TRƯỜNG ĐẠI HỌC KINH TẾ - ĐẠI HỌC ĐÀ NẴNG
KHOA THỐNG KÊ - TIN HỌC

BÁO CÁO BÀI TẬP NHÓM

ĐỀ TÀI:

ỨNG DỤNG PHƯƠNG PHÁP DECISION TREE ĐỂ KHAI PHÁ DỮ LIỆU TỈ LỆ LÀM TRÁI NGÀNH CỦA SINH VIÊN ĐẠI HỌC KINH TẾ ĐÀ NẴNG

Học phần	**:**	Kho và khai phá dữ liệu
GVHD	**:**	Nguyễn Văn Chức
Lớp	**:**	46K21.3
Nhóm	**:**	5
Sinh viên thực hiện	**:**	Nguyễn Thế Quân
		Nguyễn Ngọc Đức
		Phan Mai Tuệ Nhi
		Ngô Thị Kim Phương
		Lê Anh Thư

Đà Nẵng, 4/2023

MỤC LỤC

DANH MỤC CÁC TỪ VIẾT TẮT

SSDT:	SQL Server Data Tool
BI:	Business Intelligence
OLAP:	On-line analytical processing
CSDL:	Knowledge discovery in databases

DANH MỤC HÌNH ẢNH VÀ BẢNG BIỂU

PHẦN TRĂM ĐÓNG GÓP

Thành viên	Mã sinh viên	Mức độ hoàn thành
Nguyễn Thế Quân (Leader)	201121521336	20%
Nguyễn Ngọc Đức	201121521304	20%
Phan Mai Tuệ Nhi	201121521326	20%
Ngô Thị Kim Phương	201121521335	20%
Lê Anh Thư	201121521349	20%

1. TỔNG QUAN

1.1. Lý do hình thành đề tài

Hiện nay, bên cạnh vấn đề tuyển sinh đầu vào, số lượng - chất lượng đầu ra cùng cơ hội việc làm và lựa chọn ngành nghề của các sinh viên sau khi tốt nghiệp luôn là mối quan tâm hàng đầu của các trường đại học nói riêng và toàn xã hội nói chung. Phân tích các dữ liệu về tỉ lệ sinh viên làm trái ngành sau khi ra trường, đưa ra những dự đoán về cơ hội việc làm để từ đó có những điều chỉnh kịp thời trong quá trình định hướng và tư vấn chọn nghề và ngành học là mục tiêu quan trọng của mỗi trường đại học. Điều này hoàn toàn khả thi, nếu các trường có thể tận dụng được nguồn dữ liệu lớn của sinh viên, và áp dụng các kỹ thuật khai phá dữ liệu một cách phù hợp.

1.2. Mục tiêu của đồ án

Hiểu rõ hơn về xu hướng và tình hình việc làm của sinh viên sau khi tốt nghiệp. Việc khai phá dữ liệu này sẽ cung cấp thông tin về tỉ lệ sinh viên có việc làm trong ngành đào tạo của mình và tỉ lệ sinh viên chuyển sang ngành khác. Thông tin này có thể giúp các trường đại học cải thiện chương trình đào tạo và hỗ trợ sinh viên trong việc tìm kiếm việc làm phù hợp với năng lực và sở thích của họ. Ngoài ra, cũng có thể hỗ trợ các doanh nghiệp và nhà tuyển dụng trong việc định hướng tuyển dụng và phát triển nhân sự.

1.3. Dự kiến kết quả đạt được

- Hiểu rõ hơn về xu hướng và tình hình việc làm của sinh viên sau khi tốt nghiệp, từ đó có thể đưa ra các giải pháp phù hợp để cải thiện chất lượng đào tạo và giúp sinh viên có việc làm tốt hơn.
- Phát hiện ra những ngành đào tạo có tỉ lệ sinh viên có việc làm cao và những ngành có tỉ lệ thấp, từ đó có thể tập trung đầu tư vào các ngành có tiềm năng phát triển để giúp sinh viên có nhiều cơ hội việc làm hơn.
- Cung cấp thông tin hữu ích cho các doanh nghiệp và nhà tuyển dụng trong việc định hướng tuyển dụng và phát triển nhân sự.
- Giúp các trường đại học nắm bắt được nhu cầu thị trường và phù hợp hóa chương trình đào tạo với nhu cầu thực tế của doanh nghiệp.
- Thông qua việc phân tích và khai thác dữ liệu, có thể đưa ra các giải pháp hỗ trợ cho sinh viên sau khi tốt nghiệp, giúp họ có cơ hội tốt hơn để tìm kiếm việc làm và phát triển sự nghiệp.

2. CƠ SỞ LÝ THUYẾT

2.1. Giới thiệu về khai phá dữ liệu

2.1.1 Khái niệm

Khai thác dữ liệu (Data Mining) là quá trình tìm kiếm và khám phá các mẫu, thông tin và kiến thức mới bằng cách sử dụng các công cụ, kỹ thuật và phương pháp tính toán để phân tích dữ liệu từ các nguồn khác nhau. Quá trình này giúp đưa ra các dự đoán và quyết định thông minh dựa trên việc phân tích các dữ liệu lớn.

2.1.2 Ứng dụng thực tiễn

Khai thác dữ liệu có rất nhiều ứng dụng trong cuộc sống, ví dụ như trong lĩnh vực kinh doanh, y tế, khoa học và công nghệ, giáo dục, chính trị và xã hội. Các công cụ khai thác dữ liệu được sử dụng để tìm kiếm các mẫu và thông tin chưa biết từ các bộ dữ liệu lớn, giúp cho các tổ chức và cá nhân có thể đưa ra các quyết định thông minh và hiệu quả hơn.

2.1.3 Quy trình khai phá dữ liệu

Quy trình khai phá dữ liệu (data mining process) là một chuỗi các bước tạo ra thông tin hữu ích từ dữ liệu. Các bước này được thực hiện để xây dựng mô hình dữ liệu hoặc tìm ra quy luật và mối quan hệ giữa các đặc trưng của dữ liệu. Thông qua việc sử dụng các kỹ thuật phân tích dữ liệu, khai phá dữ liệu có thể giúp cho các tổ chức thu thập thông tin từ dữ liệu, tăng cường hiệu quả hoạt động, thu hẹp khoảng cách giữa doanh nghiệp và khách hàng.

Quy trình khai phá dữ liệu thường được chia thành các bước chính sau:

- B1: Xác định mục tiêu: Đặt ra câu hỏi cần được giải đáp hoặc mục tiêu cần đạt được.
- B2: Thu thập dữ liệu: Tìm kiếm và thu thập dữ liệu liên quan đến mục tiêu đã đề ra.
- B3: Tiền xử lý dữ liệu: Làm sạch, chuẩn hóa và biến đổi dữ liệu để chuẩn bị cho bước phân tích.
- B4: Phân tích dữ liệu: Sử dụng các kỹ thuật khai phá dữ liệu như phân tích ngữ nghĩa, phân tích tương quan, phân tích nhóm để tìm ra các mẫu, đặc trưng và quy luật trong dữ liệu.
- B5: Đánh giá và lựa chọn mô hình: Đánh giá các kết quả khai phá dữ liệu để xác định những mô hình phù hợp nhất với mục tiêu của bạn.
- B6: Triển khai và sử dụng mô hình: Áp dụng mô hình đã được lựa chọn để phát triển sản phẩm hoặc dịch vụ mới cho công ty.
- B7: Đánh giá hiệu quả: Đánh giá hiệu quả của quy trình khai phá dữ liệu bằng cách so sánh kết quả với mục tiêu ban đầu đã đề ra.

2.2 Kho dữ liệu

Kho dữ liệu (data warehouse) là một hệ thống lưu trữ dữ liệu được thiết kế để hỗ trợ việc quản lý và phân tích các thông tin liên quan đến hoạt động của một tổ chức. Kho dữ liệu được xây dựng dựa trên các nguồn dữ liệu khác nhau như các hệ thống giao dịch, các ứng dụng và các nguồn dữ liệu bên ngoài để tạo ra một kho dữ liệu toàn diện cho tổ chức.

Kho dữ liệu giúp tổ chức tổng hợp và lưu trữ các dữ liệu từ các nguồn khác nhau một cách hiệu quả, giúp các bộ phận khác nhau của tổ chức có thể truy cập và sử dụng chúng một cách dễ dàng. Nó cho phép tổ chức tập trung vào các vấn đề chiến lược và kinh doanh cốt lõi của mình, giúp các quyết định được đưa ra nhanh chóng và chính xác hơn.

Kho dữ liệu được thiết kế để hỗ trợ việc phân tích dữ liệu và đưa ra các báo cáo, dự đoán và kịch bản khác nhau để giúp tổ chức ra quyết định dựa trên cơ sở số liệu và

thông tin. Nó là một công cụ quan trọng để phân tích dữ liệu và đưa ra quyết định chiến lược trong các tổ chức lớn và phức tạp.

Dữ liệu thường được cập nhật theo đợt, không phải ngay tức khắc lúc một giao dịch xảy ra trong hệ thống nguồn.

2.2.1 Kiến trúc luồng dữ liệu

Kiến trúc luồng dữ liệu là một phương pháp thiết kế kiến trúc hệ thống dựa trên việc phân tích và mô tả luồng dữ liệu trong hệ thống. Kiến trúc luồng dữ liệu tập trung vào việc định nghĩa và quản lý các luồng dữ liệu chính trong hệ thống, từ nguồn dữ liệu đến đích dữ liệu, đồng thời xác định các thành phần hệ thống và các hoạt động xử lý dữ liệu để đảm bảo luồng dữ liệu hoạt động một cách hiệu quả.

Kiến trúc luồng dữ liệu thường được sử dụng trong các hệ thống xử lý dữ liệu lớn và phức tạp, giúp đảm bảo tính linh hoạt và mở rộng của hệ thống. Nó cung cấp một khung nhìn tổng quan về luồng dữ liệu trong hệ thống, giúp người thiết kế hiểu rõ hơn về các thành phần và các quy trình xử lý dữ liệu trong hệ thống.

Trong kiến trúc luồng dữ liệu, các thành phần của hệ thống được chia thành các phân đoạn, mỗi phân đoạn đại diện cho một bước xử lý dữ liệu trong hệ thống. Các phân đoạn này được kết nối với nhau để tạo thành các luồng dữ liệu chính trong hệ thống.

Kiến trúc luồng dữ liệu còn giúp tách rời các phần của hệ thống, đảm bảo tính độc lập và khả năng tái sử dụng. Nó cũng giúp tối ưu hóa hiệu suất hệ thống bằng cách tối ưu hóa các luồng dữ liệu và các bước xử lý dữ liệu trong hệ thống.

2.2.2 Kho dữ liệu và khai phá dữ liệu trong BI

Có ba loại ứng dụng kho dữ liệu:

- Xử lý thông tin: hỗ trợ truy vấn, phân tích thống kê cơ bản và báo cáo sử dụng crosstab, bảng, biểu đồ hoặc đồ thị.
- Xử lý phân tích: phân tích số liệu dữ liệu kho dữ liệu theo chiều sâu. Nó thường hoạt động trên dữ liệu lịch sử trong cả hai dạng tóm tắt và chi tiết.
- Khai thác dữ liệu: hỗ trợ khám phá kiến thức bằng cách tìm kiếm các mẫu ẩn và các hiệp hội, xây dựng các mô hình phân tích, thực hiện phân loại và dự đoán và trình bày các kết quả khai thác bằng các công cụ trực quan hóa.

