Audit financiar, XV, Nr. 3(147)/**2017**, *354-372*
ISSN: 1583-5812; ISSN on-line: 1844-8801

Componenta Excel *Data Mining*. Aplicații în audit și raportări financiare

Daniel HOMOCIANU,
Universitatea „Alexandru Ioan Cuza" din Iaşi (UAIC),
E-mail: daniel.homocianu@feaa.uaic.ro
Dinu AIRINEI,
Universitatea „Alexandru Ioan Cuza" din Iaşi (UAIC),
E-mail: adinu@uaic.ro

Rezumat

Argumentele de performanță în luarea deciziilor bazate pe date economice solicită uzual un management bun al multiplelor formate de date şi, de asemenea, viteză de procesare, flexibilitate, portabilitate, automatizare, putere de sugestie şi uşurință de utilizare. Lucrarea vine cu idei teoretice şi exemple practice în favoarea utilizării componentei Excel Data Mining pentru motivele menționate anterior. Cele mai multe exemple includ figuri legate la scenarii video construite de autori şi parte a unei liste on-line interactive cu optsprezece piese. Împreună ele contribuie la înțelegerea celor mai multe cerințe care trebuie îndeplinite pentru a avea exemple valide şi rezultate utile.

Cuvinte-cheie: *Date economice şi financiare, foi de calcul, Data Mining (DM), exemple.*

Clasificare JEL*: C61, D81, D83, M42*

Vă rugăm să citați acest articol astfel:
Homocianu, D. and Airinei, D. (2017), The Excel Data Mining Add-in. Applications in audit and financial reports, *Audit Financiar*, vol. XV, no. 3(147)/2017, pp. 451-468, DOI: 10.20869/AUDITF/2017/147/451

Link permanent pentru acest document:
http://dx.doi.org/10.20869/AUDITF/2017/147/451
Data primirii articolului: 17.03.2017
Data revizuirii: 13.04.2017
Data acceptării: 20.04.2017

Introducere

Lucrarea începe de la o serie de tehnici utilizate de cele mai multe instrumente de exploatare a datelor (en. *Data Mining)* pentru volume mari de date din baze de date şi prezintă avantajele utilizării foilor de calcul ca aplicații client (msdn.microsoft.com/.../dn282385.aspx). Acestea din urmă sunt foarte familiare utilizatorilor finali şi au o interfață care integrează limbaje de programare sau de configurare pentru aplicații de birou precum Visual Basic pentru Applicații (VBA) în cazul Microsoft Excel (msdn.microsoft.com/.../ee814737.aspx) şi Google Apps Script pentru foi de calcul (en. Google Sheets) (developers.google.com/.../sheets). Multe funcții şi facilitați avansate de procesare, analiză, reprezentare şi simulare bazate pe principii de interactivitate şi dinamică au impact considerabil asupra capacității utilizatorilor de a percepe, interpreta, înțelege şi gestiona informații complexe în diferite situații.

Conceptul de exploatare a datelor înseamnă în esență identificarea supervizată de tipare nedescoperite şi relații ascunse existente în seturi foarte mari de date (searchsqlserver.techtarget.com). Inmon care este un bine cunoscut „guru" în depozite de date (computerweekly.com) a dat una dintre cele mai concise definiții ale exploatării datelor (Inmon şi Linstedt, 2014) şi anume analiza unor cantități mari de date pentru a găsi tipare precum grupuri de înregistrări sau înregistrări şi dependențe neobişnuite. Inițiativele de exploatare a datelor vin de obicei de la departamentele de marketing şi vânzări cu amănuntul şi sunt potrivite pentru organizații care au baze de date de dimensiuni foarte mari (Airinei, 2002). Acest concept este strâns legat de acela de sisteme de asistare a deciziilor (en. *Decision Support Systems* - DSS) orientate spre date şi de inteligența în afaceri (en. Business Intelligence - BI) – în special acela pentru obiective strategice (dssresources.com/…id=174) care solicită volume uriaşe de date (bi-insider.com). Deşi termenul de inteligență în afaceri este cunoscut ca fiind un set de concepte şi metode de îmbunătățire a luării deciziilor care a apărut în anii '90 (Howard Dresner din grupul Gartner - dssresources.com/…/dsshistory.html), cercetările publicate în literatura de specialitate indică abordări cu 15 ani mai devreme (Cleland şi King, 1975 şi 1976), care conțin referințe clare la BI, planificatorii/ administratorii afacerii şi manageri, respectiv la luarea deciziilor.

Conform concluziilor lui Dan Power (dssresources.com/…id=199), instrumentele de exploatare a datelor includ: raționamentul bazat pe caz, vizualizarea datelor (în special grafice, arbori şi grupuri), interogări şi analize vagi (en. fuzzy), algoritmi genetici şi rețele neuronale.

Începând de acum câțiva ani suntem martorii implementării acestui concept şi a modelelor asociate prezentate nu doar în aplicații dedicate sistemelor de gestiune a bazelor şi depozitelor de date, dar şi în module ale aplicațiilor de tip foi de calcul care le folosesc aşa cum se sugerează chiar din titlul lucrării. Pare evident când ne gândim că asemenea produse dedicate au permis construirea de structuri şi modele DM plecând chiar de la un singur tabel (uzual ca agregare a mai multor tabele dintr-o bază de date).

Aplicabilitatea în audit a elementelor teoretice şi practice din acest articol, în special cel al performanței (Fraser, 1998) şi al raportărilor financiare, este justificată plecând de la o nevoie specifică de a valoriza structurile existente de date (adesea înregistrări în tabele şi tabele în baze de date) şi de a obține rapid şi cu cost minim rapoarte care pot prezenta informații clare despre relația de cauzalitate existentă între eficacitate (rezultate efective/estimate comparate cu cele propuse) şi eficiență (resurse consumate comparate cu rezultate obținute/estimate).

Exemplele concrete din acest articol susțin anumite concluzii desprinse din analiza literaturii de specialitate, şi anume:

- utilitatea abordării misiunilor de audit folosind tehnici de exploatare a datelor (Vintilescu ş.a., 2010) complementar metodelor clasice de analiză a riscurilor şi intervenției la fața locului;
- consacrarea existenței unor zone posibile de integrare a procesului de exploatare a datelor cu procesele de audit grupate pe faze (Sirikulvadhana, 2002), precum planificarea, execuția, documentarea şi finalizarea;
- utilizarea de exemple specifice (Wang şi Yang, 2009) cum ar fi rețelele neuronale pentru evaluarea riscului, detectarea erorilor şi a

fraudelor, determinarea preocupărilor pentru continuitatea activității unei firme, evaluarea dificultăților financiare, precum şi realizarea de predicții de faliment şi arbori de decizie pentru analiza de faliment, analiza de faliment bancar şi a riscului de credit);

- progresul celor mai noi instrumente software care implementează algoritmi de exploatare a datelor şi faptul că mulți utilizatori le considerau până nu demult ca nefiind foarte prietenoase (Chersan ş.a., 2013) şi necesitând abilități tehnice destul de avansate.

Lucrarea îşi propune şi eliminarea unor confuzii privind utilizarea valorilor tip ştampilă temporală (date calendaristice, părți din ea sau valori înlocuitoare) la exploatarea seriilor de timp cu date economice (de exemplu, volumul vânzărilor recunoscut ca factor de influență directă a nivelului anumitor indicatori financiari precum rezultatul din exploatare).

1. Metodologia de cercetare

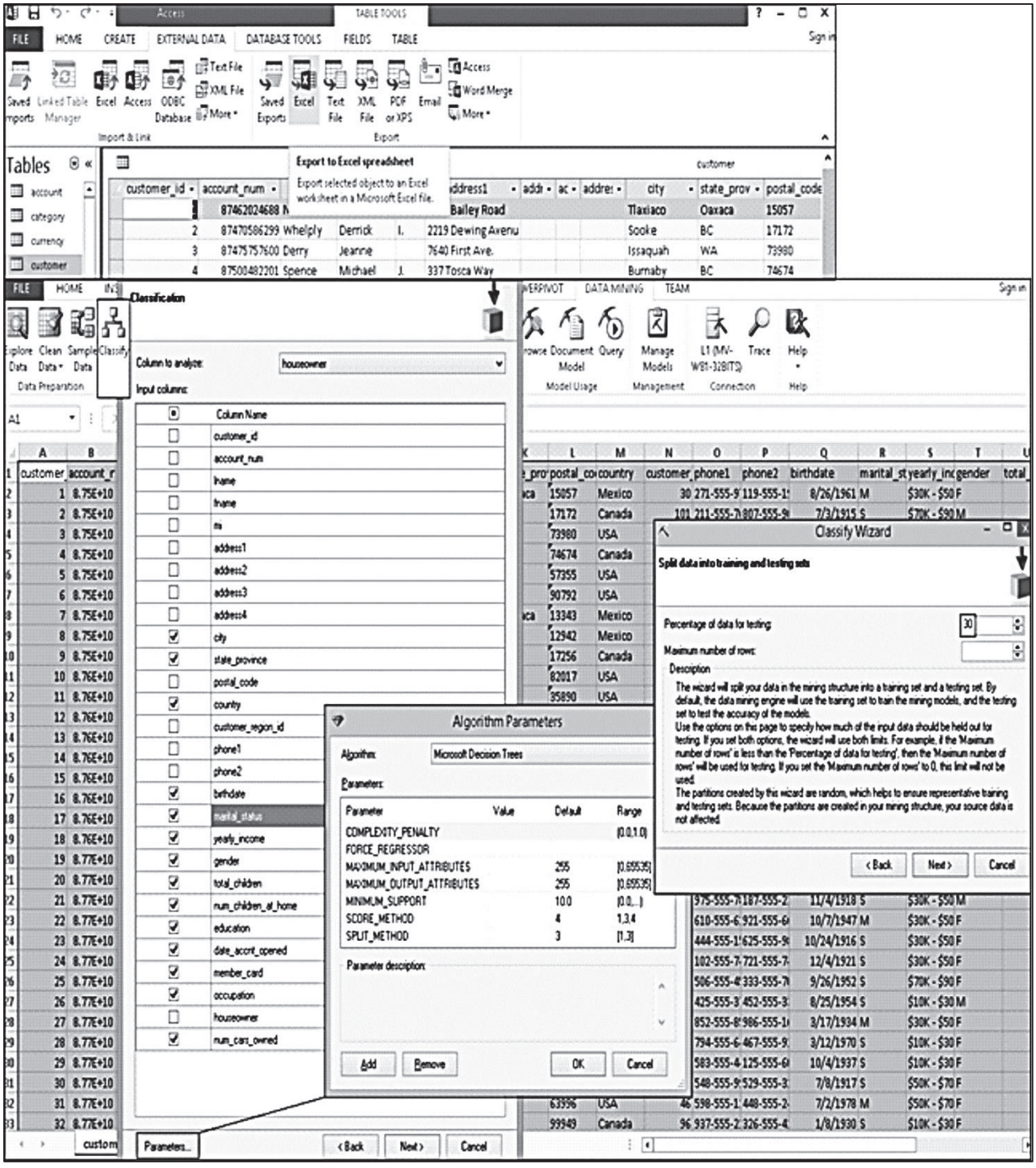
Datele sursă pentru exemplele prezentate în această lucrare provin din două mostre de baze de date Microsoft. Primul set de exemple a fost creat plecând de la un fişier bază de date Access denumit *foodmart* (sites.google.com/.../supp4excel2datamining) inițial disponibil pe CD-ul de instalare a unei versiuni anterioare de Microsoft (MS) SQL Server. Al doilea provine dintr-o bază de date mostră de tip MS SQL Server denumită *„AdventureWorksDW2012_Data.mdf"* deja instalată şi pregătită pentru utilizare într-o aplicație virtuală (y2u.be/Xs2SWtBqdzI) Windows 8.1 pe 32 biți pe care am folosit-o pentru acest articol. Această aplicație a beneficiat de licența Microsoft Imagine/fostă Dream Spark de software educațional pentru toate aplicațiile instalate în interiorul ei şi a fost optimizată pentru Oracle Virtual Box. De fapt SQL Server 2012 (sau 2008) este o componentă necesară pentru instalarea Excel *Data Mining* care este detaliată în al doilea tutorial video (lista menționată mai jos). Deşi servesc construirii de exemple şi materiale video suport corespunzătoare (tutoriale – lista creată de autori şi disponibilă la adresa goo.gl/JDDtFp), asemenea date au doar caracter instructiv în această cercetare de natură pronunțat aplicativă, asemănările cu realitatea fiind doar o coincidență.

