**ĐẠI HỌC QUỐC GIA TPHCM – TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**KHOA CÔNG NGHỆ THÔNG TIN – BỘ MÔN HỆ THỐNG THÔNG TIN**

A picture containing logo

Description automatically generated

**ĐỒ ÁN THỰC HÀNH**

**(LẦN 4)**

**Môn học: Hệ thống thông tin phục vụ trí tuệ kinh doanh**

*GVHD: Hồ Thị Hoàng Vy*

*Nguyễn Như Anh*

|  |  |  |
| --- | --- | --- |
| **Mã nhóm** | **MSSV** | **Họ và tên** |
| **TTKD-24** | 18120227 | Phạm Văn Minh Phương |
| 18120299 | Trương Công Quốc Cường |
| 18120035 | Đoàn Nguyễn Tấn Hưng |
| 18120534 | Hoàng Công Sơn |

**BẢNG PHÂN CÔNG**

**TH1**

|  |  |  |  |
| --- | --- | --- | --- |
| MSSV | Họ và tên | Phân công | Đánh giá % |
| 18120227 | Phạm Văn Minh Phương | Phân loại thuộc tính cần thiết, kiểm và sửa lỗi, làm báo cáo | 100% |
| 18120299 | Trương Công Quốc Cường | Giải thích thuộc tính từ bảng Accidents | 100% |
| 18120035 | Đoàn Nguyễn Tấn Hưng | Giải thích thuộc tính từ bảng Casualties và bảng LSOA-Postcode Mapping | 100% |
| 18120534 | Hoàng Công Sơn | Giải thích thuộc tính từ bảng Vehicles và bảng Postcodes | 100% |

**TH2**

|  |  |  |  |
| --- | --- | --- | --- |
| MSSV | Họ và tên | Phân công | Đánh giá % |
| 18120227 | Phạm Văn Minh Phương | ETL dữ liệu từ nguồn sang stage. Sửa các lỗi phát sinh. Viết báo cáo | 100% |
| 18120299 | Trương Công Quốc Cường |
| 18120035 | Đoàn Nguyễn Tấn Hưng |
| 18120534 | Hoàng Công Sơn |

**TH3**

|  |  |  |  |
| --- | --- | --- | --- |
| MSSV | Họ và tên | Phân công | Đánh giá % |
| 18120227 | Phạm Văn Minh Phương | Thiết kế NDS. Cài đặt và mô tả quy trình cài đặt các package, tạo NDS hoàn chỉnh. | 100% |
| 18120299 | Trương Công Quốc Cường | Thiết kế NDS. Viết báo cáo. |
| 18120035 | Đoàn Nguyễn Tấn Hưng | Thiết kế NDS. |
| 18120534 | Hoàng Công Sơn | Thiết kế NDS. |

**TH4**

|  |  |  |  |
| --- | --- | --- | --- |
| MSSV | Họ và tên | Phân công | Đánh giá % |
| 18120227 | Phạm Văn Minh Phương | Thiết kế NDS và DDS. Mô tả quy trình cài đặt các package ETL. Tạo model data mining và đánh giá model. Hỗ trợ OLAP và report. | 100% |
| 18120299 | Trương Công Quốc Cường | Thiết kế NDS và DDS. Cài đặt các package ETL từ Stage vào NDS, từ NDS vào DDS. Hỗ trợ OLAP và report. | 100% |
| 18120035 | Đoàn Nguyễn Tấn Hưng | OLAP và Report | 90% |
| 18120534 | Hoàng Công Sơn | OLAP và Report | 75% |

**Mục lục**

[I. Giải thích ý nghĩa thuộc tính nguồn dữ liệu 5](#_Toc92635698)

[1. UK Car Accidents (2011-2014) 5](#_Toc92635699)

[2. LSOA – Postcode Mapping 8](#_Toc92635700)

[3. Postcode 8](#_Toc92635701)

[II. Quá trình ETL dữ liệu từ nguồn vào stage 10](#_Toc92635702)

[III. Quá trình ETL dữ liệu từ Stage vào NDS 21](#_Toc92635703)

[1. Thiết kế NDS 21](#_Toc92635704)

[2. Quá trình ETL dữ liệu 21](#_Toc92635705)

[IV. Quá trình ETL dữ liệu từ NDS vào DDS 21](#_Toc92635706)

[1. Thiết kế DDS 21](#_Toc92635707)

[2. Quá trình ETL dữ liệu 21](#_Toc92635708)

[a. Chiều … 21](#_Toc92635709)

[b. Chiều … 21](#_Toc92635710)

[V. OLAP và Report 21](#_Toc92635711)

[1. Thống kê số lượng nạn nhân theo Mức Độ Nghiêm Trọng ở các Địa phương trong tất cả các năm 21](#_Toc92635712)

[2. … 21](#_Toc92635713)

[VI. Data Mining 22](#_Toc92635714)

[1. Đặt yêu cầu và phân tích 22](#_Toc92635715)

[2. Triển khai model 22](#_Toc92635716)

[3. Đánh giá model 27](#_Toc92635717)

[a. Decision Tree 27](#_Toc92635718)

[b. Dependency Network 29](#_Toc92635719)

[c. Mining Accuracy Chart 30](#_Toc92635720)

[- Chọn Input 30](#_Toc92635721)

[- Classification Matrix 31](#_Toc92635722)

[d. Model Mining Prediction 32](#_Toc92635723)

[- Khởi tạo 32](#_Toc92635724)

[- Kết quả 33](#_Toc92635725)

[VII. Kết luận chung 34](#_Toc92635726)

# Giải thích ý nghĩa thuộc tính nguồn dữ liệu

## UK Car Accidents (2011-2014)

* Bảng Accidents:

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Thuộc tính** | **Ý nghĩa** | **Kiểu dữ liệu** |
| 1 | Location Easting OSGR | mã OSGR (**Ordnance Survey Grid Reference**) phía Đông | INT |
| 2 | Location Northing OSGR | mã OSGR phía Bắc | INT |
| 3 | Longitude | Kinh độ | FLOAT |
| 4 | Latitude | Vĩ độ | FLOAT |
| 6 | Accident Severity | Mức độ độ nghiêm trọng của tai nạn | INT |
| 7 | Number of Vehicles | Số phương tiện trong tai nạn | INT |
| 8 | Number of Casualties | Số thương vong | INT |
| 9 | Date | Ngày xảy ra tai nạn | DATE |
| 10 | Day of Week | Ngày thứ mấy trong tuần | INT |
| 11 | Time | Giờ xảy ra tai nạn | TIME |
| 12 | Road type | Loại đường giao thông | INT |
| 13 | Speed limit | Giới hạn tốc độ | INT |
| 14 | Junction Detail | Chi tiết giao lộ | INT |
| 15 | Junction Control | Đèn giao thông (giá trị) | INT |
| 16 | Light Conditions | Điều kiện ánh sáng | INT |
| 17 | Weather Conditions | Điều kiện thời tiết | INT |
| 18 | Road Surface Conditions | Điều kiện mặt đường | INT |
| 19 | Special Conditions at Site | Điều kiện đặc biệt tại địa điểm | INT |
| 20 | Carriageway Hazards | Mối nguy hiểm trên đường | INT |
| 21 | Urban or Rural Area | Vùng đô thị hay ngoại ô | INT |
| 22 | LSOA of Accident Location | Mã LSOA (**Lower Layer Super Output Areas**) của vị trí tai nạn | VARCHAR |

* Bảng Casualties:

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Thuộc tính** | **Ý nghĩa** | **Kiểu dữ liệu** |
| 1 | Accident\_Index | Index của bảng | VARCHAR |
| 2 | Vehicle\_Reference | Tham chiếu tới bảng Vehicle | INT |
| 3 | Casualty\_Reference | Tham chiếu tới bảng Casualty | INT |
| 4 | Casualty\_Class | Cấp tai nạn (cấp 1, 2 hoặc 3) | INT |
| 5 | Sex\_of\_Casualty | Giới tính người bị nạn (1 hoặc 2) | INT |
| 6 | Age\_of\_Casualty | Tuổi người bị nạn | INT |
| 7 | Age\_Band\_of\_Casualty | Nhóm tuổi người bị nạn | INT |
| 8 | Casualty\_Severity | Mức độ tai nạn (mức 1, 2 hoặc 3) | INT |
| 9 | Car\_Passenger | Hành khách xe hơi | INT |
| 10 | Bus\_or\_Coach\_Passenger | Hành khách xe buýt hoặc xe khách | INT |
| 11 | Casualty\_Type | Loại tai nạn | INT |

* Bảng Vehicles:

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Tên thuộc tính** | **Ý nghĩa** | **Kiểu dữ liệu** |
| 1 | Vehicle Reference | Tham chiếu phương tiện | INT |
| 2 | Vehicle Type | Loại phương tiện | INT |
| 5 | Vehicle Location-Restricted Lane | Vị trí phương tiện - tại làn đường hạn chế | INT |
| 6 | Junction Location | Vị trí giao lộ | INT |
| 7 | Journey Purpose of Driver | Lý do tham gia giao thông của tài xế | INT |
| 8 | Sex of Driver | Giới tính tài xế | INT |
| 9 | Age of Driver | Tuổi của tài xế | INT |

## LSOA – Postcode Mapping

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Tên thuộc tính** | **Ý nghĩa** | **Kiểu dữ liệu** |
| 1 | oa11cd | Mã code OA (Output Area) năm 2011 | VARCHAR |
| 2 | lsoa11cd | Mã code LSOA (Lower Layer Output Area) năm 2011 | VARCHAR |
| 3 | msoa11cd | Mã code MSOA (Middle Layer Output Area) năm 2011 | VARCHAR |
| 4 | ladcd | Mã code LAD (Local authority districts) | VARCHAR |
| 5 | lsoa11nm | Tên của LSOA | VARCHAR |
| 6 | msoa11nm | Tên của MSOA | VARCHAR |

## Postcode

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Thuộc tính** | **Ý nghĩa** | **Kiểu dữ liệu** |
| 1 | postcode | Mã bưu chính | VARCHAR |
| 2 | easting | Tọa độ x trong hệ tham chiếu lưới | INT |
| 3 | northing | Tọa độ y trong hệ tham chiếu lưới | INT |
| 4 | latitude | Vĩ độ | FLOAT |
| 5 | longitude | Kinh độ | FLOAT |
| 6 | city | Thành phố | VARCHAR |
| 7 | county | Quận | VARCHAR |
| 8 | country\_code | Mã quốc gia | VARCHAR |
| 9 | country\_name | Tên quốc gia | VARCHAR |
| 10 | iso3166-2 | GB-Mã quốc gia (Great Britain) | VARCHAR |
| 11 | region\_code | Mã vùng | VARCHAR |

# Quá trình ETL dữ liệu từ nguồn vào stage

B1. Tạo database mới để chứa dữ liệu từ dữ liệu UK Car Accidents (2011-2014), LSOA-Postcode Mapping, Postcode

A screenshot of a computer

Description automatically generated

B2. Tạo project Integration Process Project để chạy các Data Flow Task nhằm nạp dữ liệu

A screenshot of a computer

Description automatically generated with medium confidence

B3. Mở tab Data Flow Task để tạo task nạp dữ liệu. Sau đó, kéo biểu tượng “Flat File Source” (Nguồn dữ liệu từ file .csv) và “OLE DB Destination” (Đích dữ liệu là SQL Database) vào trong giao diện.

A screenshot of a computer

Description automatically generated with medium confidence

B4. Double click vào biểu tượng Flat File Source trong giao diện Data Flow Task và tiến hành tạo kết nối tới file .csv dữ liệu gốc bằng cách nhấn vào “New”. Chọn “Browse” tới file dữ liệu gốc cần nạp. Nên đổi tên kết nối mặc định để dễ phân biệt.

