**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HỒ CHÍ MINH**Shape, square

Description automatically generated

**KHOA CÔNG NGHỆ THÔNG TIN**Ảnh có chứa văn bản, mẫu họa

Mô tả được tạo tự động

**BÁO CÁO ĐỒ ÁN CUỐI KỲ**

**ĐỀ TÀI: KHAI PHÁ, PHÂN TÍCH VÀ DỰ ĐOÁN BỆNH TIM TỪ CÁC CHỈ SỐ CÁ NHÂN.**

**Môn học: KHAI PHÁ DỮ LIỆU**

**Mã lớp học phần: DAMI330484\_22\_2\_01**

**GVHD: ThS. Nguyễn Văn Thành**

**Nhóm sinh viên thực hiện (07):**

1. **Bùi Tấn Đạt – 20133033**
2. **Lê Tuấn Nghĩa – 20133**
3. **Võ Hoàng Nguyên – 20133075**

TP Hồ Chí Minh, tháng 05 năm 2023

**DANH SÁCH THÀNH VIÊN THAM GIA**

**THỰC HIỆN ĐỀ TÀI VÀ VIẾT BÁO CÁO**

**Môn: Khai phá dữ liệu -** *HỌC KÌ II – NĂM HỌC 2022 – 2023*

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Điểm: ……………………………**

**Nhận xét của giảng viên:**

**………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………**

*Ngày … tháng 05 năm 2023*

*Giảng viên chấm điểm*

*Ths. Nguyễn Văn Thành*

# **LỜI CẢM ƠN**

Để hoàn thành tốt đề tài và bài báo cáo này, chúng em xin gửi lời cảm ơn chân thành đến giảng viên Nguyễn Văn Thành, người đã trực tiếp hỗ trợ chúng em trong suốt quá trình học. Chúng em cảm ơn thầy đã đưa ra những lời khuyên từ kinh nghiệm thực tiễn của mình để giúp đỡ và định hướng cho chúng em đi đúng với yêu cầu của đề tài đã chọn, luôn giải đáp thắc mắc và đưa ra những góp ý, chỉnh sửa kịp thời giúp chúng em khắc phục nhược điểm và hoàn thành đề tài đúng tiến độ đề ra.

Đề tài và bài báo cáo được chúng em thực hiện trong khoảng thời gian ngắn, còn gặp nhiều hạn chế trong kiến thức và kỹ thuật cũng như kinh nghiệm trong việc thực hiện một đề tài viết chương trình. Do đó, chương trình còn nhiều thiếu sót là điều không thể tránh khỏi. Vậy nên chúng em rất mong nhận được những ý kiến đóng góp quý báu của thầy để nhóm chúng em có thể hoàn thiện hơn về kiến thức, tư duy và rèn luyện kỹ năng để có thể làm tốt hơn nữa trong những lần sau.

Chúng em xin chân thành cảm ơn Thầy. Cuối lời, chúng em kính chúc Thầy luôn dồi dào sức khỏe và thành công hơn nữa trong sự nghiệp trồng người.

Nhóm chúng em xin chân thành cảm ơn!

**A. GIỚI THIỆU ĐỀ TÀI**

1. **Lý do chọn đề tài**

Từ tính ứng dụng và tính thực tiễn trong đời sống hiện này thì bệnh tim là một trong những vấn đề sức khỏe quan trọng và có tầm ảnh hưởng rộng lớn đến sức khỏe và chất lượng cuộc sống của con người. Sử dụng khai phá dữ liệu để nghiên cứu và dự đoán bệnh tim từ các chỉ số cá nhân có thể cung cấp thông tin hữu ích về khả năng mắc bệnh và tiềm năng ứng dụng trong lĩnh vực y tế. Tập dữ liệu "heart\_disease\_cleaned\_2020" là một lựa chọn lý tưởng trong môn học Khai phá dữ liệu

Và với đề tài này nó cũng có thể cung cấp cơ hội để chúng ta khám phá tri thức mới trong lĩnh vực bệnh tim và sức khỏe tim mạch. Bằng cách áp dụng các phương pháp khai phá dữ liệu, chúng ta có thể khám phá các mô hình, mối quan hệ và thông tin tiềm ẩn trong tập dữ liệu. Từ đó việc phân tích dữ liệu này có thể dẫn đến việc tìm ra các yếu tố quan trọng, các nhóm rủi ro và các khía cạnh mới liên quan đến bệnh tim, giúp mở rộng kiến thức và hiểu biết của chúng ta về lĩnh vực này.

Chính vì thế mà nhóm em đã chọn đề tài "**Khai phá, phân tích và dự đoán bệnh tim từ các chỉ số cá nhân**" để giúp chúng ta nâng cao kiến thức và kỹ năng về kho dữ liệu và đồng thời đóng góp vào việc phân tích và giải quyết vấn đề thực tế đang diễn ra trên toàn cầu.

1. **Tổng quan về DataSet**
2. **Nguồn dữ liệu được sử dụng**

Tập dữ liệu “heart\_2020\_cleaned.csv” được lấy từ Kaggle và Kaggle được biết đến là một trang web cung cấp các bộ dữ liệu, các tập hợp về dữ liệu trong nhiều lĩnh vực và bài viết về nhiều chủ đề khác nhau như dân số, sức khỏe, môi trường, kinh tế,… Đường dẫn trực tiếp đến trang tải xuống tập dữ liệu về bệnh tim. Link: <https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease>

1. **Mô tả chi tiết dữ liệu**

Ban đầu, bộ dữ liệu đến từ CDC và là một phần chính của Hệ thống giám sát yếu tố rủi ro hành vi (BRFSS), tiến hành các cuộc khảo sát qua điện thoại hàng năm để thu thập dữ liệu về tình trạng sức khỏe của cư dân Hoa Kỳ. Được thành lập vào năm 1984 với 15 tiểu bang, BRFSS hiện đang thu thập dữ liệu ở tất cả 50 tiểu bang cũng như Quận Columbia và ba vùng lãnh thổ của Hoa Kỳ. BRFSS hoàn thành hơn 400.000 cuộc phỏng vấn người lớn mỗi năm, khiến nó trở thành hệ thống khảo sát sức khỏe được tiến hành liên tục lớn nhất trên thế giới.

Bộ dữ liệu này được sử dụng để dự đoán liệu một người có khả năng mắc bệnh tim hay không dựa trên các thông số đầu vào như giới tính, tuổi tác, hút thuốc, bia rượu. Mỗi hàng trong dữ liệu cung cấp thông tin liên quan về bệnh nhân.

Bộ dữ liệu gồm: 18 cột và 255.228 dòng.

|  |  |
| --- | --- |
| **Tên biến** | **Giải thích** |
| HeartDisease | Bạn có từng báo cáo bị bệnh tim mạch vành (CHD) hoặc nhồi máu cơ tim (MI) không? |
| BMI | Chỉ số khối cơ thể (BMI) |
| Smoking | Bạn đã hút ít nhất 100 điếu thuốc trong suốt cuộc đời mình chưa? (Lưu ý: 5 gói = 100 điếu) |
| AlcoholDrinking | Người nghiện rượu nặng (nam giới trưởng thành uống hơn 14 ly mỗi tuần và phụ nữ trưởng thành uống hơn 7 ly mỗi tuần) |
| Stroke | Bạn đã bao giờ bị đột quỵ chưa? |
| PhysicalHealth | Trong 30 ngày qua, sức khỏe thể chất của bạn không tốt bao nhiêu ngày (bao gồm cả bệnh tật và thương tích)? |
| MentalHealth | Trong 30 ngày qua, sức khỏe tinh thần của bạn không tốt bao nhiêu ngày? |
| DiffWalking | Bạn có gặp khó khăn nghiêm trọng khi đi bộ hoặc leo cầu thang không? |
| Sex | giới tính |
| AgeCategory | danh mục tuổi |
| Race | chủng tộc |
| Diabetic | Bạn đã bao giờ bị bệnh tiểu đường chưa? |
| PhysicalActivity | Khác với công việc bình thường của bạn, trong 30 ngày qua, bạn có hoạt động thể chất hoặc tập thể dục không? |
| GenHealth | Nhìn chung sức khỏe của bạn như thế nào? |
| SleepTime | Trung bình bạn ngủ bao nhiêu giờ một ngày? |
| Asthma | Bạn có bị hen suyễn không? |
| KidneyDisease | Không bao gồm sỏi thận, nhiễm trùng bàng quang hoặc tiểu không kiểm soát, bạn đã bao giờ nói rằng bạn bị bệnh thận chưa? |
| SkinCancer | Bạn đã từng bị ung thư da chưa? |

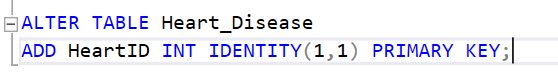
1. **Công cụ hỗ trợ**

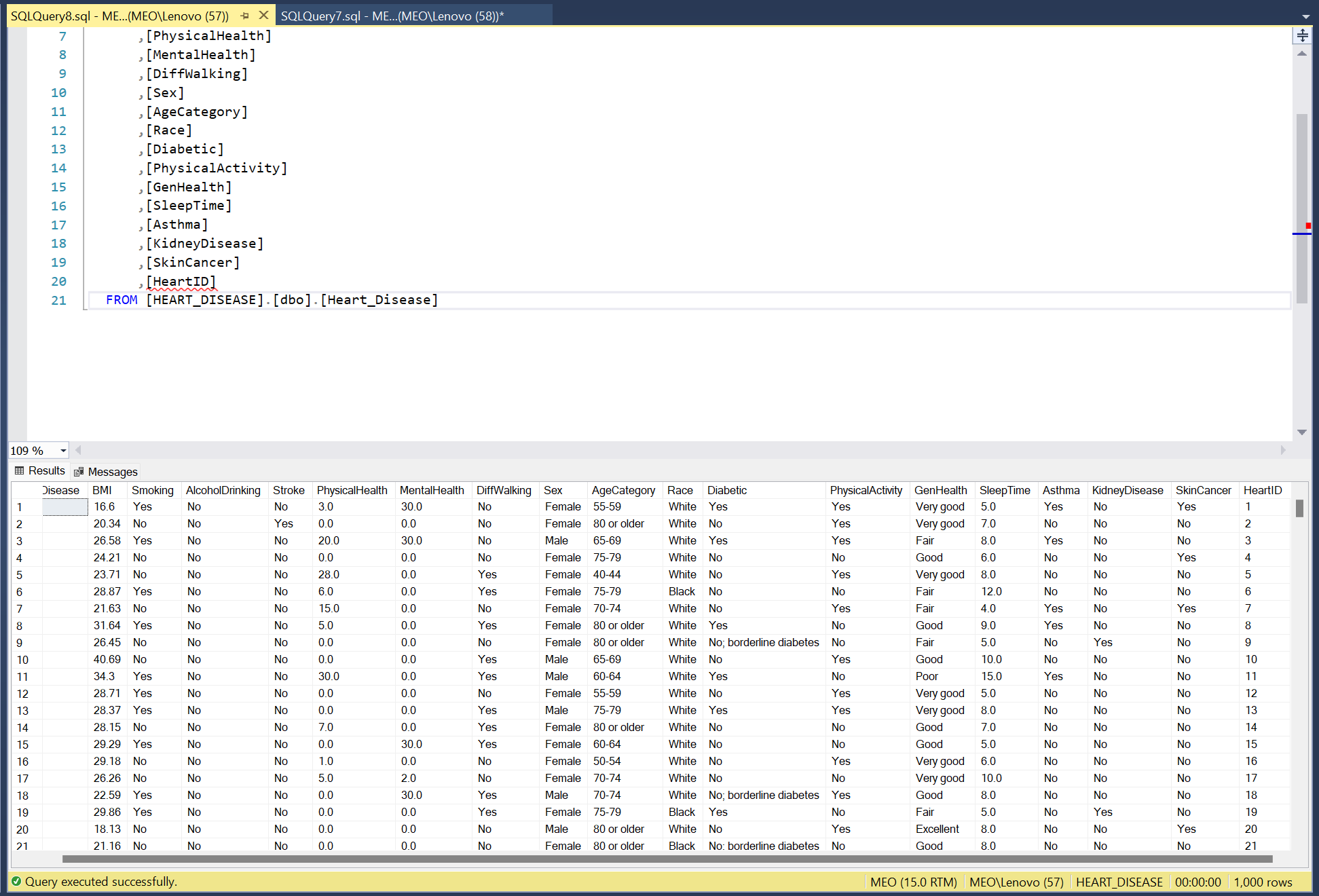
* Visual Studio 2019 hỗ trợ SSDT (SQL Server Data Tools)
* SQL Server Management Studio, SQL Server 2019
* Visual Code và ngôn ngữ Python.
* Microsoft Power BI Desktop
* Ứng dụng weka

**B. NỘI DUNG**

1. **Tiền xử lý dữ liệu**

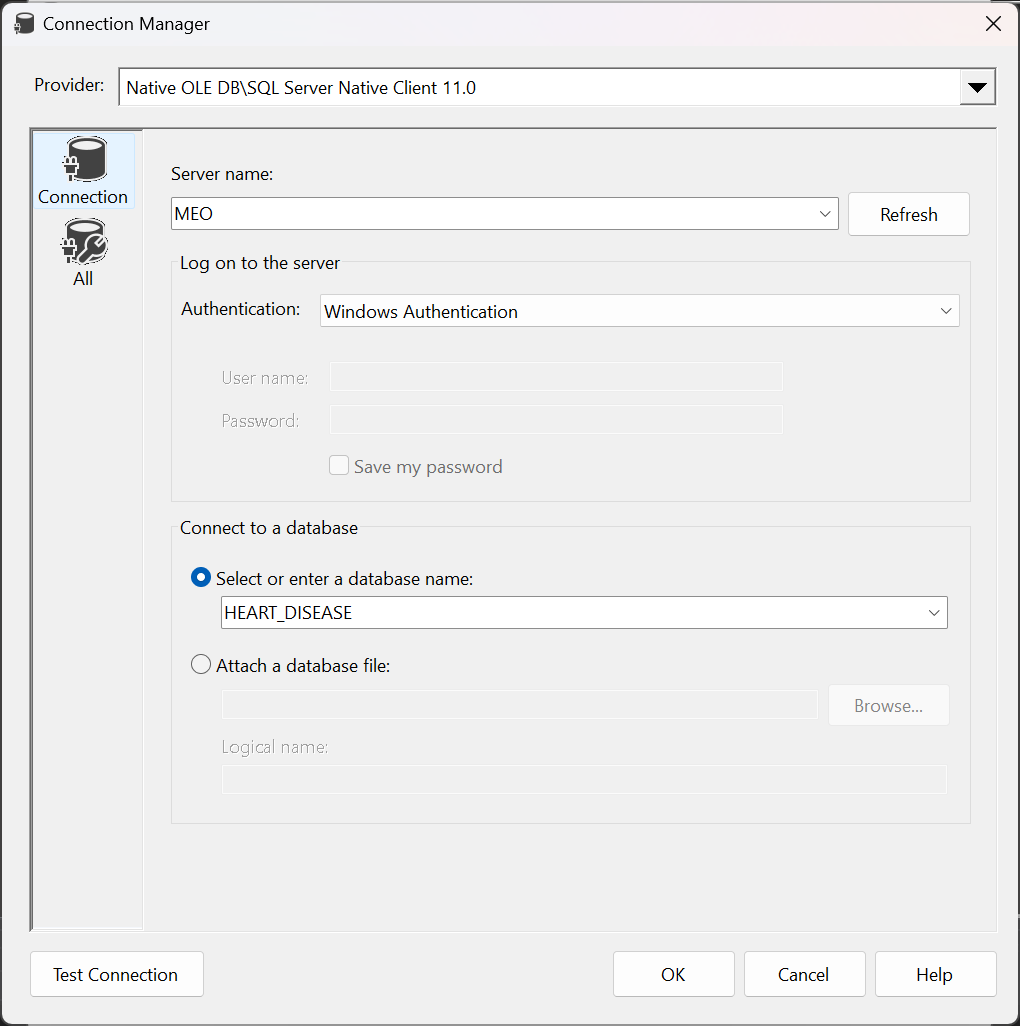
* Thêm cột **HeartID** làm khóa chính vào bảng **Heart\_Disease**

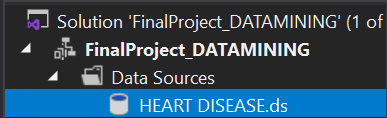




1. **Tạo Project Analysis Services Multidimensional and Data Mining Project**
   * + 1. **Tạo Data Source**

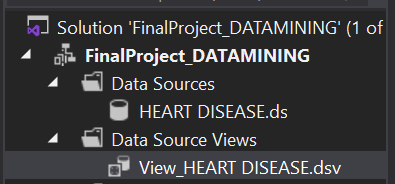
* Khởi tạo một Data Source mới có tên HEART\_DISEASE lấy dữ liệu từ database HEART\_DISEASE

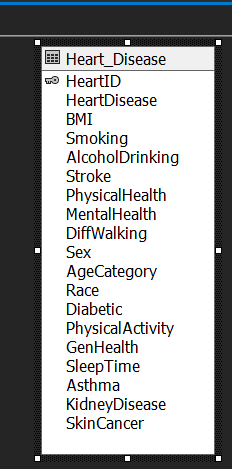
****

****

* + - 1. **Tạo Data Source View**

Tạo một Data Source View dựa trên Data Source vừa khởi tạo, chọn bảng Heart\_Disease để thực hiện các quá trình dự đoán của bài toán.

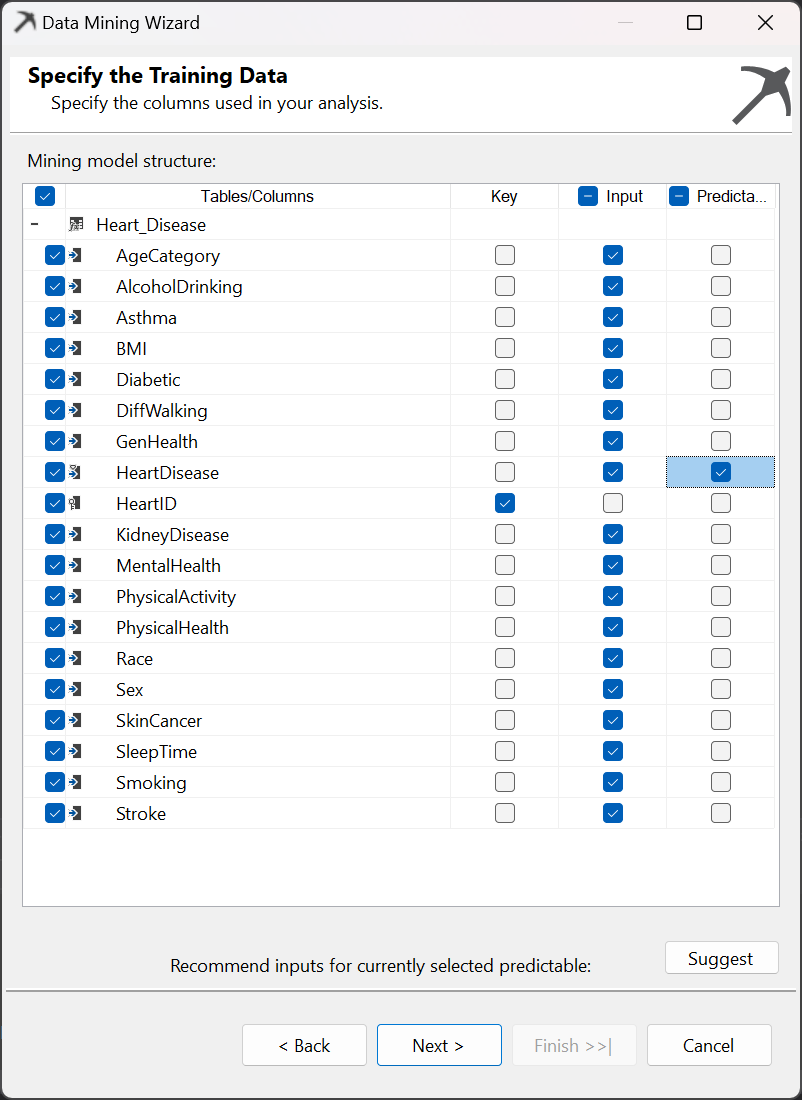




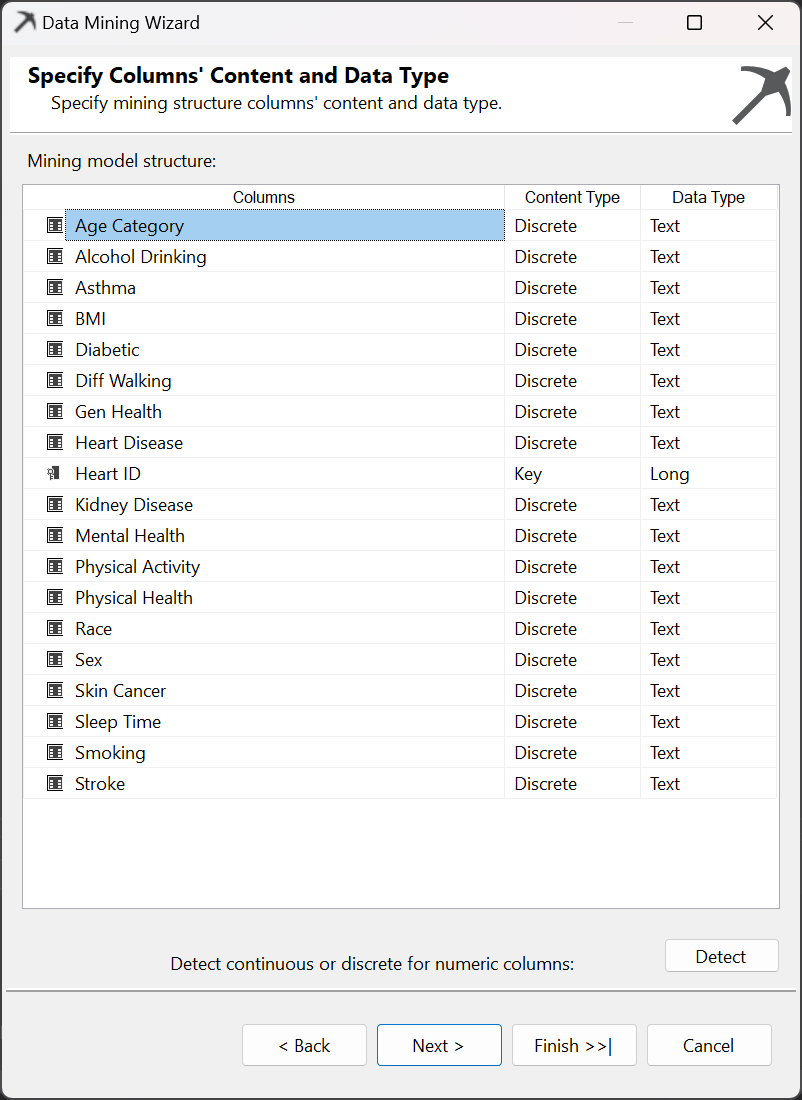
* + - 1. **Tạo Mining Structures**

1. **Association Rules**

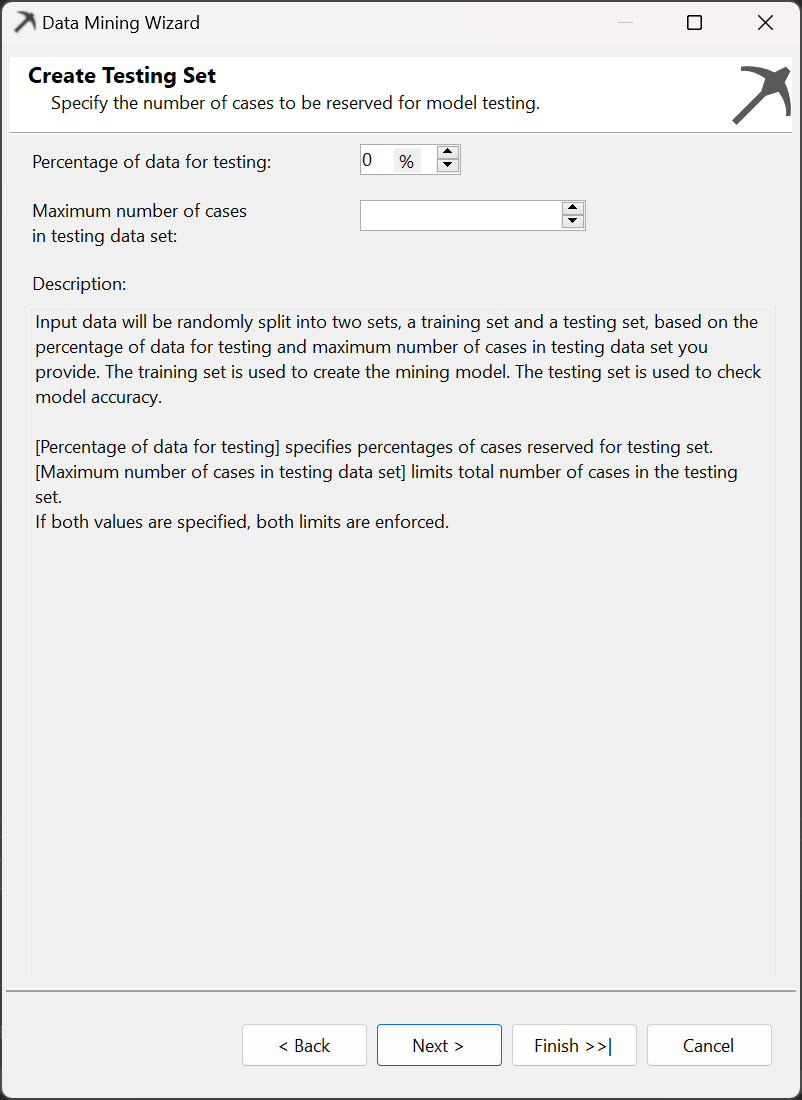
* Mục đích của Association Rule (luật kết hợp) là tìm ra các mối quan hệ giữa các đối tượng trong tập dữ liệu.
* Tạo một Mining Structure có tên **AR Heart Disease** với bảng Heart\_Disease là Case và chọn các thuộc tính như hình dưới làm Key, Input và Predictable.

****

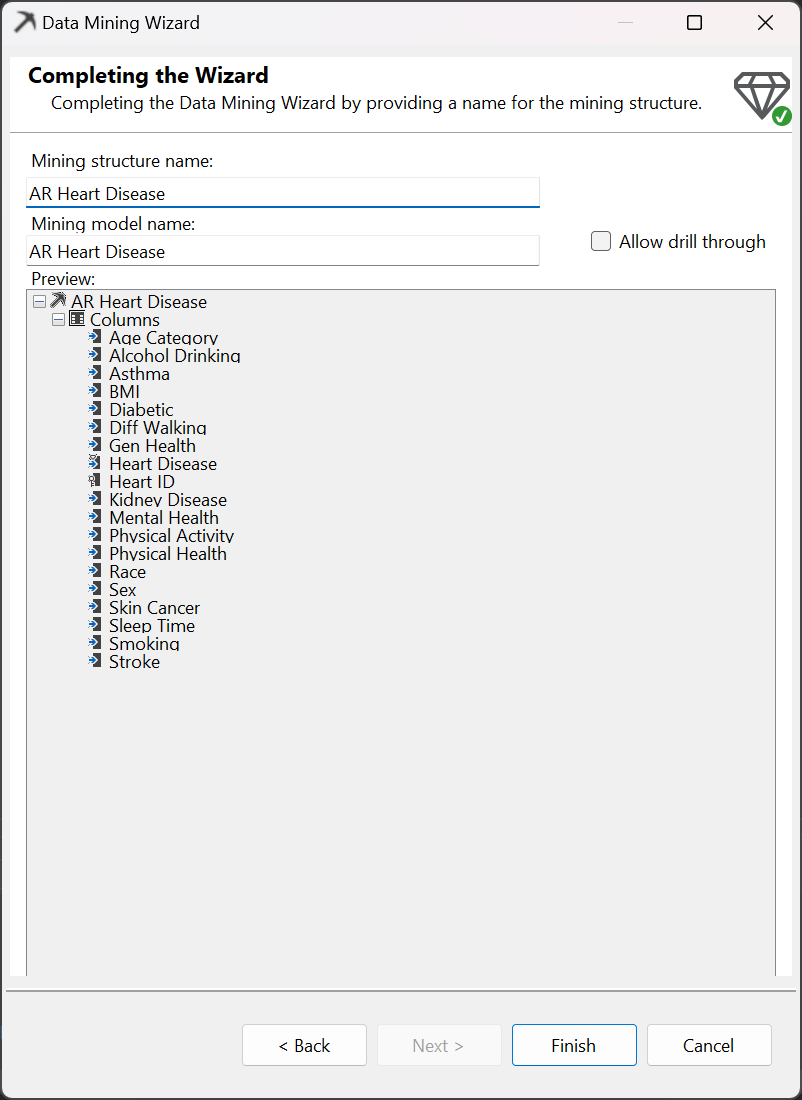
* Nhấn Next

****

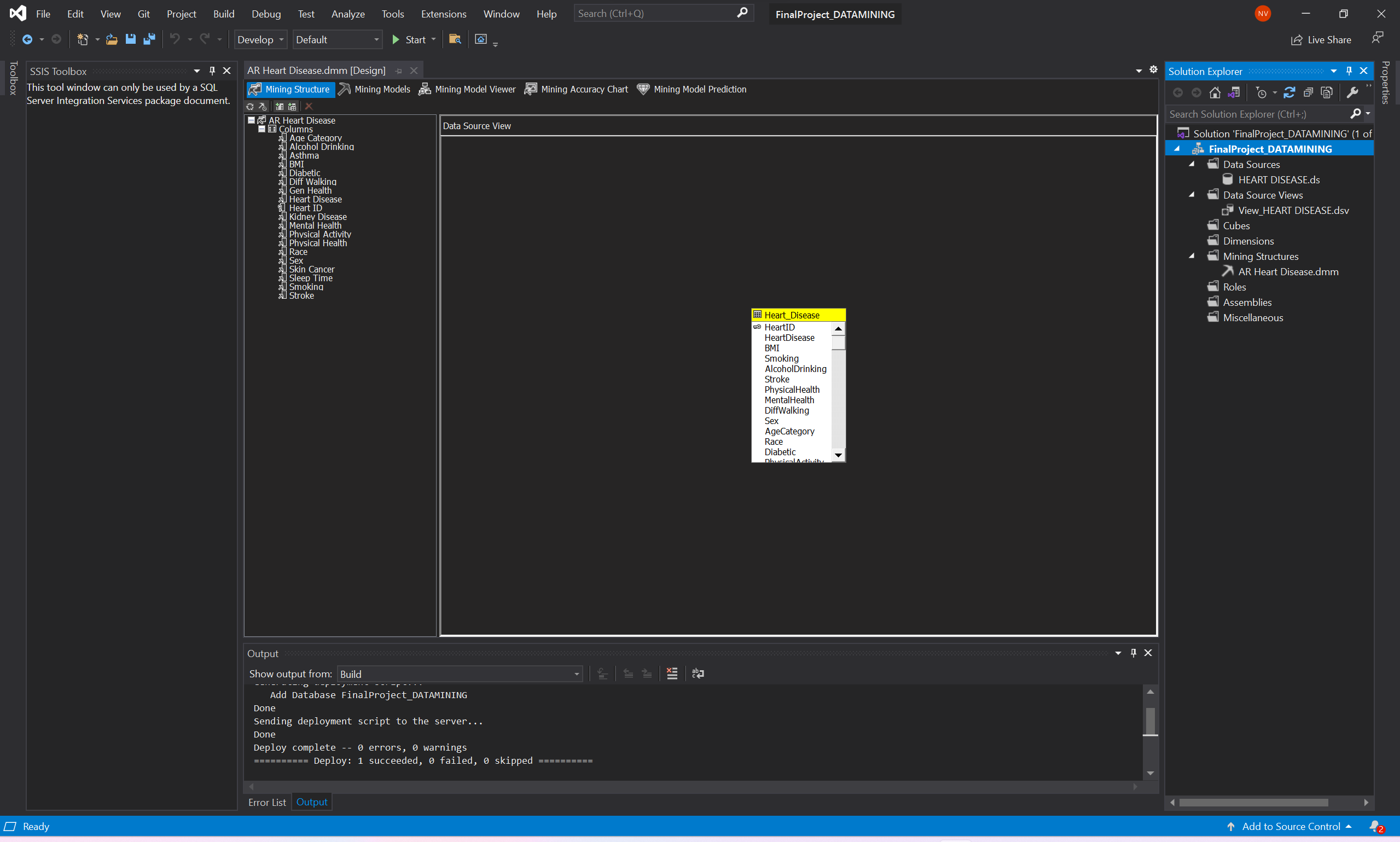
* Ở dialog Create Testing Set, ta lấy cả 100% tập dữ liệu test



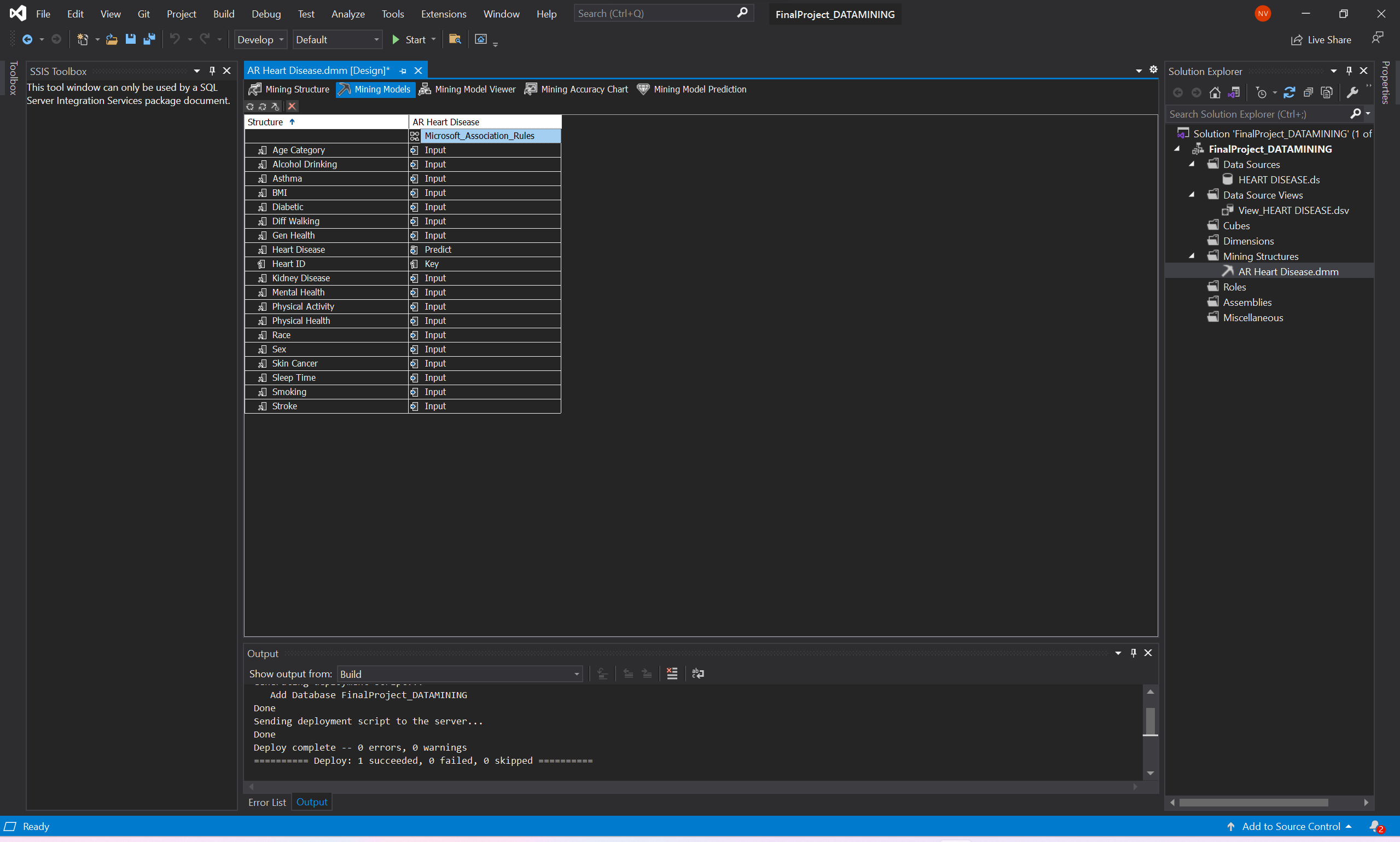
* Đặt tên cho Mining Structure và Mining Model là **AR Heart Disease** và chọn Finish để hoàn tất quá trình tạo Mining Structure.



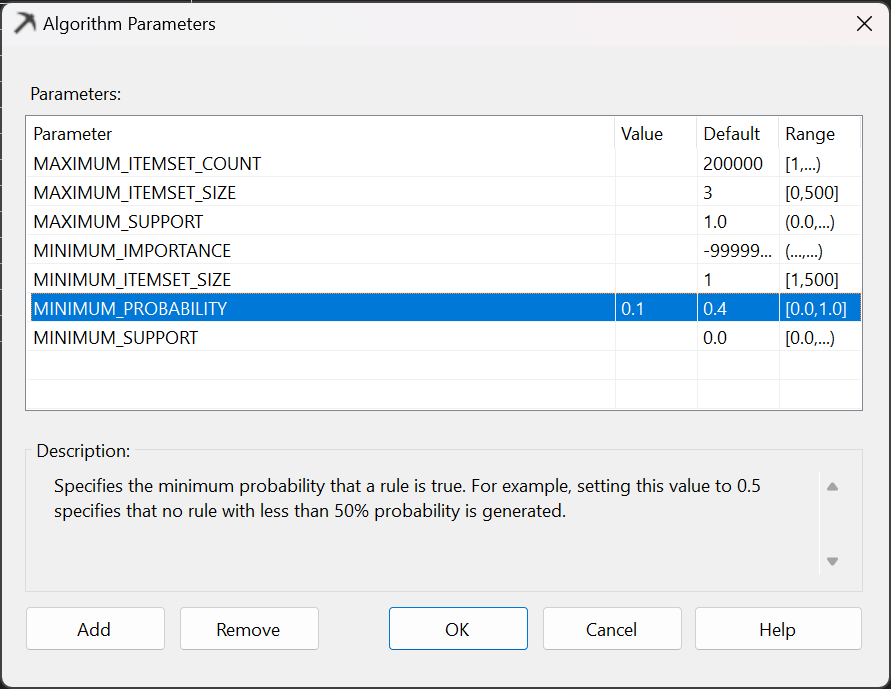
* Mining Structures sau khi tạo như sau:

****

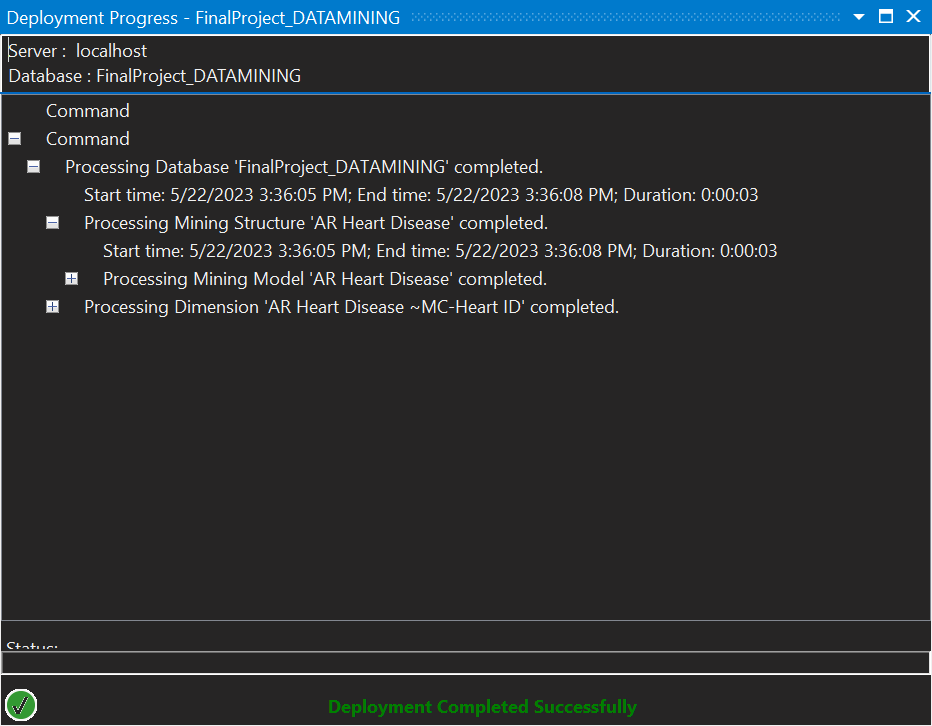
* Mining Models bao gồm các biến Input, Key và Predict như đã chọn.

****

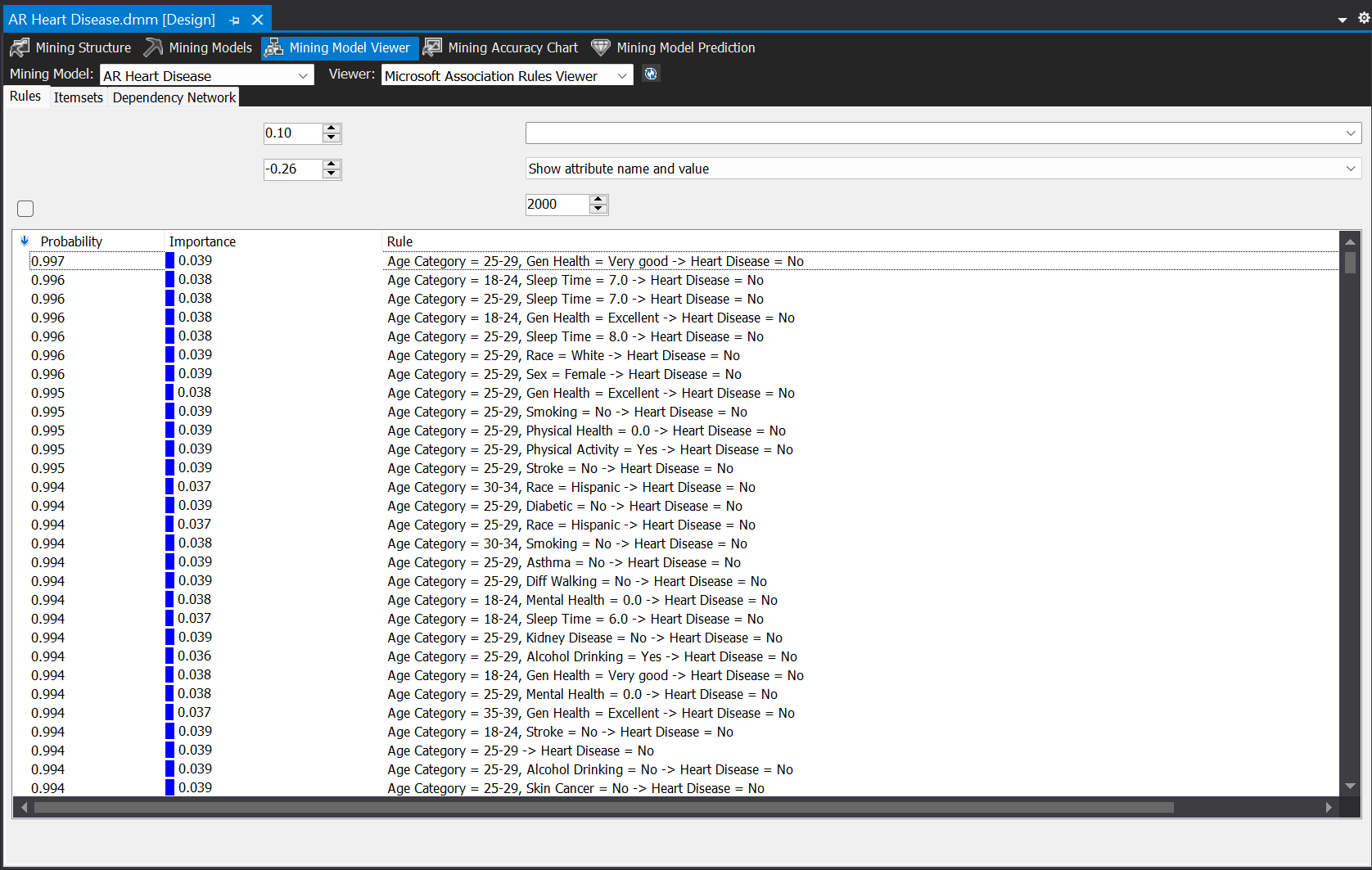
* Chỉnh sửa các thuộc tính của thuật toán cho phù hợp với thuật toán. Ở đây ta chọn MINIMUM\_PROBABILITY là 0.1

****

* Deploy Project, kết quả hiện như sau là thành công.

****

* Ở phần **Mining Model Viewer**, chọn mục **Rules**. Màn hình hiển thị các luật tìm được bằng giải thuật Association Rules với các thuộc tính mà ta lựa chọn. Trong đó:
* Probability: khả năng xảy ra của luật
* Importance: mức độ ảnh hưởng của luật
* Rule: luật tìm được

****

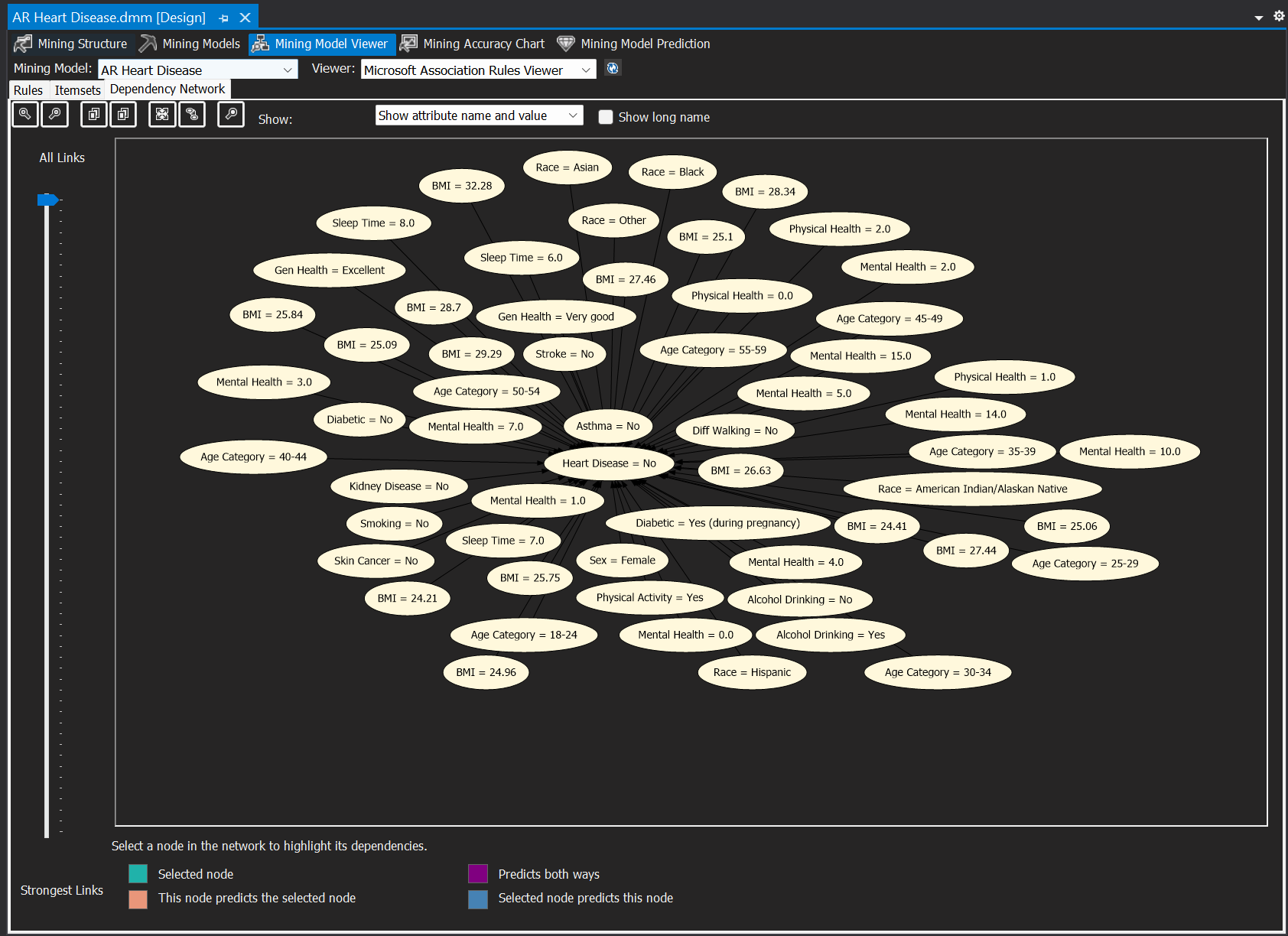
Ví dụ giải thích một Rule như sau:

*Một người có độ tuổi tầm 25-29, sức khỏe rất tốt thì 99.7% không mắc bệnh tim và mức độ ảnh hưởng của tập này trong tập dữ liệu là 3.9%.*

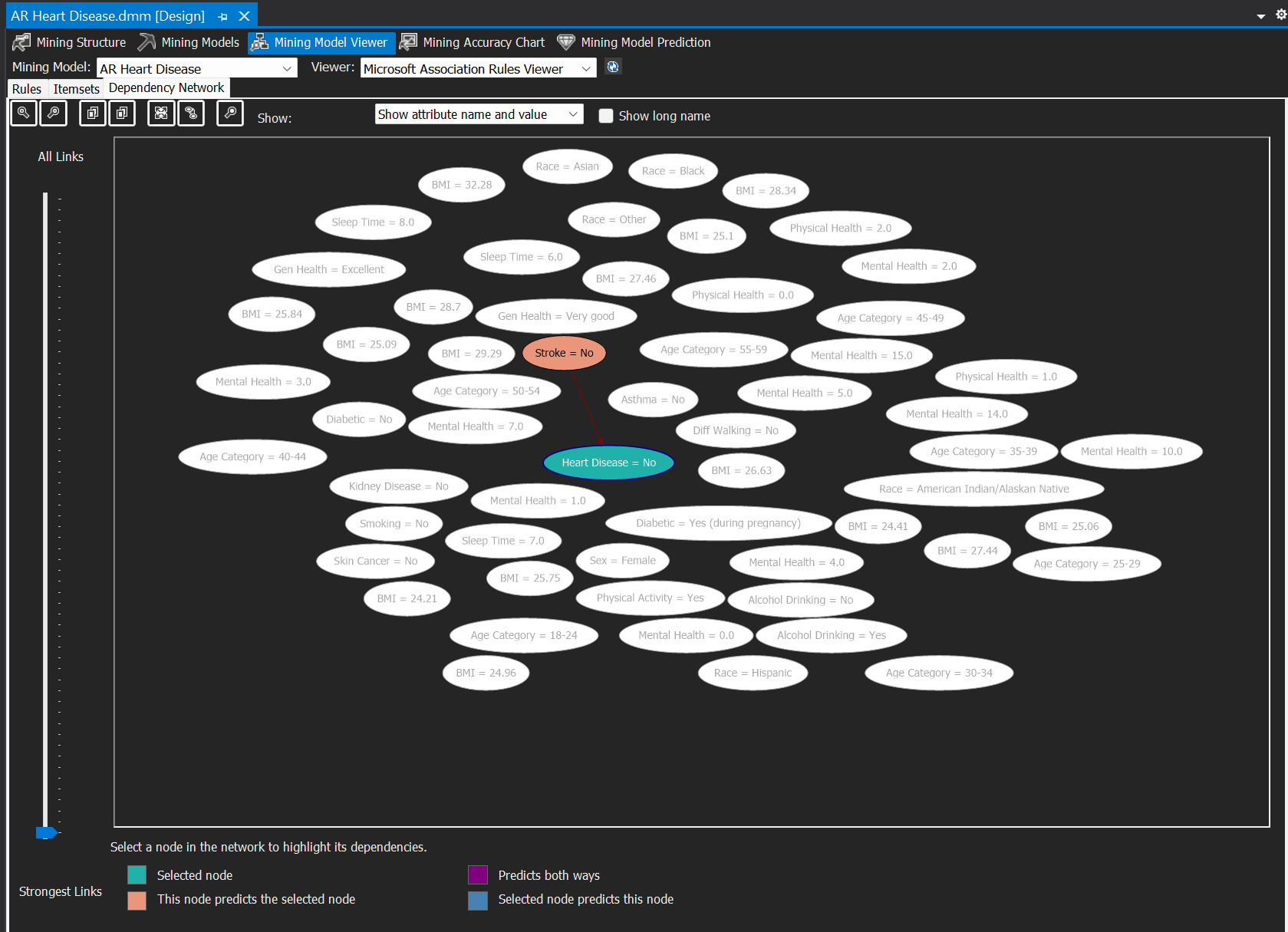
* Những Itemsets tìm được nhờ thuật toán này với các thuộc tính
* Support: độ phổ biển của Itemset trong tập dữ liệu
* Size: số lượng các phần tử của Itemset

****

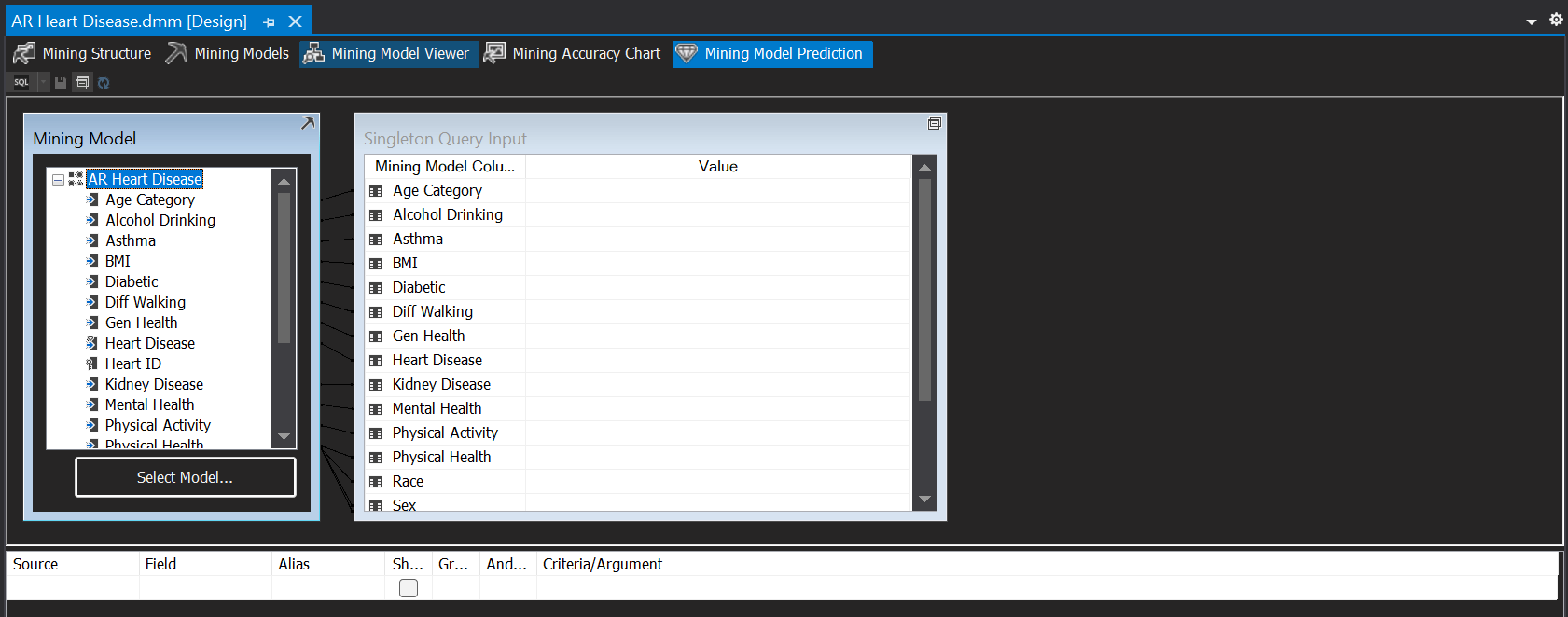
* Dependency Network (mô hình biểu diễn sự phụ thuộc và mối quan hệ của các biến trong tập dữ liệu)

****

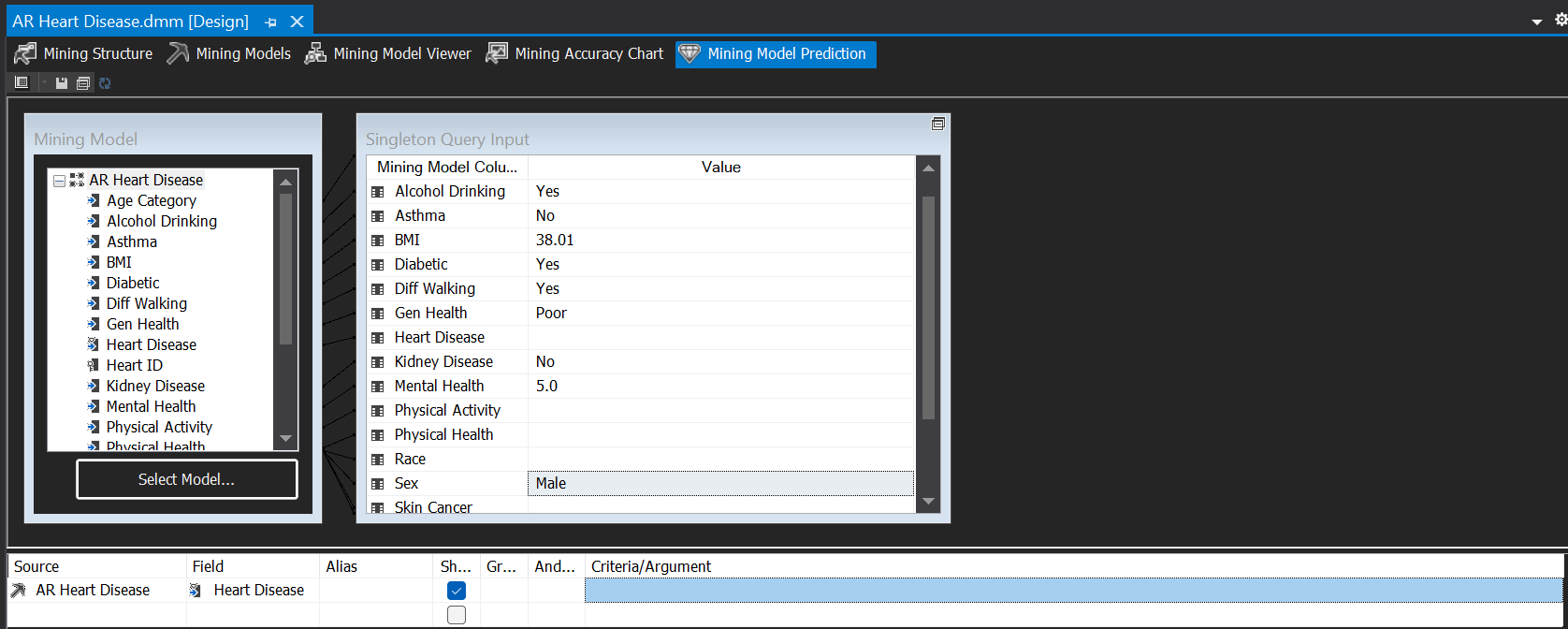
* Ở thanh All Links bên trái, càng kéo xuống thấp thì mức độ liên kết của các Node càng mạnh, hay nói cách khác là mức độ ảnh hưởng càng cao.

****

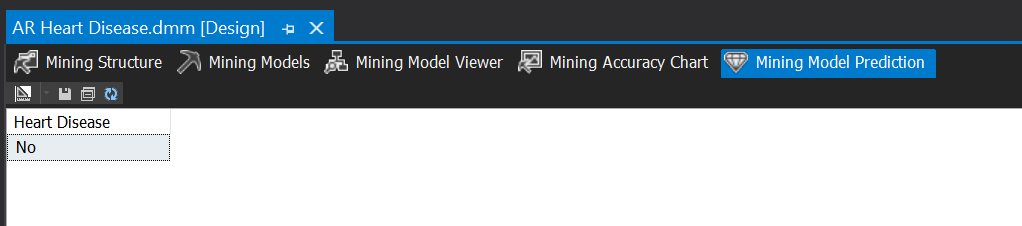
* Ở mục Mining Model Prediction, ta sẽ thử dự đoán dựa trên các biến Input.

****

* Tạo một Singleton Query với các thuộc tính được chọn như dưới, liệu một người có nghiện rượu, không hen suyễn, chỉ số BMI là 38.01, bị tiểu đường, có vấn đề sức khỏe khi đi bộ, sức khỏe không tốt, không bị bệnh thận và chỉ số sức khỏe tinh thần ở mức 5 thì có khả năng bị bệnh tim không?

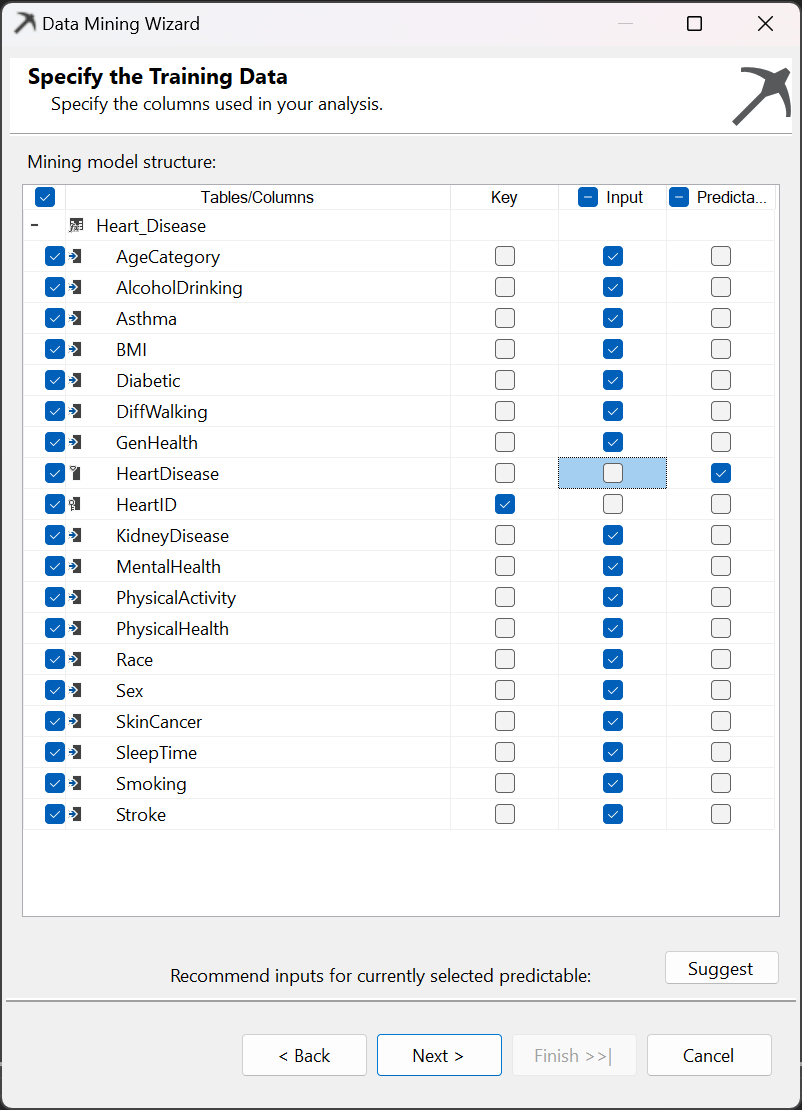


* + Kết quả trả về là không.

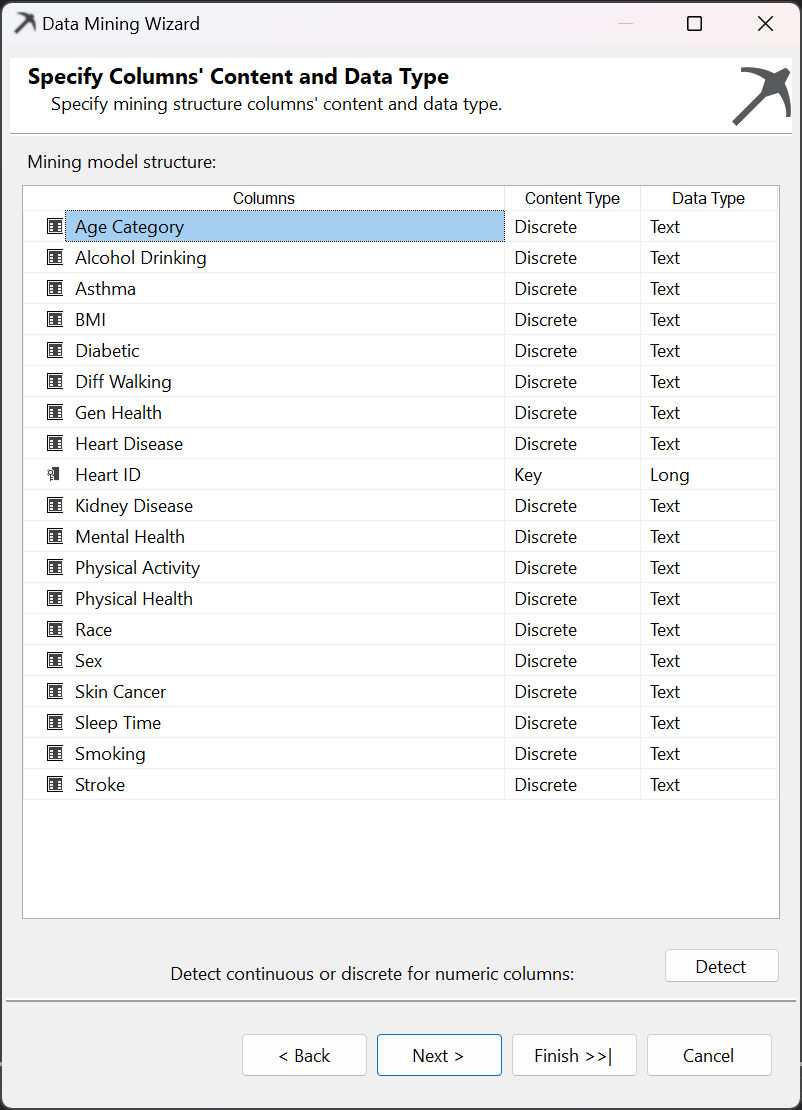


1. **Clustering (K-Mean)**

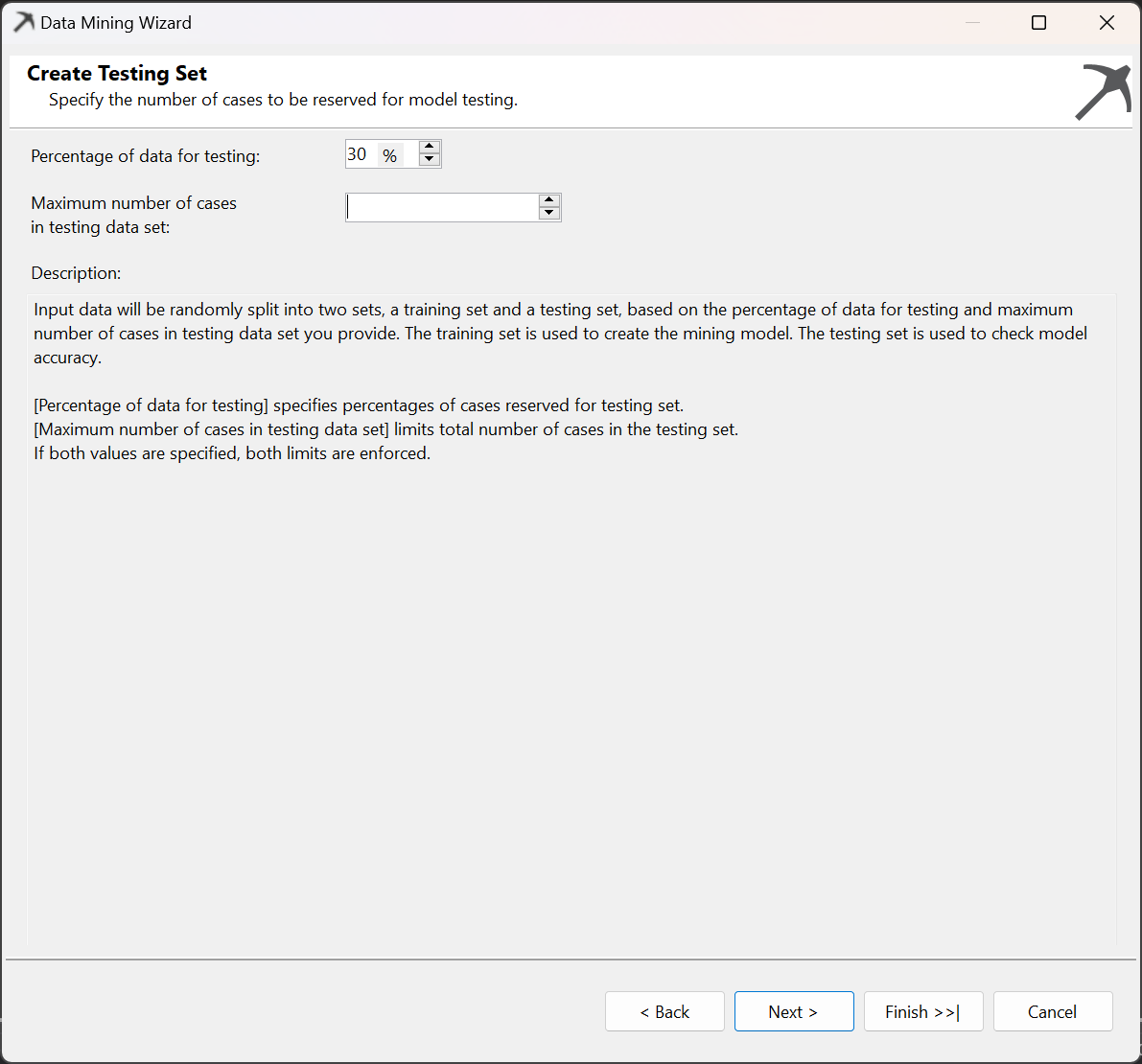
* Mục đích của thuật toán này là phân dữ liệu thành các cụm khác nhau sao cho dữ liệu trong cùng một cụm có tính chất giống nhau.
* Tạo một Mining Structure có tên **CT Heart Disease** với bảng Heart\_Disease là Case và chọn các thuộc tính như hình dưới làm Key, Input và Predictable.

****

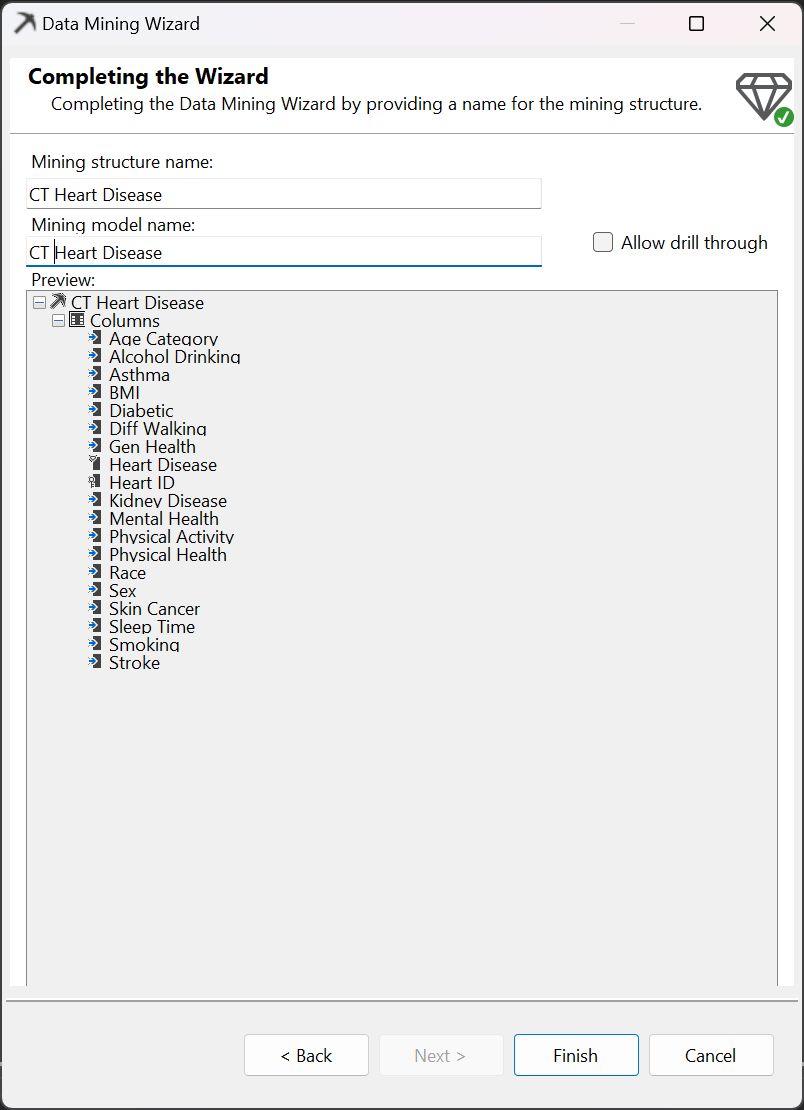
* Nhấn Next



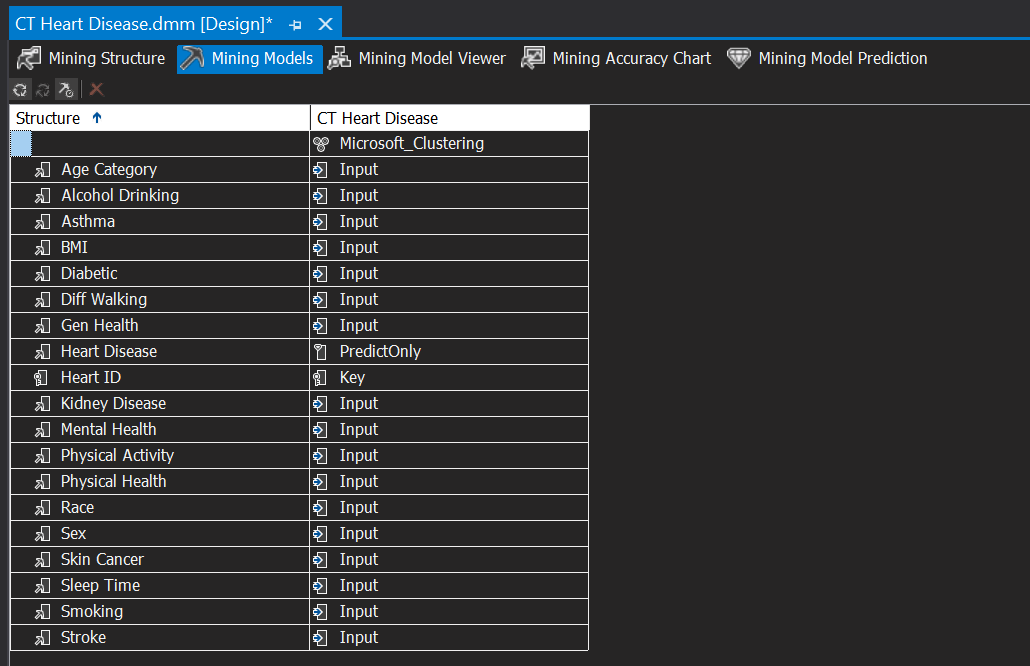
* Chọn 30% tập dữ liệu để test



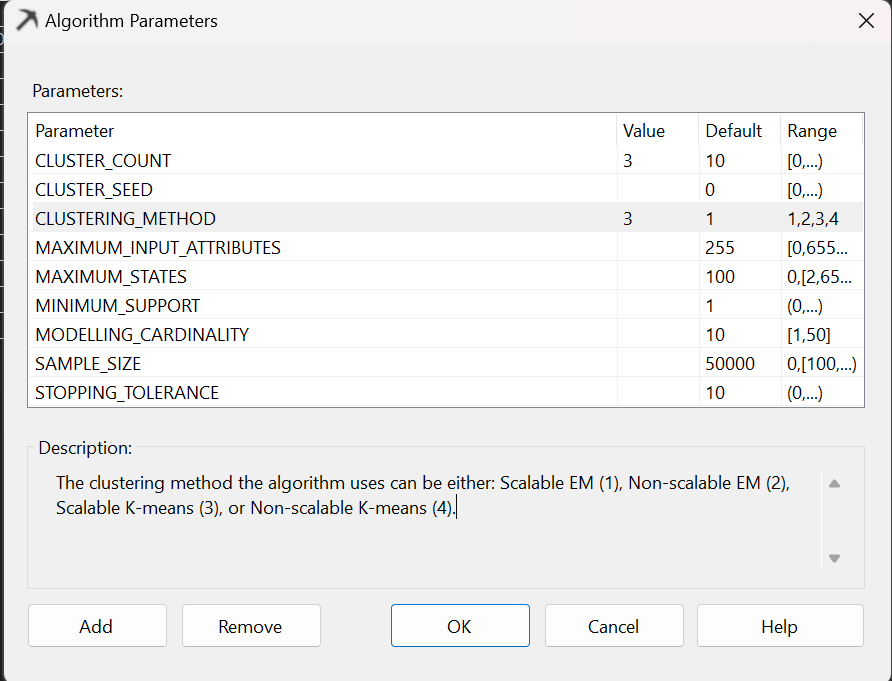
* Đặt tên cho Mining Structure và Mining Model là **CT Heart Disease**

****

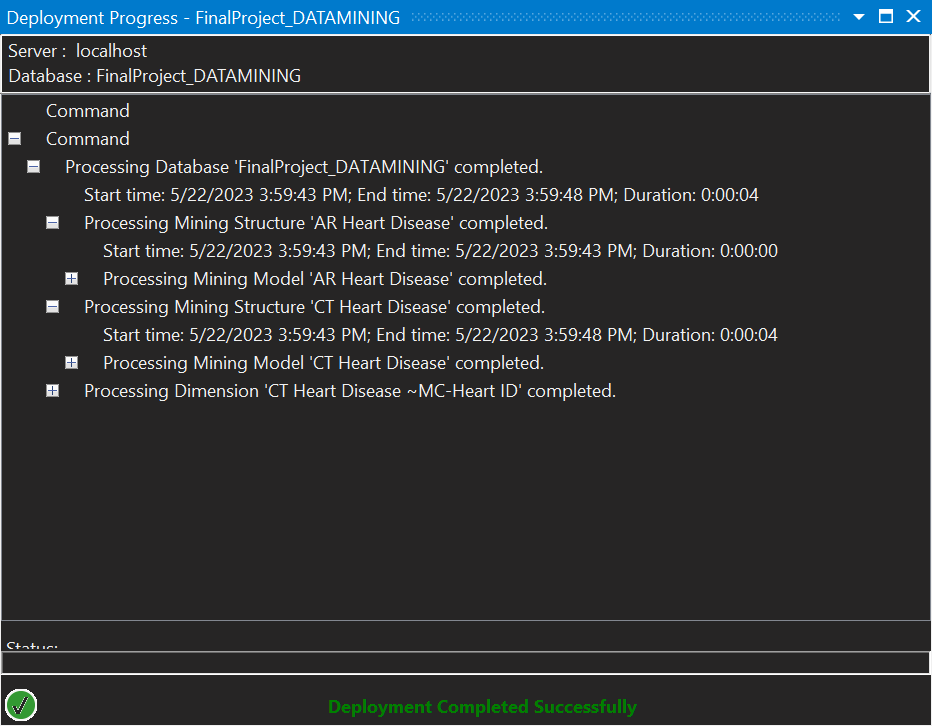
* Mining Models sau khi tạo như sau:

****

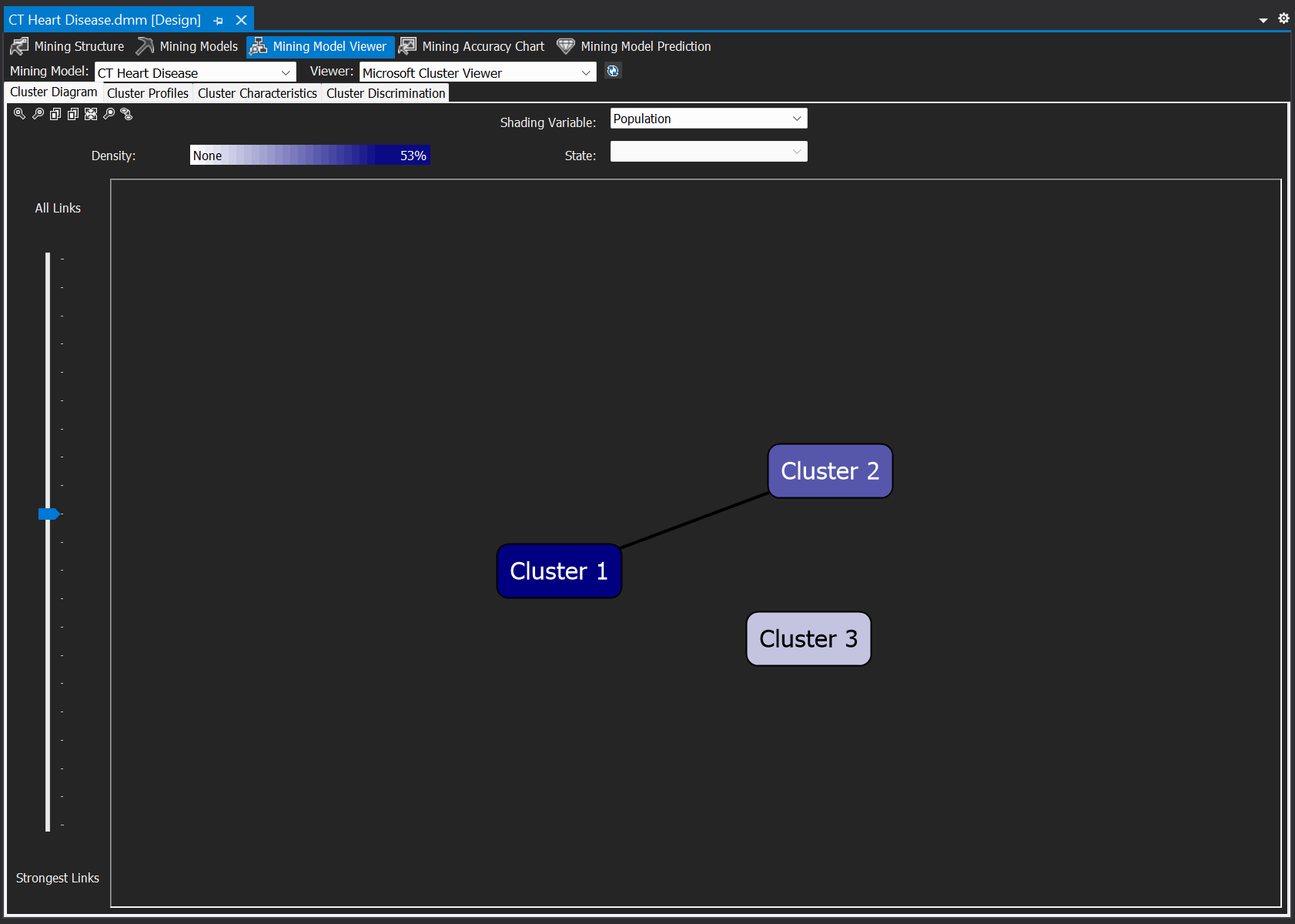
* Chọn số cụm là 3 và thuật toán KMean

****

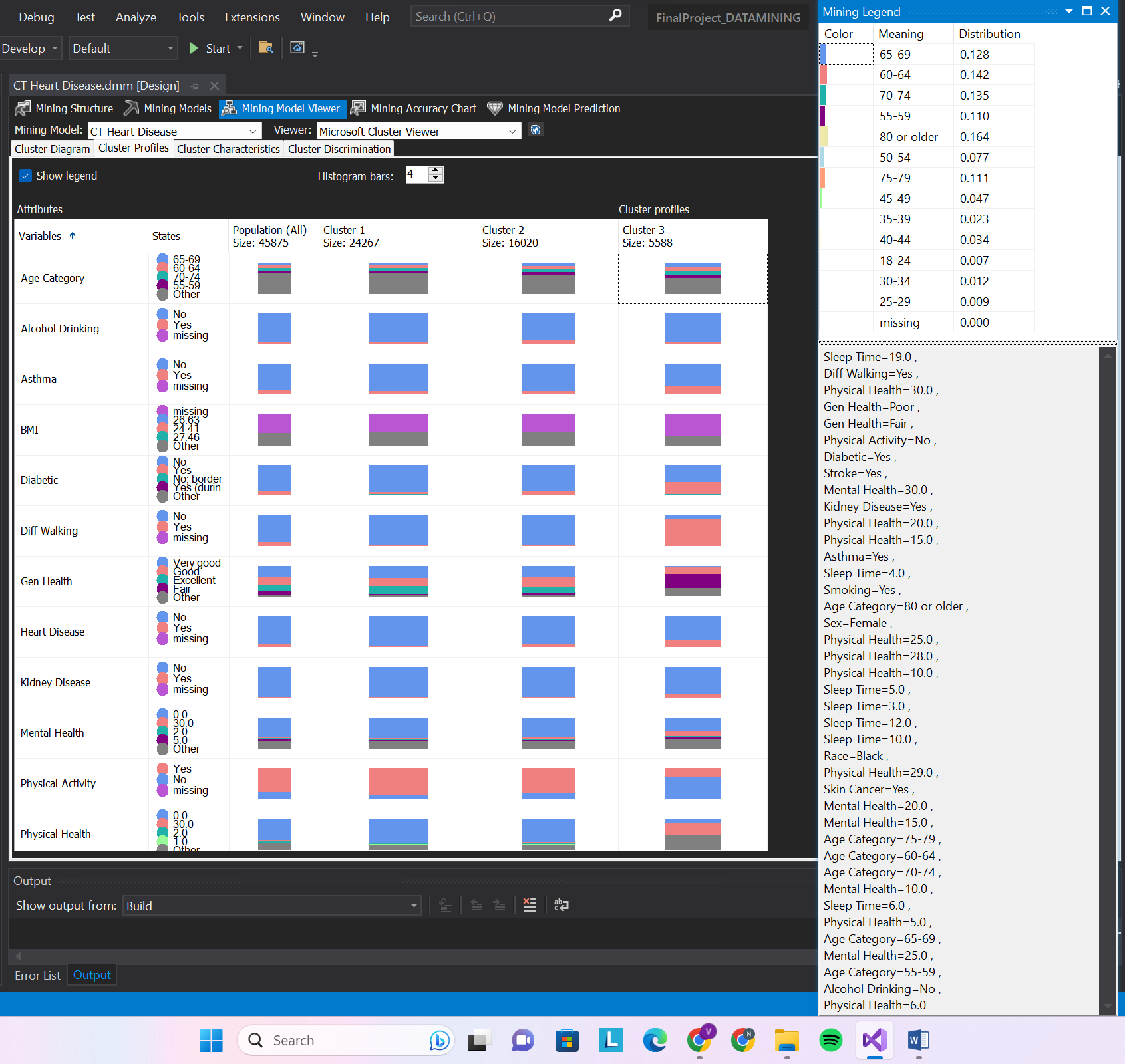
* Deploy Project, hiển thị như bên dưới là thành công.

****

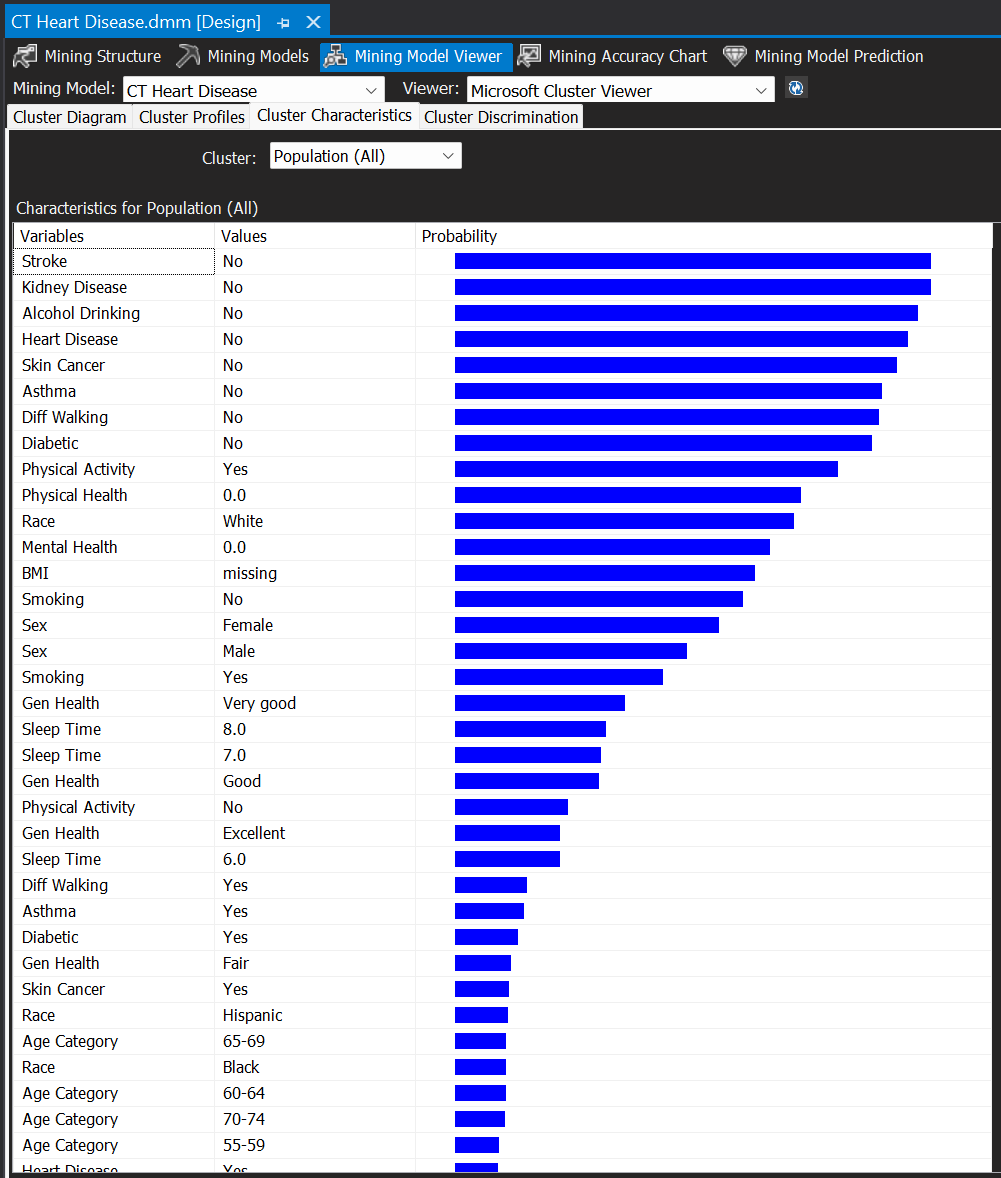
* Mining Model Viewer hiển thị như sau, cluster có màu càng đậm thì càng có nhiều phần tử.

****

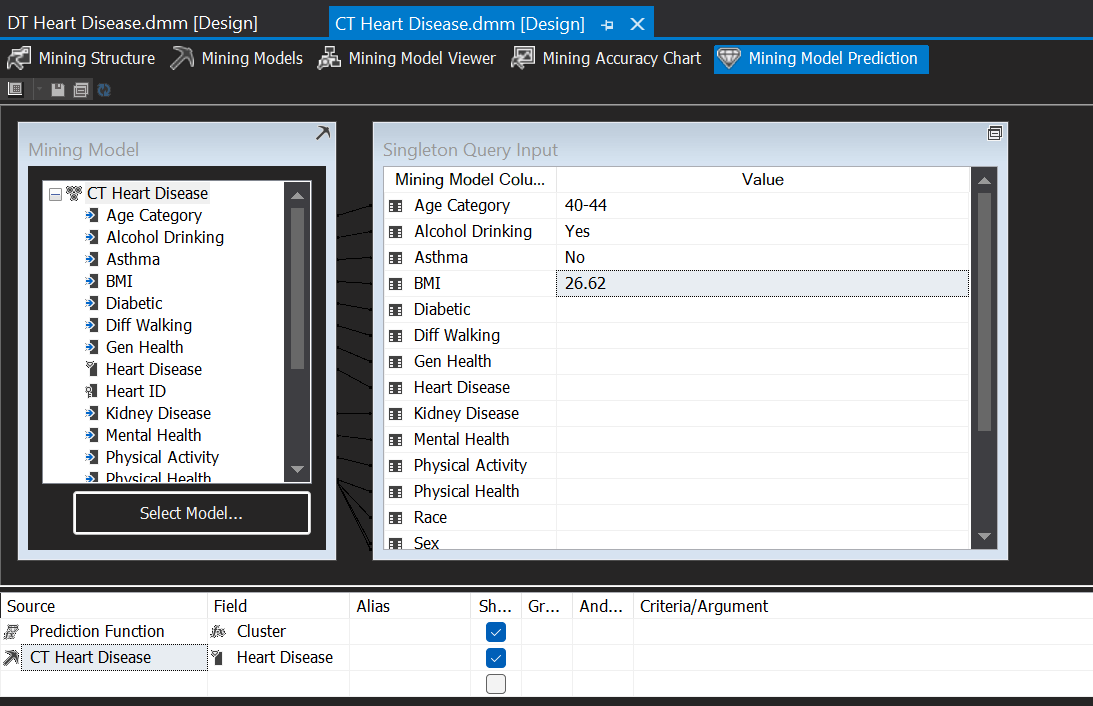
* Ở tab Cluster Profiles, sẽ hiển thị như sau, bấm vào một cluster bất kỳ sẽ hiển thị thông tin chi tiết của cluster đó, bao gồm mức độ phân bố của mỗi biến và các biến khác thuộc cùng cluster đó.

****

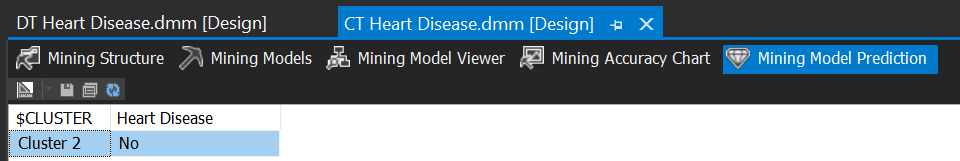
* Ở tab Cluster Characteristics, hiển thị các thuộc tính, giá trị và xác suất xuất hiện của nó trong toàn bộ cluster.



* Ở mục Mining Model Prediction, ta sẽ dự đoán dựa trên các giá trị mình lựa chọn. Ở đây ta sẽ thử dự đoán xem một người có độ tuổi từ 40-44, có nghiện rượu, không bị hen suyễn và có chỉ số BMI là 26.62 thì thuộc cluster nào và có khả năng bị bệnh tim không?

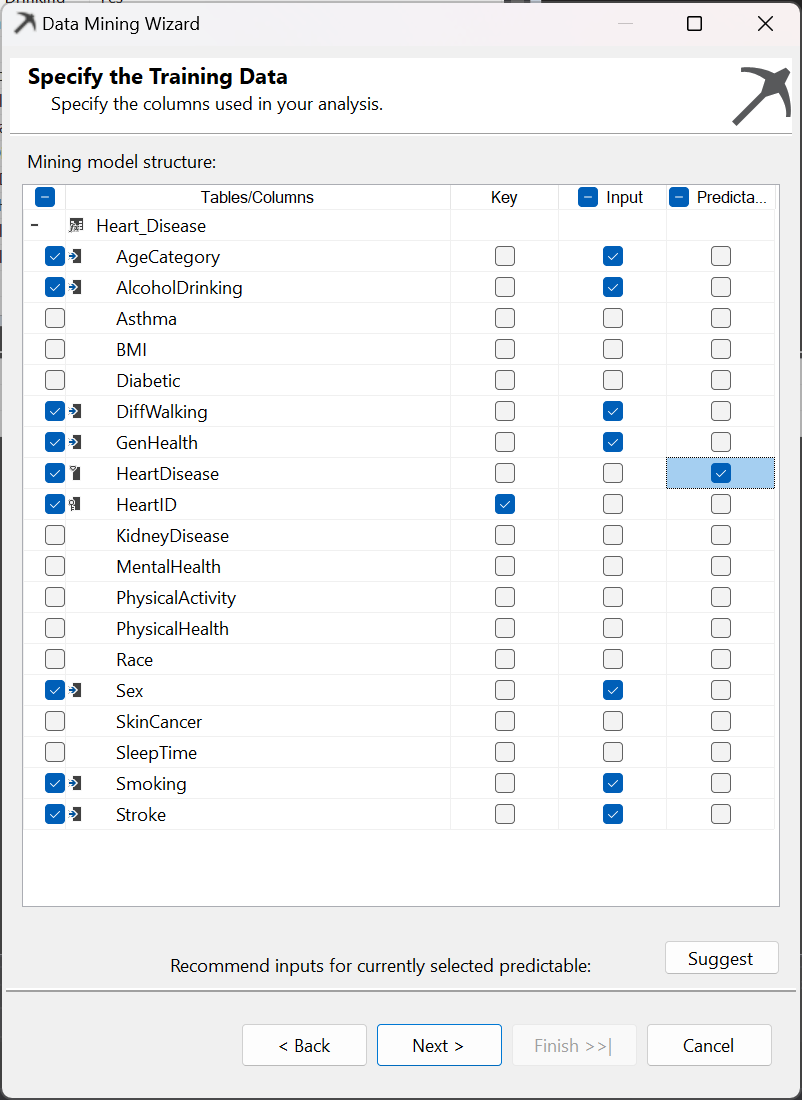


* + Kết quả trả về rằng người có những thuộc tính như trên thì không bị bệnh tim và thuộc nhóm 2.

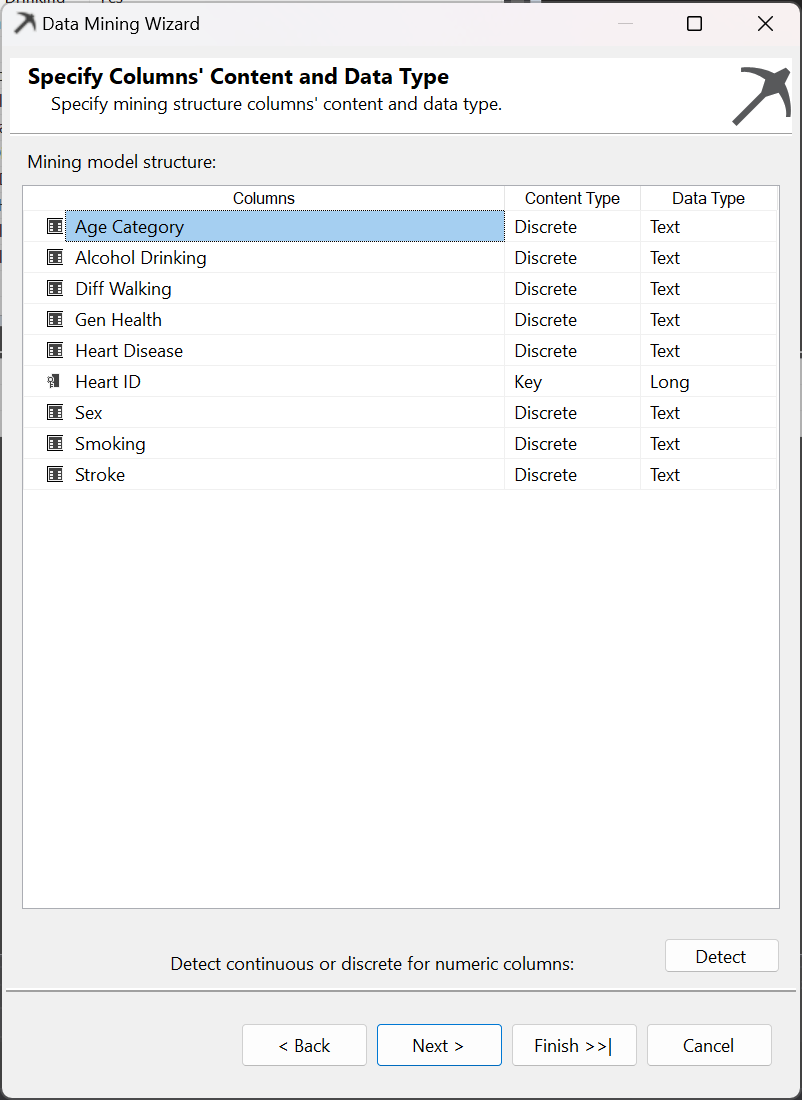


1. **Decision Trees**

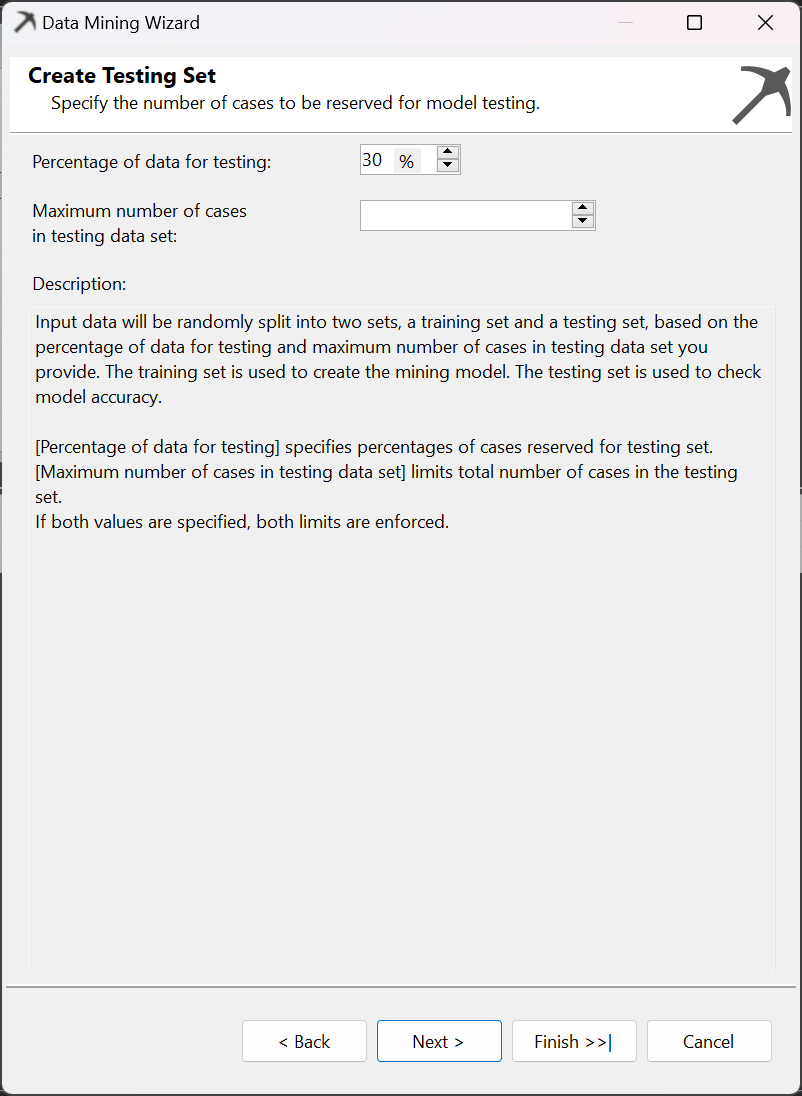
* Giải thuật này được sử dụng khá phổ biến và hiệu quả trong cả hai lớp bài toán phân loại và dự báo. Mô hình này không tồn tại phương trình dự báo, việc ta cần làm là tìm ra một cây quyết định dự báo tốt trên tập train và tập test.
* Tạo một Mining Structure có tên **DT Heart Disease** với bảng Heart\_Disease là Case và chọn các thuộc tính như hình dưới làm Key, Input và Predictable.



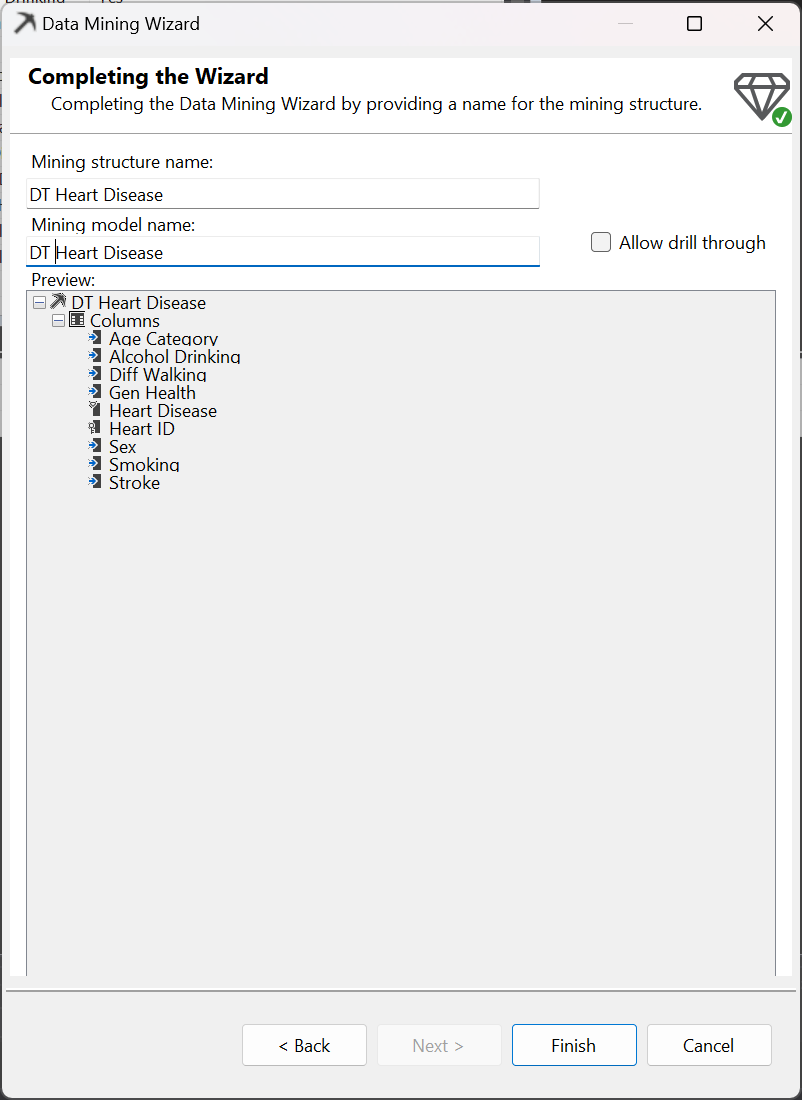
* Nhấn Next



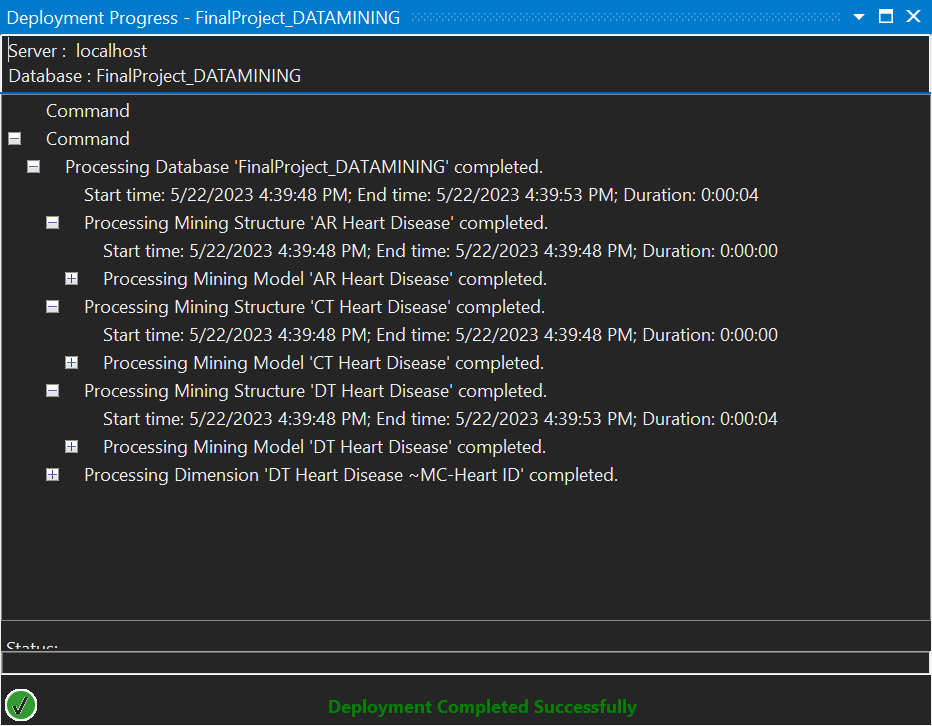
* Ta lấy 30% tập dữ liệu để test



* Đặt tên cho Mining Structure và Mining Model là **DT Heart Disease**

****

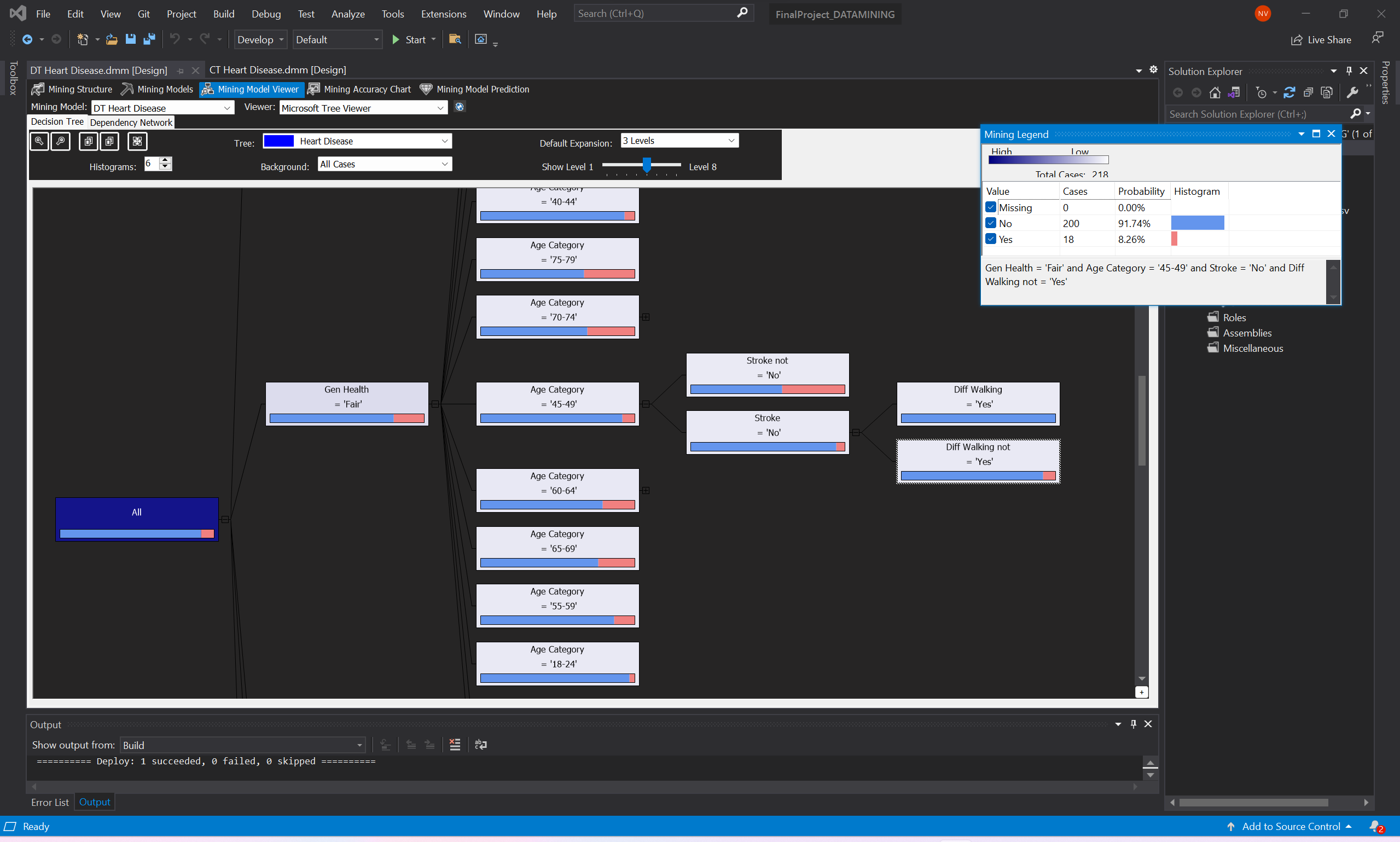
* Deploy Project, hiển thị như sau là thành công.

****

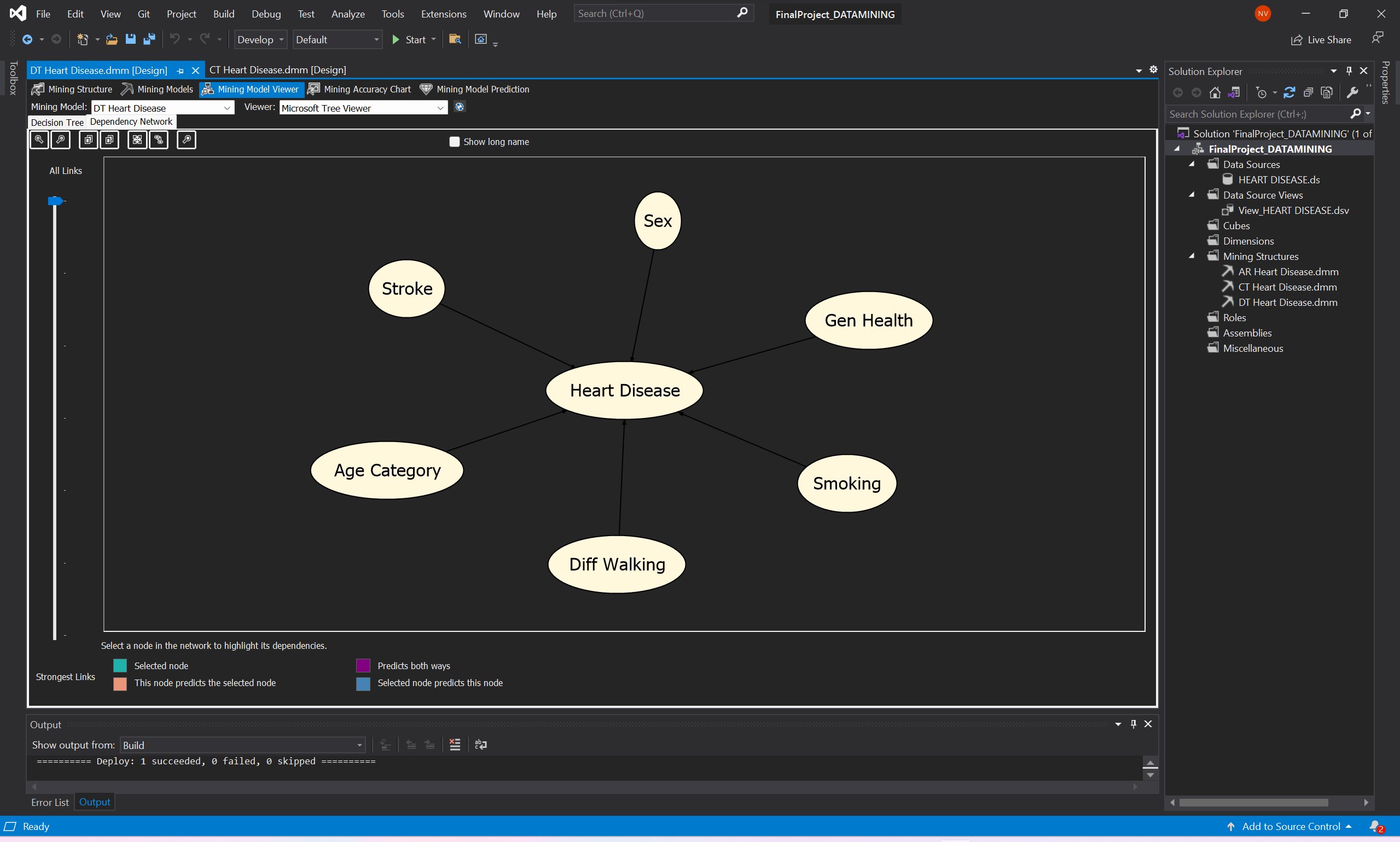
* Mining Model Viewer hiển thị như sau, mở rộng cây và chọn vào một node thuộc cây đó, màn hình hiển thị các thuộc tính trên nhánh chứa node đó và xác suất xuất hiện của nhánh đó trên tập test.

Giải thích hình bên dưới:

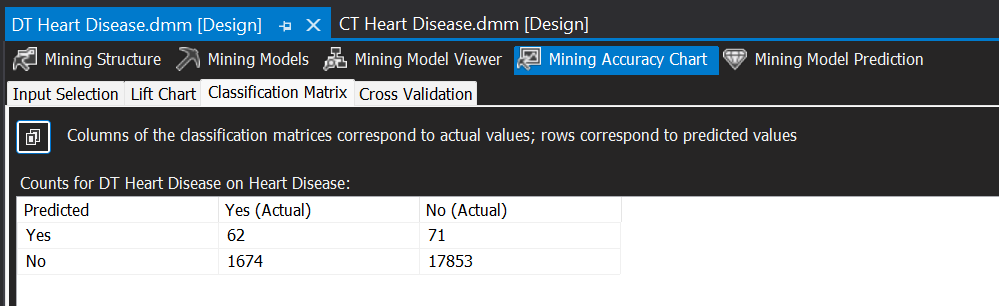
*Một người có sức khỏe trung bình, độ tuổi từ 45-49, chưa bị đột quỵ và có khó khăn trong việc đi bộ thì có 200 trường hợp không mắc bệnh tim (xác suất là 91.74%) và 18 trường hợp mắc bệnh tim (xác suất là 8.26%)*

****

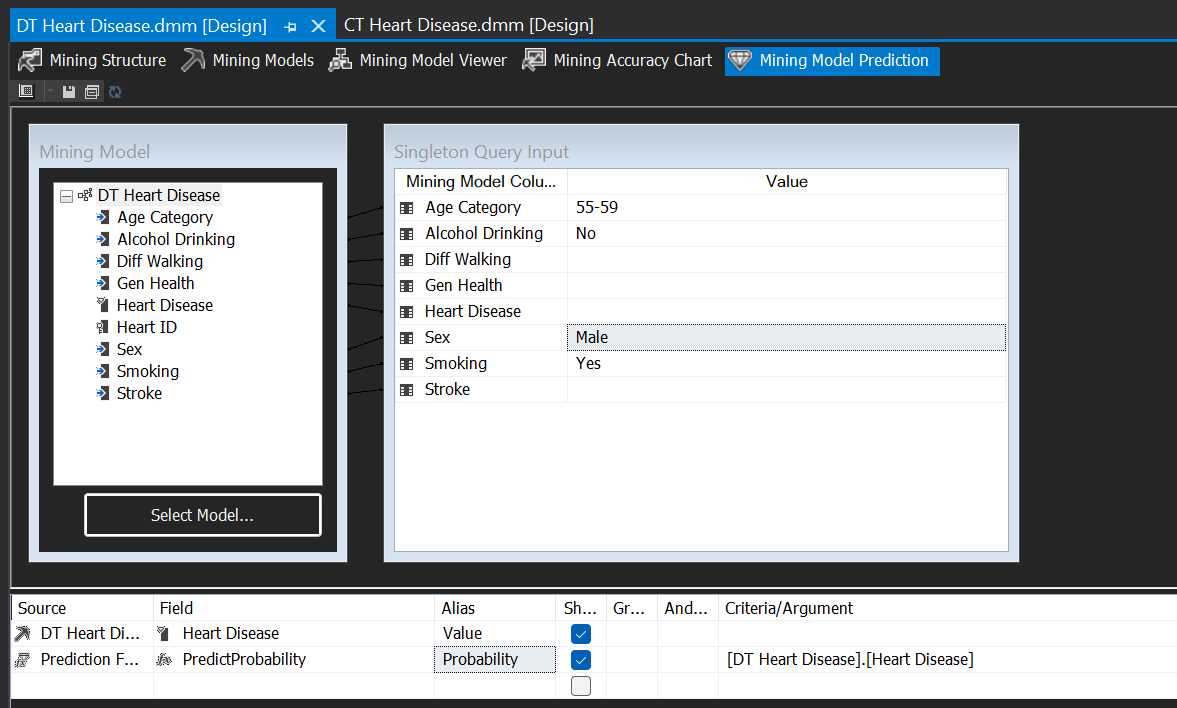
* Dependency Network (mô hình biểu diễn sự phụ thuộc và mối quan hệ của các biến Input)

****

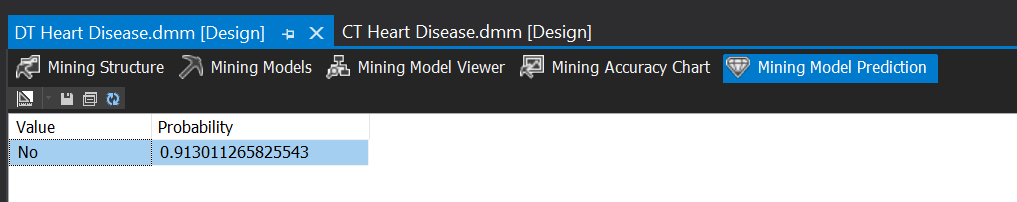
* Ở tab Mining Accuracy Chart, hiển thị rằng có 62 trường hợp thực sự bị bệnh tim thật sự được dự đoán đúng và có 17853 trường hợp thực sự không bị bệnh tim được dự đoán đúng.

****

* Ở mục Mining Model Prediction, ta thử dự đoán một người có độ tuổi từ 55-59, không nghiện rượu, giới tính nam và có hút thuốc thì có khả năng bị bệnh tim không và xác suất là bao nhiêu?



* Kết quả là người đó không bị bệnh tim với xác suất là xấp xỉ 91.3%



**III.Thực hiện thuật toán phân cụm Kmeans trên weka**

* + - 1. Kmeans:
  + Thực hiện Import data từ CSV: **Open File** 🡪 Chọn đường dẫn chứa file data và chọn loại file import là **CSV data files 🡪 Open**

A screenshot of a computer

Description automatically generated

* + Thực hiện chọn thuộc tính cho kiểu dữ liệu đưa vào: **Weka** 🡪 **filters** 🡪 **unsupervised**🡪 **attribute** 🡪 **NumericToNominal**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

* + Thực hiện lựa chọn thuật toán Kmeans: **Cluster** 🡪 **Choose** 🡪 **SimpleKmeans**

A screenshot of a computer

Description automatically generated

* + Điều chỉnh các thuộc tính cho thuật toán, ở đây chúng ta chủ yếu thay đổi số cụm tại ô **NumClusters**

A screenshot of a computer

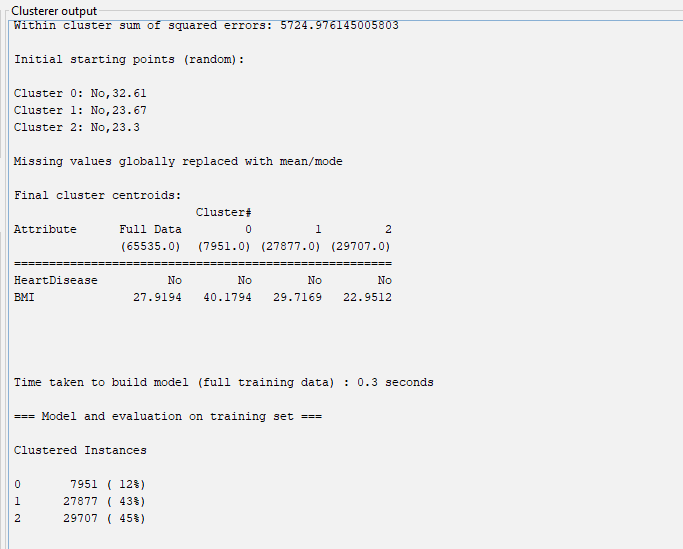
Description automatically generated

* + Sau đó chúng ta chọn “**ignore attributes**” để chọn các thuộc tính phân cụm

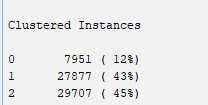
A screenshot of a computer

Description automatically generated

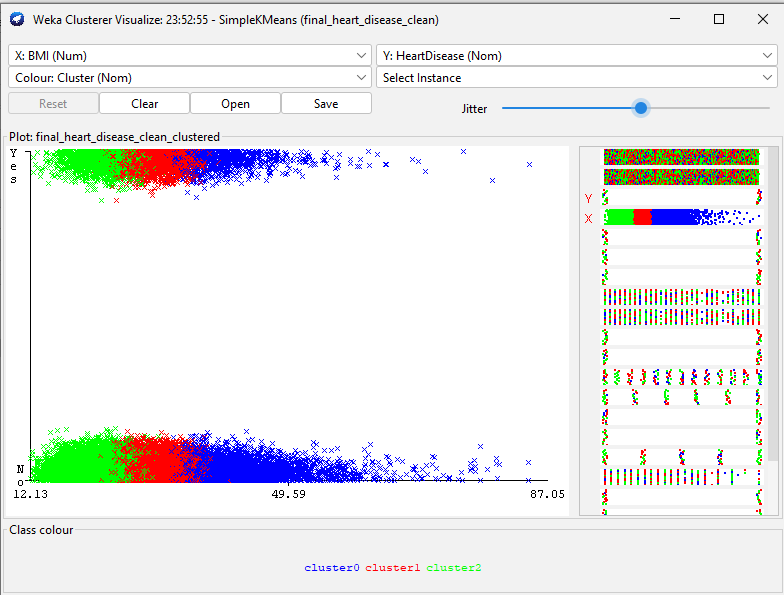
* + Bấm **Start** để thực hiện nhóm cụm



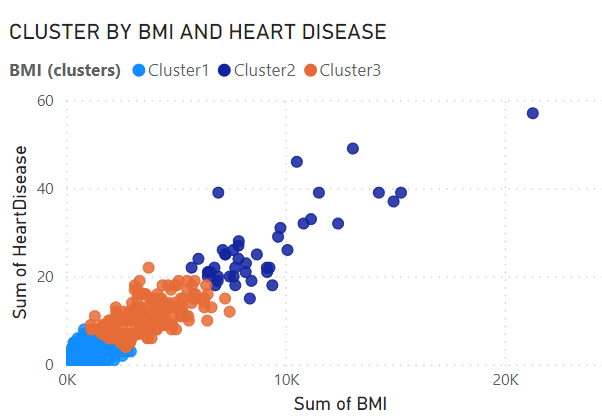
* + Sau khi nhóm cụm sẽ hiện ra 1 bảng kết quả với các cụm được chọn ra từ tập dữ liệu và số dòng tương ứng từ tập dữ liệu của các cụm



* + Để có thể trực quan hóa dữ liệu của các cụm, ta thực hiện chuột phải vào kết quả 🡪 **Visualize cluster assignments**



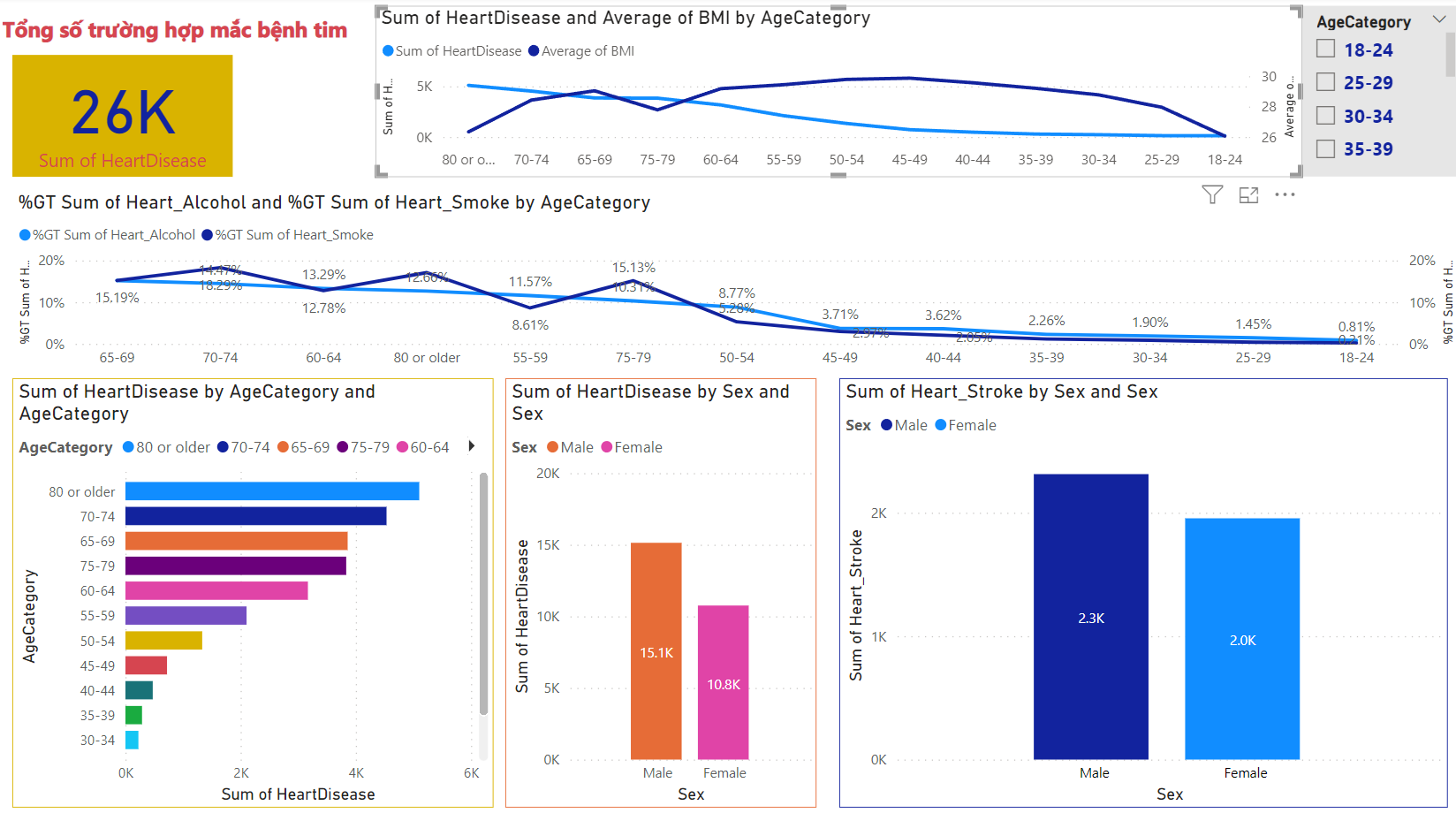
1. **Thực hiện trực quan hóa dữ liệu trên Power BI**
   * + 1. **Thực hiện giải thuật Clustering bằng thuộc tính BMI và Heart Disease**

****

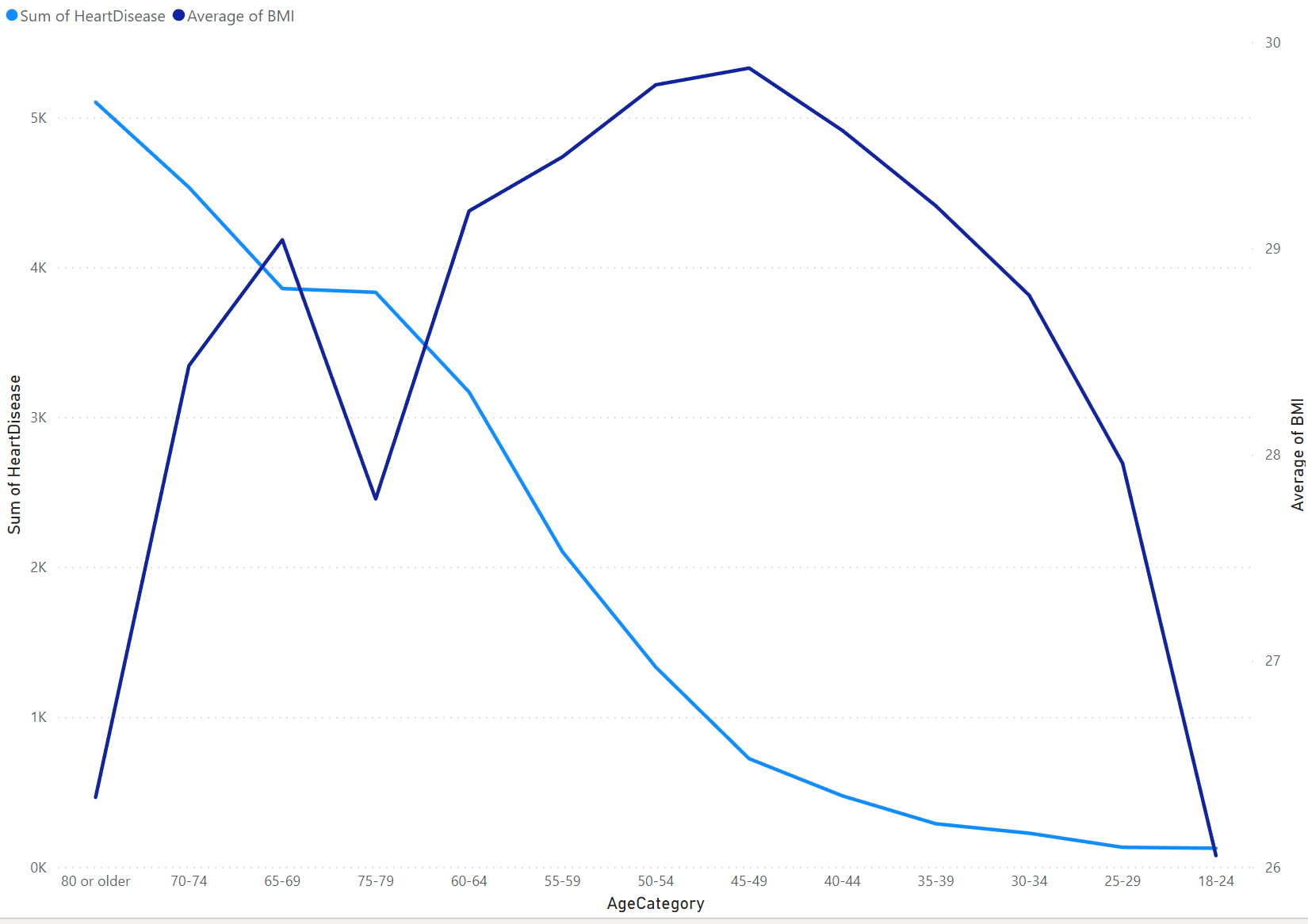
**Kết luận:** ta chia thành 3 cụm:

* Cụm 1: gồm những chỉ số BMI có từ 0 đến 5 ca mắc bệnh tim trong tập dữ liệu.
* Cụm 2: gồm những chỉ số BMI có từ 21 đến 50 ca mắc bệnh tim trong tập dữ liệu.
* Cụm 3: gồm những chỉ số BMI có từ 6 đến 20 ca mắc bệnh tim trong tập dữ liệu.
  + - 1. **Trực quan hóa dữ liệu bằng các thuộc tính khác**

Dashboard trực quan hóa dữ liệu bằng các thuộc tính trong tập dữ liệu được lọc bởi biến độ tuổi:



1. **Tương quan giữa mức độ trung bình của chỉ số BMI với tổng số ca mắc bệnh tim ở từng độ tuổi**

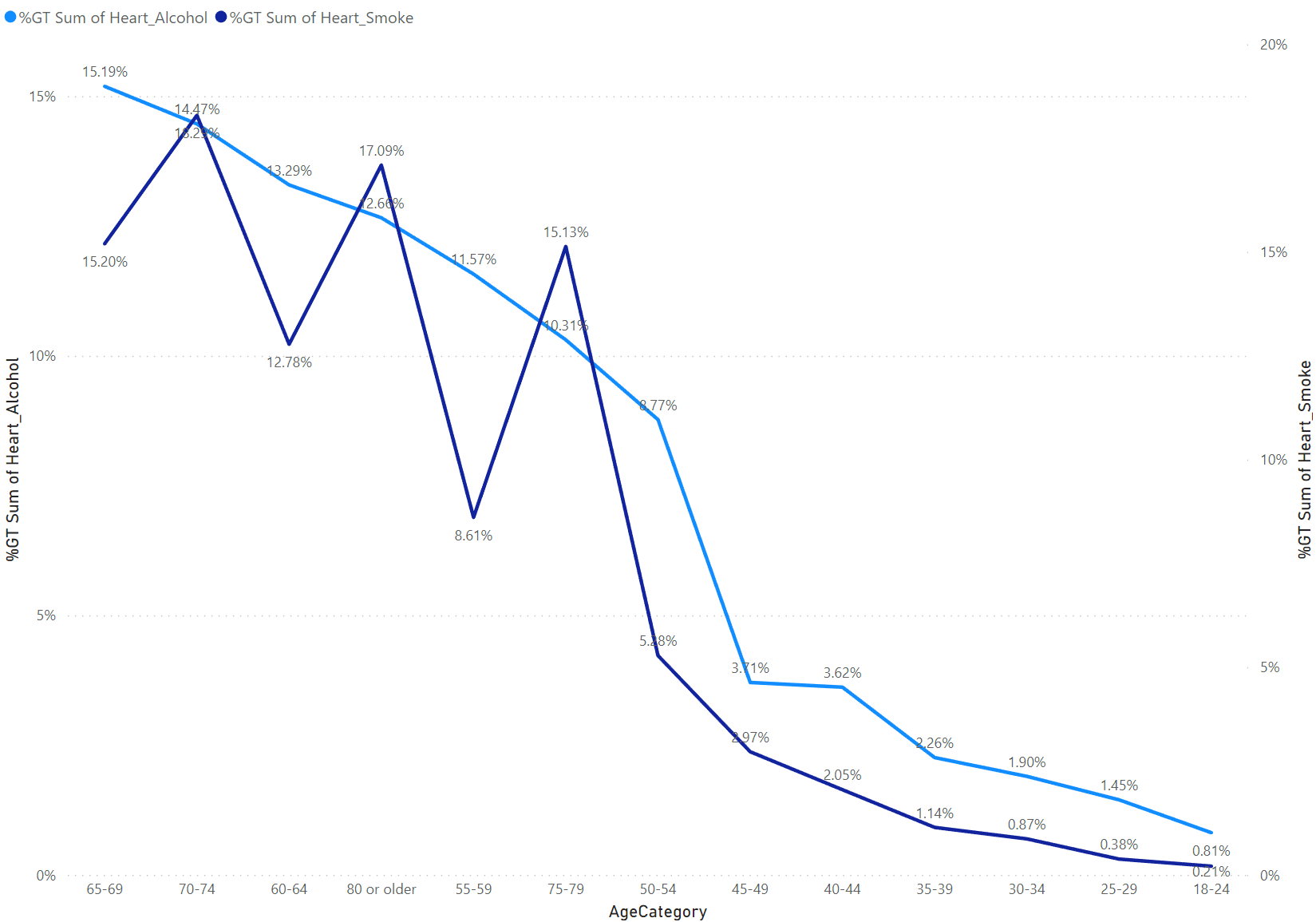


Từ biểu đồ ta có thể thấy chỉ số BMI trung bình tăng thì số lượng ca mắc bệnh tim cũng tăng ở những độ tuổi 18-24 đến 45-49. Ở những độ tuổi còn lại, có sự đối nghịch giữa trung bình BMI với số ca mắc bệnh tim, ví dụ ở độ tuổi 65-69 đến

70-74, chỉ số BMI trung bình giảm nhưng số ca mắc bệnh tim vẫn tăng.

🡪 Ta có thể kết luận rằng, chỉ số BMI có ảnh hưởng đến khả năng mắc bệnh tim nhưng chỉ là một phần nhỏ, ở một số độ tuổi nhất định.

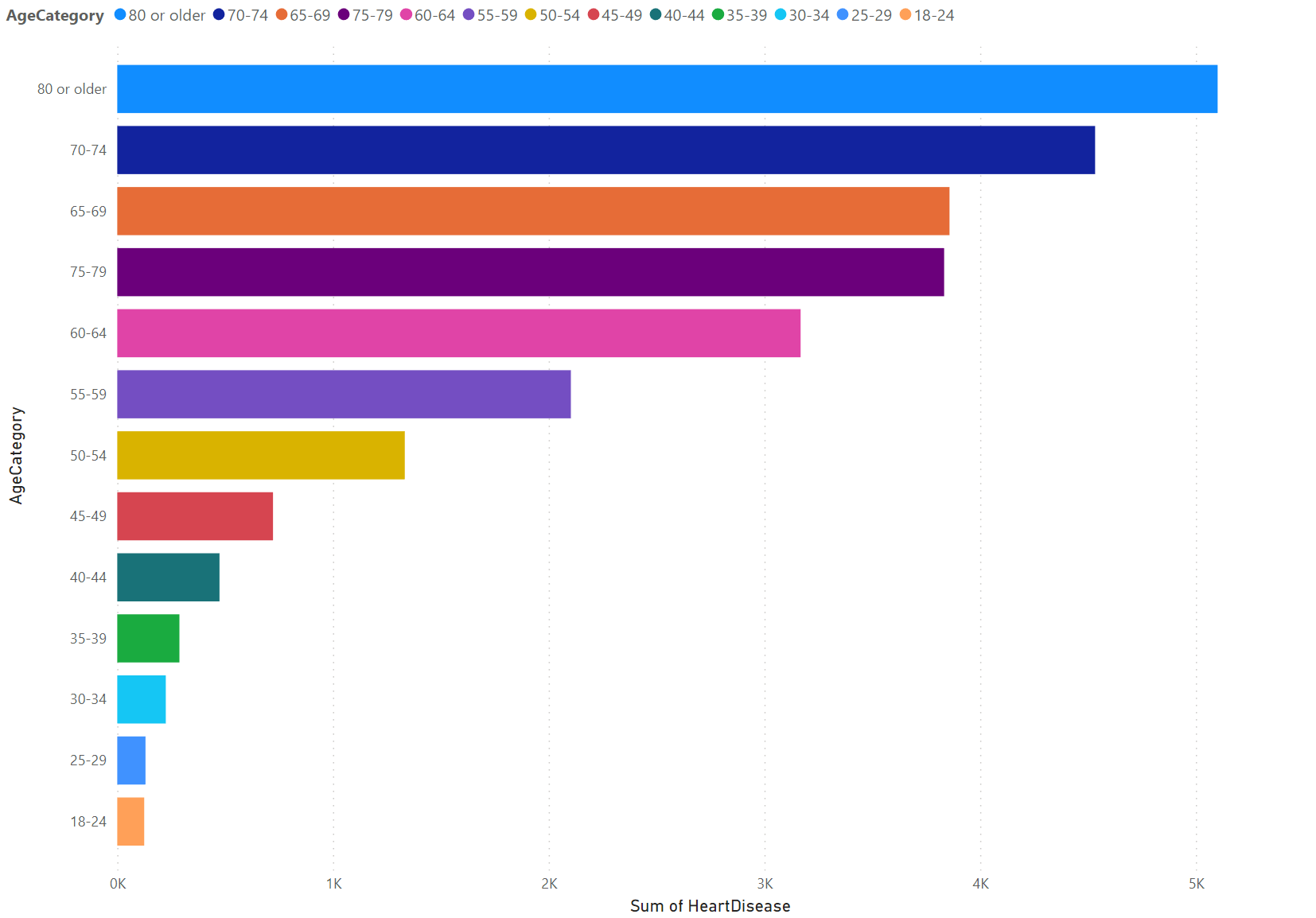
1. **Phần trăm người vừa bị bệnh tim vừa nghiện rượu và phần trăm người vừa bị bệnh tim vừa nghiện thuốc lá ở từng độ tuổi**

****

* Từ đồ thị ta có thể thấy, phần trăm người vừa nghiện rượu vừa bị bệnh tim tăng dần theo độ tuổi, còn phần trăm người vừa nghiện thuốc lá vừa bị bệnh tim tăng dần ở độ tuổi từ 18-24 đền 50-54, sau đó tăng giảm bất thường ở từng độ tuổi.
* Ở độ tuổi 45-49, có 2.97% người vừa nghiện thuốc lá vừa bị bệnh tim và 3.71% người vừa nghiện rượu vừa bị bệnh tim

🡪 Ta có thể thấy nghiện rượu và nghiện thuốc lá có ảnh hưởng đến khả năng mắc bệnh tim qua từng độ tuổi, tuy nhiên ảnh hưởng cao hay thấp tùy vào độ tuổi.

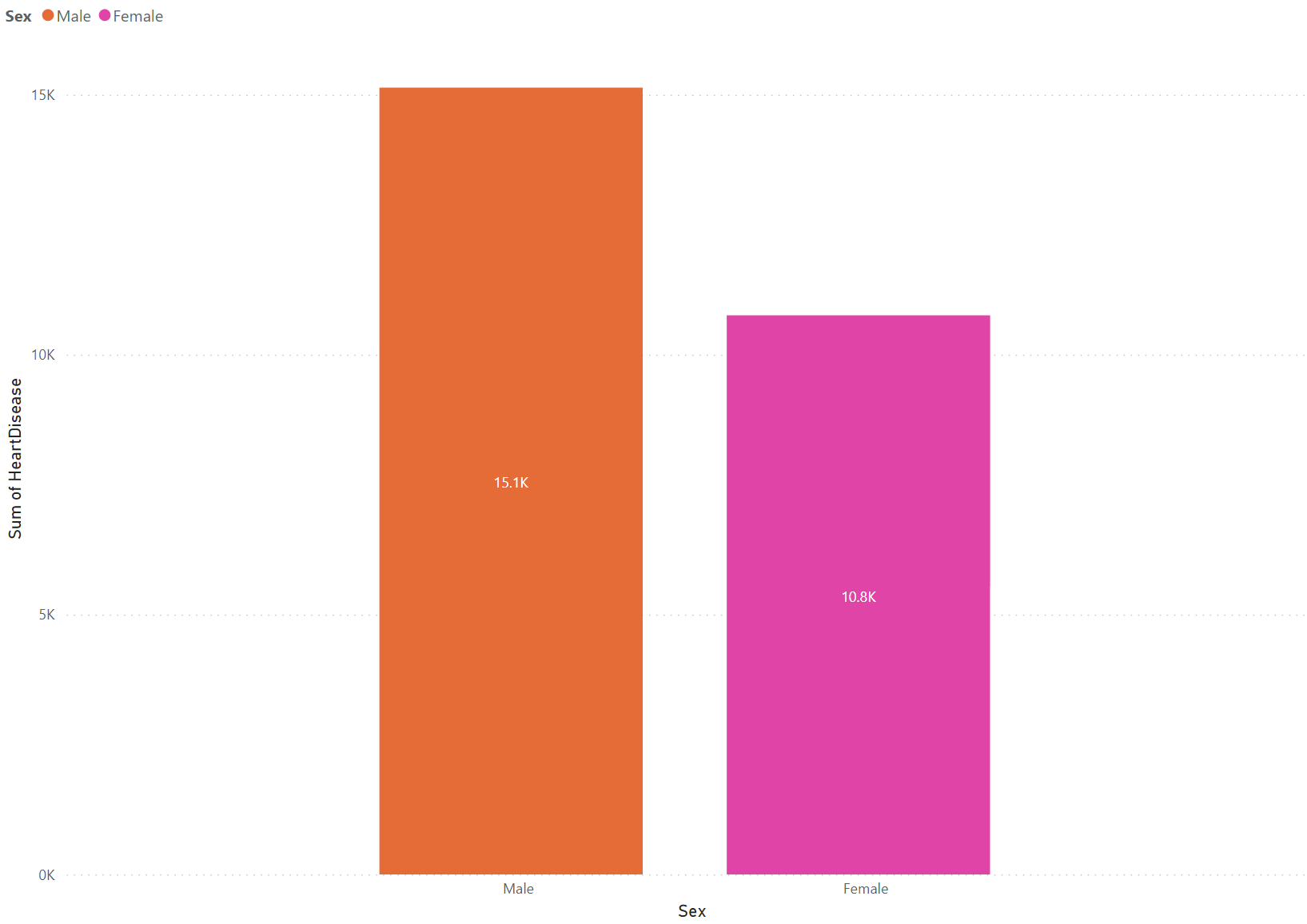
1. **Tổng số ca mắc bệnh tim ở từng độ tuổi**

****

Số ca mắc bệnh tim tăng theo độ tuổi, thấp nhất ở độ tuổi 18-24 và cao nhất ở người hơn 80 tuổi.

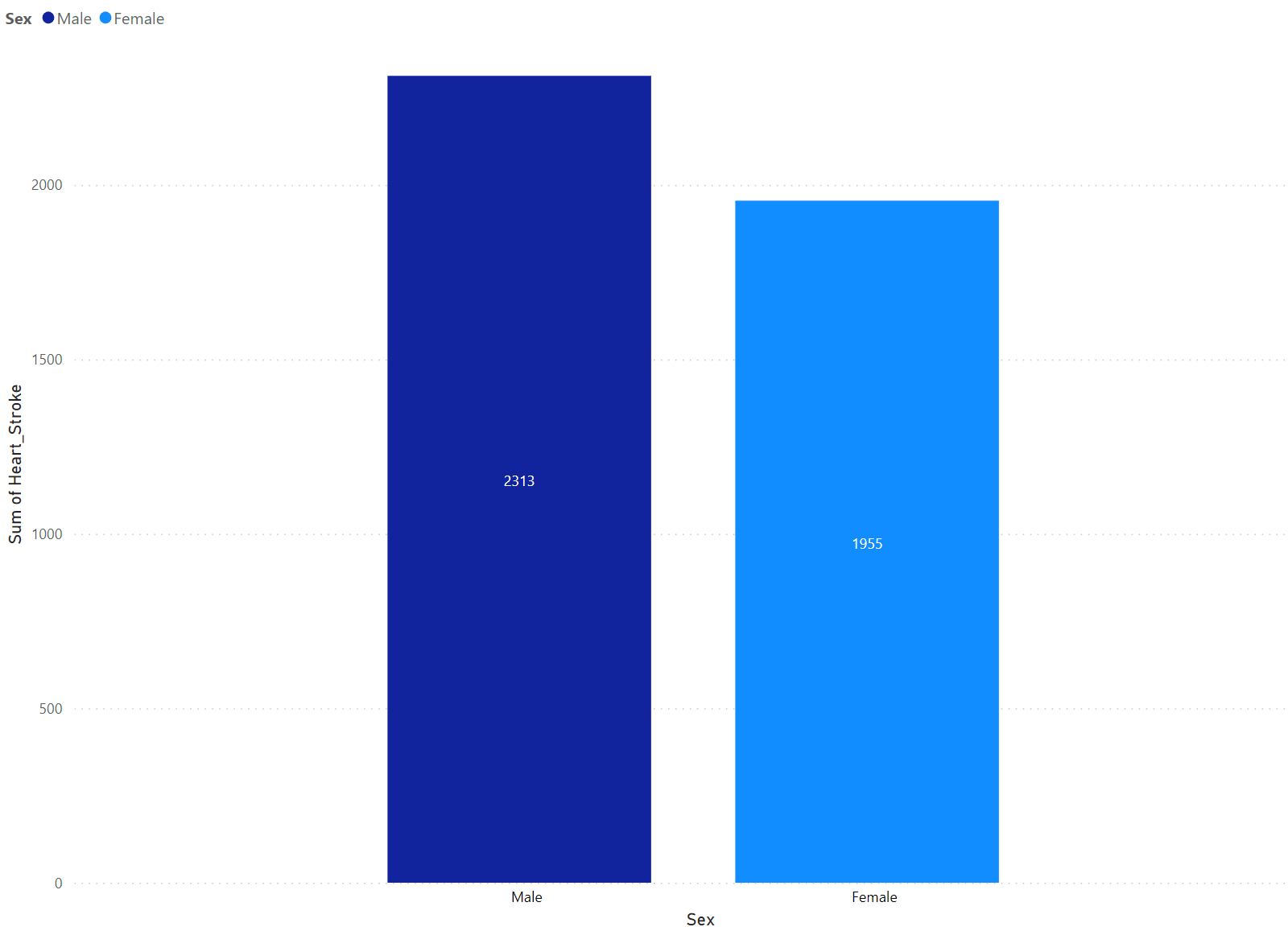
🡪 Có thể thấy tuổi càng cao thì càng có nguy cơ mắc bệnh tim cao hơn.

1. **Tổng số ca mắc bệnh tim ở từng giới tính**

****

Theo tập dữ liệu và đồ thị trực quan, ta thấy rằng số lượng nam giới mắc bệnh tim (15.1 nghìn người) cao hơn số nữ giới mắc bệnh tim (10.8 nghìn người). Tuy nhiên sự chênh lệch này không nhiều nên ta không thể kết luận rằng liệu giới tính có ảnh hưởng đến khả năng mắc bệnh tim không.

1. **Trực quan số người đã từng bị đột quỵ và mắc bệnh tim ở từng giới tính**

****

Từ đồ thị có thể thấy số lượng nam giới từng đột quỵ bị mắc bệnh tim (2313 người) cao hơn so với số lượng nữ giới từng đột quỵ bị mắc bệnh tim (1955 người). Tuy nhiên, số lượng chênh lệch cũng không nhiều nên ta không thể đưa ra kết luận gì về mối liên hệ giữa giới tính và những người từng bị đột quỵ bị mắc bệnh tim.