2. De la tipare intuitive la o analiză profundă plecând de la surse simple de date precum tabelele

Primul exemplu pe care l-am ales a fost menit să clasifice prin generarea unui arbore decizional în care variabila estimată a fost una de tip categoric, având două valori posibile (proprietar de casă/*houseowner*: Da/*Yes* – *Y* sau Nu/*No* – *N*) în funcție de alte câmpuri (vezi **Figura nr. 1**), conținând informații despre clienți (un export în Excel din tabelul cu clienți (*customer)* din baza de date *foodmart*).

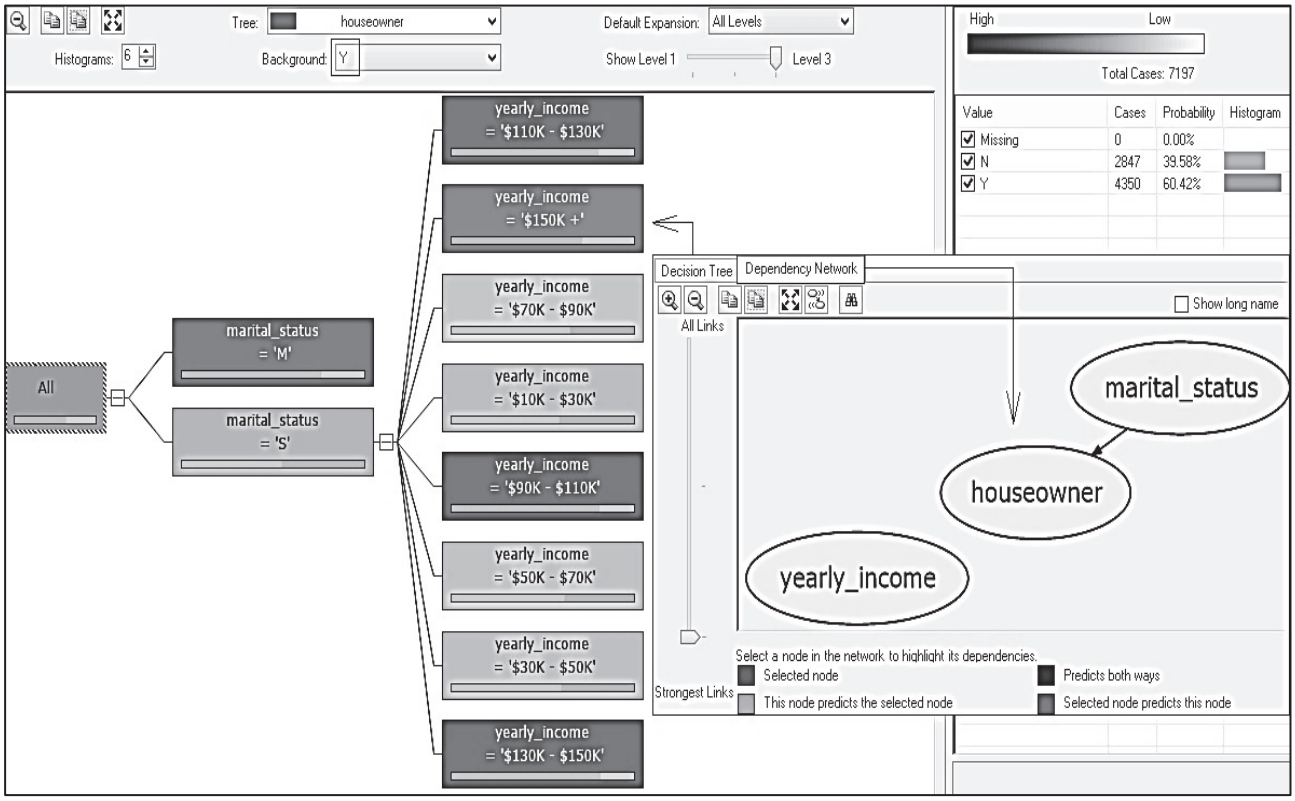
Rezultatele (**Figura nr. 2**) acestei clasificări cu reglajele implicite (algoritmul Arbori Decizionali Microsoft şi 30 procente din date pentru test) sunt: (1) un arbore decizional şi (2) o rețea de dependențe ce indică cele mai importante variabile ce influențează valoarea atributului *houseowner* şi anume: starea civilă/*marital_status* (căsătorit/*married* – *M* sau necăsătorit/*single* – *S*) şi venitul anual/*yearly_income* (opt praguri în mii/*K* de dolari/*$*: '$10K - $30K', '$30K - $50K', '$50K - $70K', '$70K - $90K', '$90K - $110K', '$110K - $130K', '$1300K - $150K', '$150K +'), în această ordine a importanței.

Figura nr. 1. Exemplu de export urmat de folosirea opțiunii de clasificare a componentei Excel Data Mining şi configurarea atributelor de intrare



Sursa: Tutorialul video creat de autori: y2u.be/Nx9xqCX1DjY

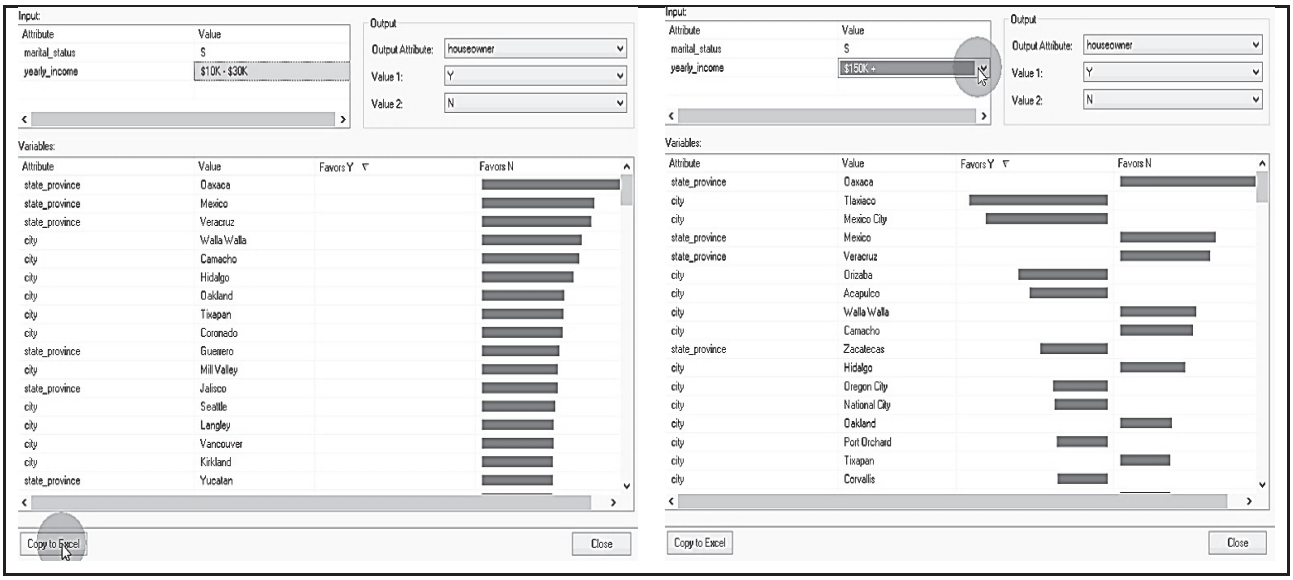
Figura nr. 2. Exemplu de rezultat al clasificării plecând de la date dintr-un tabel cu clienți şi creat folosind algoritmul Arbori Decizionali Microsoft



Sursa: Tutorialul video creat de autori: y2u.be/Nx9xqCX1DjY

Aşa cum se observă în partea stângă a **Figurii nr. 2** ramurile ce indică o probabilitate mai mare pentru Da (*Y – houseowner*) sunt mai închise, restul fiind colorate cu o nuanță mai deschisă. Putem de asemenea, observa că variabila *houseowner* depinde esențial de variabila *marital_status* (în partea dreaptă a **Figurii nr. 2** – bara de derulare pe legăturile cele mai puternice/*Strongest Links*) şi apoi de *yearly_income* (bara pe toate legăturile/*All Links*). Aceasta se poate deduce şi direct din arborele decizional în care un nod mai aproape de rădăcină exprimă un test (inf.ucv.ro) aferent atributului *marital_status*. Accesând opțiunea *marital_status='M'* (nod terminal) am obținut o probabilitate peste 74% în toate cele zece teste făcute în aceeaşi configurație (coloane de intrare respectiv de analizat, algoritm, procentaj date de test).

