Data mining on the real estate market

Tit-Liviu LEONTIN, Darie MOLDOVAN, Manuela RUSU,
Daniela SECARĂ, Corina TRIFU
Facultatea de Ştiinţe Economice şi Gestiunea Afacerilor
Universitatea „Babeş-Bolyai", Cluj-Napoca

This paper aims to give an update on the evolution of the data analyze techniques, linking them with the artificial intelligence algorithms, especially with data mining. We shall highlight a couple of fields where data mining is very useful and some software application, with a special emphasizes on the free WEKA. As example, we shall present a case study of the real estate market in Cluj-Napoca, with national and European perspectives.
Keywords: *data mining, real estate market, WEKA, decision, assessment.*

Date vechi, cunoştinţe noi

Încă de la apariţia primelor calculatoare în anii 1950, oamenii au înţeles puterea acestora de a colecţiona, manipula, clasifica, stoca şi regăsi date într-un mod mult mai rapid, mai flexibil şi mai eficient decât puterea umană. Volumul şi complexitatea informaţiilor depăşeşte capacitatea intelectuală a unei persoane; acesta este motivul pentru care sistemele informaţionale permit stocarea datelor pentru o analiză ulterioară şi oferă metodele de procesare şi filtrare a acestora pentru a extrage doar datele relevante. Mai mult, cu ajutorul algoritmilor matematici şi statistici ce implementează metode cunoscute pentru înţelegerea datelor se obţin informaţii cu înţelesuri noi.

Dar această metodă se bazează pe ingeniozitatea umană şi capacitatea de a găsi metode de interpretare a datelor, în timp ce calculatoarele rămân maşini rapide dar simple de procesare a datelor prin algoritmii implementaţi. Pe măsură ce complexitatea aplicaţiilor, numărul variabilelor ce trebuie luate în considerare şi volumul datelor primare au crescut exponenţial de la un an la altul, unii cercetători au început să-şi pună întrebări. *„Nu cumva există şi alte informaţii ce pot fi extrase din datele existente, însă care nu sunt evidente, nu le înţelegem şi nu ştim cum să le interpretăm? Dar dacă există înţelesuri ascunse, sau relaţii complexe cauză-efect între date, pe care nu le putem evidenţia cu o simplă formulă matematică? Ce se întâmplă dacă schimbările au loc mai repede decât puterea noastră de a găsi şi a verifica algoritmi noi?"*

Depăşiţi de volumul şi complexitatea datelor, cercetătorii au ales să doteze calculatoarele cu inteligenţă pentru a găsi şi extrage noi informaţii şi cunoştinţe din datele existente. Astfel, în anii 1990, psihologi, ingineri şi matematicieni şi-au înzecit eforturile de a înţelege cum funcţionează creierul uman pentru a învăţa lucruri noi. Odată definit procesul de învăţare, şi calculatoarele puteau fi dotate cu rutinele software pentru descoperirea şabloanelor repetitive în date, pentru a găsi relaţii şi interdependenţe, pentru a învăţa lucruri noi fără asistenţă umană, pentru a extrage cunoştinţe noi şi valoroase fără ca un programator sau operator să le spună cum. Astfel a fost realizată legătura cu disciplinele de inteligenţă artificială.

Majoritatea oamenilor înţeleg greşit conceptul de inteligenţă artificială; este un subiect fierbinte în anumite cercuri ale filosofiei, ingineriei şi psihologiei deoarece termenul „inteligent" poate ridica probleme privind conştiinţa proprie, mintea, comportamentul uman, credinţa în Dumnezeu şi altele. De aceea, cercetătorii au divizat domeniul inteligenţei artificiale în două [1]. Pe de o parte, cei ce se ocupă de inteligenţa artificială puternică încearcă să creeze o formă artificială de inteligenţă, capabilă de raţionament, soluţionarea problemelor, înţelegere a propriei existenţe şi gândire mai mult sau mai puţin apropiată de cea umană, însă cercetările în această direcţie au înregistrat succese de mi-

că importanță în simulări a unei inteligențe limitate bazate pe un set predefinit de reguli. Pe de altă parte, cercetătorii ce se ocupă cu inteligența artificială slabă trebuie să creeze o formă mult mai restrânsă de comportament inteligent, însă nu posedă inteligență complexă şi înțelegere a propriei existențe. În această ultimă direcție s-au obținut rezultate remarcabile în domenii de specialitate, cum ar fi limbajele de programare LISP şi Prolog, Deep Blue - supercomputerul IBM ce l-a învins în 1997 într-un joc de şah de proporții istorice pe campionul mondial Garry Kasparov [2], sistemele expert folosite în afaceri şi industrie, programe de traducere automată a documentelor dintr-o limbă în alta, recunoaşterea scrisului de mână, a vorbirii şi înțelegerea imaginilor.

