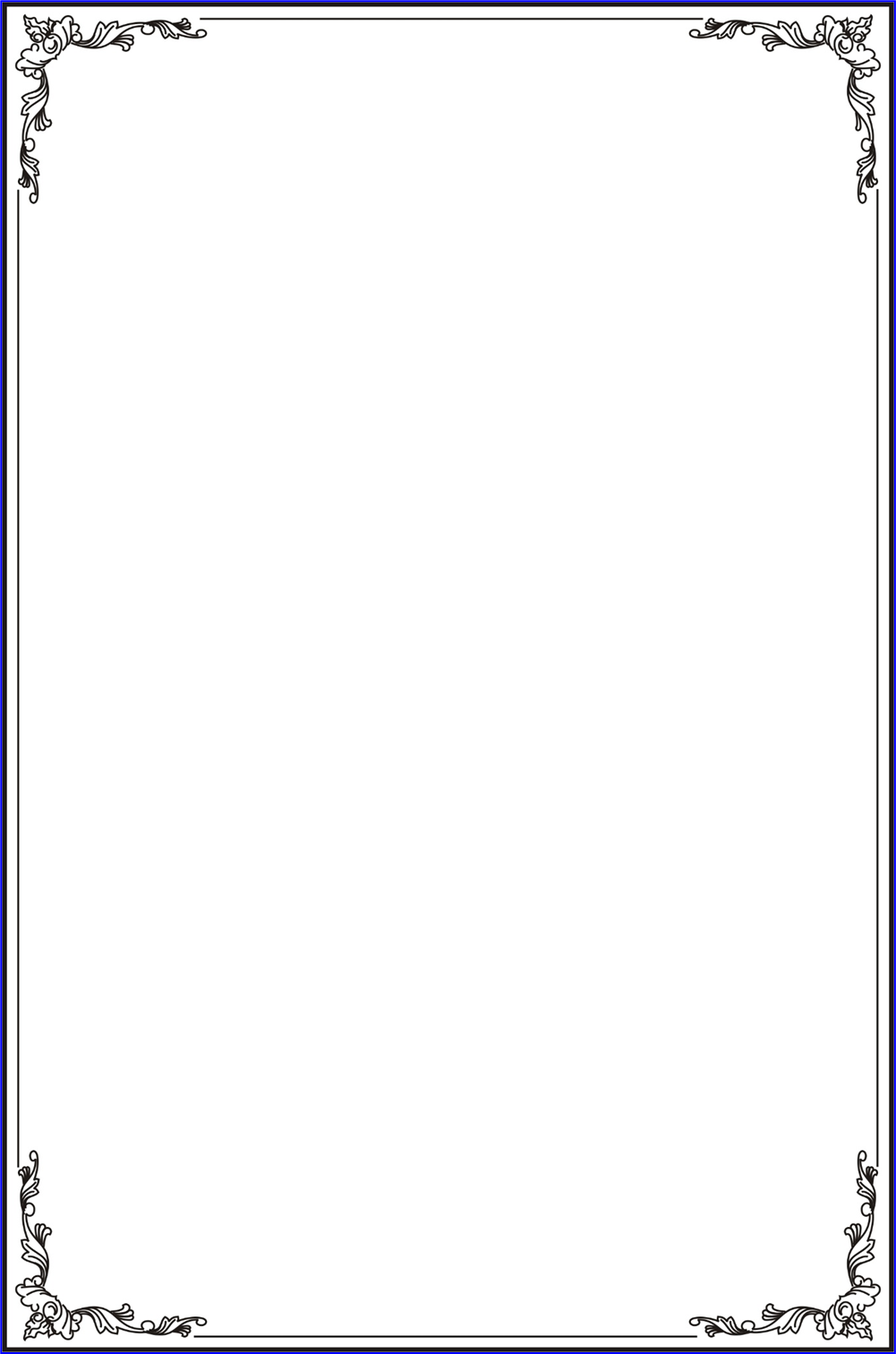
****

****

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KĨ THUẬT TP.HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**



**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**

**SV thực hiện: Lê Phước Yên MSSV: 20133119**

**Phan Thanh Tín MSSV: 20133095**

**Vũ Trung Kiên MSSV: 20133060**

**Nguyễn Đức Linh MSSV: 20133007**

**Tp. Hồ Chí Minh, tháng 5 năm 2023**

**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**

[**I.** **TỔNG QUAN ĐỀ TÀI:** 5](#_Toc135815554)

[**1.** **Bảng phân công nhiệm vụ:** 5](#_Toc135815555)

[**2.** **Lý do chọn đề tài:** 7](#_Toc135815556)

[**3.** **Giới thiệu về tập dữ liệu:** 8](#_Toc135815557)

[**3.1.** **Tập dữ liệu trong AdventureWorksDW2019:** 8](#_Toc135815558)

[**3.2.** **Tập dữ liệu về thu thập cá nhân:** 9](#_Toc135815559)

[**4.** **Công cụ sử dụng:** 9](#_Toc135815560)

[**4.1.** **Visual Studio 2019:** 9](#_Toc135815561)

[**4.2.** **Visual Studio Code:** 10](#_Toc135815562)

[**4.3.** **SQL Server 2019:** 10](#_Toc135815563)

[**4.4.** **Ngôn ngữ lập trình (Python, SQL):** 10](#_Toc135815564)

[**4.5.** **SSAS:** 10](#_Toc135815565)

[**5.** **Thuật toán sử dụng:** 11](#_Toc135815566)

[**II.** **CHUẨN BỊ DỮ LIỆU:** 11](#_Toc135815567)

[**1.** **Dữ liệu gốc:** 11](#_Toc135815568)

[**1.1.** **Tập dữ liệu trong AdventureWorksDW2019:** 11](#_Toc135815569)

[**1.2.** **Tập dữ liệu về thu thập cá nhân:** 12](#_Toc135815570)

[**2.** **Tiền xử lý dữ liệu về thu thập cá nhân:** 12](#_Toc135815571)

[**3.** **Import vào Database:** 14](#_Toc135815572)

[**3.1. Import tập dữ liệu AdventureWorksDW2019 14**](#_Toc135815573)

[**3.2. Import dữ liệu về thu nhập cá nhân 15**](#_Toc135815574)

[**4. Xử lý dữ liệu để chọn cụm tối ưu 16**](#_Toc135815575)

[**4.1.** **Tập dữ liệu trong AdventureWorksDW2019:** 16](#_Toc135815576)

[**4.2.** **Tập dữ liệu về thu thập cá nhân:** 18](#_Toc135815577)

[**III.** **THỰC HIỆN CÁC GIẢI THUẬT:** 19](#_Toc135815578)

[**1.** **Thuật toán Microsoft Clustering:** 19](#_Toc135815579)

[**1.1.** **Tạo Clustering Project với SSAS:** 19](#_Toc135815580)

[**1.2.** **Deploy Clustering Mining:** 23](#_Toc135815581)

[**2.** **So sánh độ chính xác của thuật toán Microsoft Logictic Regession và Microsoft Decision Tree trong dự đoán một người có mua xe đạp không?** 33](#_Toc135815582)

[**2.1.** **Tạo Logistic Regression Project với SSAS:** 33](#_Toc135815583)

[**2.2.** **Deploy Logistic Regression Mining:** 35](#_Toc135815584)

[**3.** **Thuật toán Microsoft Decision Tree:** 42](#_Toc135815585)

[**3.1.** **Tạo Decision Trees Project với SSAS:** 42](#_Toc135815586)

[**3.2.** **Deploy Decision Trees Mining:** 44](#_Toc135815587)

[**IV.** **TỔNG KẾT:** 46](#_Toc135815588)

[**1.** **Kết quả đạt được:** 46](#_Toc135815589)

[**2.** **Hạn chế:** 46](#_Toc135815590)

[**3.** **Phương hướng phát triển:** 47](#_Toc135815591)

[**V.** **THAM KHẢO:** 47](#_Toc135815592)

1. **TỔNG QUAN ĐỀ TÀI:**
2. **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 Logictis Regession sử dụng SSAS | 100% |  | 100% | 100% |
| Đọc kết quả và nhận xét cho thuật toán Logictis Regession | 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% |  |

1. **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.

1. **Giới thiệu về tập dữ liệu:**
   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.
* BrithDate: 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 dòng2 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.
  1. **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.

1. **Công cụ sử dụng:**
   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.

* 1. **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.

* 1. **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.

* 1. **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.

* 1. **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.

1. **Thuật toán sử dụng:**

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

* Microsoft Clustering: Được sử dụng để nhóm dữ liệu vào các nhóm (hay cụm) mang tính đồng nhất.
* Microsoft Logistic Regession: 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.

1. **CHUẨN BỊ DỮ LIỆU:**
2. **Dữ liệu gốc:**
   1. **Tập dữ liệu trong AdventureWorksDW2019:**









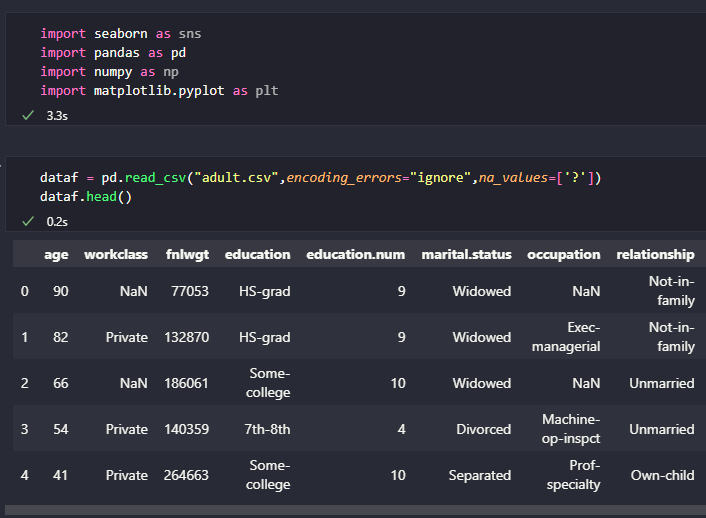
* 1. **Tập dữ liệu về thu thập cá nhân:**



