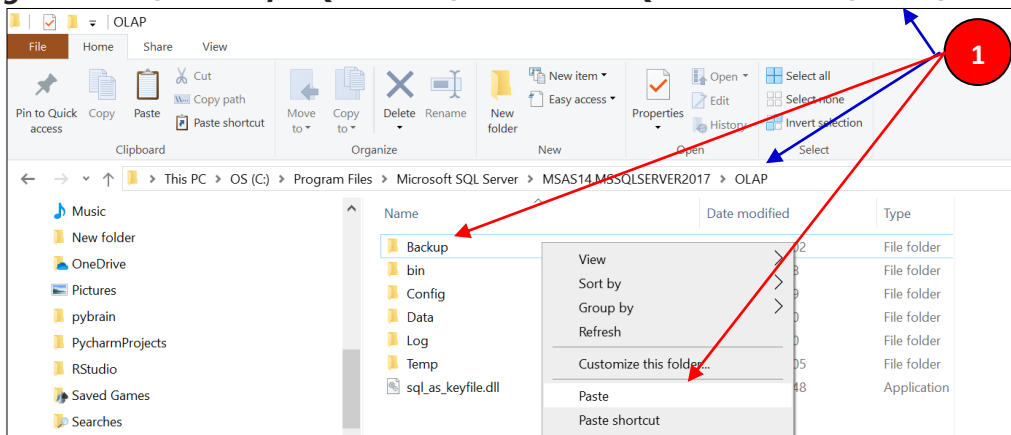


## PARTEA 3

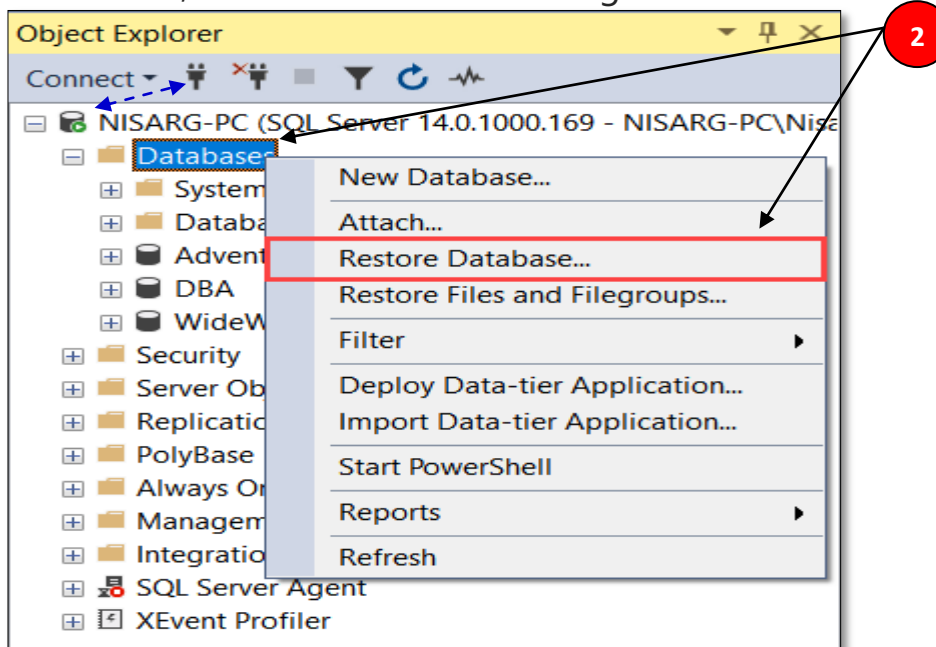
Încarcarea și instalarea pe SSMS a DWH-ului exemplu **AdventureWorksDW2017.bak** (similar e și 2014-2016) pentru lucrul cu regresia liniară multiplă, în calitate de testare și verificare a utilizării acestui algoritm.

1. La început plasăm fișierul de rezervă **AdventureWorksDW2017.bak** pe adresa

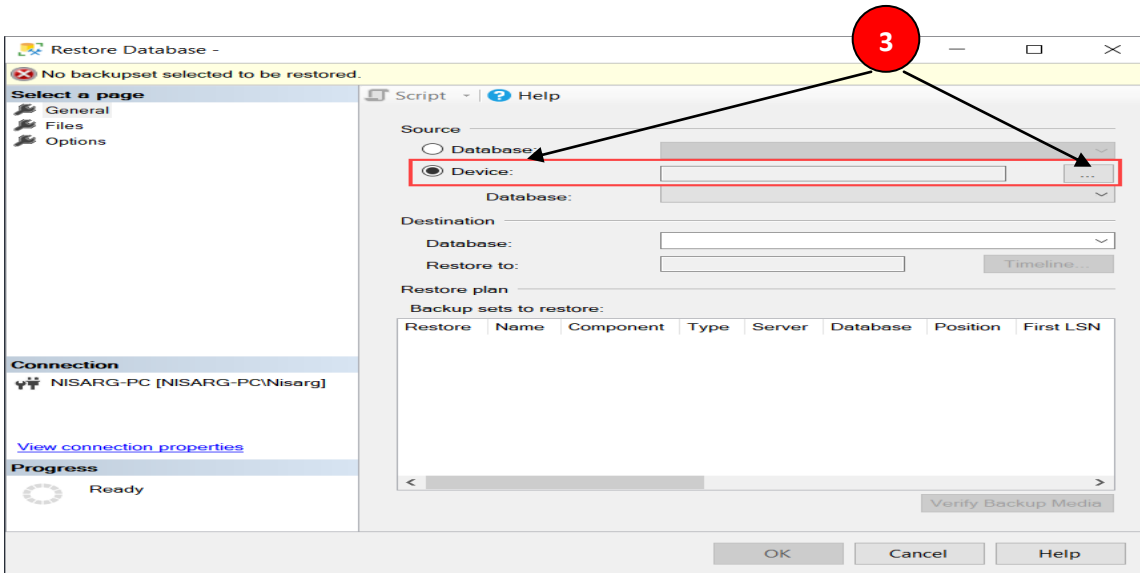
**C:\Program Files\Microsoft SQL Server\MSAS14.MSSQLSERVER2017\OLAP\Backup**