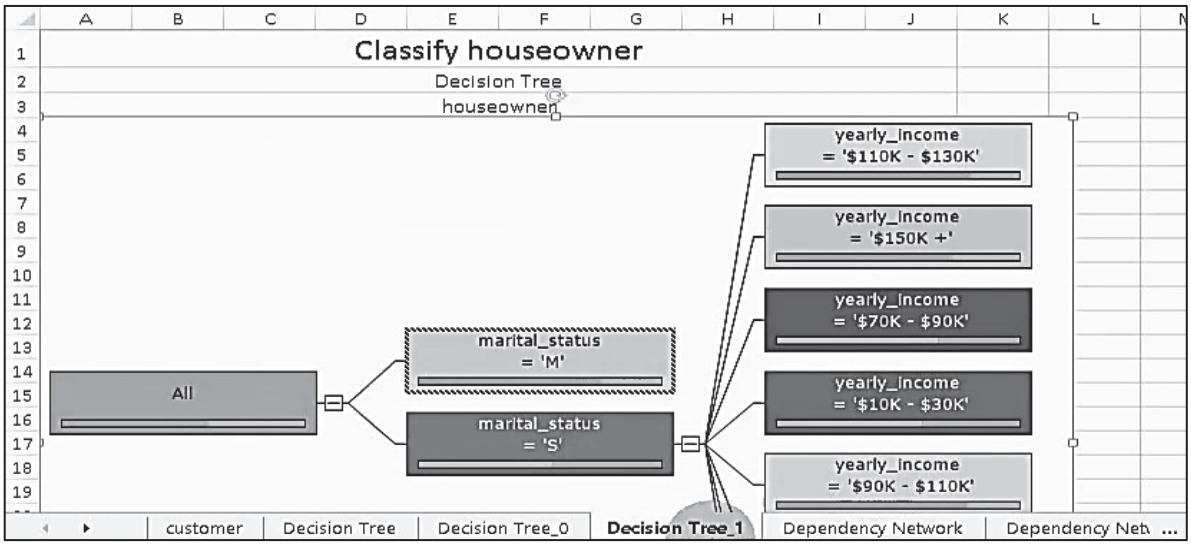
Figura nr. 3. Exemplu de analiză discriminatorie după aplicarea regresiei logistice (profs.info.uaic.ro) pentru aceleaşi condiții de mai sus şi specificând cele două variabile de intrare cu impact major

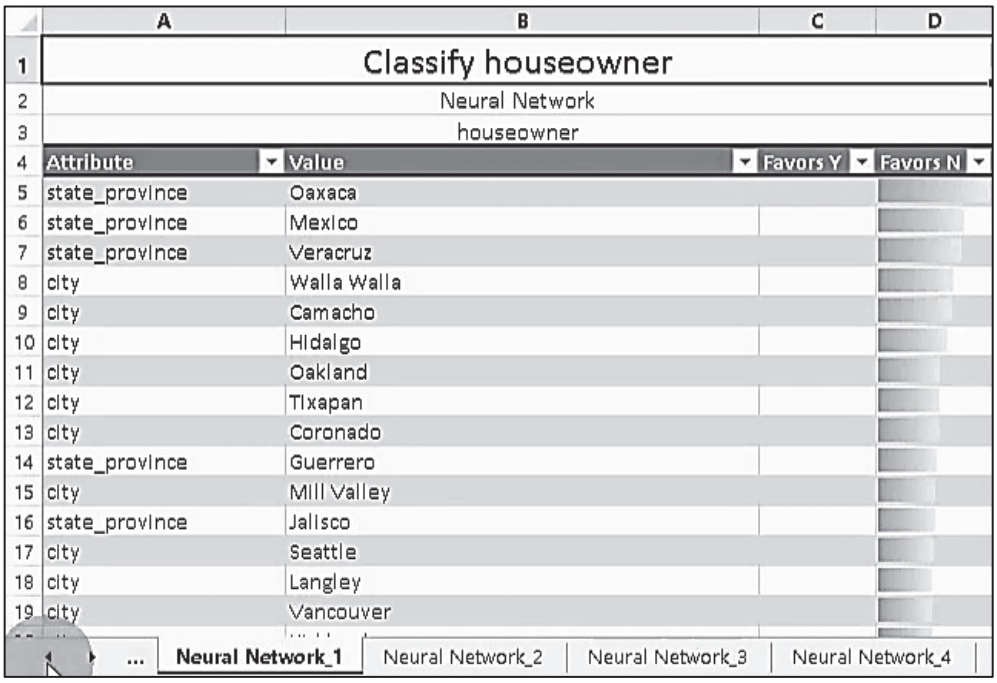


Sursa: Tutorialul video creat de autori: y2u.be/-6jzQuyTjlo

În imaginile anterioare (**Figura nr. 3**) am încercat să punem în evidență cum anume am preconizat probabilitățile pentru clienți de a face parte din cele 2 categorii de răspuns binar (onlinecourses.science.psu.edu): proprietar de casă sau nu, în funcție de anumite variabile explicative şi de valorile lor. Am realizat analiza discriminatorie redată în mod parțial mai sus (**Figura nr. 3**) plecând de la un alt algoritm, şi anume cel de regresie logistică, implementat de Microsoft, folosind o variație a algoritmului de tip rețele neuronale (msdn.microsoft.com/.../ms174828.aspx) care este mai uşor de instruit.

Figura nr. 4. Exemplu de rezultate ale funcționalității „Copiați în Excel" ("Copy to Excel")





	A	B	C	D
1	Classify houseowner			
2	Neural Network			
3	houseowner			
4	Attribute	Value	Favors Y	Favors N
5	state_province	Oaxaca		
6	state_province	Mexico		
7	state_province	Veracruz		
8	city	Walla Walla		
9	city	Camacho		
10	city	Hidalgo		
11	city	Oakland		
12	city	Tlxapan		
13	city	Coronado		
14	state_province	Guerrero		
15	city	Mill Valley		
16	state_province	Jalisco		
17	city	Seattle		
18	city	Langley		
19	city	Vancouver		

... Neural Network_1 | Neural Network_2 | Neural Network_3 | Neural Network_4

Sursa: Tutorialele video create de autori: y2u.be/Nx9xqCX1DjY şi y2u.be/-6jzQuyTjlo

Funcția "*Copy to Excel*" ne-a ajutat să trimitem rezultatele înapoi în Excel ca foi de calcul noi cu capturi de ecran (în partea stângă, **Figura nr. 4** pentru arbori decizionali) sau, mai important, seturi de date cu efecte vizuale implicând de obicei formatări condiționate realizate automat (un exemplu este analiza discriminatorie bazată pe regresie logistică – în partea dreaptă a **Figurii nr. 4**).

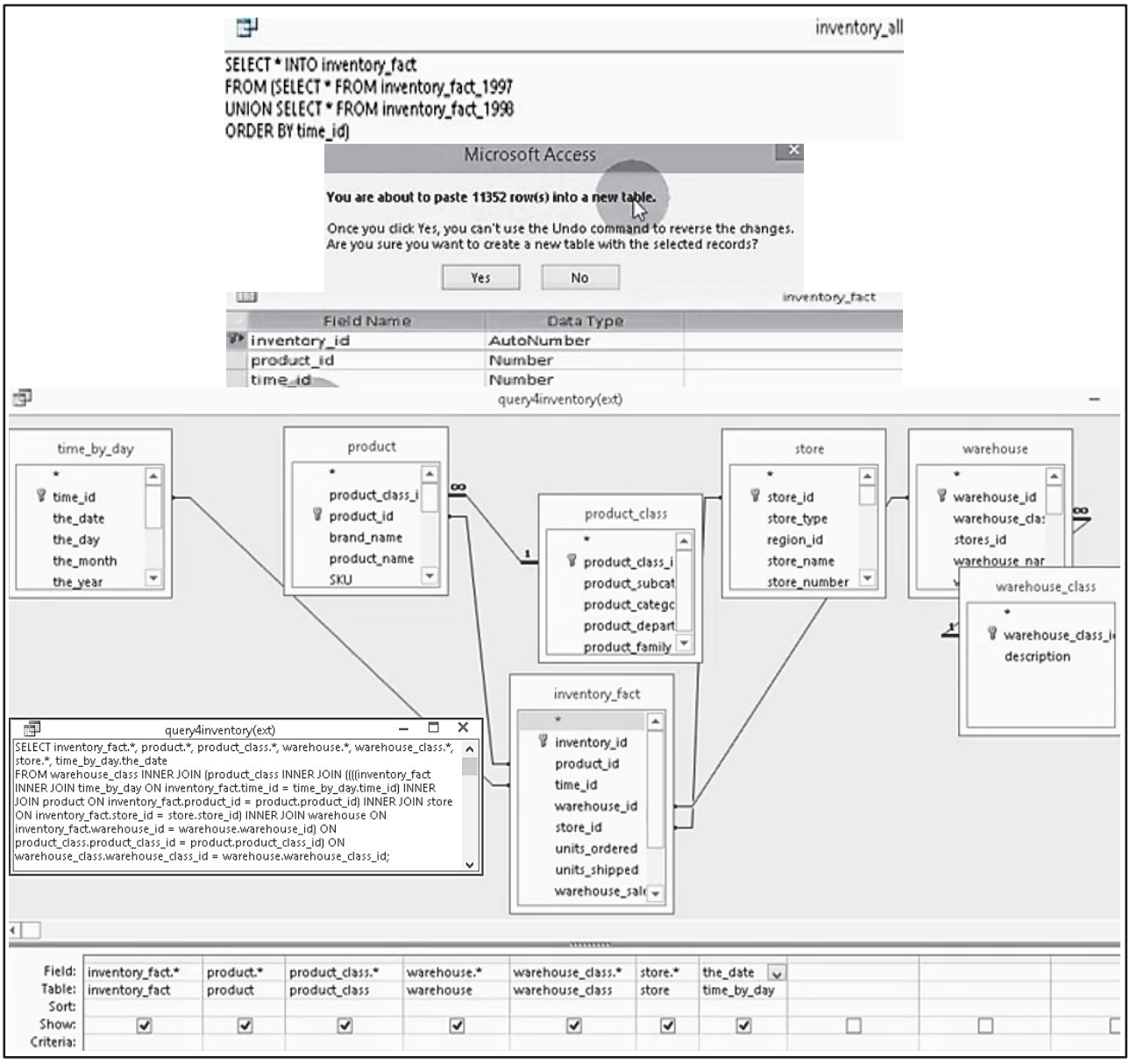
Plecând de la rezultatele descrise anterior (**Figurile nr. 1-4**) se pot dezvolta exemple similare care să rezolve inclusiv problema încadrării clientului (corespunzând fazei de acceptare/menținere în demersul de audit) într-una dintre cele două categorii: acceptabil/neacceptabil, plecând de la un istoric validat al unor astfel de decizii, în format tabelar care să includă şi multe alte atribute descriptive (zona geografică, sectorul de activitate, numărul mediu de angajați, cifra de afaceri a clientului, evoluțiile anumitor indicatori, nivelul onorariului etc.).

