**Báo Cáo Data Mining Đồ Án TH**

**1/ Xác định yêu cầu**

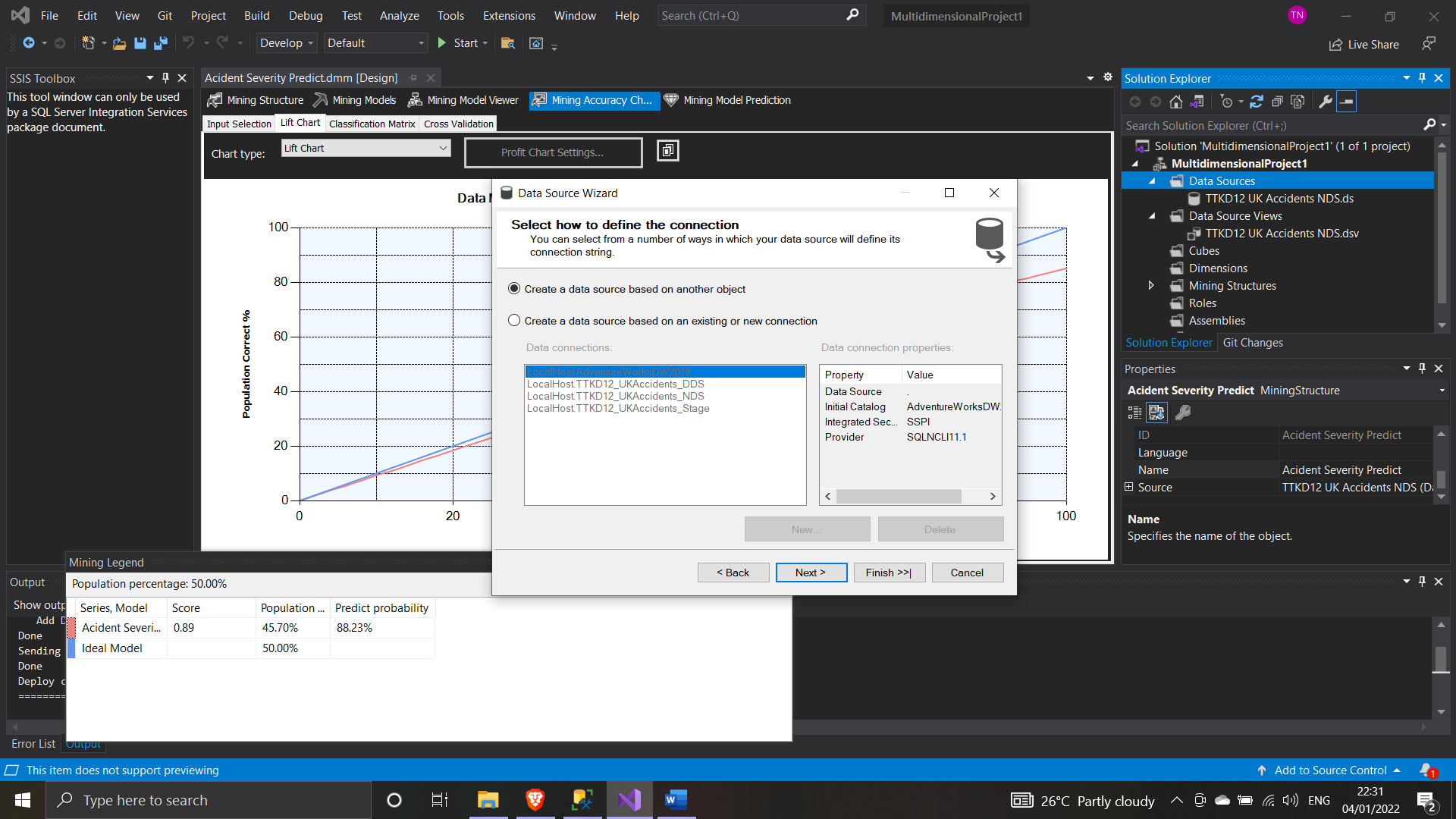
1/ Dự đoán xem mức độ nghiêm trọng của vụ tai nạn dựa trên các thuộc tính độc lập như LocalAuthorityDistrict, SpeedLimit, UrbanRural, NumberVehicles, NumberCasualties, RoadType.

2/ Dự đoán xem tai nạn xảy ra ở vùng nông thôn hay thành thị dựa trên các biến độc lập trích xuất từ DDS đã xây dựng. Hoặc dựa trên các biến độc lập được thêm vào trong bảng UrbanRural được thêm vào.

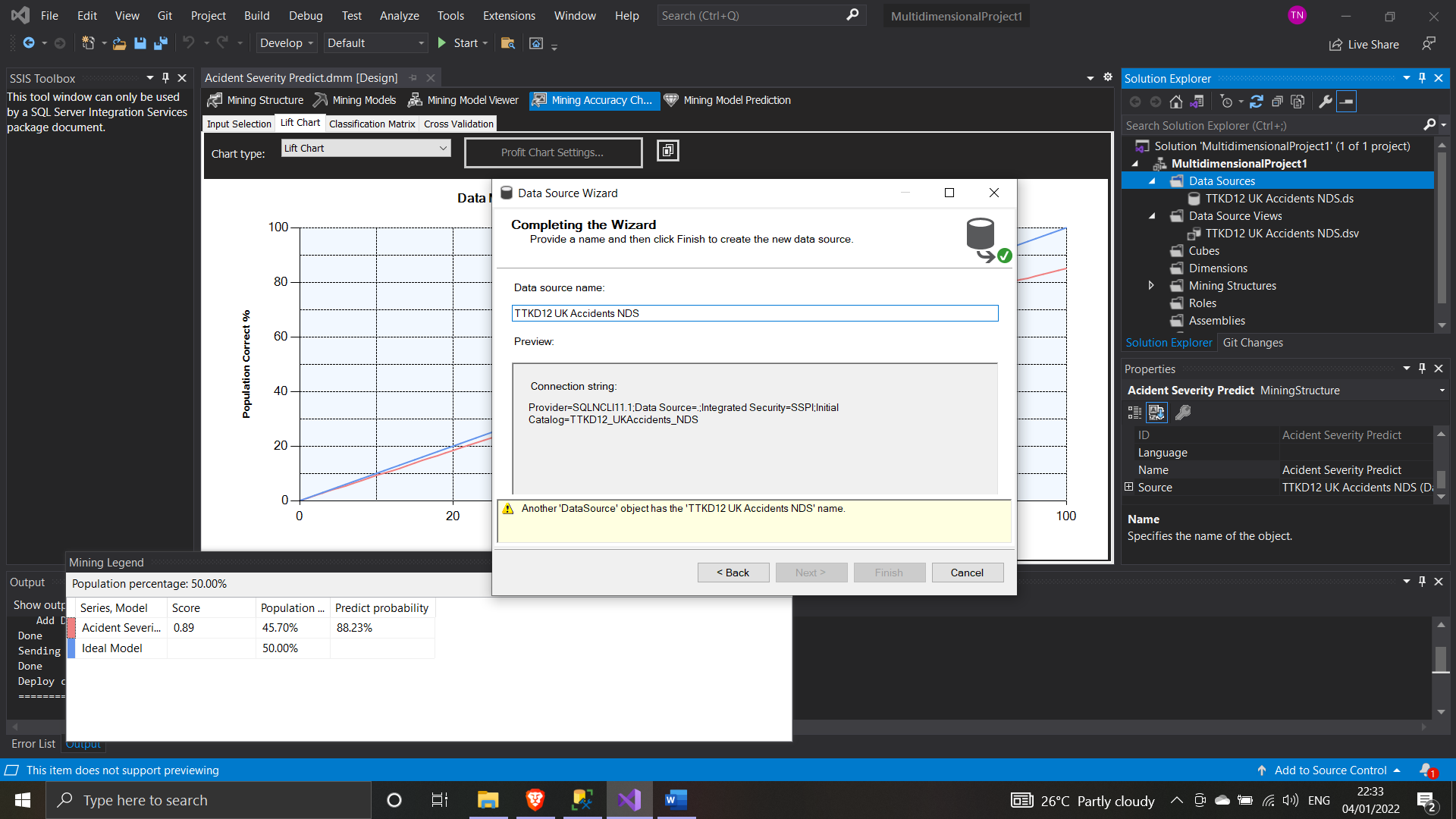
Hai thuật toán được sử dụng là thuật toán gom cụm (clustering) và cây quyết định (decision tree).

**2/ Xây dựng mô hình mining**

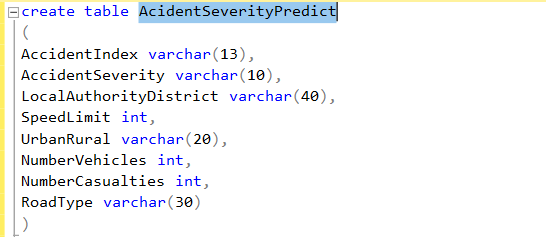
* Đầu tiên, chúng ta tạo một SSAS service trong VS2019 và click chuột phải vào data source để chọn tạo nguồn dữ liệu
* Vì lúc đầu, nhóm đã tạo một bảng khác và lưu vào một database để phục vụ cho nhu cầu mining (bảng đó không phải DDS) nên data source cần được chỉ đến database đó.



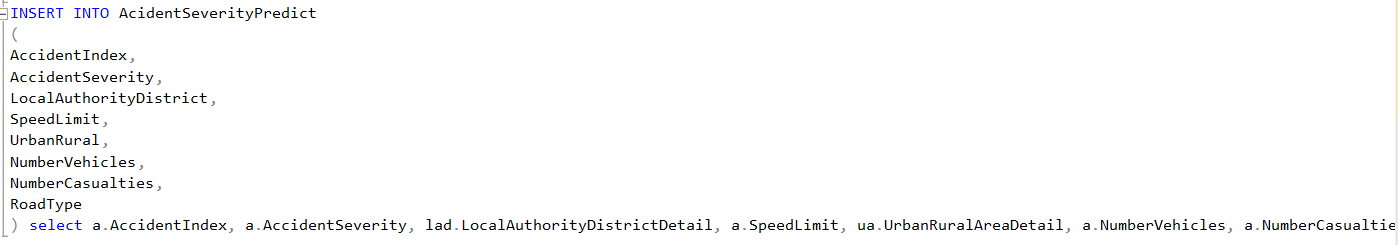
* Tạo đường dẫn đến Database và bấm chọn Finish để hoàn tất