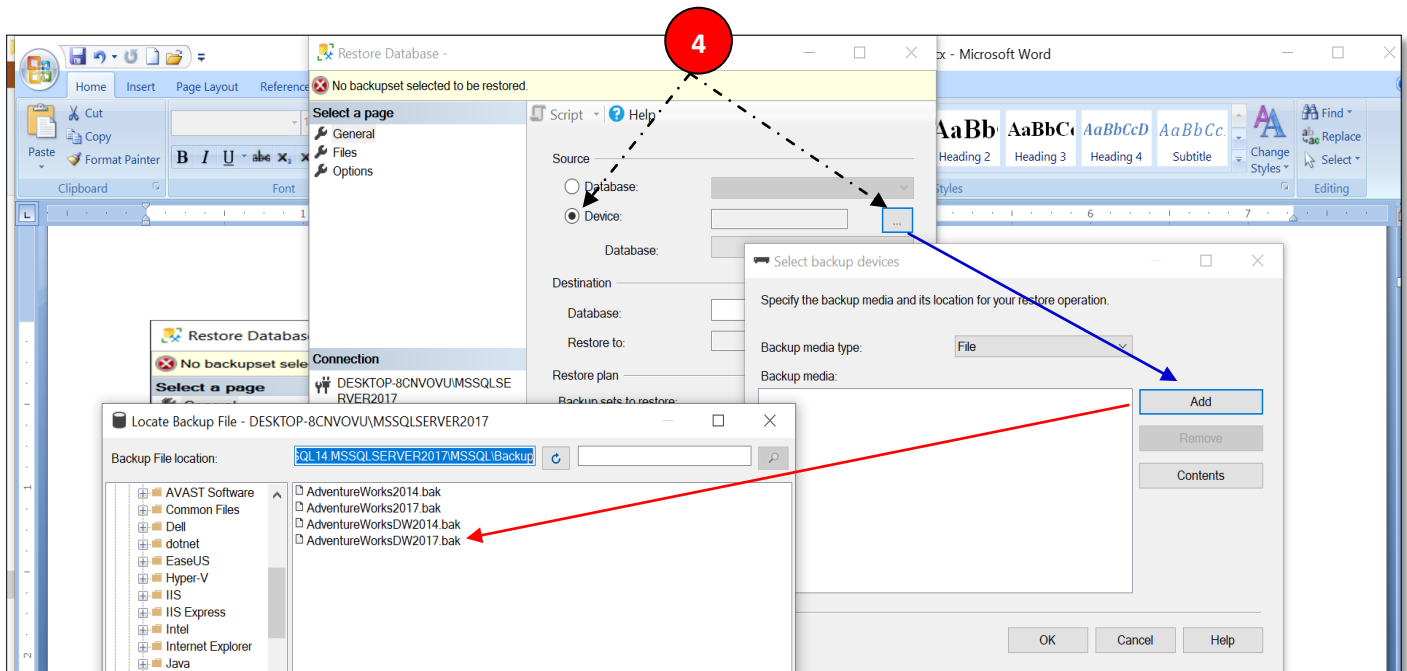
2. Lansăm **SQL Server Management Studio** și din **Object Explorer**, lansăm motorul bazei de date, apoi facem clic dreapta pe **Database** și selectăm **Restore Database**, conform următoarei imagini:



În fereastra **Restore Database**, selectăm **Device** în calitate de sursă și apoi facem click pe ellipse (...):

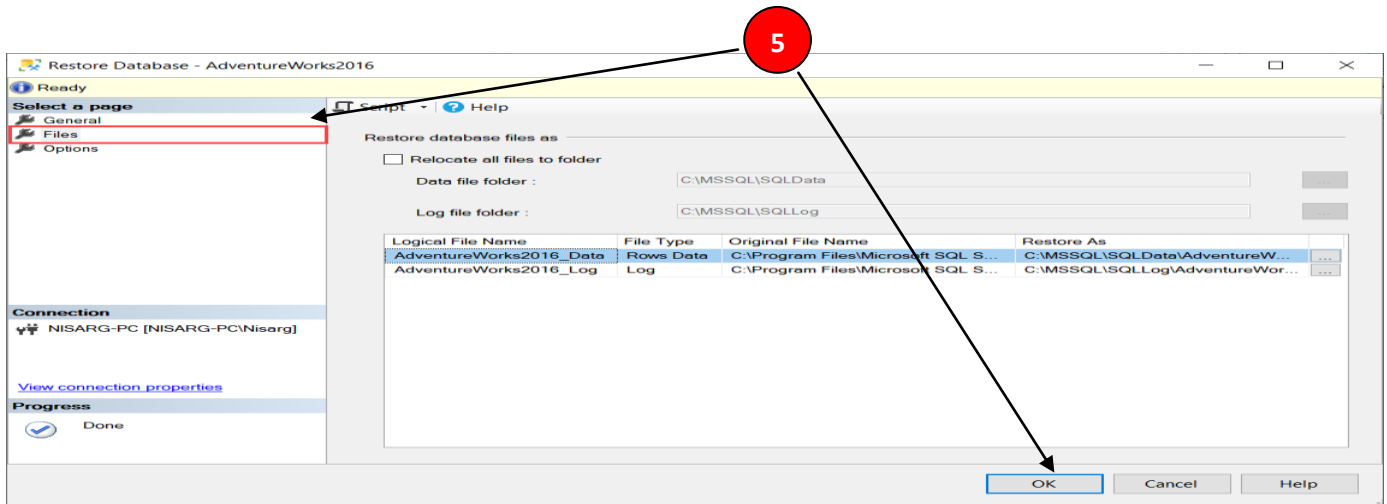


In fereastra **Locate backup devices** , selectăm suportul de rezervă făcând click pe **Add**, selectăm fișierul de rezervă *AdventureWorksDW2017.bak* . Facem clic pe **OK**:

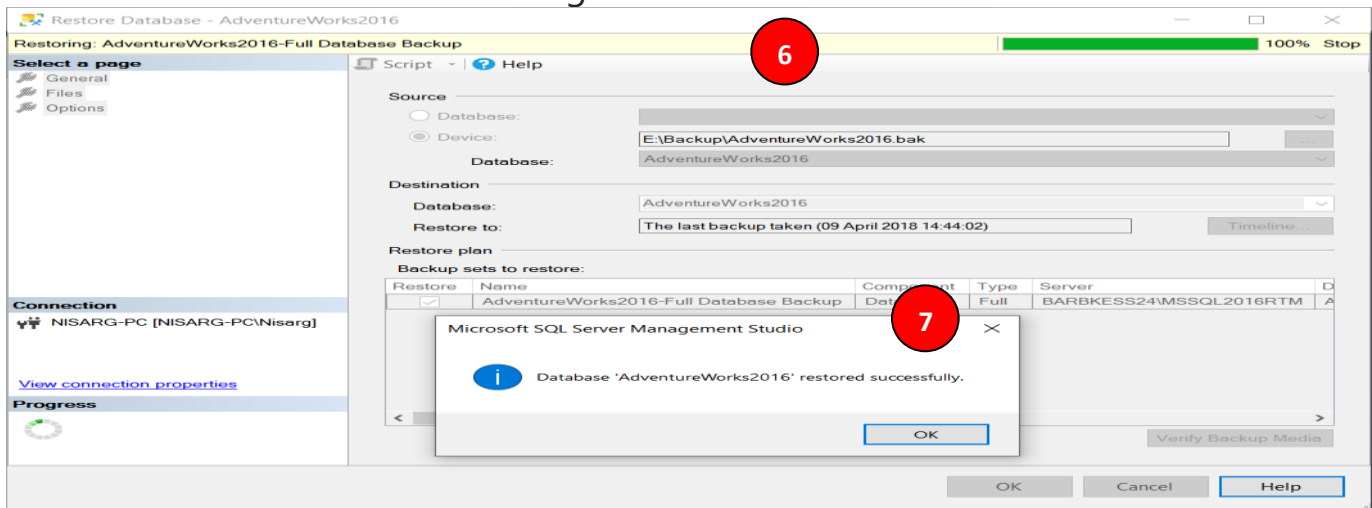


Dacă dorim să schimbăm locația fizică a fișierului de date și a fișierului log, facem clic pe panoul **Files** și modificăm locația țintă pentru fișierele de date și log.

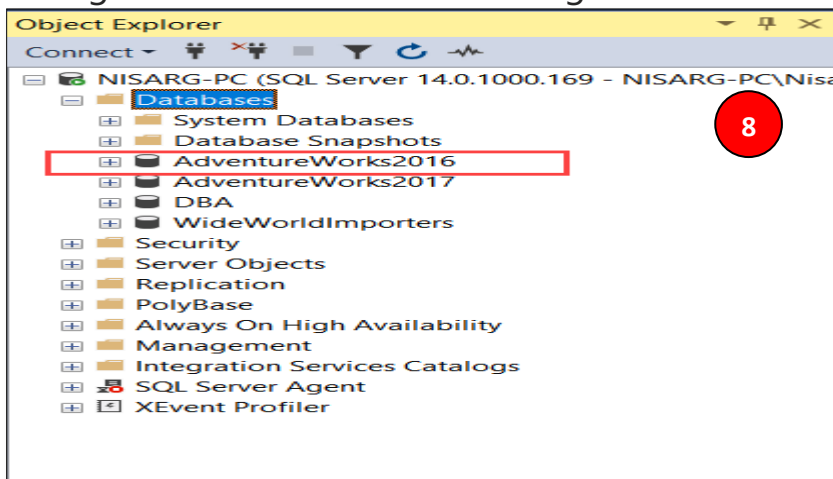
Reamintim, cea mai bună practică este să păstrați fișierele de date și fișierele de log pe unități separate.



Facem clic pe **OK**. Se va iniția procesul de restaurare a DWH/BD. Odată ce baza de date se restabilește cu succes, apare o fereastră care afirmă că DWH/BD a fost restaurată cu succes. Urmărim imaginea:



Odată ce DWH/BD a fost restaurată o putem conecta folosind SQL Server Management Studio. Urmărim imaginea:



**LA URMATORUL PAS LANȘĂM SI CONECTĂM DWH/BD DIN SSMS 2017/2019 LA VS SI APOI CONTINUĂM LUCRAREA / PASILOR CE URMEAȘĂ**



## DATA MINING

### REGRESIE LINIARĂ MULTIPLĂ MICROSOFT ÎN SQL SERVER

Regresia liniară este un algoritm al DM de extragere a datelor din seria de algoritmi de extragere a datelor SQL Server:

1. Naive Bayes,
2. arbori de decizie,
3. serii de timp,
4. reguli de asociere,
5. clustering etc

Microsoft Linear Regression este un algoritm de prognoză. In el se precaută restabilirea unui model liniar, ecuație liniară, în care sunt mai multe variabile independente si una, ce depinde de ele.

De exemplu, dacă dorim să **prezicem prețurile unei locuințe**, atunci trebuie să cunoastem mai multi factori ce il determina cum ar fi: **numărul de camere, zona locației casei, gradul de uzură și alte caracteristici ale casei.**

Aceasta înseamnă că modelul de regresie liniară poate fi reprezentat după cum urmează:

$$Y = a X_1 + b X_2 + \dots + z X_n + C ,$$

unde  $X_i, i=1,n$  sunt **variabile independente**, iar  $Y$ , variabila ce depinde de ele.

Să vedem cum putem folosi regresia liniară pe platforma Microsoft SQL Server. Vom folosi DWH-ul *AdventureWorksDW* și vizualizarea de exemplu a atributului *vTargetMail*.

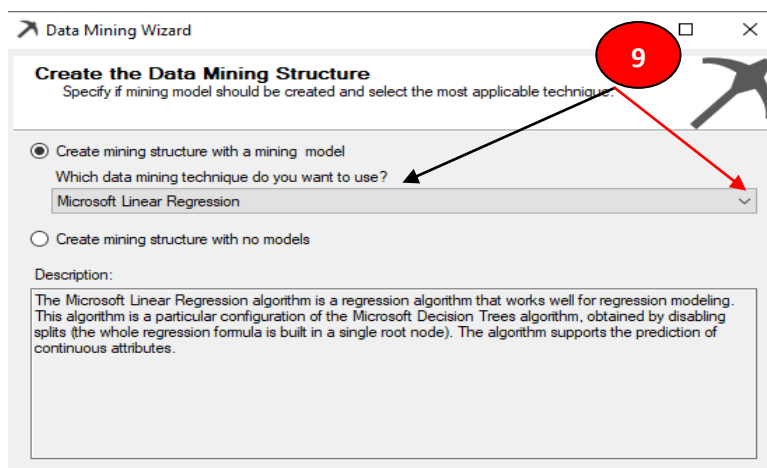
Algoritmul de extragere a datelor este unul standart si constă din 3 nivele.