3. Cumularea datelor istorice şi folosirea câmpurilor descriptive din tabelele bazei de date

Rapoartele dinamice şi interactive care răspund la multe nevoi informaționale şi cu care suntem atât de familiarizați dar şi cele statice mai vechi, ca şi fotografii ale informațiilor la momente precise ce generează mai multe întrebări decât răspunsuri (Rasmussen ş.a., 2002), pot utiliza atât date curente cât şi istorice. Prima categorie este reprezentată de date din sistemele de procesare a tranzacțiilor (TPS) care se referă în mod obişnuit la anul curent, în timp ce a doua categorie presupune în esență date ce implică o perioadă mai mare de timp. Proporția utilizării celor două categorii depinde în mod esențial de nevoile decizionale (la nivel operațional, tactic sau strategic). Pentru minimizarea redundanței şi a dependenței datelor sau din rațiuni de spațiu de stocare şi de nevoie de viteză de scriere (deshpande.mit.edu) schema unei surse de date tradiționale de tip relațional este gândită în mod uzual ca fiind formată din mai multe tabele obținute prin aplicarea principiilor normalizării (w3schools.in). Mai mult, din cauza unor motive suplimentare de performanță (nevoi de viteză de citire respectiv de scriere) datele istorice trebuie separate de cele curente. Ambele categorii includ în esență înregistrări din tabele cu tranzacții (de exemplu cheltuieli, vânzări, examene etc.) diferența fiind dată de valoarea ştampilei temporale. Aceasta explică de ce aceste tabele încărcate doar cu date istorice sunt redenumite cu un indicativ de timp, arhivate şi separate de restul sistemului tranzacțional pentru a-i îmbunătăți

performanța operațională (curentă). Când este nevoie de volume mari de date istorice pentru analize bazate pe interogări ad-hoc, sistemele trebuie să procedeze invers prin agregarea într-un singur tabel (sursă pentru un tabel de fapte într-un depozit de date) a tuturor înregistrărilor din arhivele istorice ale tabelelor cu tranzacții (de acelaşi tip ca cel rezultat). În cele mai multe cazuri, aceasta generează avantajul unui potențial crescut de identificare de tipare, dar vine şi cu dificultăți legate de punerea laolaltă a datelor într-un format comun şi consistent, în special atunci când aplicațiile şi structura sursei de date s-au modificat şi ele în timp.

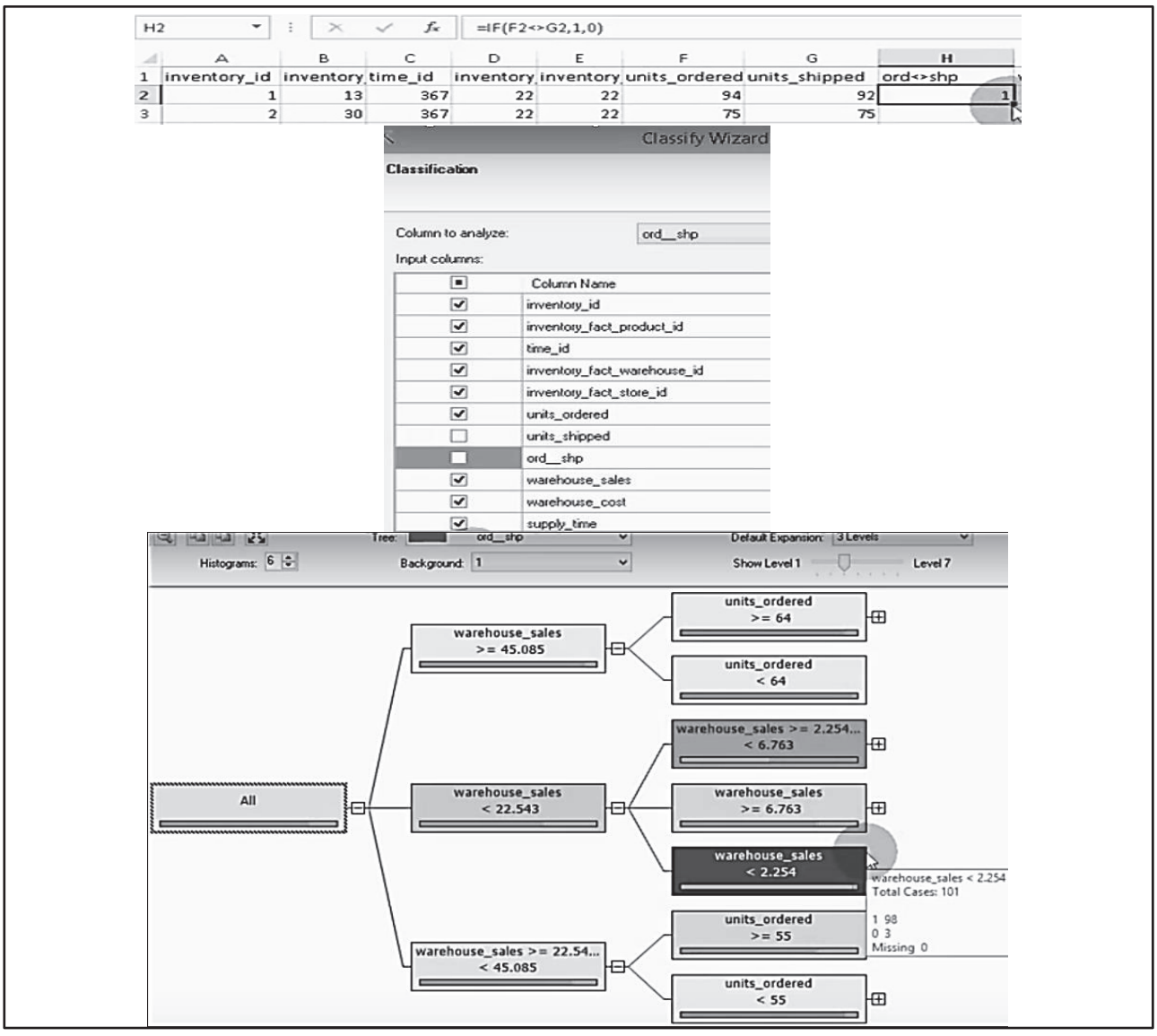
Figura nr. 5. Acumularea de date tranzacționale şi descriptive folosind două interogări SQL MS Access în cascadă

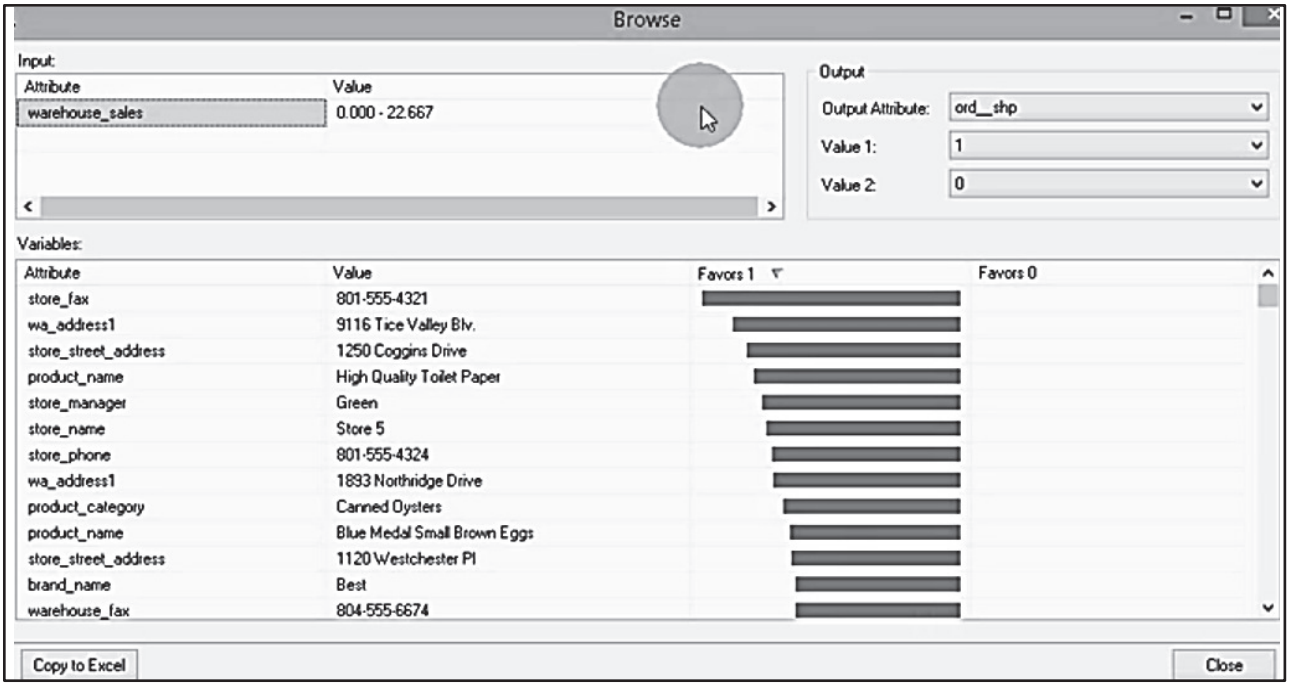


Sursa: Tutorialul video creat de autori: y2u.be/kTuYLuav3Eo

Figura nr. 5 prezintă un exemplu de acumulare de date de inventar (aplicații inclusiv în auditul transportului de marfă) în doi paşi majori, ce corespund celor două interogări SQL în Microsoft Access. Prima variantă are la bază cumularea (clauza UNION) înregistrărilor din două tabele cu tranzacții de acelaşi tip ce corespund doar celor doi ani (1997 şi 1998), respectiv adăugarea unei coloane id necesare (*inventory_id* cu valori generate automat – de tip *AutoNumber*) în tabelul persistent rezultat (clauza INTO). Cea de a doua variantă are la bază extragerea temporară a valorilor câmpurilor descriptive din toate tabelele aflate în legătură sau potrivite pentru o legătură (**Figura nr. 5** – clauza INNER JOIN) cu cel care rezultă din prima interogare de mai sus, şi anume *inventory_fact*. În acest caz datele tabelare rezultate, constând în al doilea set de doar 11.352 înregistrări, nu vor intra într-un tabel persistent al bazei de date (un fel de denormalizare - searchoracle.techtarget.com) ce este necesar în alte situații pentru economisirea de timp în detrimentul spațiului de stocare şi va servi pentru export extern (Excel) imediat după executarea/rularea interogării în sine.

