

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**  
**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HỒ CHÍ**  
**MINH**  
**KHOA CÔNG NGHỆ THÔNG TIN**



**HCMUTE**

**BÁO CÁO CUỐI KỲ**

**MÔN HỌC: KHAI PHÁ DỮ LIỆU**

**ĐỀ TÀI: ĐỀ TÀI: ANALYSIS OF POTENTIAL CUSTOMER**  
**(PHÂN TÍCH DỮ LIỆU KHÁCH HÀNG TIỀM NĂNG)**

**Giảng viên hướng dẫn: Nguyễn Văn Thành**

**Mã môn học: DAMI330484**

<b>SV thực hiện:</b>	<b>Lê Phước Yên</b>	<b>MSSV: 20133119</b>
	<b>Phan Thanh Tín</b>	<b>MSSV: 20133095</b>
	<b>Vũ Trung Kiên</b>	<b>MSSV: 20133060</b>
	<b>Nguyễn Đức Linh</b>	<b>MSSV: 20133007</b>

## LỜI CẢM ƠN

Thay mặt nhóm sinh viên chuyên ngành Công nghệ Thông tin, chúng em xin bày tỏ lời cảm ơn chân thành và sâu sắc đến Thầy Nguyễn Văn Th - người đã dành thời gian hướng dẫn, giúp đỡ và đóng góp nhiều ý kiến quý báu để giúp chúng tôi hoàn thành đề tài môn học Data Mining.

Qua đề tài này, chúng tôi đã học hỏi được nhiều kiến thức bổ ích và quý giá về các thuật toán khai thác dữ liệu, tuy nhiên vẫn còn nhiều hạn chế. Chúng tôi rất mong nhận được sự chỉ bảo, góp ý từ Thầy để áp dụng nhiều thuật toán hơn, so sánh và đánh giá kết quả tốt hơn, hoàn thiện hơn đề tài trong tương lai.

Chúng tôi hi vọng với sự hỗ trợ của Thầy, cùng với nỗ lực của chính bản thân, nhóm chúng tôi sẽ có những bước tiến dài hơn trong lĩnh vực Data Mining.

Một lần nữa, thay mặt nhóm sinh viên, tôi xin gửi lời cảm ơn chân thành nhất đến Thầy.

Trân trọng!

# MỤC LỤC

<b>I. TỔNG QUAN ĐỀ TÀI:</b>	5
1. Bảng phân công nhiệm vụ:	5
2. Lý do chọn đề tài:	7
3. Giới thiệu về tập dữ liệu:	8
3.1. Tập dữ liệu trong AdventureWorksDW2019:	8
3.2. Tập dữ liệu về thu thập cá nhân:	9
4. Công cụ sử dụng:	9
4.1. Visual Studio 2019:	9
4.2. Visual Studio Code:	10
4.3. SQL Server 2019:	10
4.4. Ngôn ngữ lập trình (Python, SQL):	10
4.5. SSAS:	10
5. Thuật toán sử dụng:	11
<b>II. CHUẨN BỊ DỮ LIỆU:</b>	11
1. Dữ liệu gốc:	11
1.1. Tập dữ liệu trong AdventureWorksDW2019:	11
1.2. Tập dữ liệu về thu thập cá nhân:	12
2. Tiền xử lý dữ liệu về thu thập cá nhân:	12
3. Import vào Database:	14
3.1. Import tập dữ liệu AdventureWorksDW2019	14
3.2. Import dữ liệu về thu nhập cá nhân	15
4. Xử lý dữ liệu để chọn cụm tối ưu	16
4.1. Tập dữ liệu trong AdventureWorksDW2019:	16
4.2. Tập dữ liệu về thu thập cá nhân:	18
<b>III. THỰC HIỆN CÁC GIẢI THUẬT:</b>	19
1. Thuật toán Microsoft Clustering:	19
1.1. Tạo Clustering Project với SSAS:	19
1.2. Deploy Clustering Mining:	23
2. So sánh độ chính xác của thuật toán Microsoft Logistic Regression và Microsoft Decision Tree trong dự đoán một người có mua xe đạp không?	33

2.1.	Tạo Logistic Regression Project với SSAS:.....	33
2.2.	Deploy Logistic Regression Mining: .....	35
3.	Thuật toán Microsoft Decision Tree:.....	42
3.1.	Tạo Decision Trees Project với SSAS: .....	42
3.2.	Deploy Decision Trees Mining:.....	44
IV.	TỔNG KẾT:.....	46
1.	Kết quả đạt được:.....	46
2.	Hạn chế: .....	46
3.	Phương hướng phát triển: .....	47
V.	THAM KHẢO:.....	47

## I. TỔNG QUAN ĐỀ TÀI:

### 1. Bảng phân công nhiệm vụ:

Công việc	Phan Thanh Tín	Lê Phước Yên	Nguyễn Đức Linh	Vũ Trung Kiên
Chọn tập dữ liệu		100%		100%
Chọn giải thuật	100%			100%
Tiền xử lý dữ liệu		100%		
Xác định các biến sử dụng để khai phá dữ liệu	100%		100%	100%
Thực hiện Elbow để chia cụm tối ưu cho tập dữ liệu		100%	100%	
Thực hiện ma trận tương quan cho tập dữ liệu		100%		
Thực hiện thuật toán Microsoft Clustering trên SSAS	100%	100%	100%	100%
Đọc kết quả và nhận xét thuật toán Microsoft Clustering	100%			100%

Thực hiện thuật toán Logistic Regression sử dụng SSAS	100%		100%	100%
Đọc kết quả và nhận xét cho thuật toán Logistic Regression	100%		100%	
Thực hiện thuật toán Decision Tree sử dụng SSAS	100%	100%	100%	100%
Đọc kết quả và nhận xét cho thuật toán Decision Tree		100%	100%	
Viết báo cáo		100%		100%
Làm slide thuyết trình	100%		100%	

## **2. Lý do chọn đề tài:**

Phân tích khách hàng tiềm năng trong cơ sở dữ liệu AdventureWorksDW2019 là một đề tài được chọn vì nó mang lại thông tin quan trọng về khách hàng và tiềm năng kinh doanh. AdventureWorksDW2019 cung cấp một tập dữ liệu rộng lớn và đa dạng về các khía cạnh của khách hàng, bao gồm tuổi, thu nhập, sở thích và hành vi mua hàng.

Bằng cách phân tích dữ liệu trong AdventureWorksDW2019, chúng ta có thể xác định những đặc điểm quan trọng của khách hàng tiềm năng, như độ tuổi, thu nhập và sở thích sản phẩm. Chúng ta có thể tìm hiểu và xác định nhóm khách hàng có tiềm năng cao để tăng cường chiến lược tiếp thị và quảng cáo, cũng như phát triển các chiến lược chăm sóc khách hàng phù hợp.

Bên cạnh đó nhóm cũng sử dụng tập dữ liệu “Adult.csv” để thực hiện khai phá dữ liệu vì tập dữ liệu có nhiều thuộc tính khác nhau như giới tính, tuổi, trình độ học vấn, tình trạng hôn nhân, số người phụ thuộc,... Cho phép nhóm phân tích mối tương quan giữa chúng với thu nhập.

Việc phân tích tập dữ liệu này có thể cho thấy những xu hướng và đặc điểm thu nhập của người dân và giúp làm sáng tỏ một số yếu tố ảnh hưởng đến thu nhập cá nhân.

### 3. Giới thiệu về tập dữ liệu:

#### 3.1. Tập dữ liệu trong AdventureWorksDW2019:

Trong cơ sở dữ liệu mẫu AdventureWorksDW2019, có một view có tên vTargetMail. View này được sử dụng để thống kê các thông tin liên quan đến chiến dịch gửi thư quảng cáo cho khách hàng.

Mô tả tập dữ liệu này:

- CustomerKey: Mã khách hàng duy nhất.
- GeographyKey: Khóa ngoại tham chiếu đến bảng Geography.
- CustomerAlternateKey: Khóa chính của bảng Customer.
- Title: Tiêu đề khách hàng.
- FirstName: Tên khách hàng.
- MiddleName: Tên đệm của khách hàng.
- LastName: Họ của khách hàng.
- BirthDate: Ngày sinh của khách hàng.
- MaritalStatus: Tình trạng hôn nhân của khách hàng.
- Suffix: Hậu tố tên của khách hàng.
- Gender: Giới tính của khách hàng.
- EmailAddress: Địa chỉ email của khách hàng.
- YearlyIncome: Thu nhập hàng năm của khách hàng.
- TotalChildren: Tổng số lượng con cái của khách hàng.
- NumberChildrenAtHome: Số lượng con cái của khách hàng đang sống tại nhà.
- EnglishEducation: Trình độ học vấn của khách hàng trong tiếng Anh.
- SpanishEducation: Trình độ giáo dục của khách hàng trong tiếng Tây Ban Nha.
- FrenchEducation: Trình độ giáo dục của khách hàng trong ngôn ngữ Pháp.
- EnglishOccupation: Nghề nghiệp của khách hàng trong ngôn ngữ Anh.
- SpanishOccupation: Nghề nghiệp của khách hàng trong ngôn ngữ Tây Ban Nha.
- FrenchOccupation: Nghề nghiệp của khách hàng trong ngôn ngữ Pháp.
- HouseOwnerFlag: Cờ cho biết khách hàng có sở hữu nhà hay không.
- NumberCarsOwned: Số lượng xe ô tô mà khách hàng sở hữu.
- AddressLine1: Địa chỉ dòng 1 của khách hàng.
- AddressLine2: Địa chỉ dòng 2 của khách hàng.
- Phone: Số điện thoại của khách hàng.
- DateFirstPurchase: Ngày khách hàng lần đầu tiên mua hàng từ công ty.
- CommuteDistance: Khoảng cách giữa địa chỉ của khách hàng và địa điểm làm việc của họ.
- Region: Khu vực của khách hàng.
- Age: Tuổi của khách hàng.
- BikeBuyer: Biến đếm nhị phân (1 hoặc 0) cho biết khách hàng đã mua xe đạp trước đó hay chưa.



### **3.2. Tập dữ liệu về thu thập cá nhân:**

Tập dữ liệu "Adult Census Income" trên Kaggle là một tập dữ liệu kinh tế - xã hội của những người dân ở Mỹ. Tập dữ liệu này được thu thập từ Cục điều tra dân số Hoa Kỳ và bao gồm hồ sơ về 32.561 người, với 15 thuộc tính định lượng và định tính. Dữ liệu này hữu ích cho các nhà nghiên cứu và các nhà phân tích để tìm hiểu về mối quan hệ giữa thu nhập và các yếu tố khác như tuổi, giới tính, trình độ học vấn, nghề nghiệp, tình trạng hôn nhân,....

Các thuộc tính trong tập dữ liệu bao gồm:

- age: Tuổi của người được khảo sát.
- workclass: Loại nghề nghiệp của người được khảo sát.
- fnlwgt: Trọng số của mỗi người được khảo sát trong mẫu.
- education: Trình độ học vấn của người được khảo sát.
- education-num: Trình độ học vấn của người được khảo sát.
- marital-status: Tình trạng hôn nhân của người được khảo sát ở dạng số.
- occupation: Nghề nghiệp của người được khảo sát.
- relationship: Mối quan hệ gia đình của người được khảo sát.
- race: chủng tộc của người được khảo sát.
- sex: Giới tính của người được khảo sát, chẳng hạn như.
- capital-gain: Thu nhập từ vốn đầu tư của người được khảo sát.
- capital-loss: Mất mát từ vốn đầu tư của người được khảo sát.
- hours-per-week: Số giờ làm việc mỗi tuần của người được khảo sát.
- native-country: Quốc gia xuất thân của người được khảo sát.
- Income: Thu nhập của người được khảo sát.

## **4. Công cụ sử dụng:**

### **4.1. Visual Studio 2019:**

Visual Studio 2019 là một môi trường phát triển tích hợp (IDE) được tạo ra bởi Microsoft. Nó được sử dụng để phát triển các ứng dụng phần mềm cho các nền tảng Windows, Android, iOS, web và đám mây. Visual Studio 2019 hỗ trợ nhiều ngôn ngữ lập trình bao gồm C++, C#, Visual Basic, F#, JavaScript, TypeScript và Python. Nó cung cấp một loạt các công cụ và tính năng giúp nhà phát triển viết, kiểm thử và triển khai mã của họ, chẳng hạn như các trình soạn thảo mã, trình gỡ lỗi, trình biên dịch và hệ thống kiểm soát phiên bản. Visual Studio 2019 cũng có một cộng đồng lớn và tích cực của các nhà phát triển tạo ra các tiện ích mở rộng và bổ sung có thể tải xuống và cài đặt để mở rộng chức năng của IDE.

#### **4.2. Visual Studio Code:**

Visual Studio Code (thường được gọi là VS Code) là một trình biên tập mã nguồn mở được phát triển bởi Microsoft. Nó hỗ trợ nhiều ngôn ngữ lập trình, bao gồm C++, C#, Java, Python, JavaScript, TypeScript, và nhiều ngôn ngữ khác. VS Code cung cấp nhiều tính năng hữu ích cho các nhà phát triển, bao gồm trình gỡ lỗi tích hợp, hỗ trợ Git, tìm kiếm và thay thế thông minh, định dạng mã tự động, và nhiều tính năng khác. VS Code cũng có một cộng đồng đông đảo và tích cực của các nhà phát triển tạo ra các tiện ích mở rộng và bổ sung có thể tải xuống và cài đặt để mở rộng chức năng của trình biên tập mã này.

#### **4.3. SQL Server 2019:**

SQL Server 2019 là một hệ quản trị cơ sở dữ liệu quan hệ được phát triển bởi Microsoft. Nó cung cấp nhiều tính năng tiên tiến và được thiết kế để xử lý các tác vụ phức tạp trong môi trường doanh nghiệp. SQL Server 2019 hỗ trợ nhiều ngôn ngữ truy vấn, bao gồm SQL, R và Python, cho phép các nhà phát triển và nhà quản trị cơ sở dữ liệu sử dụng các công cụ và ngôn ngữ yêu thích của họ để tương tác với cơ sở dữ liệu. Nó cũng cung cấp các tính năng bảo mật và quản lý dữ liệu, bao gồm mã hóa dữ liệu, xác thực người dùng, sao lưu và khôi phục dữ liệu, và nhiều tính năng khác. SQL Server 2019 cũng tích hợp với các công nghệ mới nhất, chẳng hạn như máy học, học sâu và trí tuệ nhân tạo, để phục vụ cho các tác vụ phân tích dữ liệu và khai thác dữ liệu.

#### **4.4. Ngôn ngữ lập trình (Python, SQL):**

Python là một ngôn ngữ lập trình thông dịch, đa năng và có cú pháp đơn giản. Nó được sử dụng rộng rãi trong nhiều lĩnh vực, bao gồm khoa học dữ liệu, trí tuệ nhân tạo, phát triển web, tự động hóa, và các ứng dụng máy tính khác. Python có cú pháp dễ đọc và hiểu, cộng với một thư viện phong phú, làm cho nó trở thành một ngôn ngữ lập trình phổ biến trong cộng đồng lập trình.

SQL là ngôn ngữ truy vấn cơ sở dữ liệu quan hệ được sử dụng để truy xuất, thêm, sửa đổi và xóa dữ liệu từ cơ sở dữ liệu. SQL được sử dụng rộng rãi trong các ứng dụng doanh nghiệp, quản lý cơ sở dữ liệu và phân tích dữ liệu. SQL có thể được sử dụng để tạo bảng, quản lý dữ liệu, tạo các mối quan hệ giữa các bảng, và thực hiện các phép tính và truy vấn phức tạp trên các bảng dữ liệu. Các ngôn ngữ truy vấn dữ liệu khác như NoSQL cũng đang được sử dụng rộng rãi trong môi trường lập trình hiện đại.