A screenshot of a computer

Description automatically generated

B5. Nhấp vào tab “Column” để kiểm tra xem dữ liệu được đọc lên đúng hay chưa. Nếu đã đúng, nhấn, chuyển qua tab “Advance” để thay đổi kiểu dữ liệu của dữ liệu đầu vào

A screenshot of a computer

Description automatically generated

Sau khi đã chọn đúng kiểu dữ liệu mong muốn, nhấn OK. Có thể sử dụng “Suggest Types..” để hỗ trợ xác định đúng kiểu dữ liệu.



Chú ý tích vào ô “Retain null values from the source as null values in the data flow” để giữ lại giá trị null trong kho dữ liệu.

A screenshot of a computer

Description automatically generated

B6. Double click vào tên của biểu tượng Flat File Source trên giao diện để sửa đổi tên tương ứng với nguồn dữ liệu, tránh nhầm lẫn. Sau đó, kéo mũi tên màu xanh (mũi tên output) tới biểu tượng của đích dữ liệu (OLE DB Destination)

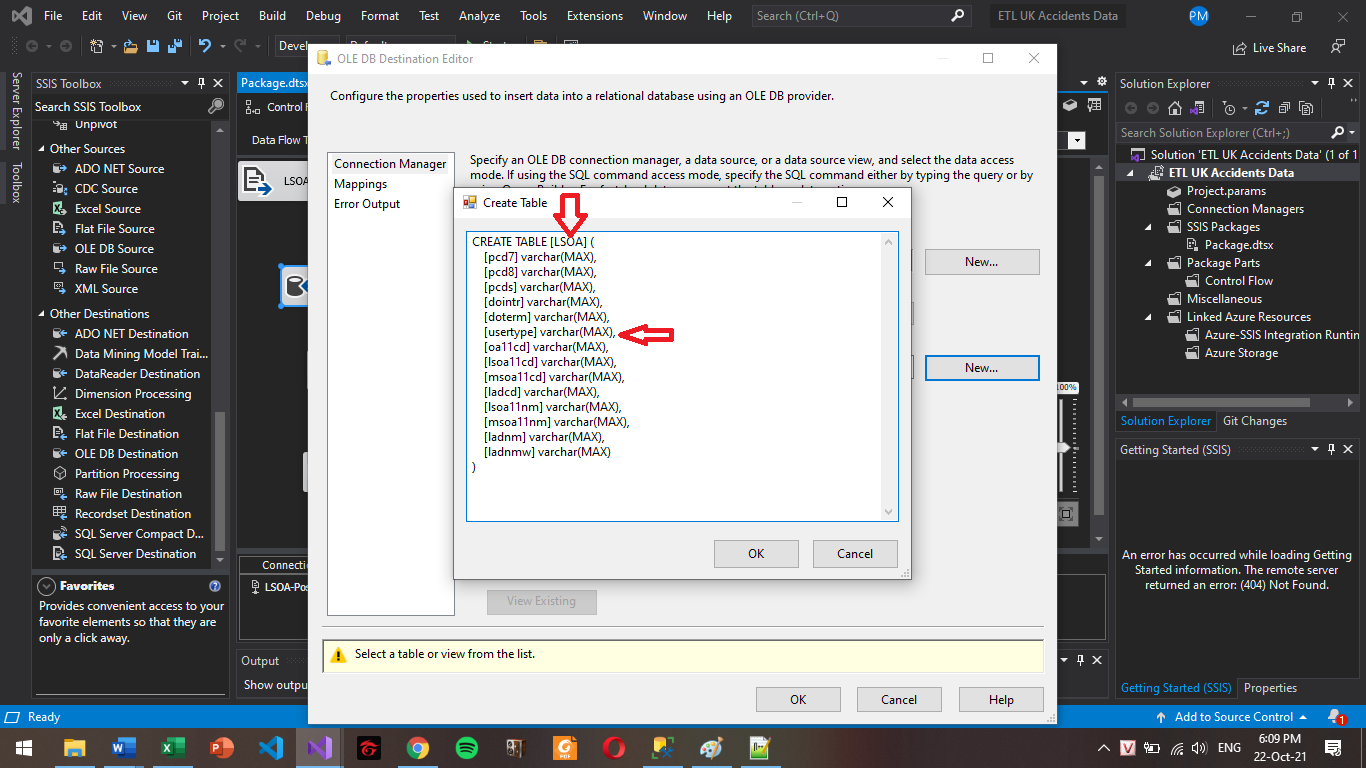
A screenshot of a computer

Description automatically generated with medium confidence

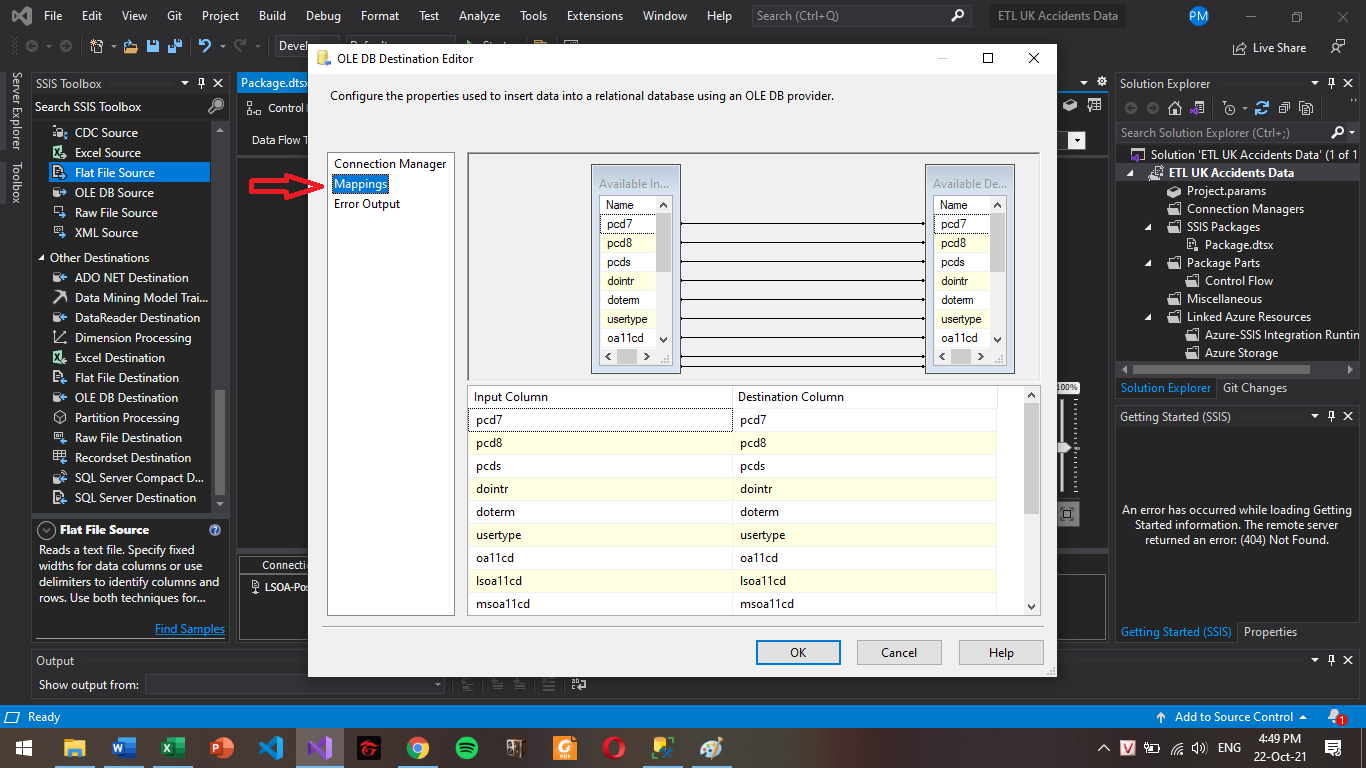
B7. Double click vào biểu tượng OLE DB Destination, chọn “New” để tạo kết nối tới Database đích tương ứng với nguồn dữ liệu được chọn. Nhấn “OK” sau khi chọn đúng giá trị ở mỗi tabA screenshot of a computer

Description automatically generated

B8. Tạo bảng trong Database để chứa dữ liệu từ nguồn. Chú ý loại kiểu dữ liệu và kích cỡ phải khớp với dữ liệu gốc để tránh xảy ra lỗi.