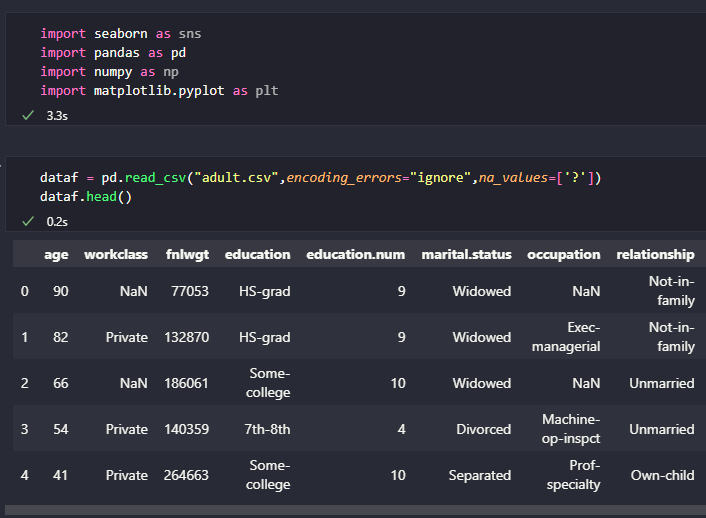


1. **Tiền xử lý dữ liệu về thu thập cá nhân:**

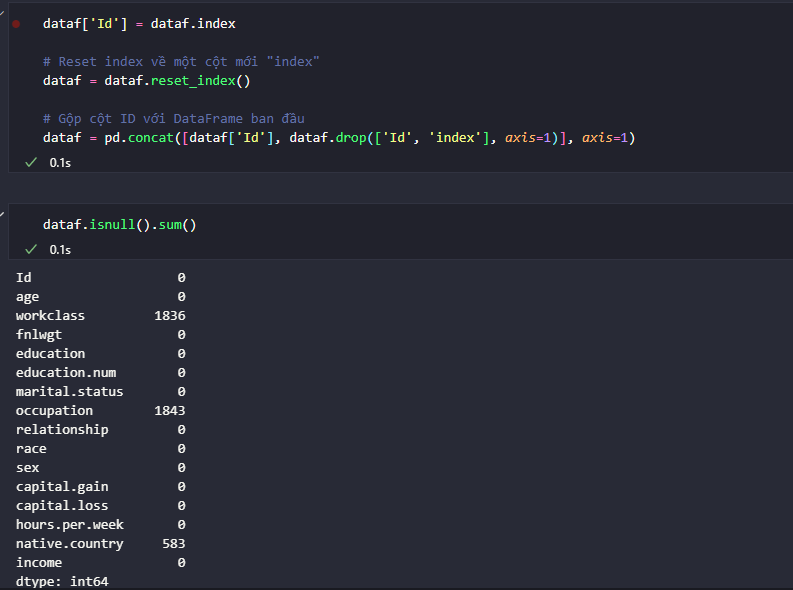
Import thư viện



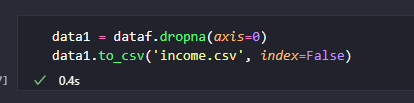
Đọc dữ liệu và quan sát tổng quan



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



Drop NA và xuất ra file CSV



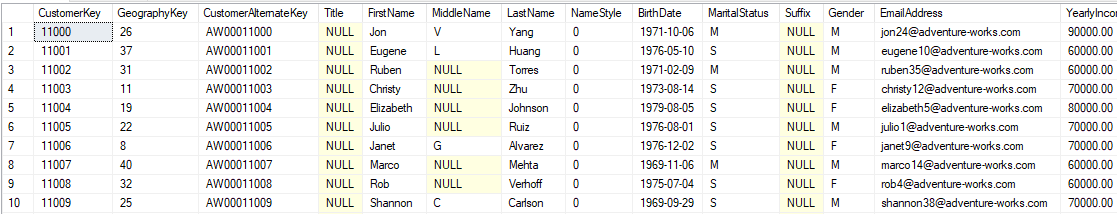
File sau khi tiền xử lý

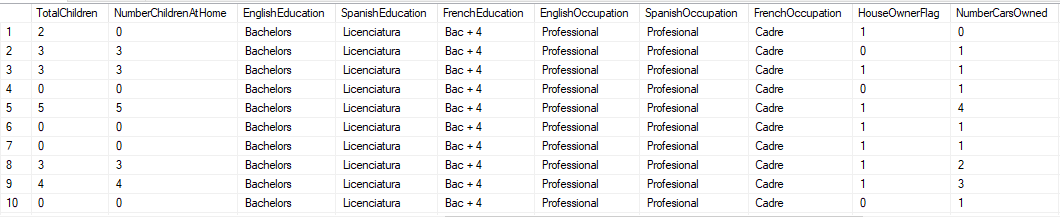


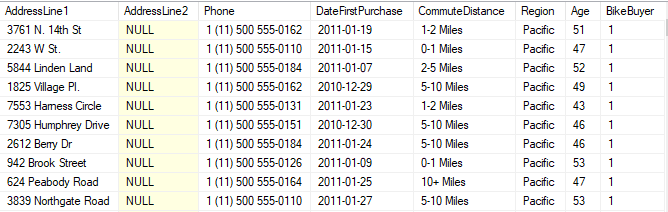


1. **Import vào Database:**

### Import tập dữ liệu AdventureWorksDW2019

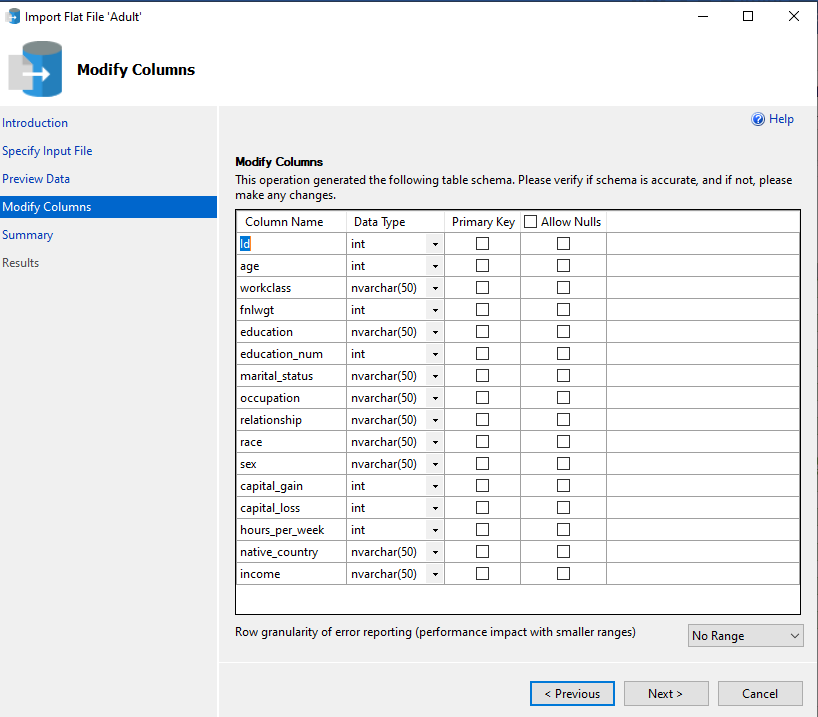




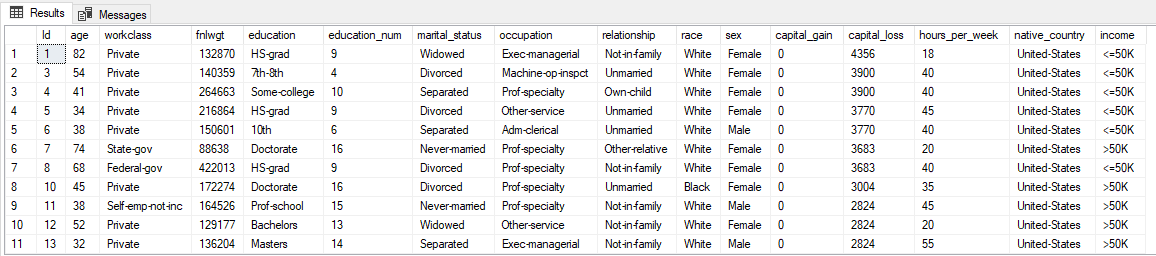


### Import dữ liệu về thu nhập cá nhân

Chọn kiểu dữ liệu



Kết quả



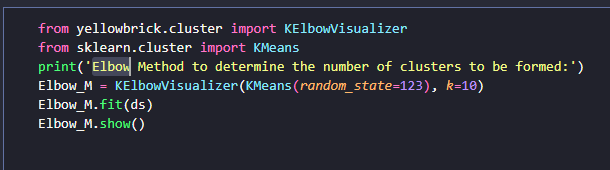
## Xử lý dữ liệu để chọn cụm tối ưu

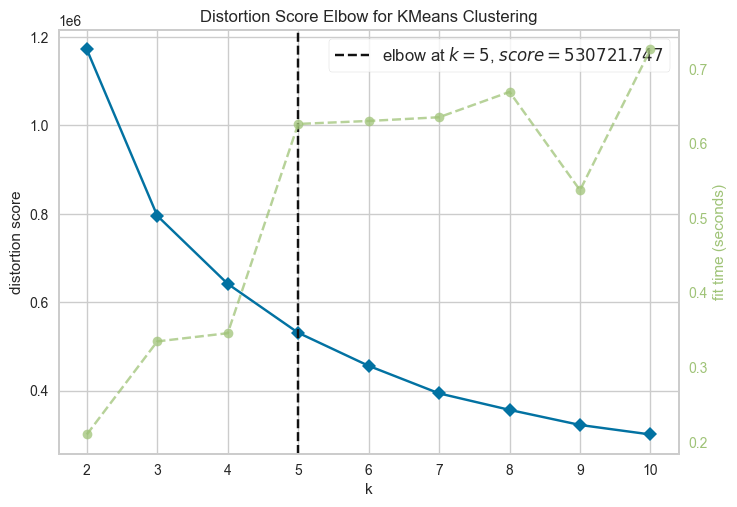
* 1. **Tập dữ liệu trong AdventureWorksDW2019:**

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



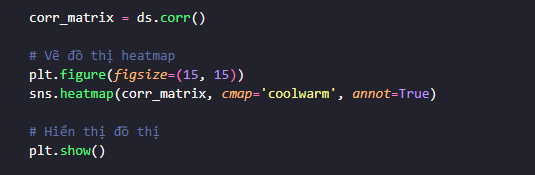
Vẽ biểu đồ Elbow

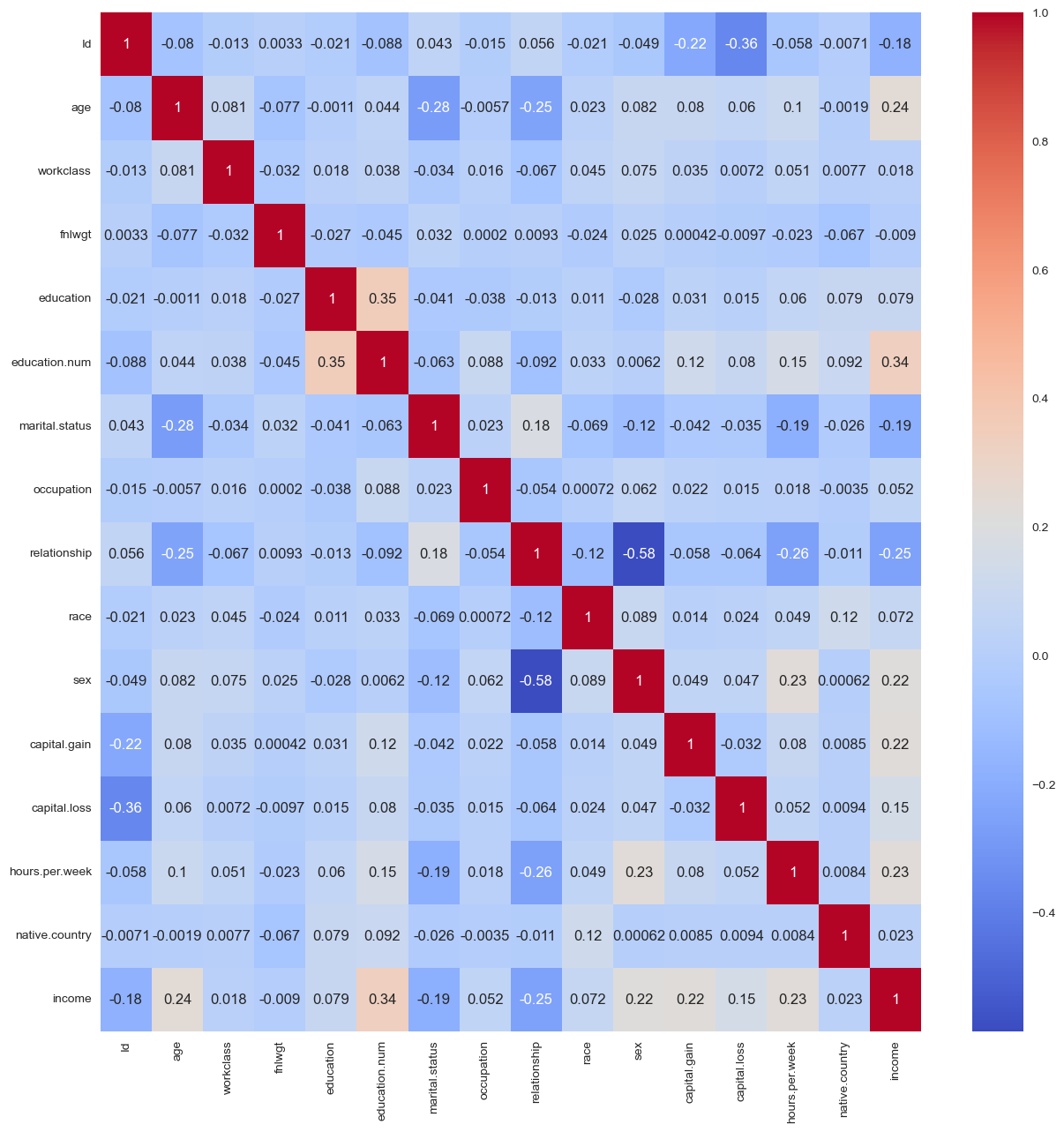




* 1. **Tập dữ liệu về thu thập cá nhân:**

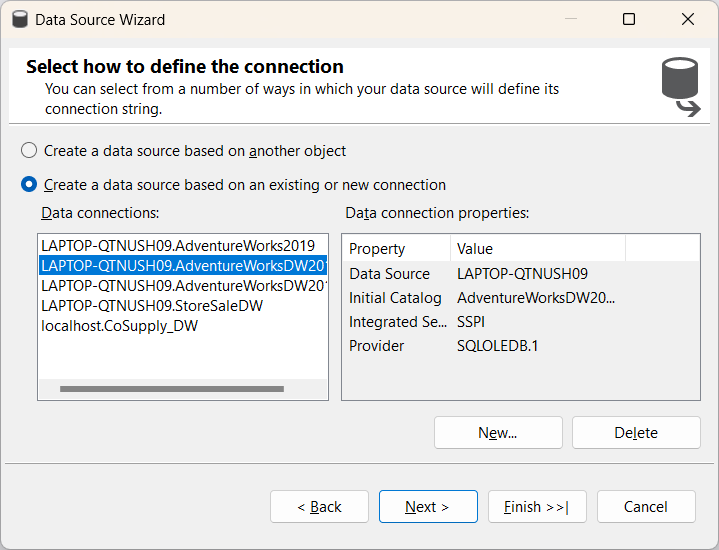
Vẽ đồ thị heatmap





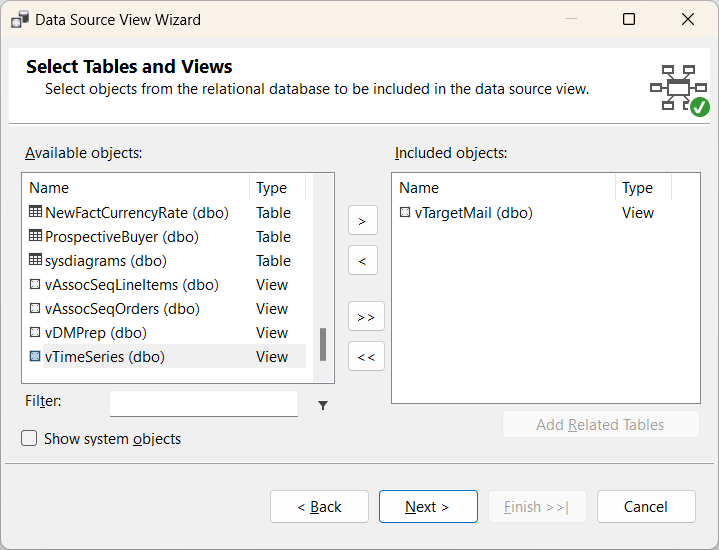
1. **THỰC HIỆN CÁC GIẢI THUẬT:**
2. **Thuật toán Microsoft Clustering:**
   1. **Tạo Clustering Project với SSAS:**
      1. **Thiết lập Data Source:**

Chọn Database là AdventureWorks2019:



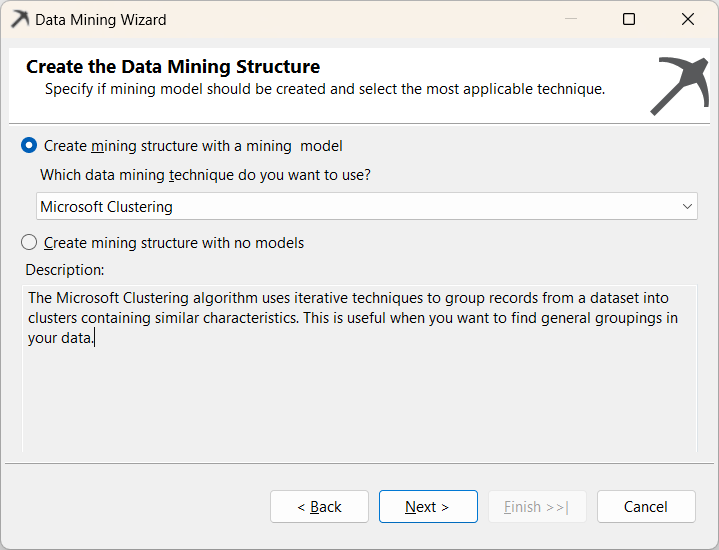
* + 1. **Thiết lập Data Source Views:**

Chọn Object là vTargetMail

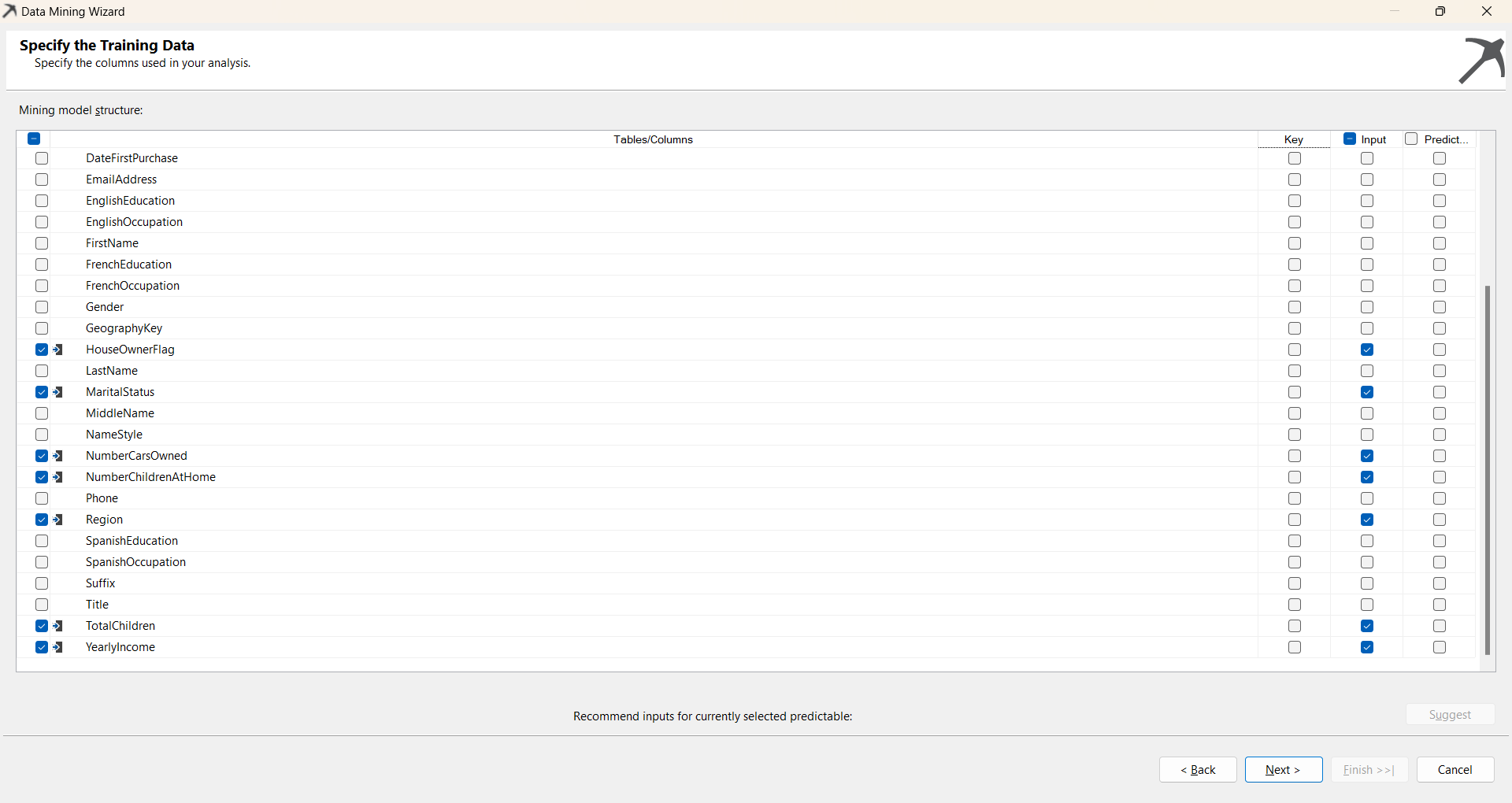


* + 1. **Thiết lập Mining Structures:**

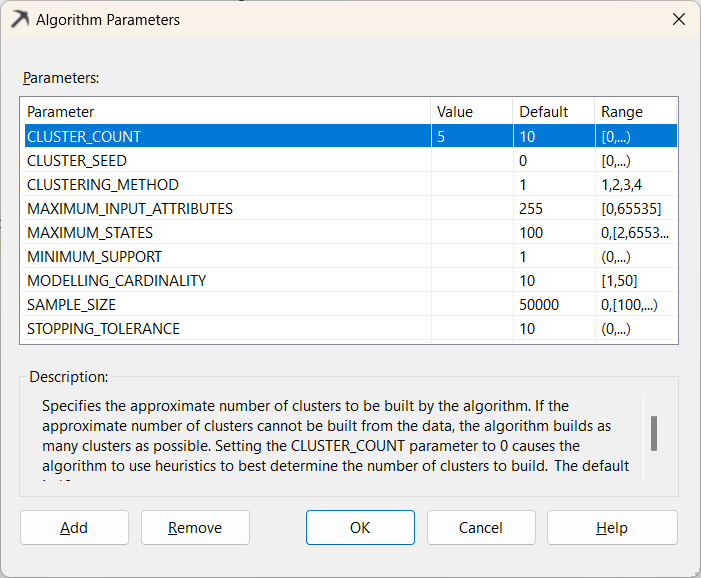
Chọn thuật toán là Clustering



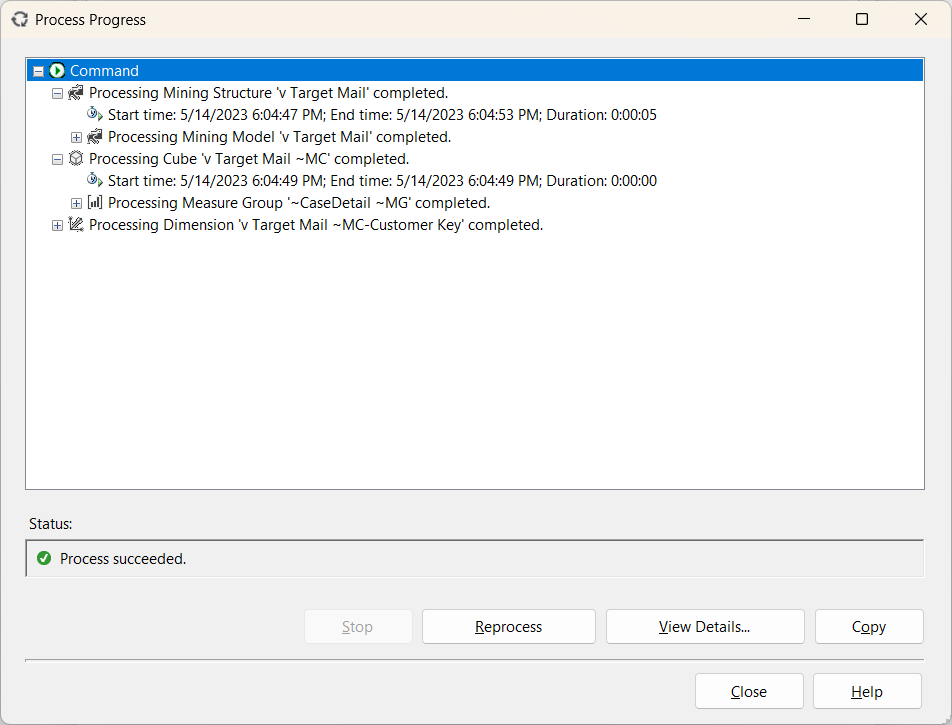
* + 1. **Thiết lập Traning Data:**



* + 1. **Thiết lập Algorithm Parameters:**



* + 1. **Processing Mining Structure:**

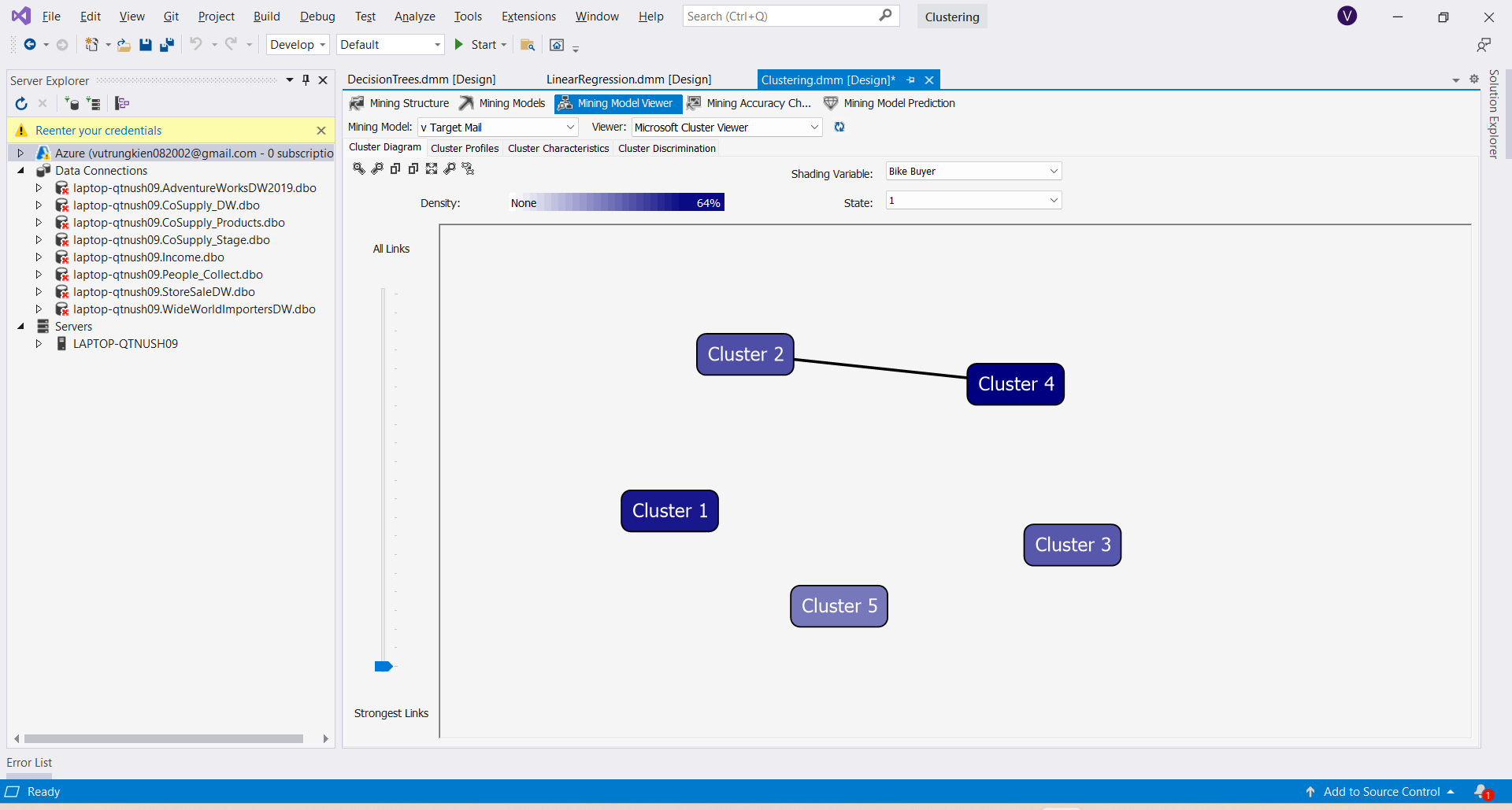


* 1. **Deploy Clustering Mining:**



* + 1. **Cluster Diagram:**

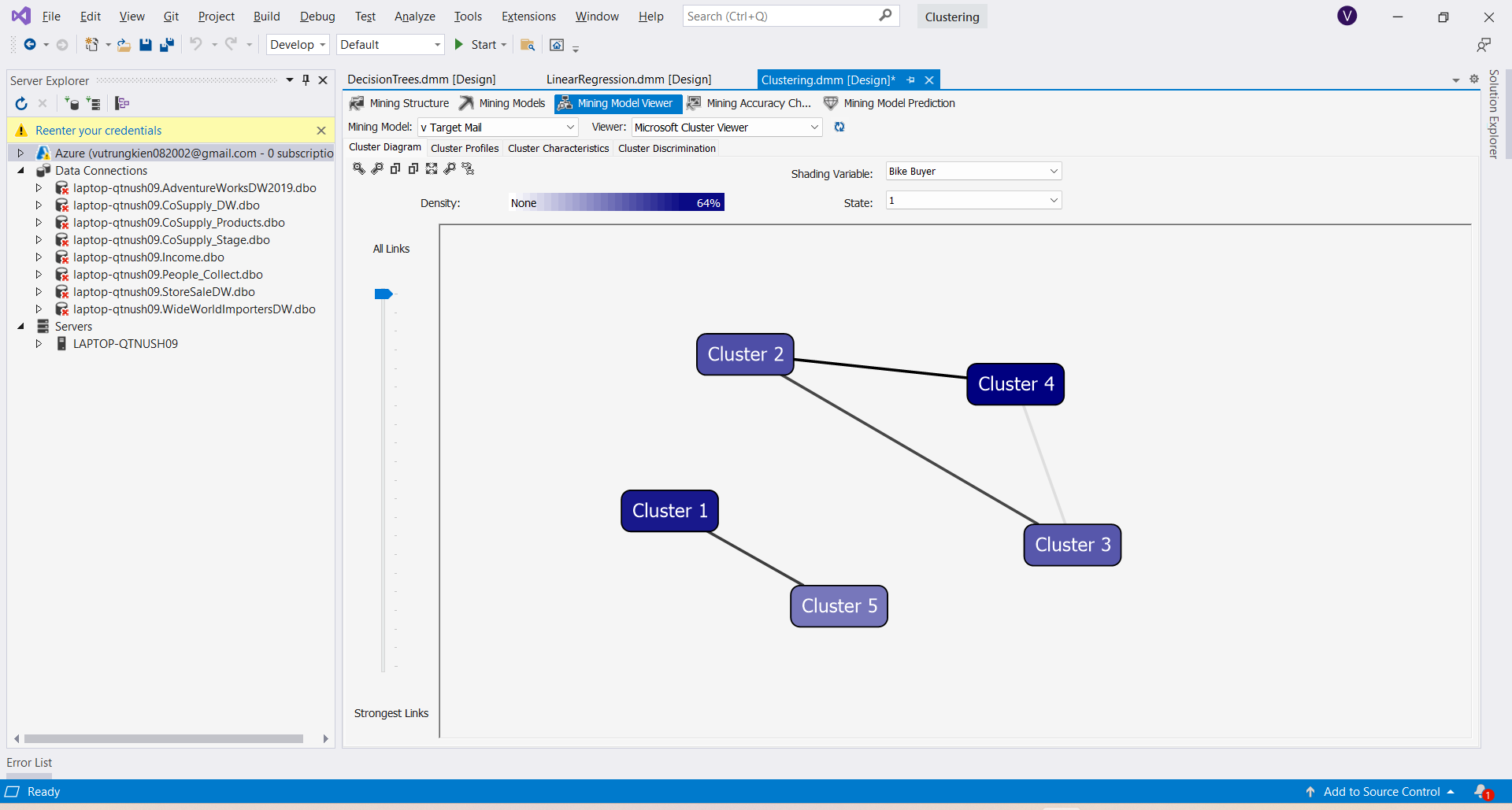
Chọn Shading Variable là Bike Buyer state 1 (những người mua xe đạp)



Tùy chọn mức liên kết mạnh nhất.

