.00	no_enrollment	Has relevent...	Graduate	Pvt Lt
4	0.844667	0.155333	0.667009	0.332991	1	0.62	15	>10000	no_enrollment	Has relevent...	Masters	Pvt Lt
5	0.850059	0.149941	0.775131	0.224869	1	0.70	12	500-999	no_enrollment	No relevent ...	Masters	NGO
6	0.848425	0.151575	0.902946	0.0970541	1	0.90	13	1000-4999	no_enrollment	Has relevent...	High School	Public
7	0.729019	0.270981	0.806209	0.193791	1	0.92	>20	50-99	no_enrollment	Has relevent...	Graduate	Pvt Lt
8	0.766438	0.233562	0.650562	0.349438	1	0.92	6	50-99	Full time cou...	No relevent ...	Graduate	Pvt Lt
9	0.814323	0.185677	0.51552	0.48448	1	0.48	>20	>10000	Full time cou...	Has relevent...	Phd	Pvt Lt
10	0.81148	0.18852	0.818857	0.181143	1	0.88	6	<10	no_enrollment	No relevent ...	Graduate	Pvt Lt
11	0.837489	0.162511	0.923833	0.0761673	1	0.90	>20	50-99	no_enrollment	Has relevent...	Phd	Pvt Lt
12	0.844862	0.155138	0.849423	0.150577	1	0.92	>20	50-99	no_enrollment	Has relevent...	Masters	Pvt Lt
13	0.853735	0.146265	0.255698	0.744302	1	0.48	1	18172.00	Full time cou...	No relevent ...	Masters	Other
14	0.796016	0.203984	0.527815	0.472185	1	0.62	2	18172.00	Full time cou...	Has relevent...	Graduate	Public
15	0.838791	0.161209	0.782757	0.217243	1	0.91	>20	50-99	no_enrollment	Has relevent...	Graduate	Pvt Lt
16	0.76727	0.23273	0.625039	0.374961	1	0.80	10	50-99	Full time cou...	Has relevent...	Graduate	Pvt Lt
17	0.762333	0.237667	0.826347	0.173653	1	0.75	7	100-500	no_enrollment	Has relevent...	Graduate	Pvt Lt
18	0.865178	0.134822	0.71307	0.28693	1	0.89	10	50-99	no_enrollment	No relevent ...	Masters	Pvt Lt
19	0.808166	0.191834	0.887043	0.112957	1	0.91	>20	50-99	no_enrollment	Has relevent...	High School	Pvt Lt
20	0.87173	0.12827	0.828942	0.171058	1	0.70	>20	100-500	no_enrollment	Has relevent...	Masters	Pvt Lt
21	0.833187	0.166813	0.50134	0.49866	1	0.70	8	1000-4999	Full time cou...	No relevent ...	Graduate	Public
22	0.860396	0.139604	0.734767	0.265233	1	0.80	9	50-99	no_enrollment	Has relevent...	Masters	Pvt Lt
23	0.825427	0.174573	0.407448	0.592552	1	0.62	3	50-99	Full time cou...	Has relevent...	Masters	Pvt Lt
24	0.763524	0.236476	0.584033	0.415967	1	0.79	3	50-99	Full time cou...	No relevent ...	Graduate	Pvt Lt
25	0.779221	0.220779	0.613578	0.386422	1	0.93	1	50-99	no_enrollment	No relevent ...	Graduate	Pvt Lt
26	0.824603	0.175397	0.777963	0.222037	1	0.85	3	50-99	no_enrollment	Has relevent...	Graduate	Pvt Lt
27	0.799219	0.200781	0.895246	0.104754	1	0.92	>20	5000-9999	no_enrollment	Has relevent...	Graduate	Pvt Lt
28	0.788282	0.211718	0.804759	0.195241	1	0.92	>20	50-99	Part time co...	Has relevent...	Graduate	Pvt Lt
29	0.809039	0.190961	0.593457	0.406543	1	0.62	7	50-99	Full time cou...	No relevent ...	High School	Pvt Lt
30	0.723704	0.276296	0.789355	0.210645	1	0.90	>20	50-99	no_enrollment	Has relevent...	Graduate	Pvt Lt
31	0.838546	0.161454	0.854941	0.145059	1	0.94	8	18172.00	no_enrollment	No relevent ...	Graduate	NGO
32	0.831112	0.168888	0.924915	0.0750846	1	0.77	18	100-500	no_enrollment	Has relevent...	Graduate	Pvt Lt
33	0.811217	0.188783	0.522604	0.477396	1	0.62	5	50-99	no_enrollment	Has relevent...	Masters	Pvt Lt
34	0.780937	0.219063	0.702704	0.297296	1	0.92	2	50-99	no_enrollment	No relevent ...	Graduate	Public
35	0.753048	0.246952	0.788456	0.211544	1	0.93	7	50-99	Full time cou...	No relevent ...	Graduate	Pvt Lt
36	0.88758	0.11242	0.815957	0.184043	1	0.92	>20	50-99	no_enrollment	Has relevent...	Masters	Pvt Lt
37	0.827726	0.172274	0.734754	0.265246	1	0.94	>20	50-99	no_enrollment	Has relevent...	Graduate	Pvt Lt
38	0.815476	0.184524	0.147173	0.852827	1	0.58	<1	50-99	Full time cou...	No relevent ...	Graduate	Pvt Lt
39	0.855464	0.144536	0.93099	0.0690096	1	0.92	>20	18172.00	no_enrollment	Has relevent...	Graduate	Funde