#### **4.5. SSAS:**

SSAS (SQL Server Analysis Services) là công cụ phân tích và báo cáo do Microsoft phát triển, nó là một phần của gói SQL Server. SSAS cho phép người dùng tạo ra các khái niệm dữ liệu ảo như khối lượng, kim tự tháp và hồ sơ được sử dụng cho việc phân tích và báo cáo dữ liệu.

Một số điểm nổi bật của SSAS:

- Cho phép phân tích dữ liệu từ nhiều nguồn: SSAS có thể kết nối và khai thác dữ liệu từ các nguồn khác nhau như SQL Server, Oracle, dữ liệu ngoài mạng hoặc file Excel.
- Tối ưu hóa hiệu suất: SSAS cho phép lưu trữ dữ liệu đã tính toán sẵn để tối ưu hóa hiệu suất khi thực hiện các phép tính phức tạp và yêu cầu nhiều dữ liệu.
- Hỗ trợ nhiều chiều: SSAS hỗ trợ việc phân tích và báo cáo dữ liệu với nhiều hơn 3 chiều thông qua khối lượng, kim tự tháp và hồ sơ.
- Tích hợp với các công cụ phân tích và báo cáo khác: SSAS tích hợp tốt với SQL Server Reporting Services, Power BI, Excel và nhiều công cụ khác để cung cấp một hệ sinh thái phân tích và báo cáo toàn diện.

## 5. Thuật toán sử dụng:

Sau khi bàn bạc thì nhóm đã quyết định đưa ra 3 thuật toán sử dụng cho việc khai phá dữ liệu:

- Microsoft Clustering: Được sử dụng để nhóm dữ liệu vào các nhóm (hay cụm) mang tính đồng nhất.
- Microsoft Logistic Regression: Thuật toán logistic regression của Microsoft thực hiện việc xác định mối quan hệ giữa biến đầu vào và biến phụ thuộc bằng cách sử dụng hàm sigmoid để dự đoán xác suất của biến phụ thuộc. Hàm sigmoid có đầu ra nằm trong khoảng từ 0 đến 1, với giá trị 0,5 được chọn là một ngưỡng để phân loại các mẫu thành hai lớp.
- Microsoft Decision Tree: Mô hình dựa trên cây quyết định được sử dụng cho cả thuật toán phân cụm và phân lớp. Nó tách dữ liệu thành nhóm nhỏ dựa trên các điều kiện của các thuộc tính.

## II. CHUẨN BỊ DỮ LIỆU:

### 1. Dữ liệu gốc:

#### 1.1. Tập dữ liệu trong AdventureWorksDW2019:

CustomerKey	GeographyKey	CustomerAlternateKey	Title	FirstName	MiddleName	LastName	NameStyle	BirthDate
11000	26	AW00011000		Jon	V	Yang	FALSE	10/6/1971
11001	37	AW00011001		Eugene	L	Huang	FALSE	5/10/1976
11002	31	AW00011002		Ruben		Torres	FALSE	2/9/1971
11003	11	AW00011003		Christy		Zhu	FALSE	8/14/1973
11004	19	AW00011004		Elizabeth		Johnson	FALSE	8/5/1979
11005	22	AW00011005		Julio		Ruiz	FALSE	8/1/1976
11006	8	AW00011006		Janet	G	Alvarez	FALSE	12/2/1976
11007	40	AW00011007		Marco		Mehta	FALSE	11/6/1969
11008	32	AW00011008		Rob		Verhoff	FALSE	7/4/1975

MaritalStatus	Suffix	Gender	EmailAddress	YearlyIncome	TotalChildren	NumberChildrenAtHome	EnglishEducation
M		M	jon24@adventure-works.com	90000	2	0	Bachelors
S		M	eugene10@adventure-works.com	60000	3	3	Bachelors
M		M	ruben35@adventure-works.com	60000	3	3	Bachelors
S		F	christy12@adventure-works.com	70000	0	0	Bachelors
S		F	elizabeth5@adventure-works.com	80000	5	5	Bachelors
S		M	julio1@adventure-works.com	70000	0	0	Bachelors
S		F	janet9@adventure-works.com	70000	0	0	Bachelors
M		M	marco14@adventure-works.com	60000	3	3	Bachelors
S		F	rob4@adventure-works.com	60000	4	4	Bachelors

SpanishEducation	FrenchEducation	EnglishOccupation	SpanishOccupation	FrenchOccupation	HouseOwnerFlag	NumberCarsOwned
Licenciatura	Bac + 4	Professional	Profesional	Cadre	1	0
Licenciatura	Bac + 4	Professional	Profesional	Cadre	0	1
Licenciatura	Bac + 4	Professional	Profesional	Cadre	1	1
Licenciatura	Bac + 4	Professional	Profesional	Cadre	0	1
Licenciatura	Bac + 4	Professional	Profesional	Cadre	1	4
Licenciatura	Bac + 4	Professional	Profesional	Cadre	1	1
Licenciatura	Bac + 4	Professional	Profesional	Cadre	1	1
Licenciatura	Bac + 4	Professional	Profesional	Cadre	1	2
Licenciatura	Bac + 4	Professional	Profesional	Cadre	1	3

AddressLine1	AddressLine2	Phone	DateFirstPurchase	CommuteDistance	Region	Age	BikeBuyer
3761 N. 14th St		1 (11) 500 555-0162	1/19/2011	1-2 Miles	Pacific	51	1
2243 W St.		1 (11) 500 555-0110	1/15/2011	0-1 Miles	Pacific	47	1
5844 Linden Land		1 (11) 500 555-0184	1/7/2011	2-5 Miles	Pacific	52	1
1825 Village Pl.		1 (11) 500 555-0162	12/29/2010	5-10 Miles	Pacific	49	1
7553 Harness Circle		1 (11) 500 555-0131	1/23/2011	1-2 Miles	Pacific	43	1
7305 Humphrey Drive		1 (11) 500 555-0151	12/30/2010	5-10 Miles	Pacific	46	1
2612 Berry Dr		1 (11) 500 555-0184	1/24/2011	5-10 Miles	Pacific	46	1
942 Brook Street		1 (11) 500 555-0126	1/9/2011	0-1 Miles	Pacific	53	1
624 Peabody Road		1 (11) 500 555-0164	1/25/2011	10+ Miles	Pacific	47	1

## 1.2. Tập dữ liệu về thu thập cá nhân:

age	workclass	fnlwtg	education	education.num	marital.status	occupation	relationship	race
90 ?		77053	HS-grad		9 Widowed	?	Not-in-family	White
82 Private		132870	HS-grad		9 Widowed	Exec-managerial	Not-in-family	White
66 ?		186061	Some-college		10 Widowed	?	Unmarried	Black
54 Private		140359	7th-8th		4 Divorced	Machine-op-inspct	Unmarried	White
41 Private		264663	Some-college		10 Separated	Prof-specialty	Own-child	White
34 Private		216864	HS-grad		9 Divorced	Other-service	Unmarried	White
38 Private		150601	10th		6 Separated	Adm-clerical	Unmarried	White
74 State-gov		88638	Doctorate		16 Never-married	Prof-specialty	Other-relative	White
68 Federal-gov		422013	HS-grad		9 Divorced	Prof-specialty	Not-in-family	White

sex	capital.gain	capital.loss	hours.per.week	native.country	income
Female	0	4356	40	United-States	<=50K
Female	0	4356	18	United-States	<=50K
Female	0	4356	40	United-States	<=50K
Female	0	3900	40	United-States	<=50K
Female	0	3900	40	United-States	<=50K
Female	0	3770	45	United-States	<=50K
Male	0	3770	40	United-States	<=50K
Female	0	3683	20	United-States	>50K
Female	0	3683	40	United-States	<=50K

## 2. Tiền xử lý dữ liệu về thu thập cá nhân:

Import thư viện

```
import seaborn as sns
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

✓ 3.3s

## Đọc dữ liệu và quan sát tổng quan

```
dataf = pd.read_csv("adult.csv",encoding_errors="ignore",na_values=['?'])
dataf.head()
```

✓ 0.2s

	age	workclass	fnlwgt	education	education.num	marital.status	occupation	relationship
0	90	NaN	77053	HS-grad	9	Widowed	NaN	Not-in-family
1	82	Private	132870	HS-grad	9	Widowed	Exec-managerial	Not-in-family
2	66	NaN	186061	Some-college	10	Widowed	NaN	Unmarried
3	54	Private	140359	7th-8th	4	Divorced	Machine-op-inspct	Unmarried
4	41	Private	264663	Some-college	10	Separated	Prof-specialty	Own-child

## Thêm ID cho tập dữ liệu và quan sát dữ liệu null

```
dataf['Id'] = dataf.index

# Reset index về một cột mới "index"
dataf = dataf.reset_index()

# Gộp cột ID với DataFrame ban đầu
dataf = pd.concat([dataf['Id'], dataf.drop(['Id', 'index'], axis=1)], axis=1)
```

✓ 0.1s

```
dataf.isnull().sum()
```

✓ 0.1s

Id	0
age	0
workclass	1836
fnlwgt	0
education	0
education.num	0
marital.status	0
occupation	1843
relationship	0
race	0
sex	0
capital.gain	0
capital.loss	0
hours.per.week	0
native.country	583
income	0

dtype: int64

## Drop NA và xuất ra file CSV

```
data1 = dataf.dropna(axis=0)
data1.to_csv('income.csv', index=False)
```

✓ 0.4s

File sau khi tiền xử lý

Id	age	workclass	fnlwgt	education	education.num	marital.status
1	82	Private	132870	HS-grad	9	Widowed
3	54	Private	140359	7th-8th	4	Divorced
4	41	Private	264663	Some-college	10	Separated
5	34	Private	216864	HS-grad	9	Divorced
6	38	Private	150601	10th	6	Separated
7	74	State-gov	88638	Doctorate	16	Never-married
8	68	Federal-gov	422013	HS-grad	9	Divorced
10	45	Private	172274	Doctorate	16	Divorced
11	38	Self-emp-not-inc	164526	Prof-school	15	Never-married

race	sex	capital.gain	capital.loss	hours.per.week	native.country
White	Female	0	4356	18	United-States
White	Female	0	3900	40	United-States
White	Female	0	3900	40	United-States
White	Female	0	3770	45	United-States
White	Male	0	3770	40	United-States
White	Female	0	3683	20	United-States
White	Female	0	3683	40	United-States
Black	Female	0	3004	35	United-States
White	Male	0	2824	45	United-States

## 3. Import vào Database:

### 3.1. Import tập dữ liệu AdventureWorksDW2019

	CustomerKey	GeographyKey	CustomerAlternateKey	Title	First Name	Middle Name	Last Name	NameStyle	BirthDate	MaritalStatus	Suffix	Gender	Email Address	YearlyIncome
1	11000	26	AW00011000	NULL	Jon	V	Yang	0	1971-10-06	M	NULL	M	jon24@adventure-works.com	90000.00
2	11001	37	AW00011001	NULL	Eugene	L	Huang	0	1976-05-10	S	NULL	M	eugene10@adventure-works.com	60000.00
3	11002	31	AW00011002	NULL	Ruben	NULL	Torres	0	1971-02-09	M	NULL	M	ruben35@adventure-works.com	60000.00
4	11003	11	AW00011003	NULL	Christy	NULL	Zhu	0	1973-08-14	S	NULL	F	christy12@adventure-works.com	70000.00
5	11004	19	AW00011004	NULL	Elizabeth	NULL	Johnson	0	1979-08-05	S	NULL	F	elizabeth5@adventure-works.com	80000.00
6	11005	22	AW00011005	NULL	Julio	NULL	Ruiz	0	1976-08-01	S	NULL	M	julio1@adventure-works.com	70000.00
7	11006	8	AW00011006	NULL	Janet	G	Alvarez	0	1976-12-02	S	NULL	F	janet9@adventure-works.com	70000.00
8	11007	40	AW00011007	NULL	Marco	NULL	Mehta	0	1969-11-06	M	NULL	M	marco14@adventure-works.com	60000.00
9	11008	32	AW00011008	NULL	Rob	NULL	Verhoff	0	1975-07-04	S	NULL	F	rob4@adventure-works.com	60000.00
10	11009	25	AW00011009	NULL	Shannon	C	Carlson	0	1969-09-29	S	NULL	M	shannon38@adventure-works.com	70000.00

	TotalChildren	NumberChildrenAtHome	EnglishEducation	SpanishEducation	FrenchEducation	EnglishOccupation	SpanishOccupation	FrenchOccupation	HouseOwnerFlag	NumberCarsOwned
1	2	0	Bachelors	Licenciatura	Bac + 4	Professional	Profesional	Cadre	1	0
2	3	3	Bachelors	Licenciatura	Bac + 4	Professional	Profesional	Cadre	0	1
3	3	3	Bachelors	Licenciatura	Bac + 4	Professional	Profesional	Cadre	1	1
4	0	0	Bachelors	Licenciatura	Bac + 4	Professional	Profesional	Cadre	0	1
5	5	5	Bachelors	Licenciatura	Bac + 4	Professional	Profesional	Cadre	1	4
6	0	0	Bachelors	Licenciatura	Bac + 4	Professional	Profesional	Cadre	1	1
7	0	0	Bachelors	Licenciatura	Bac + 4	Professional	Profesional	Cadre	1	1
8	3	3	Bachelors	Licenciatura	Bac + 4	Professional	Profesional	Cadre	1	2
9	4	4	Bachelors	Licenciatura	Bac + 4	Professional	Profesional	Cadre	1	3
10	0	0	Bachelors	Licenciatura	Bac + 4	Professional	Profesional	Cadre	0	1

AddressLine1	AddressLine2	Phone	DateFirstPurchase	CommuteDistance	Region	Age	BikeBuyer
3761 N. 14th St	NULL	1 (11) 500 555-0162	2011-01-19	1-2 Miles	Pacific	51	1
2243 W St.	NULL	1 (11) 500 555-0110	2011-01-15	0-1 Miles	Pacific	47	1
5844 Linden Land	NULL	1 (11) 500 555-0184	2011-01-07	2-5 Miles	Pacific	52	1
1825 Village Pl.	NULL	1 (11) 500 555-0162	2010-12-29	5-10 Miles	Pacific	49	1
7553 Hamess Circle	NULL	1 (11) 500 555-0131	2011-01-23	1-2 Miles	Pacific	43	1
7305 Humphrey Drive	NULL	1 (11) 500 555-0151	2010-12-30	5-10 Miles	Pacific	46	1
2612 Berry Dr	NULL	1 (11) 500 555-0184	2011-01-24	5-10 Miles	Pacific	46	1
942 Brook Street	NULL	1 (11) 500 555-0126	2011-01-09	0-1 Miles	Pacific	53	1
624 Peabody Road	NULL	1 (11) 500 555-0164	2011-01-25	10+ Miles	Pacific	47	1
3839 Northgate Road	NULL	1 (11) 500 555-0110	2011-01-27	5-10 Miles	Pacific	53	1

### 3.2. Import dữ liệu về thu nhập cá nhân

Chọn kiểu dữ liệu

Column Name	Data Type	Primary Key	<input type="checkbox"/> Allow Nulls	
<b>id</b>	int	<input type="checkbox"/>	<input type="checkbox"/>	
age	int	<input type="checkbox"/>	<input type="checkbox"/>	
workclass	nvarchar(50)	<input type="checkbox"/>	<input type="checkbox"/>	
fnlwgt	int	<input type="checkbox"/>	<input type="checkbox"/>	
education	nvarchar(50)	<input type="checkbox"/>	<input type="checkbox"/>	
education_num	int	<input type="checkbox"/>	<input type="checkbox"/>	
marital_status	nvarchar(50)	<input type="checkbox"/>	<input type="checkbox"/>	
occupation	nvarchar(50)	<input type="checkbox"/>	<input type="checkbox"/>	
relationship	nvarchar(50)	<input type="checkbox"/>	<input type="checkbox"/>	
race	nvarchar(50)	<input type="checkbox"/>	<input type="checkbox"/>	
sex	nvarchar(50)	<input type="checkbox"/>	<input type="checkbox"/>	
capital_gain	int	<input type="checkbox"/>	<input type="checkbox"/>	
capital_loss	int	<input type="checkbox"/>	<input type="checkbox"/>	
hours_per_week	int	<input type="checkbox"/>	<input type="checkbox"/>	
native_country	nvarchar(50)	<input type="checkbox"/>	<input type="checkbox"/>	
income	nvarchar(50)	<input type="checkbox"/>	<input type="checkbox"/>	

## Kết quả

Results		Messages														
	Id	age	workclass	fnlwgt	education	education_num	marital_status	occupation	relationship	race	sex	capital_gain	capital_loss	hours_per_week	native_country	income
1	1	82	Private	132870	HS-grad	9	Widowed	Exec-managerial	Not-in-family	White	Female	0	4356	18	United-States	<=50K
2		54	Private	140359	7th-8th	4	Divorced	Machine-op-inspct	Unmarried	White	Female	0	3900	40	United-States	<=50K
3	4	41	Private	264663	Some-college	10	Separated	Prof-specialty	Own-child	White	Female	0	3900	40	United-States	<=50K
4	5	34	Private	216864	HS-grad	9	Divorced	Other-service	Unmarried	White	Female	0	3770	45	United-States	<=50K
5	6	38	Private	150601	10th	6	Separated	Adm-clerical	Unmarried	White	Male	0	3770	40	United-States	<=50K
6	7	74	State-gov	88638	Doctorate	16	Never-married	Prof-specialty	Other-relative	White	Female	0	3683	20	United-States	>50K
7	8	68	Federal-gov	422013	HS-grad	9	Divorced	Prof-specialty	Not-in-family	White	Female	0	3683	40	United-States	<=50K
8	10	45	Private	172274	Doctorate	16	Divorced	Prof-specialty	Unmarried	Black	Female	0	3004	35	United-States	>50K
9	11	38	Self-emp-not-inc	164526	Prof-school	15	Never-married	Prof-specialty	Not-in-family	White	Male	0	2824	45	United-States	>50K
10	12	52	Private	129177	Bachelors	13	Widowed	Other-service	Not-in-family	White	Female	0	2824	20	United-States	>50K
11	13	32	Private	136204	Masters	14	Separated	Exec-managerial	Not-in-family	White	Male	0	2824	55	United-States	>50K

## 4. Xử lý dữ liệu để chọn cụm tối ưu

### 4.1. Tập dữ liệu trong AdventureWorksDW2019:

Chuyển đổi giá trị trong cột từ dạng chuỗi ký tự thành dạng số nguyên.