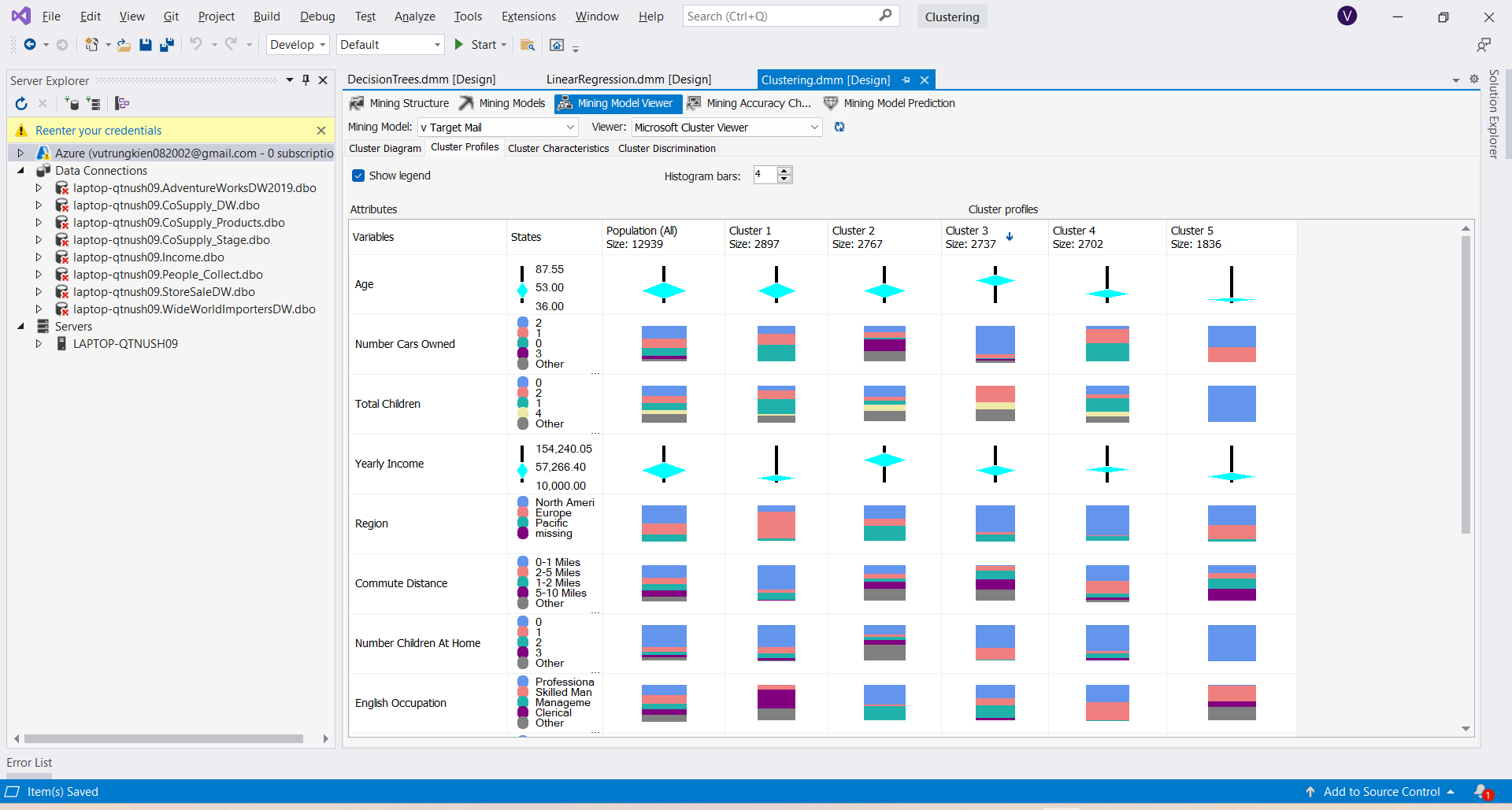


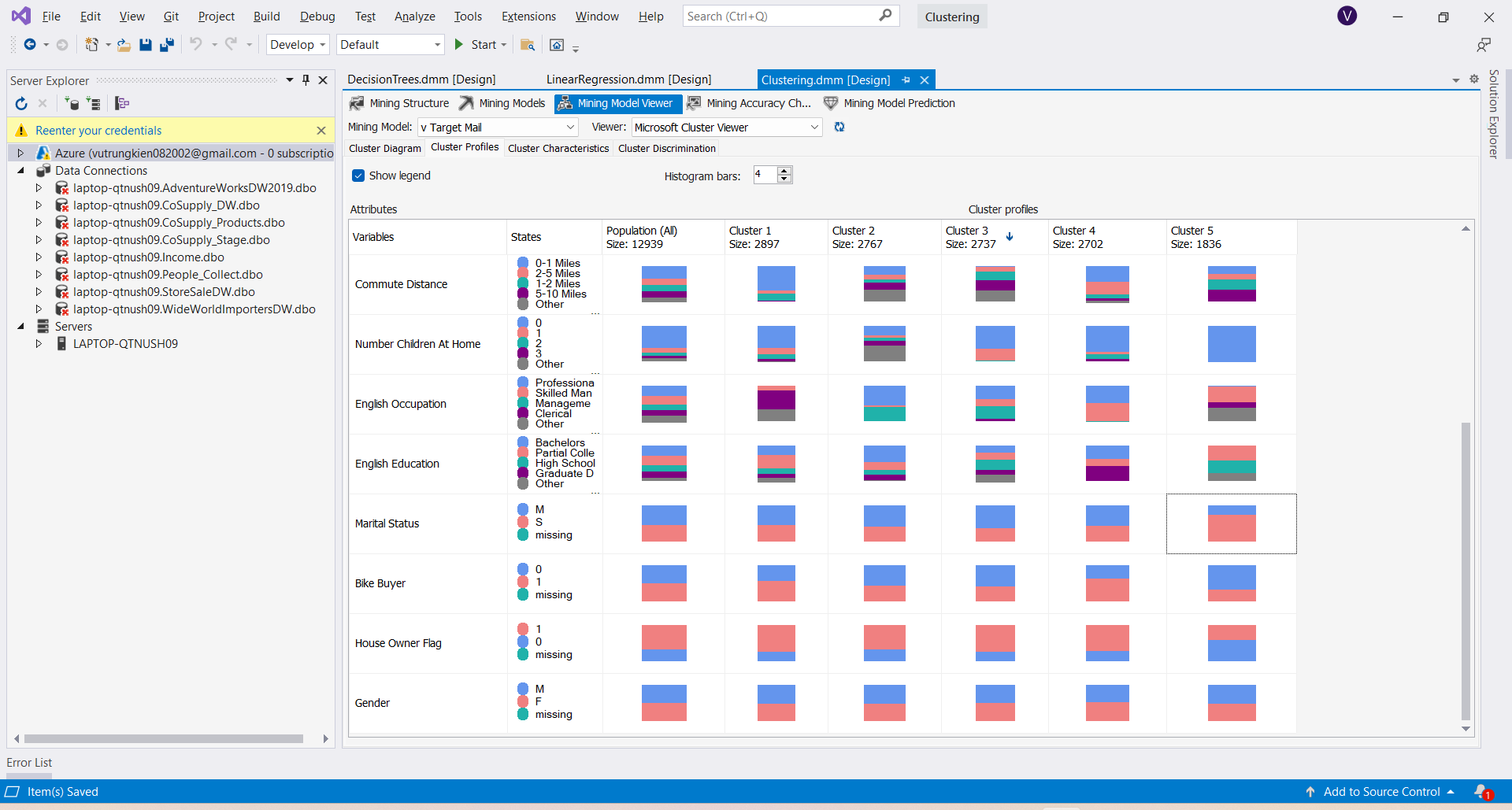
Tùy chọn mức liên kết yếu nhất.



* + 1. **Clustering Profile:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variables | States | Population (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 |

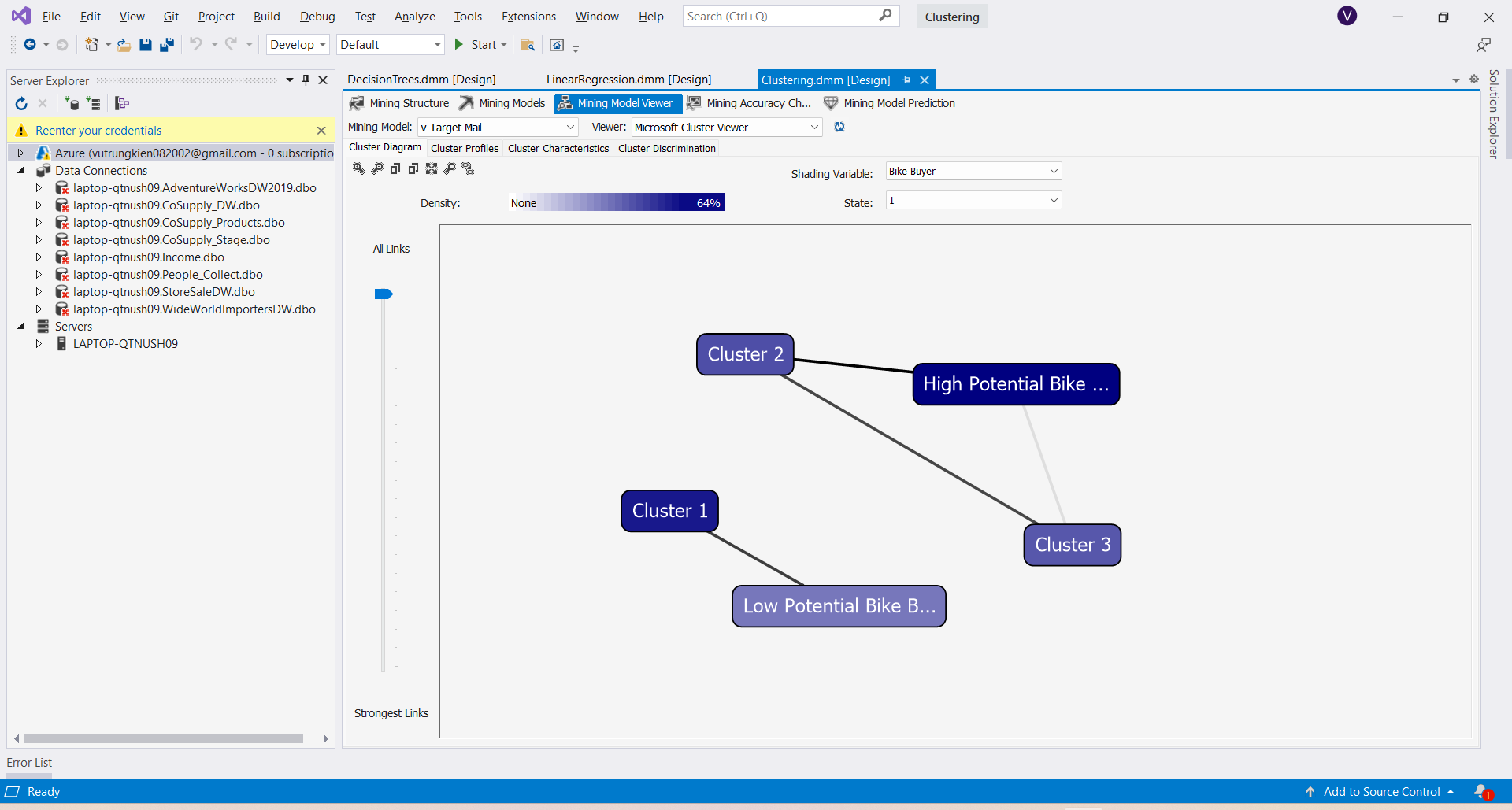




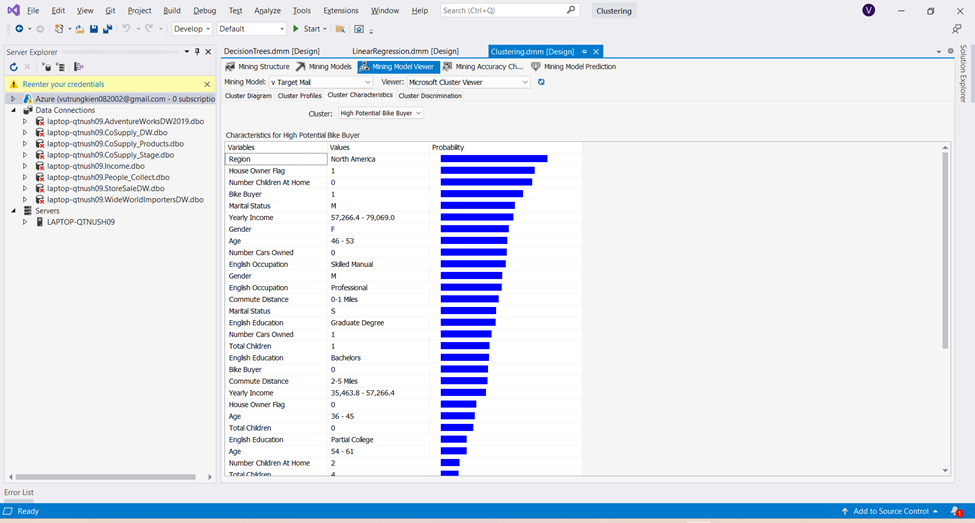
Nhận xét:

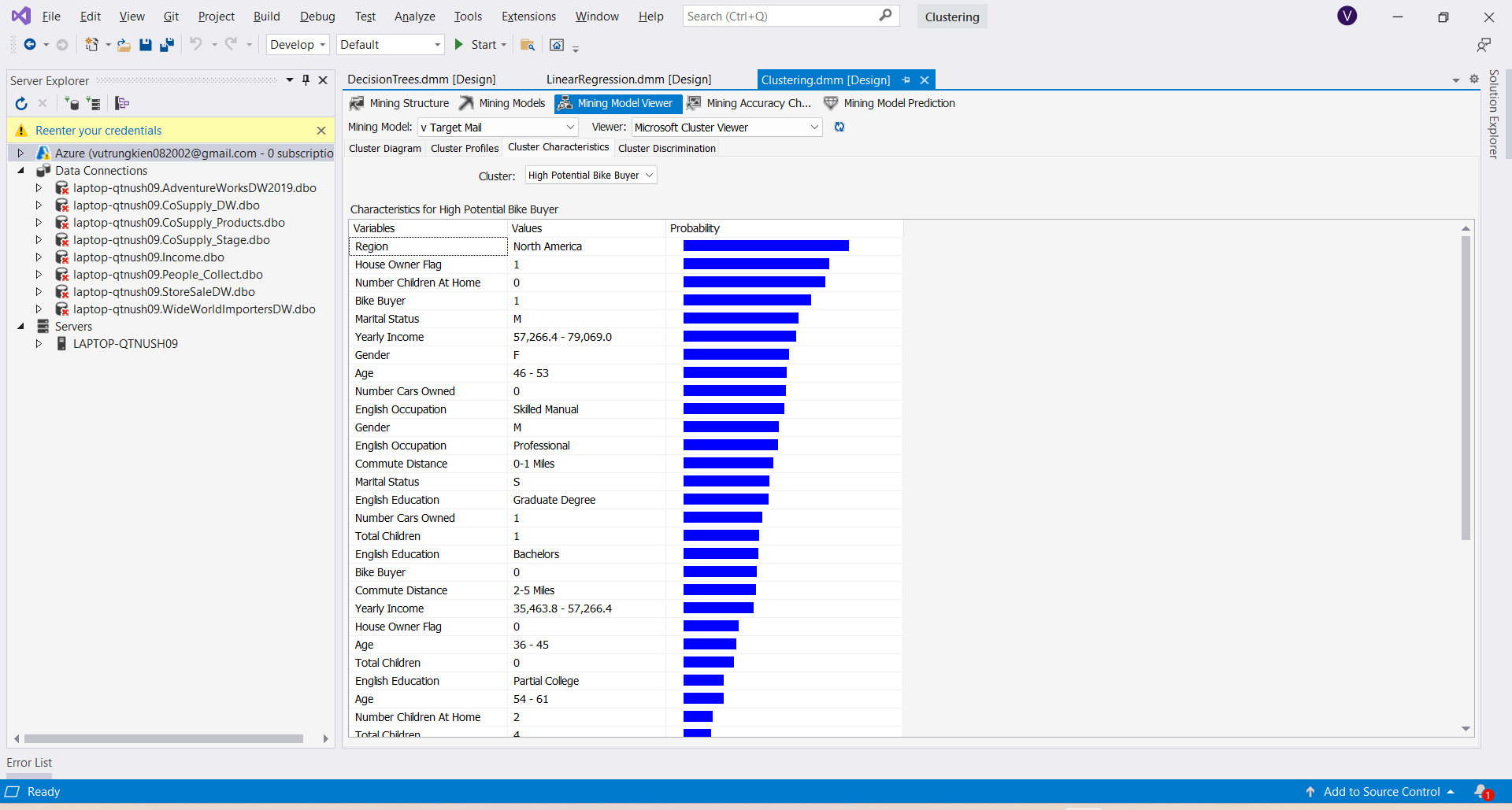
* Cluster 4 là nhóm có khả năng mua xe đạp cao nhất với 0.636. Từ những số liệu và biểu đồ trên, ta có một vài nhận xét về nhóm tiềm năng mua xe cao nhất như sau. Đa số là người ‘North America’, độ tuổi khoảng từ 49, có mức thu nhập trung bình, ‘Number Car Owned’ là 1 do đó họ có nhu cầu mua hơn, ‘English Occupation’ là ‘professional’ và ‘skilled manual’ với tỉ lệ lần lượt là 0.47 và 0.5, ‘Number Children At Home’ là 0 rất cao và ‘House Owner Flag’ là 1. Do đó có thể đặt tên Cluster 4 ‘High Potential Bike Buyer’.
* 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’.

Kết quả sau khi đổi tên các cụm



* + 1. **Cluster Characteristics:**

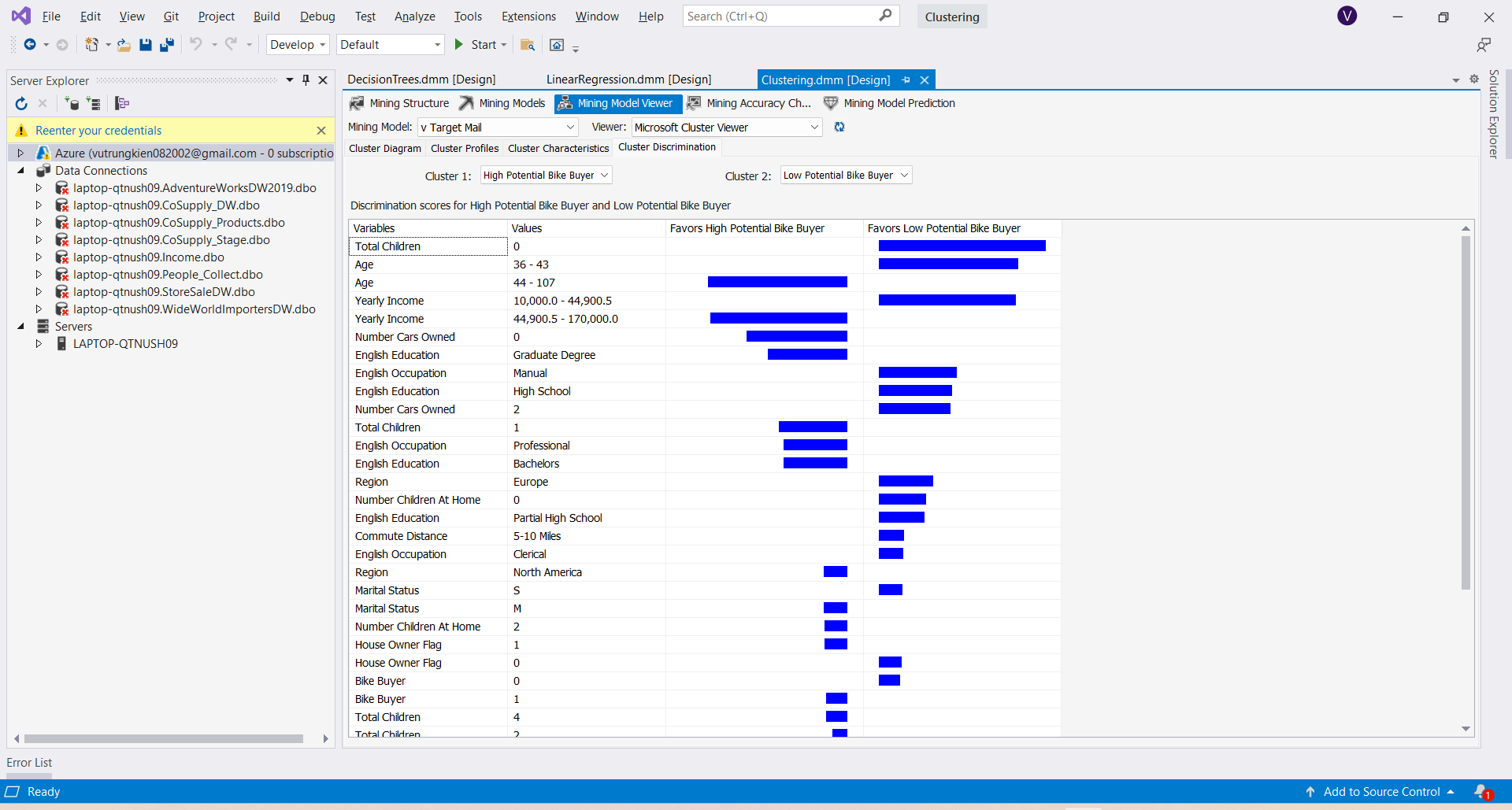
****



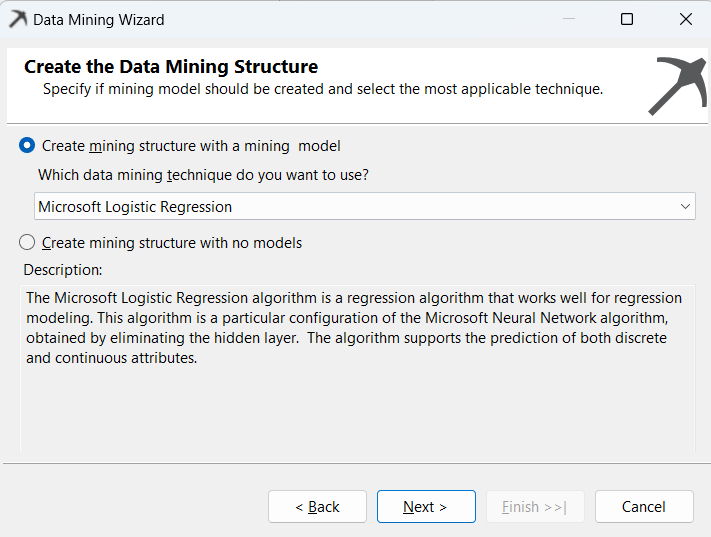
Tại đây có thể xem các đặc điểm của các cụm.

* + 1. **Cluster Discrimination:**

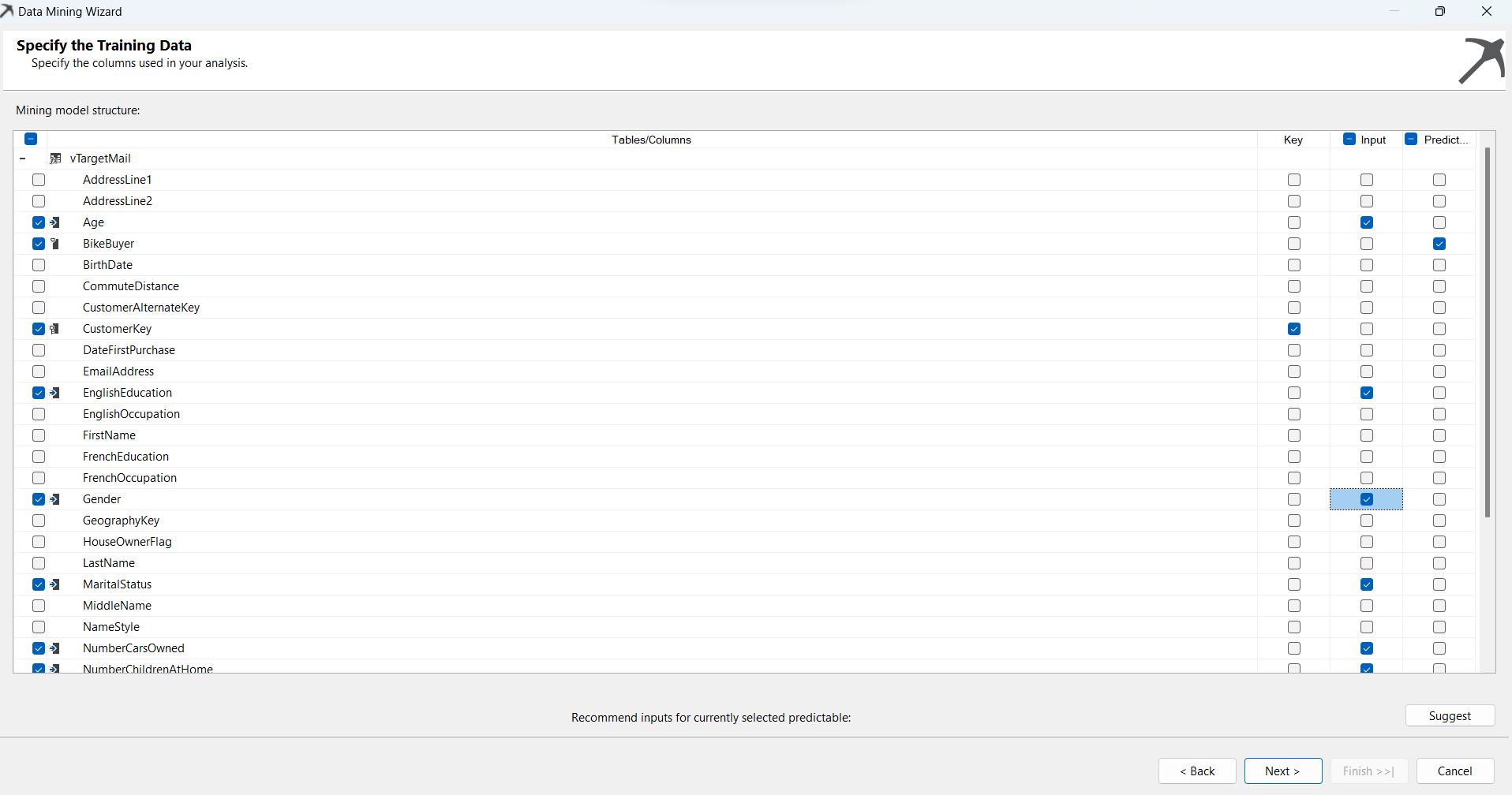
Tại mục này có thể so sánh sự khác biệt giữa hai cụm.