B9. Kiểm tra mapping dữ liệu

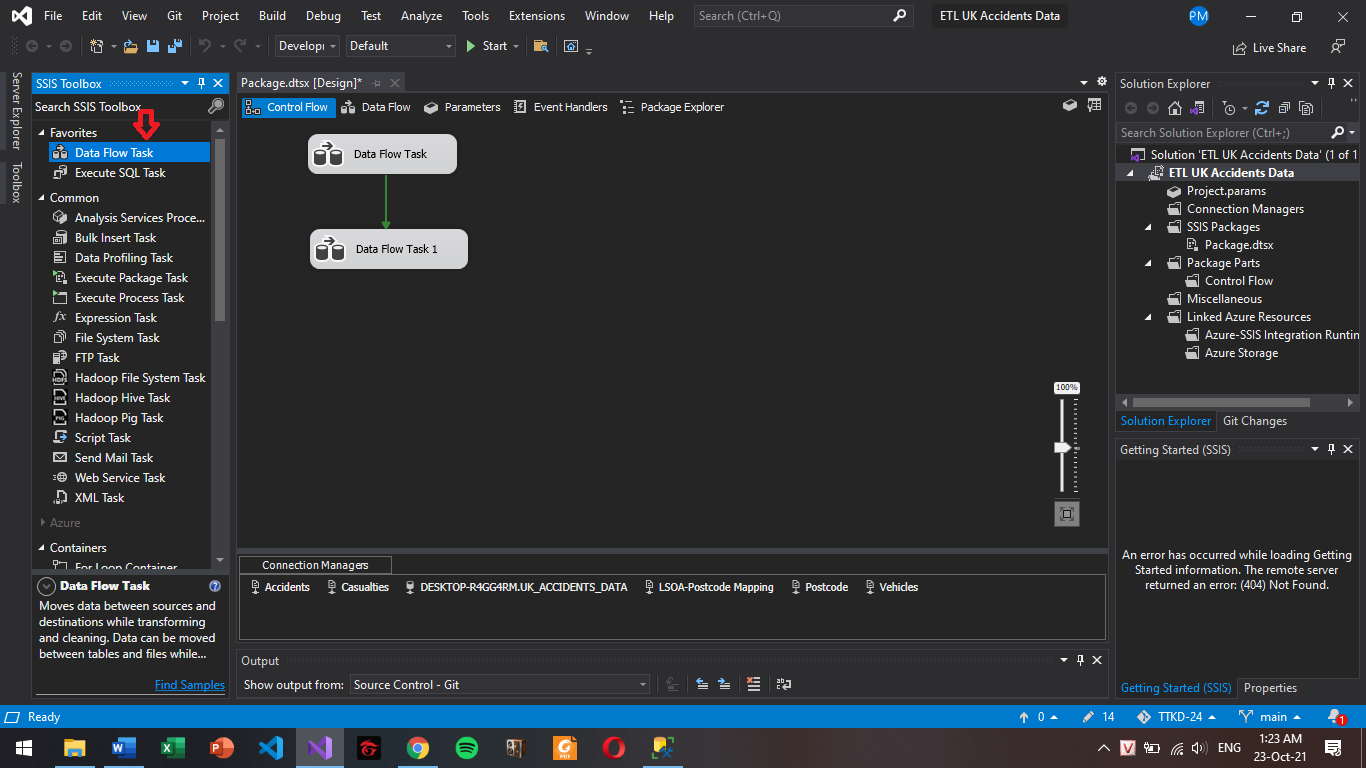


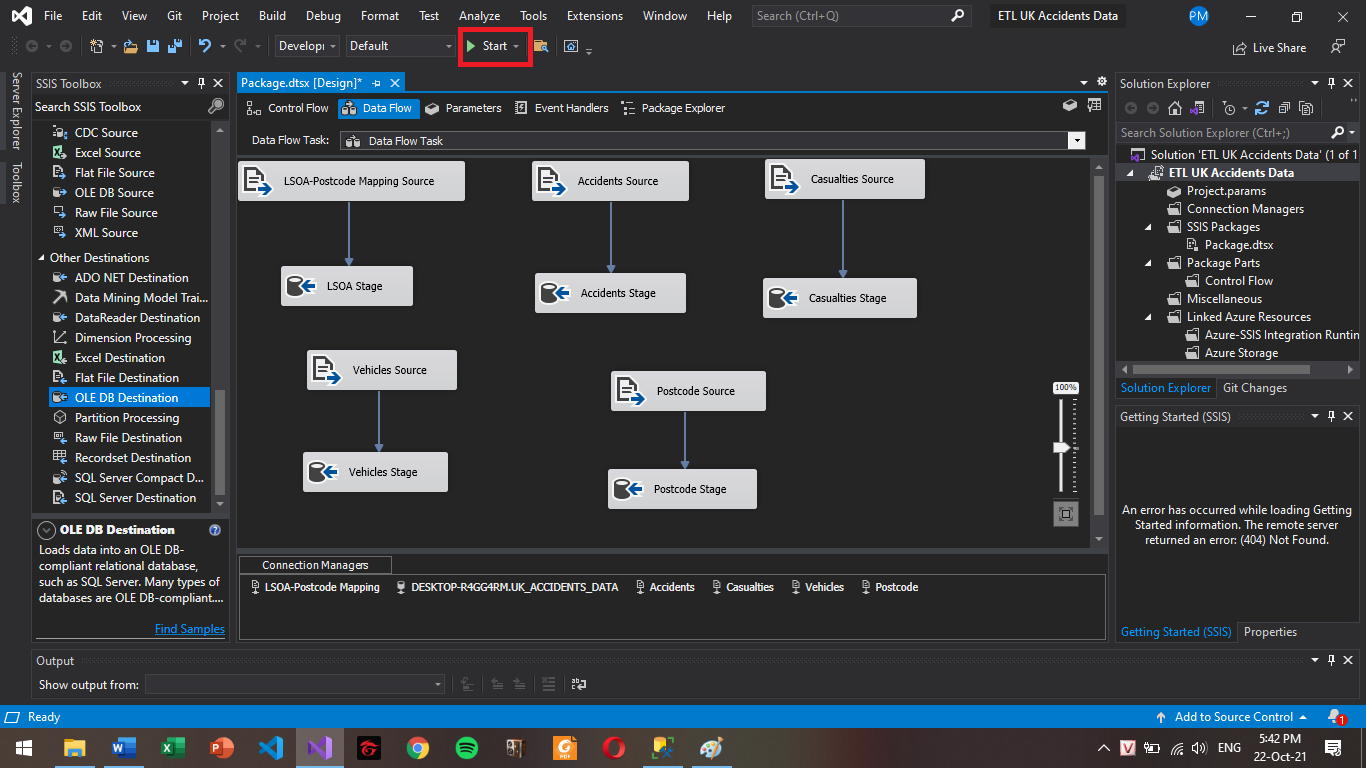
B10. Double click vào tên của biểu tượng OLE DB Destination trên giao diện để đổi lại tên tương ứng, tránh nhầm lẫn

A screenshot of a computer

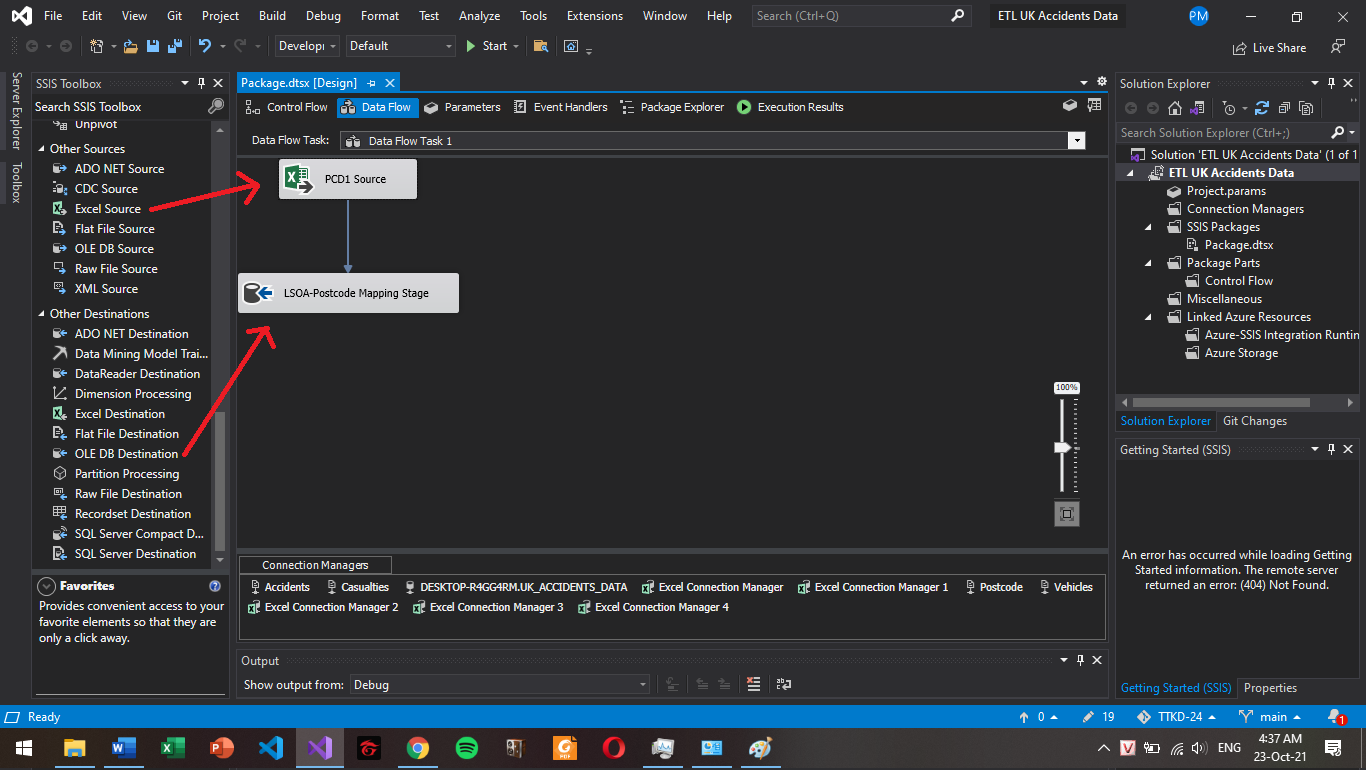
Description automatically generated with medium confidence

B11. Tạo một Data Flow Task khác và lặp lại tương tự từ B3 tới B10 với các file nguồn dữ liệu và đích dữ liệu còn lại. Nối Data Flow Task mới tạo này vào Data Flow Task ban đầu để cho Data Flow Task này chạy trước

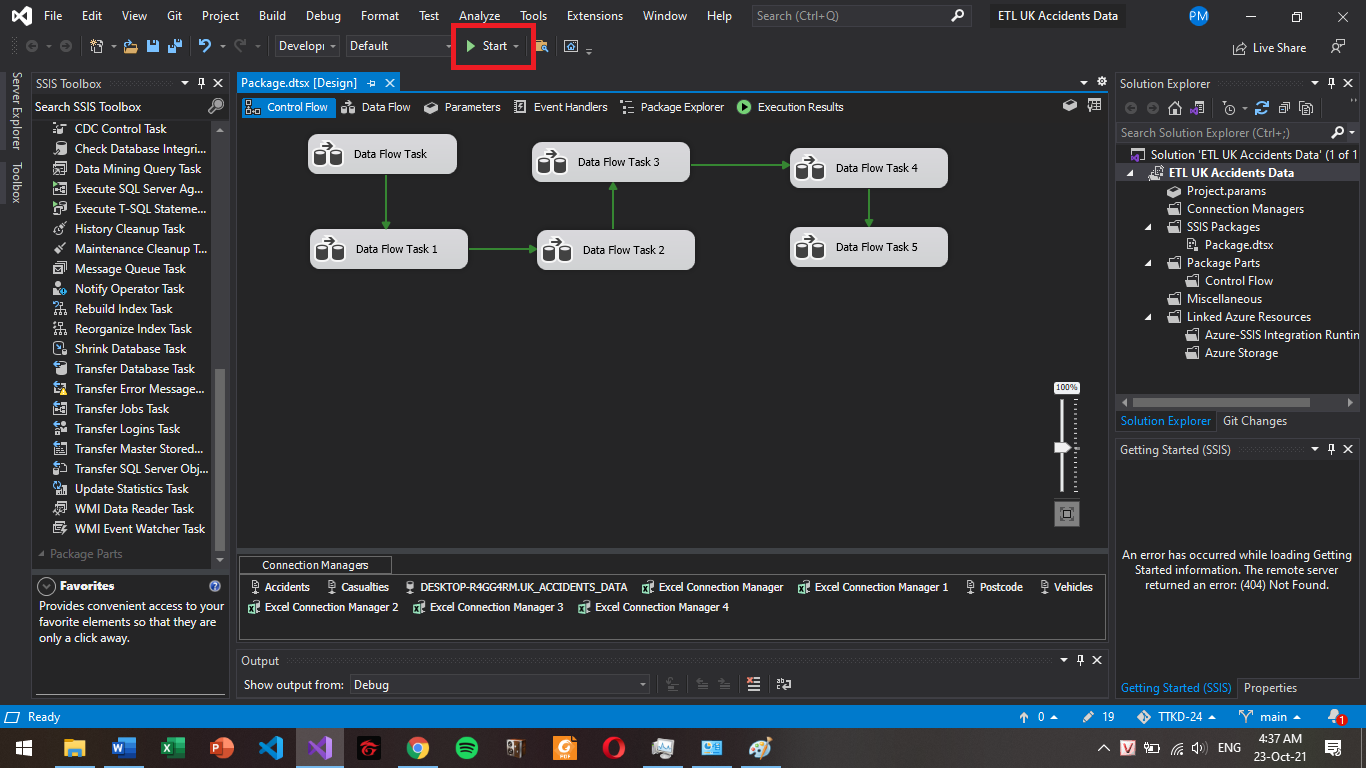




B12. Thực hiện tương tự B3-B10 nhưng với source là “Excel Source” cho các file .xlxs là nguồn dữ liệu LSOA-Postcode mapping được chia nhỏ và convert.



