# <mark>PARTEA 3</mark>

Încarcarea si instalarea pe SSMS a DWH-ului exemplu AdventureWorksDW2017.bak (similar e si 2014-2016) pentru lucrul cu regresia liniara multiplă, in calitate de testare si 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



 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:

Object Explorer	· · · · · · · · · · · · · · · · · · ·
Connect 🕶 🌹 🎽 🔳	▼ Ċ -₩
🖃 🐻 NISARG-PC (SQI	Server 14.0.1000.169 - NISARG-PC\Nise
⊟ ■ Database	New Database
🕀 💻 Databa	Attach
🕀 🗑 Advent	Restore Database
⊞ 🗎 DBA	Restore Files and Filegroups
⊞	Filter •
🗄 🛑 Server Ob	Deploy Data-tier Application
🕀 💻 Replicatic	Import Data-tier Application
🕀 💻 PolyBase	Start PowerShell
🕀 📕 Always Or	Reports
🕀 💻 Managem	
H Integratio	Refresh
	er

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

🔜 Restore Database -				l l	3	—	
😵 No backupset selected to be restored.				_	$\sim$		
Select a page	🗊 Script 🕞 😮	Help					
Mage General Files Files Ø Options	Source Databas Device Destination Database: Restore plan Backup sets Restore N	tabase:	Туре	Server Dat	abase	Position	imeline
Connection							
¥¥ NISARG-PC [NISARG-PC\Nisarg]							
View connection properties							
Progress							
C Ready	<				X	/erify Bac	> kup Media

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

	গ - 👅 🗋 📸 🗧	🔀 Restore Database -	4	- 🗆 X	x - Microsoft Word	- 🗆 X
Hom	me Insert Page Layout Reference 💈	No backupset selected to be restored.	· · · · ·			(
Paste ✓ Fo Clipboa	ut opy ormat Painter B / U * abe x, x / ard 6 Font	ielect a page © General © Files © Options	Script Pelo Source Database: Database: Database: Database:	Select backup devices	AaBb AaBbC: AaBbCc Heading 2 Heading 3 Heading 4 tyles	D AaBbCc. A AFind * Subtitle Change Styles Styles Editing
- - -	Restore Databas	Connection d DESKTOP-8CNV0VU/MSSQLSE RVFE7017	Destination Database: Restore to: Restore plan	Specify the backup media and its Backup media type: Backup media:	location for your restore operation.	
- - - - - - - - - - - - - - - - - - -		ENVOVU,MSSQLSERVER2017 4. MSSQLSERVER2017/MSSQL/Backup VentureWorks2014.bak VentureWorks2017.bak VentureWorksDW2014.bak VentureWorksDW2017.bak	- C		OK Can	Remove Contents

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.

💀 Restore Database - AdventureWorks	s2016			- 0	$\times$
🕕 Ready					
Select a page	Seript 🔹 🕜 Help				
General     Files     Options	Restore database files as				
	Data file folder :	C:\M	SSQL\SQLData		
	Log file folder :	C:\M	SSQL\SQLLog		
	Logical File Name	File Type	Original File Name	Restore As	
	AdventureWorks2016_Data	Rows Data	C:\Program Files\Microsoft SQL S	C:\MSSQL\SQLData\AdventureW	
	Adventureworks2016_Log	Log	C:Program Files Microson SGL S	C:IMSSQLISQLLogWaventurewor	
Connection					
ý掌 NISARG-PC [NISARG-PC\Nisarg]			$\backslash$		
View connection properties			\		
Progress				$\mathbf{N}$	
Oone Done				$\mathbf{A}$	

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:

롰 Restore Database - AdventureWo	rks2016						$\times$
Restoring: AdventureWorks2016-Full Da	atabase Backup					100%	Stop
Select a page ∬ General ∯ Files ∅ Options	Script - O Help Source Database: Device: Database:	E:\Backup\AdventureWorks AdventureWorks2016	s2016.bak				>
	Destination Database: Restore to: Restore plan	AdventureWorks2016 The last backup taken (09 A	spril 2018 14:44:	02)		Timeline	~
	Backup sets to restore:						
	Restore Name	016-Full Database Backup	Component	Type	Server BARBKESS24\MSSQI	2016RTM	
ψ₽ NISARG-PC [NISARG-PC\Nisarg]	Microsoft SQL Server	Management Studio	7	×			
View connection properties Progress	Database 'A	dventureWorks2016' restore	d successfully.				
$\circ$			ОК		Verify B	ackup Medi	•
				ок	Cancel	Help	

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



LA URMATORUL PAS LANSĂM SI CONECTĂM DWH/BD DIN SSMS 2017/2019 LA VS SI APOI CONTINUĂM LUCRAREA / PASILOR CE URMEAZĂ

# **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:

$$\mathbf{Y} = \mathbf{a} \mathbf{X}_1 + \mathbf{b} \mathbf{X}_2 + \ldots + \mathbf{z} \mathbf{X}_n + \mathbf{C},$$

unde X<sub>i</sub>, 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 *Adventure WorksDW* și vizualizarea de exemplu a atributului *vTargetMail*. Algoritmul de extragere a datelor este unul standart si constă din 3 nivele.

## <u>Grupul de Activități 1</u>

- 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.

Create the Data Mining Structure Specify if mining model should be created and select the most applicable technique
Create mining structure with a mining model
Which data mining technique do you want to use?
Microsoft Linear Regression
Description:
The Microsoft Linear Regression algorithm is a regression algorithm that works well for regression modelin. This algorithm is a particular configuration of the Microsoft Decision Trees algorithm, obtained by disabiling splits (the whole regression formula is built in a single root node). The algorithm supports the prediction of continuous attributes.