Từ On-Line Analytical Processing (OLAP) đến OnLine Analytical Mining (OLAM) OLAM hay còn gọi là OLAP mining: tích hợp xử lý phân tích trực tuyến (OLAP) với khai thác dữ liệu và kiến thức về khai phá dữ liệu trong cơ sở dữ liệu đa chiều. Từ On-Line Analytical Processing (OLAP) đến OnLine Analytical Mining (OLAM).

2.3 Phương pháp khai phá dữ liệu

Phương pháp phân lớp là một trong những phương pháp quan trọng trong học máy và khai phá dữ liệu, nó giúp chúng ta phân loại các đối tượng dữ liệu (như email, hình ảnh, tài liệu văn bản, sản phẩm,..) vào các lớp khác nhau dựa trên các đặc trưng của chúng. Các phương pháp phân lớp phổ biến bao gồm:

1. Cây quyết định (Decision tree): Phương pháp này sử dụng một cây để phân loại các đối tượng dữ liệu dựa trên các quyết định tại các nút của cây.

2. Hồi quy logistic (Logistic regression): Phương pháp này sử dụng một hàm logistic để dự đoán xác suất của một đối tượng dữ liệu thuộc về một lớp cụ thể.

3. Máy vector hỗ trợ (Support vector machine): Phương pháp này sử dụng một siêu mặt phẳng để phân loại các đối tượng dữ liệu.

4. Mạng nơ-ron (Neural networks): Phương pháp này sử dụng một mạng nơ-ron để phân loại các đối tượng dữ liệu.

5. Máy học tăng cường (Reinforcement learning): Phương pháp này sử dụng một hệ thống tương tác để học cách phân loại các đối tượng dữ liệu.

6. Phân tích discriminant (Discriminant analysis): Phương pháp này sử dụng một hàm phân tích discriminant để phân loại các đối tượng dữ liệu.

7. Random forest: Phương pháp này sử dụng nhiều cây quyết định để phân loại các đối tượng dữ liệu và kết hợp kết quả từ các cây để đưa ra dự đoán cuối cùng.

Mỗi phương pháp có những ưu điểm và hạn chế riêng, và việc lựa chọn phương pháp phù hợp phụ thuộc vào bộ dữ liệu và mục tiêu của bài toán phân loại.

2.4 Phương pháp nhóm chọn đề tài khai phá dữ liệu: Thuật toán Decision Tree

2.4.1 Giới thiệu Decision Tree:

Cây quyết định (Decision Tree) là một thuật toán học máy phân loại và dự đoán dữ liệu. Nó được sử dụng để đưa ra quyết định dựa trên các quy tắc được xác định từ dữ liệu huấn luyện. Cây quyết định bao gồm một tập hợp các nút biểu diễn các quyết định và các nhánh biểu diễn các kết quả của các quyết định đó.

Thuật toán cây quyết định bắt đầu bằng việc chọn thuộc tính quan trọng nhất để phân chia dữ liệu. Sau đó, nó chia dữ liệu thành các nhánh dựa trên giá trị của thuộc tính đó. Quá trình này được lặp lại cho mỗi nhánh cho đến khi đạt được một điều kiện dừng nhất định. Kết quả là một cây quyết định với các nút biểu diễn các quyết định và các nhánh biểu diễn các kết quả của các quyết định đó.

2.4.2. Ưu điểm và nhược điểm của Decision Tree

- **Ưu điểm:**
  - Dễ hiểu và diễn giải: các quyết định được biểu diễn dưới dạng cây, vì vậy người dùng có thể hiểu và giải thích các kết quả dễ dàng.
  - Tính linh hoạt: có thể được sử dụng cho các bài toán phân loại và dự đoán dữ liệu đa dạng trong nhiều lĩnh vực khác nhau.
  - Không yêu cầu chuẩn bị dữ liệu phức tạp.
  - Khả năng xử lý dữ liệu thiếu.
- **Nhược điểm:**
  - Dễ bị overfitting (quá khớp) nếu không được điều chỉnh phù hợp.
  - Không ổn định.
  - Khả năng chọn thuộc tính không tối ưu.
  - Khó xử lý các vấn đề dữ liệu liên tục.

2.5. Giới thiệu về phần mềm SQL Server Data Tool

SQL Server Data Tools (SSDT) là một phần mềm được Microsoft phát triển để giúp cho các nhà phát triển phần mềm, các chuyên gia quản trị cơ sở dữ liệu, các nhà thiết kế và triển khai cơ sở dữ liệu có thể tạo ra các ứng dụng và giải pháp cơ sở dữ liệu hiệu quả trên nền tảng Microsoft SQL Server.

Phần mềm SSDT có thể tích hợp với Visual Studio và cung cấp cho người dùng các công cụ hữu ích để phát triển cơ sở dữ liệu, như thiết kế cơ sở dữ liệu, truy vấn dữ liệu, quản lý và triển khai cơ sở dữ liệu. SSDT cho phép người dùng tạo ra các cơ sở dữ liệu mới hoặc chỉnh sửa cơ sở dữ liệu hiện có, với khả năng hỗ trợ các tính năng như kiểm tra lỗi, phân tích dữ liệu, so sánh phiên bản cơ sở dữ liệu và xây dựng các báo cáo về cơ sở dữ liệu.

Các tính năng chính của SSDT bao gồm:

- Thiết kế cơ sở dữ liệu: SSDT cho phép người dùng thiết kế các đối tượng cơ sở dữ liệu như bảng, chế độ xem, thủ tục lưu trữ và hàm, sử dụng các công cụ đồ họa hoặc mã nguồn. Các đối tượng này có thể được tạo ra từ đầu hoặc được sao chép từ cơ sở dữ liệu hiện có.
- Quản lý cơ sở dữ liệu: SSDT cho phép người dùng quản lý các đối tượng cơ sở dữ liệu, bao gồm xóa, sửa đổi và thêm mới các đối tượng. SSDT cũng cung cấp các công cụ để so sánh và đồng bộ hóa các phiên bản của cơ sở dữ liệu.
- Truy vấn cơ sở dữ liệu: SSDT cho phép người dùng viết và thực thi các truy vấn SQL trực tiếp trên cơ sở dữ liệu.
- Biên dịch và kiểm tra lỗi: SSDT cung cấp các công cụ để biên dịch và kiểm tra lỗi các đối tượng cơ sở dữ liệu, giúp người dùng phát hiện và sửa các lỗi trước khi triển khai cơ sở dữ liệu.

Tóm lại, SQL Server Data Tools là một phần mềm rất hữu ích cho các chuyên gia quản trị cơ sở dữ liệu, nhà phát triển phần mềm và các nhà thiết kế, giúp họ tạo ra các giải pháp cơ sở dữ liệu và ứng dụng hiệu quả trên nền tảng Microsoft SQL Server.

3. ỨNG DỤNG PHẦN MỀM SQL SERVER DATA TOOL

3.1 Giới thiệu về bộ dữ liệu được sử dụng

Để thực thi việc ứng dụng phần mềm SQL Data Tool vào quá trình khai phá dữ liệu để tìm ra những quy luật, định hướng và tư vấn chọn ngành nghề của sinh viên và thống kê nhóm sinh viên làm trái ngành sau khi ra trường của trường Đại học Kinh tế - Đại học Đà Nẵng. Bộ dữ liệu tập hợp các dữ liệu về sinh viên đã được thu thập trong quá trình điều tra và khảo sát với các thông tin như: Mã nhận dạng, giới tính, ngành học, năm tốt nghiệp, loại tốt nghiệp, công việc, tính phù hợp của công việc so với ngành học và lí do chọn trái ngành. Bộ dữ liệu bao gồm 1000 dòng dữ liệu và 8 cột thông tin.

Dữ liệu mẫu của bảng thông tin được sử dụng:

ID	Gender	Major	YearOfGraduation	TypeOfGraduation	CurrentCareer	CareerSuitability	ReasonWrongBranch
DUE0001	Nam	Hệ Thống Thông Tin Quản Lí	2020	Trung bình - Khá	Quản lý quán cà phê	0	Chưa tìm được công việc đúng chuyên ngành
DUE0002	Nữ	Hệ Thống Thông Tin Quản Lí	2023	Khá	Nhân viên Sales	0	Chưa tìm được công việc đúng chuyên ngành
DUE0003	Nữ	Hệ Thống Thông Tin Quản Lí	2016	Khá	Sales căn hộ	0	Thích công việc hiện tại, Chưa tìm được công việc đúng chuyên ngành, Tìm kiếm thu nhập tốt hơn, Thử
DUE0004	Nữ	Kinh Doanh Quốc Tế	2022	Giỏi	Intern	1	null
DUE0005	Nữ	Kinh Tế	2022	Giỏi	Sinh viên	1	null
DUE0006	Nữ	Kinh Doanh Quốc Tế	2022	Xuất sắc	Marketing Intern	1	null
DUE0007	Nữ	Kinh Doanh Quốc Tế	2021	Giỏi	BA	0	Tìm kiếm thu nhập tốt hơn
DUE0008	Nam	Tài Chính - Ngân Hàng	2020	Xuất sắc	Kinh doanh	0	Thích công việc hiện tại, Tìm kiếm thu nhập tốt hơn, Thử thách bản thân, Yếu tố gia đình
DUE0009	Nữ	Marketing	2021	Khá	Chuyên viên nghiên cứu thị trường	1	null
DUE0010	Nữ	Marketing	2022	Giỏi	Chuyên viên nghiên cứu thị trường	1	null
DUE0011	Nữ	Hệ Thống Thông Tin Quản Lí	2022	Khá	Nhân viên	0	Chưa tìm được công việc đúng chuyên ngành, Chuyên môn chưa đáp ứng yêu cầu tuyển dụng
DUE0012	Nam	Kinh Tế	2020	Giỏi	Kiểm toán viên	1	null
DUE0013	Nữ	Ngôn ngữ Anh	2018	Khá	Kinh doanh online	0	Thích công việc hiện tại
DUE0014	Nữ	Kế Toán	2021	Giỏi	Văn phòng	0	Chưa tìm được công việc đúng chuyên ngành, Tìm kiếm thu nhập tốt hơn
DUE0015	Nữ	Hệ Thống Thông Tin Quản Lí	2019	Khá	Tester	1	null
DUE0016	Nữ	Hệ Thống Thông Tin Quản Lí	2021	Khá	Tester	1	null
DUE0017	Nữ	Quản Trị Kinh Doanh	2022	Giỏi	Lễ tân tại một nhà hàng	1	null
DUE0018	Nam	Kế Toán	2018	Xuất sắc	Công ty	0	Thích công việc hiện tại, Chưa tìm được công việc đúng chuyên ngành
DUE0019	Nữ	Hệ Thống Thông Tin Quản Lí	2016	Giỏi	BA	1	null
DUE0020	Khác	Hệ Thống Thông Tin Quản Lí	2020	Xuất sắc	developer	0	Tìm kiếm thu nhập tốt hơn, Chuyên môn chưa đáp ứng yêu cầu tuyển dụng
DUE0021	Nam	Thương Mại Điện Tử	2022	Giỏi	Marketing	1	null
DUE0022	Nam	Quản Trị Dịch Vụ Du Lịch và Lữ Hành	2018	Khá	quản lý khách sạn	1	null
DUE0023	Nam	Quản Trị Kinh Doanh	2018	Khá	điện lực	0	Thích công việc hiện tại, Chưa tìm được công việc đúng chuyên ngành, Tìm kiếm thu nhập tốt hơn
DUE0024	Nam	Hệ Thống Thông Tin Quản Lí	2022	Khá	Kinh doanh	0	Tìm kiếm thu nhập tốt hơn
DUE0025	Nam	Luật	2016	Xuất sắc	Chuyên viên	1	null
DUE0026	Nam	Hệ Thống Thông Tin Quản Lí	2022	Khá	Sales	0	Chuyên môn chưa đáp ứng yêu cầu tuyển dụng
DUE0027	Nữ	Quản Trị Khách Sạn	2023	Giỏi	Kinh doanh	1	null