- ***Quan sát từ Excel***

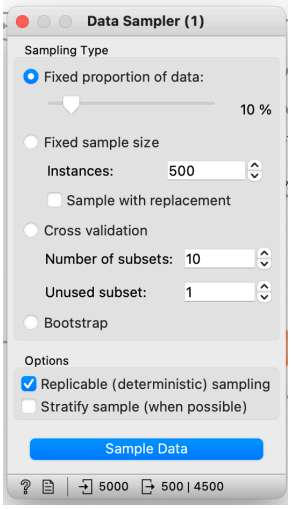


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	city_development_index	experience	company_size	enrolled_university	relevent_experience	education_level	company_type	last_new_job	major_discipline	target	Tree	SVM (1)	Logistic Regression	Tree (0)	Tree (1)	SVM (1) (0)	SVM (1) (1)	Logistic Regression (0)	Logistic Regression (1)	Fold	
2	0,48	>20	>10000	Full time course	Has relevent experience	Phd	Pvt Ltd	>4	STEM	0	1	1	1	0,432818074	0,567181926	0,738173465	0,261826535	0,492404832	0,507595168	1	
3	0,48	1	18172.00	Full time course	No relevent experience	Masters	Other	1	Other	1	1	0	1	0,367521368	0,632478632	0,765164568	0,234835432	0,281838027	0,718161973	1	
4	0,79	3	50-99	Full time course	No relevent experience	Graduate	Pvt Ltd	0	STEM	1	0	0	0	0,5234375	0,4765625	0,752911216	0,247088784	0,571023303	0,428976697	1	
5	0,92	>20	5000-9999	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	>4	STEM	0	0	0	0	0,880733945	0,119266055	0,766104975	0,233895025	0,883337409	0,116662591	1	
6	0,9	>20	50-99	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	>4	STEM	0	0	0	0	0,771784232	0,228215768	0,765358236	0,234641764	0,777819234	0,222180766	1	
7	0,92	2	50-99	no_enrollment	No relevent experience	Graduate	Public Sector	1	STEM	0	0	0	0	0,669387755	0,330612245	0,776372484	0,223627516	0,718710059	0,281289941	1	
8	0,65	5	1000-4999	no_enrollment	Has relevent experience	Masters	Pvt Ltd	4	STEM	0	0	0	0	0,888461538	0,111538462	0,792548284	0,207451716	0,7559742	0,24402258	1	
9	0,62	<1	50-99	no_enrollment	No relevent experience	High School	Pvt Ltd	0	STEM	0	1	0	0	0,49122807	0,50877193	0,74799071	0,25200929	0,511955667	0,488044333	1	
10	0,83	>20	>10000	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	>4	STEM	0	0	1	0	0,888888889	0,111111111	0,73519824	0,26480176	0,829527616	0,170472384	1	
11	0,55	5	50-99	no_enrollment	Has relevent experience	Masters	NGO	1	STEM	0	1	0	1	0,459731544	0,540268456	0,754081107	0,243591893	0,484155085	0,515844915	1	
12	0,92	6	>10000	Part time course	Has relevent experience	Masters	Pvt Ltd	>4	STEM	0	0	0	0	0,888888889	0,111111111	0,759947626	0,240052374	0,910693124	0,089306876	1	
13	0,79	8	100-500	Full time course	Has relevent experience	Phd	Pvt Ltd	>4	STEM	0	0	0	0	0,907597536	0,092402464	0,761405435	0,238594565	0,804619867	0,195380133	1	
14	0,92	13	50-99	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	1	STEM	0	0	0	0	0,669387755	0,330612245	0,774770959	0,225229041	0,756083549	0,243916451	1	
15	0,91	15	50-99	no_enrollment	Has relevent experience	Masters	Pvt Ltd	1	STEM	0	0	0	0	0,824561404	0,175438596	0,781464631	0,218535369	0,825384616	0,174615384	1	
16	0,9	7	500-999	no_enrollment	No relevent experience	Masters	Pvt Ltd	1	STEM	0	0	0	0	0,906593407	0,093406593	0,77856012	0,22143988	0,869052532	0,130947468	1	
17	0,91	16	18172.00	no_enrollment	Has relevent experience	Masters	Pvt Ltd	1	STEM	0	0	0	0	0,912162162	0,087837838	0,844547935	0,155452065	0,949368407	0,050631593	1	
18	0,84	<1	50-99	no_enrollment	No relevent experience	Graduate	Pvt Ltd	1	Business Degree	1	0	0	1	0,824561404	0,175438596	0,764206738	0,235793262	0,379108115	0,620891885	1	
19	0,76	12	50-99	no_enrollment	Has relevent experience	Graduate	Funded Startup	1	STEM	0	0	0	0	0,824561404	0,175438596	0,76591801	0,23408199	0,773797516	0,226202484	1	
20	0,8	>20	50-99	no_enrollment	Has relevent experience	Masters	Pvt Ltd	1	STEM	1	0	0	0	0,824561404	0,175438596	0,785662934	0,214337066	0,751331261	0,248668739	1	
21	0,69	16	18172.00	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	>4	STEM	0	0	0	0	0,912162162	0,087837838	0,77632653	0,22367347	0,808590555	0,191409445	1	
22	0,94	2	500-999	Full time course	No relevent experience	High School	Pvt Ltd	1	STEM	0	0	0	0	0,906593407	0,093406593	0,770195213	0,229804787	0,878936479	0,121063521	1	
23	0,84	2	100-500	no_enrollment	Has relevent experience	Masters	Pvt Ltd	1	Humanities	0	0	0	0	0,907597536	0,092402464	0,77821286	0,22178714	0,894200524	0,105799476	1	
24	0,62	6	50-99	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	1	STEM	1	1	1	1	0,459731544	0,540268456	0,682749295	0,317250705	0,455188448	0,544811552	1	
25	0,62	14	50-99	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	>4	STEM	0	1	1	1	0,432818074	0,567181926	0,734736098	0,265263902	0,330609137	0,669390863	1	
26	0,91	15	50-99	no_enrollment	Has relevent experience	Masters	Pvt Ltd	1	STEM	0	0	0	0	0,824561404	0,175438596	0,781464631	0,218535369	0,825384616	0,174615384	1	
27	0,92	13	500-999	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	1	STEM	0	0	0	0	0,906593407	0,093406593	0,771636103	0,228363897	0,862891949	0,137108051	1	
28	0,62	11	1000-4999	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	5	STEM	1	1	0	0	0,432818074	0,567181926	0,768764365	0,231235635	0,63258972	0,36741028	1	
29	0,95	6	50-99	Full time course	No relevent experience	Graduate	Pvt Ltd	1	STEM	0	0	1	0	0,5	0,5	0,727647327	0,272352673	0,674749746	0,325250254	1	
30	0,91	12	>10000	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	4	Arts	0	0	0	0	0,888888889	0,111111111	0,76874977	0,23125023	0,880226438	0,119773562	1	
31	0,9	5	100-500	no_enrollment	No relevent experience	Graduate	Public Sector	>4	STEM	0	0	0	0	0,907597536	0,092402464	0,758346788	0,241653212	0,856953591	0,143046409	1	
32	0,89	6	50-99	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	>4	STEM	1	0	1	0	0,771784232	0,228215768	0,711742304	0,288257696	0,760948396	0,239051604	1	
33	0,92	5	50-99	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	1	No Major	1	0	0	0	0,669387755	0,330612245	0,768681929	0,231318071	0,705635742	0,294364258	1	
34	0,92	10	50-99	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	1	Humanities	0	0	0	0	0,669387755	0,330612245	0,773539756	0,226460244	0,805402693	0,194597307	1	
35	0,93	13	100-500	no_enrollment	Has relevent experience	Masters	Pvt Ltd	5	STEM	0	0	0	0	0,907597536	0,092402464	0,773114989	0,226885011	0,916715056	0,083284944	1	
36	0,62	5	18172.00	Full time course	Has relevent experience	Graduate	Pvt Ltd	1	STEM	0	1	0	0	0,459731544	0,540268456	0,750182568	0,249817432	0,563466925	0,436533075	1	
37	0,92	4	50-99	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	2	STEM	0	0	0	0	0,711764706	0,288235294	0,779517733	0,220482267	0,721036724	0,278963276	1	
38	0,92	13	5000-9999	Full time course	Has relevent experience	Phd	Pvt Ltd	>4	Other	0	0	0	0	0,880733945	0,119266055	0,765557855	0,234442145	0,868269665	0,131730335	1	
39	0,92	10	>10000	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	1	STEM	0	0	0	0	0,888888889	0,111111111	0,764703234	0,235296766	0,89648119	0,10351881	1	
40	0,62	10	1000-4999	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	1	STEM	0	1	0	0	0,459731544	0,540268456	0,775755566	0,224244434	0,68699377	0,31300623	1	
41	0,91	5	50-99	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	5	STEM	0	0	0	0	0,711764706	0,288235294	0,778367365	0,221632635	0,752872148	0,247127852	1	
42	0,92	10	18172.00	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	4	STEM	0	0	0	0	0,912162162	0,087837838	0,779554153	0,220445847	0,914017483	0,085982517	1	
43	0,91	6	18172.00	no_enrollment	Has relevent experience	Graduate	Funded Startup	1	STEM	0	0	0	0	0,912162162	0,087837838	0,747417401	0,252582599	0,926171299	0,073828701	1	
44	0,93	4	50-99	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	1	STEM	0	0	0	0	0,669387755	0,330612245	0,784022889	0,215977111	0,775600563	0,224399437	1	
45	0,92	6	100-500	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	4	STEM	0	0	1	0	0,907597536	0,092402464	0,743931627	0,256068373	0,902654638	0,097345362	1	
46	0,78	6	100-500	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	2	STEM	0	0	1	0	0,907597536	0,092402464	0,693471956	0,306528044	0,791467791	0,208532209	1	
47	0,92	10	100-500	no_enrollment	Has relevent experience	Graduate	NGO	1	STEM	0	0	0	0	0,907597536	0,092402464	0,771262671	0,228737329	0,929234844	0,070765156	1	
48	0,92	4	50-99	no_enrollment	No relevent experience	Graduate	Pvt Ltd	0	STEM	0	0	0	0	0,857142857	0,142857143	0,764911986	0,235088014	0,786663028	0,213336972	1	
49	0,9	10	18172.00	Full time course	Has relevent experience	Graduate	Early Stage Startup	2	STEM	0	0	0	0	0,912162162	0,087837838	0,767721739	0,232278261	0,841379119	0,158620881	1	
50	0,62	10	>10000	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	4	Other	1	1	0	0	0,432818074	0,567181926	0,760566092	0,239433908	0,625239615	0,374760385	1	
51	0,74	4	100-500	no_enrollment	Has relevent experience	Graduate	Pvt Ltd	3	Other	0	0	0	0	0,907597536	0,092402464	0,763096284	0,236903716	0,752975925	0,247024075	1	