```
# create an instance of LabelEncoder
from sklearn.calibration import LabelEncoder

le = LabelEncoder()

# copy original dataset
ds = df.copy()
# select categorical columns
categorical_cols = ds.select_dtypes(include=['object']).columns.tolist()

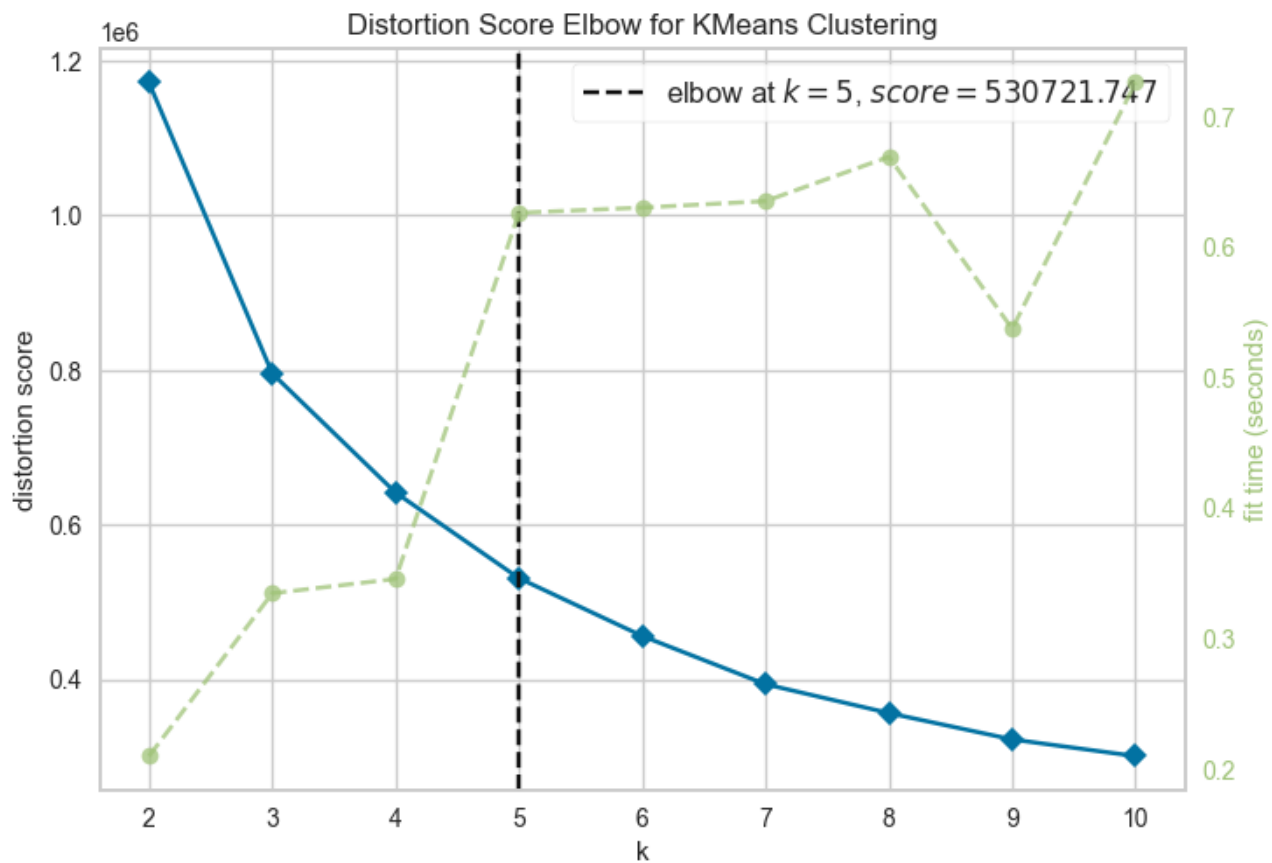
# label encode each column in the list
for col in categorical_cols:
    ds[col] = le.fit_transform(ds[col])

print(ds.info())
```

Vẽ biểu đồ Elbow

```
from yellowbrick.cluster import KElbowVisualizer
from sklearn.cluster import KMeans
print('Elbow Method to determine the number of clusters to be formed:')
Elbow_M = KElbowVisualizer(KMeans(random_state=123), k=10)
Elbow_M.fit(ds)
Elbow_M.show()
```





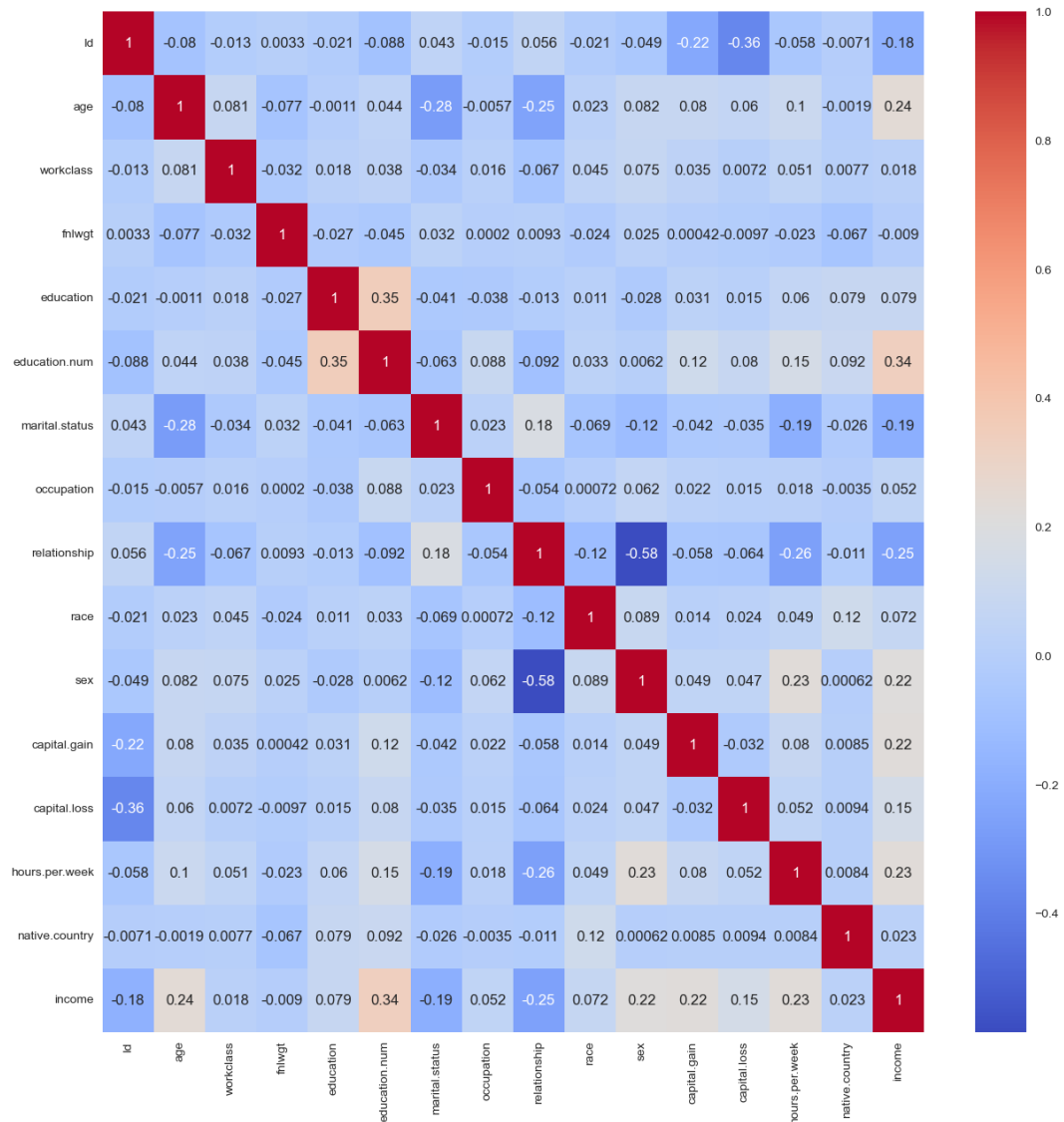
## 4.2. Tập dữ liệu về thu thập cá nhân:

Vẽ đồ thị heatmap

```
corr_matrix = ds.corr()