Hình 1. Dữ liệu khảo sát từ thông tin trái ngành

3.2 Giai đoạn tiền xử lý dữ liệu

Do bảng dữ liệu mẫu này, có nhiều các cột không có giá trị, nên nhóm tiến hành bỏ Null vào những cột đó và những cột có biến phân loại thì sẽ bỏ số Yes, No để bộ dữ liệu về nghề nghiệp sau khi ra trường trở nên chính xác hơn trong quá trình khởi chạy các thuật toán Decision Tree. Với các giá trị được giữ lại sau khi sàng lọc lược bỏ và bảng mô tả dữ liệu công việc sau khi ra trường như sau:

STT	Tên dữ liệu	Tên tiếng Việt	Kiểu dữ liệu
1	*ID*	Mã nhận dạng	Int
2	*Gender*	Giới tính	Nvarchar
3	*Major*	Ngành học	Nvarchar
4	*YearOfGraduation*	Năm tốt nghiệp	Datetime
5	*TypeOfGraduation*	Loại tốt nghiệp	Nvarchar
6	*CurrentCareer*	Công việc hiện tại	Nvarchar
7	*CareerSuitability*	Tính phù hợp công việc	Nvarchar
8	*ReasonWrongBranch*	Lý do chọn trái ngành	Nvarchar

Hình 2. Bảng mô tả dữ liệu thông tin sinh viên làm trái ngành

3.3 Triển khai thuật toán cây quyết định (Decision Tree Algorithm)

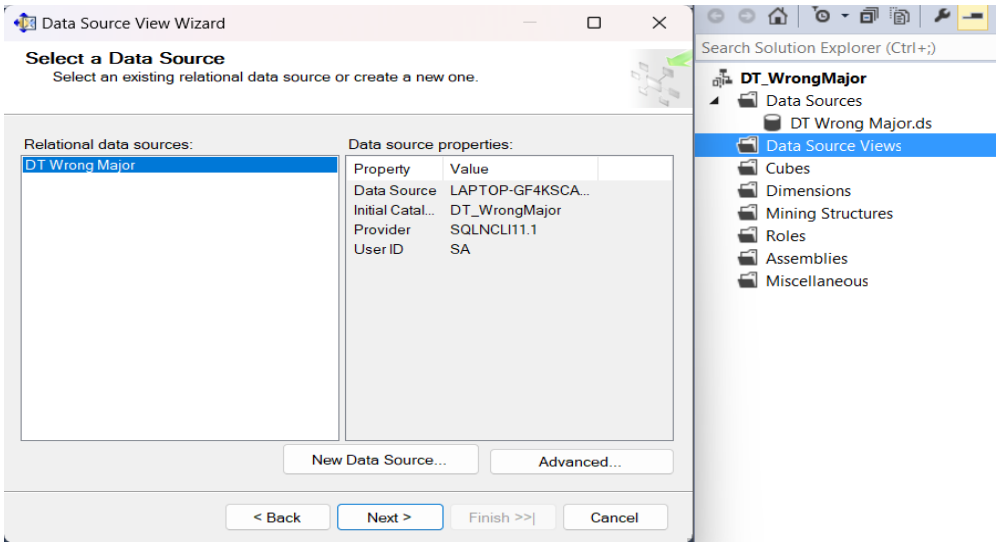
3.3.1 Khởi chạy thuật toán cây ra quyết định trên SQL Server Data Tool

Để tiến hành khởi chạy thuật toán, SSDT bắt buộc chúng ta phải khởi tạo khung nhìn dữ liệu nguồn. Chúng ta có thể dựa vào khung nhìn dữ liệu nguồn (Data

Source View) để tại nên các cấu trúc khai phá dữ liệu hoặc có thể thêm các cột vào bảng, tổng hợp, tính toán các thông tin bên trong bộ dữ liệu. Bằng cách sử dụng chế độ khung nhìn dữ liệu nguồn, chúng ta có thể lựa chọn được cái dữ liệu mà liên quan đến thuật toán mình sẽ sử dụng và hơn thế nữa như là thay đổi cấu trúc bảng, chỉnh sửa các mối quan hệ giữa các bảng mà không bị thay đổi cấu trúc ban đầu của CSDL.

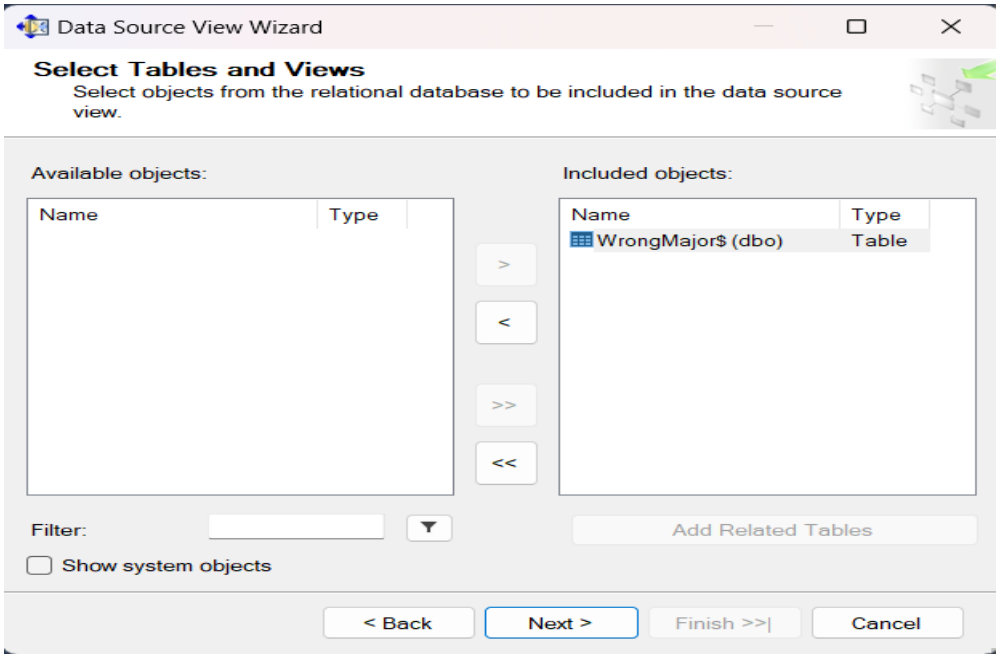
Các thao tác tạo khung nhìn dữ liệu nguồn

B1. Tại khung cửa sổ Solution Explorer, phải chuột vào Data Source Views, và chọn New Data Source View. Sau đó hiện lên cửa sổ Data Source View Wizard, chọn Data Source đã tạo như Hình 3, tiếp theo chọn Next



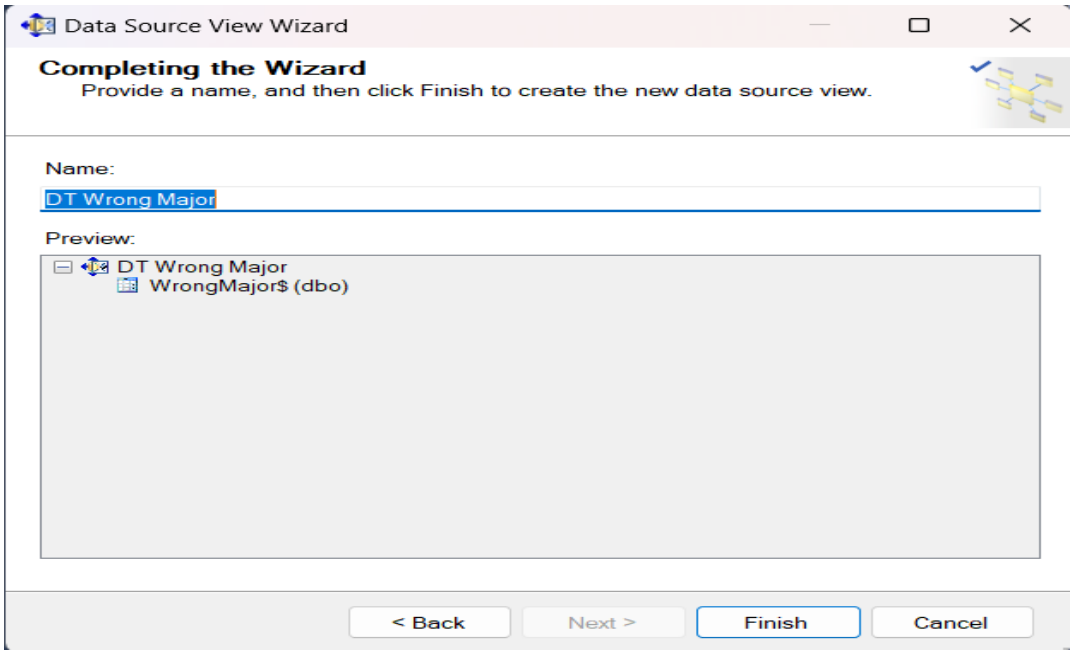
Hình 3. Thao tác khởi tạo Data Source View

B2. Sau đó hiện lên hộp thoại Select Tables and Views, tại đây chúng ta có thể lựa chọn các đối tượng – các thực thể, các khung nhìn và chọn vào nút " > " để di chuyển một đối tượng qua khung Included Objects, chọn " >> ". Ngoài ra, chúng ta còn có thể tìm kiếm tên các đối tượng tại ô tìm kiếm Filter để tìm kiếm nhanh các đối tượng mà ta mong muốn. Và chọn Next



Hình 4. Tìm kiếm và lựa chọn các đối tượng cần thiết cho quá trình chạy thuật toán

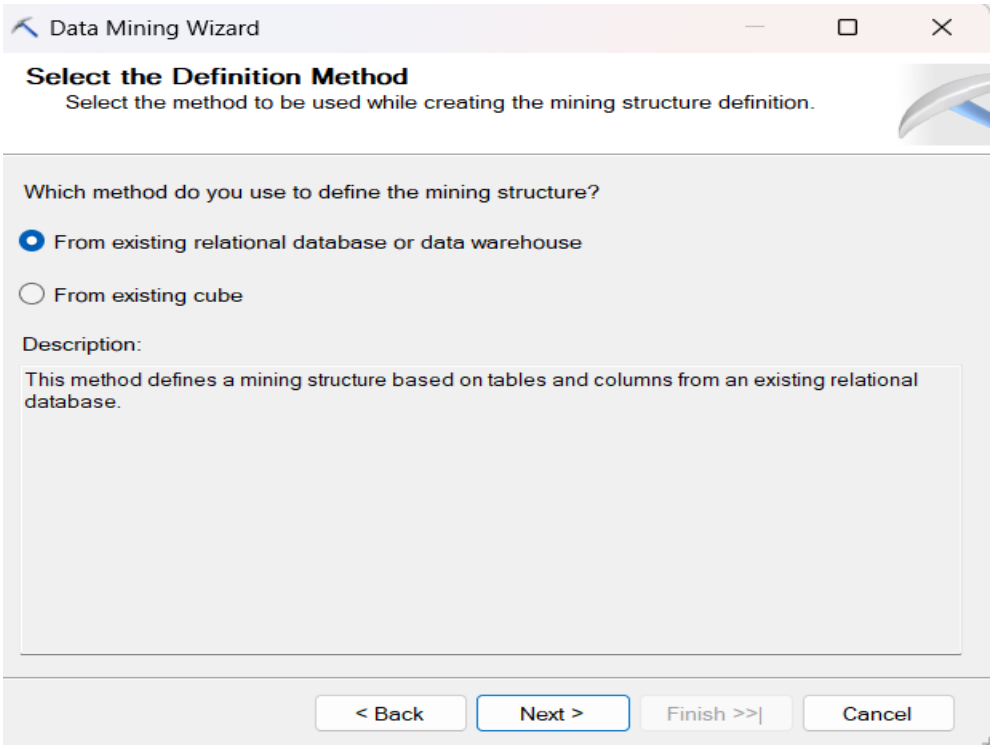
B3. Sau khi đã lựa chọn xong các đối tượng cần thiết, sẽ hiển thị lên hộp thoại Completing the Wizard, tại đây chúng ta sẽ thay đổi tên của Data Source View và chọn Finish