B13. Nối thứ tự các task và click “Start”



**Kết quả chạy thành công:**

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

Graphical user interface, text

Description automatically generated

# Quá trình ETL dữ liệu từ Stage vào NDS

## Thiết kế NDS

## Quá trình ETL dữ liệu

# Quá trình ETL dữ liệu từ NDS vào DDS

## Thiết kế DDS

## Quá trình ETL dữ liệu

### Chiều …

### Chiều …

# OLAP và Report

## Thống kê số lượng nạn nhân theo Mức Độ Nghiêm Trọng ở các Địa phương trong tất cả các năm

## …

# Data Mining

## Đặt yêu cầu và phân tích

Yêu cầu: Sử dụng mô hình dự đoán mức độ nghiêm trọng của các tai nạn

Dữ liệu hiện có: Thông tin liên quan đến thời điểm và địa điểm tai nạn xảy ra (Buổi xảy ra tai nạn, tại vùng nông thôn hay đô thị, giới hạn tốc độ, loại đường, mức độ nghiêm trọng…)

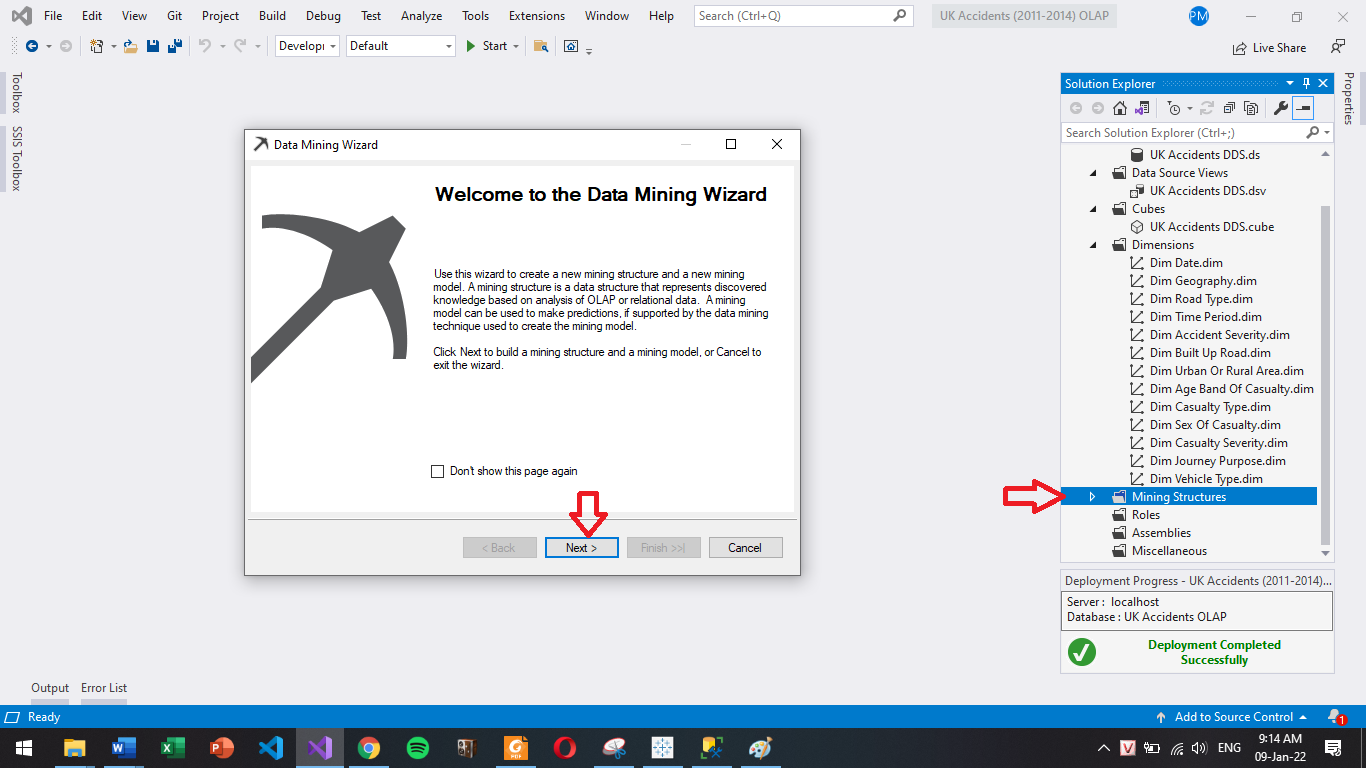
Mục tiêu: Tìm hiểu liệu các yếu tố trên có ảnh hưởng gì đến mức độ nghiêm trọng của tai nạn khi tai nạn xảy ra hay không và sử dụng các yếu tố trên để dự đoán mức độ nghiêm trọng các tai nạn

Model mining phù hợp dự tính: Decision Tree

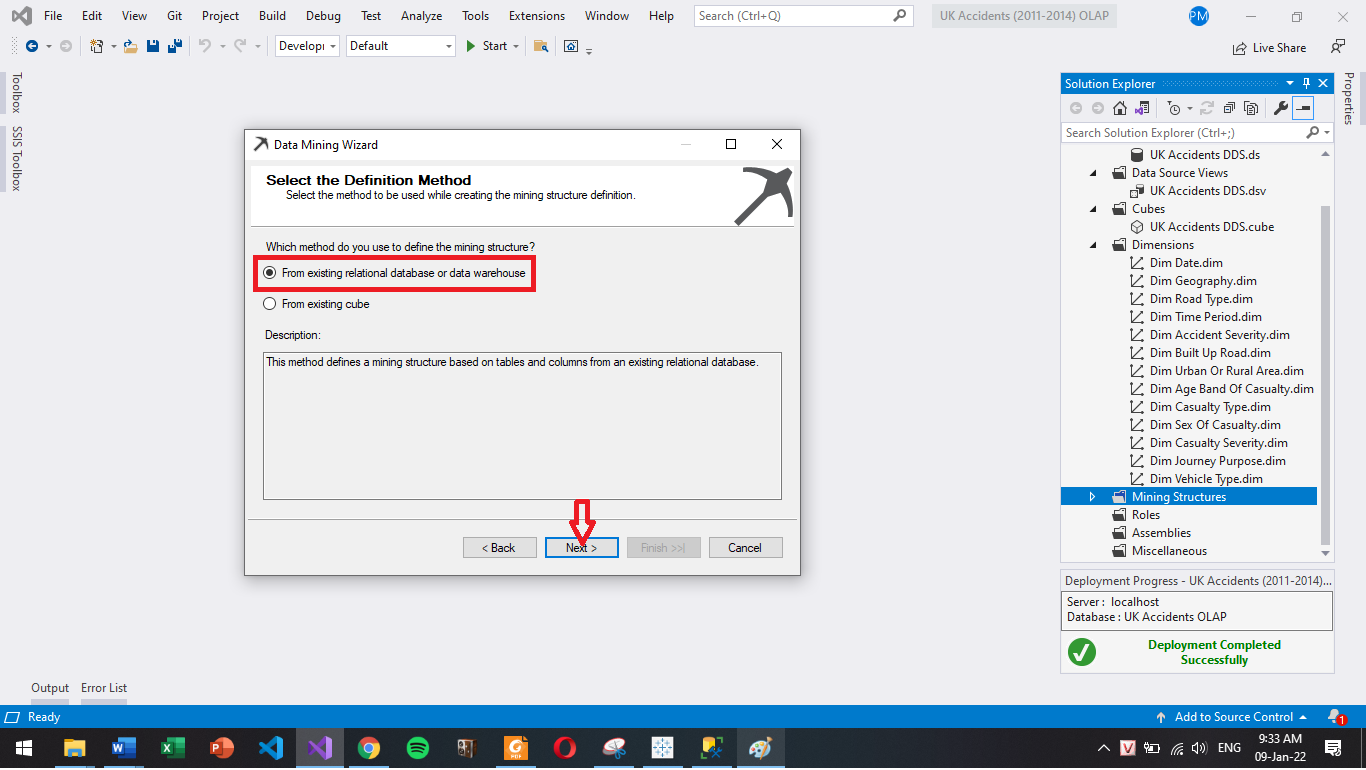
## Triển khai model

Sử dụng project Analysis Services Multidimensional and Data Mining đã kết nối tới nguồn dữ liệu, tạo Data Source View và làm theo các bước sau:

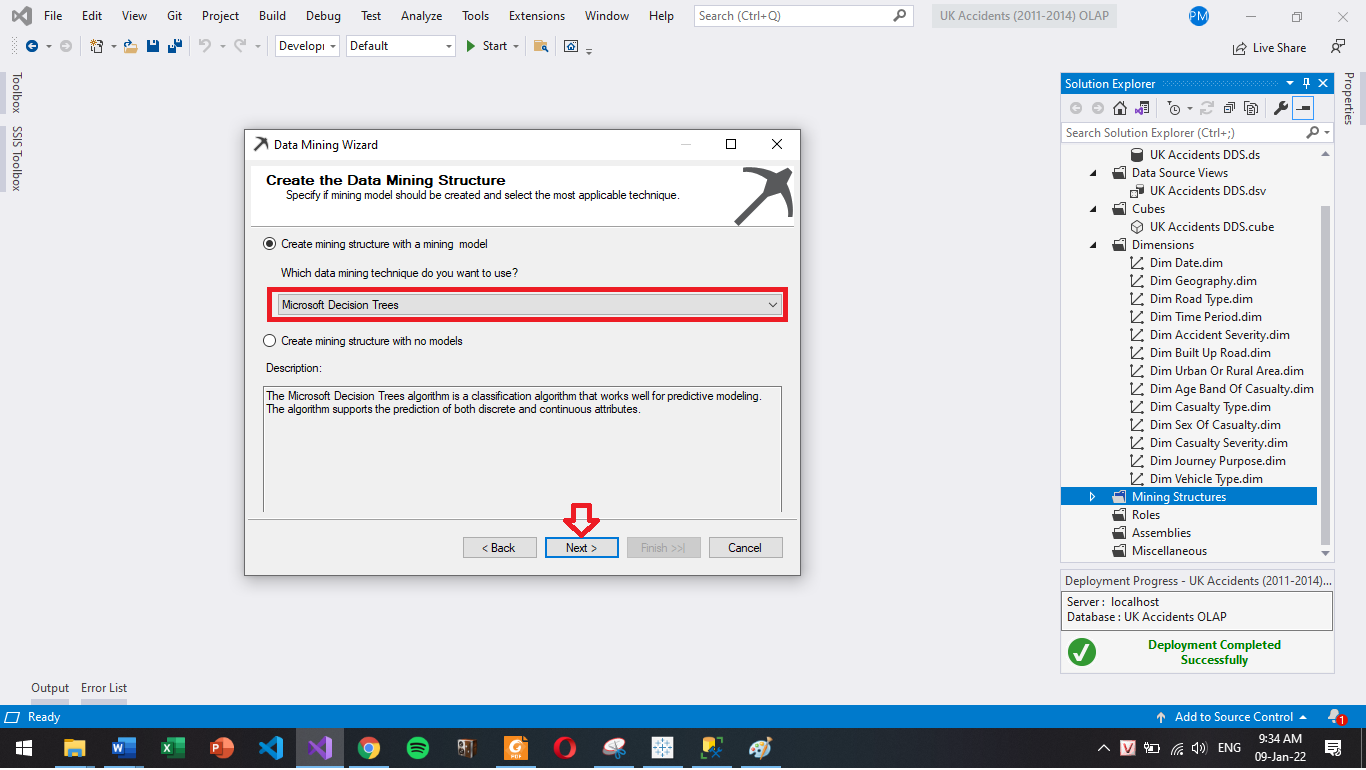
Bước 1: Chuột phải vào Mining Structure, chọn “New Mining Structure”. Chọn “Next”.