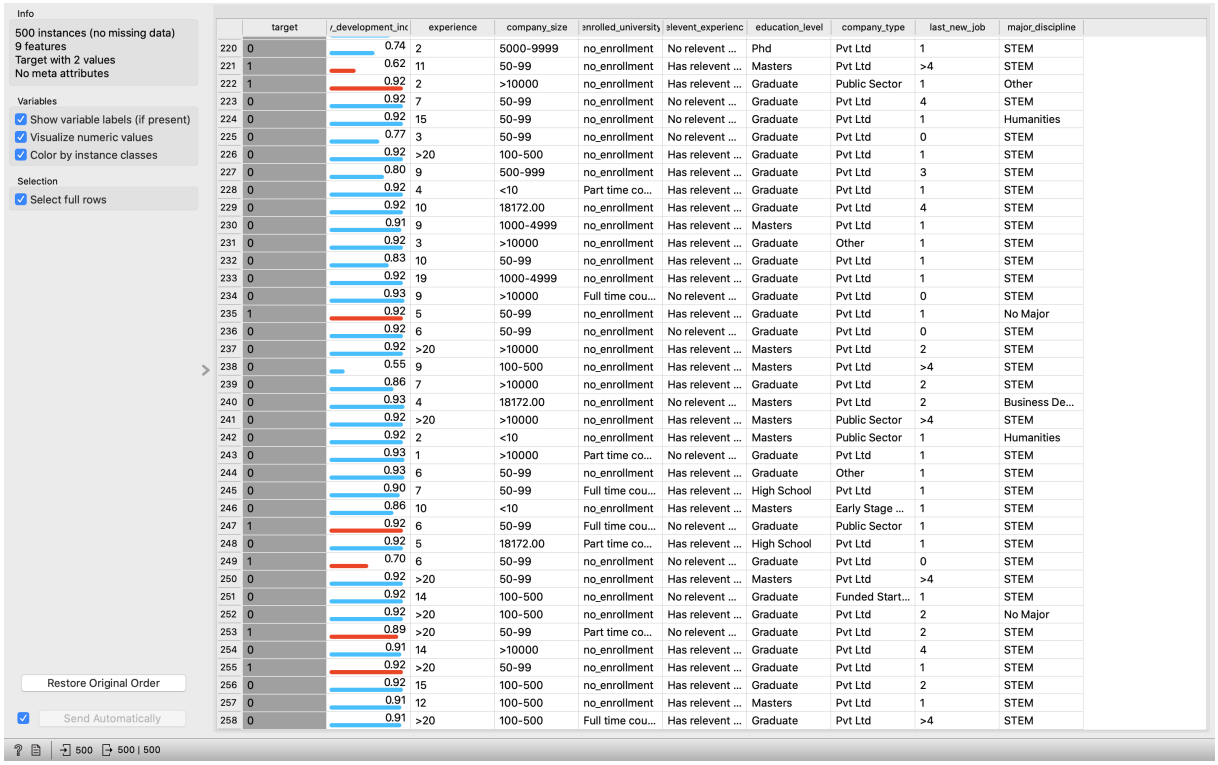
3. Dự báo từ kết quả phân lớp dữ liệu đạt được.

3.1. Trích xuất 10% dữ liệu

Sau khi đã đánh giá các phương pháp, ta chọn ngẫu nhiên 500 ứng viên (10% trong bộ dữ liệu) để tiến hành dự báo. Sử dụng tiện ích **Data Sampler** tương tự bước tiền xử lý, tuy nhiên lần này số mẫu lấy ra ngẫu nhiên chỉ còn 500.



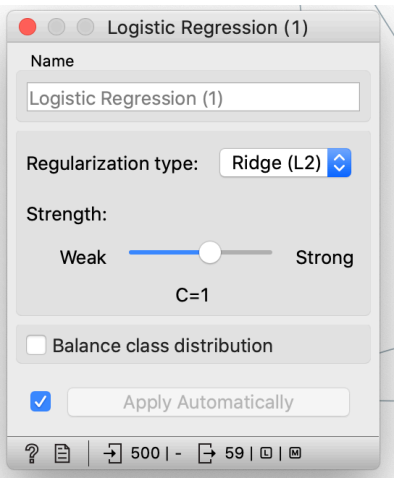
❖ ***Ta lưu tập dữ liệu chỉ gồm 500 dữ liệu***



	target	r_development_inc	experience	company_size	enrolled_university	relevent_experienc	education_level	company_type	last_new_job	major_discipline
220	0	0.74	2	5000-9999	no_enrollment	No relevent ...	Phd	Pvt Ltd	1	STEM
221	1	0.62	11	50-99	no_enrollment	Has relevent ...	Masters	Pvt Ltd	>4	STEM
222	1	0.92	2	>10000	no_enrollment	Has relevent ...	Graduate	Public Sector	1	Other
223	0	0.92	7	50-99	no_enrollment	No relevent ...	Graduate	Pvt Ltd	4	STEM
224	0	0.92	15	50-99	no_enrollment	No relevent ...	Graduate	Pvt Ltd	1	Humanities
225	0	0.77	3	50-99	no_enrollment	No relevent ...	Graduate	Pvt Ltd	0	STEM
226	0	0.92	>20	100-500	no_enrollment	Has relevent ...	Graduate	Pvt Ltd	1	STEM
227	0	0.80	9	500-999	no_enrollment	Has relevent ...	Graduate	Pvt Ltd	3	STEM
228	0	0.92	4	<10	Part time co...	Has relevent ...	Graduate	Pvt Ltd	1	STEM
229	0	0.92	10	18172.00	no_enrollment	Has relevent ...	Graduate	Pvt Ltd	4	STEM
230	0	0.91	9	1000-4999	no_enrollment	Has relevent ...	Masters	Pvt Ltd	1	STEM
231	0	0.92	3	>10000	no_enrollment	Has relevent ...	Graduate	Other	1	STEM
232	0	0.83	10	50-99	no_enrollment	Has relevent ...	Graduate	Pvt Ltd	1	STEM
233	0	0.92	19	1000-4999	no_enrollment	Has relevent ...	Graduate	Pvt Ltd	1	STEM
234	0	0.93	9	>10000	Full time cou...	No relevent ...	Graduate	Pvt Ltd	0	STEM
235	1	0.92	5	50-99	no_enrollment	Has relevent ...	Graduate	Pvt Ltd	1	No Major
236	0	0.92	6	50-99	no_enrollment	No relevent ...	Graduate	Pvt Ltd	0	STEM
237	0	0.92	>20	>10000	no_enrollment	Has relevent ...	Masters	Pvt Ltd	2	STEM
238	0	0.55	9	100-500	no_enrollment	Has relevent ...	Masters	Pvt Ltd	>4	STEM
239	0	0.86	7	>10000	no_enrollment	Has relevent ...	Graduate	Pvt Ltd	2	STEM
240	0	0.93	4	18172.00	no_enrollment	No relevent ...	Masters	Pvt Ltd	2	Business De...
241	0	0.92	>20	>10000	no_enrollment	Has relevent ...	Masters	Public Sector	>4	STEM
242	0	0.92	2	<10	no_enrollment	Has relevent ...	Masters	Public Sector	1	Humanities
243	0	0.93	1	>10000	Part time co...	No relevent ...	Graduate	Pvt Ltd	1	STEM
244	0	0.93	6	50-99	no_enrollment	Has relevent ...	Graduate	Other	1	STEM
245	0	0.90	7	50-99	Full time cou...	Has relevent ...	High School	Pvt Ltd	1	STEM
246	0	0.86	10	<10	no_enrollment	Has relevent ...	Masters	Early Stage ...	1	STEM
247	1	0.92	6	50-99	Full time cou...	No relevent ...	Graduate	Public Sector	1	STEM
248	0	0.92	5	18172.00	Part time co...	Has relevent ...	High School	Pvt Ltd	1	STEM
249	1	0.70	6	50-99	no_enrollment	No relevent ...	Graduate	Pvt Ltd	0	STEM
250	0	0.92	>20	50-99	no_enrollment	Has relevent ...	Masters	Pvt Ltd	>4	STEM
251	0	0.92	14	100-500	no_enrollment	No relevent ...	Graduate	Funded Start...	1	STEM
252	0	0.92	>20	100-500	no_enrollment	Has relevent ...	Graduate	Pvt Ltd	2	No Major
253	1	0.89	>20	50-99	Part time co...	No relevent ...	Graduate	Pvt Ltd	2	STEM
254	0	0.91	14	>10000	no_enrollment	Has relevent ...	Graduate	Pvt Ltd	4	STEM
255	1	0.92	>20	50-99	no_enrollment	Has relevent ...	Graduate	Pvt Ltd	1	STEM
256	0	0.92	15	100-500	no_enrollment	Has relevent ...	Graduate	Pvt Ltd	2	STEM
257	0	0.91	12	100-500	no_enrollment	Has relevent ...	Masters	Pvt Ltd	1	STEM
258	0	0.91	>20	100-500	Full time cou...	Has relevent ...	Graduate	Pvt Ltd	>4	STEM

✶ Phân lớp dữ liệu mới theo thuật toán Logistic Regression

Đưa bộ dữ liệu vừa trích xuất vào tiện ích Logistic Regression để tiến hành phân lớp theo các thông số như sau:

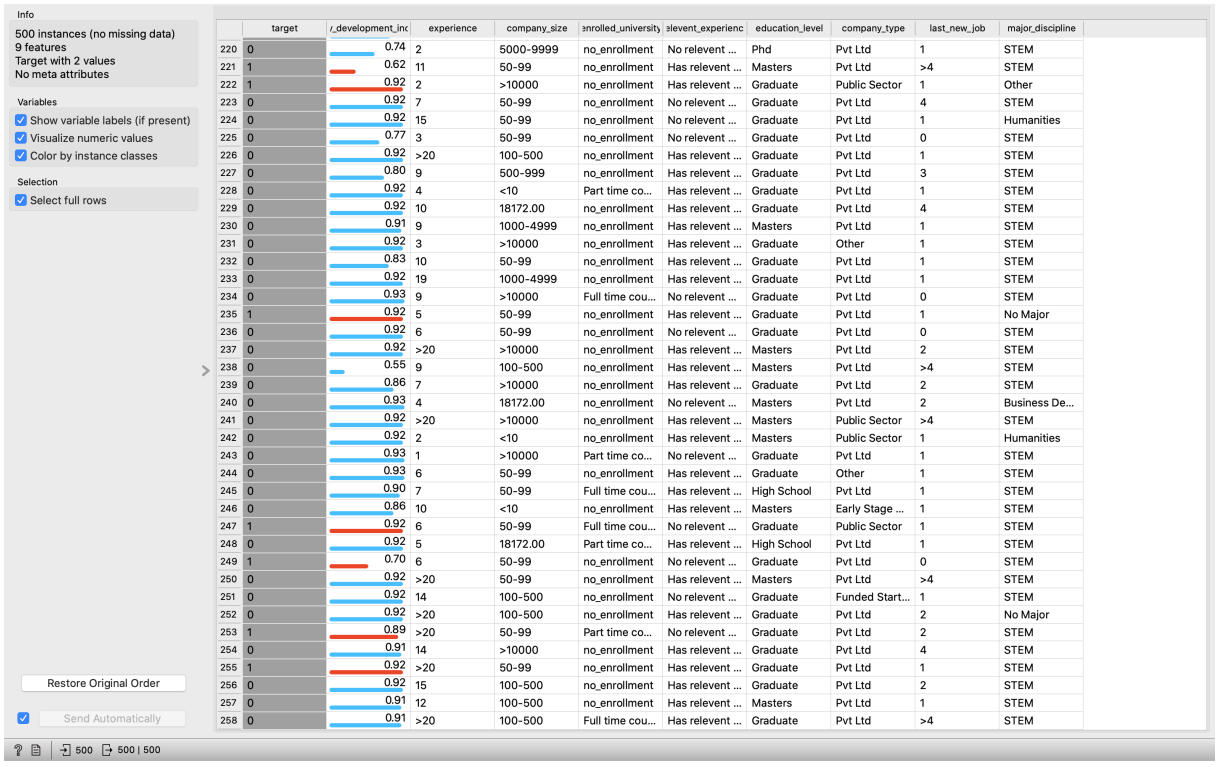


3.2. Predictions

Sử dụng **Prediction** để hiển thị mô hình dự báo của bộ dữ liệu. Gồm:

- **Predicted class**: dạng xem cung cấp thông tin về lớp được dự đoán.
- **Predicted probabilities**: chế độ xem cung cấp thông tin về các xác suất được dự đoán bởi (các) bộ phân loại.
- **Draw distribution bars**: cung cấp hình ảnh trực quan về các xác suất.

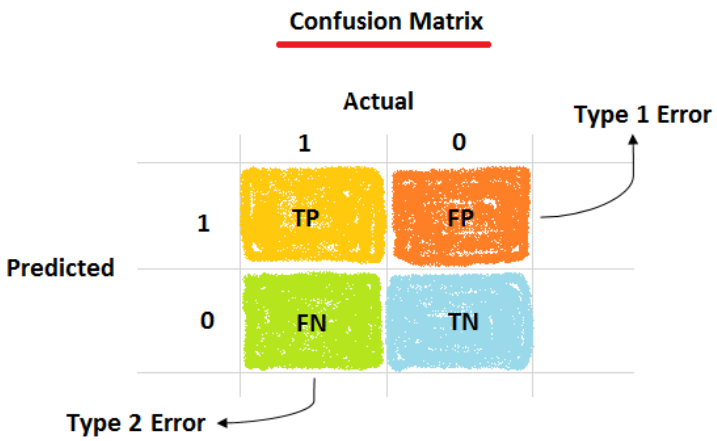
Kết hợp với mô hình hồi quy Logistic Regression được lựa chọn là phương pháp tốt nhất để dự báo cho bộ dữ liệu trên, và kết quả đạt được như sau:



	target	r_development_inc	experience	company_size	enrolled_university	relevent_experienc	education_level	company_type	last_new_job	major_discipline
220	0	0.74	2	5000-9999	no_enrollment	No relevent ...	Phd	Pvt Ltd	1	STEM
221	1	0.62	11	50-99	no_enrollment	Has relevent ...	Masters	Pvt Ltd	>4	STEM
222	1	0.92	2	>10000	no_enrollment	Has relevent ...	Graduate	Public Sector	1	Other
223	0	0.92	7	50-99	no_enrollment	No relevent ...	Graduate	Pvt Ltd	4	STEM
224	0	0.92	15	50-99	no_enrollment	No relevent ...	Graduate	Pvt Ltd	1	Humanities
225	0	0.77	3	50-99	no_enrollment	No relevent ...	Graduate	Pvt Ltd	0	STEM
226	0	0.92	>20	100-500	no_enrollment	Has relevent ...	Graduate	Pvt Ltd	1	STEM
227	0	0.80	9	500-999	no_enrollment	Has relevent ...	Graduate	Pvt Ltd	3	STEM
228	0	0.92	4	<10	Part time co...	Has relevent ...	Graduate	Pvt Ltd	1	STEM
229	0	0.92	10	18172.00	no_enrollment	Has relevent ...	Graduate	Pvt Ltd	4	STEM
230	0	0.91	9	1000-4999	no_enrollment	Has relevent ...	Masters	Pvt Ltd	1	STEM
231	0	0.92	3	>10000	no_enrollment	Has relevent ...	Graduate	Other	1	STEM
232	0	0.83	10	50-99	no_enrollment	Has relevent ...	Graduate	Pvt Ltd	1	STEM
233	0	0.92	19	1000-4999	no_enrollment	Has relevent ...	Graduate	Pvt Ltd	1	STEM
234	0	0.93	9	>10000	Full time cou...	No relevent ...	Graduate	Pvt Ltd	0	STEM
235	1	0.92	5	50-99	no_enrollment	Has relevent ...	Graduate	Pvt Ltd	1	No Major
236	0	0.92	6	50-99	no_enrollment	No relevent ...	Graduate	Pvt Ltd	0	STEM
237	0	0.92	>20	>10000	no_enrollment	Has relevent ...	Masters	Pvt Ltd	2	STEM
238	0	0.55	9	100-500	no_enrollment	Has relevent ...	Masters	Pvt Ltd	>4	STEM
239	0	0.86	7	>10000	no_enrollment	Has relevent ...	Graduate	Pvt Ltd	2	STEM
240	0	0.93	4	18172.00	no_enrollment	No relevent ...	Masters	Pvt Ltd	2	Business De...
241	0	0.92	>20	>10000	no_enrollment	Has relevent ...	Masters	Public Sector	>4	STEM
242	0	0.92	2	<10	no_enrollment	Has relevent ...	Masters	Public Sector	1	Humanities
243	0	0.93	1	>10000	Part time co...	No relevent ...	Graduate	Pvt Ltd	1	STEM
244	0	0.93	6	50-99	no_enrollment	Has relevent ...	Graduate	Other	1	STEM
245	0	0.90	7	50-99	Full time cou...	Has relevent ...	High School	Pvt Ltd	1	STEM
246	0	0.86	10	<10	no_enrollment	Has relevent ...	Masters	Early Stage ...	1	STEM
247	1	0.92	6	50-99	Full time cou...	No relevent ...	Graduate	Public Sector	1	STEM
248	0	0.92	5	18172.00	Part time co...	Has relevent ...	High School	Pvt Ltd	1	STEM
249	1	0.70	6	50-99	no_enrollment	No relevent ...	Graduate	Pvt Ltd	0	STEM
250	0	0.92	>20	50-99	no_enrollment	Has relevent ...	Masters	Pvt Ltd	>4	STEM
251	0	0.92	14	100-500	no_enrollment	No relevent ...	Graduate	Funded Start...	1	STEM
252	0	0.92	>20	100-500	no_enrollment	Has relevent ...	Graduate	Pvt Ltd	2	No Major
253	1	0.89	>20	50-99	Part time co...	No relevent ...	Graduate	Pvt Ltd	2	STEM
254	0	0.91	14	>10000	no_enrollment	Has relevent ...	Graduate	Pvt Ltd	4	STEM
255	1	0.92	>20	50-99	no_enrollment	Has relevent ...	Graduate	Pvt Ltd	1	STEM
256	0	0.92	15	100-500	no_enrollment	Has relevent ...	Graduate	Pvt Ltd	2	STEM
257	0	0.91	12	100-500	no_enrollment	Has relevent ...	Masters	Pvt Ltd	1	STEM
258	0	0.91	>20	100-500	Full time cou...	Has relevent ...	Graduate	Pvt Ltd	>4	STEM