Hình 5. Thay đổi tên và hoàn thành thao tác tạo Data Source View

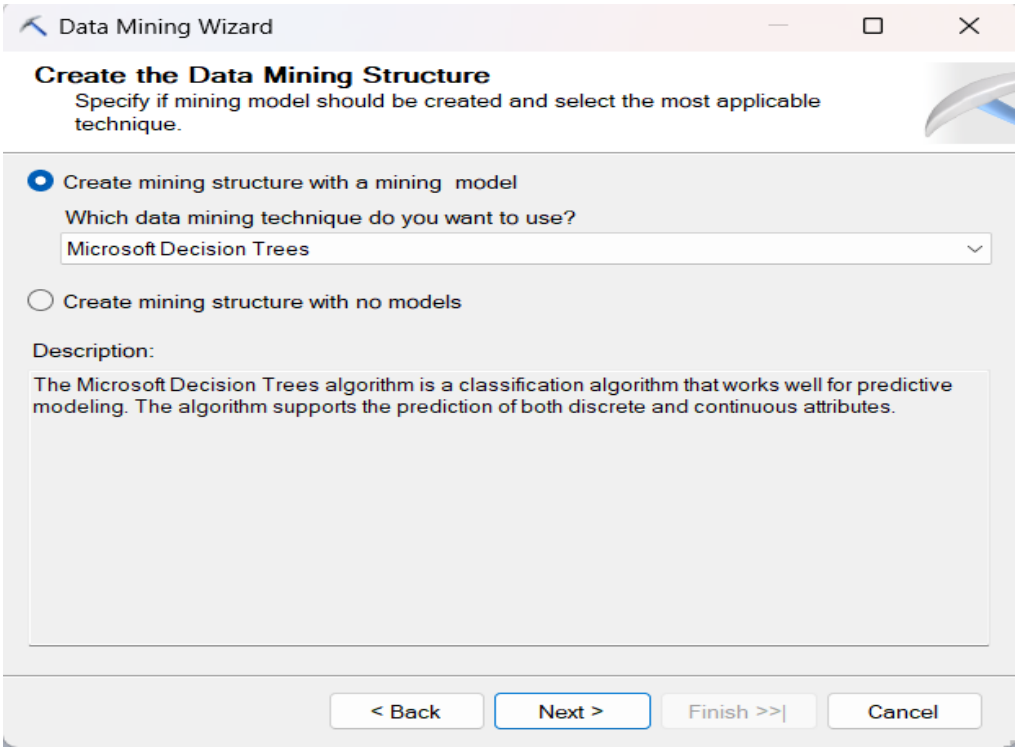
Các thao tác tạo cấu trúc khai phá

B1. Tại Solution Explorer, chọn Mining Structures, phải chuột chọn New Mining Structure, hộp thoại Select the Definition Method hiện lên và chọn tiếp "From existing relational database or data warehouse", và chọn Next



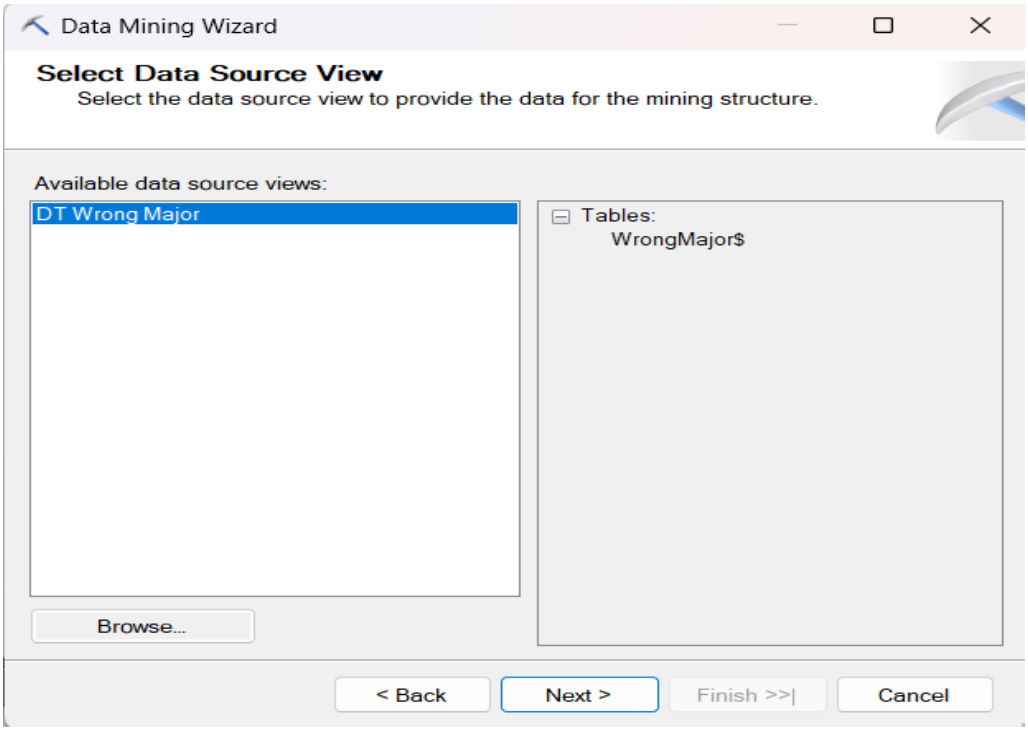
Hình 6. Hộp thoại lựa chọn phương thức để khai báo cấu trúc

B2. Hộp thoại Create the Data Mining Structure, tại đây chúng ta lựa chọn phương thức khai phá dữ liệu ở bên dưới. Để chạy thuật toán cây ra quyết định này, chúng em lựa chọn Microsoft Decision Trees và chọn Next



Hình 7. Lựa chọn thuật toán cần được sử dụng

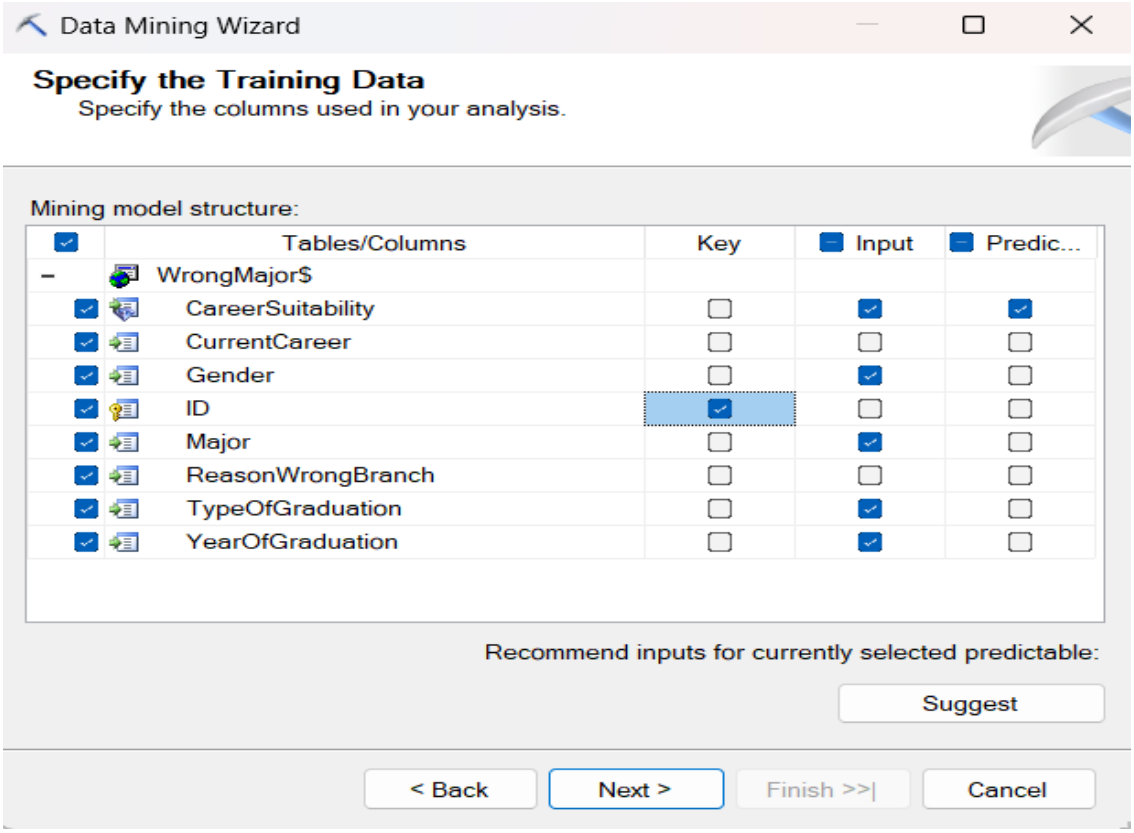
B3. Hộp thoại Select Data Source View hiện lên, tại đây chúng ta chọn Data Source View mà chúng ta sẽ chọn Data Source View đa tạo từ trước. Sau đó chọn Next