Figura nr. 6. Rezultatele folosirii consecutive a 2 modele Data Mining - câmp țintă derivat cu 2 valori posibile





Sursa: Tutorialele video create de autori: y2u.be/4nOMMRoC2BU şi y2u.be/wce_aoTTsbw

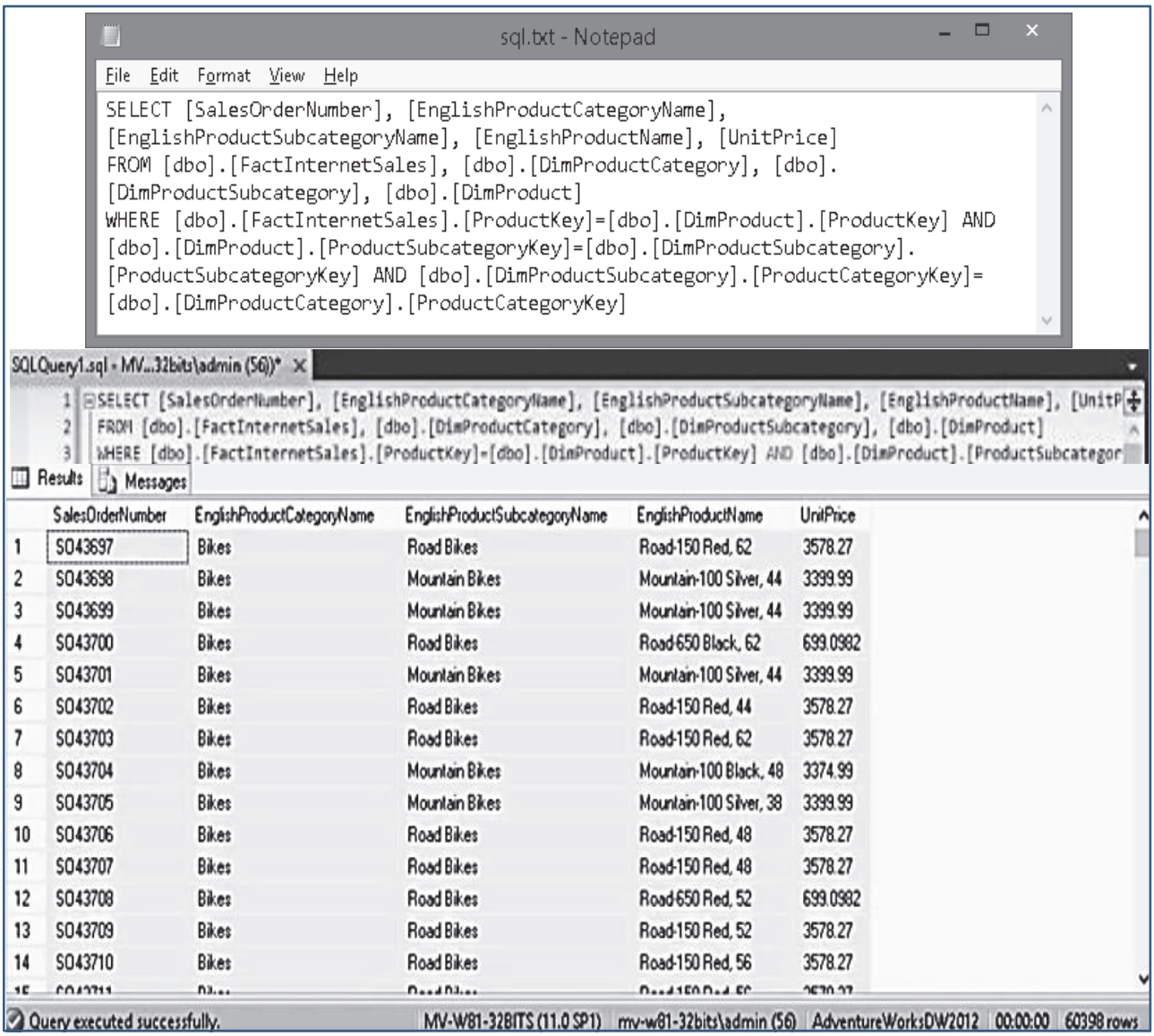
Mai mult, din motive legate de viteza de proiectare am ales toate câmpurile sursă fără selectarea lor explicită, dar indicându-le prin folosirea celui mai flexibil caracter de căutare şi anume asteriscul / "*" după denumirea tabelului (**Figura nr. 5**), atât în modul SQL cât şi în cel de proiectare asistată/*design* (techrepublic.com). Din aceleaşi motive prezentate mai sus, noua coloană derivată necesară pentru analize (**Figura nr. 6** - atribut de ieşire pentru ambele modele: clasificare-sus şi regresie logistică-jos) a fost definită ulterior direct în Excel folosind funcția IF (partea superioară a fig.6, ord<>shp ca 1 sau 0, adică unitățile comandate/*units_ordered* şi unitățile expediate/*units_shipped* sunt diferite, respectiv egale).

4. Reguli de asociere pentru identificarea de modele comportamentale

În teoria şi practica depozitelor de date şi a modelării multidimensionale exemplele de mai jos amintesc de schema „fulg de nea", însemnând că sursa unei dimensiuni (perspectivă de analiză bazată pe coloane descriptive, organizate în ierarhii) nu este reprezentată doar de un singur tabel, ci de mai multe tabele aflate în legătură (relații de timpul „unu la mai multe": de exemplu categoria de produs/*product category*, subcategoria/*product subcategory* şi produsul/*product* – **Figura nr. 7**) şi capabile să asigure suportul unei analize pentru mai mult decât un singur câmp descriptiv pe dimensiune. Pentru a putea aplica algoritmul regulilor de asociere, în cazul de mai jos am avut nevoie şi de valori repetitive pentru câmpul Numărul comenzii de vânzare/*SalesOrderNumber* de asociat la diferite categorii, subcategorii sau nume de produse.

Motivul principal pentru acumularea acestor date descriptive din multiplele tabele ale bazei de date din exemplul de mai sus (**Figura nr. 7**) este de a determina reguli de asociere în comportamentul de cumpărare de tipul „Dacă achiziționez produsul X, voi cumpăra şi produsul Y." (tabelul sursă *FactInternetSales*) precum şi cele mai importante dependențe (**Figura nr. 8**).

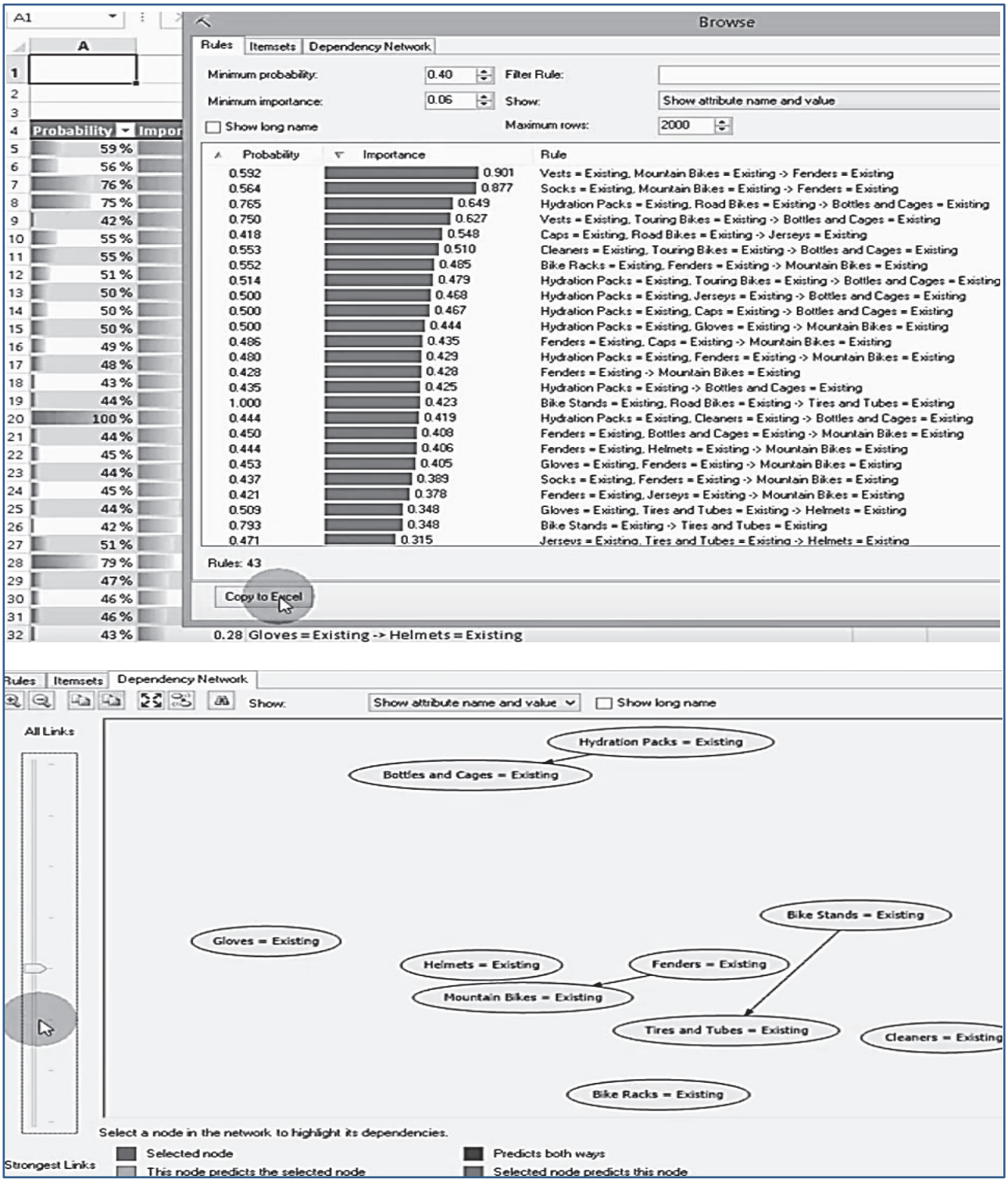
Figura nr. 7. Acumularea ambelor tipuri de date: tranzacționale despre vânzări şi descriptive despre produse folosind o singură interogare MS SQL Server



Sursa: Tutorialul video creat de autori: y2u.be/2rW2wK77HD8

Rezultatele din **Figurile nr. 7** şi **8** ne pot face să înțelegem de ce aplicațiile algoritmilor de identificare de reguli de asociere pot contribui în audit şi la detectarea şi prevenirea fraudelor. Cu titlu de exemplu, dacă setul de date de intrare ar avea atribute precum: Identificator daună, Tip produs de asigurare, Nume asigurat, Asigurator, Nume agent constatator şi Soluționare (daună totală sau parțială şi respingere), iar algoritmul ar identifica „DACĂ asigurare Casco, Asigurat X şi Agent constatator Y ATUNCI daună totală" ca asociere cu probabilitate şi importanță mari, aceasta nu ar însemna neapărat o alarmă de fraudă, dar ar merita măcar efortul de a face investigații suplimentare.