# Vẽ đồ thị heatmap
plt.figure(figsize=(15, 15))
sns.heatmap(corr_matrix, cmap='coolwarm', annot=True)

# Hiển thị đồ thị
plt.show()
```



### III. THỰC HIỆN CÁC GIẢI THUẬT:

#### 1. Thuật toán Microsoft Clustering:

##### 1.1. Tạo Clustering Project với SSAS:

##### 1.1.1. Thiết lập Data Source:

Chọn Database là AdventureWorks2019:

**Data Source Wizard**

**Select how to define the connection**  
You can select from a number of ways in which your data source will define its connection string.

☐ Create a data source based on another object

☒ Create a data source based on an existing or new connection

**Data connections:**

LAPTOP-QTNUSH09.AdventureWorks2019
LAPTOP-QTNUSH09.AdventureWorksDW2019
LAPTOP-QTNUSH09.AdventureWorksDW2008
LAPTOP-QTNUSH09.StoreSalesDW
localhost.CoSupply_DW

**Data connection properties:**

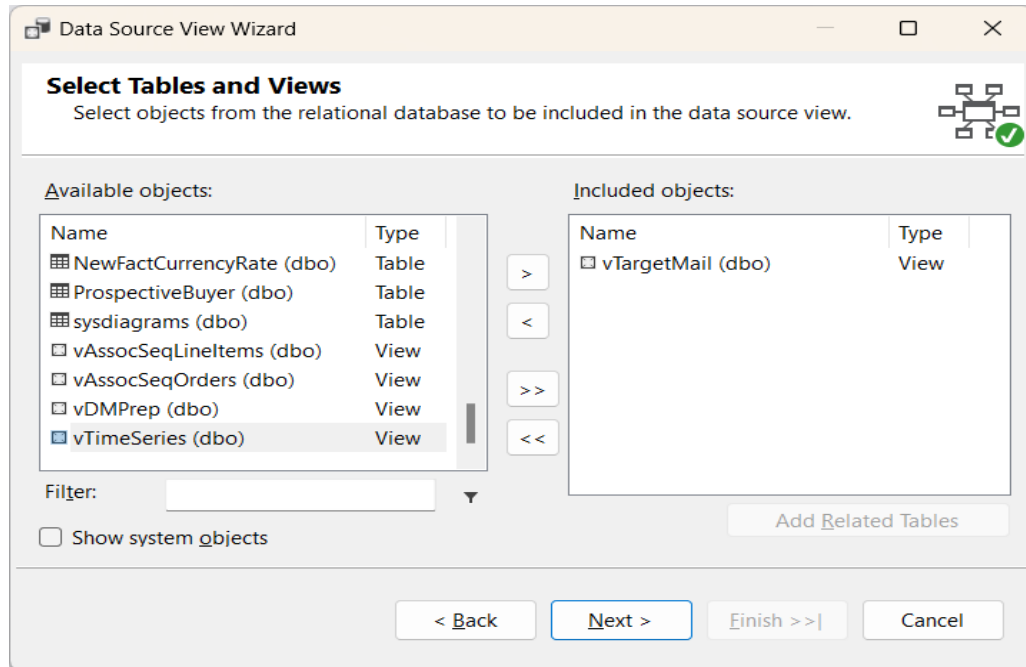
Property	Value
Data Source	LAPTOP-QTNUSH09
Initial Catalog	AdventureWorksDW2019
Integrated Security	SSPI
Provider	SQLOLEDB.1

New... Delete

< Back Next > Finish >>| Cancel

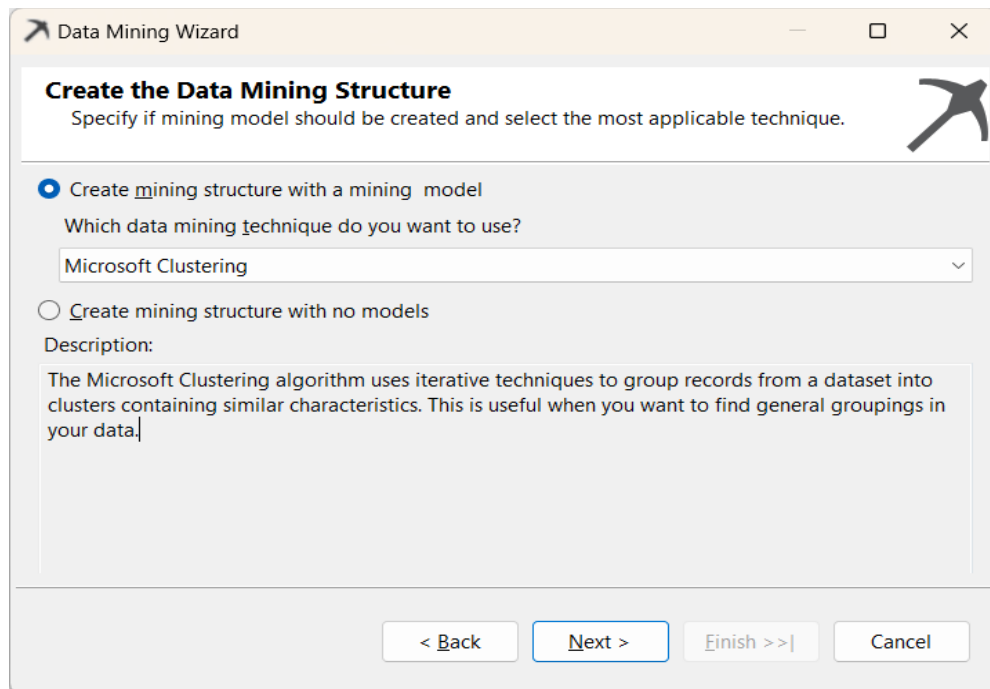
### 1.1.2. Thiết lập Data Source Views:

Chọn Object là vTargetMail



### 1.1.3. Thiết lập Mining Structures:

Chọn thuật toán là Clustering



## 1.1.4. Thiết lập Training Data:

Data Mining Wizard

**Specify the Training Data**  
Specify the columns used in your analysis.

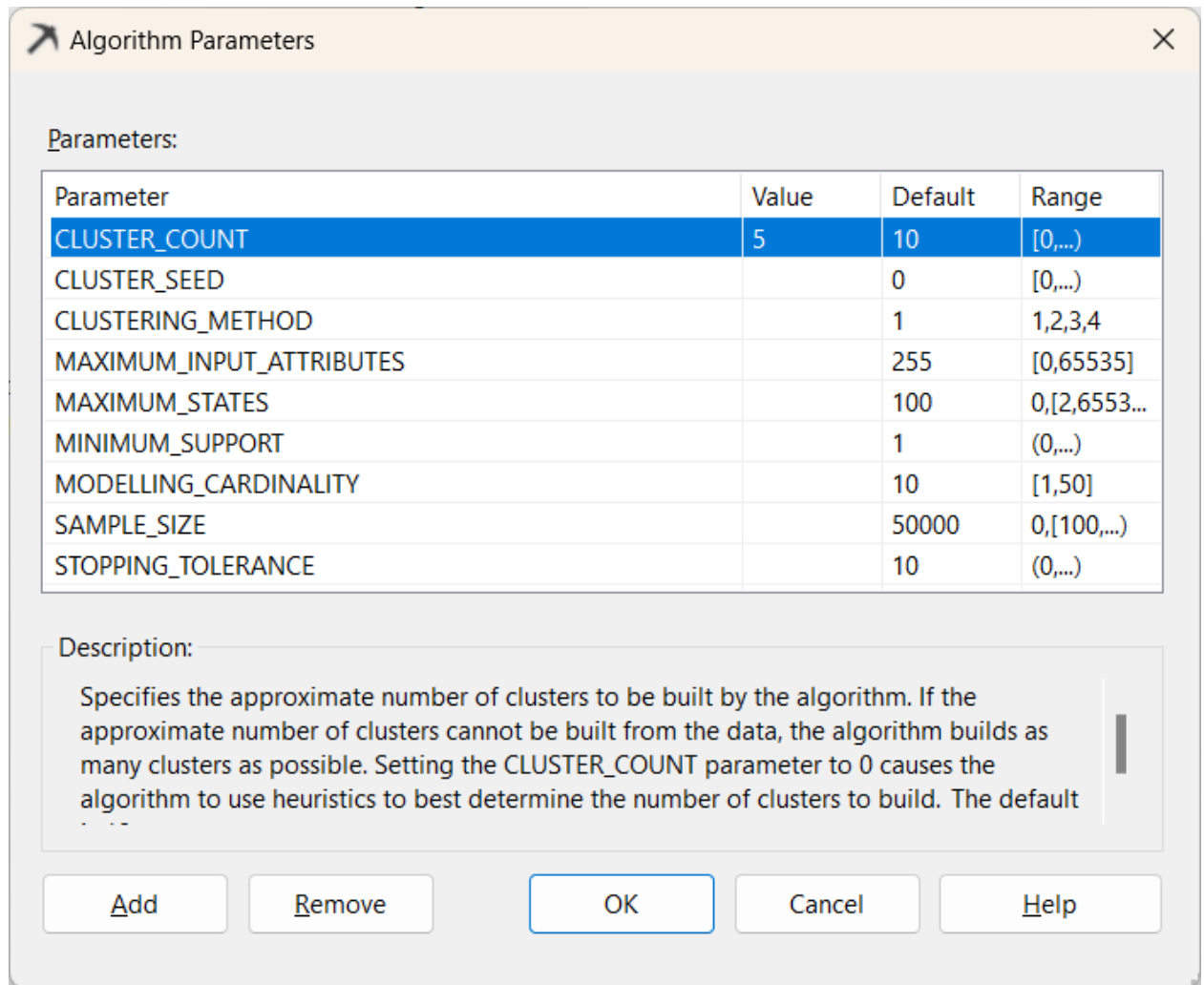
Mining model structure:

Tables/Columns	Key	Input	Predict
<input type="checkbox"/> DateFirstPurchase	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/> EmailAddress	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/> EnglishEducation	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/> EnglishOccupation	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/> FirstName	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/> FrenchEducation	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/> FrenchOccupation	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/> Gender	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/> GeographyKey	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/> HouseOwnerFlag	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/> LastName	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/> MaritalStatus	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/> MiddleName	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/> NameStyle	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/> NumberCarsOwned	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/> NumberChildrenAtHome	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/> Phone	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/> Region	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/> SpanishEducation	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/> SpanishOccupation	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/> Suffix	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/> Title	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/> TotalChildren	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/> YearlyIncome	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>

Recommend inputs for currently selected predictable: Suggest

< Back Next > Finish >> Cancel

### 1.1.5. Thiết lập Algorithm Parameters:

The image shows a software dialog box titled "Algorithm Parameters". It contains a table of parameters with columns for Parameter, Value, Default, and Range. The "CLUSTER\_COUNT" parameter is highlighted in blue and has a value of 5. Below the table is a description of the CLUSTER\_COUNT parameter. At the bottom are buttons for "Add", "Remove", "OK", "Cancel", and "Help".

Algorithm Parameters

Parameters:

Parameter	Value	Default	Range
CLUSTER_COUNT	5	10	[0,...)
CLUSTER_SEED		0	[0,...)
CLUSTERING_METHOD		1	1,2,3,4
MAXIMUM_INPUT_ATTRIBUTES		255	[0,65535]
MAXIMUM_STATES		100	0,[2,6553...
MINIMUM_SUPPORT		1	(0,...)
MODELLING_CARDINALITY		10	[1,50]
SAMPLE_SIZE		50000	0,[100,...)
STOPPING_TOLERANCE		10	(0,...)

Description:

Specifies the approximate number of clusters to be built by the algorithm. If the approximate number of clusters cannot be built from the data, the algorithm builds as many clusters as possible. Setting the CLUSTER\_COUNT parameter to 0 causes the algorithm to use heuristics to best determine the number of clusters to build. The default

Add

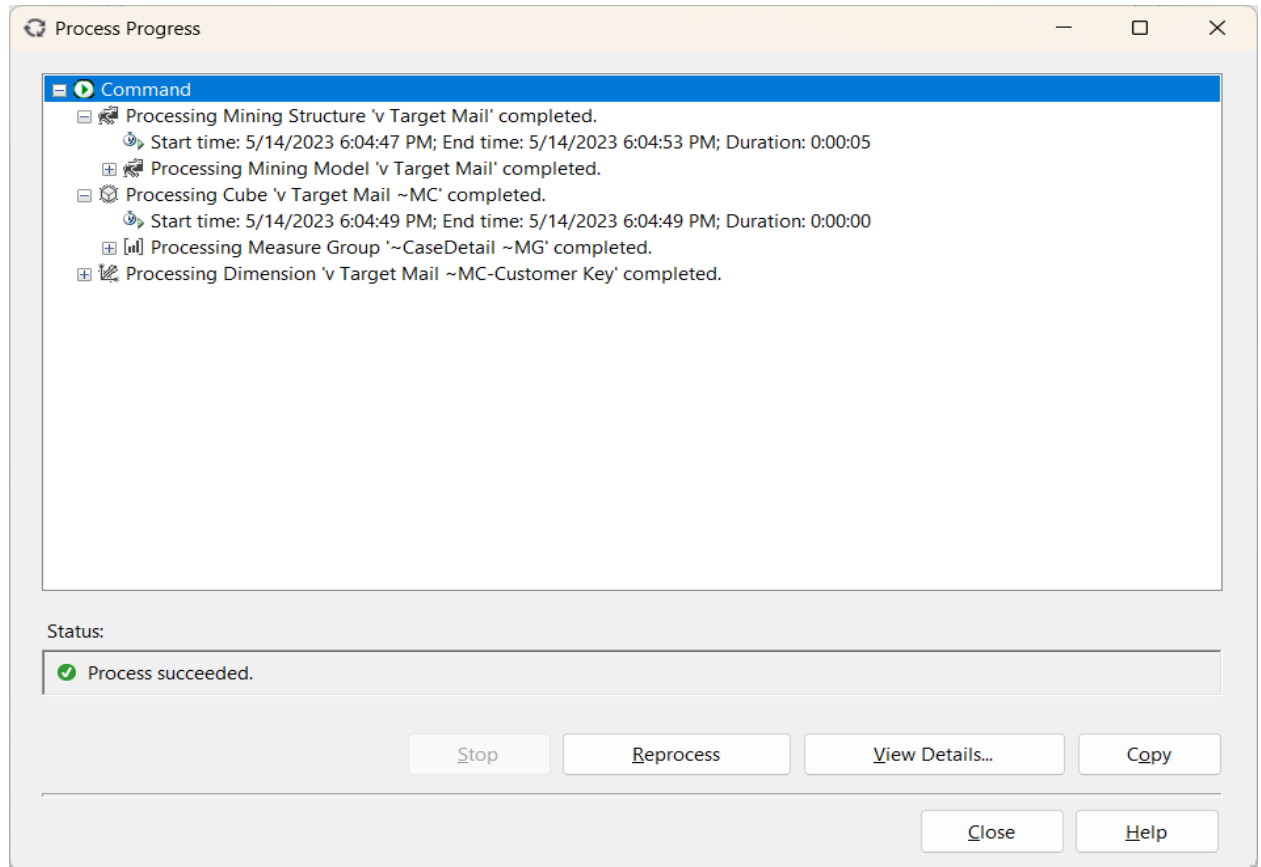
Remove

OK

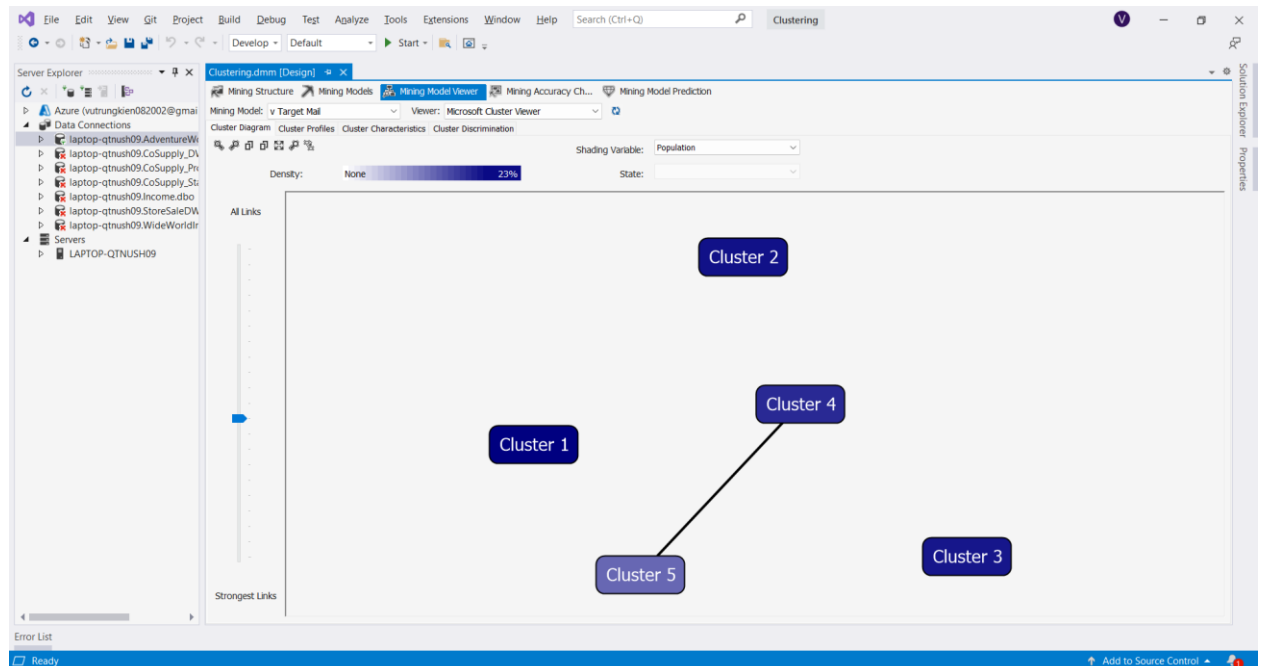
Cancel

Help

### 1.1.6. Processing Mining Structure:

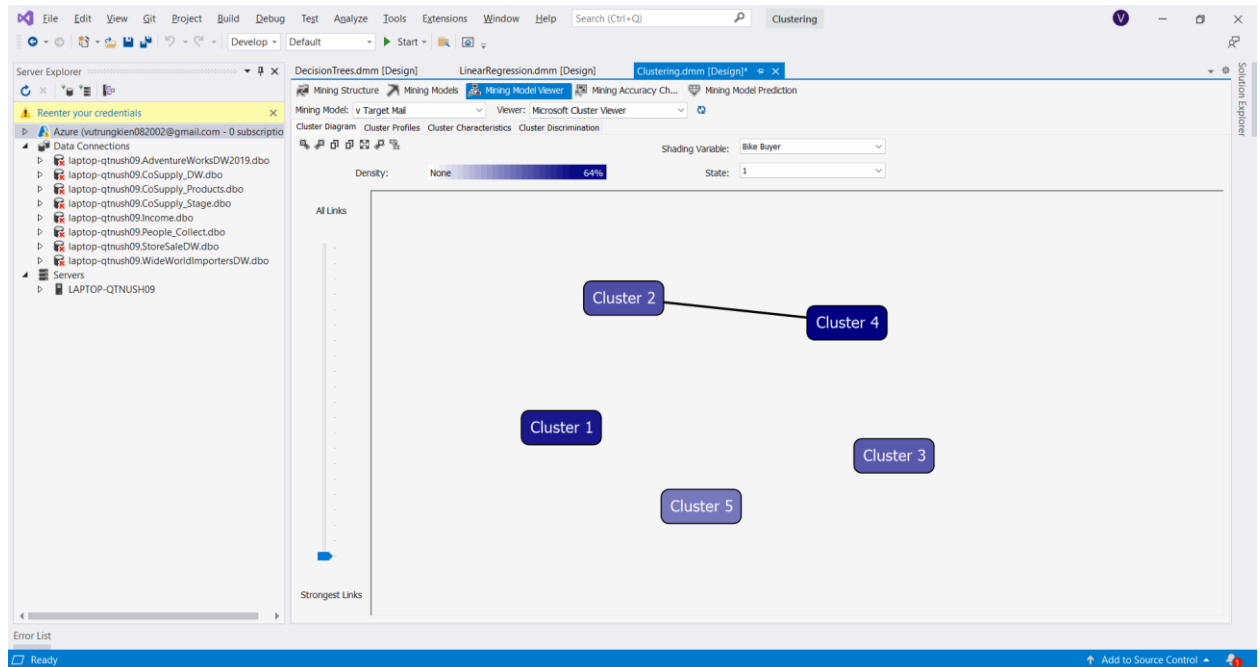


### 1.2. Deploy Clustering Mining:

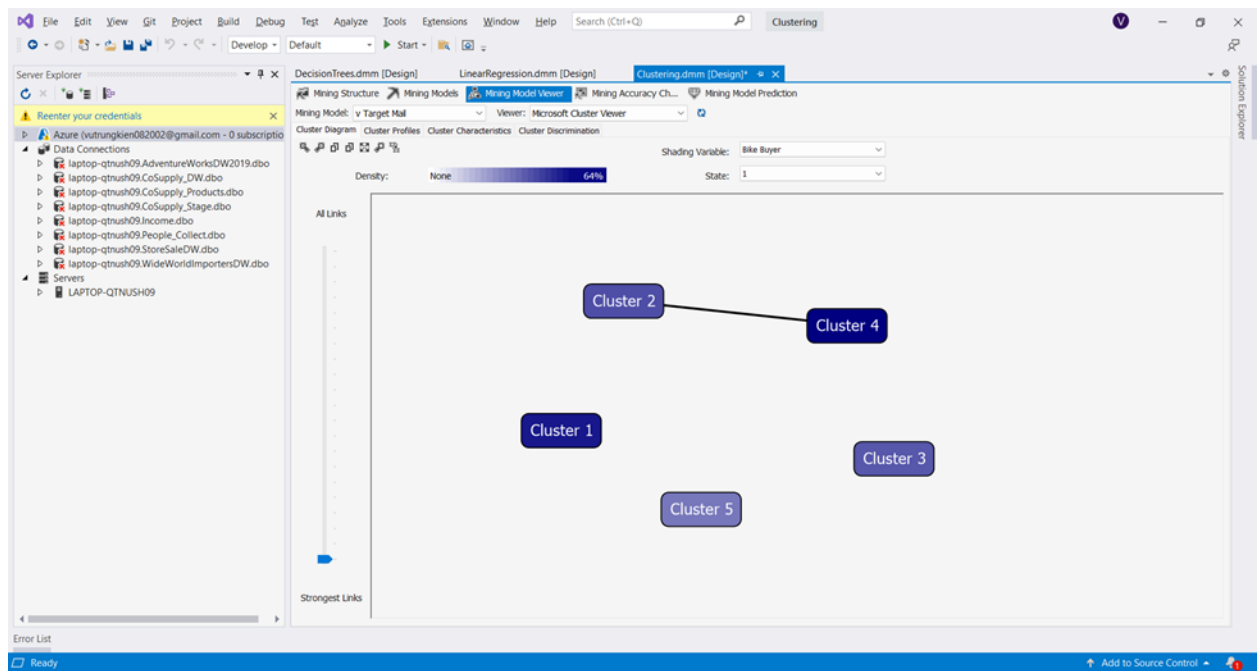


## 1.2.1. Cluster Diagram:

Chọn Shading Variable là Bike Buyer state 1 (những người mua xe đạp)