Hình 8. Lựa chọn Data Source View cần cho quá trình khai phá

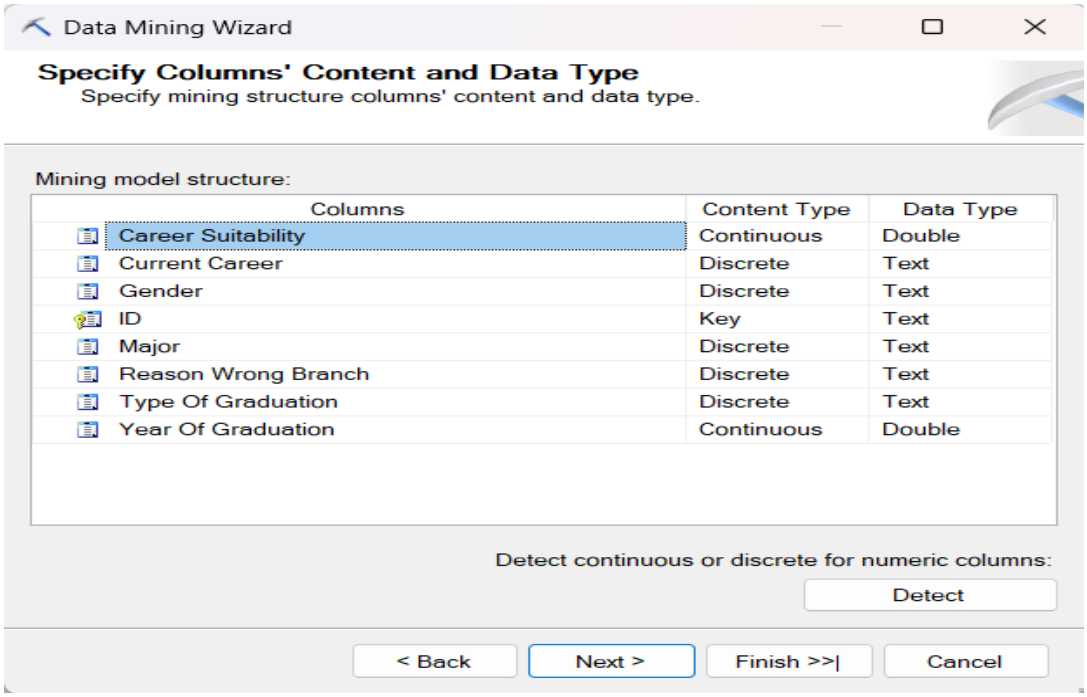
B4. Tại bước này chúng ta, lựa chọn thuộc tính Case tại **MajorWrong**, và chọn Next

B5. Hiện lên hộp thoại Specify the Training Data và lựa chọn thuộc tính khóa (Key) là ID, và thuộc tính dự đoán (Predictable) là CareerSuitablity. Sau đó truyền vào các dữ liệu (Input): ***Gender (Giới tính), Major (Ngành học), YearOfGraduation(Năm tốt nghiệp), TypeOfGraduation(Loại tốt nghiệp), CurrentCareer(Công việc hiện tại), CareerSuitability(Tính phù hợp công việc), ReasonWrongBranch(Lý do chọn trái ngành)***



Hình 9. Hộp thoại truyền vào các thuộc tính trên Data Source View

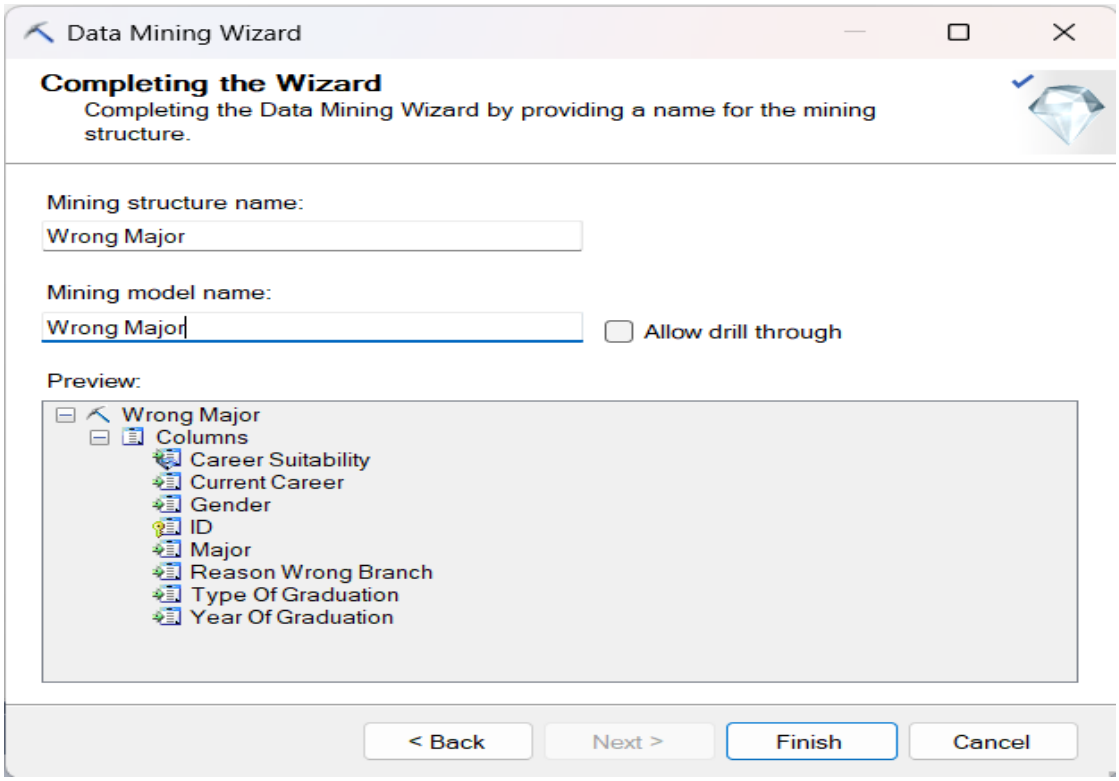
B6. Tiếp theo, hộp thoại lựa chọn các giá trị các cột cùng với các loại dữ liệu, tại đây chúng ta có thể lựa chọn các kiểu dữ liệu – mà có thể chạy ra thuật toán. Hoặc có thể chọn Detect để hệ thống tự động chỉnh. Và chọn Next để qua bước tiếp theo



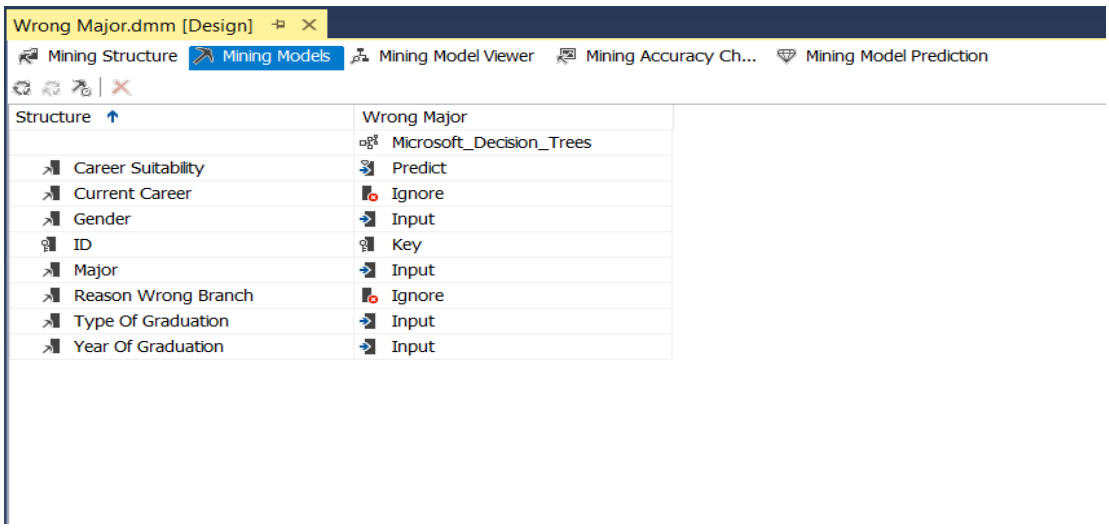
Hình 10. Hộp thoại chỉnh sửa các thuộc tính của CSDL

B7. Sau đó lựa chọn phần trăm dữ liệu để thực nghiệm (Percentage of data for testing) = 0% và số lượng lớn nhất các trường hợp để thực nghiệm trong bộ dữ liệu (Maximum Number of cases in testing data set) = 1000, sau đó chọn Next

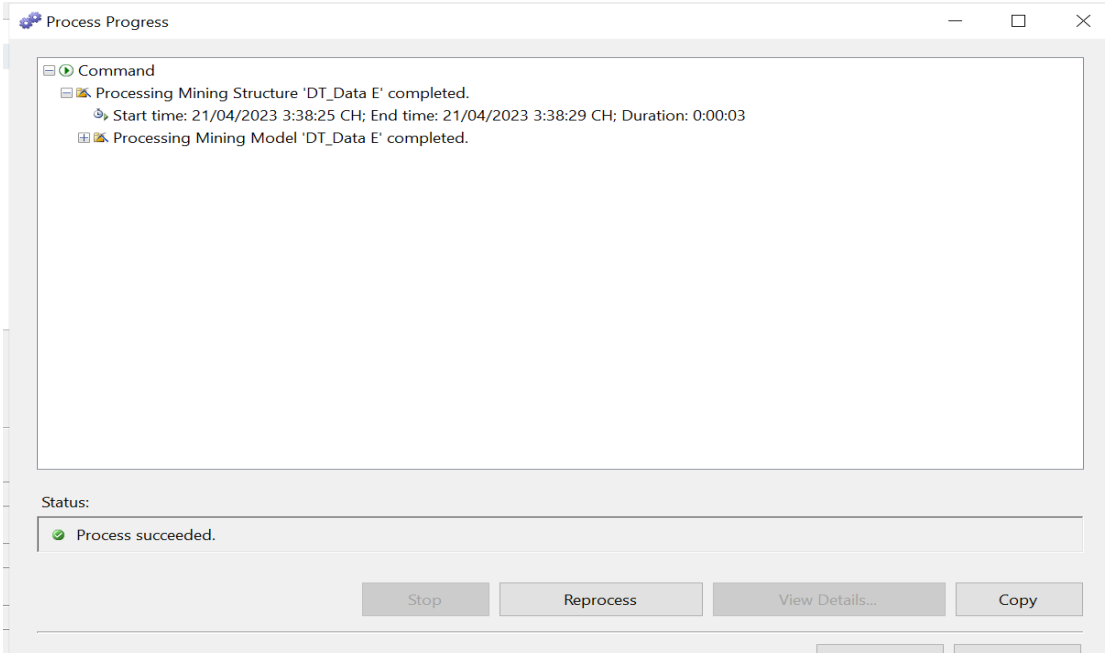
B8. Sau khi nhấn Next ở B7. Tiếp theo sẽ hiện lên hộp thoại Completing the Wizard tại đây bao gồm 2 khung: Mining Structure Name và Mining Model Name. Chúng ta sẽ điền tên của từng thuộc tính và chọn vào ô Allow drill through để có thể xuyên qua các cấu trúc khác để lấy thêm dữ liệu khi cần thiết, sau đó chọn Finish. Ngay sau đó chúng ta đã có một Mining Structure với tên là chúng ta đã đặt ở thẻ Mining Structure trong Solution Explorer



Hình 11. Hộp thoại đặt tên cấu trúc khai phá và tên của mẫu khai phá



Hình 12. Quá trình lựa chọn và khởi chạy cấu trúc khai phá (1)