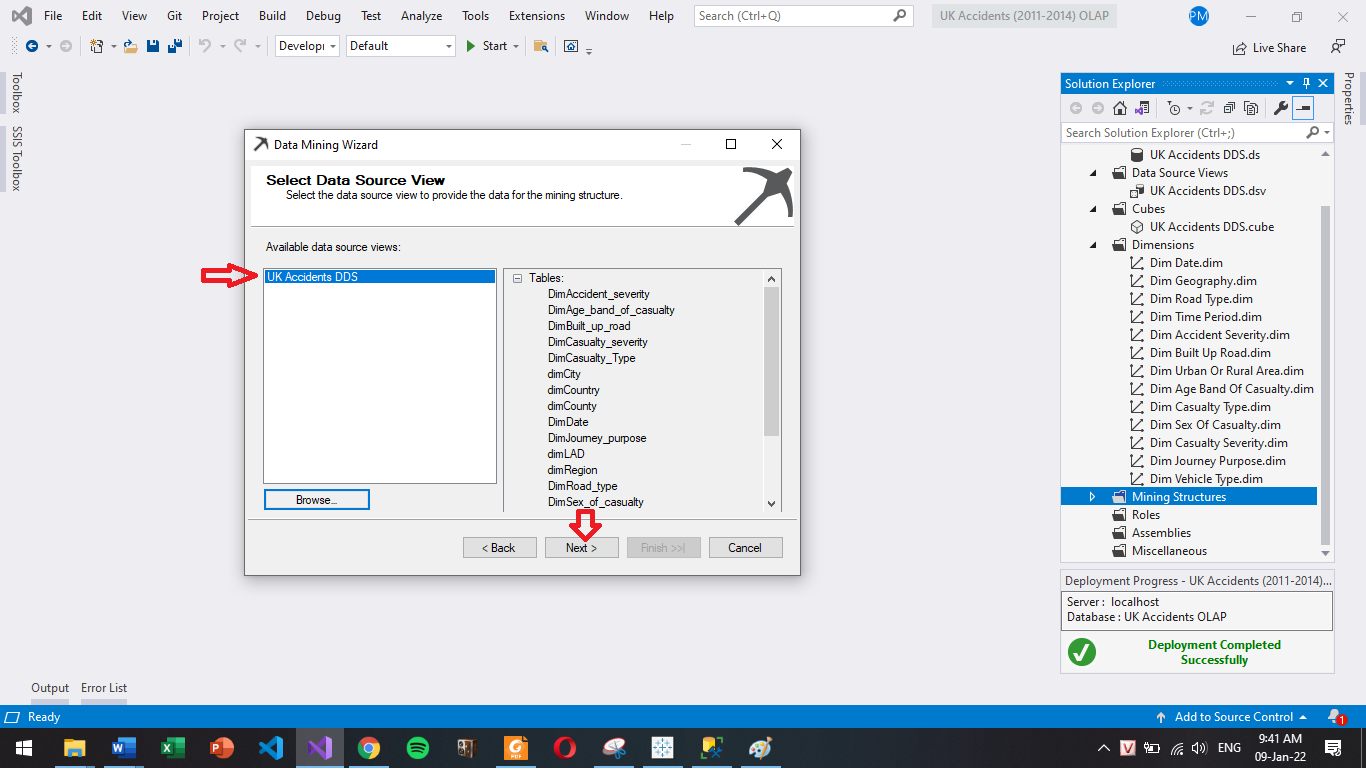
Bước 2: Chọn sử dụng data warehouse hoặc CSDL quan hệ có sẵn



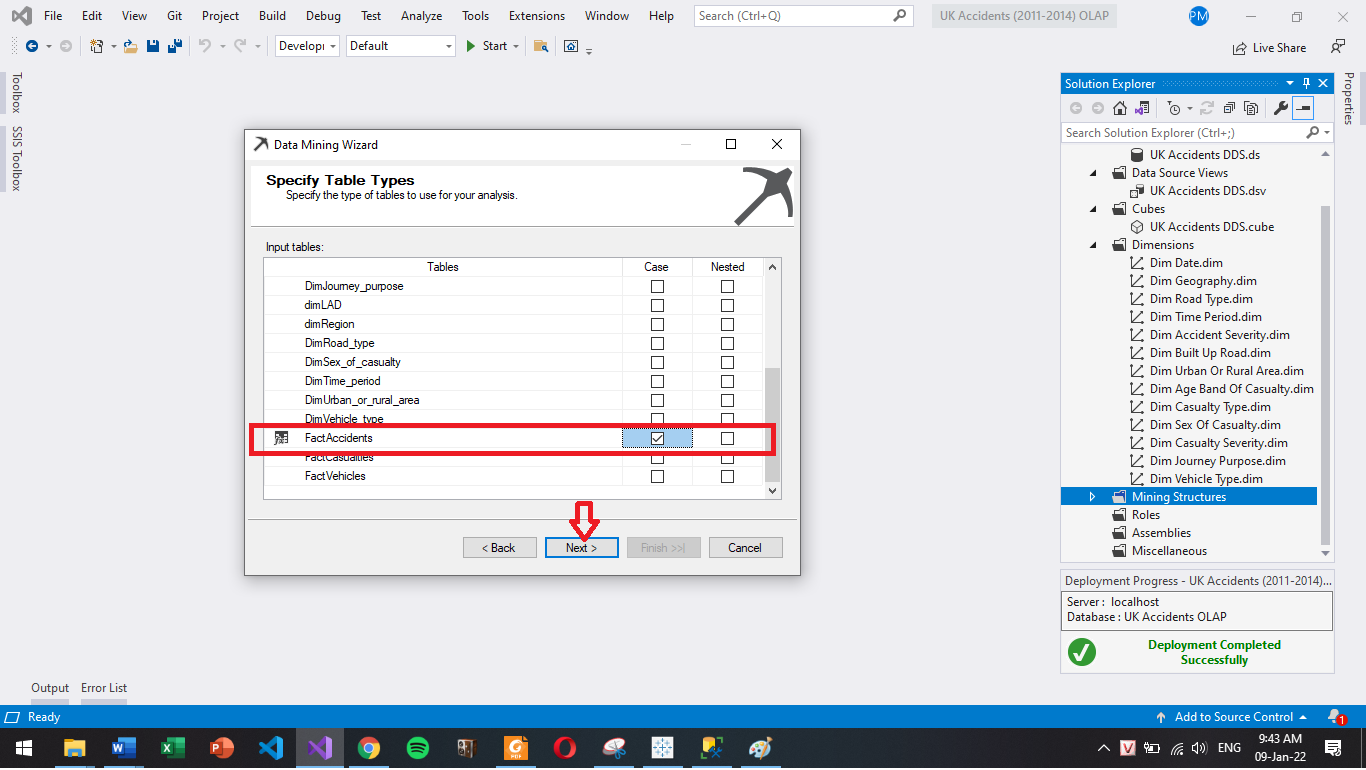
Bước 3: Chọn cấu trúc mining “Decision Tree”



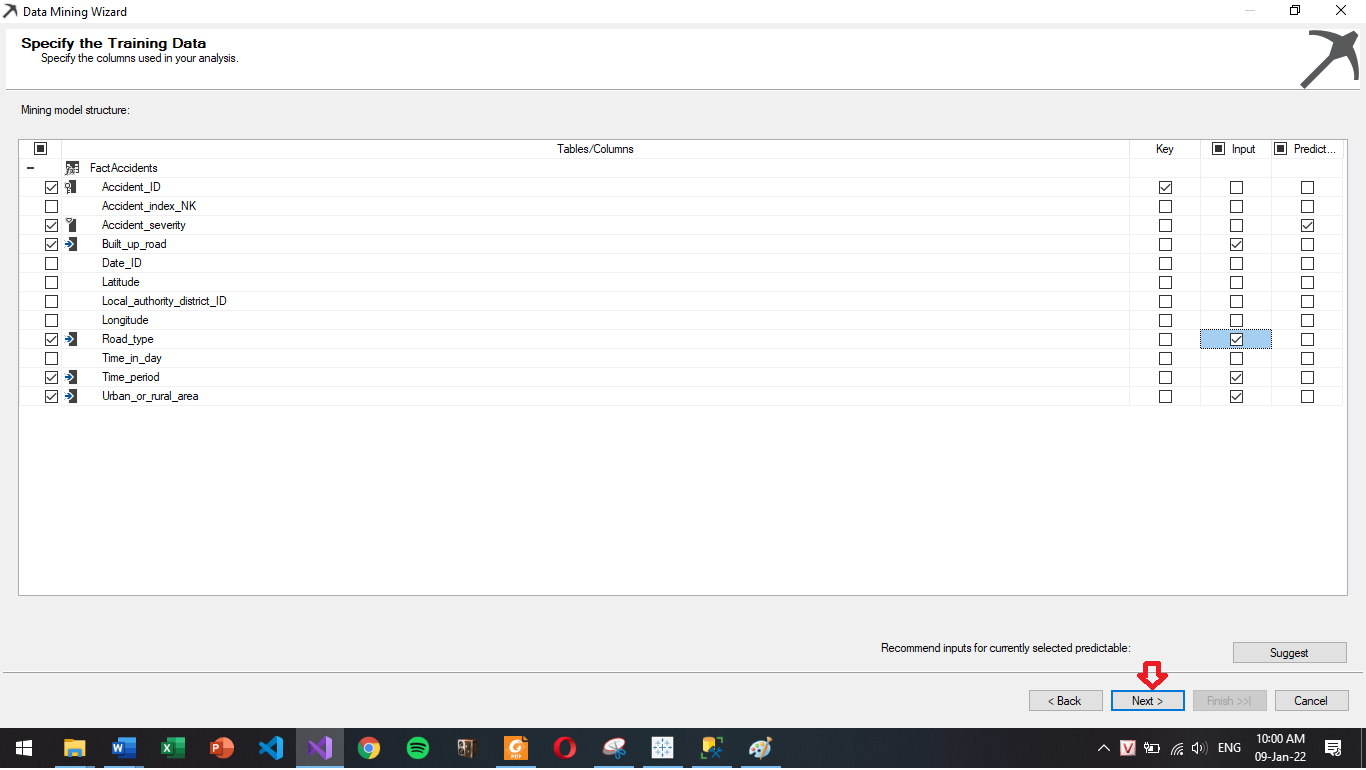
Bước 4: Chọn Data Source View



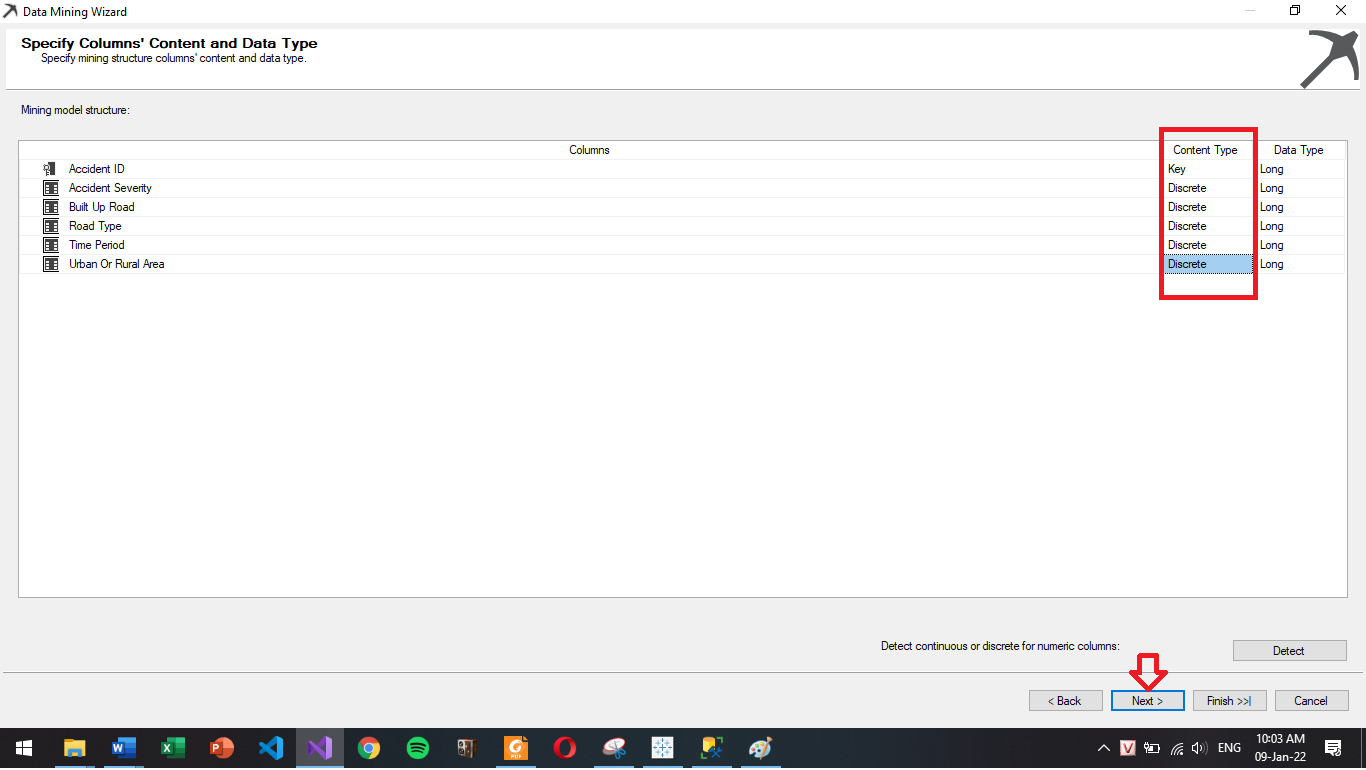
Bước 5: Chọn bảng case là FactAccidents



Bước 6: Chọn các thuộc tính input (Built\_up\_road, Road\_type, Time\_period, Urban\_or\_rural\_area) và thuộc tính để dự đoán (trường hợp này là Accident\_severity)