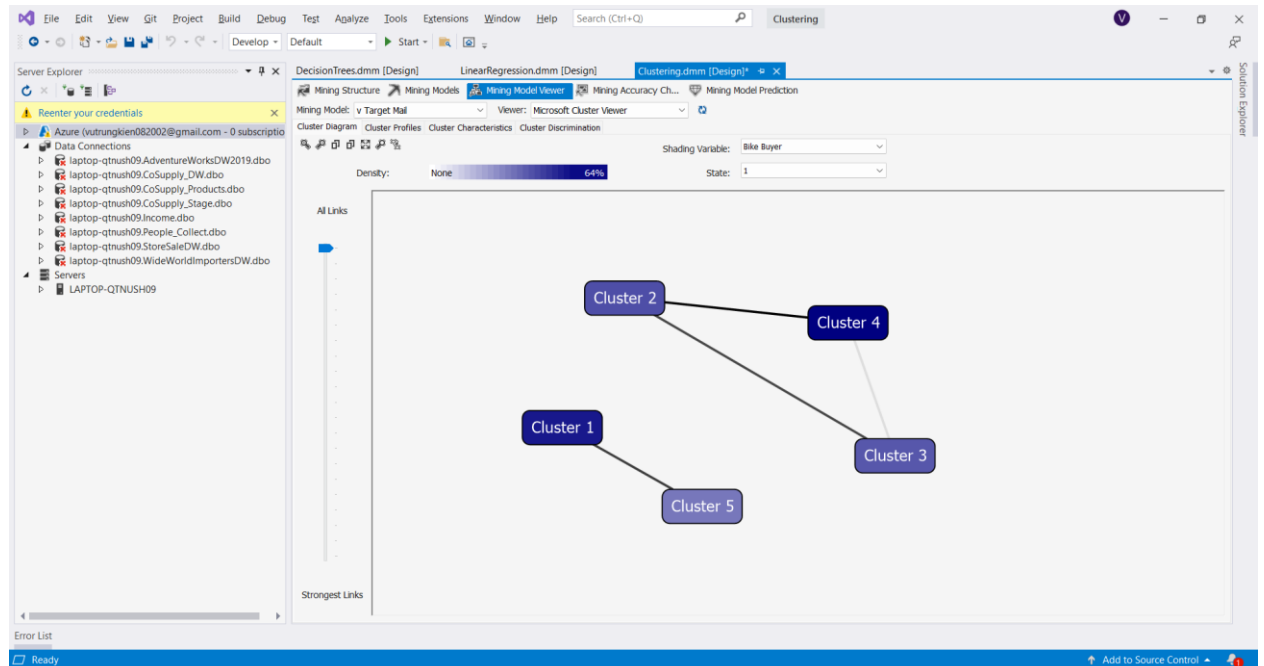


Tùy chọn mức liên kết mạnh nhất.





Tùy chọn mức liên kết yếu nhất.



### 1.2.2. Clustering Profile:

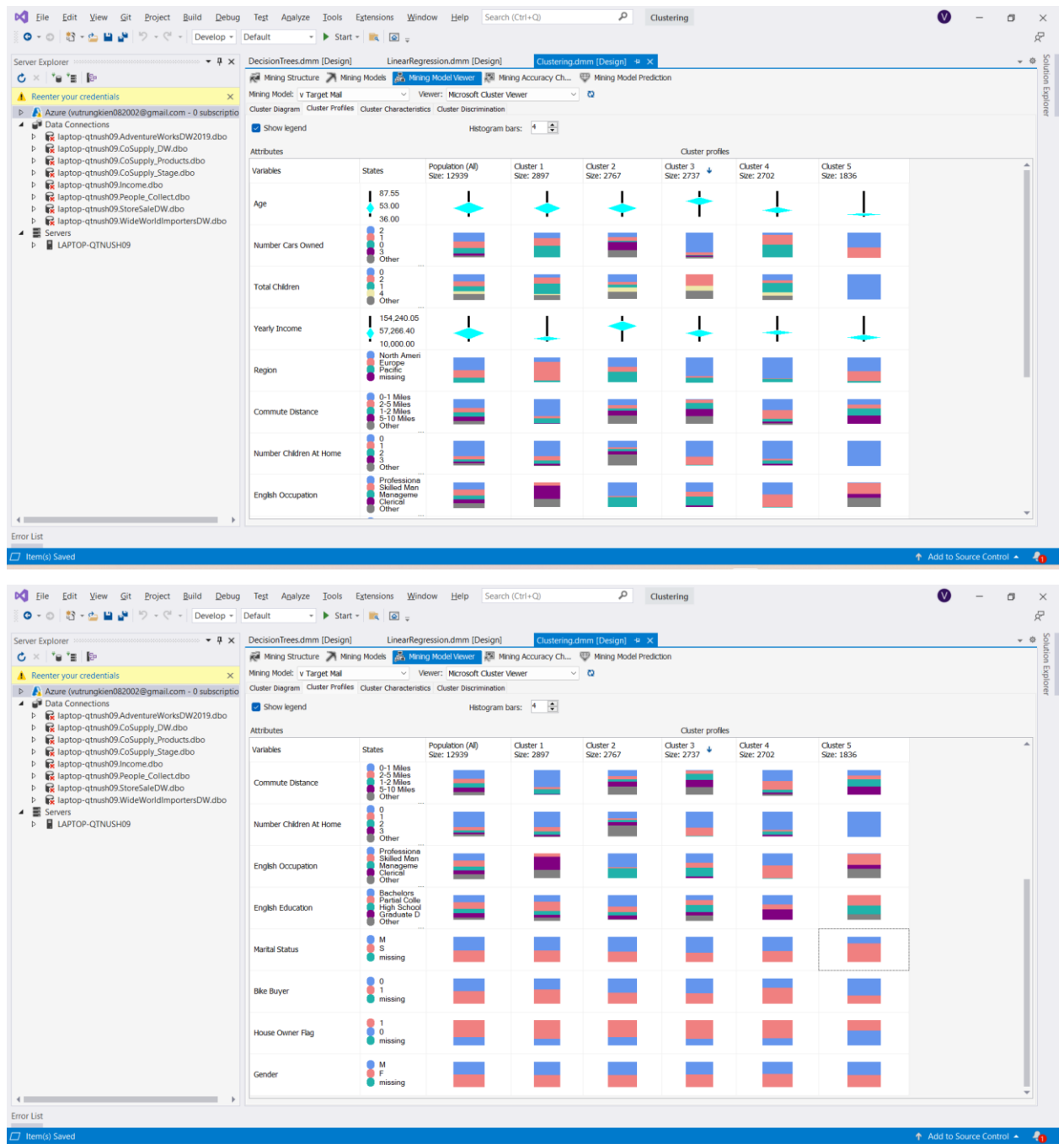
Variables	States	Popula tion (All)	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Size		12939	2897	2767	2737	2702	1836
Age	Mean	53.00	53.21	53.21	67.20	49.11	40.67
Age	Deviation	11.52	11.21	9.65	7.38	5.55	2.64
Bike Buyer	0	6576	0.426	0.561	0.583	0.364	0.663
Bike Buyer	1	6363	0.574	0.439	0.417	0.636	0.337
Bike Buyer	missing	0	0.000	0.000	0.000	0.000	0.000
Commute Distance	0-1 Miles	4429	0.674	0.254	0.027	0.446	0.228
Commute Distance	2-5 Miles	2276	0.098	0.135	0.142	0.360	0.164

Commute Distance	1-2 Miles	2266	0.202	0.088	0.236	0.115	0.277
Commute Distance	5-10 Miles	2212	0.025	0.201	0.292	0.079	0.331
Commute Distance	10+ Miles	1756	0.002	0.323	0.303	0.000	0.000
Commute Distance	missing	0	0.000	0.000	0.000	0.000	0.000
English Education	Bachelors	3700	0.271	0.458	0.207	0.374	0.005
English Education	Partial College	3575	0.372	0.223	0.206	0.203	0.413
English Education	High School	2288	0.168	0.132	0.297	0.000	0.351
English Education	Graduate Degree	2264	0.123	0.150	0.137	0.424	0.000
English Education	Partial High School	1112	0.067	0.037	0.153	0.000	0.231
English Education	missing	0	0.000	0.000	0.000	0.000	0.000
English Occupation	Professional	3849	0.000	0.542	0.369	0.470	0.040
English Occupation	Skilled Manual	3185	0.140	0.056	0.209	0.500	0.439
English Occupation	Management	2154	0.000	0.402	0.343	0.026	0.000
English Occupation	Clerical	2090	0.528	0.000	0.079	0.004	0.151

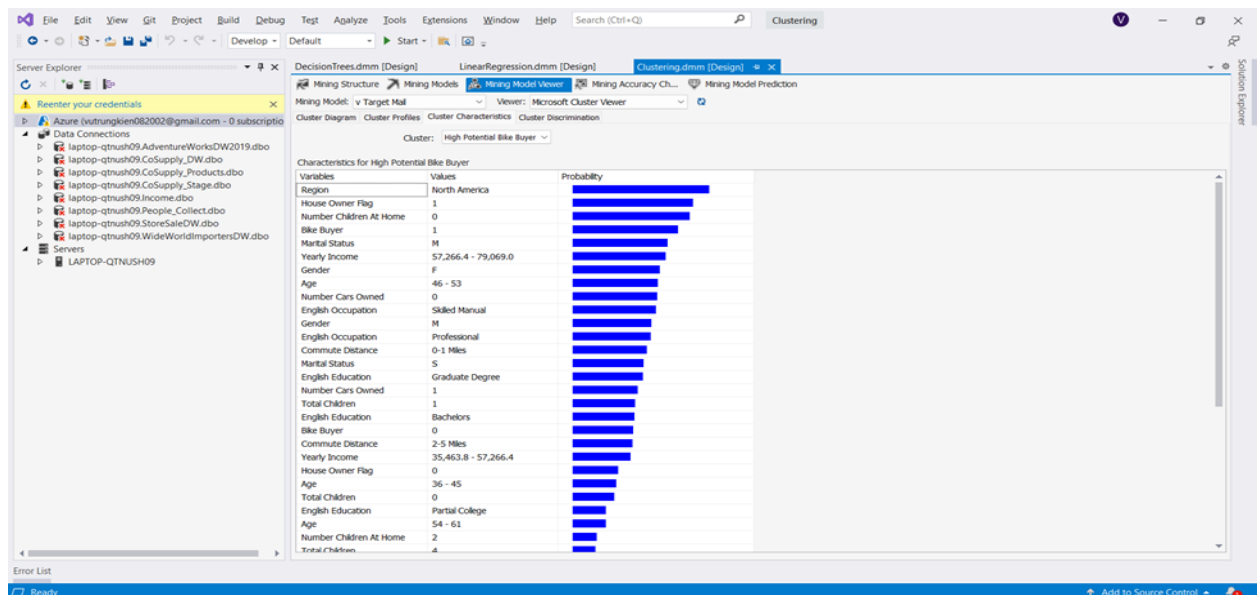
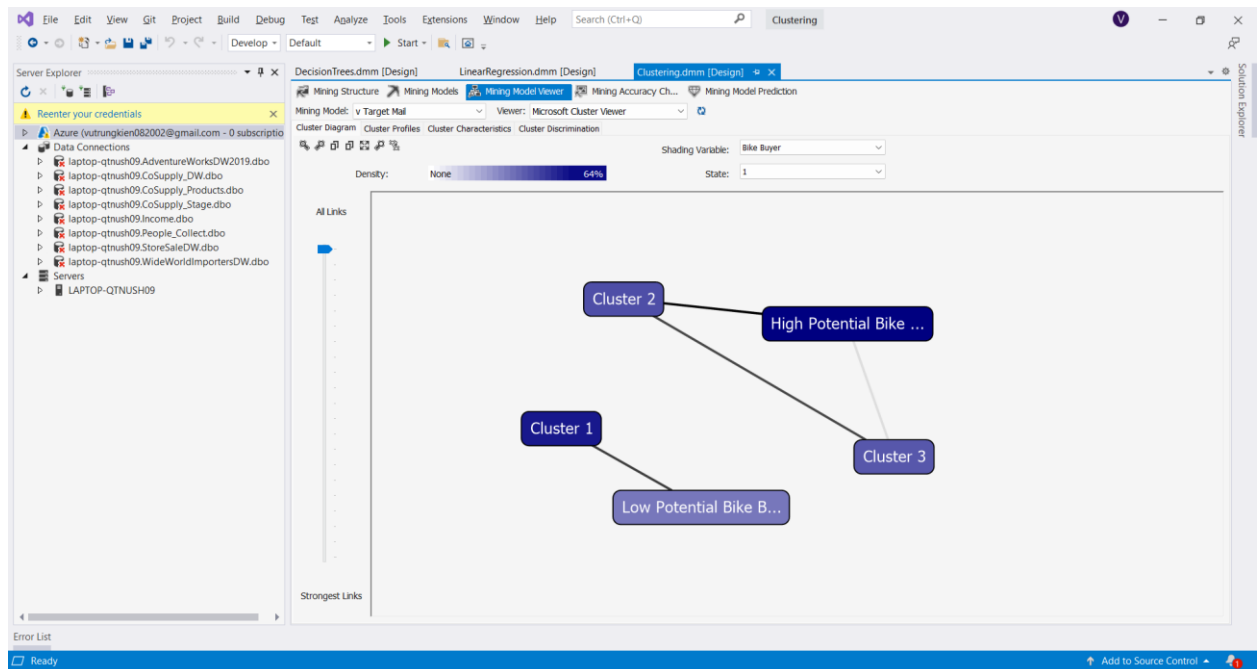
English Occupation	Manual	1661	0.332	0.000	0.000	0.000	0.370
English Occupation	missing	0	0.000	0.000	0.000	0.000	0.000
Gender	M	6558	0.515	0.513	0.507	0.476	0.527
Gender	F	6381	0.485	0.487	0.493	0.524	0.473
Gender	missing	0	0.000	0.000	0.000	0.000	0.000
House Owner Flag	1	8730	0.731	0.673	0.733	0.723	0.422
House Owner Flag	0	4209	0.269	0.327	0.267	0.277	0.578
House Owner Flag	missing	0	0.000	0.000	0.000	0.000	0.000
Marital Status	M	6957	0.548	0.584	0.625	0.571	0.261
Marital Status	S	5982	0.452	0.416	0.375	0.429	0.739
Marital Status	missing	0	0.000	0.000	0.000	0.000	0.000
Number Cars Owned	2	4524	0.221	0.179	0.784	0.099	0.590
Number Cars Owned	1	3444	0.306	0.166	0.114	0.391	0.410

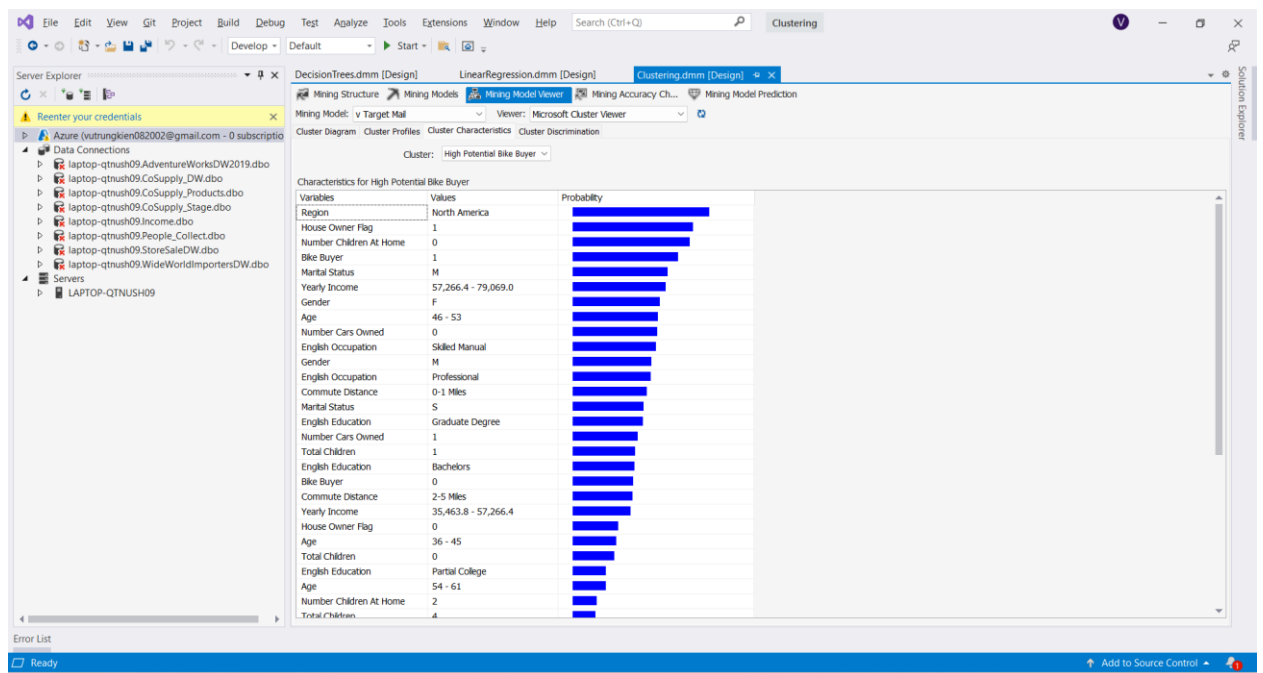
Number Cars Owned	0	2963	0.465	0.050	0.034	0.509	0.000
Number Cars Owned	3	1137	0.007	0.326	0.055	0.000	0.000
Number Cars Owned	4	871	0.000	0.279	0.013	0.000	0.000
Number Cars Owned	missing	0	0.000	0.000	0.000	0.000	0.000
Number Children At Home	0	7779	0.597	0.270	0.620	0.706	0.996
Number Children At Home	1	1749	0.172	0.075	0.321	0.067	0.000
Number Children At Home	2	1154	0.129	0.090	0.042	0.148	0.001
Number Children At Home	3	833	0.071	0.133	0.017	0.068	0.000
Number Children At Home	4	755	0.027	0.214	0.000	0.012	0.000
Number Children At Home	5	669	0.004	0.218	0.000	0.000	0.003

Number Children At Home	missing	0	0.000	0.000	0.000	0.000	0.000
Region	North America	6578	0.173	0.375	0.731	0.823	0.533
Region	Europe	3875	0.742	0.209	0.068	0.040	0.403
Region	Pacific	2486	0.085	0.416	0.201	0.137	0.064
Region	missing	0	0.000	0.000	0.000	0.000	0.000
Total Children	0	3673	0.129	0.301	0.002	0.252	0.990
Total Children	2	2617	0.252	0.112	0.459	0.123	0.006
Total Children	1	2553	0.410	0.112	0.001	0.376	0.000
Total Children	4	1597	0.056	0.180	0.208	0.137	0.000
Total Children	3	1519	0.140	0.117	0.196	0.091	0.001
Total Children	5	980	0.013	0.179	0.134	0.021	0.003
Total Children	missing	0	0.000	0.000	0.000	0.000	0.000
Yearly Income	Mean	57,266.40	27,033.44	98,791.83	57,291.59	61,456.95	32,816.68
Yearly Income	Deviation	32,324.55	11,076.79	27,762.16	19,785.44	12,211.35	15,243.82



- Ngược lại là Cluster 5. Đây là nhóm khách hàng có ít tiềm năng nhất. Những đặc điểm của cụm này là 'Number Children At Home' và 'Total Children' là 0 với tỉ lệ 0.996 và 0.99. Độ tuổi thấp khoảng 40. Có mức thu nhập dưới trung bình khoảng 32000. Đa số là người độc thân. Ta có thể đặt tên Cluster 5 là 'Low Potential Bike Buyer'
- Tương tự cho các cụm còn lại, ta có thể dựa vào số liệu và biểu đồ để có thể đặt tên và đưa ra những đặc điểm của các cụm này.
- Ở đây nhóm chúng em tập trung quan sát cho 2 cụm là 'High Potential Bike Buyer' và 'Low Potential Bike Buyer'.

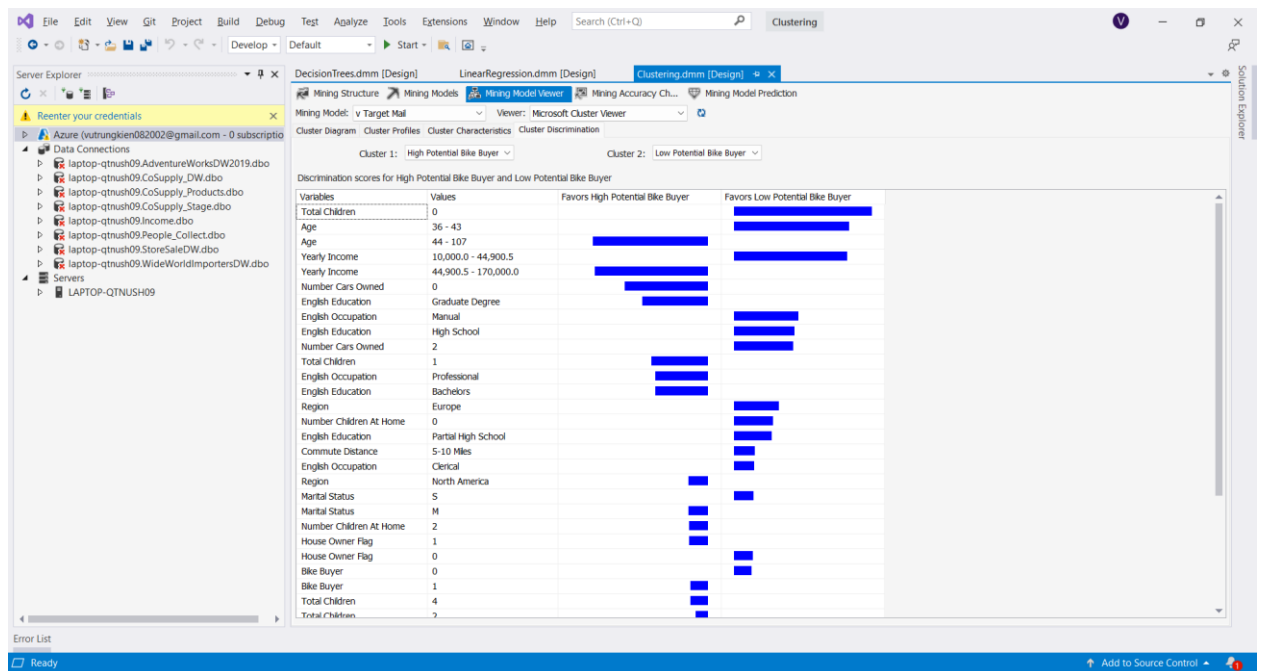




Tại đây có thể xem các đặc điểm của các cụm.

## 1.2.4. Cluster Discrimination:

Tại mục này có thể so sánh sự khác biệt giữa hai cụm.