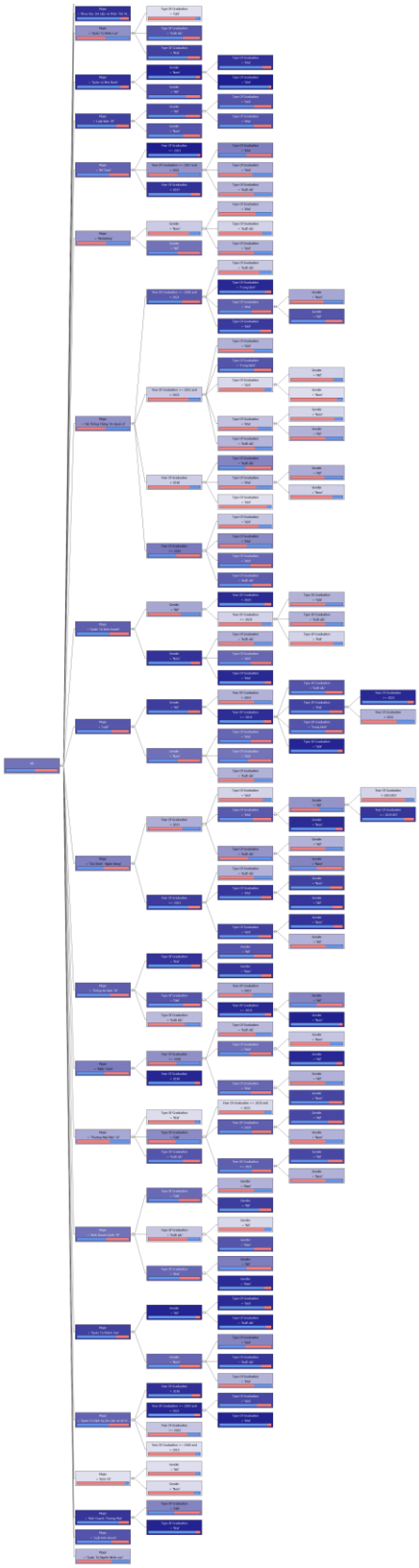
Hình 13. Quá trình lựa chọn và khởi chạy cấu trúc khai phá (2)

3.3.2 Kết quả của thuật toán cây ra quyết định

Sau khi đã thực hiện các thao tác khởi chạy trên thuật toán, tại đây chúng ta sẽ được xem kết quả của quá trình khởi chạy thuật toán. Bằng việc chọn vào thẻ Mining Model Viewer Với các thuộc tính:

- Tree: CareerSuitability

- Background: Yes (Làm đúng ngành) - No (Làm trái ngành)



Hình 14. Cây ra quyết định sau khi khởi chạy với thuộc tính là CareerSuitability

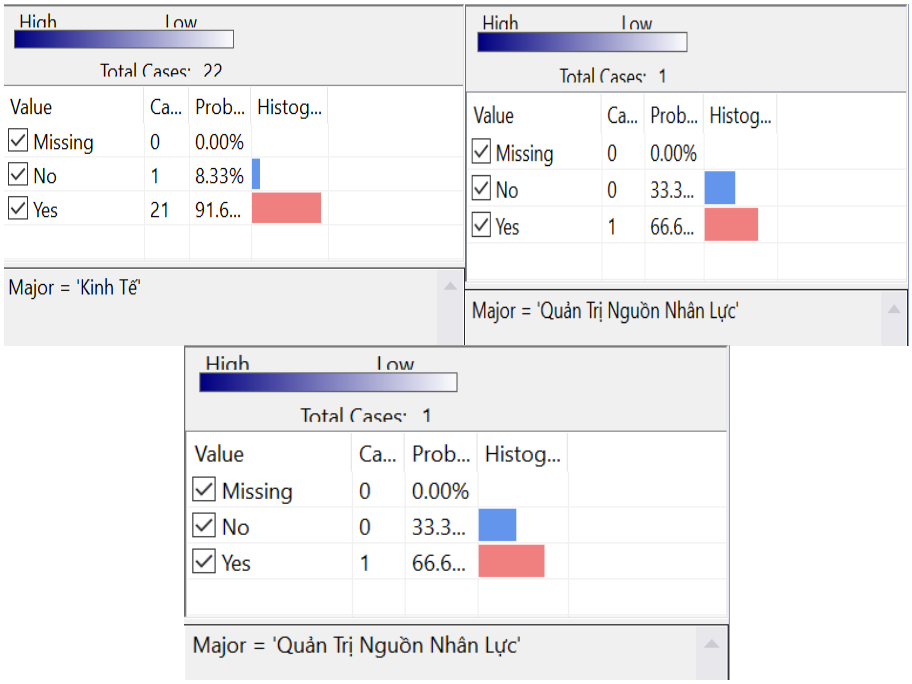
Số lượng sinh viên làm trái ngành

Số lượng sinh viên làm đúng ngành

Với gam màu xanh dương từ nhạt tới đậm biểu thị tỷ lệ sinh viên ra trường làm việc đúng ngành – trái ngành trong cây. Màu xanh càng đậm có nghĩa là tỉ trọng sinh viên làm việc trái ngành càng cao và thanh màu xanh nhạt trong từng các lá biểu hiện cho phần trăm các sinh viên làm trái ngành và màu hồng ứng với số sinh viên làm việc đúng ngành.

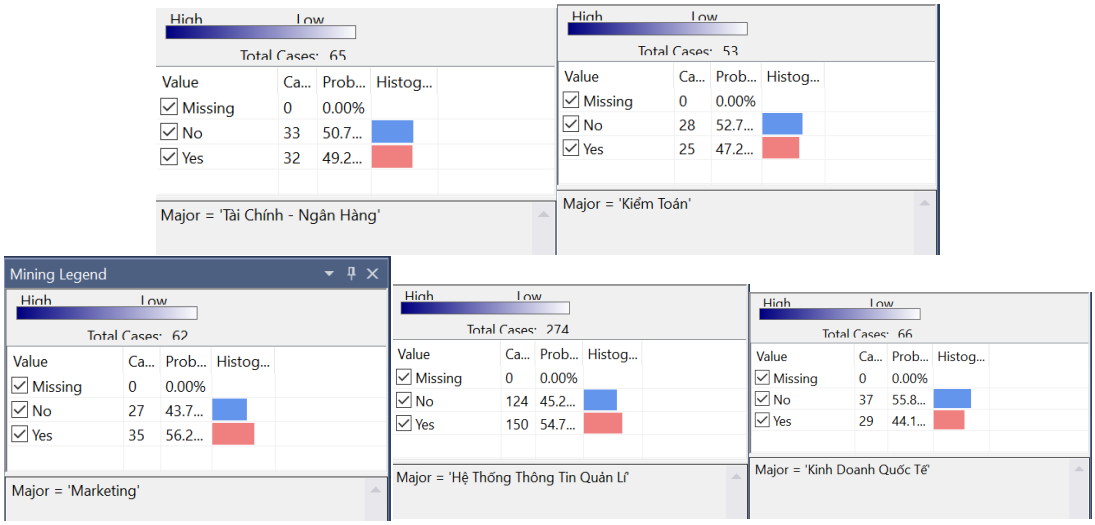
Tại cây ra quyết định này, chúng ta có thể thấy được một số ngành học có tỉ lệ sinh viên ra trường làm việc đúng ngành hay trái ngành cao, nhóm chúng em chia thành 3 nhóm:

- **Nhóm 1- Đúng ngành:** Kinh tế, Thương mại điện tử, Quản trị nguồn nhân lực



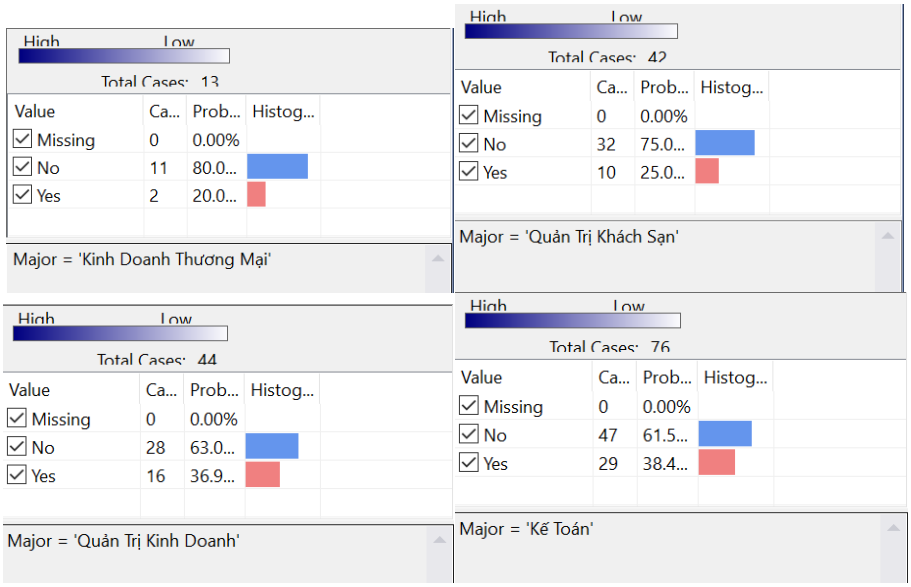
Hình 15. Tỉ trọng làm đúng ngành của sinh viên

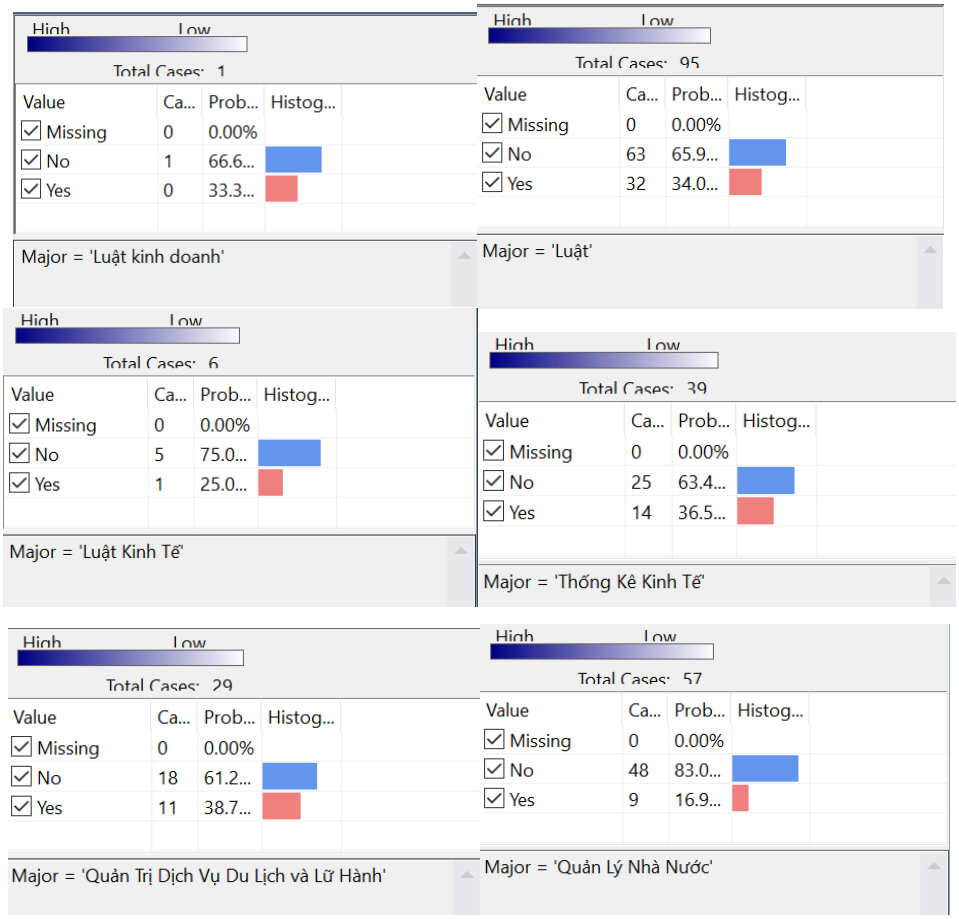
- **Nhóm 2 - Đúng ngành Trái ngành cân bằng:** Tài chính – Ngân hàng, Kiểm toán, Marketing, Hệ thống thông tin quản lý, Kinh doanh quốc tế.