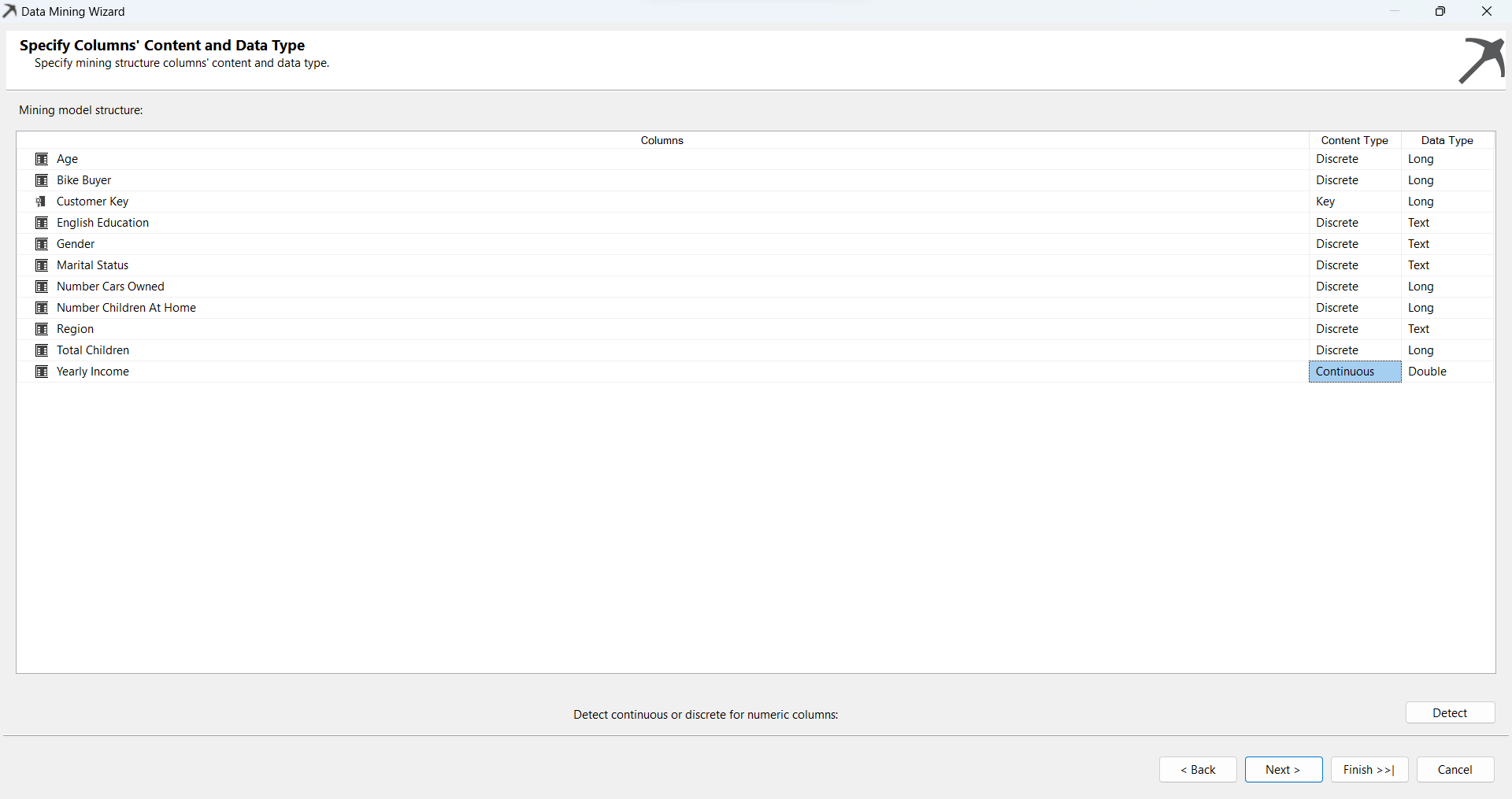
1. **So sánh độ chính xác của thuật toán Microsoft Logictic Regession và Microsoft Decision Tree trong dự đoán một người có mua xe đạp không?**
   1. **Tạo Logistic Regression Project với SSAS:**
      1. **Thiết lập Mining Structure:**

****

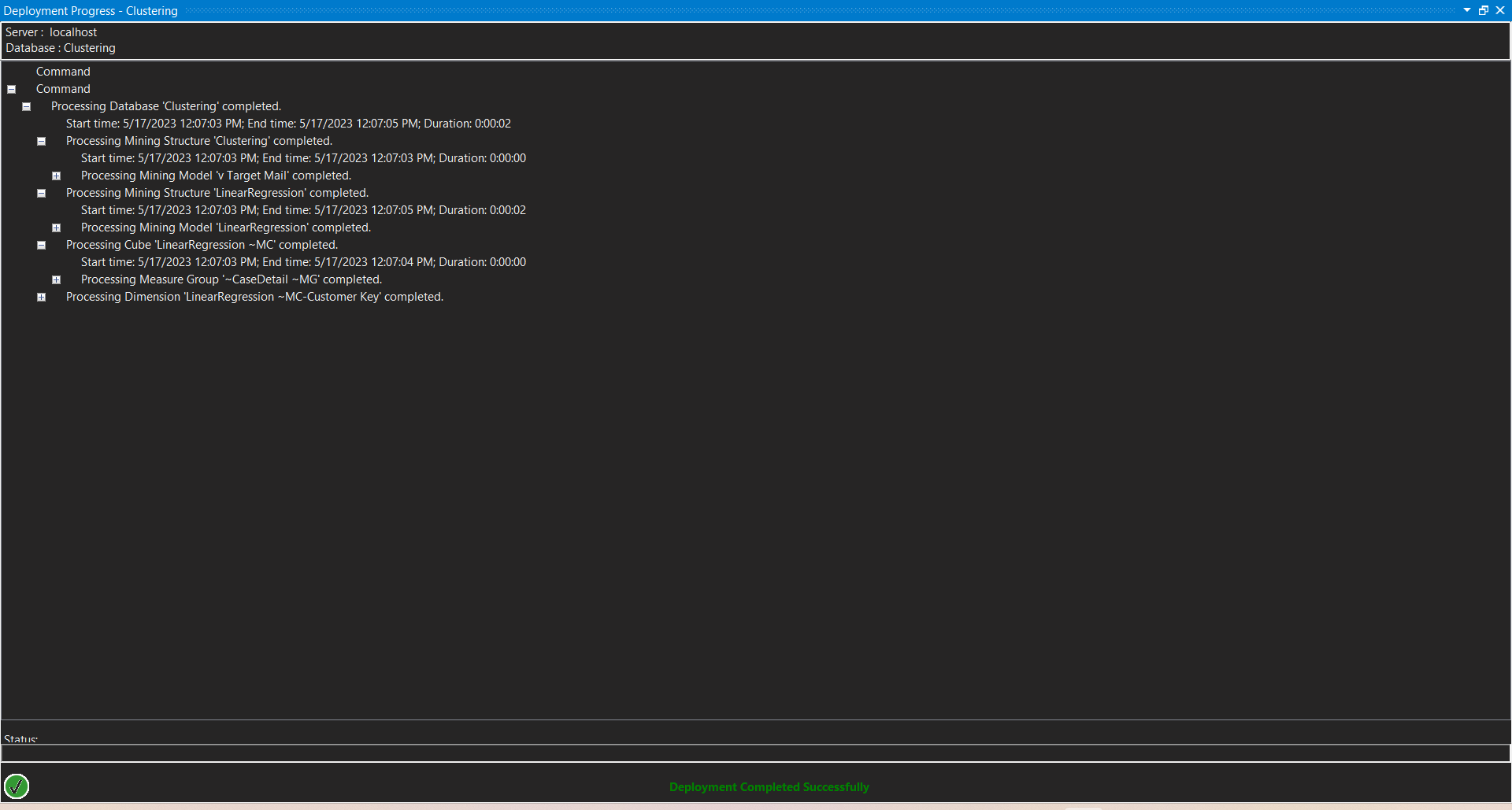
* + 1. **Thiết lập Traning Data:**

****

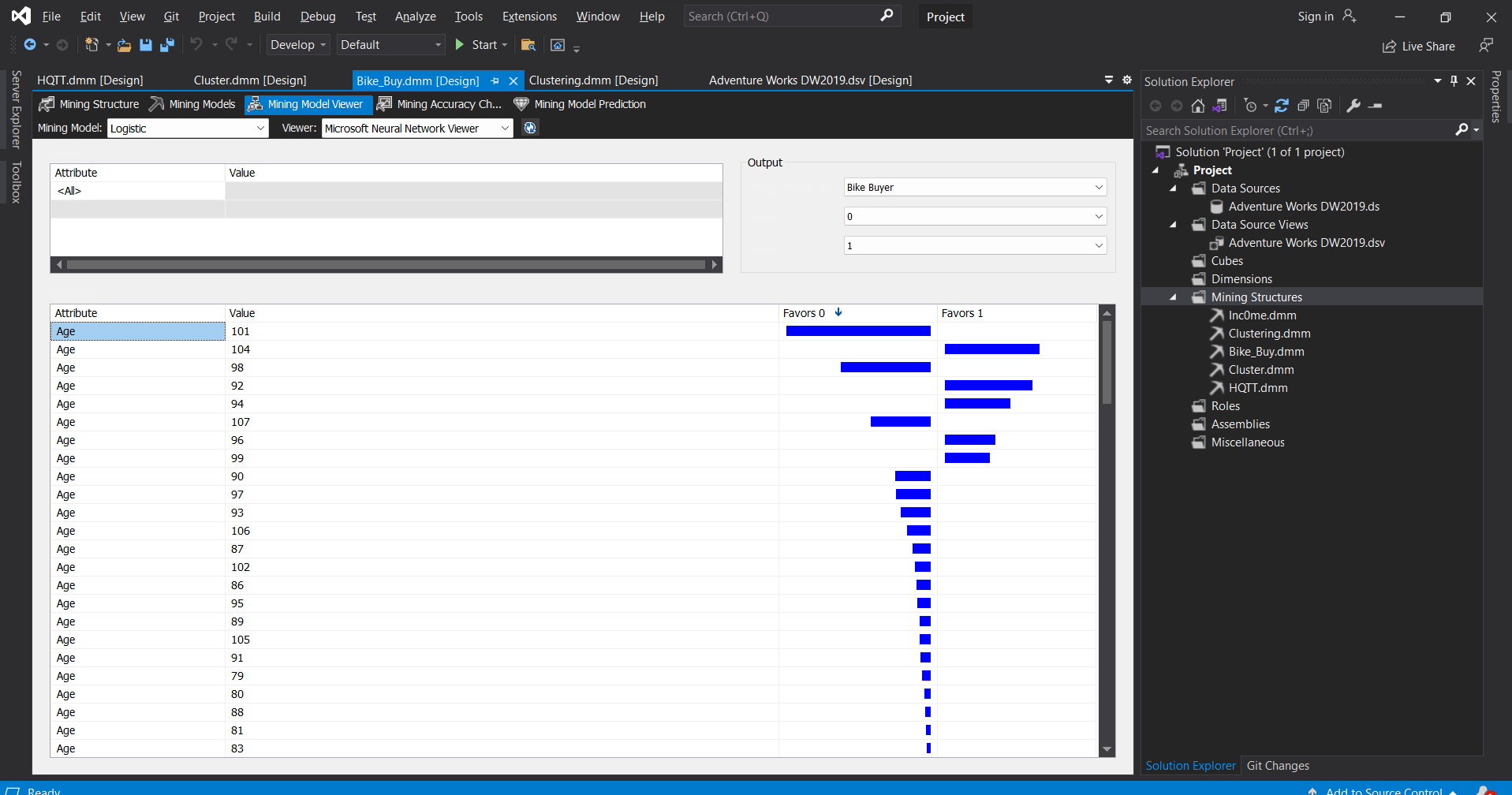
Ở đây dự đoán một người có mua xe đạp hay không



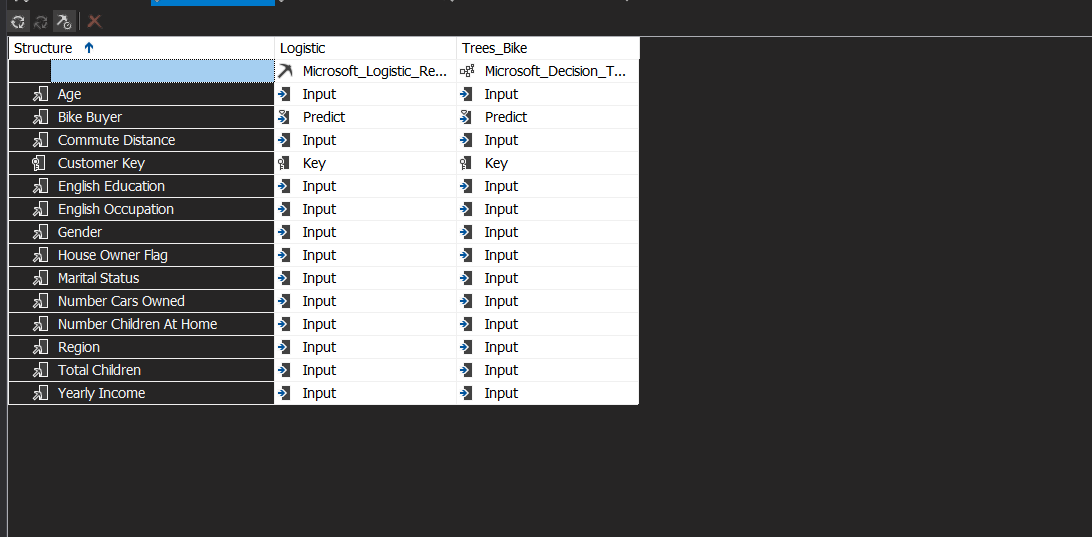
* + 1. **Processing Mining Structure:**



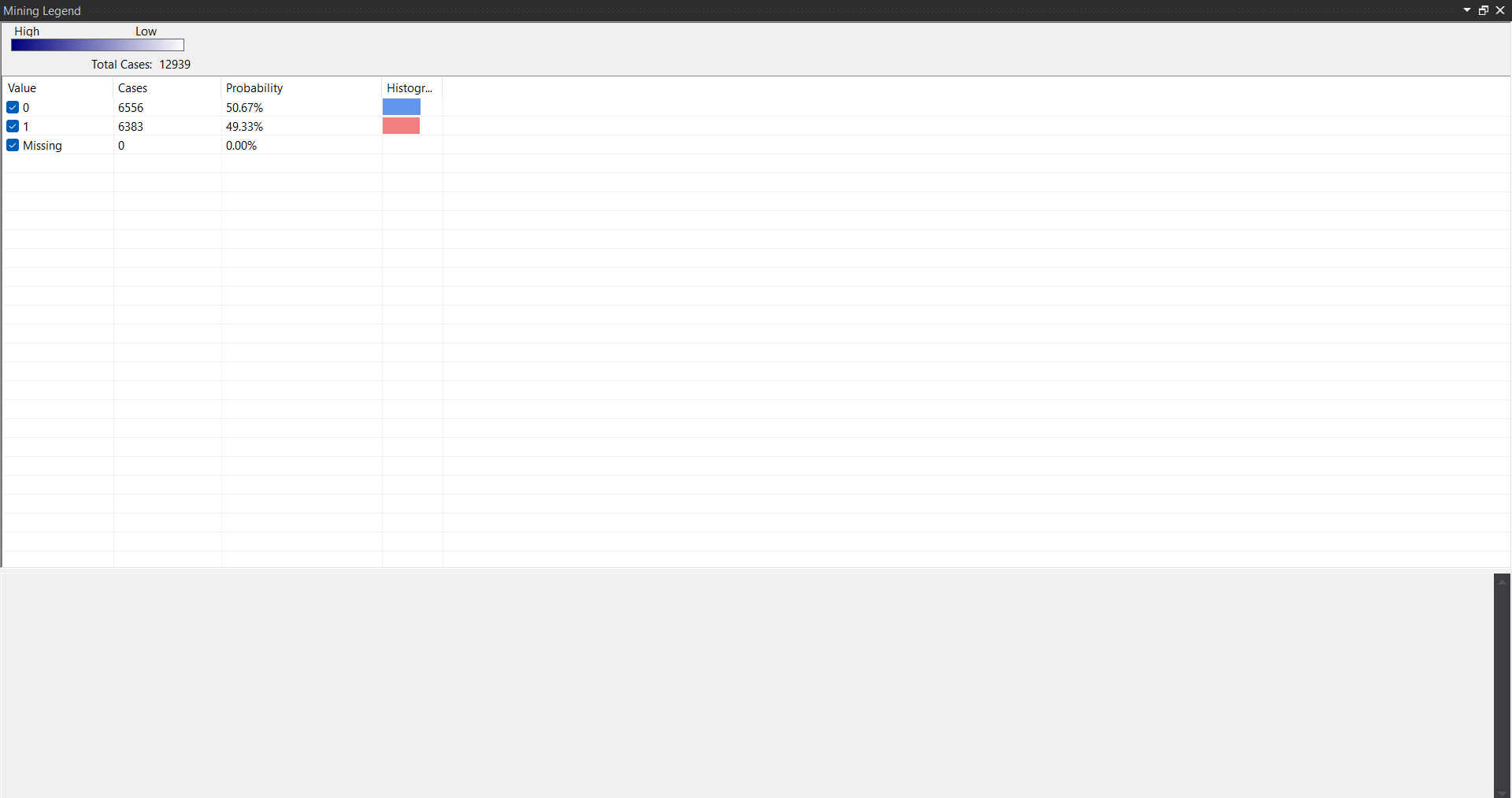
* 1. **Deploy Logistic Regression Mining:**