#### Grupul de Activități 1

1. Conectarea serverului
2. Conectarea DWH
3. Crearea unui proiect in BI SSAS DM

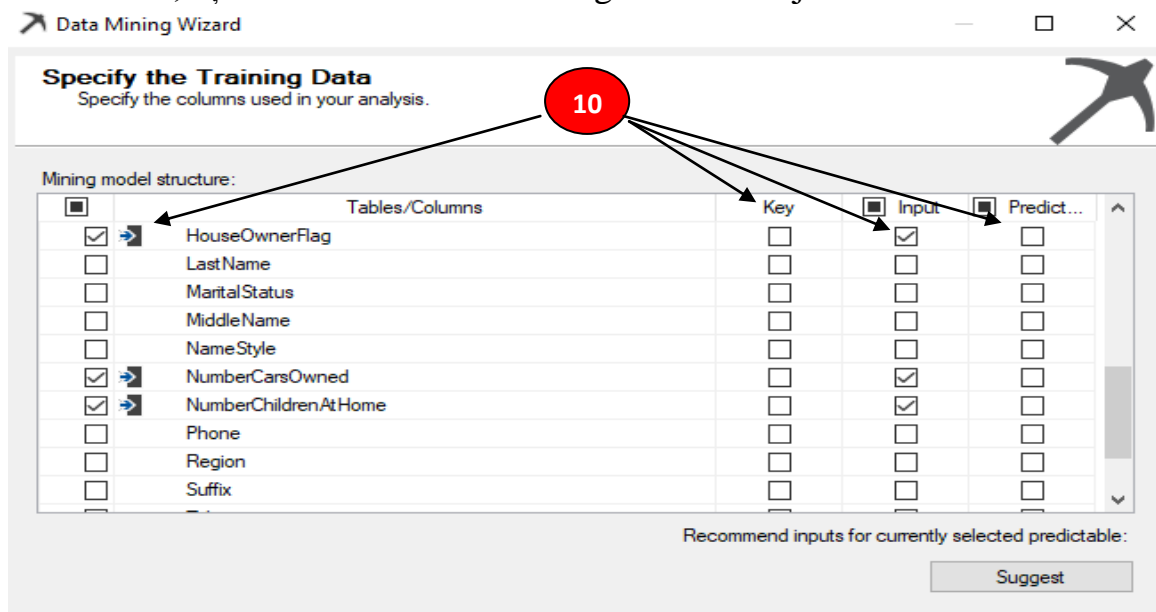
#### Grupul de Activități 2

1. Crearea unei surse de date – **DataSource** / in cazul nostru **sursa** este *AdventureWorksDW*/
2. Crearea pentru această sursă de date, o vedere a sursei de date –**DataSourceView**, in care pentru vizualizare a sursei de date selectăm vizualizarea *vTargetMail*.
3. Apoi selectam **Datamining Structure** si in Wizard-ul /Asistentul/ care se deschide alegem Microsoft Linear Regression, in calitate de algoritm de extragere a datelor, așa cum se arată în imaginea de mai jos.



În acest algoritm se folosește *Microsoft Decision Trees tehnic / Tehnica Microsoft a arborilor de decizie*. Spre deosebire de arborii de decizie, regresia liniară are doar un singur nod, prin care se verifică rezultatele regresiei liniare cu arbori de decizie pe care îl vom precăuta la sfârșitul lucrării de laborator.

**Tabelul vTargetMail** va fi tabelul de *Case/Caz/Studiu de caz* și ne va oferi să alegem atributele relevante, așa cum este indicat în imaginea de mai jos/ce urmează:



**Cheia/Key** pentru *Client* este aleasă în calitate de cheie din algoritmul de pe ecranul de mai sus.

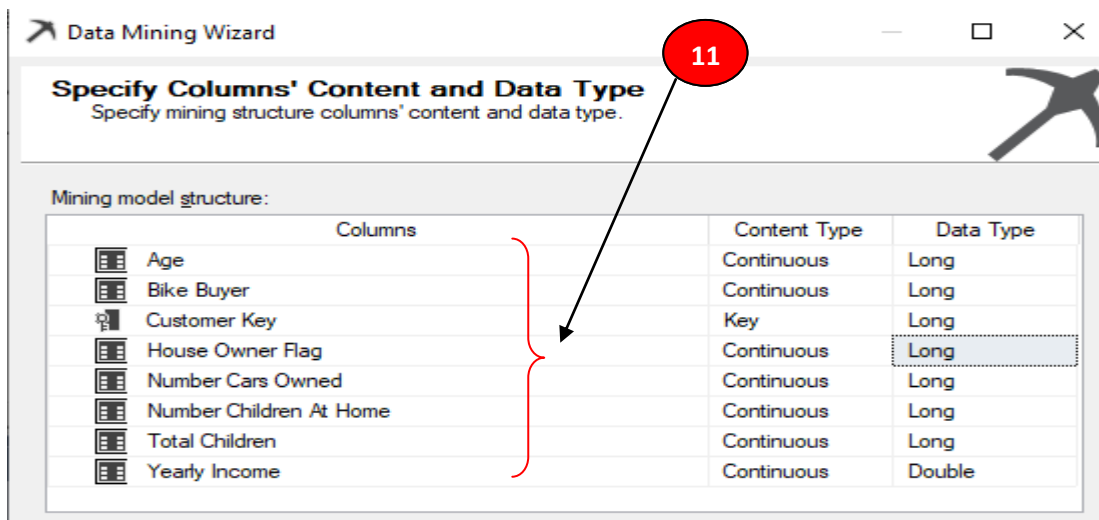
**Notă:** În regresia liniară Microsoft, toate intrările ar trebui să fie numerice; coloana de text nu trebuie selectată.

Prin urmare, în selecția de mai sus, *Vârsta, BikeBuyer, HouseOwnerFlag, /nu se recomanda!! nu este numeric!! NumberCarsOwned, NumberChildrenatHome, TotalChildren* sunt selectate ca atribute de *Intrare/Input*.

Aceasta este o limitare majoră în regresia liniară Microsoft, care nu se află în tehnicile de regresie liniară standard.

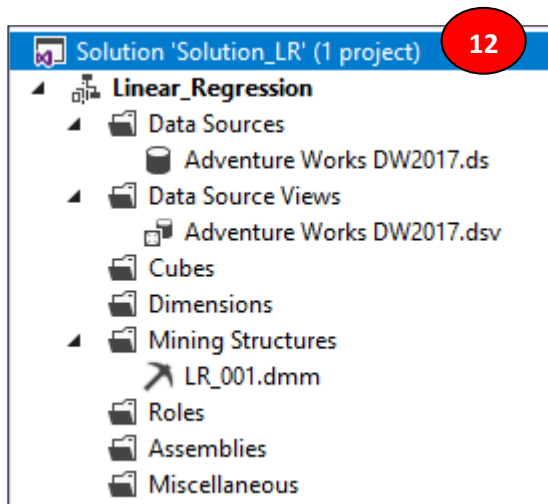
În alți algoritmi putem selecta de exemplu un **cimp textual**, în calitate de coloana pentru **predicție/forecast**, cum ar fi de exemplu **Cumpărătorul de biciclete/ Bike Buyer**. Cu toate acestea, în Regresia liniară Microsoft, trebuie să prezicem **Venitul anual/YearlyIncome**.

Deși există **Tipuri de conținut/Content types** implicite, există cazuri în care trebuie să le schimbăm. Acestea pot fi modificate din imaginea ce urmează.



**Notă:** În mod implicit, dacă **House Owner Flag** este selectat cu tipul de date implicit ca **text**, dar care în acest algoritm trebuie modificat la tipul de date **Long**. */El de fapt poate fi și omis din start/*

În celelalte ecrane din asistentul de extragere a datelor, sunt utilizate setările implicite. Acesta este **Exploratorul de soluții/ Solution Explorer** pentru extragerea datelor în **Microsoft Linear Regression** prin algoritmul Data Mining.

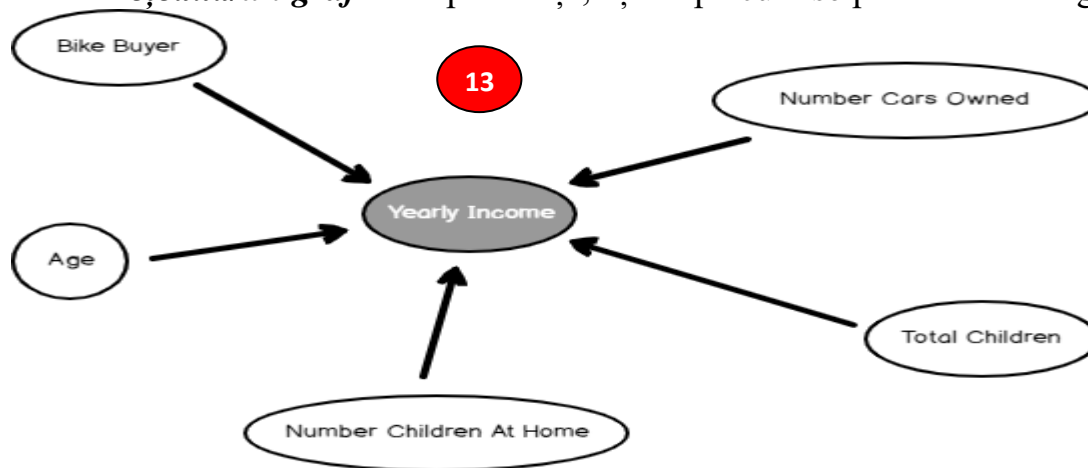


Următorul pas este procesarea *Structurii de extragere a datelor/Data Mining Structure*. Ar putea apare un mesaj de avertizare care să spună că nu există o împărțire/divizare în arbori de decizie. Acest avertisment poate fi ignorat pentru regresia liniară; nu va fi nici o împărțire/divizare pentru arborii de decizie.

După procesarea *Structurii de extragere a datelor/Data Mining Structure*, putem urmări rezultatele.

### Vizualizarea rezultatelor

În majoritatea algoritmilor Data Mining SQL Server, inclusiv în regresia liniară, putem urmări *rețeaua/un graf* de dependență, așa după cum se prezintă în imaginea de mai jos.



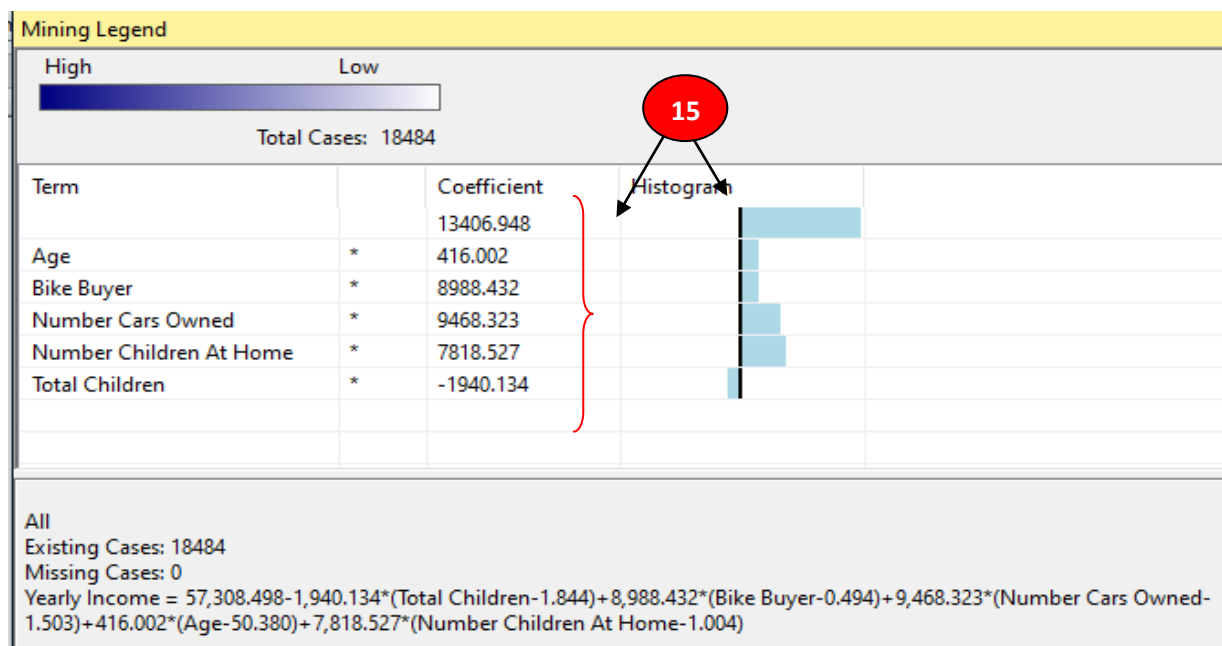
Rețeaua de dependență arată care sunt cele mai dependente attribute pentru a prezice *VenitAnnual/YearlyIncome*. Prin glisarea glisorului în partea stângă, putem afla semnificația acestor attribute.

În *Microsoft Linear Regression*, o altă vizualizare disponibilă este *Tree View*. Dar, așa cum s-a indicat anterior, este o vizualizare a unui arbore cu un singur nod.



Din acest *Tree View*, putem obține *ecuația de regresie liniară*, care este scopul final al algoritmului *Microsoft Linear Regression Data Mining*

Următoarea imagine arată ecuația de regresiei liniare.



Aceasta este ecuația și pur și simplu trebuie să înlocuim valorile relevante pentru a prezice **YearlyIncome**.

$$\begin{aligned}
 \text{YearlyIncome} &= 57.308.498 \\
 &- 1.940.134 * (\text{Total copii}-1.844) \\
 &+ 8.988.432 * (\text{Cumpărător de biciclete}-0.494) \\
 &+ 9.468.323 * (\text{Număr mașini deținute}-1.503) \\
 &+ 416.002 * (\text{Age}-50.380) \\
 &+ 7,818.527 * (\text{Număr de Copii La Acasă}-1.004)
 \end{aligned}$$

16

Să vedem cum putem prezice **YearlyIncome** utilizând modelul construit, prin funcția de predicție obținută mai sus.

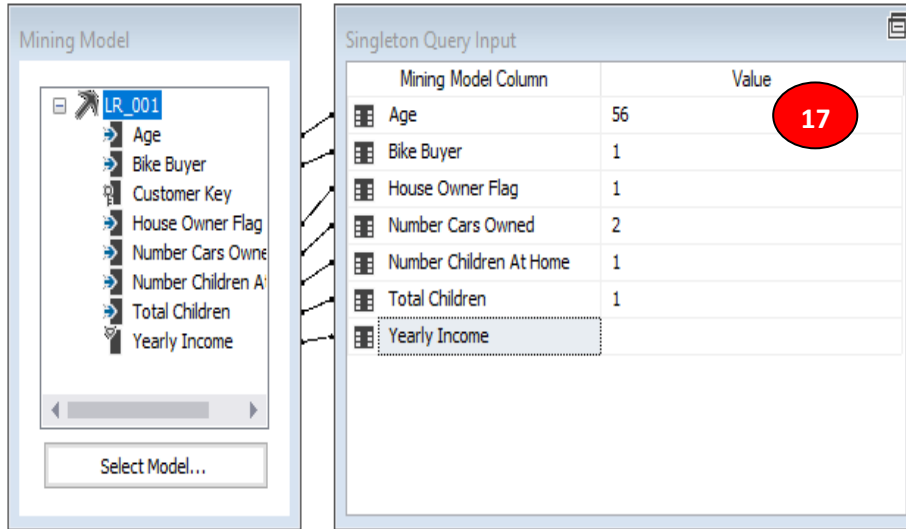
### Prognozarea.

Un aspect important al oricărui *Algoritm de Extragere a Datelor/Data Mining* este de a **prognoza utilizarea modelului construit**.

Să vedem cum putem efectua prognozarea folosind modelul Microsoft Linear Regression.

Acest lucru se poate face din fila/eticheta **Prognozarea cu Modelul Data Mining/Mining Model Prediction tab**, așa cum se arată în imaginea de mai jos. În exemplul următor, sunt furnizate anumite valori pentru o instanță de prognoză a **venitului annual /YearlyIncome**.





| Source              | Field              | Alias                | Show                                | Group | And/Or | Criteria/Argument        |
|---------------------|--------------------|----------------------|-------------------------------------|-------|--------|--------------------------|
| Prediction Function | Predict            | Predicted YearIncome | <input checked="" type="checkbox"/> |       |        | [LR_001].[Yearly Income] |
| Prediction Function | PredictProbability | Probability          | <input checked="" type="checkbox"/> |       |        | [LR_001].[Yearly Income] |

Din fila/eticheta rezultatelor, pot fi vizualizate rezultatele așa cum se arată în imaginea de mai jos.

| Predicted YearIncome | Probability       |
|----------------------|-------------------|
| 70506.5056044393     | 0.999945905009196 |

Aceleși rezultate pot fi obținute și cu ajutorul SQL Server Management Studio prin executarea interogării **DMX**. Următorul ecran arată interogarea și rezultatul acesteia.

```

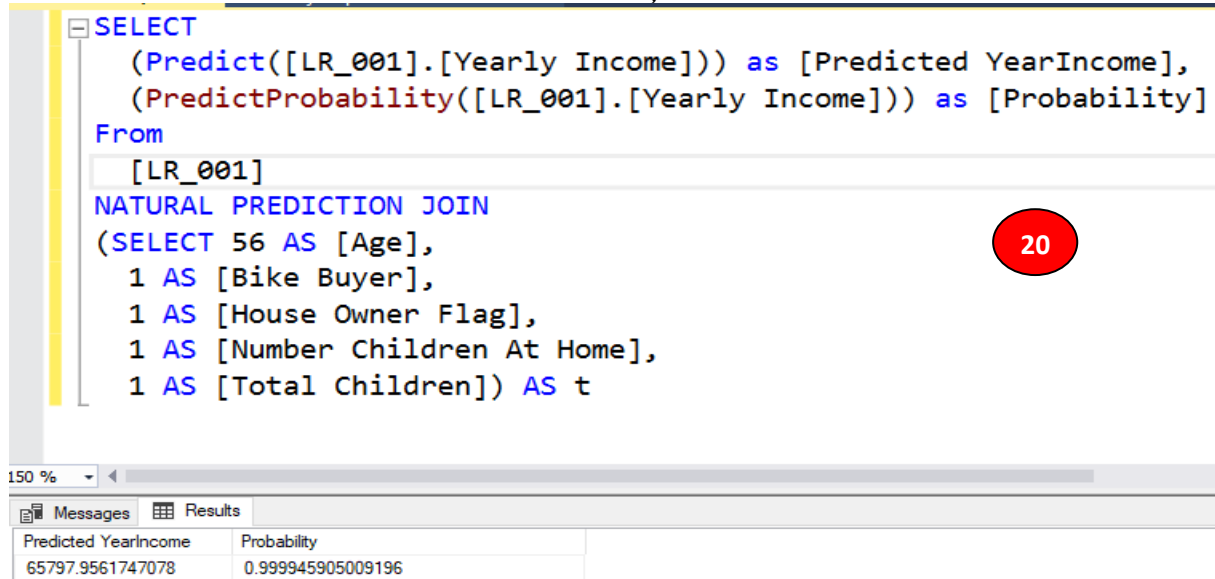
SELECT
    (Predict([LR_001].[Yearly Income])) as [Predicted YearIncome],
    (PredictProbability([LR_001].[Yearly Income])) as [Probability]
From
    [LR_001]
NATURAL PREDICTION JOIN
    (SELECT 56 AS [Age],
     1 AS [Bike Buyer],
     1 AS [House Owner Flag],
     2 AS [Number Cars Owned],
     1 AS [Number Children At Home],
     1 AS [Total Children]) AS t

```

| Predicted YearIncome | Probability       |
|----------------------|-------------------|
| 70506.5056044393     | 0.999945905009196 |

Este important să menționăm că, dacă nu avem anumite atribute, putem totuși obține rezultatele. Următoarea imagine arată valoarea de predicție/prognoză a modelului de regresie liniară atunci când Numărul de Mașini/Number Car Owned nu este cunoscut.

```
SELECT
(Predict([LR_001].[Yearly Income])) as [Predicted YearIncome],
(PredictProbability([LR_001].[Yearly Income])) as [Probability]
From
[LR_001]
NATURAL PREDICTION JOIN
(SELECT 56 AS [Age],
1 AS [Bike Buyer],
1 AS [House Owner Flag],
1 AS [Number Children At Home],
1 AS [Total Children]) AS t
```

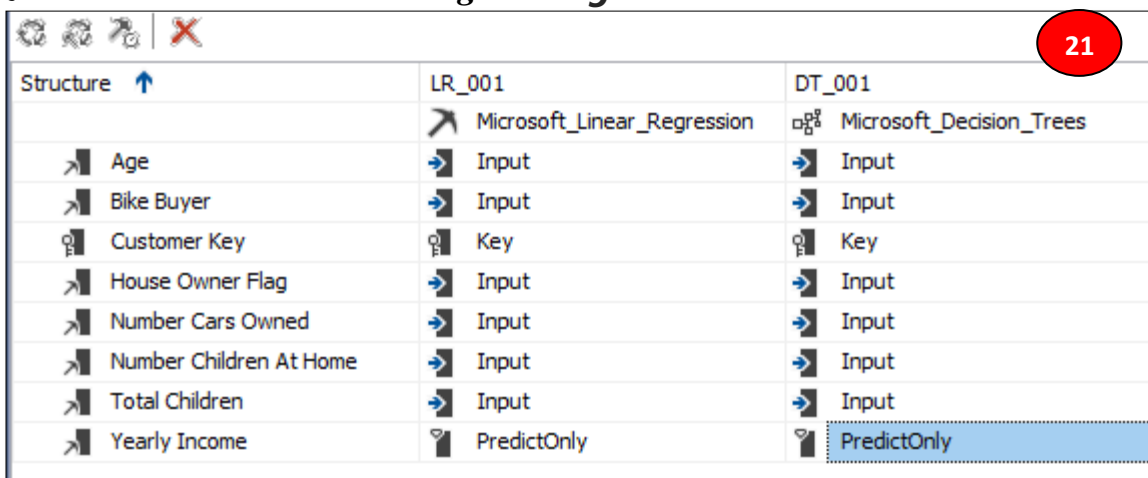


**Notă:** Când lipsește un atribut, partea respectivă a celui atribut va fi ignorată din întreaga ecuație.

Să verificăm/validăm ecuația regresiei liniare cu ajutorul tehnicii arborelui de decizie/ Decision Tree.

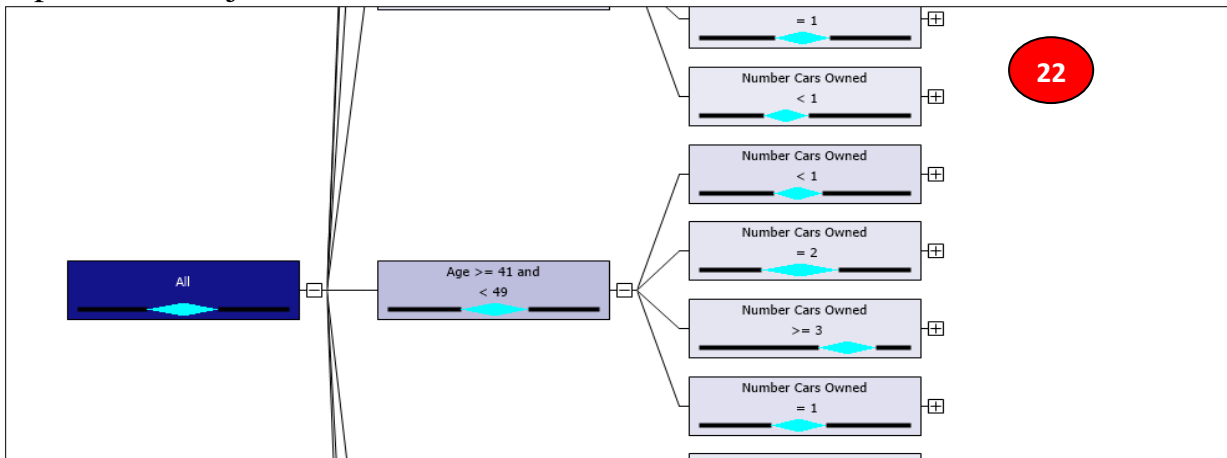
**Verificarea/Validarea ecuației regresiei liniare cu ajutorul arborelui de decizie**

Adăugarea unei alte tehnici de extragere a datelor în SQL Server este mult mai simplă. Puteți adăuga un alt Data Mining Model la atributele existente din fila/eticheta **Prognozarea cu Modelul Data Mining/Mining Model Prediction tab**



| Structure               | LR_001                      | DT_001                   |
|-------------------------|-----------------------------|--------------------------|
|                         | Microsoft_Linear_Regression | Microsoft_Decision_Trees |
| Age                     | Input                       | Input                    |
| Bike Buyer              | Input                       | Input                    |
| Customer Key            | Key                         | Key                      |
| House Owner Flag        | Input                       | Input                    |
| Number Cars Owned       | Input                       | Input                    |
| Number Children At Home | Input                       | Input                    |
| Total Children          | Input                       | Input                    |
| Yearly Income           | PredictOnly                 | PredictOnly              |

După procesarea Structurii Data Mining, putem observa Arborele Deciziei după cum se prezintă mai jos.

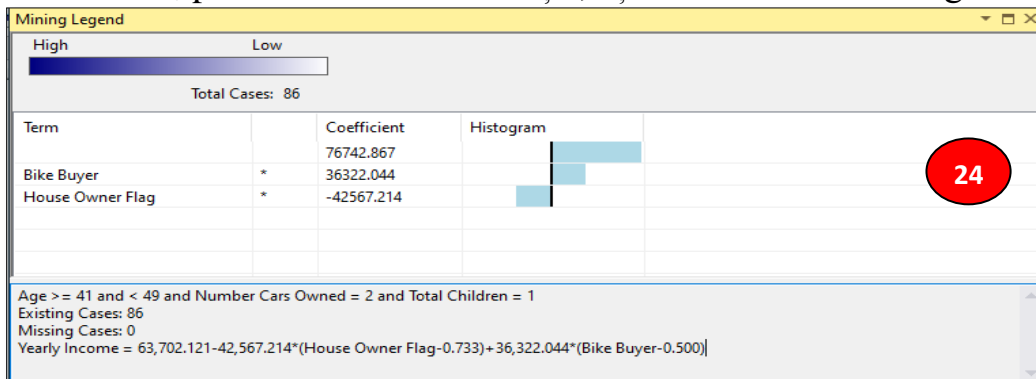


Să urmărim ecuația de la nodul principal. Ea arată după cum urmează:

$$\text{Venit anual} = 57,308.215 + 9,468.574 * (\text{Număr autoturisme proprietate}-1.503) + 415.816 * (\text{Age}-50.384) + 8,988.666 * (\text{biciclete Cumparator}-0,494) + 7,817.585 * (\text{Număr de Copii La Acasă}-1.004) - 1,939.209 * (\text{Total Copii}-1.844)$$

Putem descoperi că de fapt am obținut aceeași ecuație care a fost obținută din regresia liniară mai sus.

Mai mult decât acea ecuație, arborele de decizie are avantajul suplimentar de a avea o ecuație de nod cu adevărat “deosebită”/”smart one”. În Arborele de Decizie, dacă facem un clic pe fiecare nod, putem identifica o ecuație, așa cum se arată în imaginea de mai jos.



Acest lucru înseamnă că ecuația **VenitulAnual/YearlyIncome** va fi după cum urmează

$$63.702.121-42.567.214 * (\text{House Owner Flag } -0.733 ) + 36.322.044 * (\text{Bike Buyer } -0.500)$$

este valabilă pentru

$$\text{Age} \geq 41 \text{ și } < 49 \text{ și Number of cars owned} = 2 \text{ și Total children} = 1.$$

Următorul tabel prezintă ecuațiile diferite diferite noduri din *Arbore de Decizie*.

| Set de date  | Ecuatie   |
|--|---|
| Vârsta >= 73 și <81 și Total copii = 3   | Venit anual = 56.936.254-4.193.080 * (Cumpărător de biciclete-0.121) -20.137.503 * (Număr mașini deținute-1.994) -1 936.065 * (Vârsta-75.146)                               |
| Total copii = 3 ani și vârsta = 76   | Venit anual = 58.000.000-8.884.447 * (Cumpărător de biciclete-0.100)  |
| Vârsta = 73 ani și Total copii = 3   | Venit anual = 56.998.501 + 4.498.500 * (Cumpărător de biciclete-0.333)  |
| Vârsta >= 73 și <81 și Total copii = 2 și numărul copiilor acasă = 3                                   | Venit anual = 121.037.417 + 2.108.061 * (Vârsta-75.667) + 14.848.268 * (Cumpărător de biciclete-0.333)  |
| Vârsta >= 49 și <51 și total copii >= 4 și numărul de mașini deținute = 2 și numărul copiilor acasă <3 | Venit anual = 62.553.618-18.114.343 * (Număr copii acasă-1.897) + 5.525.516 * (Cumpărător biciclete-0.793) + 6.861.981 * (Vârsta-49.759) - 14.461.923 * (Total copii-4.017) |

Aceasta înseamnă că Arborii de Decizie sunt mai exacti decât regresia liniară Microsoft.

## Parametrii modelului

Urmează să înțelegem că, fiecare algoritm de Extragere a Datelor/Data Mining își are proprii săi parametric pentru a se potrivi cu datele și mediile de lucru instalate de Dvs pe calculator.

Algorithm Parameters

Parameters:

| Parameter                 | Value | Default | Range     |
|---------------------------|-------|---------|-----------|
| FORCE_REGRESSOR           |       |         |           |
| MAXIMUM_INPUT_ATTRIBUTES  |       | 255     | [0,65535] |
| MAXIMUM_OUTPUT_ATTRIBUTES |       | 255     | [0,65535] |
|                           |       |         |           |
|                           |       |         |           |
|                           |       |         |           |
|                           |       |         |           |

## **FORCE\_REGRESSOR**

Algoritmul Microsoft Linear Regression detectează în mod automat atributele cele mai potrivite și generează ecuația liniară. În această încercare, se poate să se renunțe la anumite atribute.

Cu toate acestea, se poate de forțat orice atribut care se dorește a fi inclus în ecuație, prin includerea lui în parametrii FORCE\_REGRESSOR. Dacă există mai multe atribute, le putem include toate, cum ar fi {Atribut 1}, {Atribut 2}.

## **REZUMAT**

În această lucrare, am discutat despre Regresia Liniară ca un algoritm de prognoză din instrumental Data Mining Microsoft Sql Server.

Algoritmul Regresiei Lineare are la bază construirea Arborilor de Decizie, ca în cele din urmă să identificăm că Arborii de Decizie pot fi folosiți și ca algoritmi pentru efectuarea regresiei liniare.