Hình 16. Tỉ trọng đúng ngành và trái ngành cân bằng

- **Nhóm 3 -Trái ngành:** Kinh doanh thương mại, Quản trị khách sạn, Quản trị kinh doanh, Kế toán, Luật Kinh tế, Luật, Luật kinh doanh, Thống kê kinh tế, Quản trị Dịch vụ Du lịch và Lữ hành, Quản lý Nhà Nước.



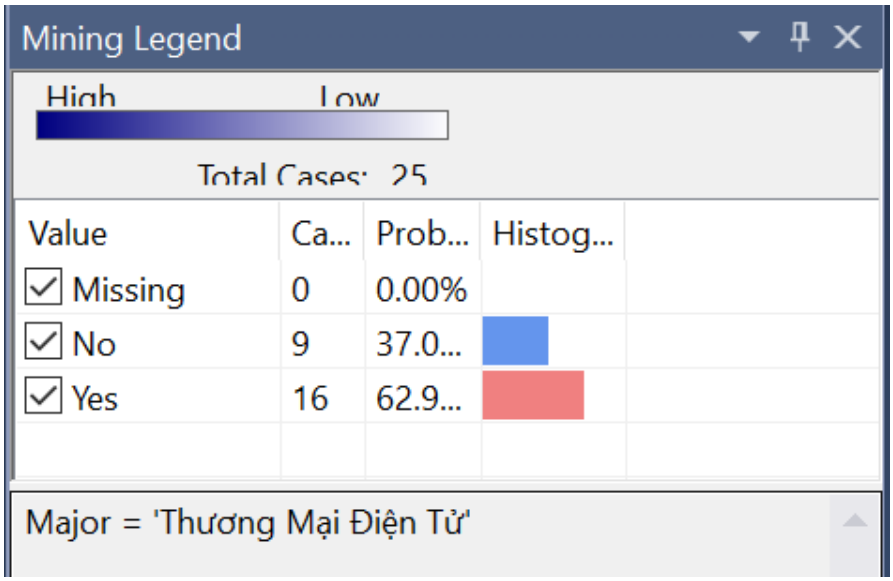


Hình 17. Tỉ trọng làm trái ngành

3.3.3 Phân tích kết quả của thuật toán

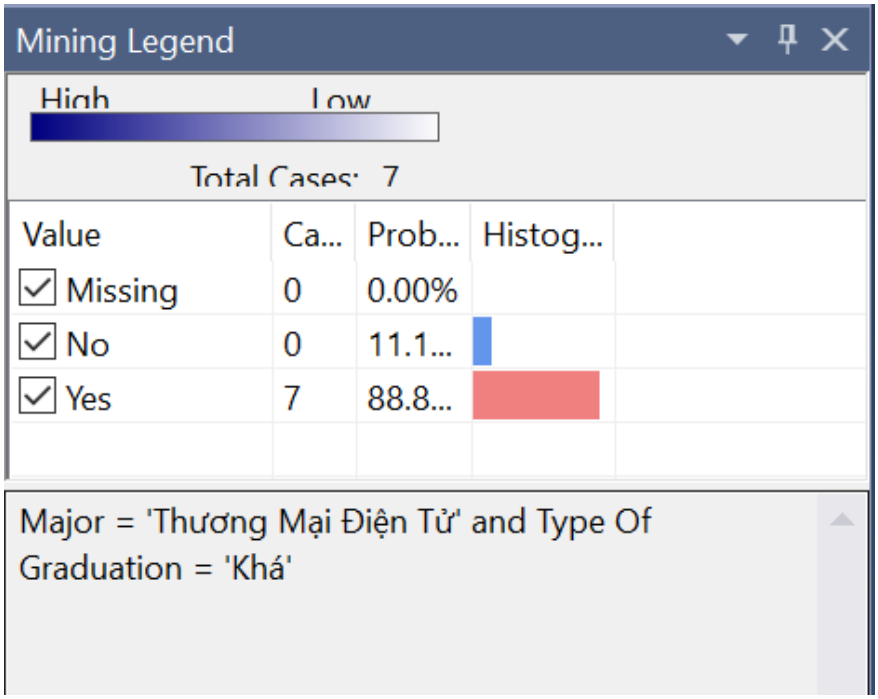
❖ **Nhóm 1 - Đúng ngành**: Lựa chọn ngành Thương mại điện tử để phân tích

Sinh viên ngành Thương mại điện tử có tỷ lệ làm đúng ngành là 62,9%, cụ thể:



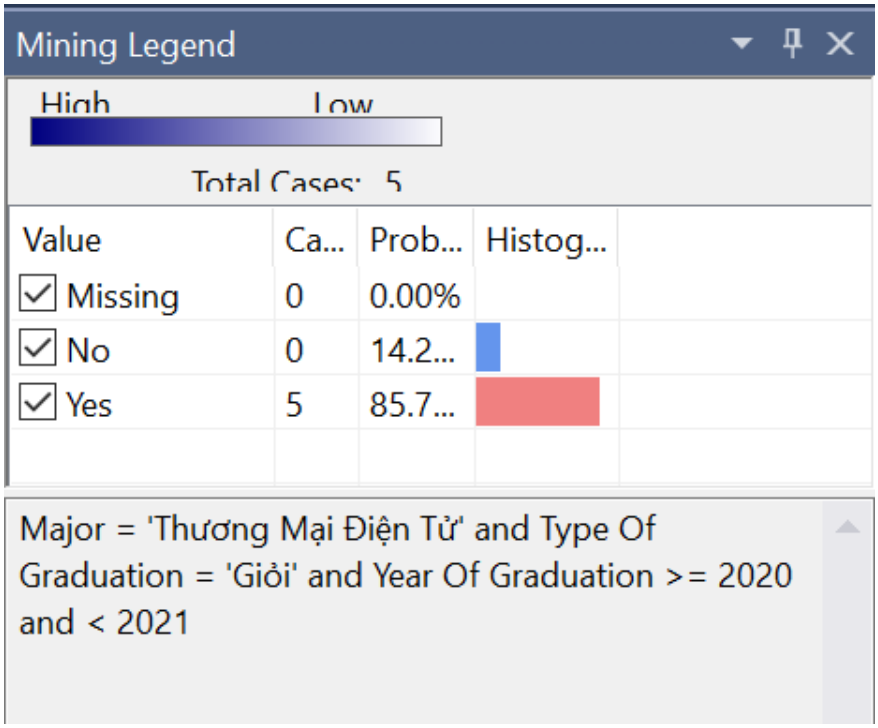
Hình 18. Tỉ trọng làm đúng ngành thuộc ngành Thương mại điện tư

- Sinh viên tốt nghiệp loại Khá có tỉ lệ làm đúng ngành cao với 88,8%.



Hình 19. Tỉ trọng làm đúng ngành thuộc ngành Thương mại điện tử theo TypeOfGraduation

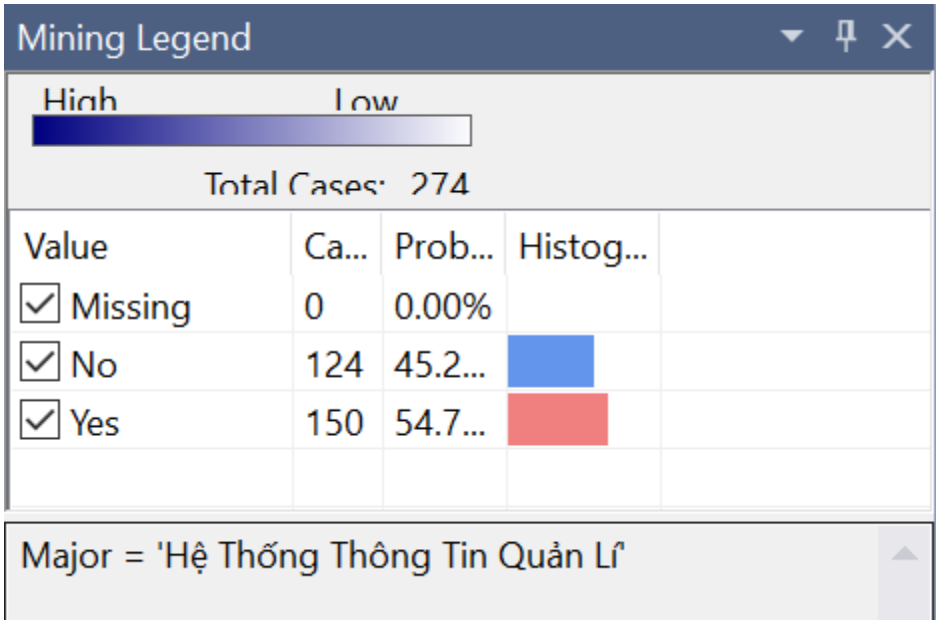
- Tốt nghiệp loại Giỏi từ năm 2020-2021 có tỉ lệ làm đúng ngành là 85,7%.



Hình 20. Tỉ trọng làm đúng ngành thuộc ngành Thương mại điện tử theo TypeOfGraduation và YearOfGraduation

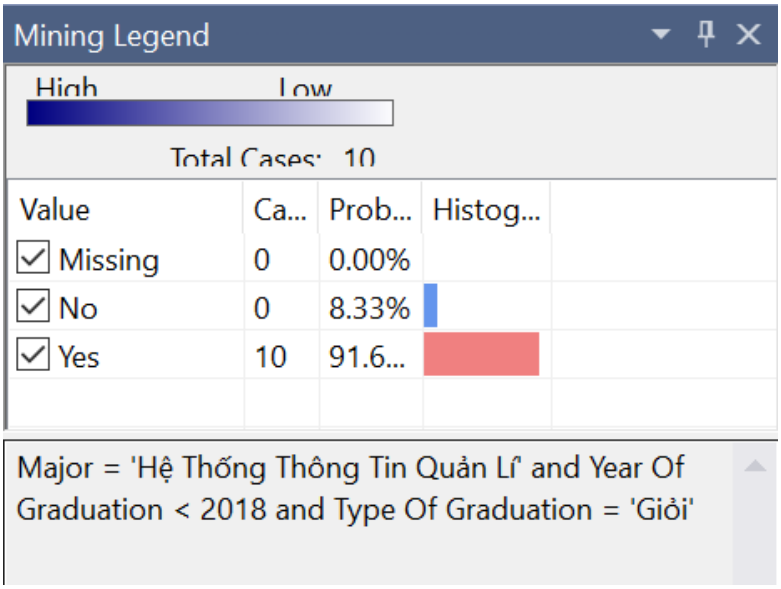
❖ **Nhóm 2 - Cân bằng:** Lựa chọn ngành Hệ thống thông tin quản lý để phân tích

Ngành Hệ thống thông tin quản lý có 54,7% tỉ lệ làm trái ngành và 45,2% tỉ lệ làm đúng ngành, cụ thể:



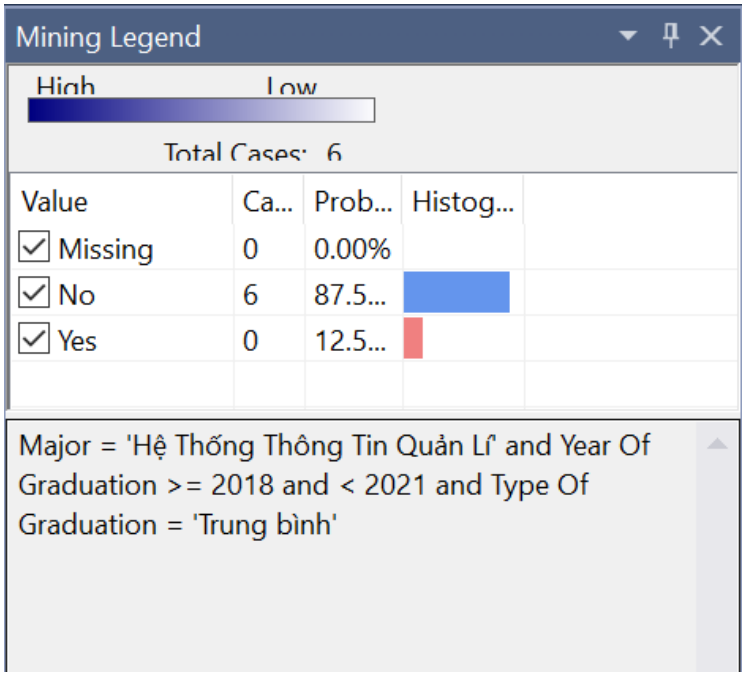
Hình 21. Tỉ trọng làm đúng ngành và trái ngành cân bằng ngành thuộc ngành Hệ thống thông tin quản lí

- Sinh viên ngành Hệ thống thông tin quản lý tốt nghiệp trước năm 2018 có học lực giỏi thì có tỷ lệ làm việc đúng ngành cao với 91,6%.