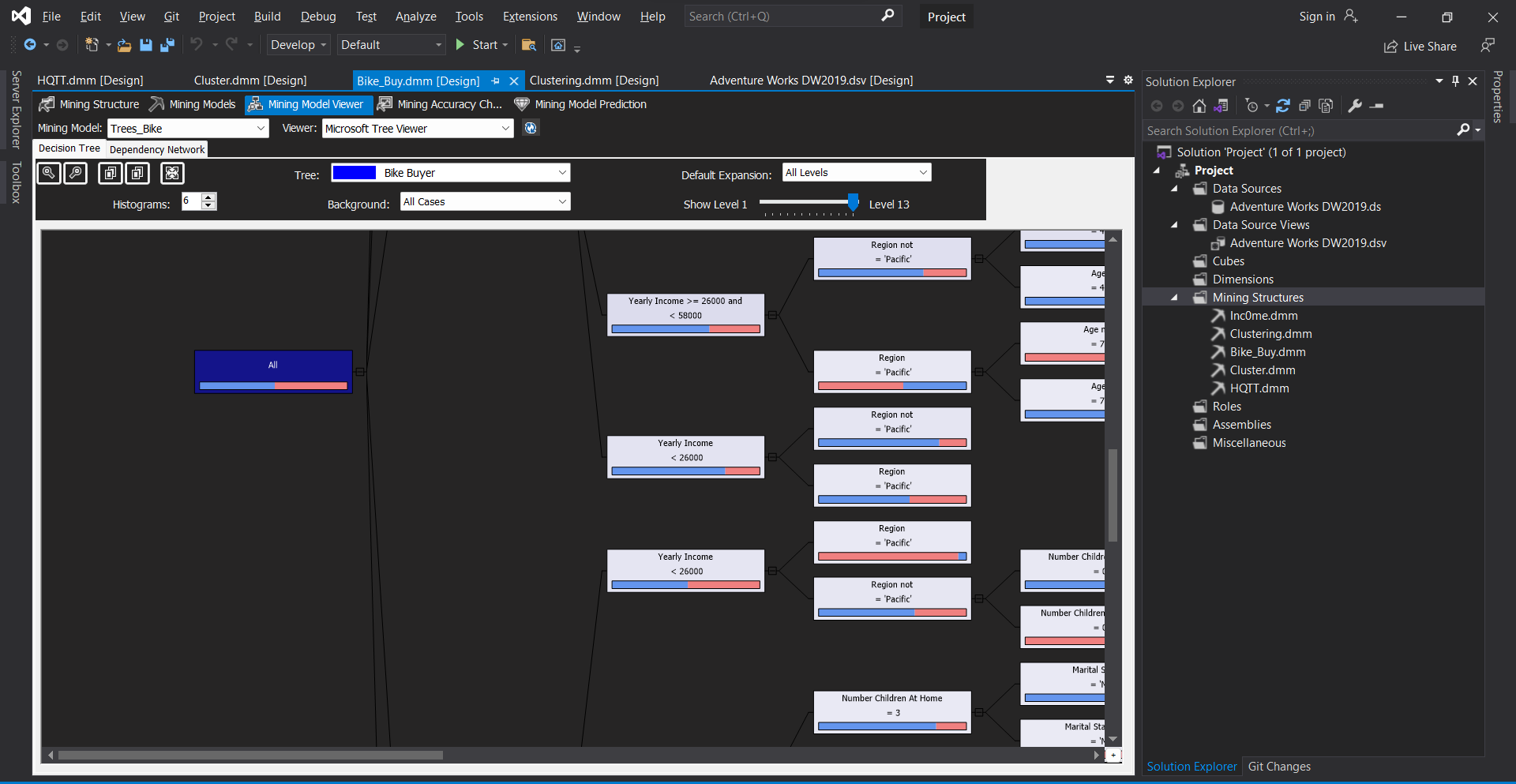
****

* 1. **Tạo Decision Trees Mining:**

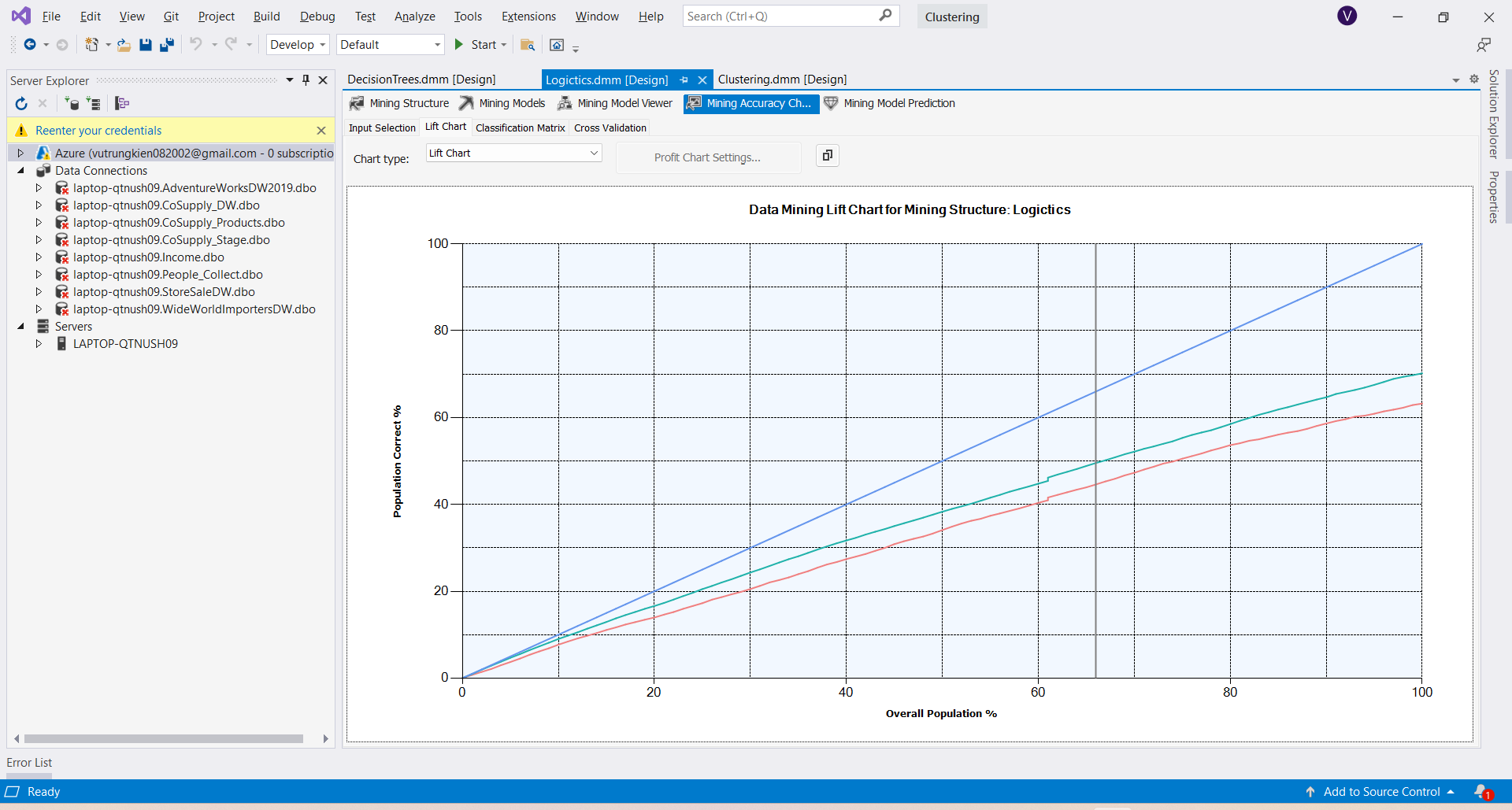


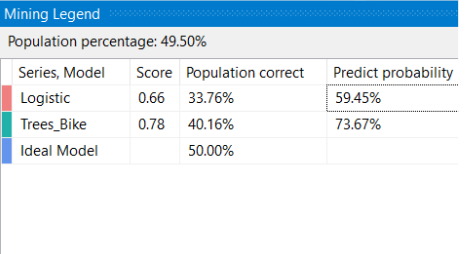
Mining model Viewer





* 1. **So sánh độ chính xác giữa 2 mô hình trong việc dự đoán mua xe đạp:**
     1. **Đánh giá qua biểu đồ lift chart:**

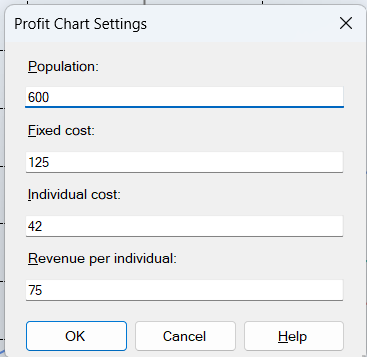




Trong 2 mô hình, mô hình tốt nhất là mô hình gần với mô hình lý tưởng nhất và có số điểm đấnh giá và phần trăm dự đoán cao hơn. Điều này có nghĩa là mô hình tốt nhất cho tập dữ liệu trên để người mua xe đạp là thuật toán cây quyết định.

* + 1. **Đánh giá qua profit chart:**

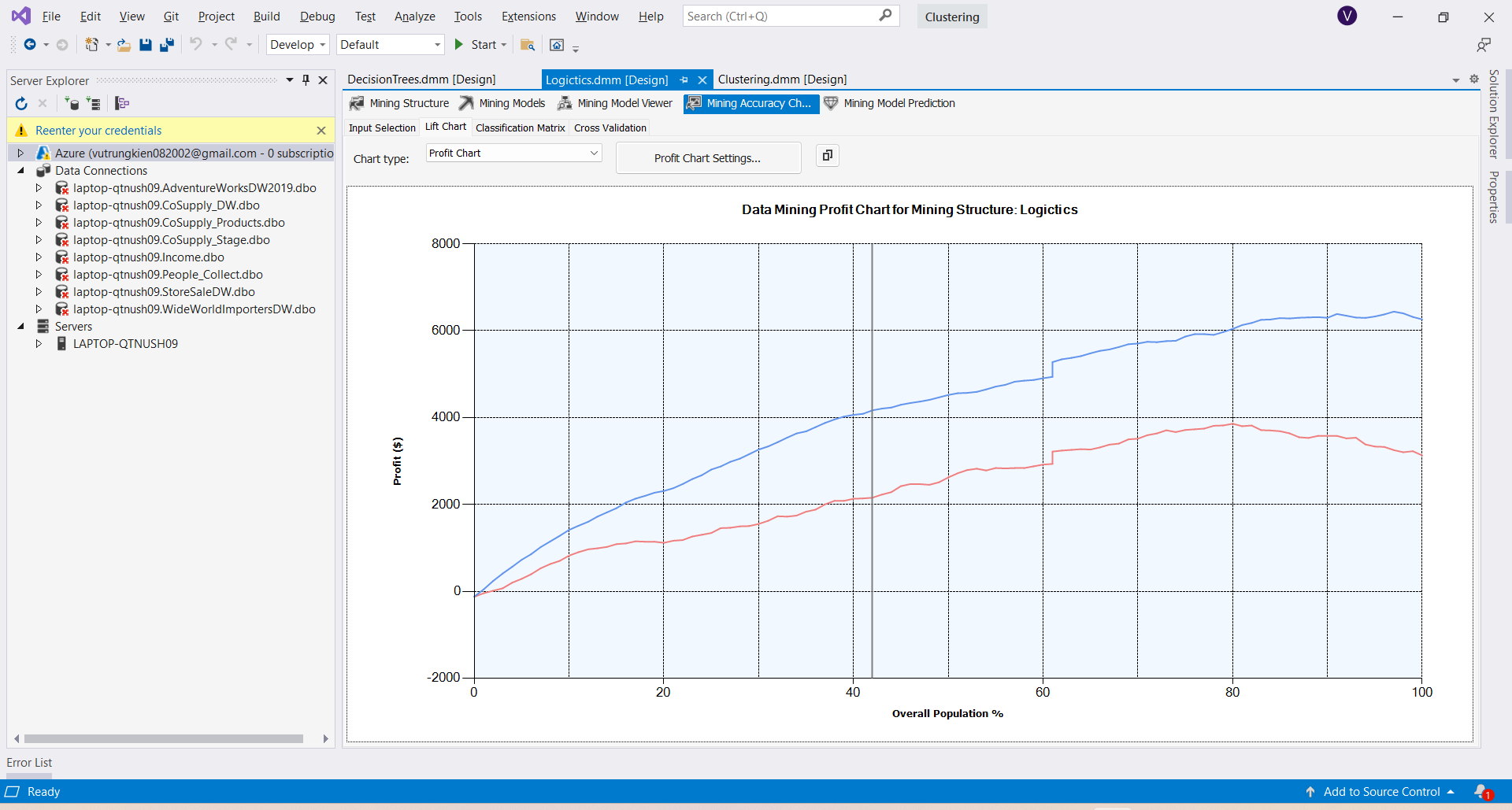
Biểu đồ Lợi nhuận có phần độc đáo trong các công cụ của Microsoft. Giả sử chúng ta đang xem xét một chương trình khuyến mãi để cải thiện người mua xe đạp. Đối với một chiến dịch tiếp thị, có bốn thông số population, fixed cost (Chi phí cố định), individual cost (chi phí cá nhận) và revenue per individual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

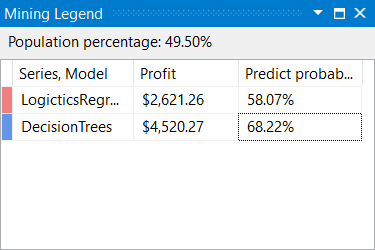


* Giải thích các tham số:

|  |  |  |
| --- | --- | --- |
| Setting | Value | Comments |
| **Population** | 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. |
| **Fixed cost** | 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. |
| **Individual cost** | 42 | Chi phí cho mỗi đơn vị cho chiến dịch gửi thư được nhắm mục tiêu. |
| **Revenue per individual** | 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:



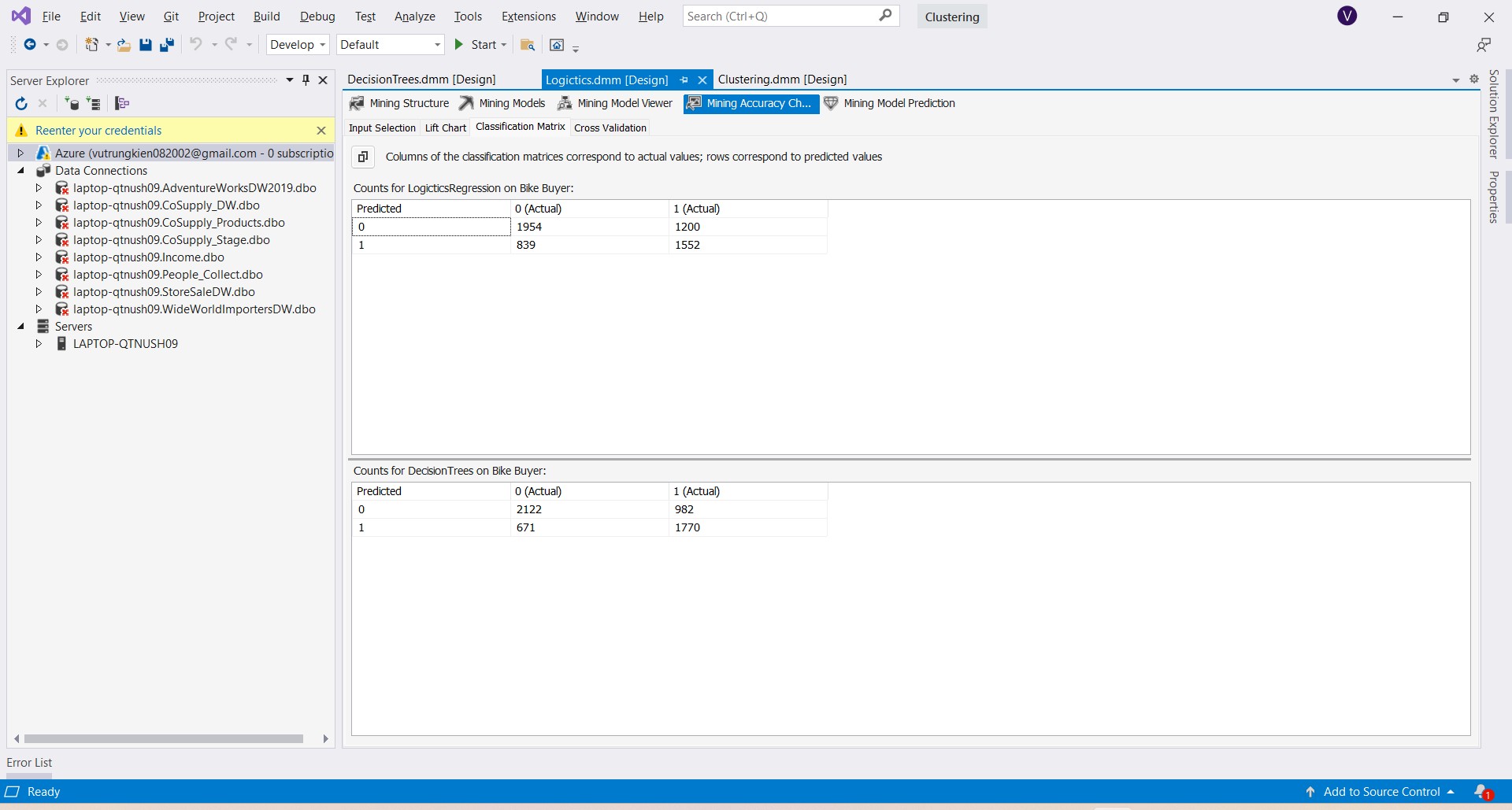


Trục Y của biểu đồ biểu thị lợi nhuận, trong khi trục X biểu thị tỷ lệ phần trăm khách hàng đã được liên hệ bởi chiến dịch gửi thư được nhắm mục tiêu.

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.

* + 1. **Ma Trận Phân Loại**

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

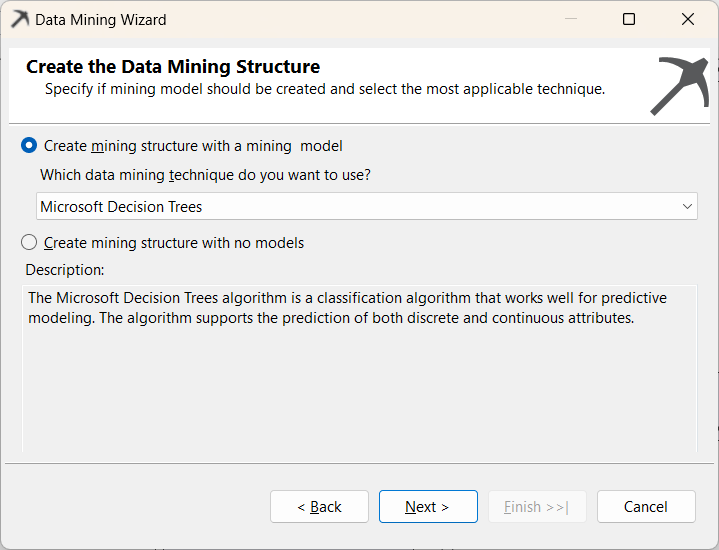


Ta thấy tỉ lệ dự đoán chính xác số trường hợp mua xe trong thuật toán cây quyết định là 71,7% cao hơn tỉ lệ dự đoán trong thuật toán hồi quy logistic 62.5%

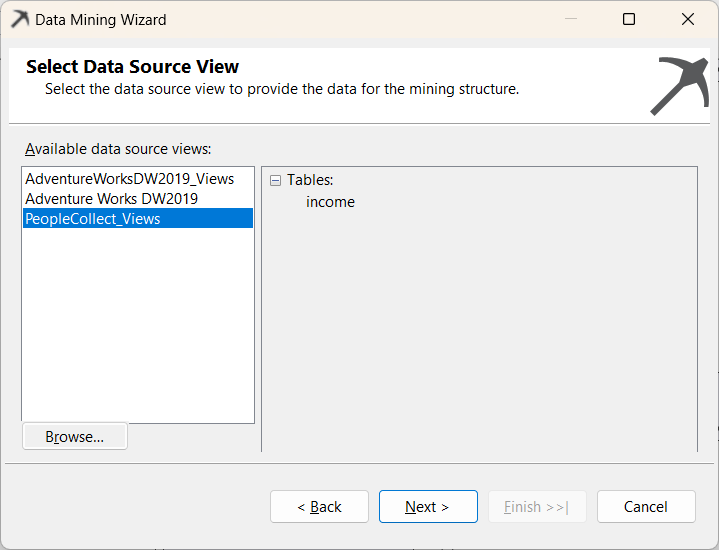
Tỉ lệ dự đoán chính xác số trường hợp không mua xe trong thuật toán cây quyết định là 70,3% cao hơn tỉ lệ dữ đoán trong thuật toán hồi quy logistic 62.1%

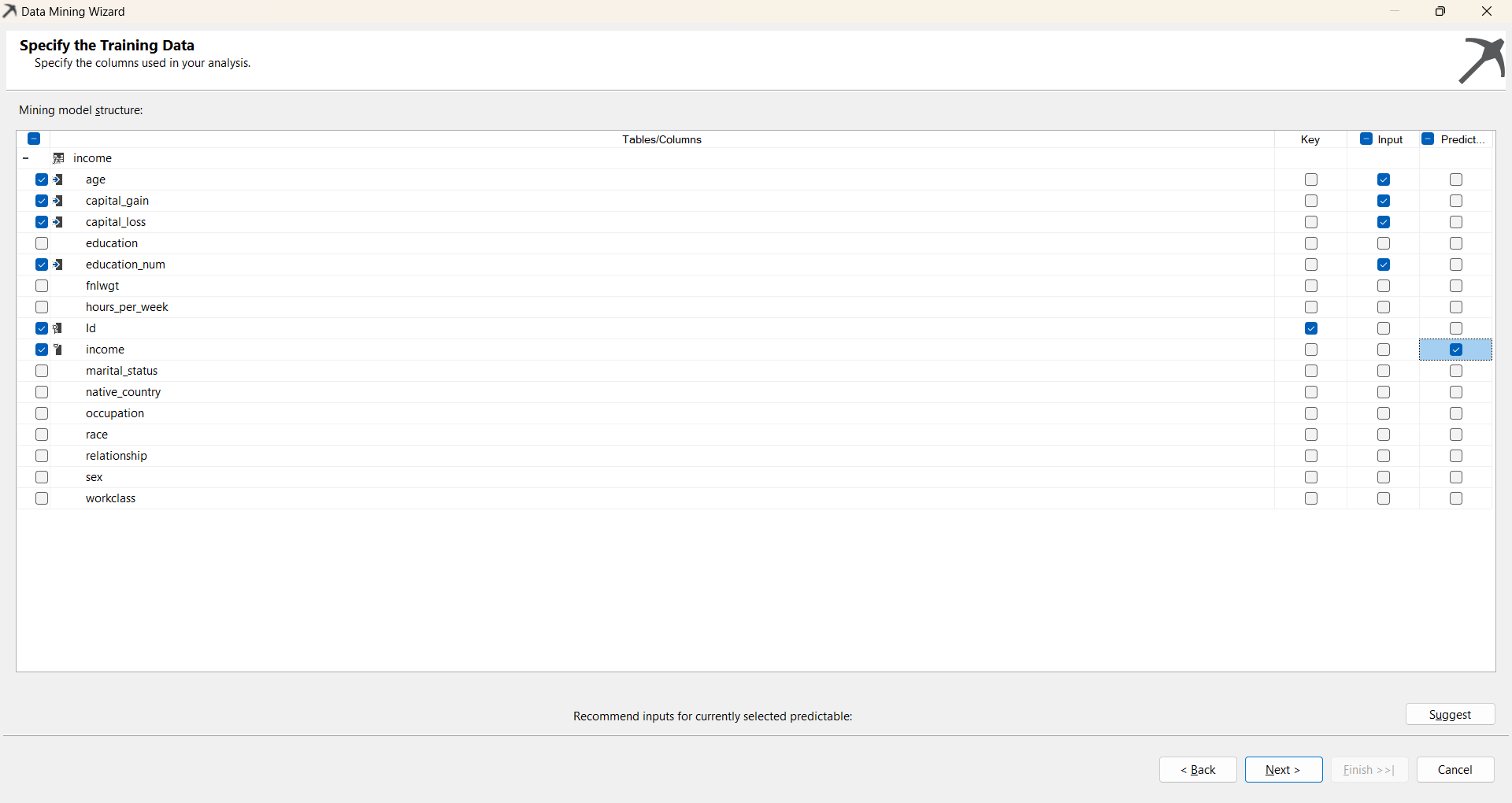
Vì vậy theo ma trận phân loại thì thuật toán cây quyết định vẫn tối ưu hơn cho dự đoán người mua xe đạp

1. **Thuật toán** **Microsoft** **Decision Tree:**
   1. **Tạo Decision Trees Project với SSAS:**
      1. **Thiết lập Mining Structure:**



* + 1. **Thiết lập Traning Data:**





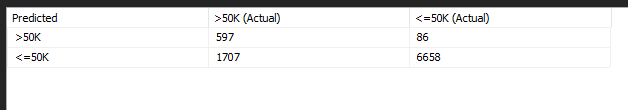
* 1. **Deploy Decision Trees Mining:**
     1. **Decision Trees Diagram**



Nhận xét:

* Khoảng Education num < 9 : thu nhập dưới 50K $.
* Khoảng Education num ∈ [9; 11):
* Nếu người đó có số tuổi Age<33 hoặc là Age >= 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 >= 2000,3 và Capital < 1742,4 thì có thể xác định họ thuộc nhóm có thu nhập dưới 50K $.
* Khoảng Education num ∈ [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>=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 ∈ [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 >= 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 >=15: có Age < 33, Capital Gain < 9999,9 và Capital Loss < 1742,4 thì khả năng cao họ có thu nhập dưới 50K $.

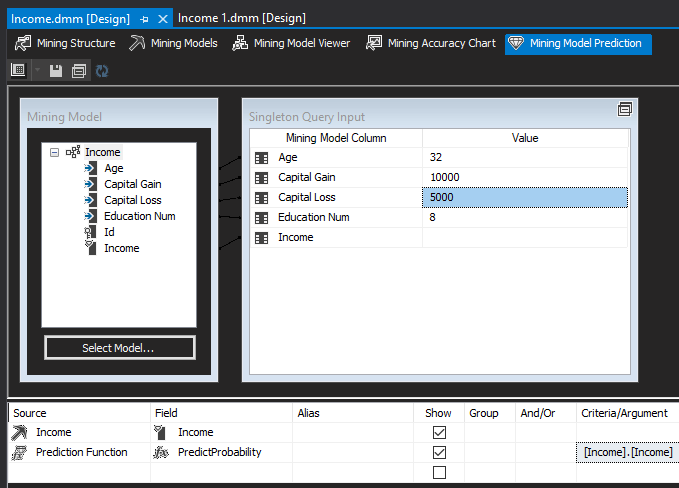
Thực hiện xem đánh giá mô hình dựa vào confusion matrix:



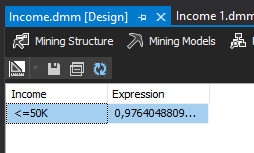
Độ chính xác của mô hình 80.34%

Thực hiện dự đoán cho income:

Thử với Education num<9

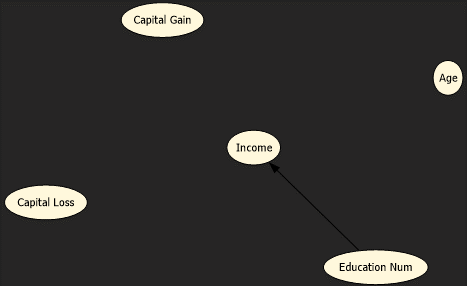


Kết quả cho ra 97.64% thì người này thuộc thu nhập dưới 50K $.



* + 1. **Dependency Network:**

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



1. **TỔNG KẾT:**
2. **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.

1. **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.

1. **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.

1. **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>