3.3. Các ước lượng và bảng ma trận nhầm lẫn

Confusion Matrix (ma trận nhầm lẫn hay ma trận lỗi) là một bố cục bảng cụ thể cho phép hình dung hiệu suất của một thuật toán, nơi mà kết quả truy xuất ra có thể là một bảng gồm đối tượng chia thành 2 hoặc nhiều hơn 2 phân lớp. Bảng Confusion Matrix cho thấy 4 sự kết hợp khác nhau gồm 2 thành phần chính là: predicted (dự báo) và actual values (giá trị thực tế). Hình ảnh bảng ma trận nhầm lẫn:



❖ ***Phân tích từ bảng ta có True - False và Positive - Negative mang ý nghĩa:***

- **True/False:** những gì ta dự đoán sẽ xảy ra (True) hoặc sẽ không xảy ra (False)
- **Positive/Negative:** những gì ta dự đoán sẽ xảy ra và thực tế là hoàn toàn đúng (Positive) hoặc dự đoán sẽ không xảy ra và thực tế đúng là không xảy ra (Negative)

✳ ***Các trường hợp True***

- **TP - True Positive:** Dự đoán ta đoán là xảy ra (Positive) và thực tế là hoàn toàn xảy ra (True Positive)
- **TN - True Negative:** Dự đoán ta đoán là không xảy ra (Negative) và thực tế đúng là không xảy ra (True Negative)

▸ ***Với "True" thì tất cả những dự đoán ta đoán đều đúng với thực tế cho thấy***

✳ **Các trường hợp False**

- FP - False Positive: Dự đoán ta đoán là sẽ xảy ra (Positive) và thực tế là sai, không xảy ra (False Positive)
- FN - False Negative: Dự đoán ta đoán là không xảy ra (Negative) và thực tế là sai, đã xảy ra (False Negative)

▸ ***Với "False" thì có nghĩa những gì ta dự đoán đều sai với thực tế cho thấy***

▸ ***FP và FN trong thống kê còn được xem là Xác suất mắc sai lầm loại I và Xác suất mắc sai lầm loại II***

Bên cạnh đó, bảng ma trận nhầm lẫn Confusion Matrix này thể hiện những số lượng ước tính tương đương với số liệu kết quả tính được từ phương pháp Logistic Regression để ước lượng cho Recall, Precision, Specificity, Accuracy và AUR-ROC từ đó, ta có thể phân tích dự báo chênh lệch giữa Predicted và Actual.

❖ ***Cơ sở đưa ra dự báo ước lượng từ các công thức sau***

- **Precision:** tỷ lệ giữa dự đoán ta đoán sẽ xảy ra là đúng sự thật sẽ xảy ra trong thực tế

$$\textbf{Precision} = \frac{TP}{TP + FP}$$

- **Recall Sensitivity:** trong số những trường hợp đã xảy ra trong thực tế thì có bao nhiêu trường hợp ta dự đoán đúng được thực hiện bởi mô hình của ta ứng dụng

$$\textbf{Recall} = \frac{TP}{TP + FN}$$

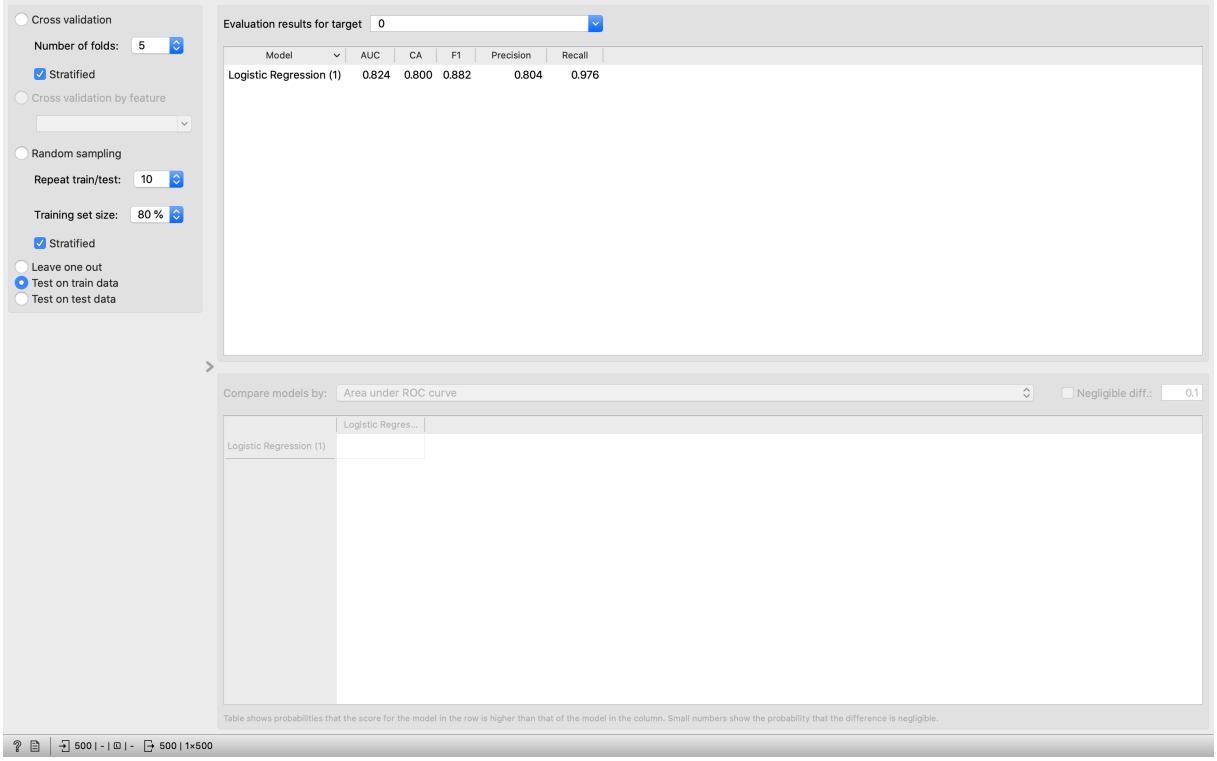
- **Accuracy:** tỷ lệ dự đoán đúng trong số tất cả những dự đoán của chúng ta đặt ra. Nhưng Accuracy chỉ thật sự đúng để đánh giá mô hình khi dự đoán ta đưa ra so với thực tế là gần như bằng nhau.

$$\textbf{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

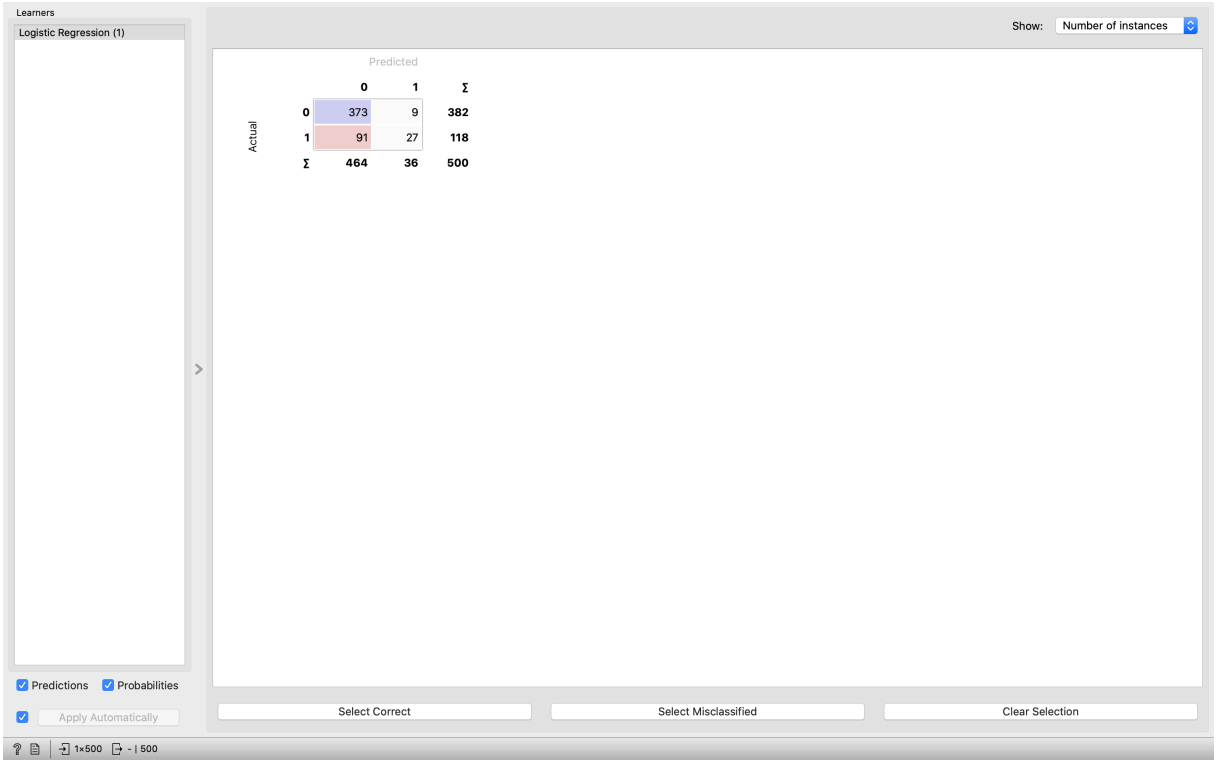
- **Specificity:** cho biết có bao nhiêu kết quả là không đến từ dự đoán của chúng ta. Và ta chọn specificity khi chúng ta quan tâm đến giá trị của TN (True Negative) và không muốn cảnh báo sai về những trường hợp FP.