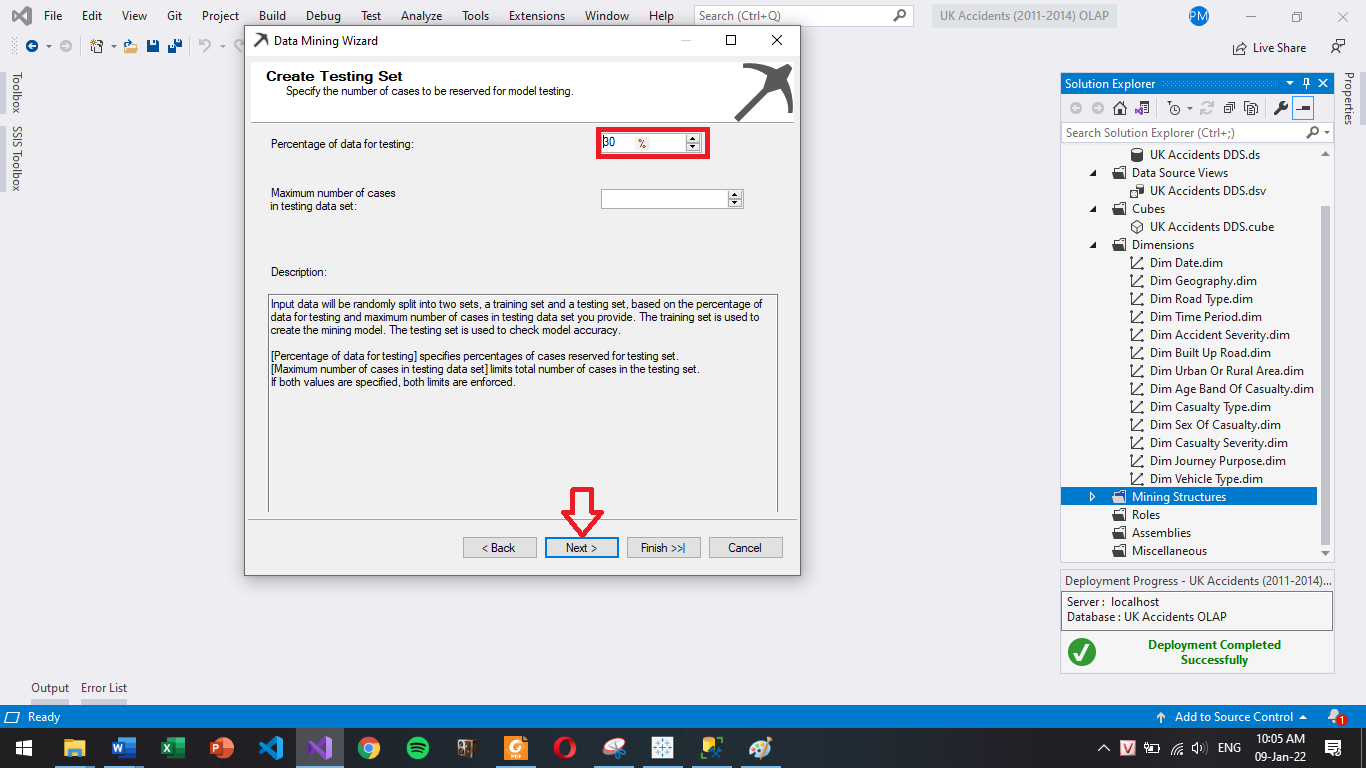


Bước 7: Xác định loại thuộc tính cho các thuộc tính đã chọn là rời rạc (Discrete) hay liên tục (Continuos)

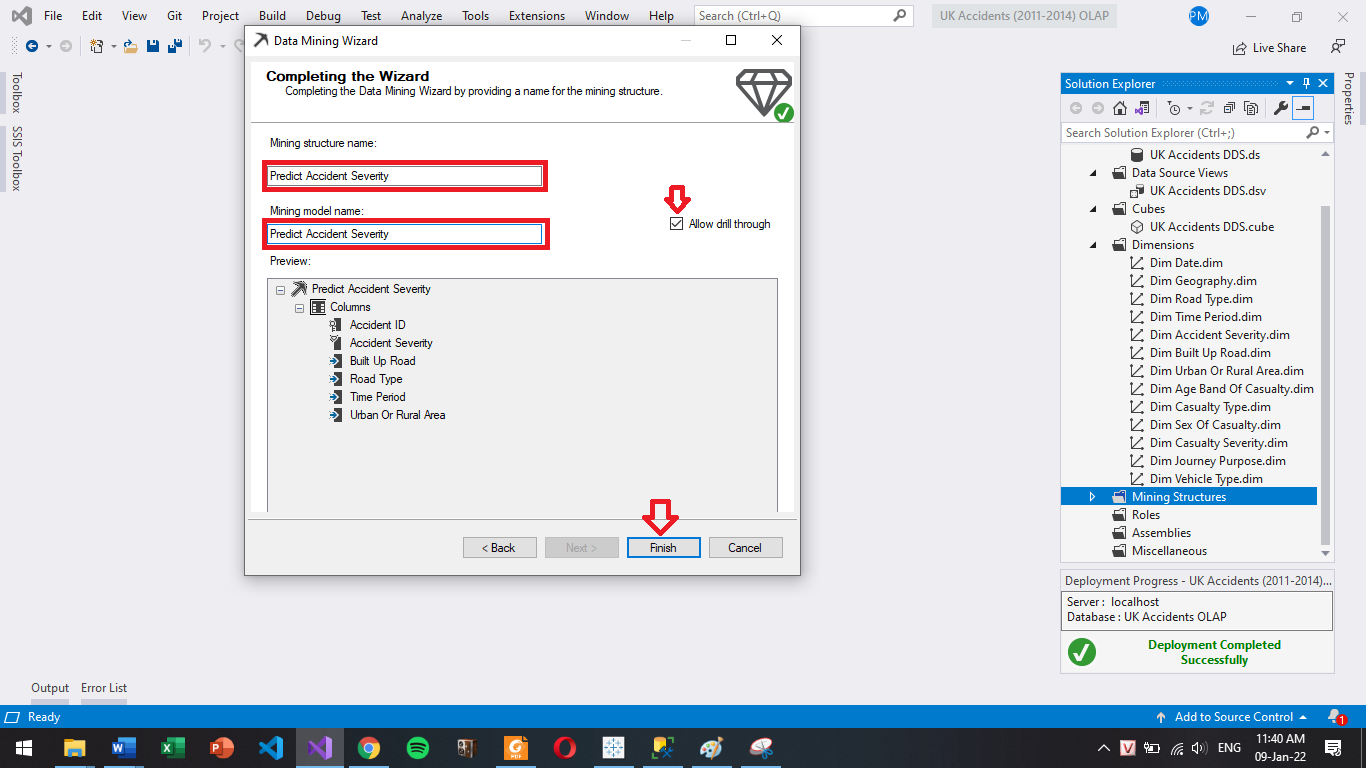


Tất cả các thuộc tính trên đều thuộc loại rời rạc vì chúng đều là thuộc tính có tập giá trị giới hạn và được xác định trước

Bước 8: Xác định tỉ lệ dữ liệu sẽ được dùng cho việc test độ chính xác của model. Mặc định là 30%



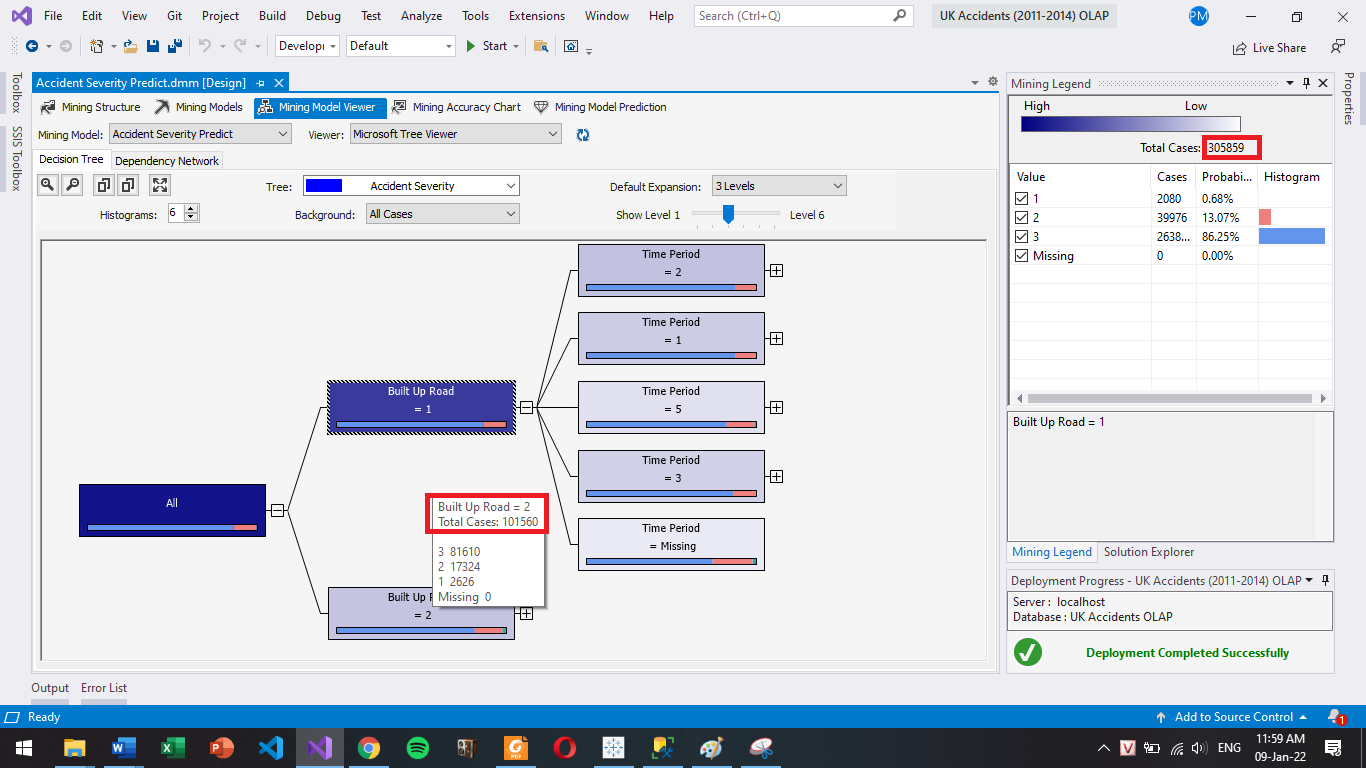
Bước 9: Đặt tên cho Mining Structure và Mining Model



## Đánh giá model

### Decision Tree

Dựa trên các thuộc tính đã chọn, ta có một cây quyết định 6 bậc. Ta có thể chọn các cách hiển thị khác nhau, dưới đây là hiển thị của số lượng tai nạn thỏa mãn các yếu tố được viết trên các node của cây. Màu càng đậm tức là có càng nhiều tai nạn thỏa yếu tố đó. Có thể thấy màu của node “Built Up Road = 1” đậm hơn hẳn node cùng cấp còn lại bởi vì số lượng tai nạn xảy ra có yêu tố Built Up Road = 1 (tức đường có tốc độ giới hạn dưới 50mph) nhiều gấp 3 lần so với số lượng tai nạn xảy ra có yếu tố Bult Up Road = 2 (tức đường có tốc độ giới hạn lớn hơn 50mph).