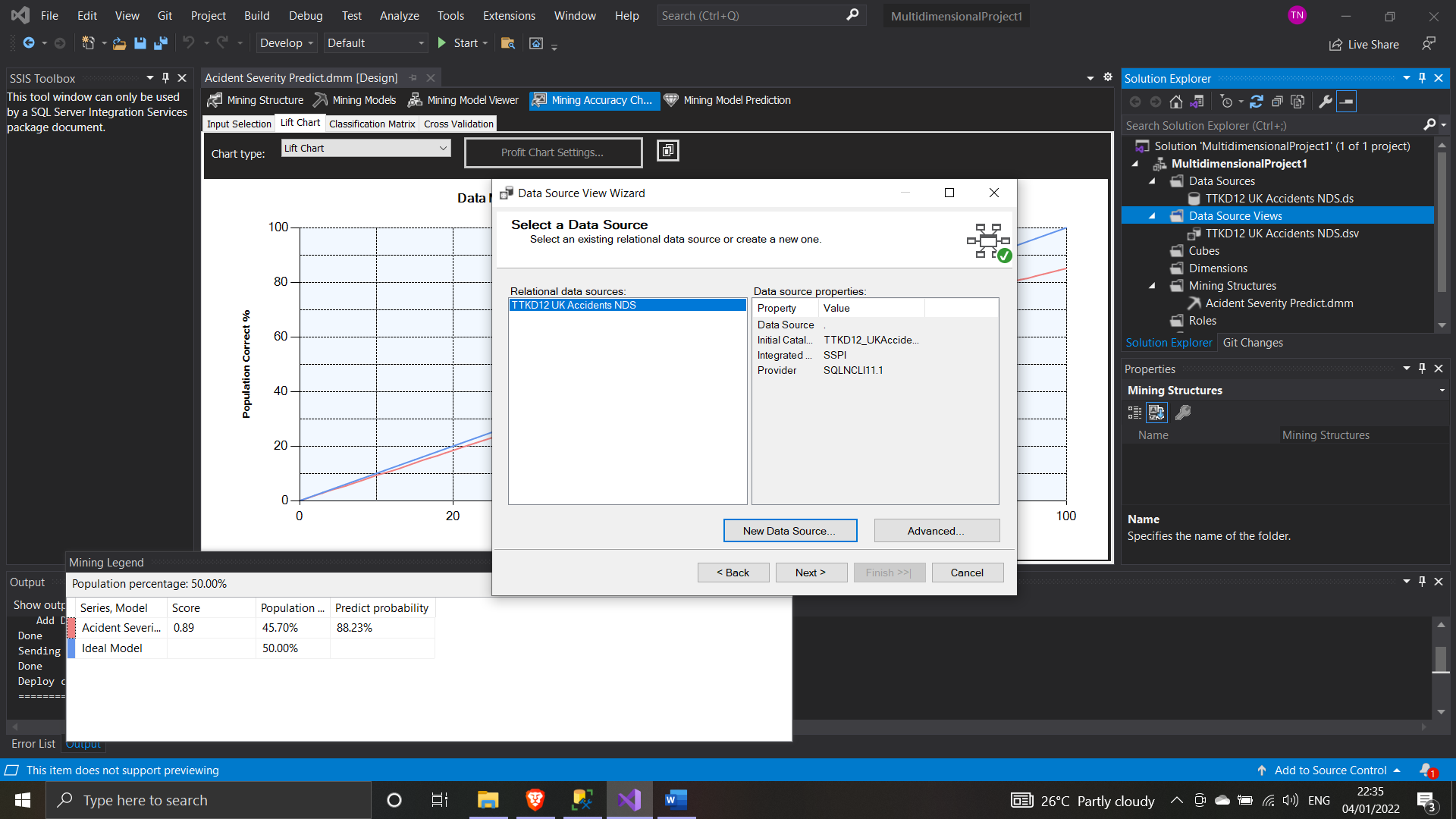
* Tiếp theo, chúng ta tạo một bảng AcidentSeverityPredict trong SQL Server với cấu trúc như sau:



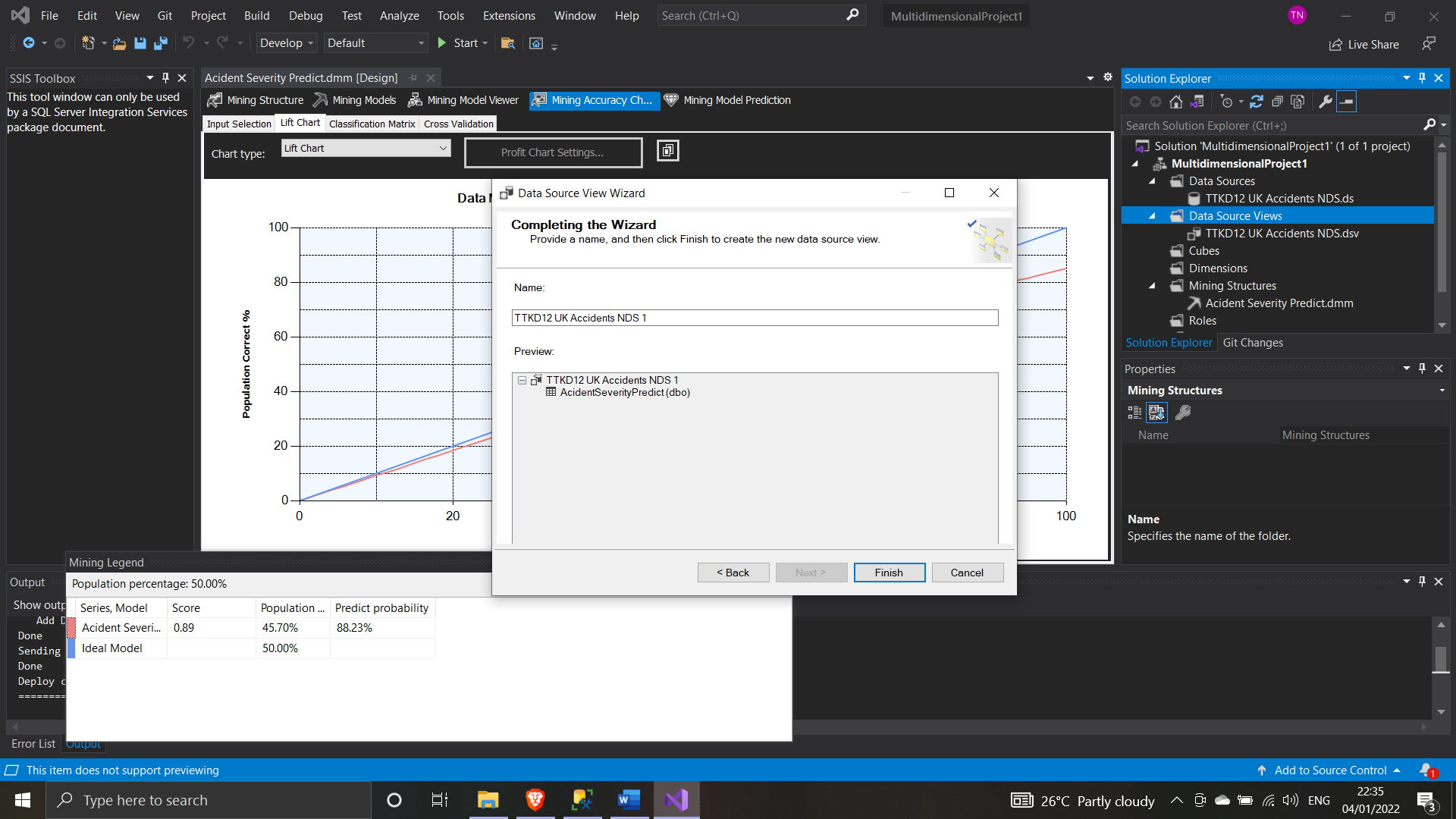
* Sau đó, tiến hành đổ dữ liệu cần thiết vào bảng vừa tạo như sau:



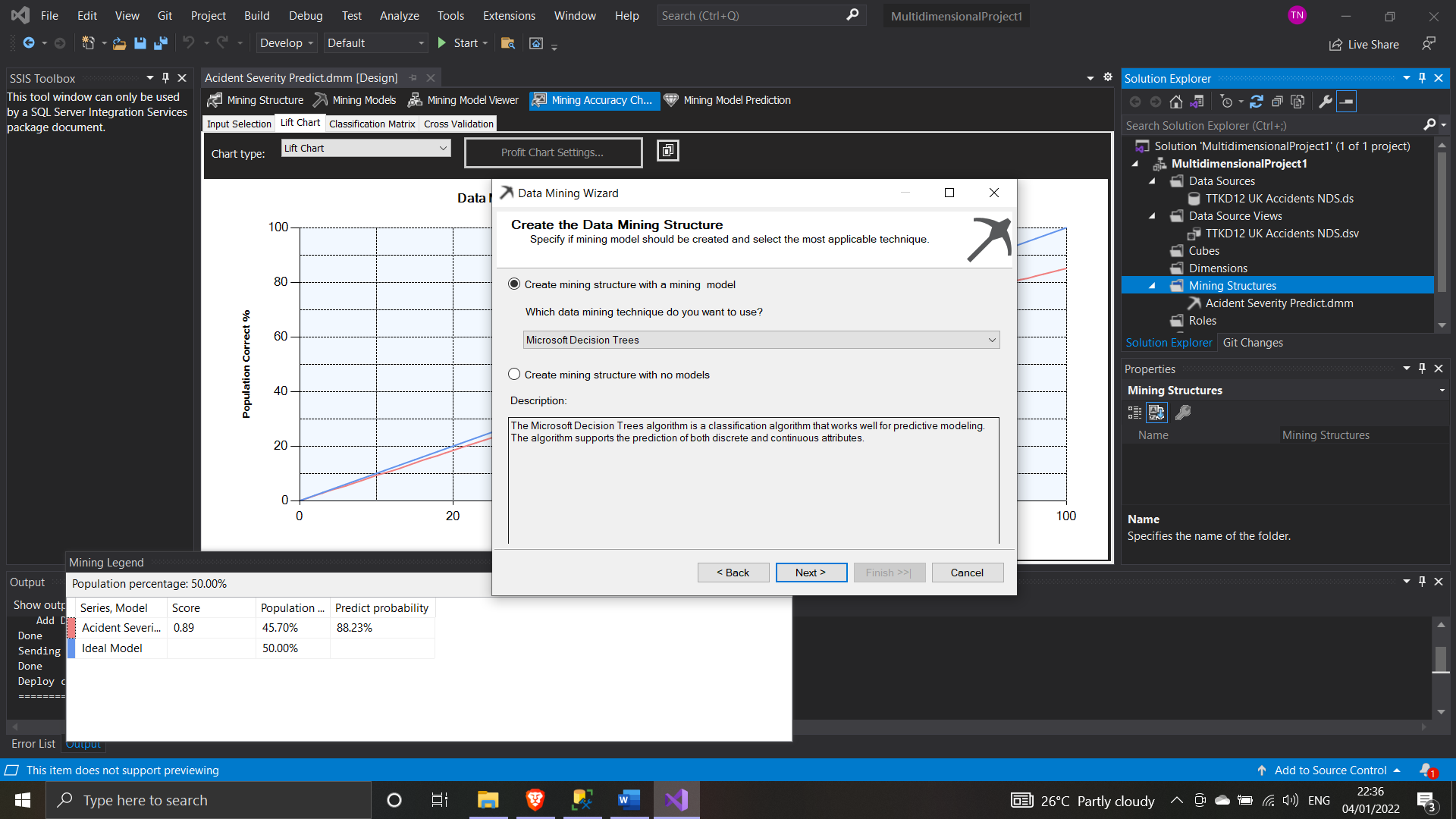
* Tiếp đến, chúng ta cần tạo một Data Source View để View dữ liệu của bảng AcidentSeverityPredict. Click chuột phải vào Data Source View để tạo mới



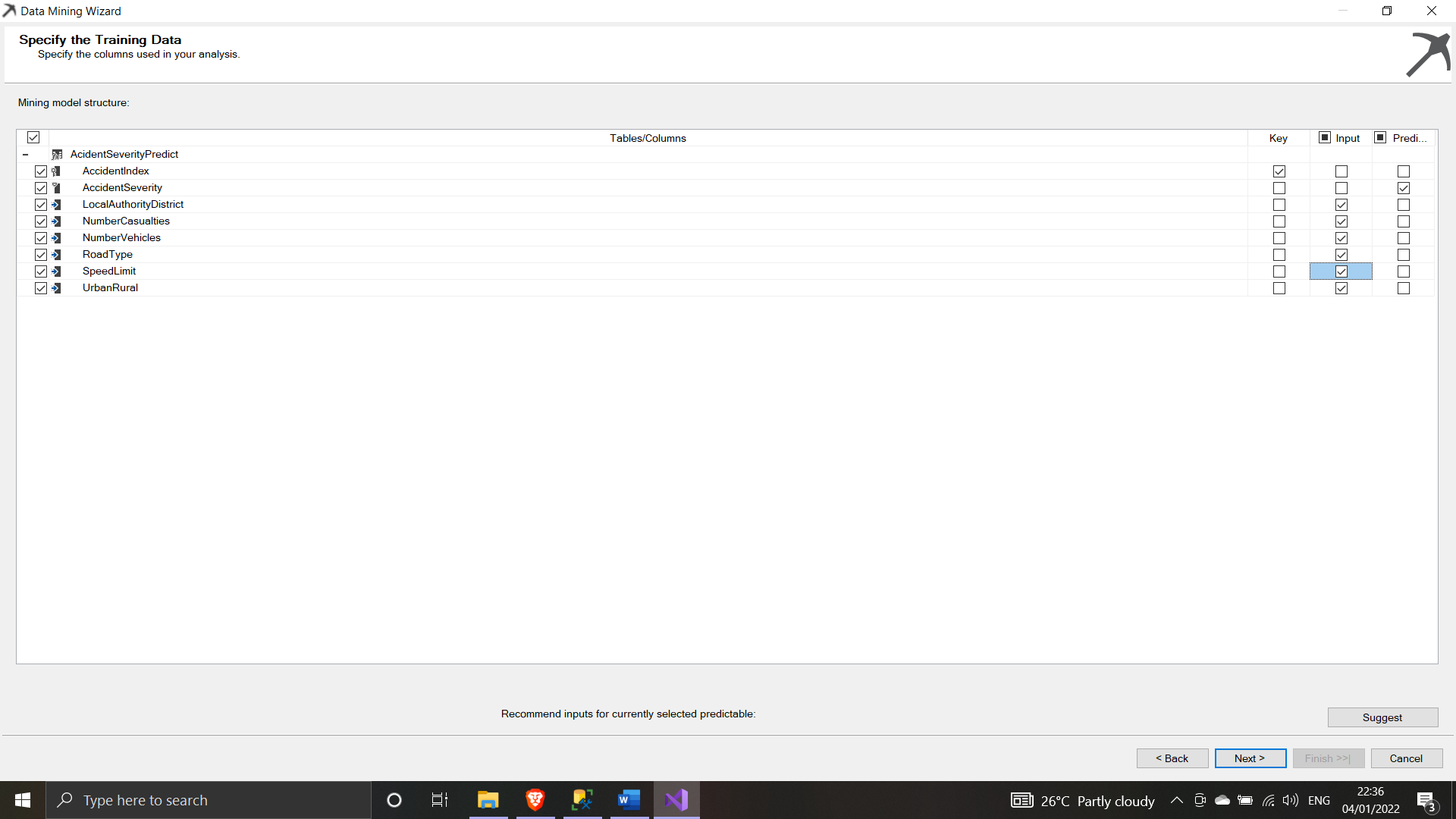
* Sau đó, trỏ đến bảng AcidentSeverityPredict. Chọn Finish



* Bây giờ chúng ta sẽ tạo Mining Structure. Đầu tiên ta cần xác định thuật toán. Thuật toán được chọn cho mô hình là Decision Tree.



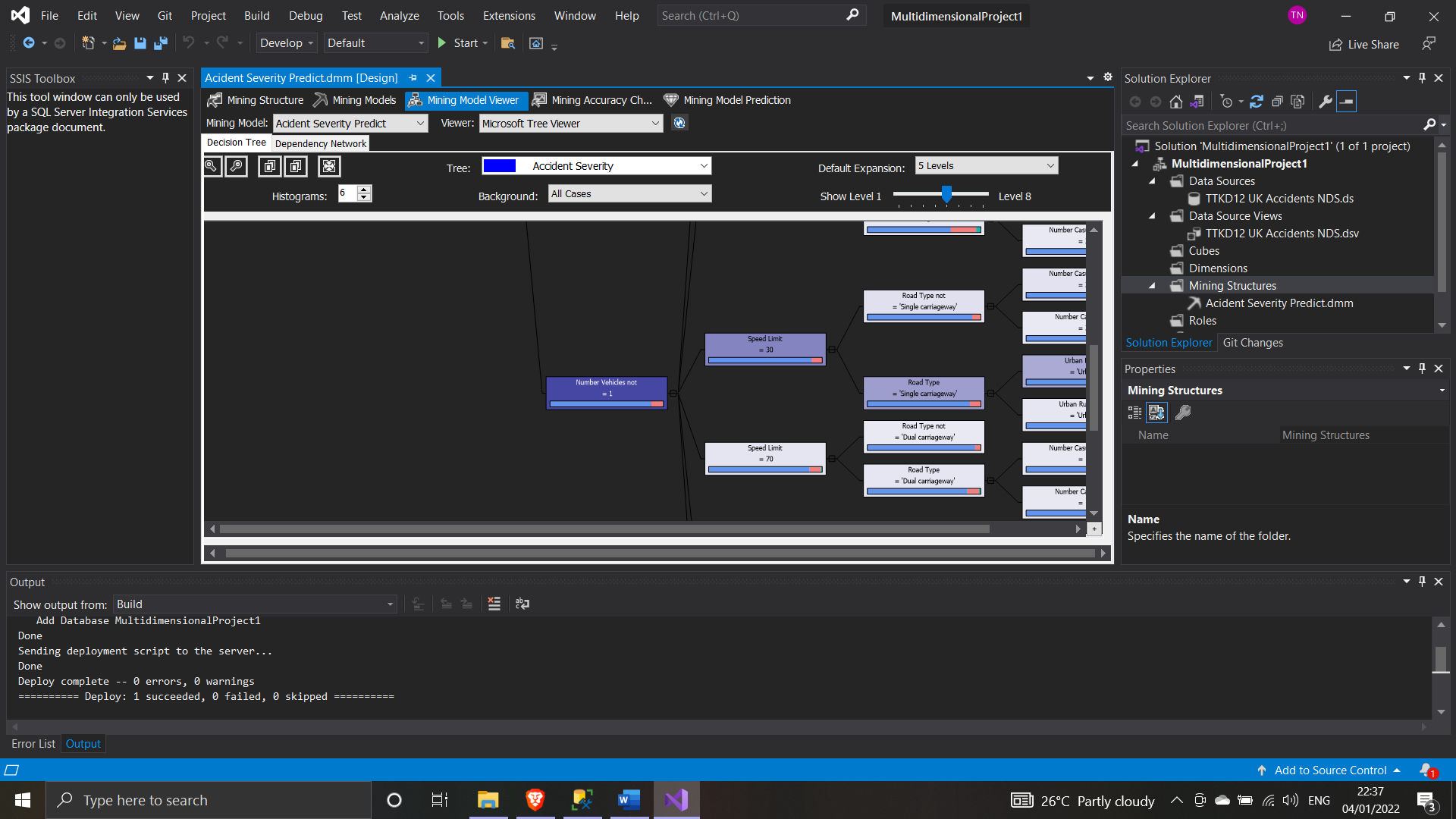
* Xác định thuộc tính cần Predict là thuộc tính nào. Ở đây ta chọn AccidentSeverity. Còn các thuộc tính khác sẽ là thuộc tính Input.



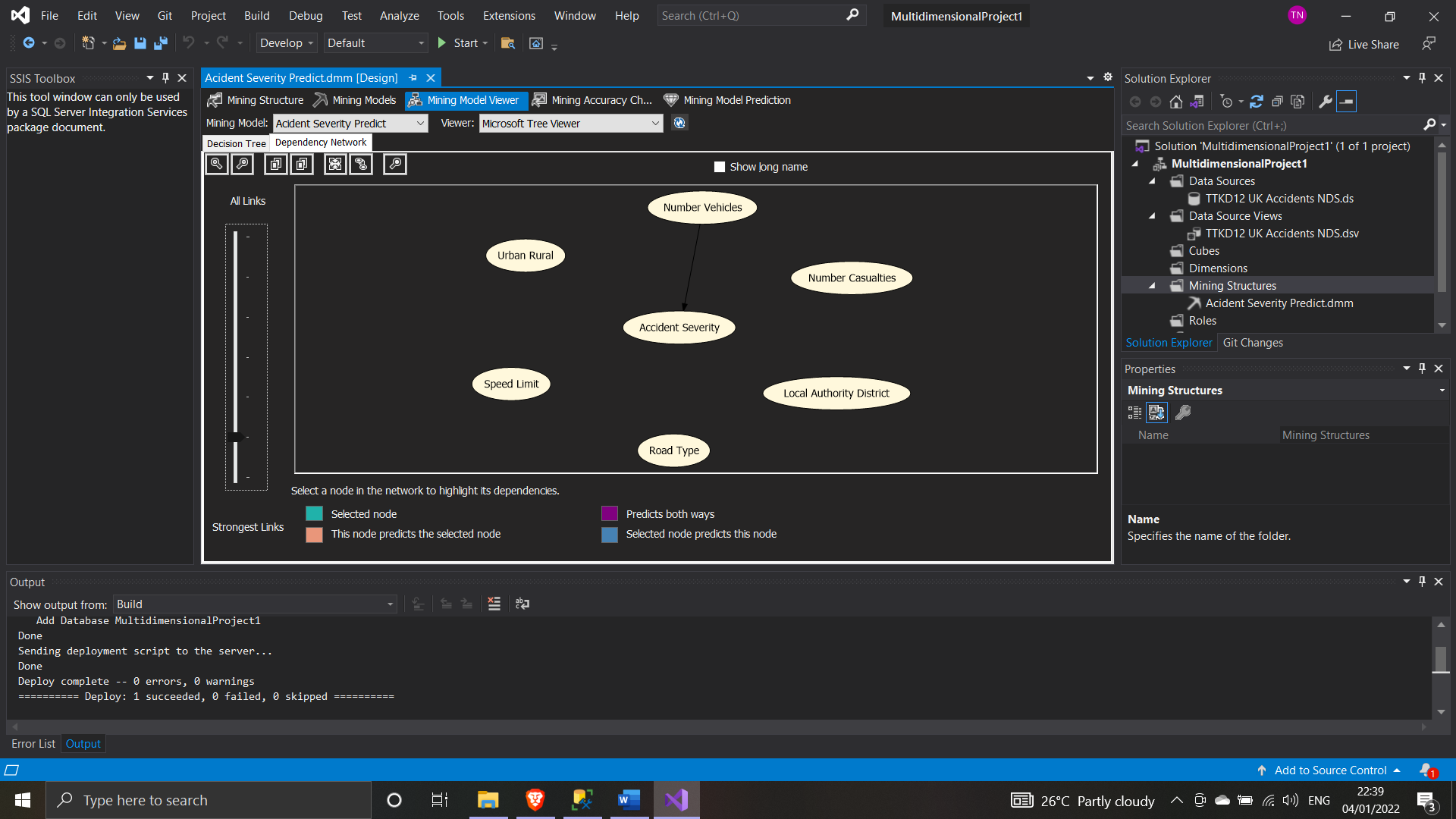
* Cuối cùng, ta xác định phần trăm dữ liệu dùng cho tập test và độ lớn của mỗi test case. Sau đó bấm finish.

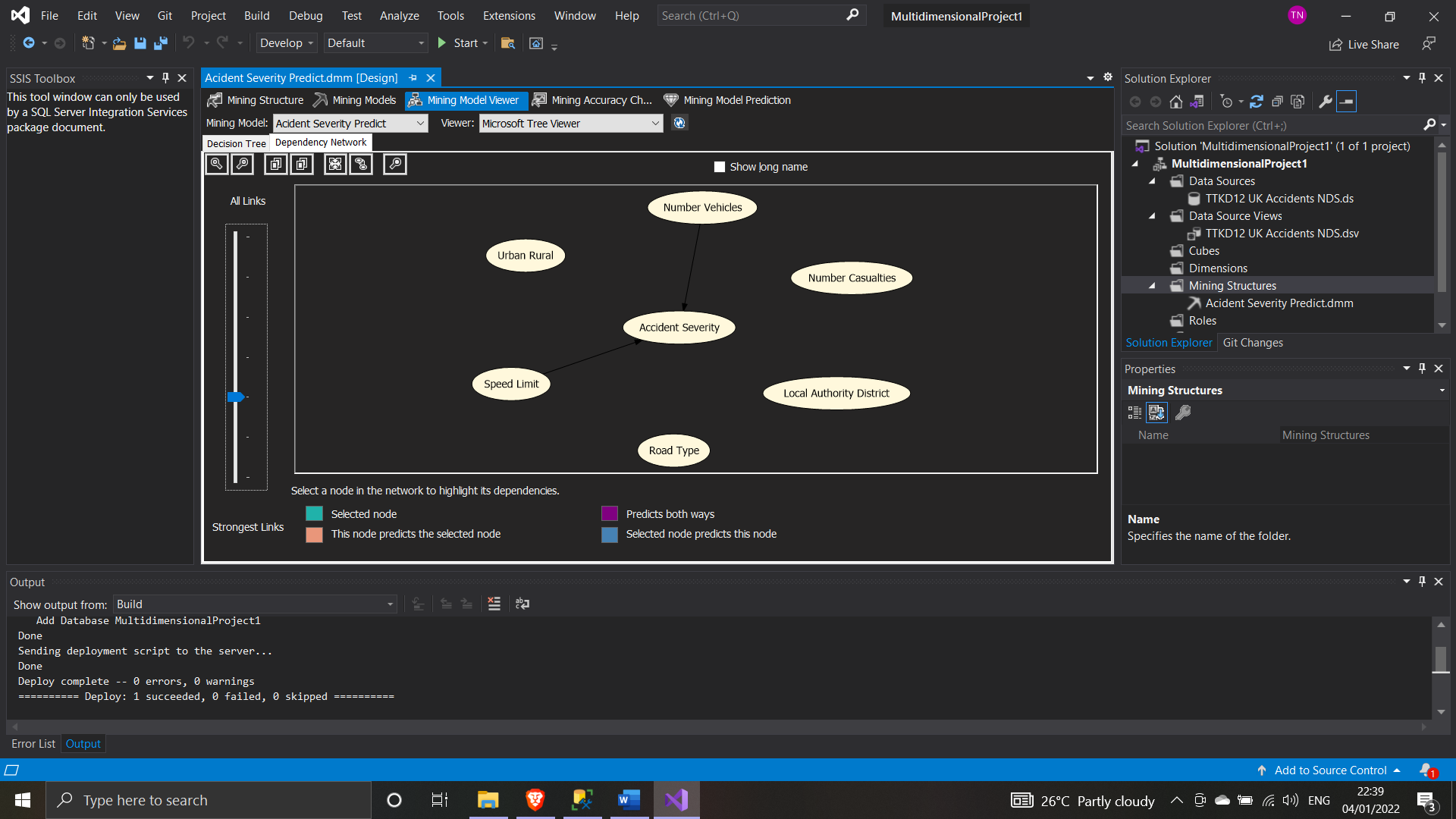


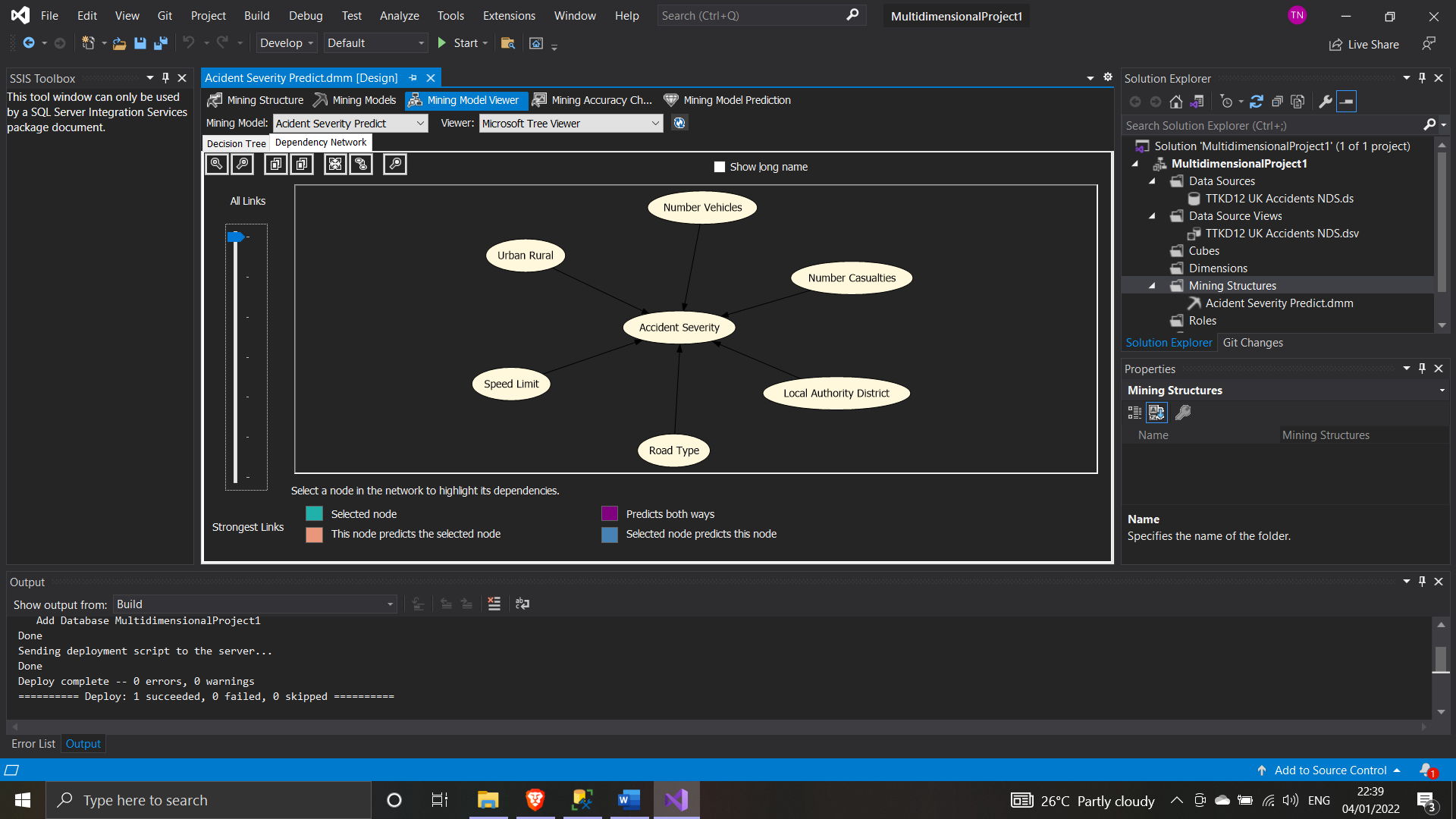
* Sau đó, ta chọn vào tab Mining Model Viewer và tiến hành build mô hình. Đây là kết quả thu được sau khi build thành công.



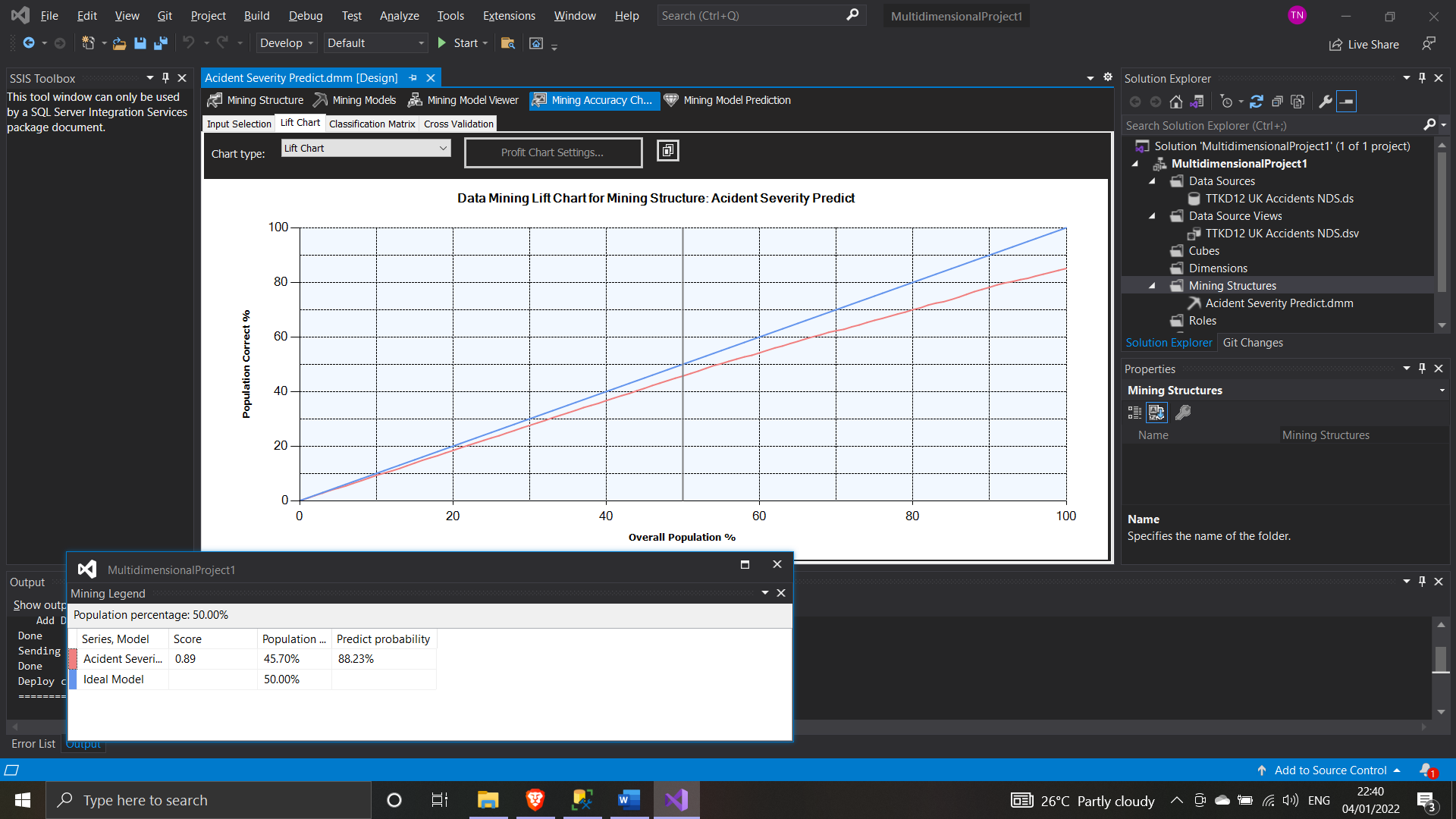
* Chuyển sang tab Denpendancy Network để xem sự phụ thuộc. Ta thấy được rằng số lượng phương tiện tham gia sẽ có ảnh hưởng lớn nhất, tiếp đến là Giới hạn tốc độ, và theo sau là các thuộc tính còn lại



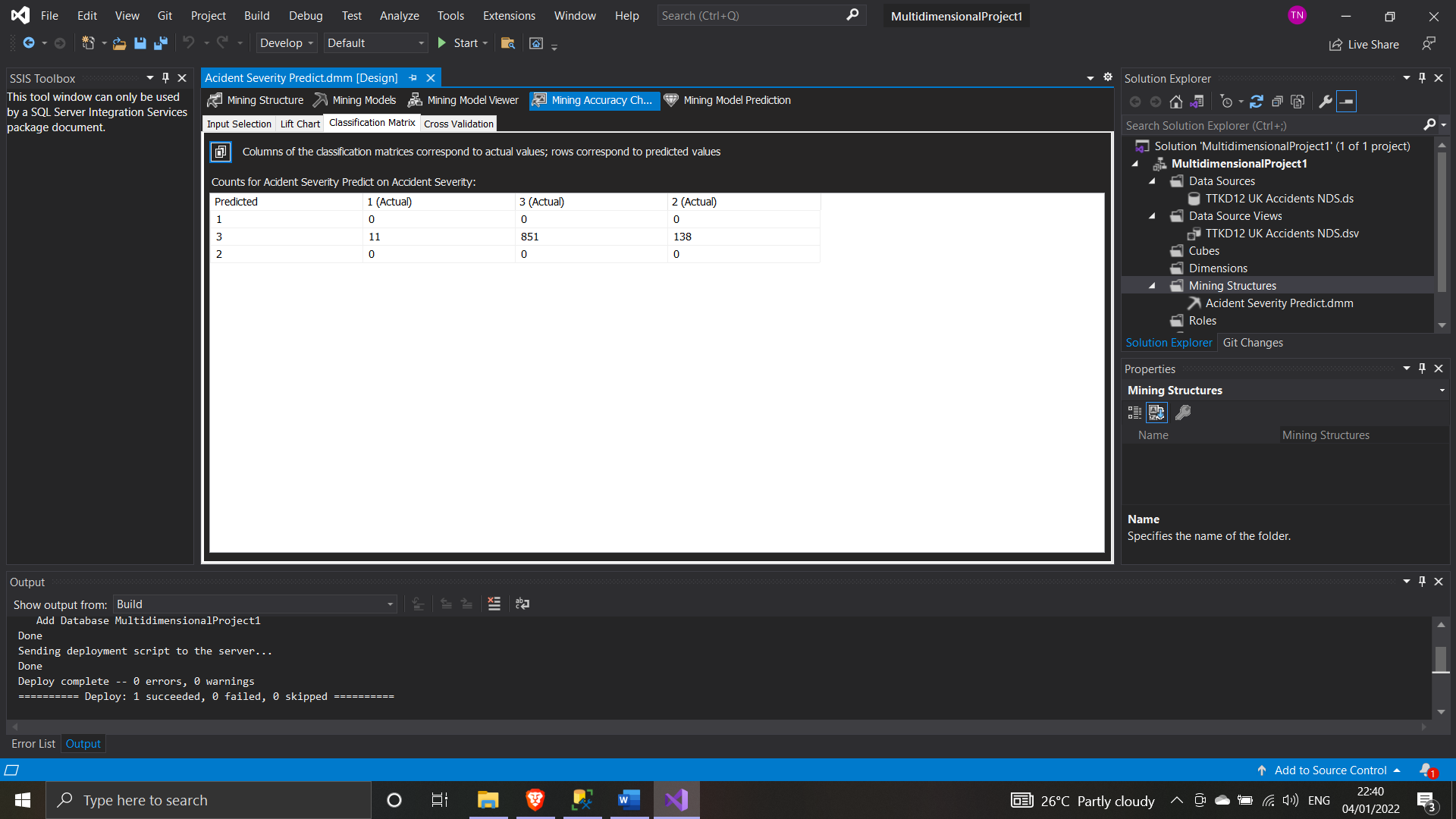




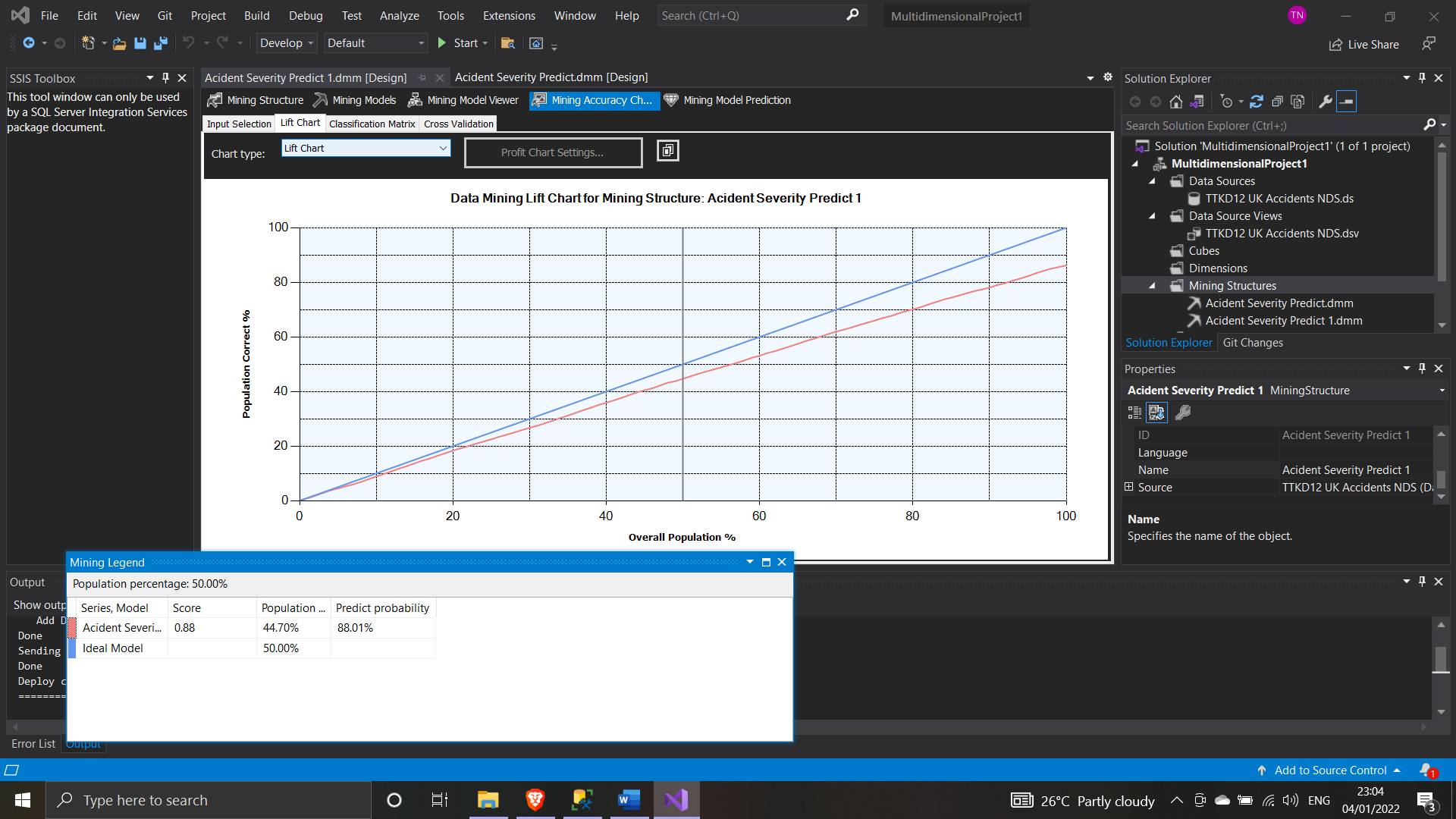
* Tab Mining Accuracy Chart sẽ cho ta biết được độ tin cậy của dự đoán mà ta vừa xây dựng. Ở đây, độ tin cậy của dự đoán mà mô hình này đạt được là 88.23%

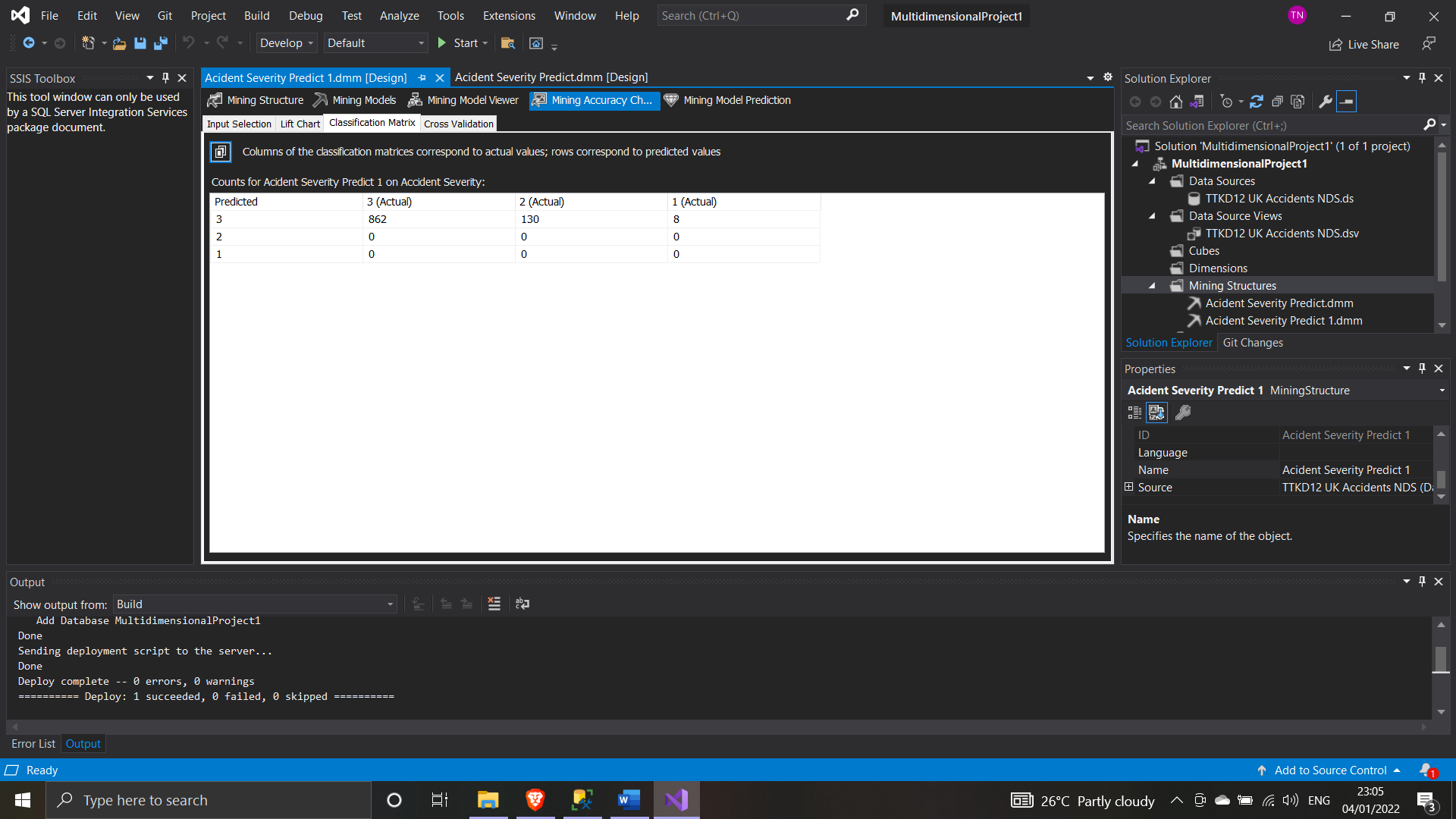


* Ta có thể hiểu ma trận phân loại này như sau: với giá trị được dự đoán là 3 thì thực tế, có 11 giá trị là 1, 851 giá trị là 3 và 138 giá trị là 2. Do đó, phần trăm phân loại chính xác của thuật toán này là 851/ (851+138+11) = 0.851 = 85.1%. Tương tự cho mô hình xây dựng theo thuật toán clustering là 86.2%



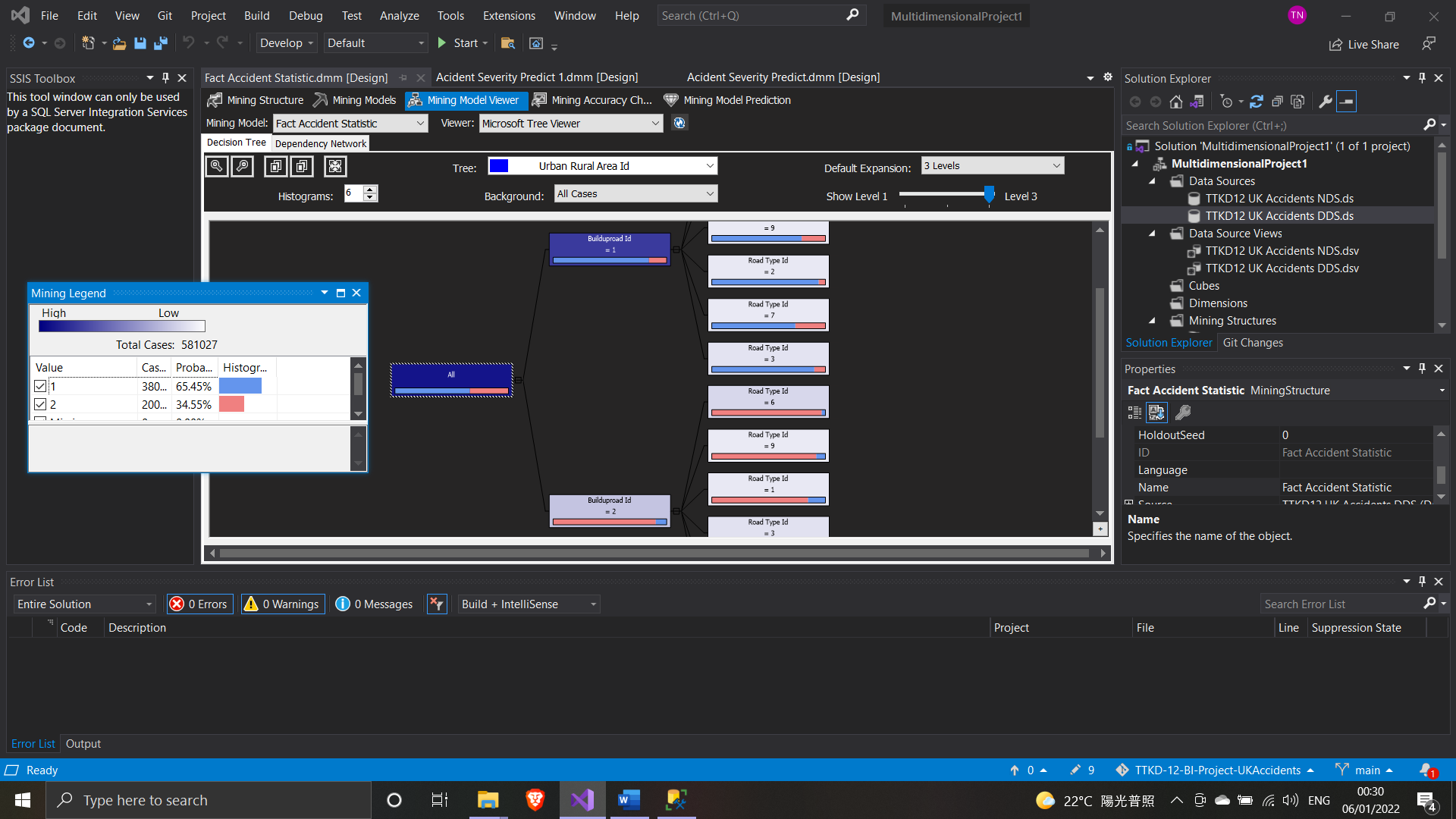
* Tương tự như trên, ta xây dựng mô hình mining theo thuật toán Clustering. Độ tin cậy của dự đoán mà mô hình Clustering đạt được là 88.01%



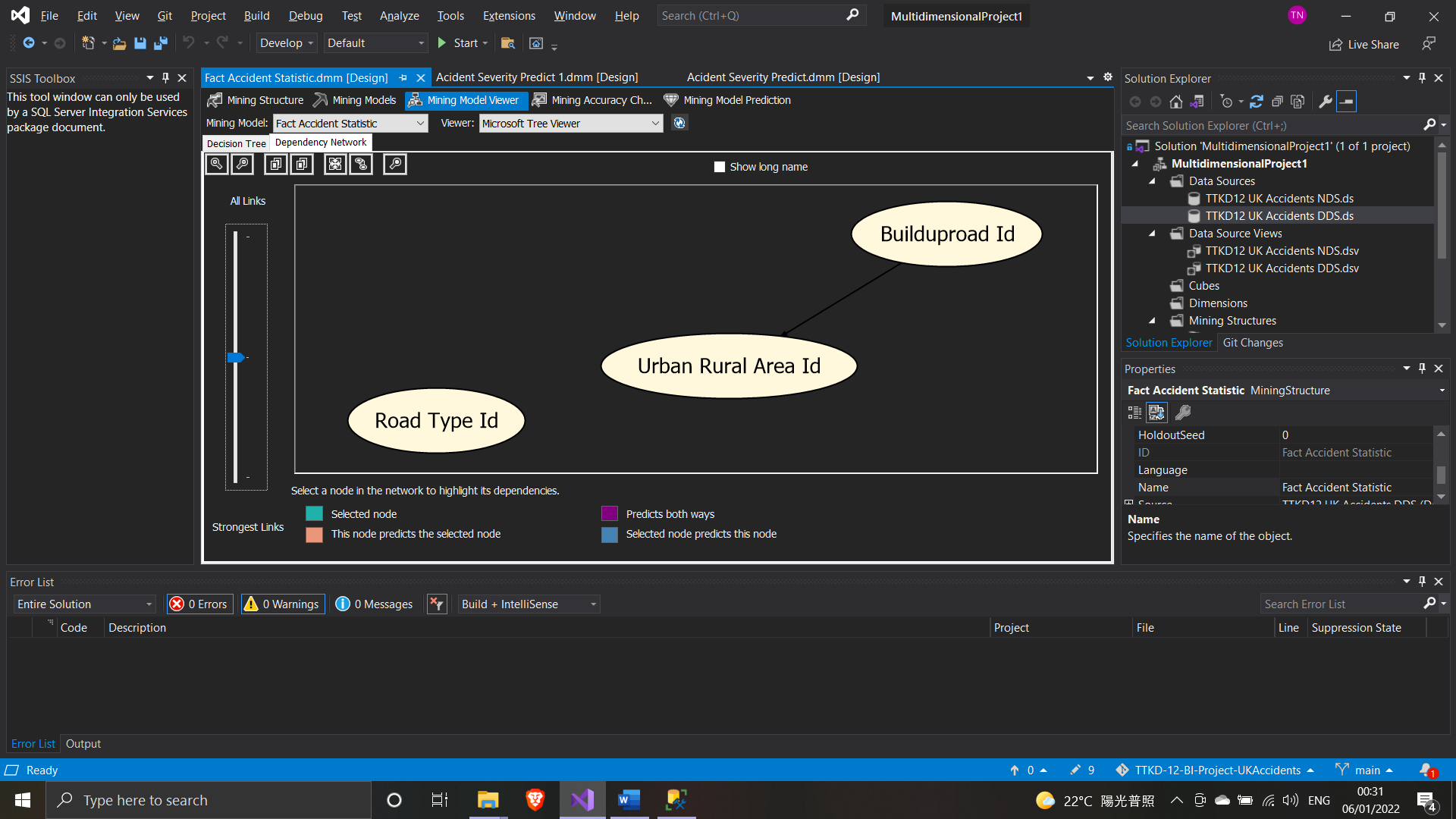


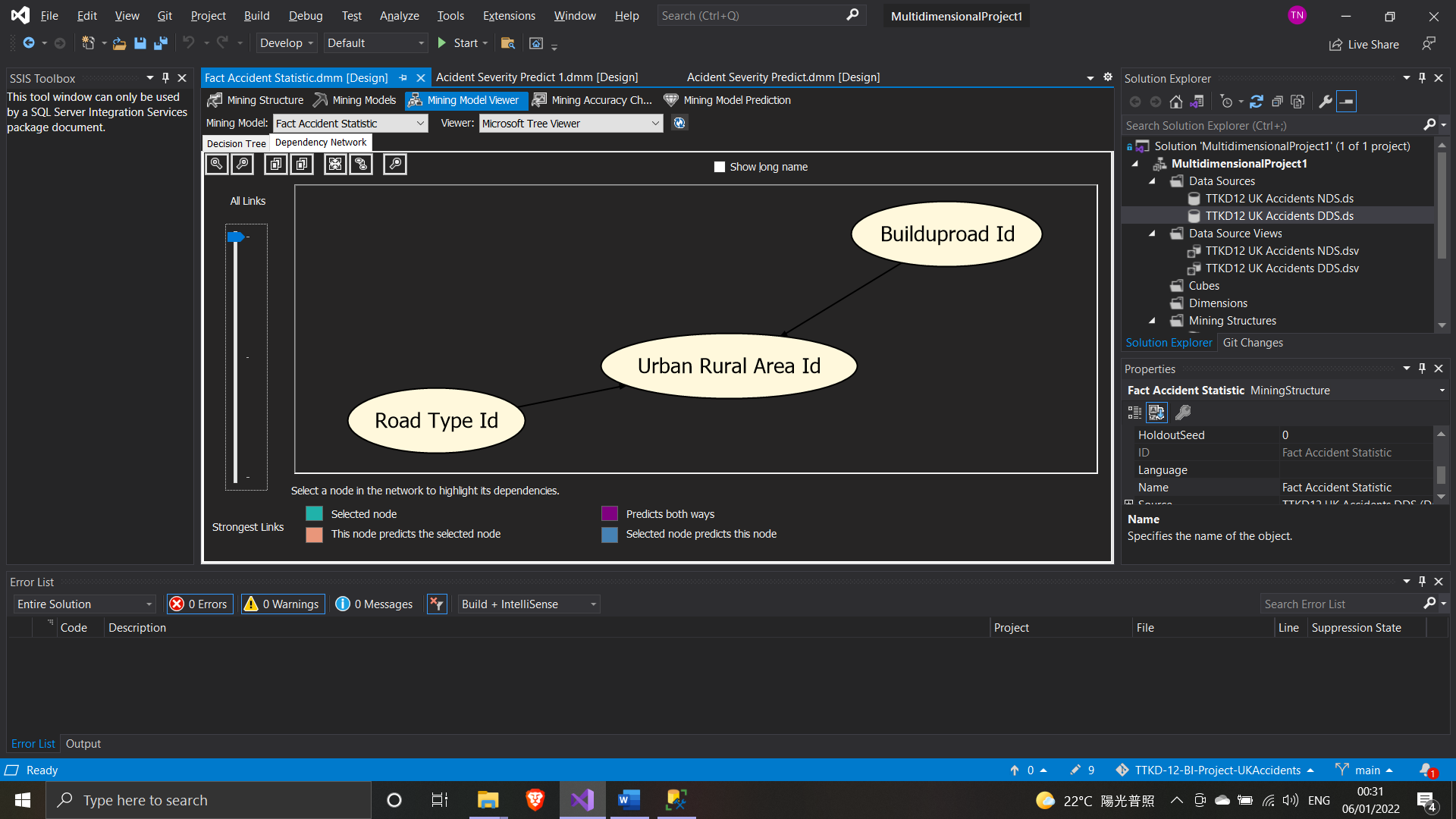
|  |  |  |  |
| --- | --- | --- | --- |
| Loại thuật toán | Độ tin cậy của dự đoán | Lift Score | Độ chính xác |
| Decision Tree | 88.23% | 0.89 | 85.1% |
| Clustering | 88.01% | 0.88 | 86.2% |

* Tương tự ta thực hiện cho mô hình dự đoán lần 2.

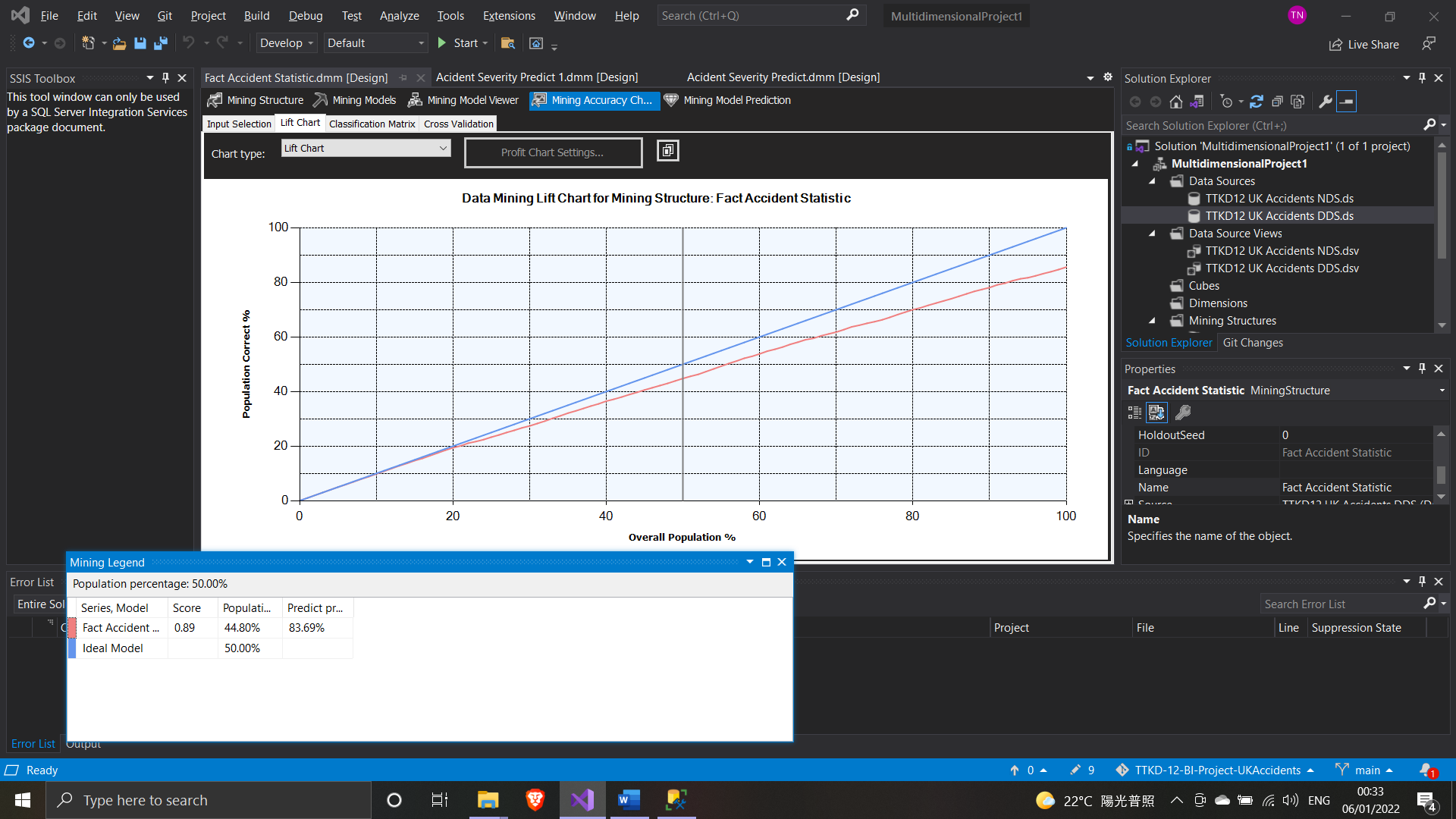


* Denpendancy Network của yêu cầu 2. Tại đây ta thấy việc xác định tai nạn xảy ra ở nông thôn hay thành thị phụ thuộc khá lớn vào biến độc lập BuildUpRoad (có liên quan đến tốc độ giới hạn). Tiếp theo đó là phụ thuộc vào biến RoadType.

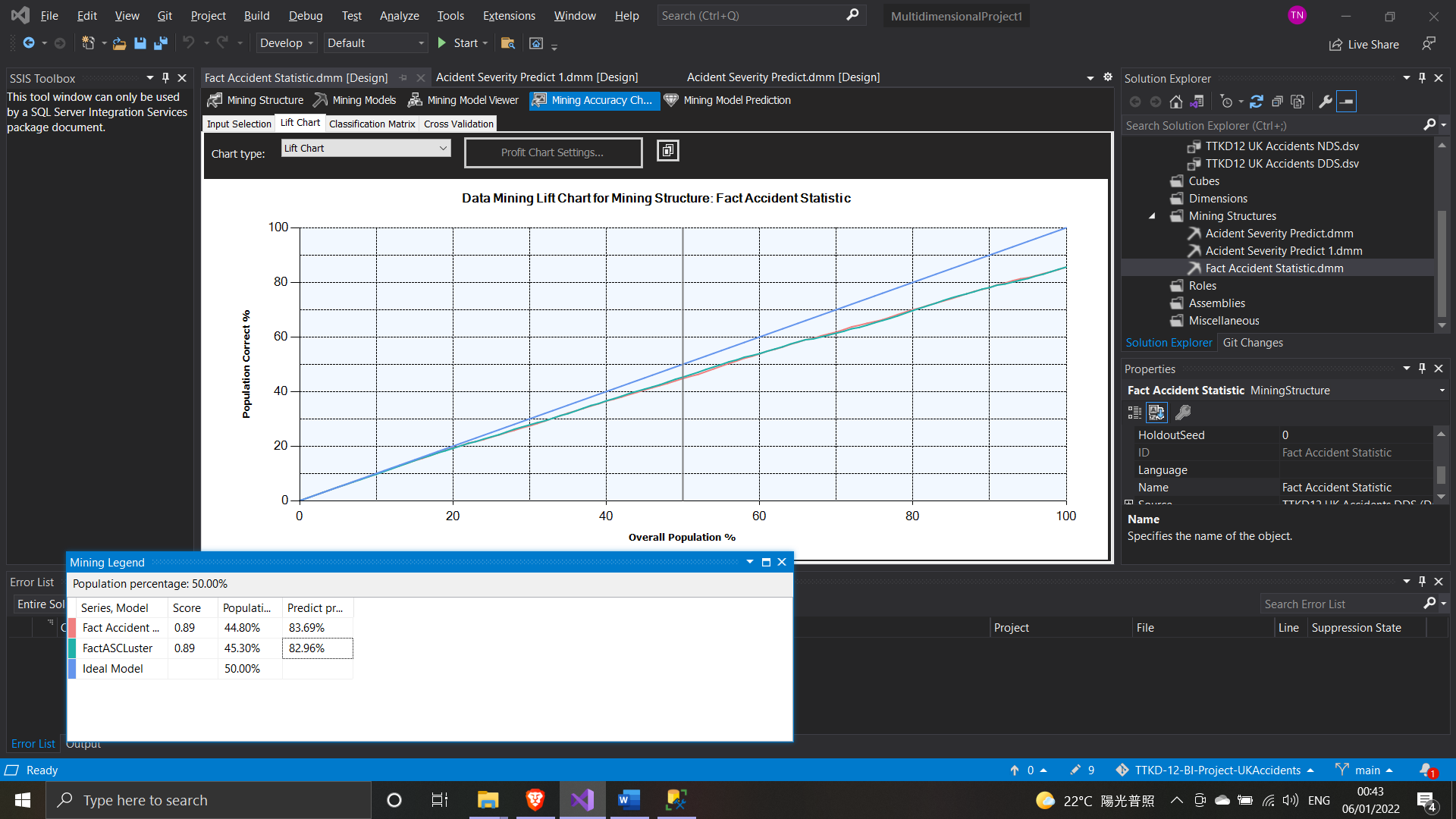




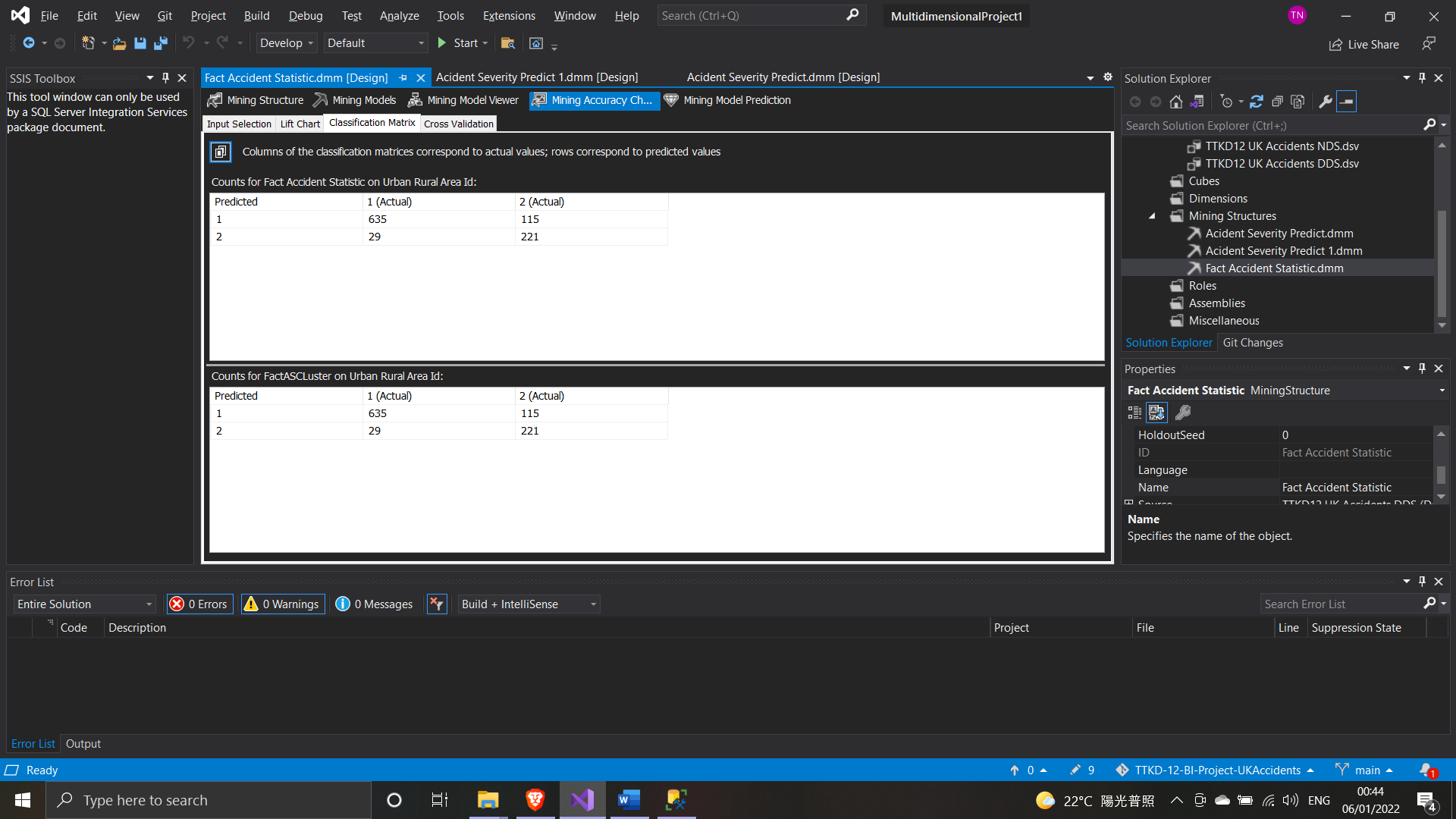
* Mô hình này độ tin cậy là 83.69% và có được lift score là 0.89



* Độ chính xác của mô hình này là (635+221)/1000 = 0.856 = 85.6%.
* Với mô hình được xây dựng theo thuật toán clustering, độ tin cậy đạt 82.96%, thấp hơn một chút so với mô hình xây dựng bằng thuật toán Decision Tree.

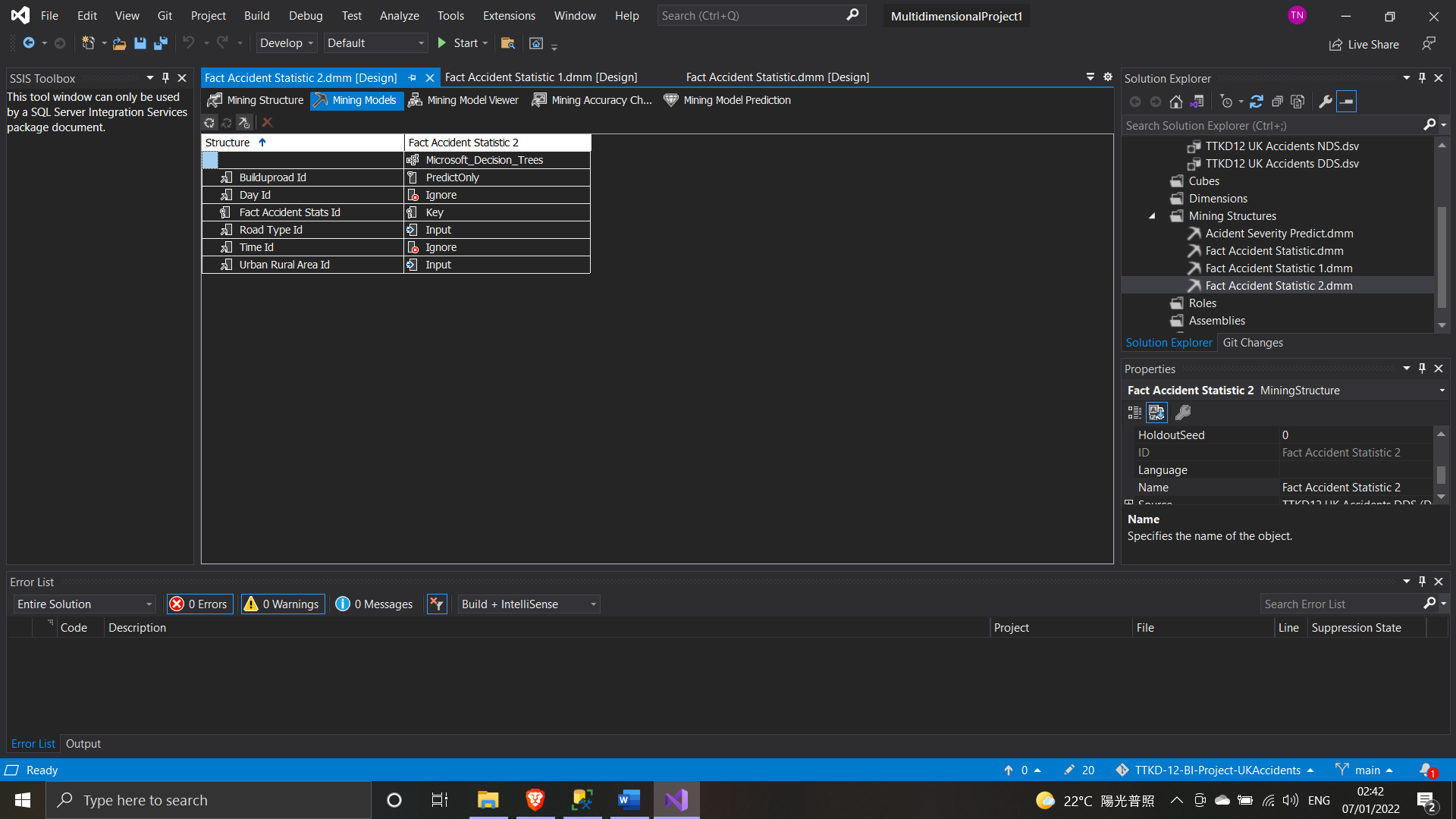


* Tuy nhiên, độ chính xác đạt tương tự với mô hình decision tree.

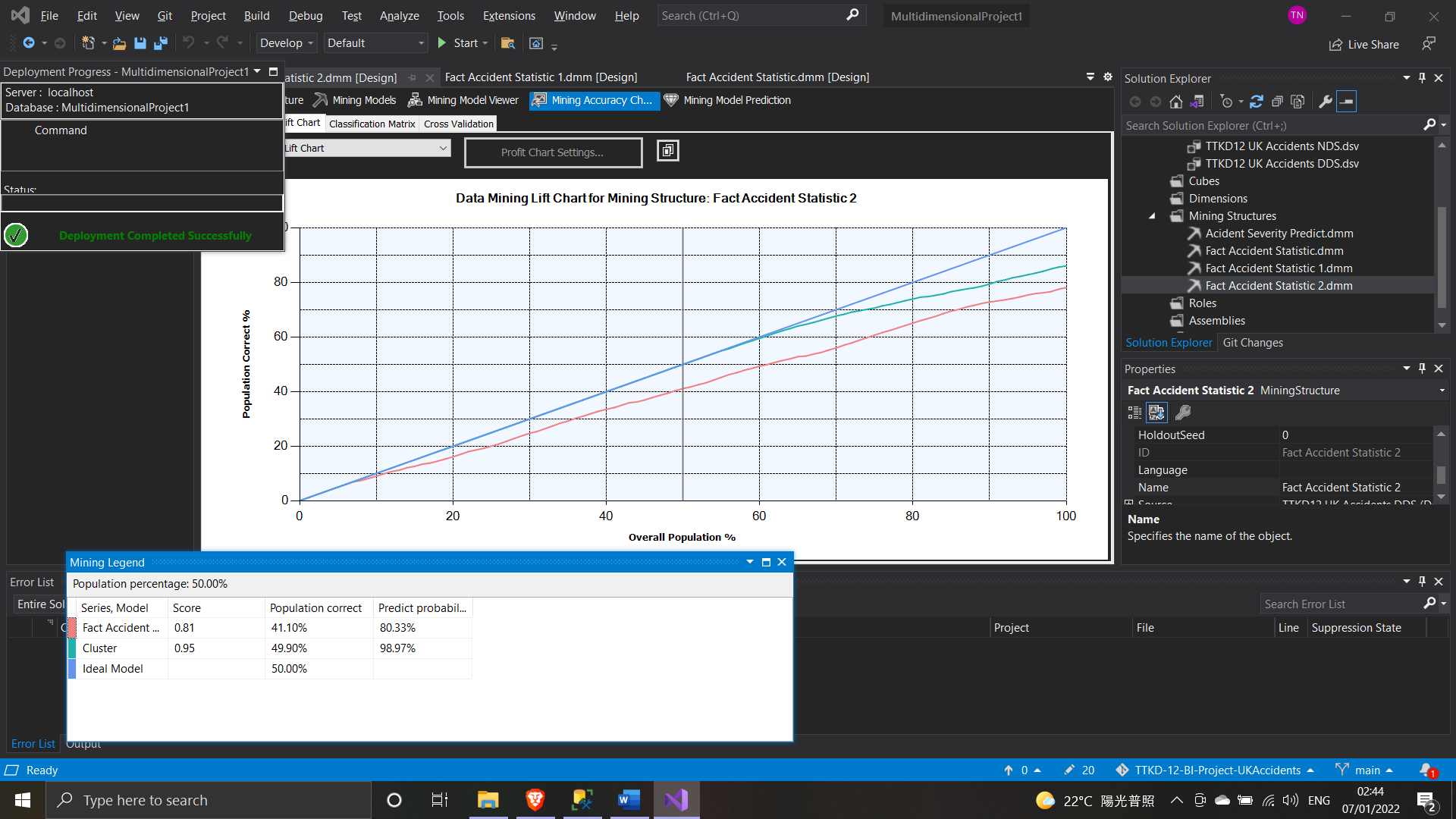


|  |  |  |  |
| --- | --- | --- | --- |
| Loại thuật toán | Độ tin cậy của dự đoán | Lift Score | Độ chính xác |
| Decision Tree | 83.69% | 0.89 | 85.6% |
| Clustering | 82.96% | 0.89 | 85.6% |

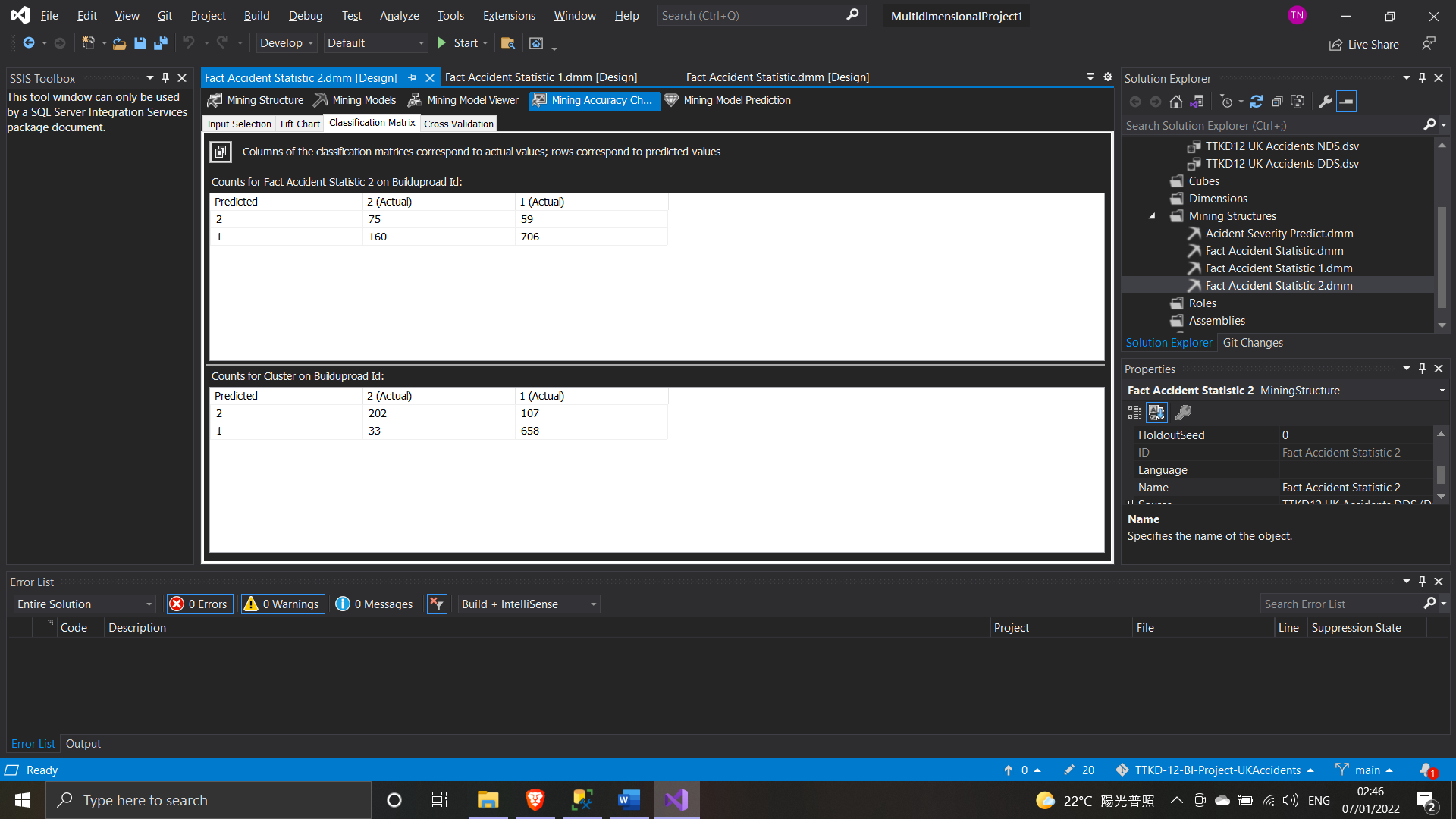
* Bây giờ, chúng ta sẽ tiến hành xây dựng một mô hình dự đoán khác với yêu cầu là dự đoán xem đoạn đường xảy ra tai nạn có đang được xay dựng hay không dựa trên loại đường và khu vực xảy ra tai nạn.



* Tiến hành xem độ tin cậy của dự đoán. Độ tin cậy mà ta có được với thuật toán Decision Tree là 80.33% và với thuật toán Clustering là 98.97%. Điểm của chúng lần lượt là 0.81 và 0.95



* Tiếp theo ta sẽ check ma trận phân loại để tính độ chính xác của mô hình dự đoán. Với thuật toán Decision Tree, độ chính xác của mô hình là 78.1% và mô hình Clustering là 86%.



|  |  |  |  |
| --- | --- | --- | --- |
| Loại thuật toán | Độ tin cậy của dự đoán | Lift Score | Độ chính xác |
| Decision Tree | 80.33% | 0.81 | 78.1% |
| Clustering | 98.97% | 0.95 | 86% |

**Hết**