❖ ***Quan sát Test and Score sthu được từ bộ dữ liệu Job Change từ Orange Data Mining***

Với 500 dữ liệu rút trích 10% đem đi dự báo trên, ước tính qua chỉ riêng Logistic Regression (phương pháp hiệu quả nhất được thấy thông qua Test and Score), ta nối đồng thời dự báo 500 dữ liệu qua Test and Score, chọn **Test on train data** và ước lượng target muốn đạt được set = 0 để thấy các mức độ Recall, Precision, Accuracy tính toán tương ứng:



❖ ***Quan sát Matrix Confusion thu được từ 500 dữ liệu***



4. Phân tích rút ra từ dự báo mô hình 10% dữ liệu

❖ ***Quan sát từ bảng ma trận với số lượng mẫu là 500 ứng viên, ta rút ra các kết quả sau:***

	Dự đoán không tìm việc mới	**Dự đoán tìm việc mới**
Thực tế không tìm việc mới	373 (TP)	9 (FP)
Thực tế tìm việc mới	91 (FN)	27 (TN)

- Có **373 ứng viên** được **dự báo đúng** là **không có xu hướng tìm việc mới** (True Positive)
- Có **91 ứng viên** được **dự báo sai** là **sẽ không tìm việc mới** nhưng thực tế là họ **có xu hướng tìm việc mới** (False Negative)
- Có **9 ứng viên** được **dự báo sai** là **sẽ tìm việc mới** nhưng thực tế họ lại **không có xu hướng tìm việc mới**. (False Positive)
- Có **27 ứng viên** ta **dự báo đúng** là **có xu hướng tìm việc mới** và thực tế là họ **có xu hướng tìm công việc mới** (True Negative)
  - *Trong ma trận nhầm lẫn này, trong số **382 mẫu thực tế không tìm việc mới**, hệ thống đã đánh giá rằng có **373/382 mẫu khớp với thực tế** (9 mẫu dự báo sai thực tế) và trong số **118 mẫu thực tế** có xu hướng đi tìm việc mới thì chỉ có **27/118 mẫu khớp với thực tế đưa ra** (91 mẫu dự báo sai thực tế)*

❖ ***Quan sát từ Test Score, ta rút ra các kết luận sau:***

$$\textbf{Recall} = \frac{TP}{TP + FN} = \frac{373}{373 + 91} = 0.8$$

$$\textbf{Precision} = \frac{TP}{TP + FP} = \frac{373}{373 + 9} = 0.976$$

$$\textbf{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{373 + 27}{500} = 0.8$$

Ta thấy, **Precison cao hơn Recall (0.976 > 0.8)** và căn cứ theo định nghĩa, Precision cho ta biết những dự đoán của chúng ta có độ chính xác là bao nhiêu so với thực tế. Do đó, ta thấy có **373 là True Positive ứng viên đã được dự đoán chính xác** trên tổng số 382 ứng viên được dự đoán sẽ không có xu hướng tìm công việc mới trên thực tế. Vậy nên, ta đã dự đoán đúng 373 trường hợp này **tương ứng mức độ chính xác là 97,6%**. Và chỉ có **9 trường hợp là False Positive ứng viên đã bị dự đoán sai**, tức là họ được dự đoán sẽ đi tìm công việc mới nhưng trên thực tế thì họ lại không có xu hướng tìm việc mới. Vậy nên, ta đã dự đoán sai 9 trường hợp này **tương ứng mức độ sai lầm là 2,6% (Sai lầm loại I)**.

- ***Ta sẽ có 2,6% quyết định là sai khi không thu nhận 9 ứng viên này vào làm dù họ không có xu hướng tìm công việc mới mà thật sự mong muốn gắn bó với công việc này***

dài lâu. Tuy nhiên, mức độ sai phạm là rất nhỏ bé so với những dự đoán chính xác ta có được từ mô hình so với thực tế khi ta thu nhận 373 ứng viên thật sự có mong muốn gắn bó với công việc nhà khoa học dữ liệu mà công ty đang tuyển dụng. Có nghĩa là dự báo ta đưa ra mức độ chinh xác là cao với thực tế.

Recall thấp hơn Precision và căn cứ theo định nghĩa Recall cho biết tỷ lệ dự đoán với mô hình ta sử dụng. Do đó, ta dự đoán có **373 là True Positive ứng viên đã được dự đoán chính xác** trên tổng số 464 trường hợp trong mô hình sẽ không đi tìm công việc mới trên thực tế. Vậy nên, ta đã dự đoán đúng 373 trường hợp **tương ứng mức độ chính xác là 80,3%**. Và chỉ có **91 trường hợp là False Negative ứng viên bị dự đoán sai**, tức là họ có dự định không muốn làm việc lâu dài với công ty nhưng công ty lại dự báo là họ thật sự muốn làm việc lâu dài ở vị trí nhà khoa học dữ liệu. Vậy nên, ta đã dự đoán sai 91 trường hợp này **tương ứng mức độ sai lầm là 19,7% (Sai lầm loại II)**

- ***Ta sẽ có 19,7% quyết định là sai khi thu nhận 91 ứng viên có xu hướng tìm việc mới sau một thời gian làm ở vị trí nhà khoa học dữ liệu tại công ty. Tuy nhiên ta thấy 19,7% dự báo sai vẫn ít hơn so với 80,3% dự báo đúng. Có nghĩa là mô hình ta đang áp dụng là đúng đắn và có tính hiệu quả dùng để dự báo.***

Bên cạnh đó, Accuracy không được sử dụng trong mô hình phân lớp với dữ liệu được rút trích 10% vì Accuracy theo góc nhìn của CM, chỉ nên sử dụng trong trường hợp lượng người được khảo sát có và không có xu hướng gắn bó với công việc dài lâu là ngang bằng nhau. Nếu như số lượng người được khảo sát có sự cân bằng giữa thực tế và dự báo, thì ta sẽ sử dụng Accuracy để đánh giá mức độ hiệu quả của mô hình cũng như độ chính xác so với thực tế mà không cần đến Precision hay Recall.

CHƯƠNG V. HƯỚNG PHÁT TRIỂN ĐỀ TÀI

Với bộ dữ liệu Job Changes Of Data Scientist được xử lý thông qua phần mềm Orange Data Mining bằng phương pháp phân cụm và phân lớp. Kết quả sau cùng ta thu được là dự báo liên quan đến xác suất trong các nhân viên ứng tuyển vị trí nhà khoa học dữ liệu thì có bao nhiêu nhân viên thực sự mong muốn được làm lâu dài tại vị trí ứng tuyển của công ty. Ta thấy, với 10 % dữ liệu được trích xuất từ 5000 dữ liệu, cứ 500 ứng viên ứng tuyển thì sẽ có 373 ứng viên được dự đoán là không có xu hướng tìm công việc mới thực sự có mong muốn được hợp tác lâu dài tại vị trí Data Scientists. Ta chỉ sai lệch 2,4% (Sai lầm loại 1) dẫn đến vô tình loại nhầm 9 ứng viên thực sự mong muốn làm việc lâu dài nhưng ta dự đoán họ có xu hướng tìm việc mới. Trên 500 ứng viên đó thì ta mắc sai lầm loại II khi ta nhận 91 ứng viên ta dự đoán không tìm công việc mới nhưng thực tế là họ tìm công việc mới. ***Như vậy với 10% ứng viên trích xuất ra, ta sẽ nhận chính xác 373 ứng viên và nhận lầm 91 ứng viên, từ chối 36 ứng viên trong đó có 9 ứng viên là từ chối nhầm và 27 ứng viên là từ chối đúng thực tế. Precision và Recall đóng vai trò quan trọng trong dự báo này vì nó cho thấy những gì ta dự báo ở mô hình so với thực tế là chính xác với tỷ lệ cao, mô hình đang sử dụng là đúng đắn.***

Nếu xét trên mô hình gồm 19 800 nghìn dữ liệu, ta sẽ thấy đa số các ứng viên khi tham gia ứng tuyển vị trí công ty có mong muốn làm việc lâu dài với công ty hơn là đi tìm công việc mới (14 382/19 800 có target = 0 ứng viên không có xu hướng tìm việc mới). Tuy nhiên, với mô hình ta đưa ra, các ứng viên được nhận vào làm vẫn sẽ có tình trạng nhận lầm do dự báo sai thực tế, dẫn đến việc họ không thật sự muốn làm việc lâu dài nhưng lại đậu ứng tuyển và trong tương lai, những ứng viên này sẽ có tác động đến số lượng nhân sự công ty khi chẳng may họ xin nghỉ việc để tìm công việc/công ty mới. Muốn giảm thiểu các tác động liên quan đến nhân sự như đề cập, ta cần phân tích về một trong các yếu tố sẽ ảnh hưởng đến một ứng viên với quyết định làm lâu dài từ khảo sát đưa ra. Có tất cả 9 thuộc tính thu được gồm: chỉ số phát triển của thành phố đang sinh sống, giới tính, các kinh nghiệm liên quan đến vị trí, trình độ học vấn, khoá học tham gia ở đại học, chuyên ngành đăng ký học, số kinh nghiệm làm việc theo năm, số lượng nhân viên trong công ty hiện tại, loại hình công ty, sự khác biệt về số nằm giữa công việc trước đây và công việc hiện tại, số giờ đào tạo đã hoàn thành xong. Xếp theo Rank, chúng tôi xin đề cập 3 trong số 9 thuộc tính (top 3) sẽ ảnh hưởng nhiều nhất đến các quyết định hợp tác của ứng viên. Đầu tiên là chỉ số phát triển của thành phố đang sinh sống có mức độ ảnh hưởng cao nhất đến một trong các quyết định của nhân viên. Theo nhận xét từ bộ dữ liệu Job Change, những ứng viên đến từ thành phố phát triển cao với chỉ số phát triển từ 0.7 đến 0.95 sẽ có xu hướng muốn làm công việc này lâu dài hơn (target = 0 chiếm 67,8%) so với các thành phố có chỉ số phát triển thấp từ 0.45 đến 0.5 thì ứng viên thường có xu hướng tìm công việc khác (target = 1 chiếm 32,2%). Nguyên nhân là do các ứng viên sinh sống ở các thành phố phát triển thì họ đã sớm được trải nghiệm và gần như là có thâm niên lâu dài hoặc các kinh nghiệm liên quan đến công nghệ hay các ngành nghề tương ứng Data Sientists. Mặt khác thì số năm kinh nghiệm làm việc cũng sẽ ảnh hưởng đến quyết định hợp tác dài lâu của nhân viên. Từ vị trí thứ 2 xếp theo Orange và dữ liệu tính toán từ Excel, những nhân viên có hơn 20 năm kinh nghiệm hoặc từ 10 đến 19 năm kinh nghiệm sẽ mong muốn được làm việc với công ty lâu dài và không có xu hướng bỏ việc, đi tìm công việc mới như các nhân viên mới vào làm có thâm niên làm việc dưới 5 năm kinh nghiệm. Có thể nói, kinh nghiệm làm việc là một trong các yếu tố công ty dùng để đánh giá sự bền bỉ của nhân viên

theo đuổi ngành nghề họ ứng tuyển vị trí. Thêm vào đó, Company Size tức là số lượng nhân viên trong công ty hiện tại cũng sẽ ảnh hưởng đến quyết định làm việc lâu dài. Theo bộ dữ liệu nhận xét, đối với những công ty có trên 10.000 nhân viên trở lên thì ứng viên sẽ ít tự tin ứng tuyển vào vị trí làm việc bởi vì tỷ lệ cạnh tranh là cao hơn so với mô hình công ty nhỏ lẻ dưới 500 nhân sự. Nếu được nhận vào làm ở những công ty có mô hình rộng lớn, sự sa thải hằng năm có thể sẽ cao hơn những công ty nhỏ lẻ nhưng điều này lại phụ thuộc vào mức độ chăm chỉ làm việc, tỷ lệ tuyển dụng hằng năm,… còn đối với những công ty có mô hình nhỏ lẻ, sa thải hằng năm của nhân viên có thể sẽ thấp hơn.

Nhằm hạn chế tốt nhất ở sai lầm loại II tức là tuyển 91 ứng viên vào vị trí nhà khoa học dữ liệu nhưng thực tế thì những người này lại không có xu hướng mong muốn làm việc lâu dài với công ty thì ta cũng rất cần để ý mức độ ảnh hưởng của sai lầm loại I và loại II. Nếu chấp nhận quá nhiều sai lầm loại II, ta sẽ tuyển nhầm nhân viên và ảnh hưởng đến nhân sự, hiệu suất công ty. Do đó, nhóm chúng tôi xin đưa ra những hướng phát triển, giải pháp nhằm khắc phục tình trạng trên sau:

- ***Hạn chế tuyển dụng sai người ngay từ ban đầu***

  Ngay từ những ngày thử việc, các nhà tuyển dụng nên giới thiệu cho nhân viên về cách quản lý, tổ chức của công ty, giúp nhân viên có sự đánh giá về khả năng phù hợp của mình với doanh nghiệp, từ đó đưa ra quyết định gắn bó hay không ngay từ đầu.

- ***Nắm bắt nhu cầu và mong muốn của người lao động***

  Tổ chức các cuộc khảo sát ý kiến của nhân viên về công ty và công việc hiện tại, tiếp thu góp ý để cải thiện môi trường làm việc và cách quản lý nhân viên.

- ***Có các chế độ đãi ngộ và lợi ích cạnh tranh***

  Một số nhân viên sẽ xem lương thưởng là yếu tố thúc đẩy họ tìm kiếm một môi trường làm việc tốt hơn với chế độ lương thưởng cao hơn. Một chế độ đãi ngộ tốt, lương thưởng phù hợp sẽ động viên nhân viên làm việc hiệu quả, gắn bó lâu dài với công ty.

- ***Lịch làm việc đa dạng***

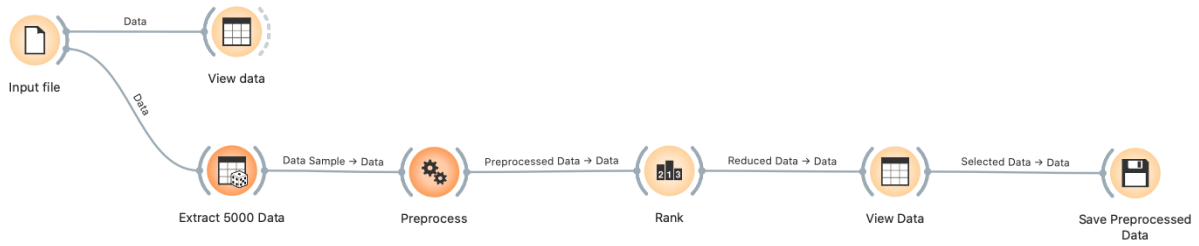
  Cho phép các nhân viên làm việc sớm hoặc muộn hơn tùy theo lịch trình của họ, ngoài ra có thể cho phép làm việc tại nhà mà vẫn đem lại hiệu quả.

- ***Phân bổ công việc phù hợp với năng lực cá nhân,***

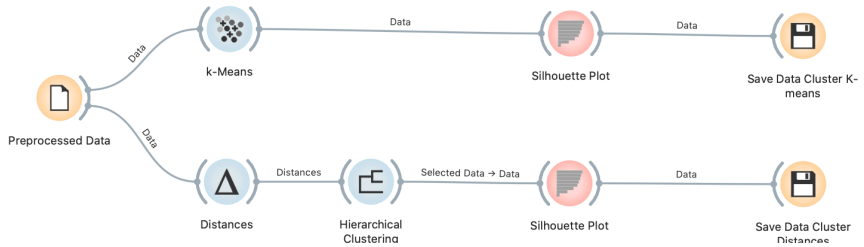
  Giúp nhân viên hoàn thành tốt công việc ứng với sở trường của mình, nếu không chỉ mang lại tổn hại cho doanh nghiệp và nhân viên nghỉ việc hàng loạt.

MÔ HÌNH DỮ LIỆU ORANGE DATA MINING

1. Tiền xử lý dữ liệu

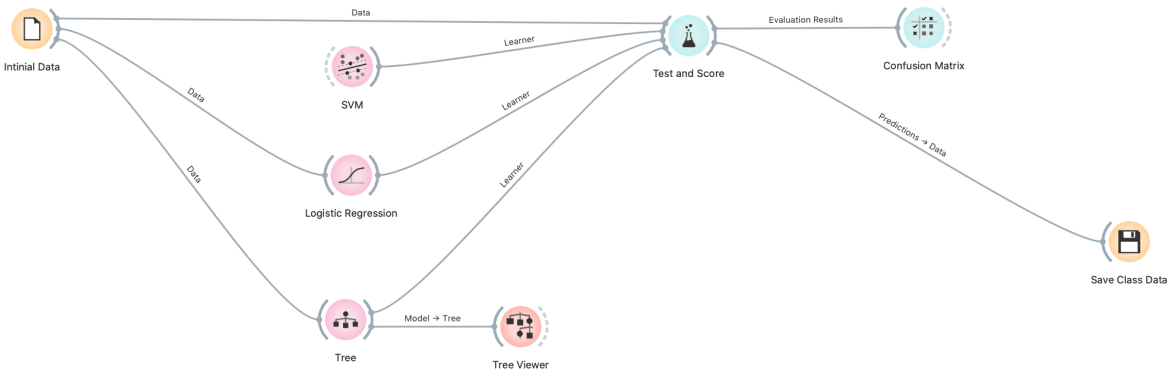


2. Phân cụm bộ dữ liệu

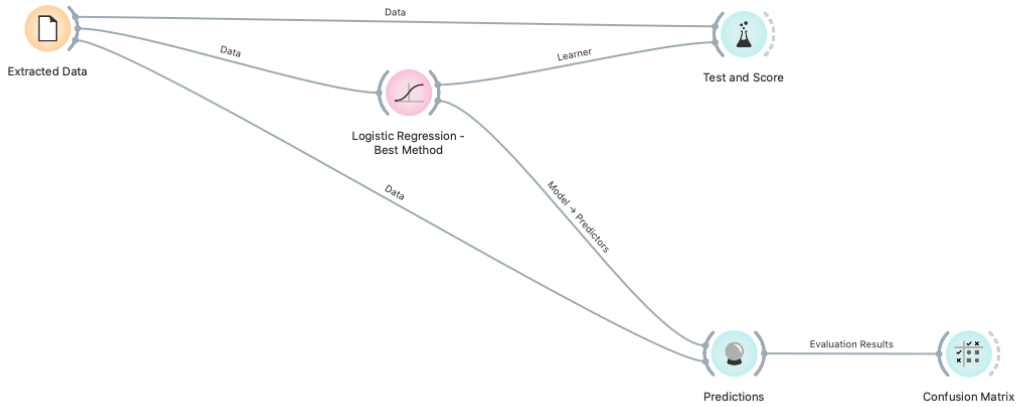


3. Phân lớp bộ dữ liệu

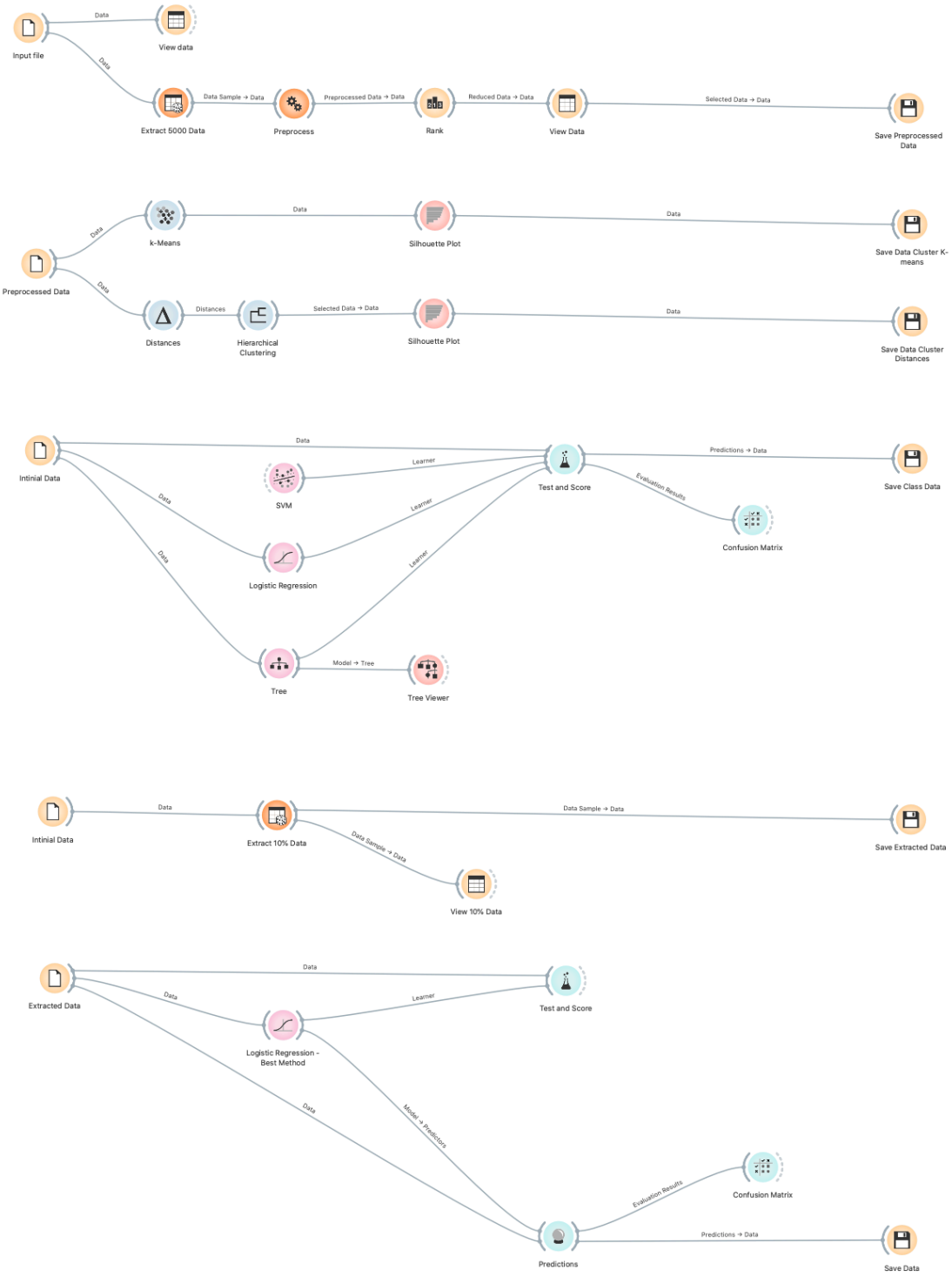
❖ ***Tìm phương pháp phân lớp tốt nhất***



❖ ***Trích xuất 10% dữ liệu dự báo***



4. Mô hình Orange Data Mining tổng thể



✳ **Kiểm tra đạo văn báo cáo cuối kỳ đề tài "Ứng dụng công cụ Orange Data Mining phân tích xu hướng thay đổi công việc Data Scientists từ bộ dữ liệu khảo sát Job Change of Data Scientists"**

bài KHDL

ORIGINALITY REPORT

8%	7%	2%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	text.123docz.net Internet Source	3%
2	www.slideshare.net Internet Source	3%
3	Banking Academy Publication	1%

Exclude quotes	On	Exclude matches	< 100 words
Exclude bibliography	On		

TÀI LIỆU THAM KHẢO

[1] Big Data and Data Science Firm, "*Analytics Job Change Of Data Scientists*", 2020. [Trực tuyến]. Địa chỉ: https://www.kaggle.com/datasets/arashnic/analytics-job-change-of-data-scientists. [Truy cập 18/10/2020]

[2] S. Narkhede, "*Understanding Confusion Matrix*", 2018. [Trực tuyến]. Địa chỉ: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62. [Truy cập 9/5/2018]

[3] Bài giảng học phần Khoa Học Dữ Liệu, "*Orange Data Mining*" khoa Công Nghệ Thông Tin Kinh Doanh, Trường Đại học Kinh tế Tp.HCM, 2022.

[4] C. Lew, "*How to hold employees accountable without micromaning them*", 2021. [Trực tuyến]. Địa chỉ: https://knowyourteam.com/blog/2021/10/27/how-to-hold-employees-accountable-without-micromanaging-them/. [Truy cập: 27/10/2021]

[5] Orange Workflows, "*Classification Tree*". [Trực tuyến]. Địa chỉ: https://orangedatamining.com/workflows/Classification/.

[6] Orange Workflows, "*Clustering*". [Trực tuyến]. Địa chỉ: https://orangedatamining.com/workflows/Clustering/.

[7] Orange Workflows, "*Hierarchical Clustering*". [Trực tuyến]. Địa chỉ:https://orangedatamining.com/workflows/Hierarchical-Clustering/.

[8] IBM Cloud Education, "*What is Data Sciene*", 2022. [Trực tuyến]. Địa chỉ: https://www.ibm.com/cloud/learn/data-science-introduction. [Truy cập: 2/8/2022]

[9] Orange Data Mining, "*Orange Visual Programming Documentation*", Release 3, 2/8/2019.