Figura nr. 8. Rezultatele aplicării algoritmilor cu reguli de asociere Microsoft (opțiunea asociază/associate)

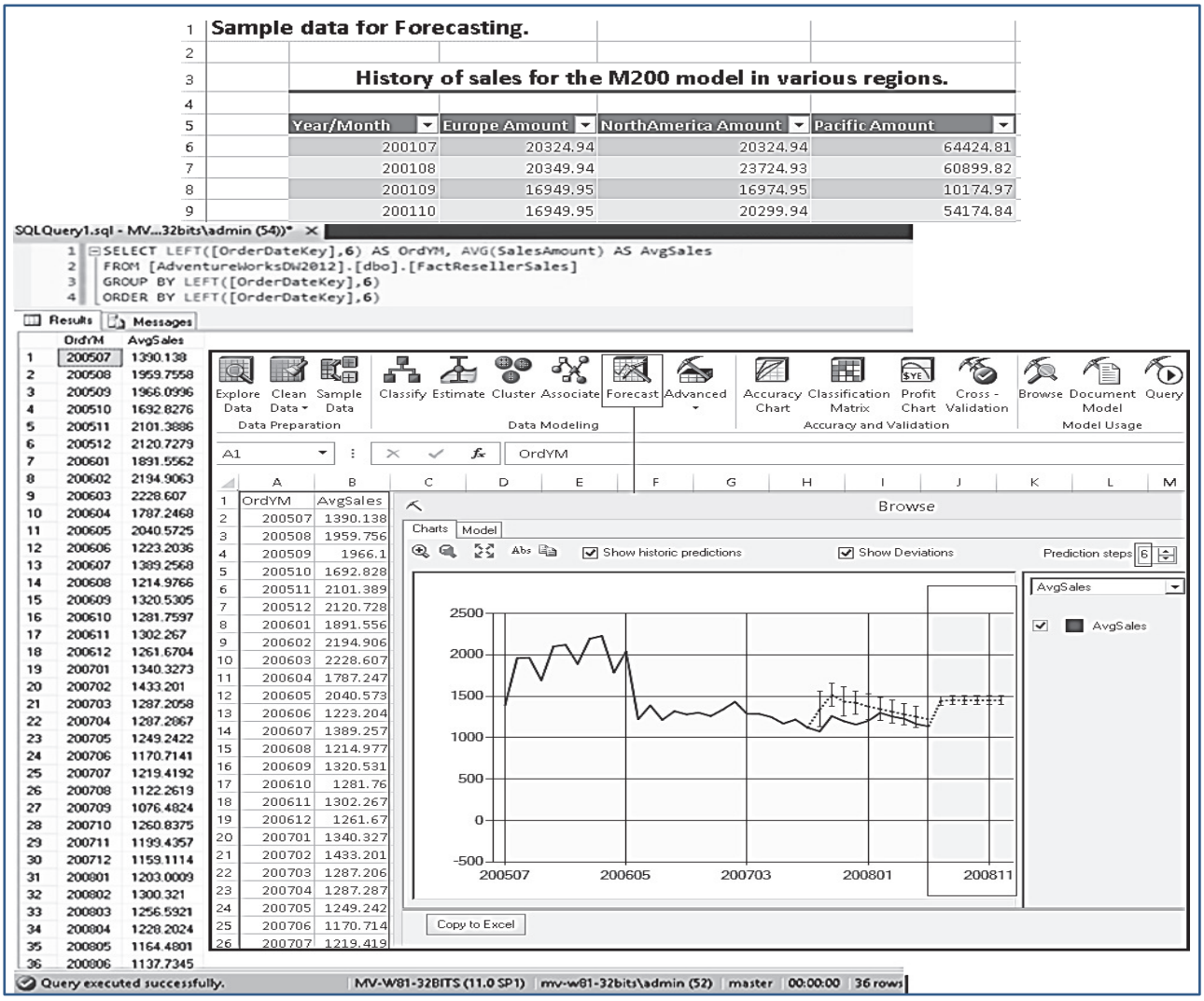


Sursa: Tutorialul video creat de autori: y2u.be/3_8E01hnSD0

5. Previziuni plecând de la date istorice agregate

Pentru mai multe date istorice decât cele din exemplul precedent (**Figura nr. 5**) am luat în considerare crearea unui scenariu special de previziune mai apropiat de realitate. Am pornit de la zero cu un nou exemplu care solicită date pe 36 de luni, din patru ani calendaristici, de data aceasta folosind o interogare simplă SQL pentru un singur tabel, dar cu clauze ORDER BY şi GROUP BY pentru a obține rezultate sortate şi valori agregate, cum ar fi: sume, valori medii, număr total de apariții, număr de apariții pentru o condiție specificată etc. În cazul nostru acestea au fost medii lunare pe mai mulți ani, combinate într-un singur câmp numeric derivat, prin trecerea de la stânga la dreapta în ordinea specifică: de la ani la luni, corespunzând celei de la unități mai mari la unități mai mici (**Figura nr. 9** – exact ca în mostra de date a Microsoft care este furnizată la instalarea componentei *Data Mining*).

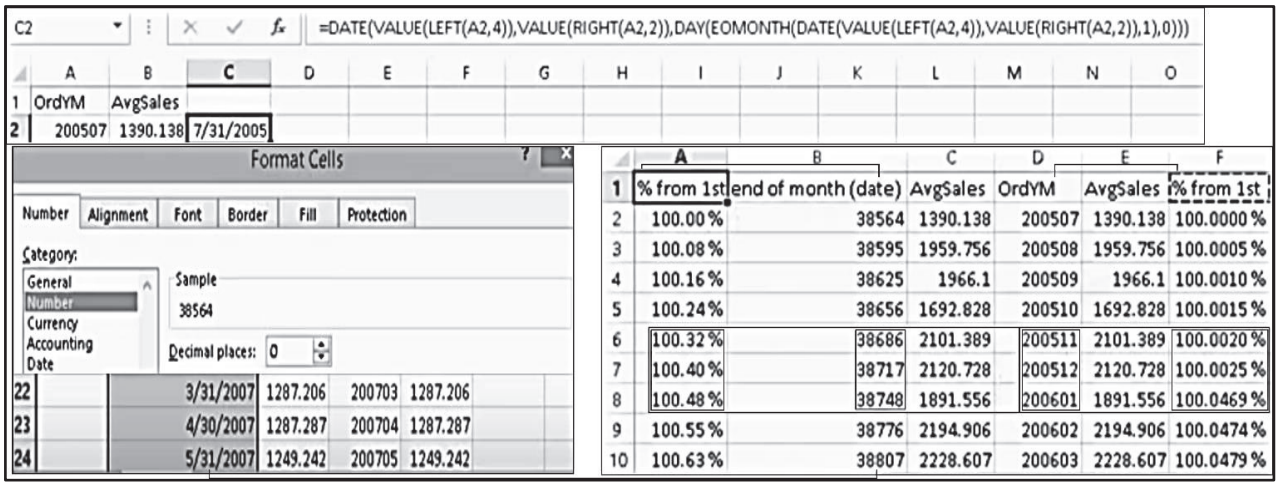
Figura nr. 9. Agregarea de date istorice (stilul ştampilei temporale din mostra Microsoft) cu o interogare SQL Server (clauza GROUP BY) şi o copiere simplă a rezultatelor ce vor fi folosite pentru previziune (componenta Excel* Data Mining*)



Sursa: Tutorialele video create de autori: y2u.be/RjTwGROD0TI şi y2u.be/qHJ3Zm3JBT4

După paşii descrişi mai sus (**Figura nr. 9**) şi alte câteva operații de prelucrare (**Figura nr. 10**) vom ajunge la un set de date potrivit pentru previziuni implementate folosind algoritmii de serii de timp Microsoft ca o combinație a algoritmilor ARIMA (medie mobilă auto-regresivă şi integrată - optimizată pentru creşterea acurateței în predicțiile pe termen lung) şi ARTXP (arbori de auto-regresie cu predicție încrucişată optimizați pentru estimarea următoarei valori probabile într-o serie de timp - msdn.microsoft.com/.../bb677216.aspx).

Figura nr. 10. Derivarea şi explicarea ştampilelor temporare corecte ca date complete stocate intern (Excel) drept numere în surse de date în format potrivit ca suport pentru previziuni nedistorsionate