De la inteligență artificială la *data mining*

Machine learning este un domeniu al inteligenței artificiale ce implică dezvoltarea unor metode de a crea aplicații software pentru analiza seturilor de date pentru a permite calculatoarelor să „învețe" din experiența anterioară [3]. *Machine learning* foloseşte intensiv statistica matematică, din moment ce ambele studiază interpretarea datelor. În funcție de tipul rezultatelor, algoritmii de *machine learning* sunt organizați în mai multe categorii, cum ar fi *supervised learning* care generează funcții ce pun în corespondență datele de intrare şi ieşire, sau *unsupervised learning* în care algoritmii generează modele bazându-se pe datele existente.

Recunoaşterea formelor sau clasificarea duce cu un pas înainte statistica matematică şi *machine learning* prin efectuarea unei acțiuni în funcție de tipul datelor, folosind metode cum ar fi rețelele neuronale şi rețelele bayesiene şi având aplicații tipice în clasificarea textului şi recunoaşterea vorbirii şi a imaginilor.

Aflându-se sub aceeaşi „umbrelă" cu inteligența artificială pentru numeroasele înțelesuri într-o gamă largă de contexte, *data mining* este cunoscută şi sub numele de descoperirea cunoştințelor în bazele de date şi foloseşte tehnici computaționale bazate pe inteligență artificială, *machine learning*, recunoaşterea formelor şi statistică matematică. *Data mining* este un proces analitic ce explorează un număr foarte mare de date în căutarea unor şabloane sau relații între variabile [4], apoi generalizează aceste rezultate într-un model, formulă sau arbore de decizie şi, în cele din urmă, verificarea corectitudinii modelului generat prin testarea lui pe setul de date existent sau al unuia nou. De cele mai multe ori, cu predilecție în aplicațiile economice, scopul acestui algoritm este de a găsi metode de previziune a evenimentelor şi rezultatelor viitoare pe baza datelor existente, de a ajuta managerii să ia decizii repede şi corect. În alte situații, *data mining* poate fi folosită pentru a verifica dacă există o corelație ascunsă între datele de intrare şi rezultate, sau dacă o asemenea corespondență este improbabilă.

Domenii de utilizare şi aplicații software pentru *data mining*

Data mining este o unealtă foarte puternică, folosită în cele mai variate domenii. Iată câteva exemple:

- **Piața țintă**. Obiectivul este acela de a folosi data mining pe baza rezultatelor unei campanii de direct-mail din trecut pentru a se identifica respondenții cei mai promițători, combinând date demografice şi geografice culese la acea dată. Avantajele trebuie să fie o mai bună rată a respondenților şi costuri mai scăzute ale noii campanii. Grupul financiar Fleet a investit 38 de milioane de dolari în depozite de date pentru a-şi înnoi infrastructura privind clienții [5]. *Data mining* a fost folosit pentru a prevedea probabilitatea de răspuns în ceea ce priveşte interesul pentru un credit ipotecar al 20.000 de clienți dintr-o bază de date de 15 milioane, a-i găsi pe cei profitabili şi a-i depista pe cei neprofitabili chiar dacă ar răspunde favorabil.
- **Depistarea clienților ce au tendința de a renunța la serviciile firmei**. Obiectivul este de a preveni atât pierderea clienților existenți cât şi atragerea de clienți predispuşi la a renunța uşor la serviciile firmei. Soluția *data mining* este identificarea pe baza unor caracteristici comune a clienților pe cale să renunțe, folosind rețele neuronale şi analize de serii cronologice. France Telecom a implementat în acest sens un Customer Profiling

System (CPS), având ca rezultat detectarea rapidă a clienților cu tendința de renunțare prin compararea anumitor caracteristici ale acestora cu caracteristici ale celor care renunțaseră deja [5].

▪ **Detectarea fraudelor**. Fraudele măresc costurile sau reduc veniturile. Folosind regresii şi rețele neuronale pentru a determina caracteristicile esențiale ale cazurilor de fraudă pentru a le preveni pe viitor, se pot obține importante creşteri ale profitului, evitând clienții indezirabili. Automobile Insurance Bureau of Massachusetts, o companie de asigurări auto, a folosit rapoarte vechi ale experților în legătură cu fraudele depistate [5]. Mai multe caracteristici (peste 60), cum sunt tipul accidentului, tipul tratamentului aplicat, gravitate, au fost codificate într-o bază de date. Folosind metode specifice de *data mining* au fost găsite caracteristici comune cazurilor de fraudă.

▪ **Analiza riscului**. Riscul de a acorda credite unor clienți cu potențial de a deveni răuplatnici poate fi diminuat considerabil construind funcții de separare a clienților pe baza anumitor caracteristici [5].

▪ **Sisteme de recomandare**. Vizitatorii şi clienții magazinelor virtuale pe Internet, de exemplu amazon.com, evaluează produsele prezentate. Informațiile astfel obținute sunt folosite pentru a recomanda produsele către alți vizitatori. Folosind tehnica numită „filtrare colaborativă" [5], companiile şi-au crescut veniturile prin *cross-selling* şi *up-selling*.

Un studiu online efectuat de site-ul de specialitate KDnuggets.com în luna august 2004 privind domeniile de utilizare a *data mining* [6] a relevat următoarele: industria bancară este lider cu 13%, urmată de marketing direct, detecția fraudelor şi cercetarea ştiințifică cu câte 9%; alte domenii sunt biotehnologie cu 8%, asigurări şi medicină cu 7% fiecare, comerț electronic şi telecomunicații cu 6% fiecare, investiții financiare, manufactură, vânzare cu amănuntul şi securitate cu 4% fiecare. Un alt studiu al aceluiaşi site, efectuat în luna septembrie 2004, promovează algoritmii ce produc arbori de decizie pe primul loc între categoriile de algoritmi sugerate.

Câteva pachete software comerciale folosite în *data mining* sunt: Oracle9i Data Mining pentru Oracle9i Database Enterprise Edition, Oracle Data Mining Suite [7], Teradata Warehouse Mining [8], SAS Enterprise Miner [9], Crystal Analysis and Decisions [10], Clementine [11], SmartDiscovery [12], Monarch [13], Statistica [14], InfoBase [15], PolyAnalyst [16].

Câteva aplicații software gratuite, destinate mediului de cercetare şi educație, sunt: IBM Intelligent Miner [17], WEKA [18], ADaM [19], YALE [20]. Alte produse software pentru *data mining*, grupate pe categorii, pot fi găsite online pe site-ul www.KDnuggets.com

Mediul de programare WEKA

WEKA (the Waikato Environment for Knowledge Analysis) este un soft gratuit pus la dispoziție de catedra de specialitate a Universității Waikato din Hamilton, Noua Zeelandă [18]. Mediul de programare WEKA permite aplicarea tehnicilor de învățare automată asupra problemelor practice şi integrează diverse unelte pentru învățarea automată ce pot fi utilizate într-un mediu de lucru uzual, caracterizat de o interfață omogenă. Utilizatorii pot folosi gama largă de tehnici de învățare automată pentru extragerea unor informații utile din baze de date foarte mari. Trebuie precizat faptul că WEKA poate fi utilizat în orice domeniu de interes, având astfel un avantaj major asupra celorlalte aplicații de *data mining*, mai ales asupra celor comerciale care sunt destinate unui singur domeniu de activitate.

WEKA conține *unelte* pentru preprocesarea datelor, iar pentru clasificarea acestora se utilizează arbori de decizie, regresie, clusterizare, reguli de asociere si vizualizare. Aplicația este dezvoltată în Java, iar codul sursă este deschis, eliberat sub licență GNU General Public License. Acesta este un mare avantaj al sistemului WEKA spre deosebire de alte aplicații, deoarece permite modificarea sistemului de către utilizatori în modul în care aceştia au nevoie de el, eventual cu dezvoltarea de noi tehnici de învățare automată şi implementarea de algoritmi proprii. De asemenea, la fel de important e faptul că sistemul poate fi utilizat pe mai multe platforme: Unix, Linux şi Microsoft Windows.

Ultima *versiune* pusă la dispoziția utilizatorilor este WEKA 3.4.3 şi poate fi instalată atât pe platforma Windows cât şi pe alte platforme: Linux, Unix, etc. Trebuie menționat că pentru MacOS X nu este disponibilă deocamdată decat versiunea WEKA 3.4.2. Pentru a rula WEKA trebuie să existe instalată pe sistem maşina virtuală Java 1.4. O versiune anterioară a WEKA este WEKA 3.0, ce se bazează pe lucrul în linie de comandă.

La *lansarea WEKA* apare fereastra „GUI Chooser" care permite utilizatorilor să opteze pentru lucrul în linie de comandă („CLI") sau pentru deschiderea lucrului în interfața grafică („Explorer"). WEKA Explorer pune la dispoziție în interfața grafică pachetele sistemului, şi anume:

- Preprocessing, în cadrul căruia se pot deschide seturile de date atât sub forma fişierelor ARFF cât şi dintr-o bază de date anume; de asemenea, se poate realiza o filtrare nesupravegheată a datelor cu unul din filtrele puse la dispoziție;
- Classify, ce permite alegerea şi rularea oricărui algoritm de clasificare din cele 6 categorii de algoritmi definite;
- Cluster, în cadrul căruia se poate alege şi rula metoda de clusterizare a datelor;
- Associate, ce permite setarea unei reguli de asociere a datelor şi aplicarea acesteia;
- