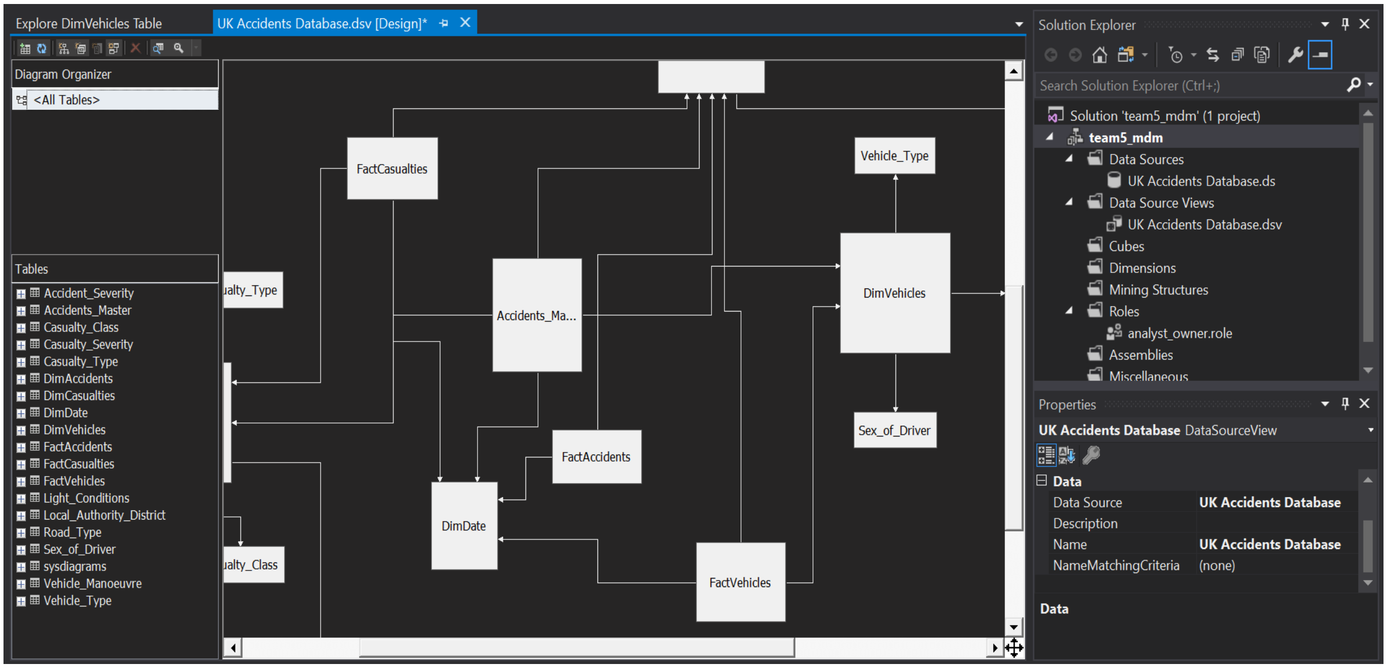
1. Connection to analysis service database

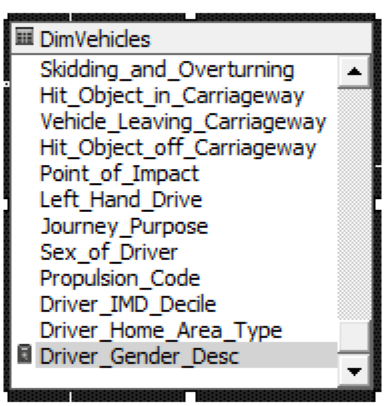
Create a data connection to the UK_Accidents_Database database on stwssbsql01.ad.okstate.edu and create a Data Source View that has all the tables in the UK_Accidents_Database relational database.

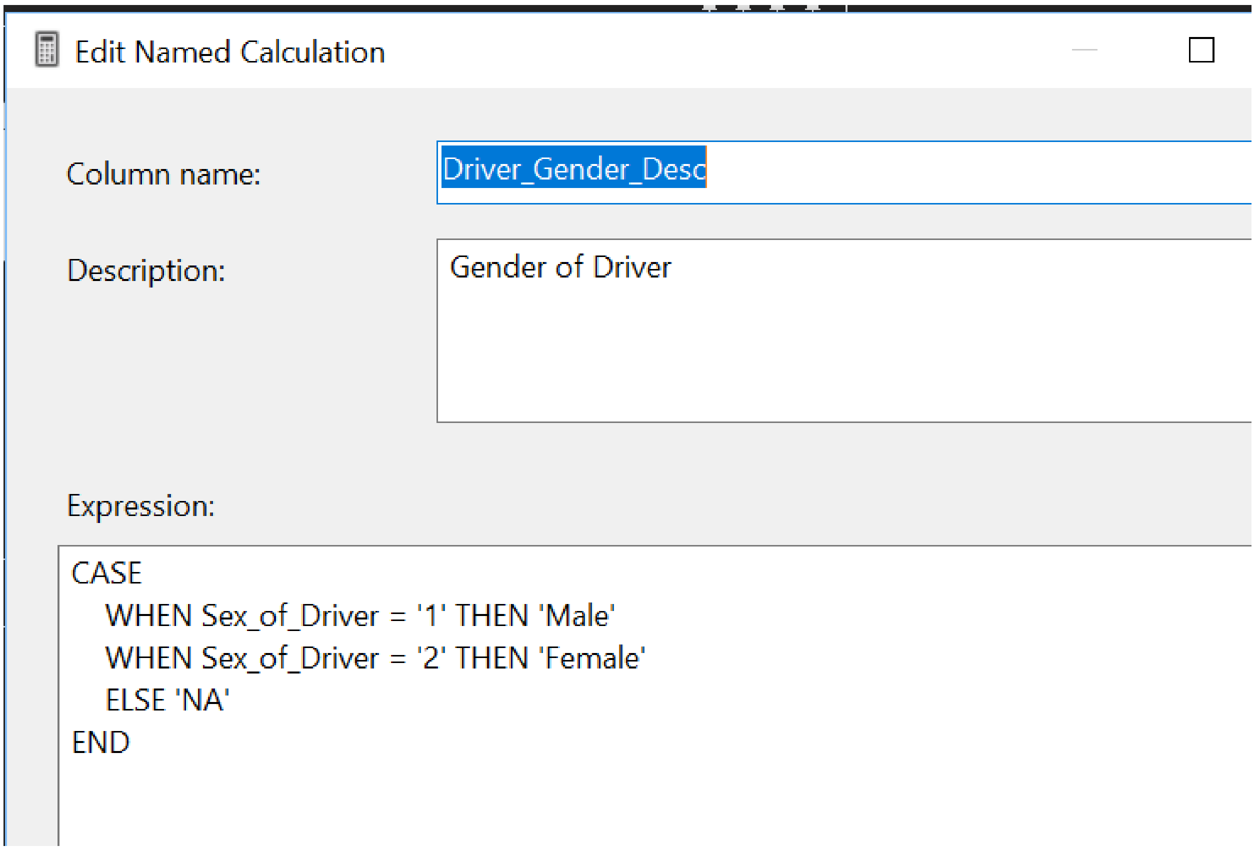


2. OLAP Cube design and use

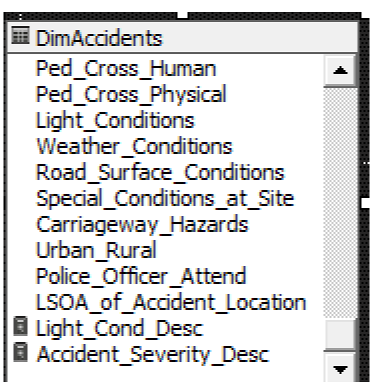
- Create the named calculations:

1. Driver's Gender description





2. Light condition description (as screenshot on section 3)
3. Accident severity description



Column name: Light_Cond_Desc

Description: Description of the light conditions

Expression:

```
CASE
WHEN Light_Conditions = '1' THEN 'Daylight'
WHEN Light_Conditions = '4' THEN 'Darkness - lights lit'
WHEN Light_Conditions = '5' THEN 'Darkness - lights unlit'
WHEN Light_Conditions = '6' THEN 'Darkness - no lighting'
WHEN Light_Conditions = '7' THEN 'Darkness - lighting unknown'
ELSE 'Data missing'
END
```

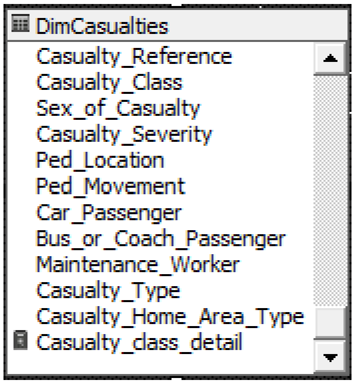
Column name: Accident_Severity_Desc

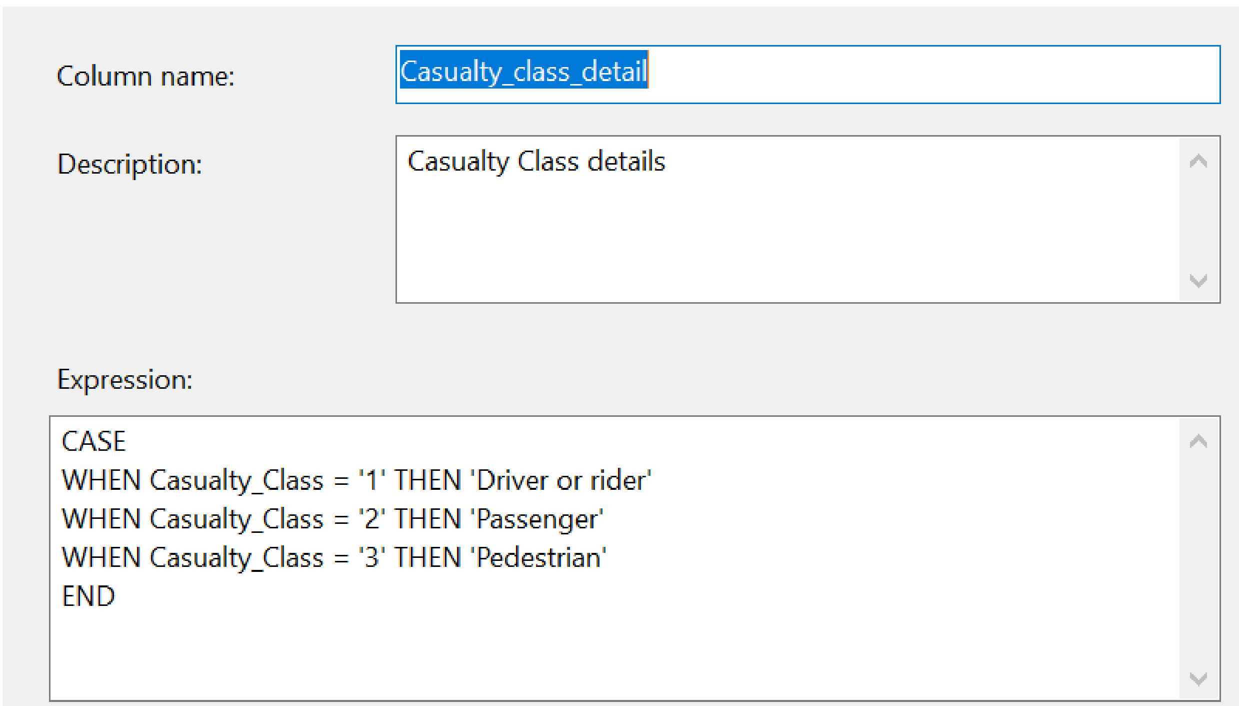
Description: Description of Accident Severity

Expression:

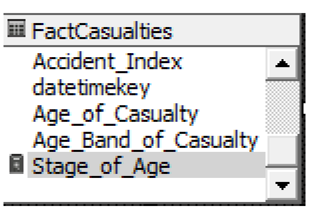
```
CASE
WHEN Accident_Severity = '1' THEN 'Fatal'
WHEN Accident_Severity = '2' THEN 'Serious'
WHEN Accident_Severity = '3' THEN 'Slight'
END
```

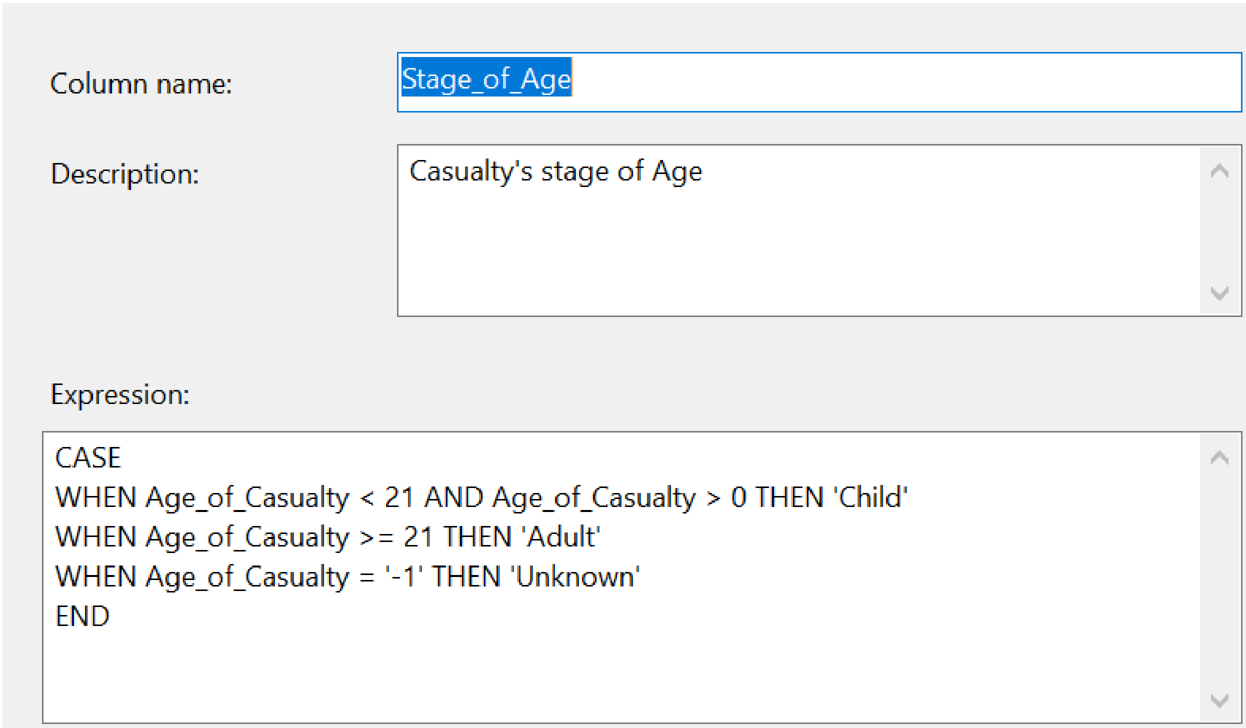
4. Casualty class details:





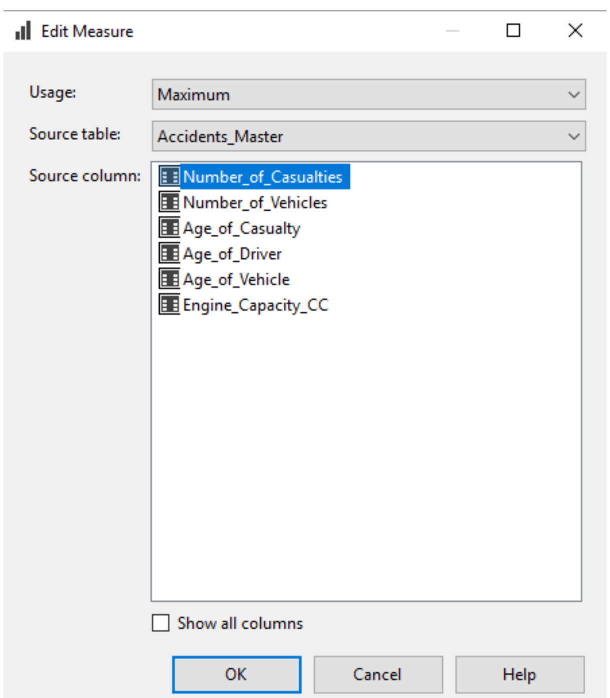
5. Casualty stage of Age: <21 Child; >=21 Adult



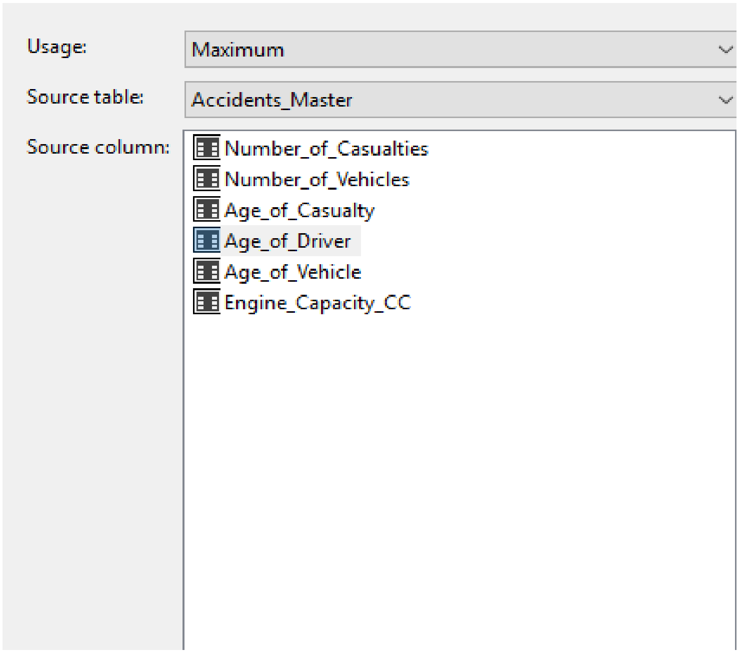


- Create 2 new measures

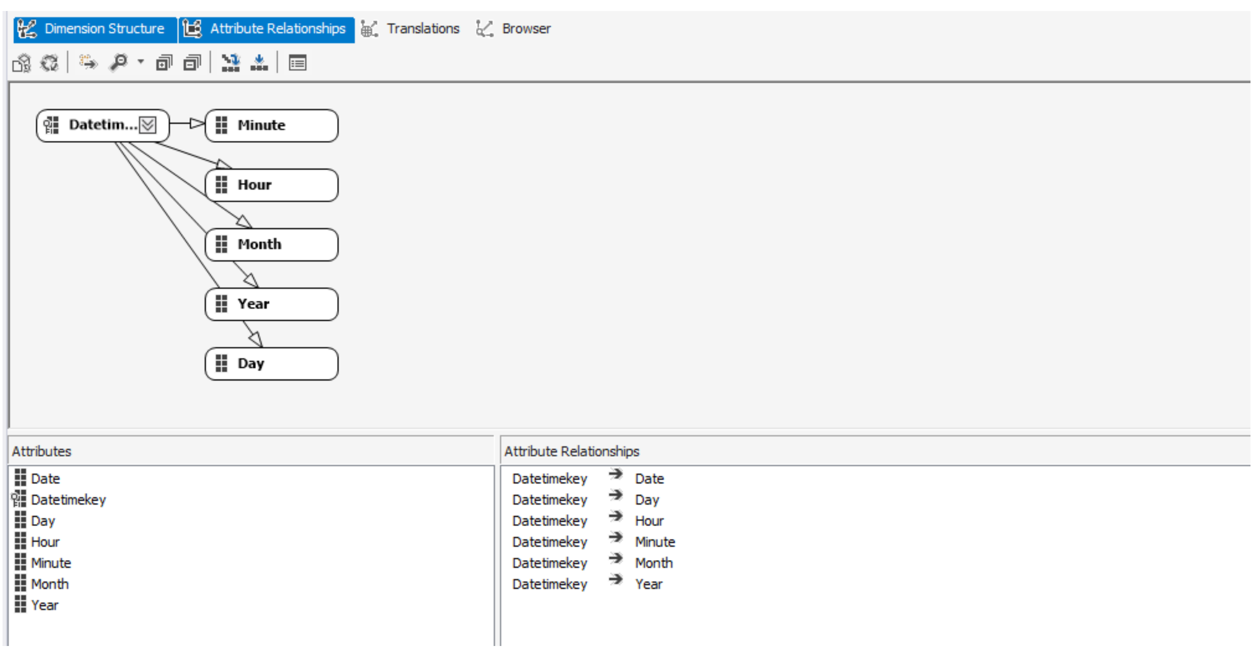
Maximum number of casualty

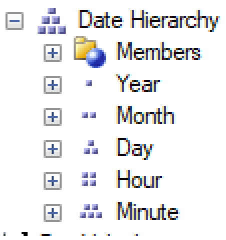


Maximum Age of Drivers

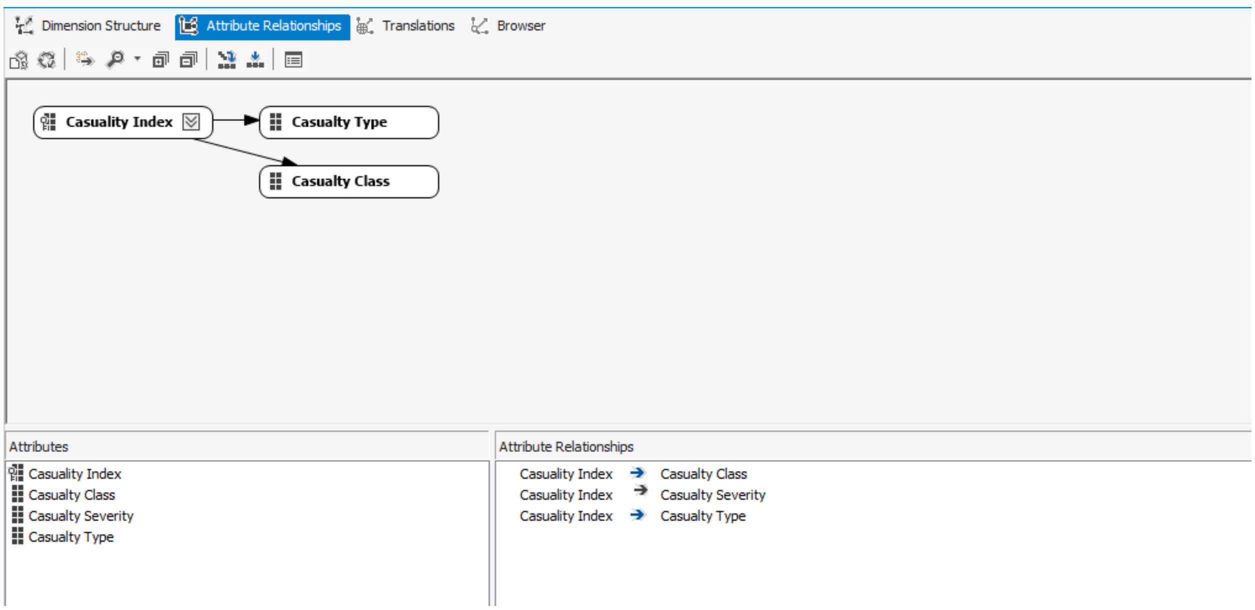


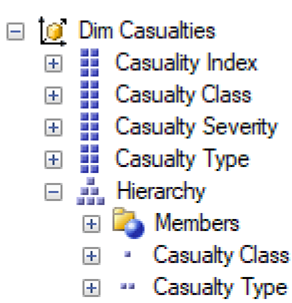
- Create a hierarchy for the Date Dimension





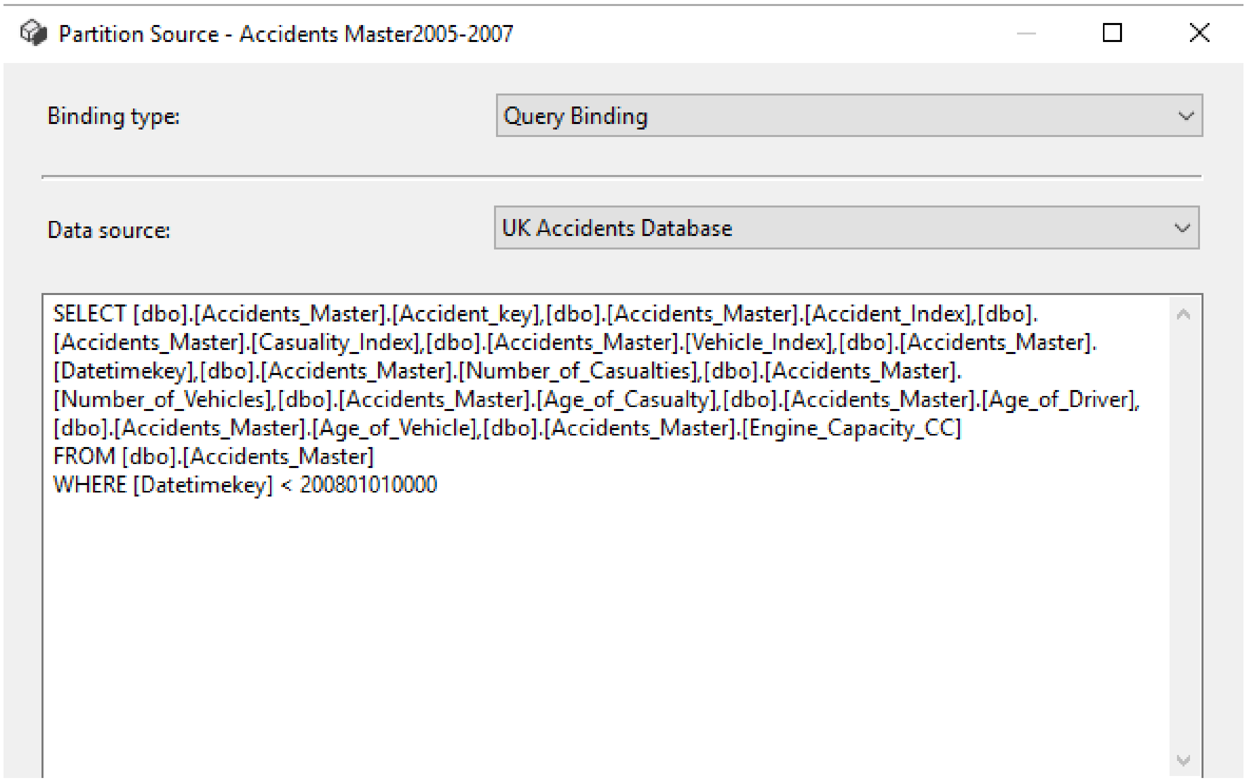
Create our custom hierarchy



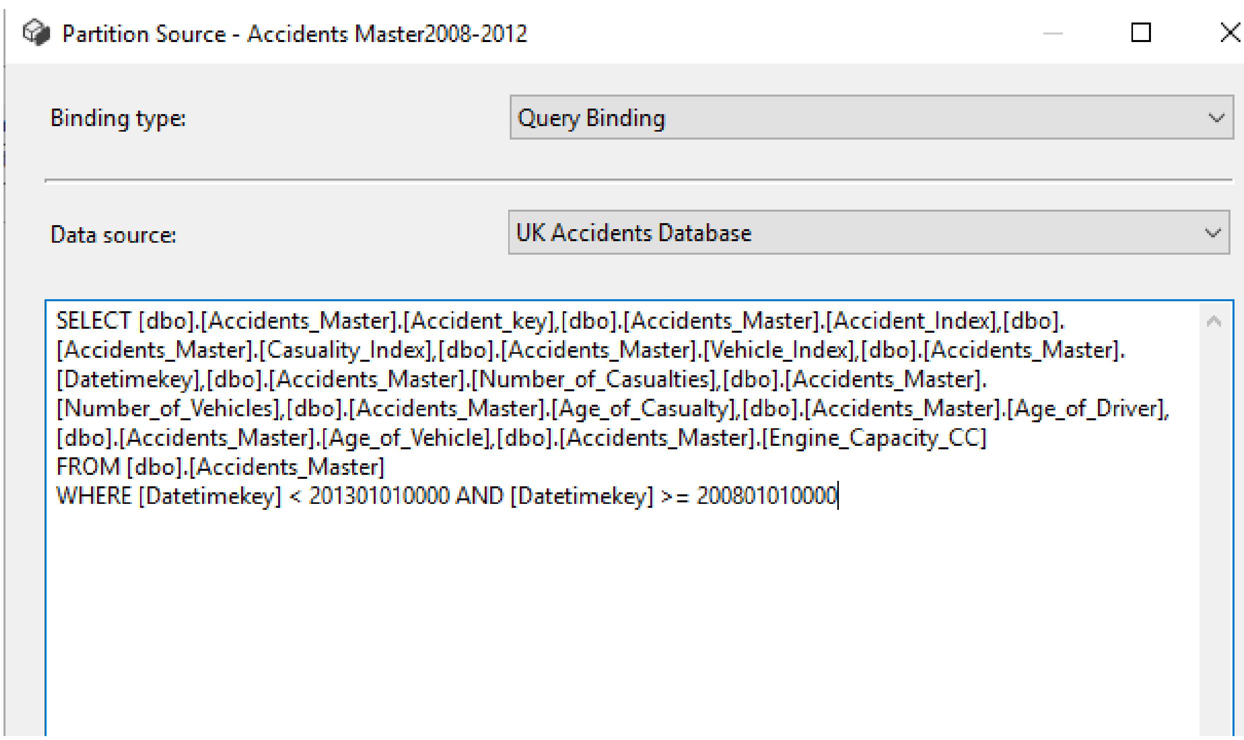


Partition and Aggregation

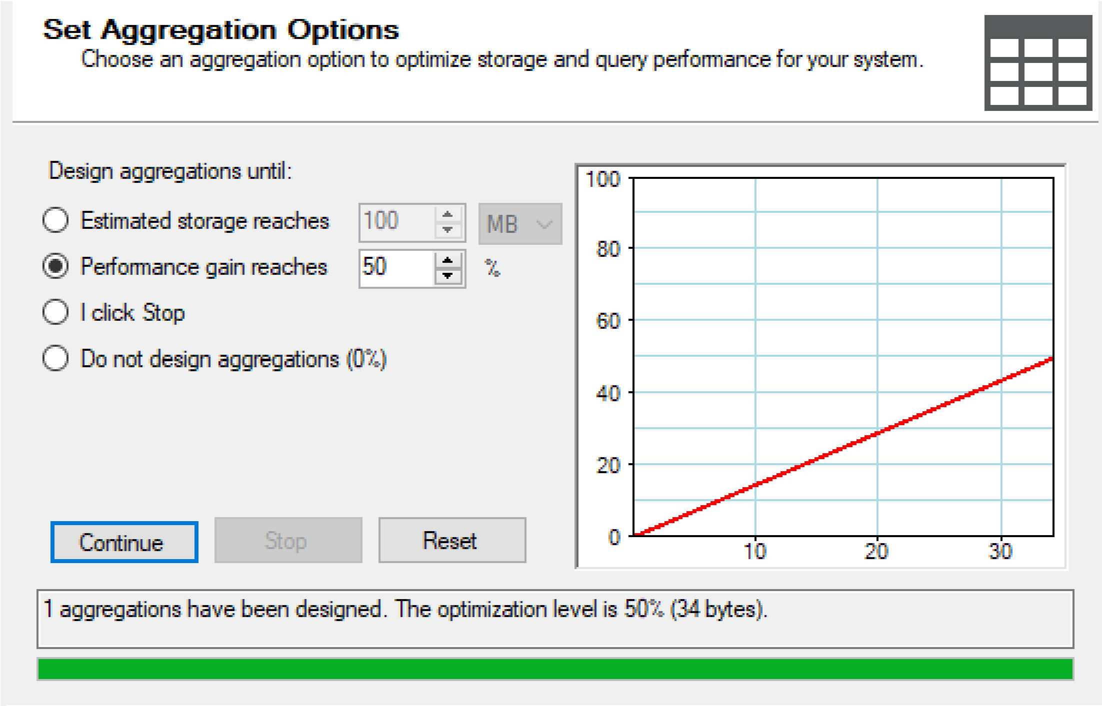
Accidents Master 2005-2007



Accidents Master 2008-2012



Create 2 aggregations for 50% percent improving in performance



Cube Structure | Dimension Usage | Calculations | KPIs | Actions | Partitions | Aggregations | Perspectives | Translations | Browser

	Aggregations	Estimated Partition...	Partitions
Accidents Master (2 Aggregation Designs)			
AggregationDesign50percent20052007	2	1506361	Accidents Master 2005-2007
AggregationDesign50percent20082012	2	2046047	Accidents Master 2008-2012

```
//1.   Top 6 months among all years with the most number of casualties (TopCount)
Select [Measures].[Number Of Casualties] on 0,
TopCount(([Dim Date].[Year].children,[Dim Date].[Month].children), 6, [Measures].[Number
of Casualties]) on 1
From [UK Accidents Database]
```

Messages | Results

		Number Of Casualties
2005	12	98494
2005	11	96708
2006	7	96497
2005	10	96078
2007	10	92313
2007	8	92035

```
//2. Number of vehicles with casualty severity type 3 (IIF)
with member [Measures].[Multicar] AS
iif([Measures].[Number Of Vehicles] > 1, "Multicar Crash", "Single Car Crash")
```

```
select {[Measures].[Number Of Casualties], [Measures].[Multicar]} on 0,
[Dim Vehicles].[Sex Of Driver].[Sex Of Driver].members on 1
From [UK Accidents Database]
```

	Number Of Casualties	Multicar
-1	84	Multicar Crash
1	4964611	Multicar Crash
2	2082483	Multicar Crash
3	275906	Multicar Crash
Unknown	(null)	Single Car Crash

```
//3. Number of casualties per year
select [Measures].[Number Of Casualties] on 0,
[Dim Date].[Year].members on 1
From [UK Accidents Database]
```

	Number Of Casualties
All	7323084
2005	1060832
2006	1004696
2007	993009
2008	904923
2009	857946
2010	814998
2011	932853
2012	753827
2013	(null)
2014	(null)
2015	(null)

```
//4. Bottom six months for number of casualties (BottomCount, Filter, Not IsEmpty)
Select [Measures].[Number Of Casualties] on 0,
BottomCount(Filter([Dim Date].[Month].members, Not IsEmpty ([Measures].[Number of
Casualties])
), 6, [Measures].[Number Of Casualties]) on 1
From [UK Accidents Database]
```

	Number Of Casualties
2	528925
1	559728
3	562577
4	581769
12	583622
5	602829

```
//5. Number of casualties in the first four years (Head)
select [Measures].[Number Of Casualties] on 0,
Head([Dim Date].[Year].members, 4) on 1
from [UK Accidents Database]
```

	Number Of Casualties
All	7323084
2005	1060832
2006	1004696
2007	993009

```
//6. Number of vehicles in accidents during the last four years (Tail)
select [Measures].[Number Of Vehicles] on 0,
Tail([Dim Date].[Year].members, 4) on 1
from [UK Accidents Database]
```

	Number Of Vehicles
2012	847437
2013	(null)
2014	(null)
2015	(null)

```
//7. Number of Casualties during the February (Extract)
select [Measures].[Number Of Casualties] on 0,
Extract({[Dim Date].[Month].&[2]}, [Dim Date].[Month]) on 1
From [UK Accidents Database]
```

	Number Of Casualties
2	528925

```
//8. Number of casualties by month ordered by number of casualties descending (Order)
select [Measures].[Number Of Casualties] on 0,
Order ([Dim Date].[Year].children, [Measures].[Number Of Casualties], ASC) on 1
From [UK Accidents Database]
```

	Number Of Casualties
2013	(null)
2014	(null)
2015	(null)
2012	753827
2010	814998
2009	857946
2008	904923
2011	932853
2007	993009
2006	1004696
2005	1060832

```
//9. Number of casualties per year except 2015 (Except)
select [Measures].[Number Of Casualties] on 0,
Except([Dim Date].[Year].[Year],[Dim Date].[Year].&[2015]) on 1
From [UK Accidents Database]
```

	Number Of Casualties
2005	1060832
2006	1004696
2007	993009
2008	904923
2009	857946
2010	814998
2011	932853
2012	753827
2013	(null)
2014	(null)

```
//10. Maximum age of driver per year
select [Measures].[Maximum Age Of Drive] on 0,
[Dim Date].[Year].[Year] on 1
From [UK Accidents Database]
```

	Maximum Age Of Drive
2005	99
2006	98
2007	98
2008	98
2009	99
2010	99
2011	99
2012	99
2013	(null)
2014	(null)
2015	(null)

10 Functions Used

1. Head
2. Tail
3. Order
4. Not IsEmpty
5. Filter
6. Except
7. Extract
8. BottomCount
9. TopCount
10. IIF

3. Choosing models for casualty severity prediction

We will predict the severity of casualties in accidents using Casualty Severity as the target variable. Casualty Severity have 3 values: 1 – Fatal; 2- Serious; 3-Slight. The predictors will have discrete values, as tables below:

- Casualty Type

code	label
0	Pedestrian
1	Cyclist
2	Motorcycle 50cc and under rider or passenger
3	Motorcycle 125cc and under rider or passenger
4	Motorcycle over 125cc and up to 500cc rider or passenger
5	Motorcycle over 500cc rider or passenger
8	Taxi/Private hire car occupant
9	Car occupant
10	Minibus (8 - 16 passenger seats) occupant
11	Bus or coach occupant (17 or more pass seats)
16	Horse rider
17	Agricultural vehicle occupant
18	Tram occupant
19	Van / Goods vehicle (3.5 tonnes mgw or under) occupant
20	Goods vehicle (over 3.5t. and under 7.5t.) occupant
21	Goods vehicle (7.5 tonnes mgw and over) occupant
22	Mobility scooter rider
23	Electric motorcycle rider or passenger
90	Other vehicle occupant
97	Motorcycle - unknown cc rider or passenger
98	Goods vehicle (unknown weight) occupant

- Ped location

code	label
0	Not a Pedestrian
1	Crossing on pedestrian crossing facility
2	Crossing in zig-zag approach lines
3	Crossing in zig-zag exit lines
4	Crossing elsewhere within 50m. of pedestrian crossing
5	In carriageway, crossing elsewhere
6	On footway or verge
7	On refuge, central island or central reservation
8	In centre of carriageway - not on refuge, island or central reservation
9	In carriageway, not crossing
10	Unknown or other
-1	Data missing or out of range

- Sex of Casualty

code	label
1	Male
2	Female
3	Not known
-1	Data missing or out of range

- Car Passenger: number of car passengers
- Bus or Coach Passenger: number of bus or Coach passenger
- Casualty Class Detail

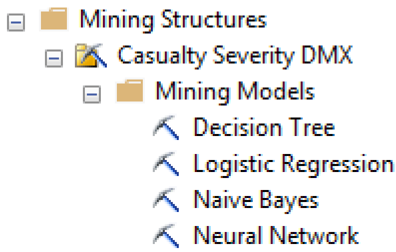
code	label
1	Driver or rider
2	Passenger
3	Pedestrian

We will use 4 data mining techniques: Decision Tree, Logistic Regression, Naïve Bayes and Neural Network to create 4 prediction models. First, we create the mining structure as below:

Mining Structure | Mining Models | Mining Model Viewer | Mining Accuracy Chart | Mining Model Prediction

Structure	Decision Tree	Logistic Regression	Neural Network	Naive Bayes
	Microsoft_Decision_Trees	Microsoft_Logistic_Regression	Microsoft_Neural_Network	Microsoft_Naive_Bayes
Bus Or Coach Passenger	Input	Input	Input	Input
Car Passenger	Input	Input	Input	Input
Casuality Index	Key	Key	Key	Key
Casualty Class Detail	Input	Input	Input	Input
Casualty Severity	PredictOnly	PredictOnly	PredictOnly	PredictOnly
Casualty Type	Input	Input	Input	Input
Ped Location	Input	Input	Input	Input
Sex Of Casualty	Input	Input	Input	Input

In these mining structures, we used a maximum of 1000 cases and30 percent test data.

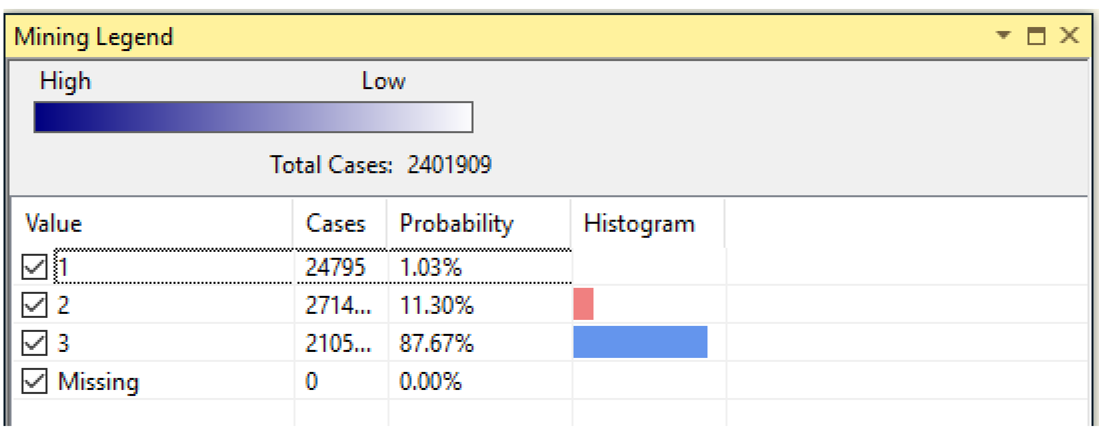


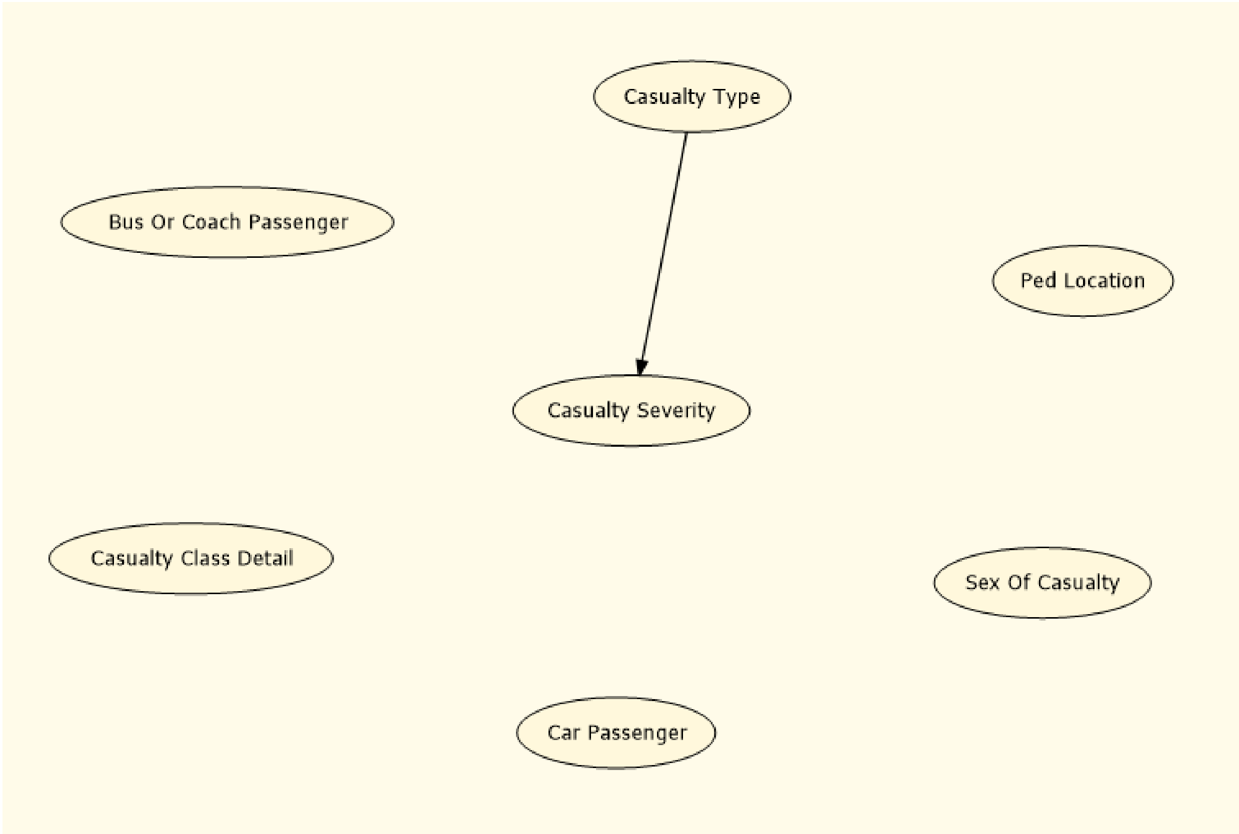
4. Models assessment and findings

Decision Tree model

As you can see, the probability of casualties which have fatal injuries after accident is 1.03%, serious condition is 11.30% and slight condition is 87.67% which has the most cases.

The most important predictor is Casualty Type.



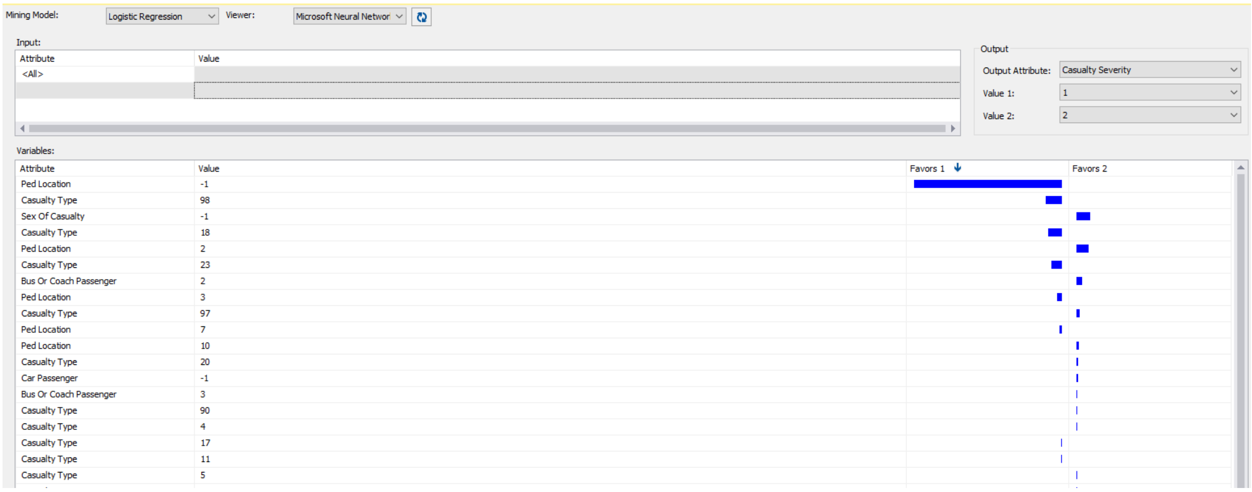


Logistic Regression model

Concluding from the output, casualties which have highest probability of fatal injuries are Ped Location = -1. But this value means missing or unknown, so it makes no sense.

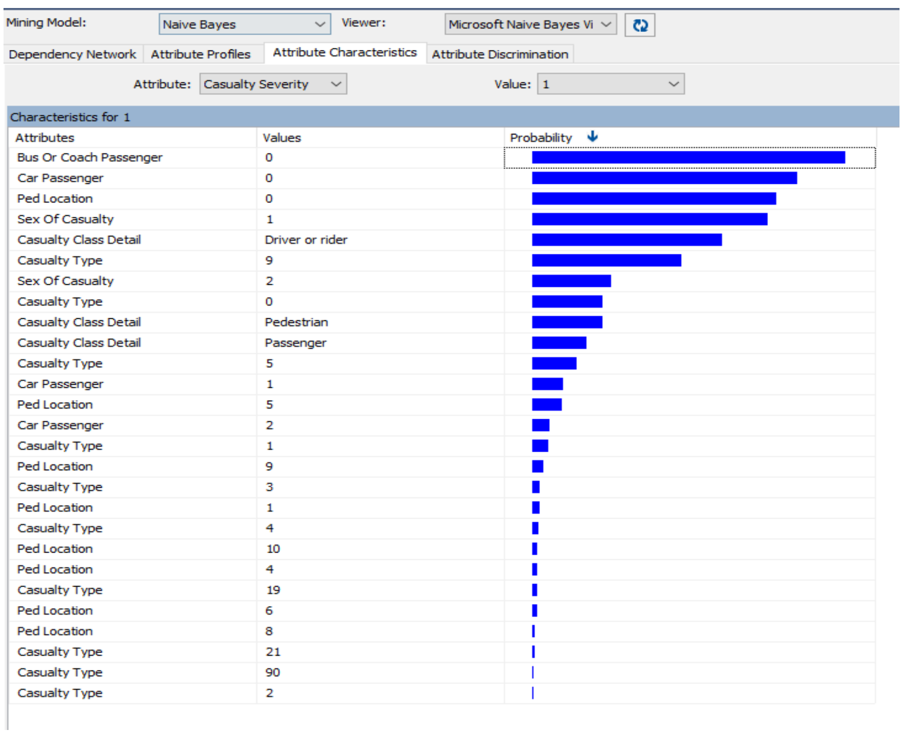
Assessing the model result, there are some conclusions as below:

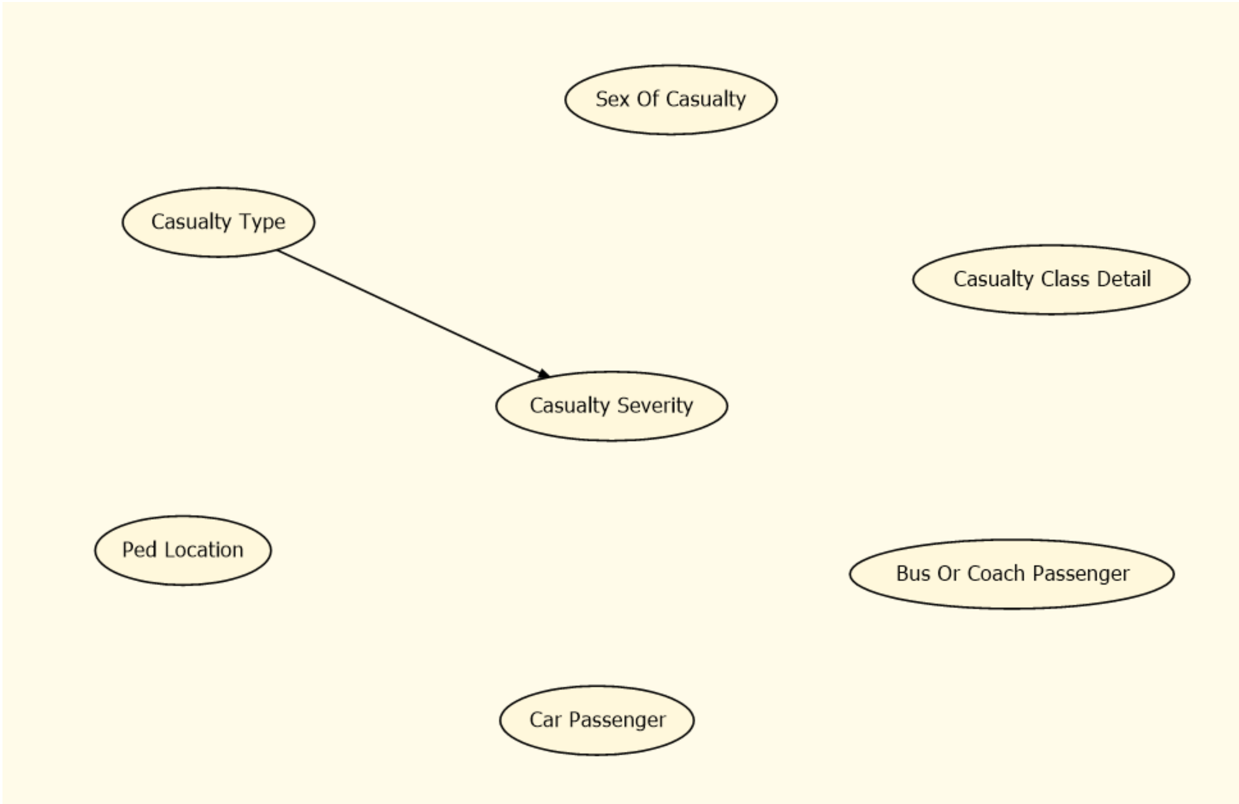
- Casualty Type = 18 (Tramp Occupant) will have highest probability of fatal injuries with 99.98% probability, then Ped Location = 3 (crossing in zig-zag exit line) with 94.30% probability of fatal injuries.
- Casualty Type = 97 (motor cycle – unknow cc or passenger) will have highest probability of serious injuries.
- Casualty Type = 23 (Electric motor cycle rider) and Ped location = 2 (crossing in zig-zag approach line) will have highest probability of slight injuries with 99.99% and 99.40% respectively.



Naïve Bayes model

Casualty Type is the most important predictor for Casualty Severity. The characters of casualties which have highest probabilities of fatal injuries are: No bus – coach passenger (99%), no car passenger (84%), not a pedestrian, sex = male, driver or rider, casual type = 9 (car occupant)



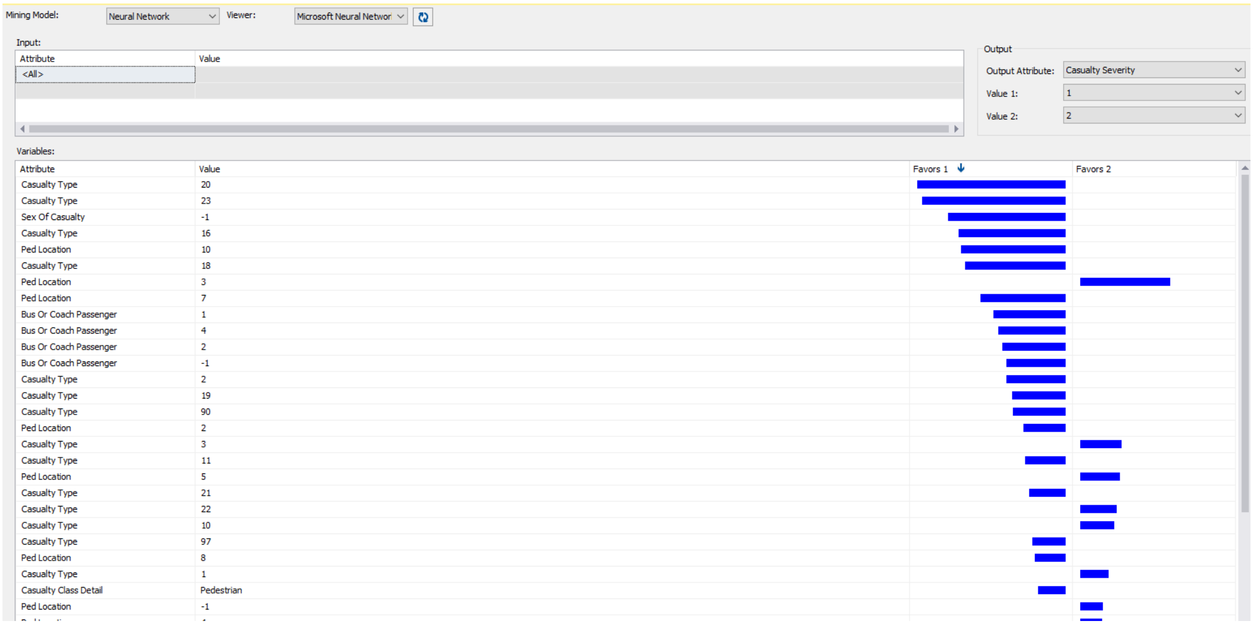


Neural Network model

Excluding the unknown values (-1) for predictors, the casualties with casual type = 18 (tramp occupant) will have the highest probability of fatal injuries (34.09%).

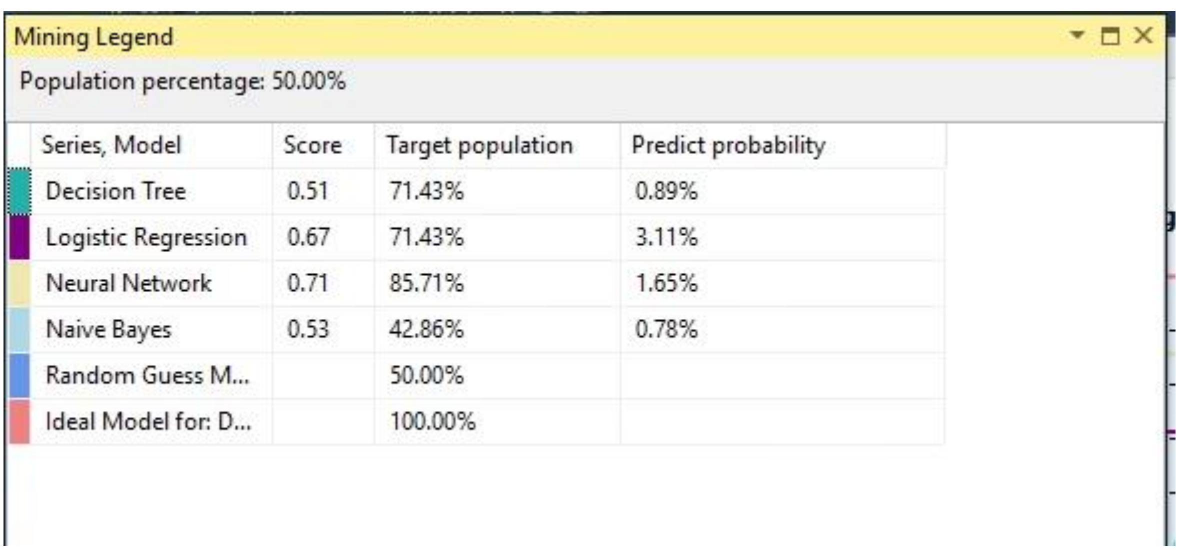
Casualties with casual type = 98 (Goods vehicle – unknown weight occupant) will have the highest probability of serious injuries (64.92%)

Casualties with casual type = 20 (Goods vehicle 3.5-7.5t occupant) and casual class detail = passenger have the highest probability of slight injuries ~94%



5. Models comparison and conclusion:

Lift score: using the mining structure test cases, we will assess the lift scores of all 4 models for predicting Fatal Injuries (Severity = 1). Neural Network has highest score here.



Mining Legend

Population percentage: 50.00%

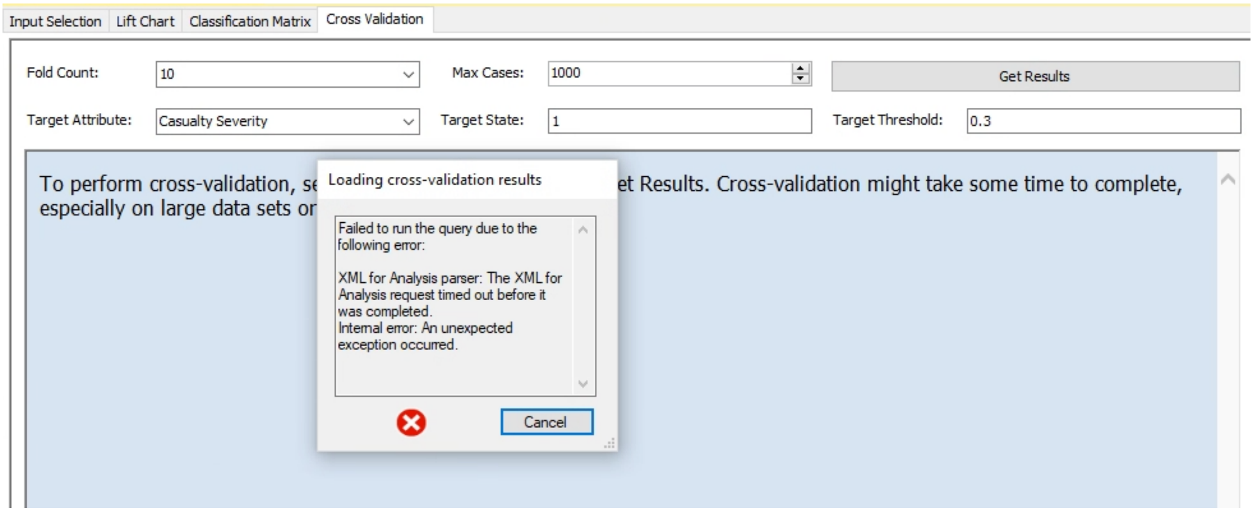
Series, Model	Score	Target population	Predict probability
Decision Tree	0.51	71.43%	0.89%
Logistic Regression	0.67	71.43%	3.11%
Neural Network	0.71	85.71%	1.65%
Naive Bayes	0.53	42.86%	0.78%
Random Guess M...		50.00%	
Ideal Model for: D...		100.00%	

Classification Matrix:

Counts for Decision Tree on Casualty Severity => correct classification percentage = 881/1000 = 88.1%				
	Predicted	1 (Actual)	2 (Actual)	3 (Actual)
	1	0	0	0
	2	0	0	0
	3	7	112	881
Counts for Logistic Regression on Casualty Severity => correct classification percentage = 880/1000 = 88%				
	Predicted	1 (Actual)	2 (Actual)	3 (Actual)
	1	0	1	0
	2	0	0	1
	3	7	111	880
Counts for Neural Network on Casualty Severity => correct classification percentage = 881/1000 = 88.1%				
	Predicted	1 (Actual)	2 (Actual)	3 (Actual)
	1	0	0	1
	2	0	1	0
	3	7	111	880
Counts for Naive Bayes on Casualty Severity => correct classification percentage = 840/1000 = 84%				
	Predicted	1 (Actual)	2 (Actual)	3 (Actual)
	1	0	0	0
	2	0	12	53
	3	7	100	828

Basically, all 4 models have the same correct classification percentage.

We try to run the cross validation with fold count = 10 but the dataset is too large and timeout occurred.



⇨ Conclusion: the best prediction model here is Neural Network with highest lift score, and the result shows that you target for 50% population of casualties, you will correctly identify 85.71% of fatal injuries, a very good score. This model also has 88.1% percent of correct classification, which is a decent result.