În acest algoritm se folosește *Microsoft Decision Trees technic / 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 il vom precăuta la sfârșitul lucrării de laborator.

**Tabelul** *vTargetMail* va fi tebelul de *Case/Caz/Studiu de caz* și ne va oferi să alegem atributele relevante, așa cum este indicat in 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, Vârstă, 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 lineară standard.

În alti algoritmi putem selecta de exemplu un **cimp textual**, in 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 annual/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.

A Data Mining Wizard	-	
Specify Columns' Content and Data T Specify mining structure columns' content and data ty	pe.	7
Mining model <u>s</u> tructure:		
Columns	Content Type	Data Type
Age	Continuous Lo	ng
Bike Buyer	Continuous Lo	ng
P Customer Key	Key Lo	ng
House Owner Flag	Continuous Lo	ng
Number Cars Owned	Continuous Lo	ng
Number Children At Home	Continuous Lo	ng
Total Children	Continuous Lo	ng

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

În celelalte ecrane din asistentul de extragere a datelor, sunt utilizate setările implicite. Acesta este *Exploratorul de soluții/ Soluțion Explorer* pentru extragerea datelor in *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 d *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 atribute 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 acestTree 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.

Mining Legend		
High	Low	
Total (	Cases: 18	3484
Term		Coefficient Histogra
		13406.948
Age	*	416.002
Bike Buyer	*	8988.432
Number Cars Owned	*	9468.323
Number Children At Home	*	7818.527
Total Children	*	-1940.134
All		
Existing Cases: 18484		
Missing Cases: 0	040 124*/	Tetal Children 1 944) - 9 009 422*/Pike Puwer 0 404) - 9 469 222*/Number Care Owned
1.503)+416.002*(Age-50.380)+	7 818.527	iotal Children - 1.044)+ 8,988.452" (Bike Buyer-0.494)+ 9,488.525" (Number Cars Owned *(Number Children At Home-1.004)
1.503)+416.002*(Age-50.380)+	7,818.527	/*(Number Children At Home-1.004)

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

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



Să vedem cum putem prezice **YearlyIncome** utilizînd 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*.

Mining Model		Sing	gleton Query Input			٥		
Image: Select Mode	yer er Key Dwner Flag Cars Owne Children Ar hildren Income		Mining Model Column Age Bike Buyer House Owner Flag Number Cars Owned Number Children At Home Total Children Yearly Income	Value 56 1 1 2 1 1	17			
Source	Field			Alias	Show	Group	And/Or	Criteria/Argument
Prediction Function	í Predict			Predicted YearIncome	$\checkmark$			[LR_001].[Yearly Income]
Prediction Function	🎄 PredictProt	ability		Probability	$\checkmark$			[LR_001].[Yearly Income]

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

🔊 - 💾 🗇 🔯	18
Predicted YearIncome	Probability
70506.5056044393	0.999945905009196

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

	<b>□</b> SELECT	
	(Predict([LR_001].[Yearly Income])) as [Predicted YearIncome],	
	(PredictProbability([LR_001].[Yearly Income])) as [Probability	1
	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	
150 %	• 4	
B <u>r</u>	Messages III Results	
Pre	cted YearIncome Probability	
70	0.999945905009196	

Este important să mentionăm 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.



Notă: Când lipsește un atribut, partea respectivă a acelui 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* 

R R B 🗙		21
Structure 🔨	LR_001	DT_001
	Microsoft_Linear_Regression	마람 Microsoft_Decision_Trees
Age	Input	Input
> Bike Buyer	Input	Input
စ္နဲ့ Customer Key	စ္နဲ့ Key	စူ 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 prezinta mai jos.



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

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

Mining Legend					<b>-</b> □ ×
High	Low				
Tc	otal Cases: 86				
Term		Coefficient 76742.867	Histogram		
Bike Buyer	*	36322.044			24
House Owner Flag	*	-42567.214			
Age >= 41 and < 49 and N Existing Cases: 86 Missing Cases: 0 Yearly Income = 63,702.12	lumber Cars O 1-42,567.214*(	wned = 2 and Total House Owner Flag-	Children = 1 0.733)+36,322.044*(Bike Buy	er-0.500)	<u></u>

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

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

este valabilă pentru

Age > = 41 și <49 și Number of cars owned = 2 și Total children = 1.

Set de date	Ecuație 25
Vârsta> = 73 și <81 și Total	Venit anual = 56.936.254-4.193.080 *
copii = 3	(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 =	Venit anual = 58.000.000-8.884.447 *
76	(Cumpărător de biciclete-0.100)
Vârsta = 73 ani și Total copii	Venit anual = 56.998.501 + 4.498.500 *
= 3	(Cumpărător de biciclete-0.333)
Vârsta> = 73 și <81 și Total	Venit anual = 121.037.417 + 2.108.061 *
copii = 2 și numărul copiilor	(Vârsta-75.667) + 14.848.268 * (Cumpărător de
acasă = 3	biciclete-0.333)
Vârsta> = 49 și <51 și total	Venit anual = 62.553.618-18.114.343 * (Număr
copii> = 4 și numărul de	copii acasă-1.897) + 5.525.516 * (Cumpărător
mașini deținute = 2 și	biciclete-0.793) + 6.861.981 * (Vârsta-49.759) -
numărul copiilor acasă <3	14.461.923 * (Total copii-4.017)

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

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 prpriii săi parametric pentru a se potrivi cu datele și mediile de lucru instalate de Dvs pe calculator.

Algorithm Parameters			
Parameters:			26
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 doreste 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 acastă 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.