Hình 22. Tỉ trọng làm đúng ngành và trái ngành cân bằng thuộc ngành Hệ thống thông tin quản lí theo TypeOfGraduation và YearOfGraduation (1)

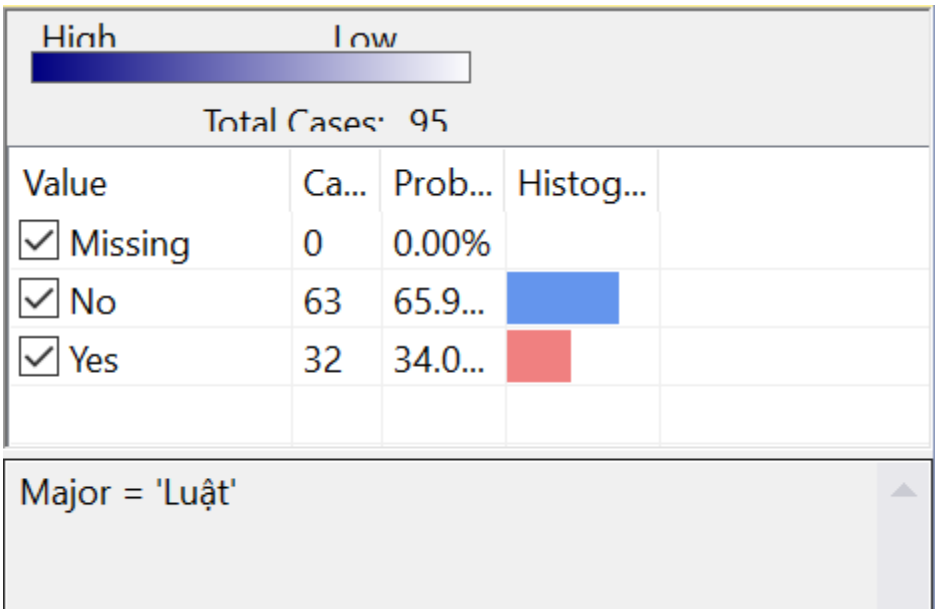
- Sinh viên ngành Hệ thống thông tin quản lý tốt nghiệp từ năm 2018 đến năm 2021 có học lực trung bình thì có tỷ lệ trái ngành cao với 87,5%.



Hình 23. Tỉ trọng làm đúng ngành và trái ngành cân bằng thuộc ngành Hệ thống thông tin quản lí theo TypeOfGraduation và YearOfGraduation (2)

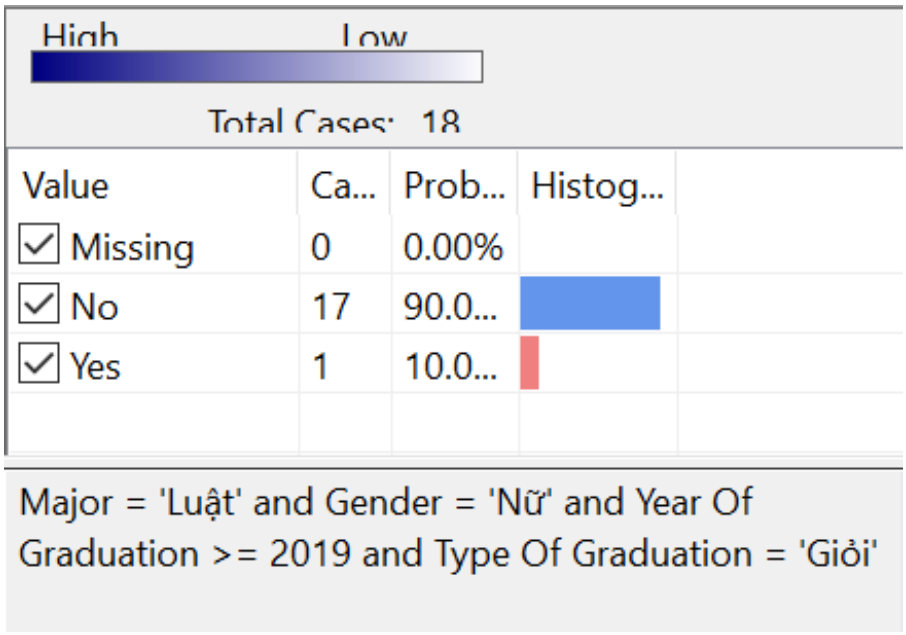
❖ **Nhóm 3 - Trái ngành**: Lựa chọn ngành Luật

Ngành Luật có 65,9% tỷ lệ làm trái ngành, cụ thể:



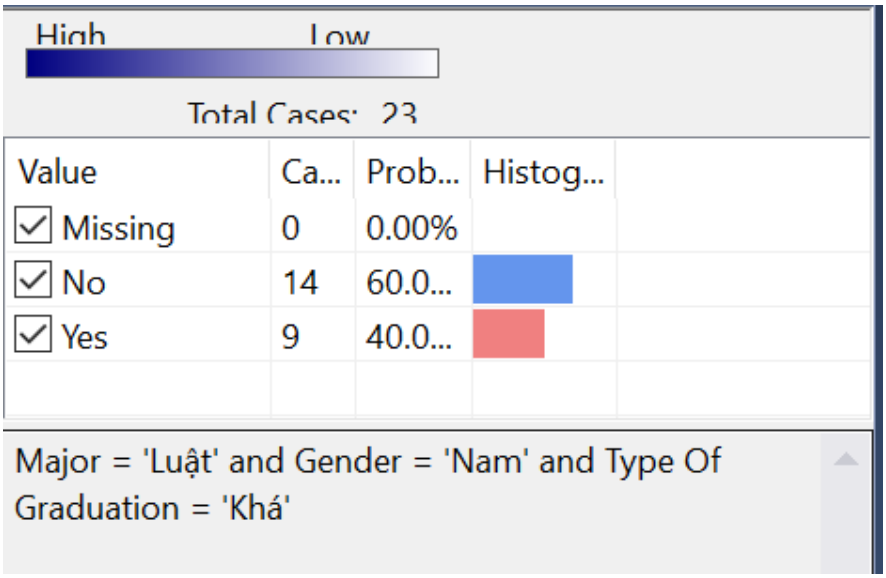
Hình 24. Tỉ trọng làm trái ngành thuộc ngành Luật

- Sinh viên nữ tốt nghiệp sau năm 2019 học lực Giỏi có tỉ lệ làm trái ngành cao với 90%



Hình 25. Tỉ trọng làm trái ngành thuộc ngành Luật theo Gender, TypeOfGraduation và YearOfGraduation (1)

- Sinh viên nam có học lực Khá có tỷ lệ thất nghiệp cao với 60%



Hình 26. Tỉ trọng làm trái ngành thuộc ngành Luật theo Gender, TypeOfGraduate và YearOfGraduation (2)

4. KẾT LUẬN

4.1. Kết quả đạt được

- Tìm hiểu được công cụ SQL Server Data Tool

- Hiểu được lý thuyết của Khai phá dữ liệu.

- Vận dụng kiến thức đã học để xây dựng thuật toán cây ra quyết định để thấy được tổng quan tỷ lệ trái ngành của sinh viên trường Đại học kinh tế.

- Đọc được kết quả từ thuật toán cây ra quyết định

4.2. Hạn chế

- Dữ liệu thu thập không đủ lớn nên cây quyết định không được chi tiết vì vậy tính chính xác khi chạy thuật toán còn chưa cao.

- Chưa trình bày đầy đủ các thao tác cài đặt, cập nhật phần mềm.

4.3. Hướng phát triển và kết luận

4.3.1. Kết luận:

Qua kết quả khai phá dữ liệu nghề nghiệp của sinh viên sau khi tốt nghiệp bằng thuật toán Decision Tree nhóm chúng em đưa ra kết luận là với những sinh viên thuộc nhóm ngành 1 (nhóm đúng ngành) đề xuất nên tiếp tục phát huy trong kế hoạch đào tạo và định hướng ngành học. Những sinh viên thuộc nhóm ngành 2 (nhóm có tỷ lệ cân bằng giữa đúng và trái ngành), đề xuất nên xem lại kế hoạch định hướng nghề nghiệp của ngành để có hướng đi tốt hơn. Những sinh viên thuộc nhóm ngành 3 (nhóm trái ngành), đề xuất nên học hỏi công tác định hướng của nhóm ngành 1 và xem lại cách định hướng để điều chỉnh nhằm giúp sinh viên nhóm ngành 3 có chuẩn đầu ra tốt hơn.

4.3.2. Hướng phát triển:

Chúng em vừa trình bày một cách tiếp cận khai phá dữ liệu để phân tích dữ liệu tỷ lệ làm trái ngành của sinh viên trường ĐH Kinh tế sau khi tốt nghiệp từ năm 2017-2023. Các bước thực hiện bao gồm thu thập dữ liệu Cựu sinh viên của trường, tiếp theo là xây dựng cơ sở dữ liệu, tiền xử lý dữ liệu và xây dựng mô hình cây quyết định để đưa ra được mô hình trực quan hóa. Kết quả thu được có thể cung cấp thông tin hữu ích cho việc định hướng sinh viên khi còn ngồi trên giảng đường đại học. Trong tương lai chúng em dự định mở rộng nghiên cứu và phát triển cho nhiều cựu sinh viên của Trường ĐH Kinh tế hơn. Ngoài ra, cần phải tham khảo thêm nhiều ý kiến khác của các chuyên gia để góp phần nâng cao độ tin cậy trong việc tìm ra những lập luật quan trọng.

TÀI LIỆU THAM KHẢO

[1]. Video bài giảng học phần Kho và Khai Phá Dữ Liệu (Data Warehouse and Data Mining Tutorials) của thầy Nguyễn Văn Chức, Trường Đại học Kinh Tế - Đại học Đà Nẵng

Link youtube: https://www.youtube.com/@chucnguyenvan6177

[2]. Introduction to Data Mining and Knowledge Discovery - Third Edition, by Two Crows Corporation.

[3]. Một số tài liệu từ Microsoft và Internet
https://vjol.info.vn/index.php/pyu/article/view/51440/42282?fbclid=IwAR1k_O0FHM-O4YlhKZl0gRuAd1VopnINTnPwSkHzUqBcGIH1X_WeJN8sums

http://www.mssqltips.com/sqlservertutorial/2000/sql-server-analysis-ervicesssas/