## 2. So sánh độ chính xác của thuật toán Microsoft Logistic Regression và Microsoft Decision Tree trong dự đoán một người có mua xe đạp không?

### 2.1. Tạo Logistic Regression Project với SSAS:

#### 2.1.1. Thiết lập Mining Structure:

**Data Mining Wizard**

**Create the Data Mining Structure**  
Specify if mining model should be created and select the most applicable technique.

☒ Create mining structure with a mining model

Which data mining technique do you want to use?

Microsoft Logistic Regression

☐ Create mining structure with no models

Description:

The Microsoft Logistic Regression algorithm is a regression algorithm that works well for regression modeling. This algorithm is a particular configuration of the Microsoft Neural Network algorithm, obtained by eliminating the hidden layer. The algorithm supports the prediction of both discrete and continuous attributes.

< Back   Next >   Finish >> |   Cancel

#### 2.1.2. Thiết lập Training Data:

**Data Mining Wizard**

**Specify the Training Data**  
Specify the columns used in your analysis.

Mining model structure:

Tables/Columns	Key	Input	Predict...
vTargetMail			
AddressLine1			
AddressLine2			
Age			
BikeBuyer			
BirthDate			
CommuteDistance			
CustomerAlternateKey			
CustomerKey			
DateFirstPurchase			
EmailAddress			
EnglishEducation			
EnglishOccupation			
FirstName			
FrenchEducation			
FrenchOccupation			
Gender			
GeographyKey			
HouseOwnerFlag			
LastName			
MaritalStatus			
MiddleName			
NameStyle			
NumberCarsOwned			
NumberChildrenAtHome			

Recommend inputs for currently selected predictable: Suggest

< Back   Next >   Finish >> |   Cancel

Ở đây dự đoán một người có mua xe đạp hay không

Data Mining Wizard

**Specify Columns' Content and Data Type**  
Specify mining structure columns' content and data type.

Mining model structure:

Columns	Content Type	Data Type
Age	Discrete	Long
Bike Buyer	Discrete	Long
Customer Key	Key	Long
English Education	Discrete	Text
Gender	Discrete	Text
Marital Status	Discrete	Text
Number Cars Owned	Discrete	Long
Number Children At Home	Discrete	Long
Region	Discrete	Text
Total Children	Discrete	Long
Yearly Income	Continuous	Double

Detect continuous or discrete for numeric columns: Detect

< Back Next > Finish >>| Cancel

### 2.1.3. Processing Mining Structure:

Deployment Progress - Clustering

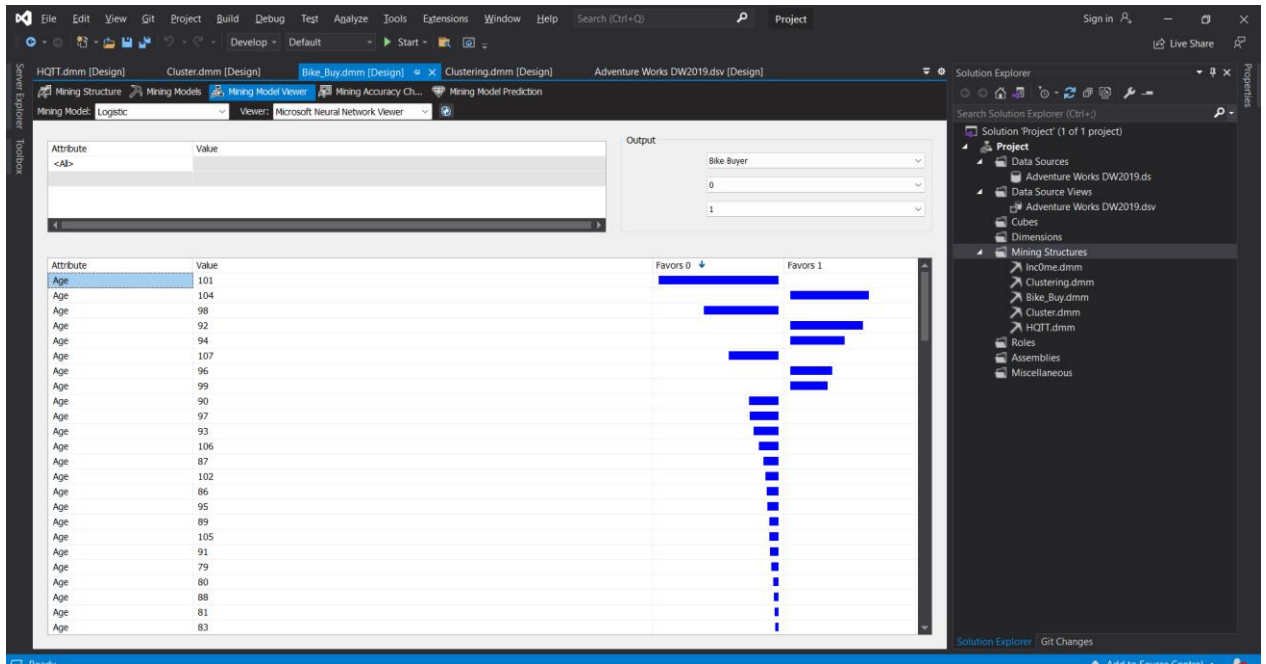
Server : localhost  
Database : Clustering

Command

- Processing Database 'Clustering' completed.  
Start time: 5/17/2023 12:07:03 PM; End time: 5/17/2023 12:07:05 PM; Duration: 0:00:02
- Processing Mining Structure 'Clustering' completed.  
Start time: 5/17/2023 12:07:03 PM; End time: 5/17/2023 12:07:03 PM; Duration: 0:00:00
  - Processing Mining Model 'v-Target Mail' completed.
- Processing Mining Structure 'LinearRegression' completed.  
Start time: 5/17/2023 12:07:03 PM; End time: 5/17/2023 12:07:05 PM; Duration: 0:00:02
  - Processing Mining Model 'LinearRegression' completed.
- Processing Cube 'LinearRegression -MC' completed.  
Start time: 5/17/2023 12:07:03 PM; End time: 5/17/2023 12:07:04 PM; Duration: 0:00:00
  - Processing Measure Group 'CaseDetail -MG' completed.
  - Processing Dimension 'LinearRegression -MC-Customer Key' completed.

Status: Deployment Completed Successfully

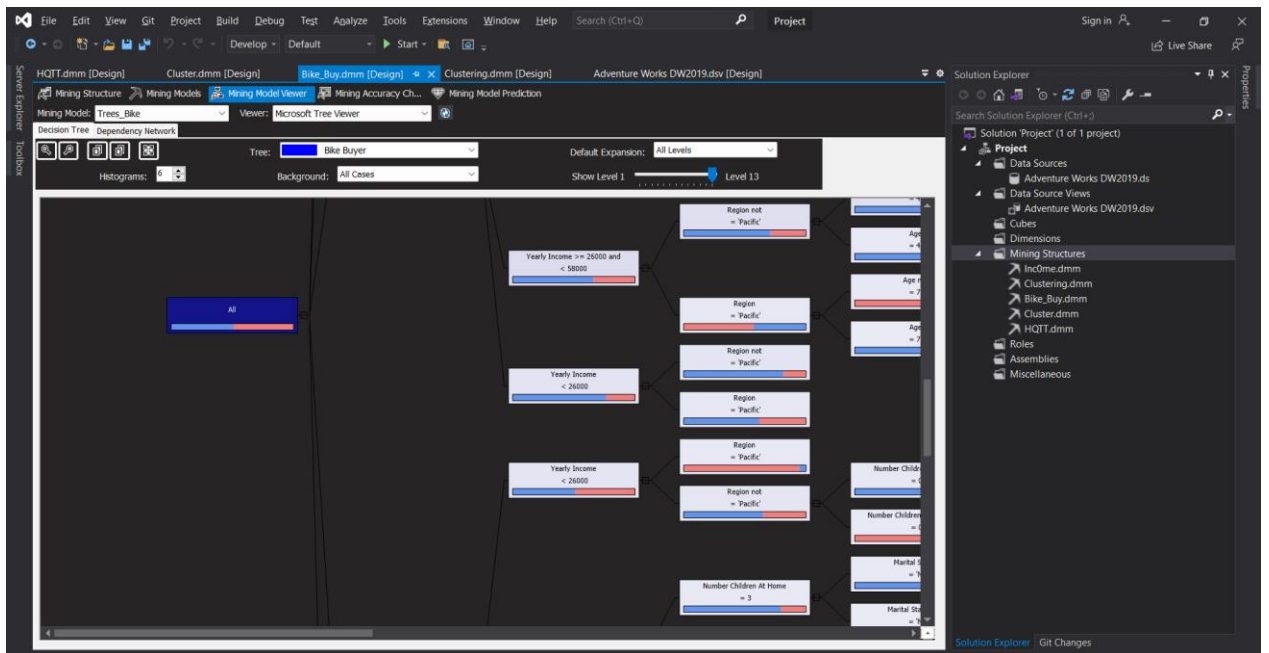
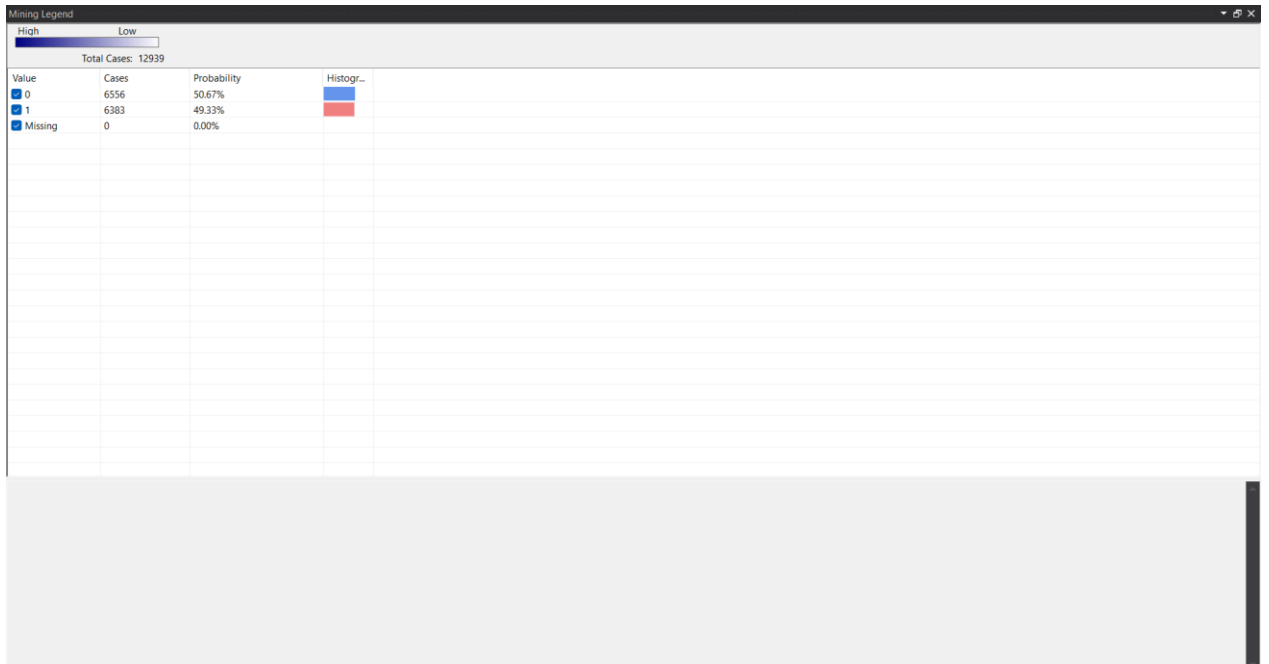
## 2.2. Deploy Logistic Regression Mining:



## 2.3. Tạo Decision Trees Mining:

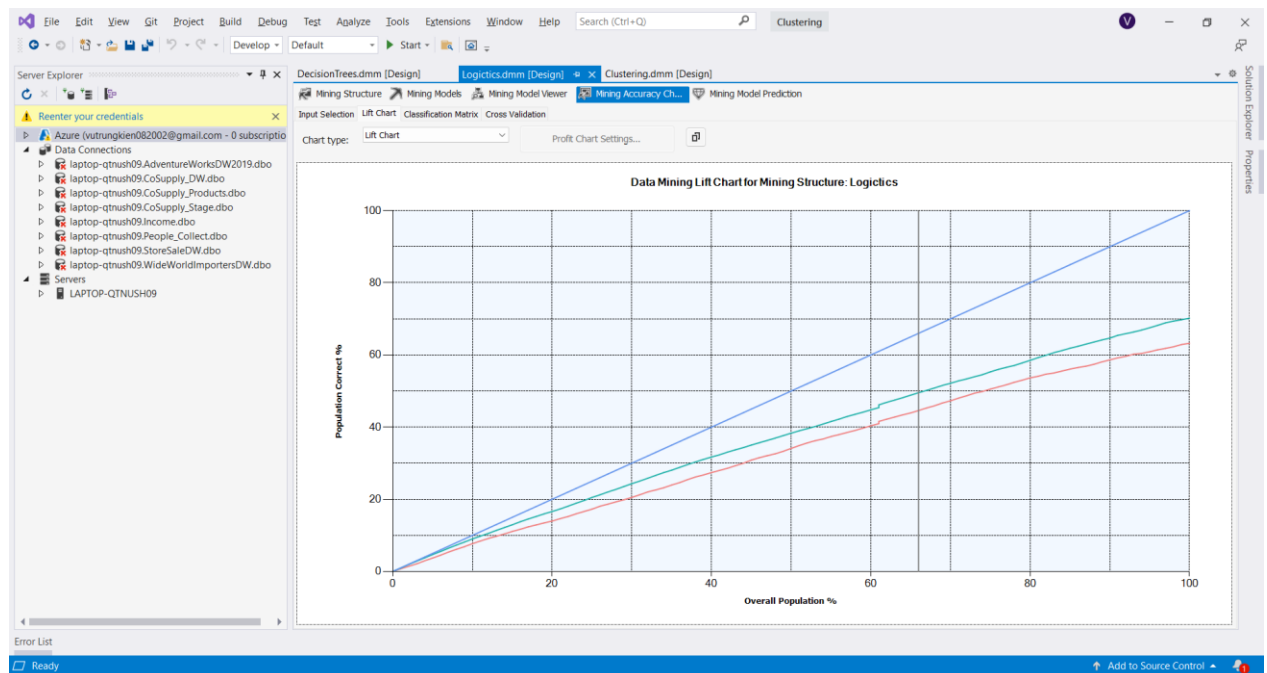
Structure	Logistic	Trees_Bike
Age	Microsoft_Logistic_Re...	Microsoft_Decision_T...
Bike Buyer	Input	Input
Commute Distance	Predict	Predict
Customer Key	Input	Input
English Education	Key	Key
English Occupation	Input	Input
Gender	Input	Input
House Owner Flag	Input	Input
Marital Status	Input	Input
Number Cars Owned	Input	Input
Number Children At Home	Input	Input
Region	Input	Input
Total Children	Input	Input
Yearly Income	Input	Input

# Mining model Viewer



## 2.4. So sánh độ chính xác giữa 2 mô hình trong việc dự đoán mua xe đạp:

### 2.4.1. Đánh giá qua biểu đồ lift chart:

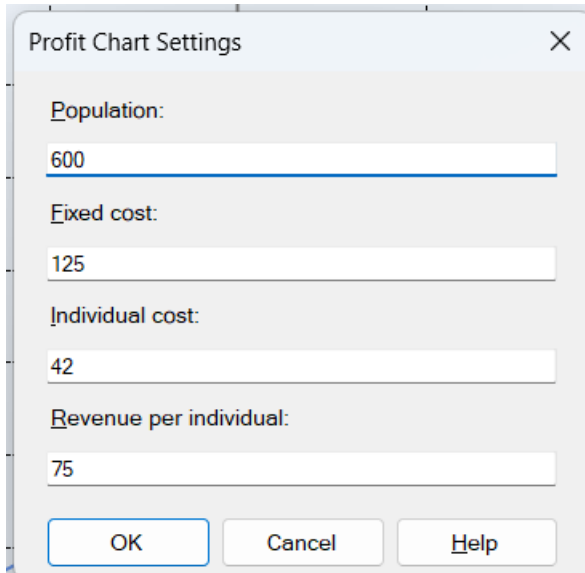


Mining Legend			
Population percentage: 49.50%			
Series, Model	Score	Population correct	Predict probability
Logistic	0.66	33.76%	59.45%
Trees_Bike	0.78	40.16%	73.67%
Ideal Model		50.00%	

Trong 2 mô hình, mô hình tốt nhất là mô hình gần với mô hình lý tưởng nhất và có số điểm đánh giá và phần trăm dự đoán cao hơn. Điều này có nghĩa là mô hình tốt nhất cho tập dữ liệu trên để người mua xe đạp là thuật toán cây quyết định.

### 2.4.2. Đánh giá qua profit chart:

Biểu đồ Lợi nhuận có phần độc đáo trong các công cụ của Microsoft. Giả sử chúng ta đang xem xét một chương trình khuyến mãi để cải thiện người mua xe đạp. Đối với một chiến dịch tiếp thị, có bốn thông số population, fixed cost (Chi phí cố định), individual cost (chi phí cá nhân) và revenue per individual (doanh thu dự kiến). Những thông số đó có thể được nhập như sau.



Profit Chart Settings

Population: 600

Fixed cost: 125

Individual cost: 42

Revenue per individual: 75

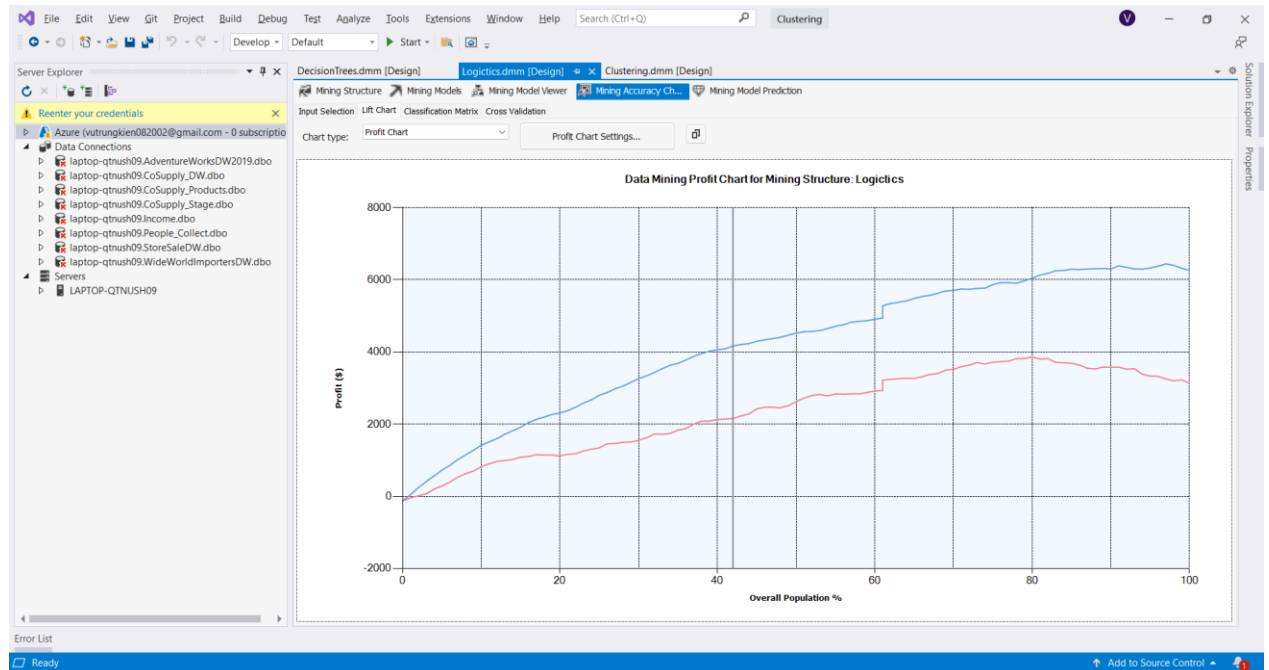
OK Cancel Help

- Giải thích các tham số:

Setting	Value	Comments
<b>Population</b>	600	Cơ sở dữ liệu có thể chứa nhiều khách hàng, nhưng để tiết kiệm chi phí gửi thư, có thể chọn chỉ nhắm mục tiêu 600 khách hàng có nhiều khả năng phản hồi nhất.
<b>Fixed cost</b>	125	Chi phí một lần để thiết lập chiến dịch gửi thư được nhắm mục tiêu cho 600 người. Điều này có thể bao gồm việc in ấn, hoặc chi phí thiết lập một chiến dịch e-mail.

<b>Individual cost</b>	42	Chi phí cho mỗi đơn vị cho chiến dịch gửi thư được nhắm mục tiêu.
<b>Revenue per individual</b>	75	Nhập một giá trị đại diện cho số tiền lãi hoặc thu nhập có thể được mong đợi từ một kết quả thành công. Trong trường hợp này, chúng tôi giả định rằng việc gửi danh mục qua thư sẽ dẫn đến việc mua phụ kiện hoặc xe đạp với giá trung bình là 400 đô la.

## Hiểu thị biểu đồ lợi nhuận:



Mining Legend		
Population percentage: 49.50%		
Series, Model	Profit	Predict probab...
LogitcticsRegr...	\$2,621.26	58.07%
DecisionTrees	\$4,520.27	68.22%

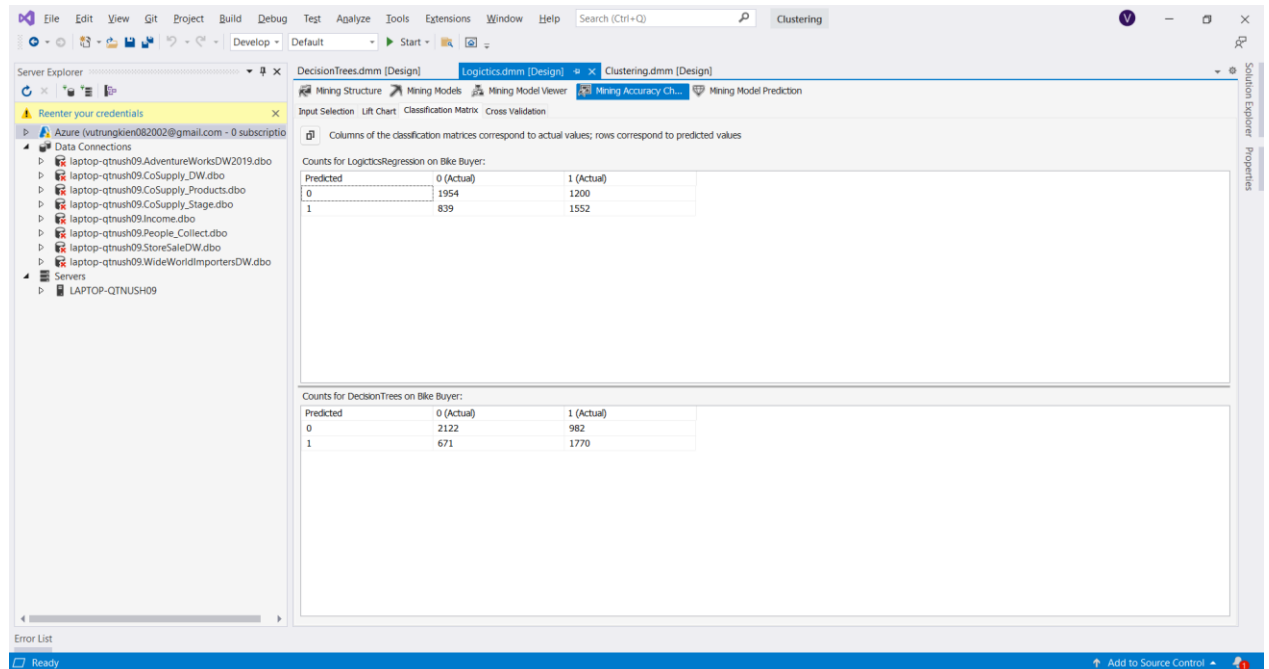
Trục Y của biểu đồ biểu thị lợi nhuận, trong khi trục X biểu thị tỷ lệ phần trăm khách hàng đã được liên hệ bởi chiến dịch gửi thư được nhắm mục tiêu.

Ngay cả biểu đồ lợi nhuận cũng chỉ ra rằng thuật toán Cây quyết định tốt hơn. Trong cây quyết định, 49.5% Population sẽ kiếm được lợi nhuận tối đa là 5.518 đô la. Điều này có nghĩa là tốt hơn là chỉ nhắm mục tiêu 45% người mua có thể có xác suất mua cao nhất.



### 2.4.3. Ma Trận Phân Loại

Mô hình hồi quy Logistic, Mô hình cây quyết định:



Ta thấy tỉ lệ dự đoán chính xác số trường hợp mua xe trong thuật toán cây quyết định là 71,7% cao hơn tỉ lệ dự đoán trong thuật toán hồi quy logistic 62.5%

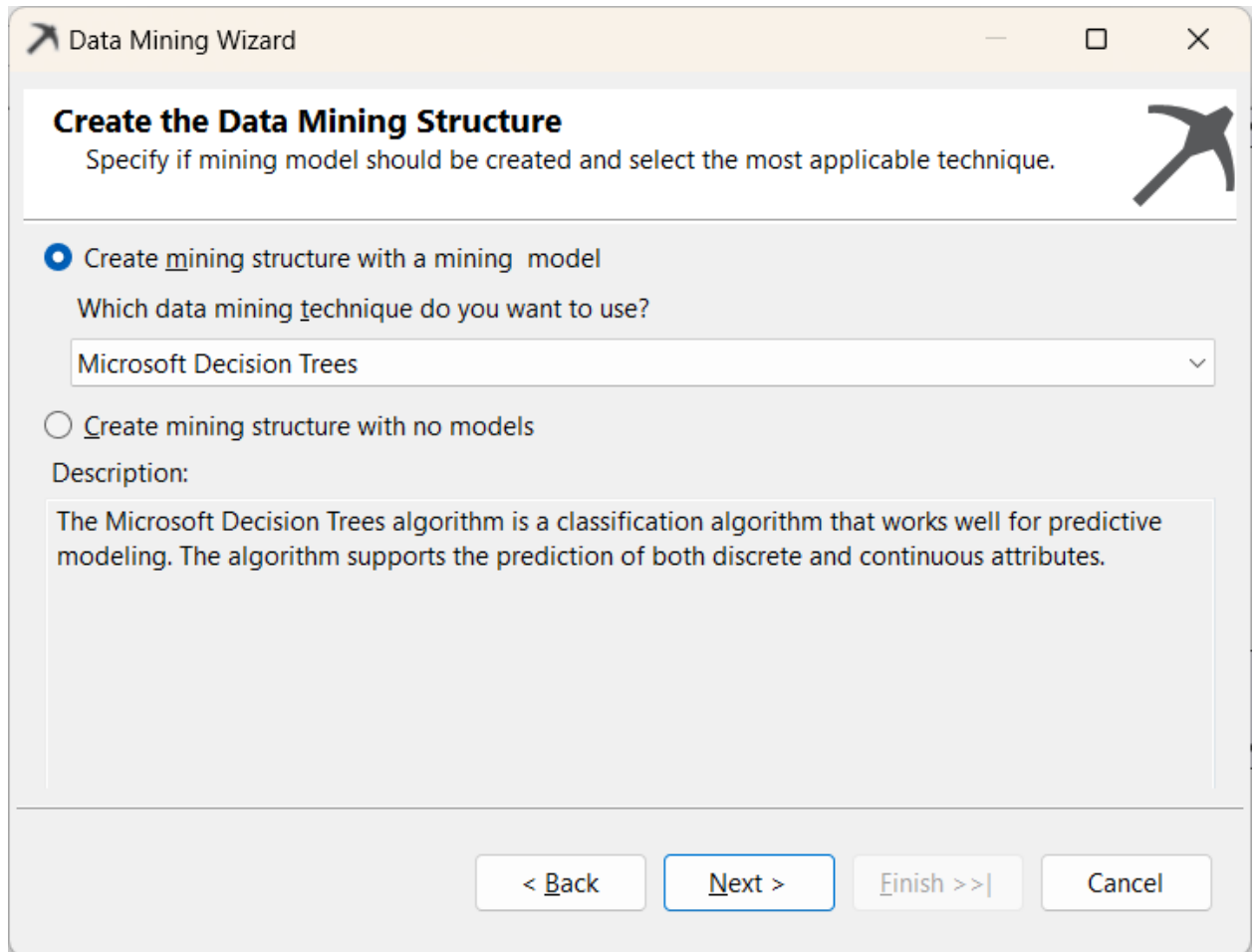
Tỉ lệ dự đoán chính xác số trường hợp không mua xe trong thuật toán cây quyết định là 70,3% cao hơn tỉ lệ dự đoán trong thuật toán hồi quy logistic 62.1%

Vì vậy theo ma trận phân loại thì thuật toán cây quyết định vẫn tối ưu hơn cho dự đoán người mua xe đạp

### 3. Thuật toán Microsoft Decision Tree:

#### 3.1. Tạo Decision Trees Project với SSAS:

##### 3.1.1. Thiết lập Mining Structure:



The screenshot shows the 'Data Mining Wizard' window. The title bar reads 'Data Mining Wizard'. The main heading is 'Create the Data Mining Structure', with a subtitle 'Specify if mining model should be created and select the most applicable technique.' and a pickaxe icon. There are two radio button options: 'Create mining structure with a mining model' (selected) and 'Create mining structure with no models'. Below the first option is a dropdown menu labeled 'Which data mining technique do you want to use?' with 'Microsoft Decision Trees' selected. Below the second option is a 'Description:' section containing text about the Microsoft Decision Trees algorithm. At the bottom are four buttons: '< Back', 'Next >' (highlighted with a blue border), 'Finish >>|', and 'Cancel'.

**Create the Data Mining Structure**  
Specify if mining model should be created and select the most applicable technique.

☒ Create mining structure with a mining model  
Which data mining technique do you want to use?  
Microsoft Decision Trees

☐ Create mining structure with no models  
Description:  
The Microsoft Decision Trees algorithm is a classification algorithm that works well for predictive modeling. The algorithm supports the prediction of both discrete and continuous attributes.

< Back   Next >   Finish >>|   Cancel

### 3.1.2. Thiết lập Training Data:

**Data Mining Wizard**

**Select Data Source View**  
Select the data source view to provide the data for the mining structure.

Available data source views:

- AdventureWorksDW2019\_Views
- Adventure Works DW2019
- PeopleCollect\_Views**

Browse...

Tables:  
income

< Back   Next >   Finish >>|   Cancel

**Data Mining Wizard**

**Specify the Training Data**  
Specify the columns used in your analysis.

Mining model structure:

Tables/Columns	Key	Input	Predict...
income			
age	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
capital_gain	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
capital_loss	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
education	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
education_num	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
fnlwgt	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
hours_per_week	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Id	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