Tuy nhiên, khi xét về độ nghiêm trọng tai nạn thì có thể thấy rằng tỉ lệ tai nạn nghiêm trọng và tai nạn chết người trên đường có tốc độ giới hạn > 50mph nhiều hơn so với trên đường có tốc độ giới hạn < 50mph

Graphical user interface, application, table

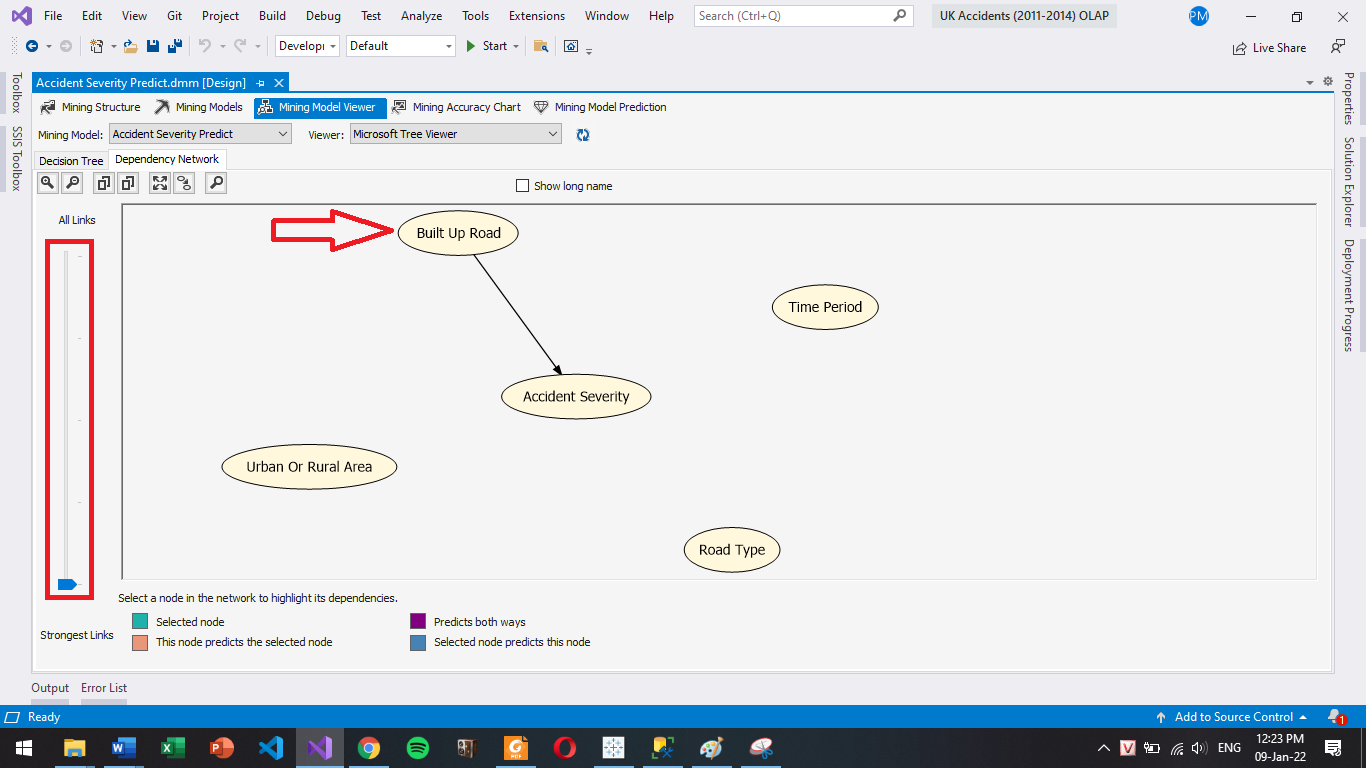
Description automatically generatedGraphical user interface, application, table

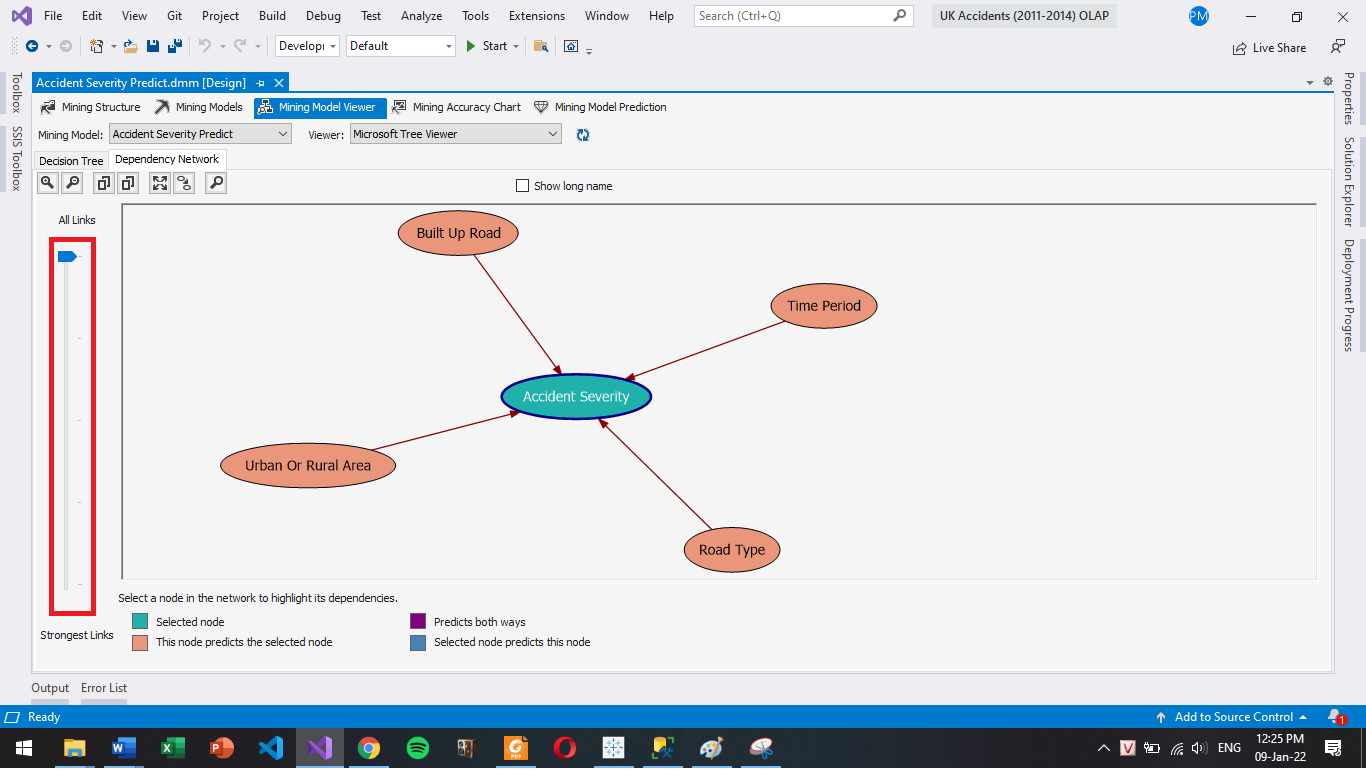
Description automatically generated

### Dependency Network

Sử dụng dependency network được tạo ra bởi model, ta có thể xác định xem yếu tố nào có ảnh hưởng (liên kết) mạnh hơn tới mức độ nghiêm trọng của tai nạn

Kéo thanh hiển thị các phụ thuộc trong hình, ta có thể xem model cho rằng yếu tố nào sẽ ảnh hưởng mạnh hơn tới mức độ nghiêm trọng của tai nạn





Theo như model, ta có các yếu tố ảnh hưởng (liên kết) mạnh với độ nghiêm trọng của tai nạn theo thứ tự như sau:

Built Up Road > Road Type > Time Period > Urban or Rural Area

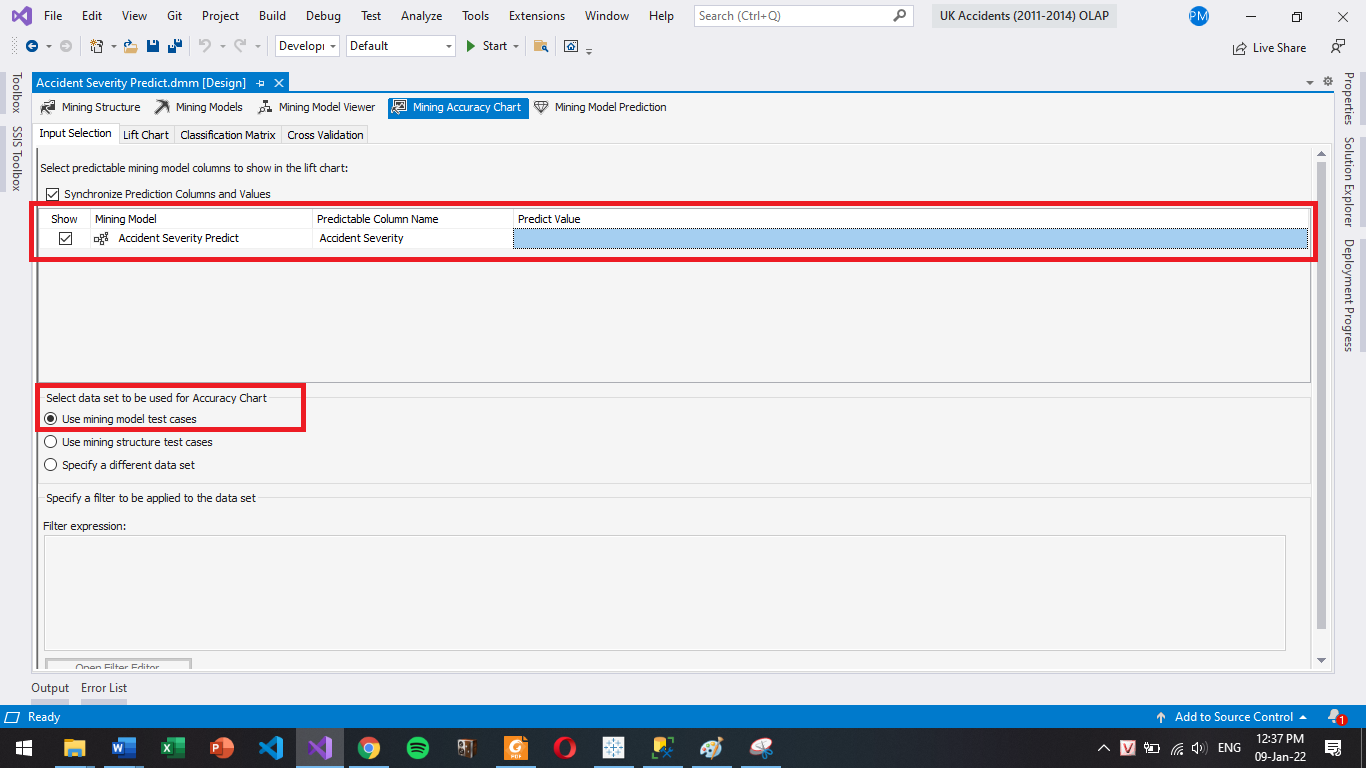
(Tốc độ giới hạn trên đường > Loại đường > Buổi xảy ra tai nạn > Nơi xảy ra tai nạn là Vùng đô thị hay Nông thôn )