Sursa: Tutorialul video creat de autori: y2u.be/e0SkDwG9mNY

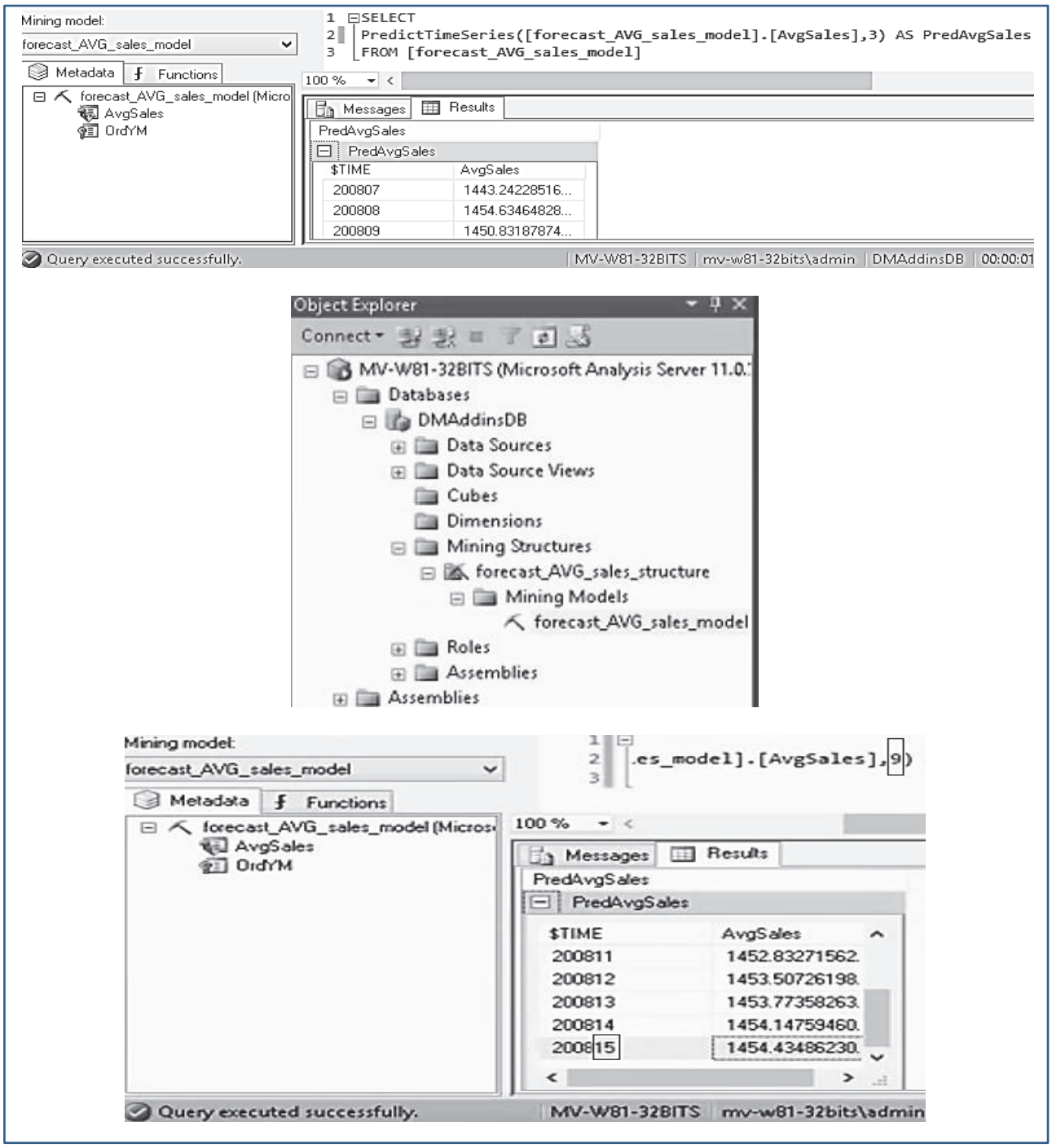
Ne-am gândit, de asemenea, la derivarea automată a etichetelor ştampilelor temporale corecte (formatul LL/ZZ/AAAA tradus într-un număr întreg – **Figura nr. 10**) şi am prezentat mai multe detalii despre comportamentul lor comparativ la obținerea de funcții de trend şi previziunea rezultatelor cu această componentă Excel *Data Mining* (ultimele trei tutoriale din lista menționată anterior).

6. Suport pentru interogarea de modele *Data Mining* persistente

În primul rând, persistent în acest context, se referă la un model definit astfel încât să fie procesat şi stocat pe server (*SQL Server Analysis Services*/servicii de analiză – modul diferit de motorul de baze de date/*Database Engine*) şi disponibil pentru interogare (**Figura nr. 11**).

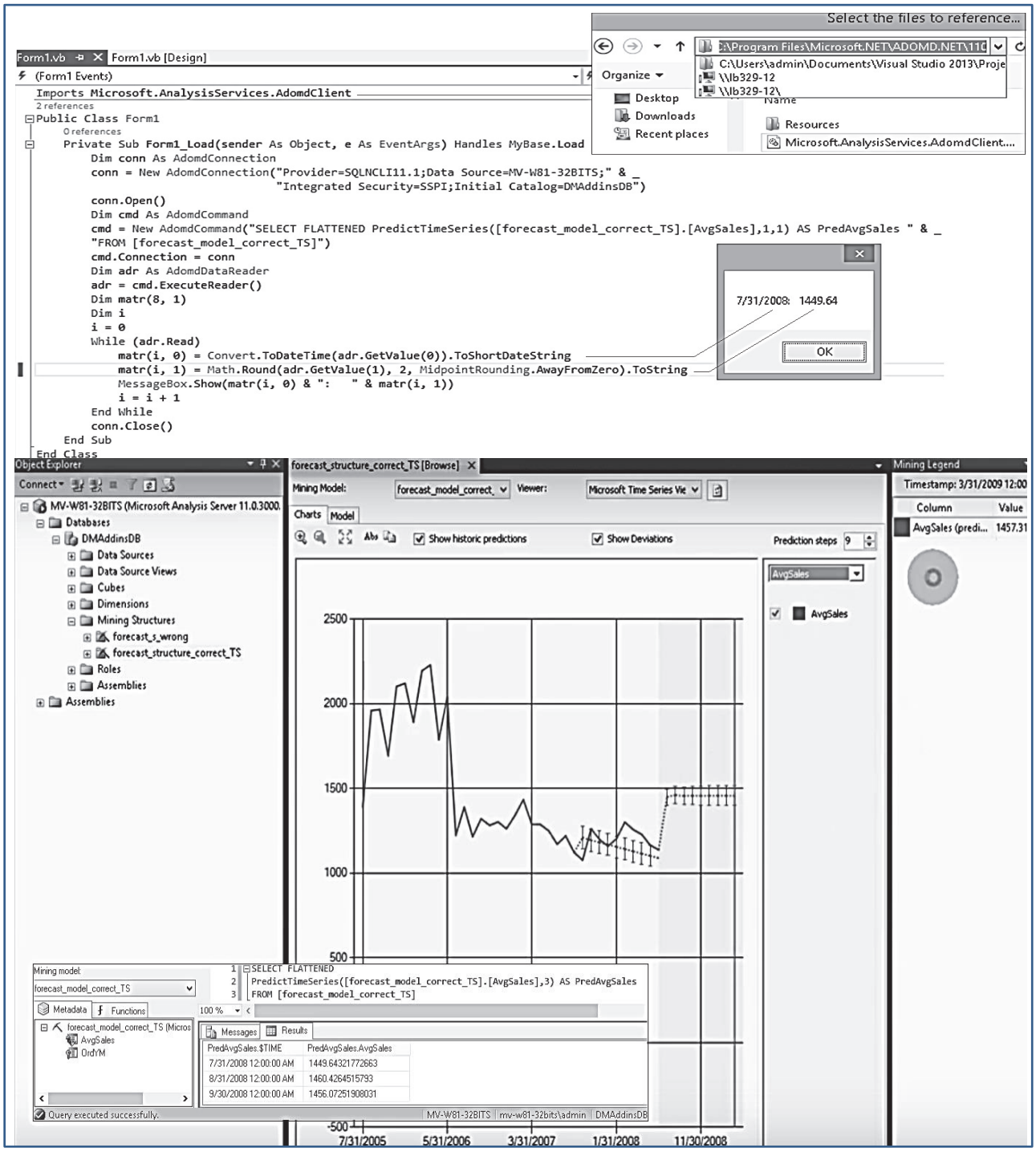
Componenta *Data Mining* din Excel oferă multe avantaje față de utilizarea directă a *SQL Server Analysis Service*. Printre altele, putem menționa aici: viteza de utilizare a mediului tabelar şi a setului de formule din Excel, posibilitatea multor exporturi/importuri în/din foi de calcul tabelar plecând de la diferite formate de baze de date şi de a implica indirect multiple tabele sursă folosind limbajul structurat de interogare (SQL), posibilitățile de exploatare a structurilor şi modelelor obținute: direct (opțiunea *copy to Excel*), cu interogări (extensia DMX a SQL – **Figurile nr. 12** şi **13**) sau programatic?? (**Figura nr. 12**). Ultimele două sunt condiționate de activarea persistenței la definirea modelelor (opțiunea de folosire a unui model temporar/*use temporary model* nebifată – **Figura nr. 13** comparativ cu **Figurile nr. 1** şi **2**).

Figura nr. 11. Exemple de interogări Data Mining eXtensions (DMX) pentru previziunea vânzărilor (SQL Server Analysis Services) pe baza unui model DM de serii de timp plecând de la o sursă de date în format greşit (ştampilă temporală text: 200815 / a 15-a lună în 2008)



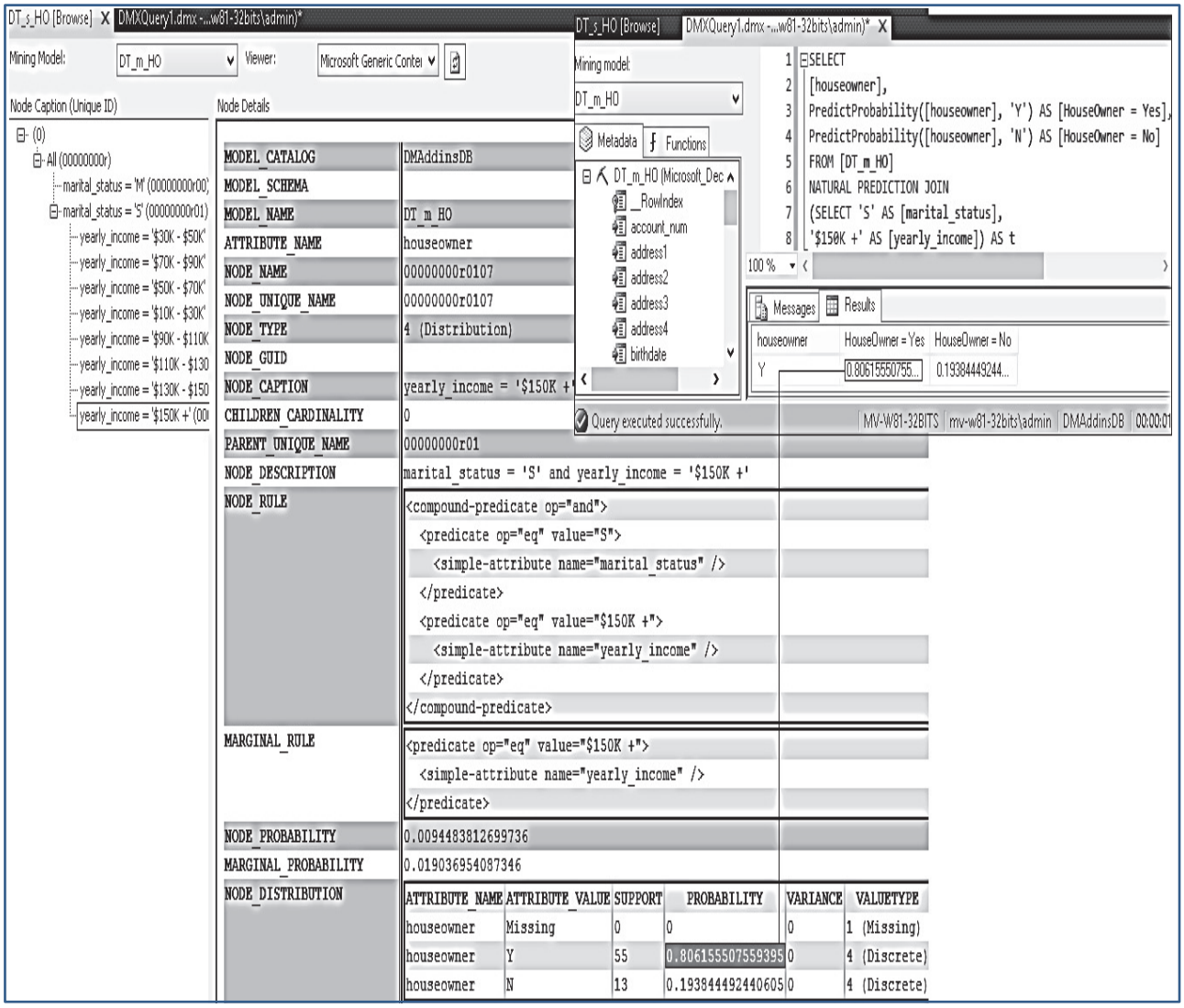
Sursa: Tutorialul video creat de autori: http://y2u.be/qHJ3Zm3JBT4

Figura nr. 12. Exemplu brut de interogare programatică a unui model Data Mining bine definit folosind o interogare DMX în Visual Basic (VB).NET precedată de testarea a mare parte din ea în SQL Server Analysis Services



Sursa: Proiecția autorilor obținută după încercări de dezvoltare cu VB şi SQL Server

Figura nr. 13. Exemplu de interogare de predicție DMX (proprietar casă/houseowner) pe baza unui model de clasificare persistent cu arbori de decizie (vizualizarea generică a conținutului în fundal)



Sursa: Proiecția autorilor obținută după încercări de dezvoltare cu SQL Server

Acest ultim avantaj ne aminteşte că generarea programatică (Airinei şi Homocianu, 2009) de tablouri de bord şi tabele de scoruri Excel folosind reprezentări sugestive, indicatori de alertă şi formatări dinamice cu suport pentru BI s-a simplificat mult începând cu versiunea 2007 a pachetului Microsoft Office. Combinarea acesteia cu abilitatea de a determina programatic modele comportamentale şi de a genera valori previzionate plecând de la instrumente de înaltă performanță şi uşor de folosit precum această componentă *Data Mining,* disponibilă pentru Office 2010, 2013 şi 2016, promite mult în termeni de productivitate. Toate aceste progrese au fost definite după mai mulți ani de utilizare de tehnologii dedicate şi acum bine-cunoscute (de exemplu, SQL Server testat de autori de la începutul anilor 2000).

Pentru produsele de tip foi de calcul tabelar (dssresources.com/.../sshistory.html) precum: VisiCalc, Lotus 1-2-3, Microsoft Excel, Microsoft Works Spreadsheet, Sun Open Office Spreadsheets, Polaris

Office Sheet şi Google Sheets experiența medie a utilizatorilor finali este de până la zeci de ani. Mai mult, uşurința de utilizare a acestor aplicații chiar şi doar ca instrumente de interfață pentru conectarea la datele din baze de date şi depozite de date şi afişarea lor a fost un motiv obiectiv pentru a continua cu testarea componentei *Data Mining* care a condus la conceperea acestui articol.

Folosind o modalitate de raportare care se identifică cu o secvență de paşi care împrumută numele de la cele optsprezece tutoriale suport şi, de asemenea, unele tehnici anterior definite şi anume: E2P4CAFR (Homocianu, 2015), ACCORD/CADRE (Homocianu şi Airinei, 2014) şi S-DOT (Homocianu şi Airinei, 2014) se poate ajunge în etape, dar cu un număr minim de paşi de urmat, la anumite reprezentări care sunt dinamice, interactive, sugestive, bazate pe cauzalitate şi înrădăcinate în realitatea curentă şi în istoria definită de datele stocate în sursele de date ale organizației.

Concluzii

Putem concluziona că posibilitățile componentei Excel *Data Mining* sunt peste aşteptările unui analist de afaceri, oferind avantajul integrării tiparelor de clasificare identificate, a regulilor de asociere şi a predicțiilor cu suportul pentru conectivitate la formate variate de date. Validările de date, reprezentările grafice avansate, referențieri geografice, formatările condiționate automate şi indicatorii cheie de performanță (KPI), tabelele şi graficele de tip pivot şi *power pivot*, rezolvarea automată de probleme de optimizare (*solver*) şi limbajul DAX (expresii de analiză a datelor/*Data Analysis eXpressions*) împreună cu limbajul tradițional de formule sporesc şansele definirii de tablouri de bord fundamentate pe simulări, analize şi modele *Data Mining* cu adevărat utile pentru personalul de audit interesat de monitorizarea performanței.

Sperăm că am identificat multe motivații reale pentru alegerea acestei componente Microsoft pentru pachetul Office ca un instrument *Data Mining* aproape în timp real, dincolo de multe alte recomandări disponibile în literatura şi practica de specialitate.

Dincolo de exemple substanțiale de lucru cu aplicații software bine cunoscute, disponibile pentru o gamă largă de utilizatori, care asigură metode avansate de analiză, interogare şi reprezentare a datelor curente şi istorice specifice instrumentelor suport pentru *Data Mining* şi *Business Intelligence*, lucrarea furnizează şi o scurtă descriere teoretică necesară înțelegerii unei modalități rapide de generare de rapoarte complexe şi dinamice precum tablourile de bord fundamentate pe analize şi modele *Data Mining* plecând în special de la date despre vânzări şi date financiare.

Tutorialele video dezvoltate de autori, integrate într-o listă, la care s-a făcut referire în mod succesiv în această lucrare, demonstrează încercările de îmbogățire a modului de raportare anterior menționat şi de asigurare a minimizării numărului de paşi necesari atunci când se încearcă implementarea de exemple similare.

În ansamblu, articolul încearcă să transmită prin exemple clare anumite caracteristici dorite cum ar fi: viteză, simplitate, capacitate de sinteză, transparență, flexibilitate şi disponibilitate în raportarea fundamentată pe exploatarea datelor, ca elemente cheie de performanță în pregătirea situațiilor financiare şi sprijinirea activităților de audit.

BIBLIOGRAFIE

1. Airinei, D. (2002), *Depozite de date*, Editura Polirom, Iaşi.
2. Airinei, D. şi Homocianu, D. (2009), The Geographical Dimension of DSS Applications, *Analele Ştiintifice ale Universității „Alexandru Ioan Cuza" din Iaşi*, Tome LVI, pp. 637-642.
3. Chersan, I.C., Carp, M. şi Mironiuc, M. (2013), Data mining – o provocare pentru auditorii financiari, *Audit Financiar*, vol. XI, nr. 10, pp. 57-64.
4. Cleland, D.I. şi King, W.R. (1975), Competitive Business Intelligence Systems, *Business Horizons Journal*, vol. 18, nr. 6, pp. 19-28, DOI 10.1016/0007-6813(75)90036-1.
5. Fraser, L.E. (1998), Public Sector Audit - Business Integration and Causal Analysis, *Quality Audit Conference*, February 26-27, 1998, Louisville, KY, vol. 7.
6. Homocianu, D. (2015), Excel Power Pivot's Applications in Audit and Financial Reports, *Audit Financiar*, vol. XIII, nr. 11, pp. 127-138.
7. Homocianu, D. şi Airinei, D. (2014), Business Intelligence facilities with applications in audit and

financial reporting, *Audit Financiar*, vol. XII, nr. 9, pp. 17-29.

8. Homocianu, D. şi Airinei, D. (2014), Consolidating source data in audit reports, *Audit Financiar*, vol. XII, nr. 8, pp. 10-19.
9. Inmon, W.H. şi Linstedt, D. (2014), *Data architecture: A primer for the data scientist. Big Data, Data Warehouse and Data Vault*, Morgan Kaufmann, MA.
10. Rasmussen, N.H., Goldy, P.S., Solli, P.O. (2002), *Financial business intelligence – trends, technology, software selection and implementation*, John Wiley and Sons, Inc., New York, pp. 98-99.
11. Sirikulvadhana, S. (2002), *Data mining as a financial auditing tool (thesis), Swedish School of Economics and Business Administration*, pp. 49-57, disponibil online la adresa: https://pdfs.semanticscholar.org /2612/f764664796f911e9ff9a79b7bb9de84bf16c.pdf, accesat pe data de 15.03.2017.
12. Vintilescu Belciug, A., Crețu, D. şi Gegea, C. (2010), Utilizarea tehnicilor de data mining ca metodă complementară în audit, *Audit Financiar*, vol. VIII, nr. 7, pp. 30-35.
13. Wang, J. şi Yang, J.G.S. (2009), Data Mining Techniques for Auditing Attest Function and Fraud Detection, *Journal of Forensic & Investigative Accounting*, vol. 1, no. 1, pp. 1-24.
14. http://bi-insider.com/business-intelligence/operational-bi-vs-strategic-bi/
15. http://dssresources.com/faq/index.php?action=artikel&id=174
16. http://dssresources.com/faq/index.php?action=artikel&id=199
17. http://dssresources.com/history/dsshistory.html
18. http://dssresources.com/history/sshistory.html
19. http://inf.ucv.ro/documents/rstoean/5.%20Arbori%20de%20decizie.pdf
20. http://profs.info.uaic.ro/~val/statistica/StatWork_10.pdf
21. http://searchoracle.techtarget.com/tip/Optimizing-database-performance-part-2-Denormalization-and-clustering
22. http://searchsqlserver.techtarget.com/definition/data-mining
23. http://www.computerweekly.com/tip/Inmon-vs-Kimball-Which-approach-is-suitable-for-your-data-warehouse
24. http://www.techrepublic.com/article/10-tips-for-using-wildcard-characters-in-microsoft-access-criteria-expressions/
25. http://www.w3schools.in/dbms/database-normalization/
26. http://y2u.be/Xs2SWtBqdzI
27. https://deshpande.mit.edu/portfolio/project/hybrid-dbms-optimized-read-intensive-applications
28. https://developers.google.com/apps-script/guides/sheets
29. https://msdn.microsoft.com/en-us/library/bb677216.aspx
30. https://msdn.microsoft.com/en-us/library/dn282385.aspx
31. https://msdn.microsoft.com/en-us/library/ms174828.aspx
32. https://msdn.microsoft.com/en-us/library/office/ee814737.aspx
33. https://onlinecourses.science.psu.edu/stat504/node/149
34. https://sites.google.com/site/supp4excel2dataminin g2017af/d