income	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
marital_status	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
native_country	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
occupation	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
race	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
relationship	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
sex	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
workclass	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Recommend inputs for currently selected predictable: Suggest

< Back   Next >   Finish >>|   Cancel

### 3.2.1. Decision Trees Diagram



Nhận xét:

- Khoảng Education num  $< 9$  : thu nhập dưới 50K \$.
- Khoảng Education num  $\in [9; 11)$ :
- Nếu người đó có số tuổi Age  $< 33$  hoặc là Age  $\geq 65$  thì họ có thu nhập dưới 50K \$ là xác suất rất cao.
- Từ độ tuổi  $[33, 41)$  nếu người đó có số Education num = 9 hoặc 10 kết hợp với điều kiện Capital Loss  $< 1742,4$  (ngoài điều kiện này thì ngược lại) thì họ có thu nhập dưới 50K \$ là xác suất rất cao.
- Từ độ tuổi  $[41, 65)$  nếu họ thuộc vào điều kiện Capital Gain  $< 6999,93$  và Capital Loss  $\geq 2000,3$  và Capital  $< 1742,4$  thì có thể xác định họ thuộc nhóm có thu nhập dưới 50K \$.

- Khoảng Education num  $\in [11; 13)$ :
- Nếu người đó có số tuổi Age < 33 thì họ có thu nhập dưới 50K \$ là xác suất rất cao.
- Tuy nhiên đối với người có số tuổi Age  $\geq 33$  mà thuộc vào các khoảng Capital Gain < 6999,93 và Capital Loss < 1742,4 thì xác suất họ có thu nhập dưới 50K \$ là khá cao.
- Khoảng Education num  $\in [13, 15)$ :
- Nếu người có Age < 28 thì họ có thu nhập dưới 50K \$.
- Nếu người có Age thuộc [28, 33) và Capital Gain < 9999,9 thì xác suất khá cao họ thuộc nhóm có thu nhập dưới 50K \$.
- Nếu người có Age  $\geq 33$  với Capital Gain < 2999,79 và Capital Loss thuộc [1568,16 ; 1742,4) thì họ là người có thu nhập dưới 50K \$.
- Khoảng Education num  $\geq 15$ : có Age < 33, Capital Gain < 9999,9 và Capital Loss < 1742,4 thì khả năng cao họ có thu nhập dưới 50K \$.

Thực hiện xem đánh giá mô hình dựa vào confusion matrix:

Predicted	>50K (Actual)	<=50K (Actual)
>50K	597	86
<=50K	1707	6658

Độ chính xác của mô hình 80.34%

Thực hiện dự đoán cho income:

Thử với Education num < 9

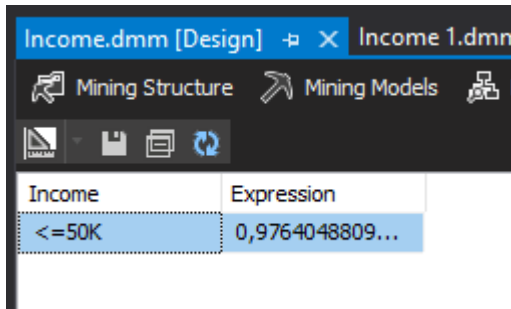
The screenshot shows the Orange Data Mining software interface. The main window is titled "Income.dmm [Design]". The "Mining Model Prediction" tab is active. On the left, the "Mining Model" tree shows the "Income" model selected. On the right, the "Singleton Query Input" table is displayed with the following data:

Mining Model Column	Value
Age	32
Capital Gain	10000
Capital Loss	5000
Education Num	8
Income	

Below the table, there is a table with columns Source, Field, Alias, Show, Group, And/Or, and Criteria/Argument. The table shows the Income field being predicted with a probability.

Source	Field	Alias	Show	Group	And/Or	Criteria/Argument
Income	Income		<input checked="" type="checkbox"/>			
Prediction Function	PredictProbability		<input checked="" type="checkbox"/>			[Income].[Income]

Kết quả cho ra 97.64% thì người này thuộc thu nhập dưới 50K \$.

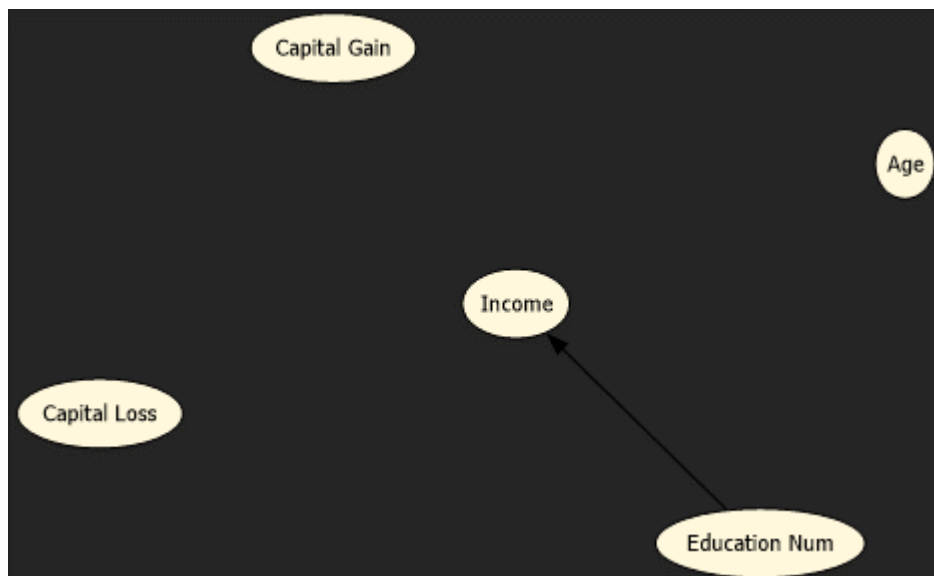


The screenshot shows the 'Mining Structure' pane in SSAS. It displays a table with two columns: 'Income' and 'Expression'. The 'Income' column has a value of '<=50K' and the 'Expression' column has a value of '0,9764048809...'. The table is part of a larger interface with tabs for 'Income.dmm [Design]' and 'Income 1.dmm'.

Income	Expression
<=50K	0,9764048809...

### 3.2.2. Dependency Network:

Chỉnh ở mức liên kết mạnh nhất thì ta thấy rằng Education num gây ảnh hưởng đến việc thu nhập.



## IV. TỔNG KẾT:

### 1. Kết quả đạt được:

- Học hỏi thêm các kiến thức về cách sử dụng công cụ SSAS, Python.
- Nắm rõ các kiến thức cơ bản về các thuật toán mà nhóm sử dụng để thực hiện đề tài.
- Khai phá được nhiều những tri thức mới trong tập dữ liệu.

### 2. Hạn chế:

- Am hiểu về các giải thuật chưa chuyên sâu.
- Chưa tối ưu tham số cho thuật toán.
- Kết luận còn khá mơ hồ.
- Chưa thực hiện trực quan hoá trên Power BI.

### **3. Phương hướng phát triển:**

- Hoàn thiện các thuật toán đưa các thuật toán đến mức tối ưu nhất.
- Trực quan hoá để có cái nhìn khách quan hơn kết quả khi thực thi các thuật toán.
- Áp dụng thêm nhiều giải thuật mới và chọn ra thuật toán ổn định nhất.

### **V. THAM KHẢO:**

- Nguồn tài liệu từ các slide và ví dụ của thầy Nguyễn Văn Thành.
- Nguồn tài liệu hướng dẫn:

<https://www.sqlshack.com/the-association-rule-mining-in-sql-server/>

- Nguồn Youtube hướng dẫn:

<https://www.youtube.com/watch?v=o8tEEBy5zjQ>