### Mining Accuracy Chart

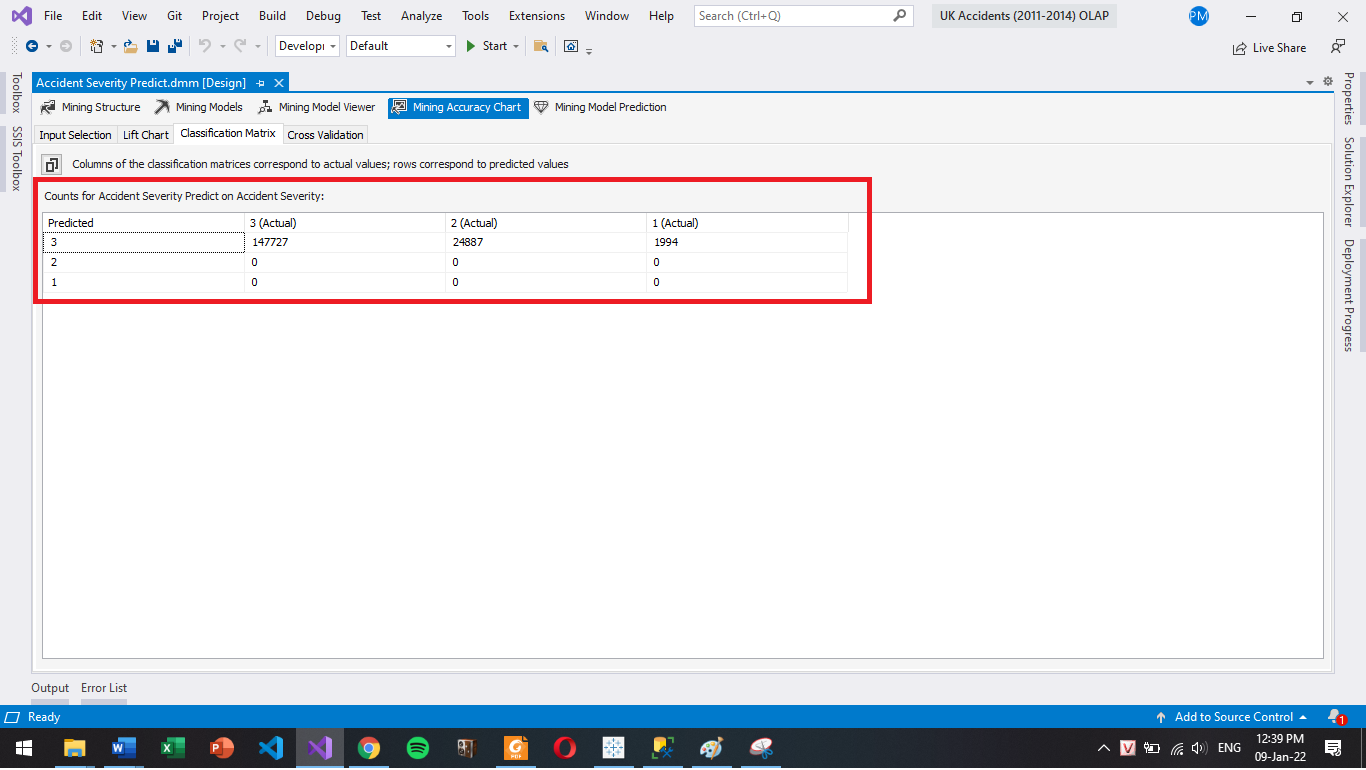
#### - Chọn Input

Chọn Predicttable Column Name là Accident Severity và Predict Value là 1, 2, 3 (tương ứng Fatal – Chết người, Serious – Nghiêm trọng, Slight – Nhẹ) để dự đoán độ nghiêm trọng của tai nạn.

Chọn sử dụng bộ dữ liệu test (30% tập dữ liệu gốc) để kiểm tra model.



#### - Classification Matrix



Model dự đoán toàn bộ độ nghiêm trọng là 3 (Slight). Trong đó số trường hợp dự đoán đúng là 147727, tổng số trường hợp đoán sai là 24887 + 1994 = 26881.

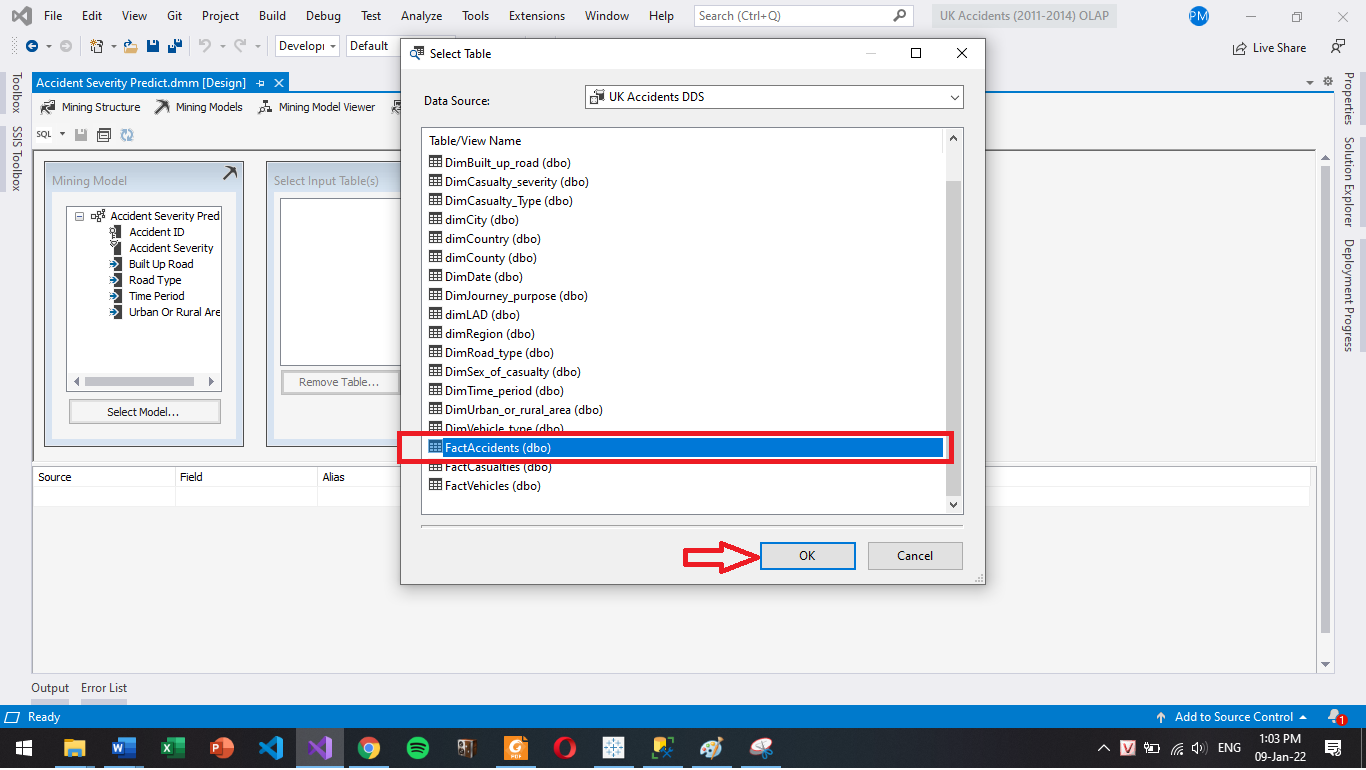
🡪 Dữ liệu dùng để train model có quá nhiều record có thuộc tính Accident Severity = 3 khiến cho model luôn dự đoán Accident Severity = 3 bất kể giá trị các thuộc tính input có là gì

### Model Mining Prediction

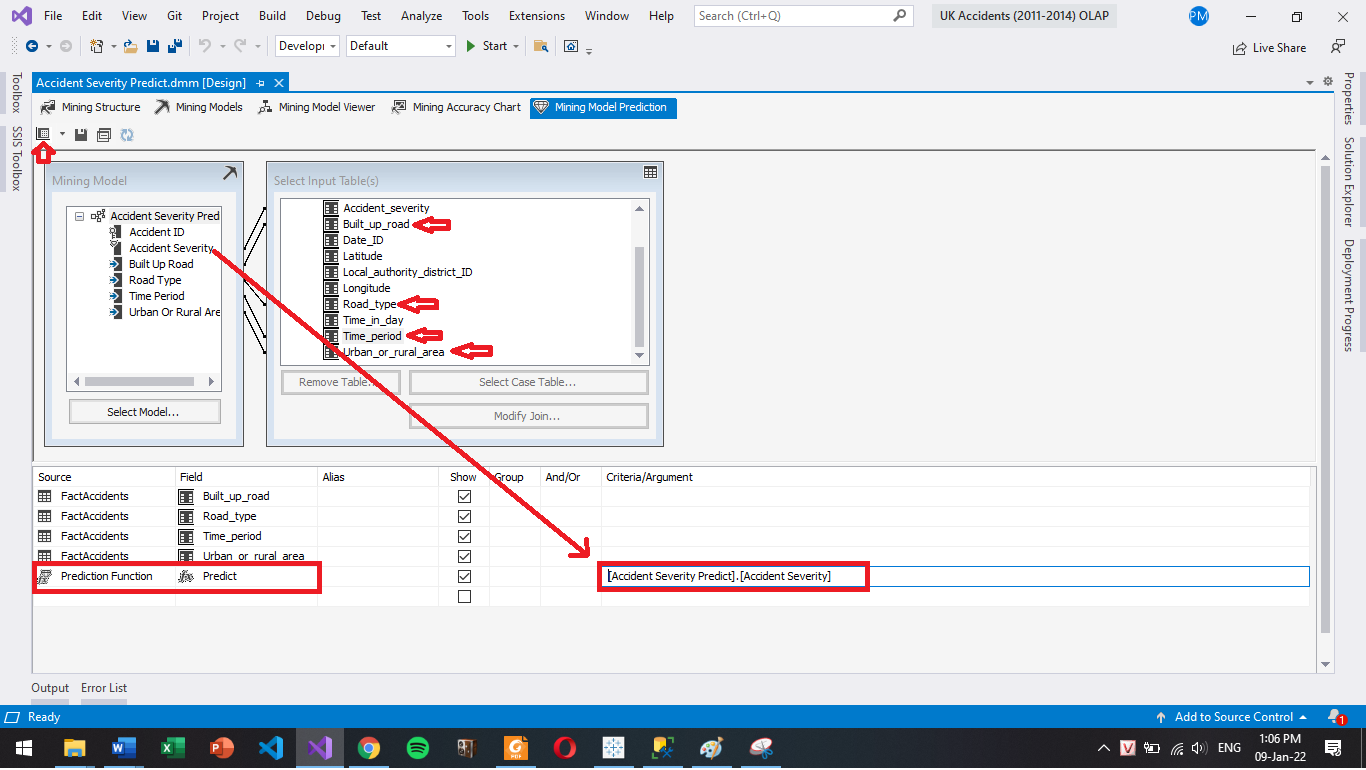
#### - Khởi tạo

Bước 1: Chọn tab “Model Mining Prediction” để sử dụng model dự đoán độ nghiêm trọng của tai nạn. Chọn “Select Case table” để chọn ra bảng dữ liệu input, vì không có dữ liệu mới nên ta sử dụng lại bảng FactAccidents





Bước 2: Chọn các thuộc tính input từ Case table. Chọn thuộc tính được dự đoán (tham chiếu từ model) bỏ vào mục Criteria/Argument.



#### - Kết quả

Chuyển sang Result View để xem kết quả dự đoán

Graphical user interface, Word

Description automatically generated

Model dự đoán toàn bộ Accident Severity là 3. Dường như model này sẽ không có nhiều lợi ích khi sử dụng nó để dự đoán mức độ nghiêm trọng của một tai nạn vì dữ liệu dùng để train model này không cân đối.

# Kết luận chung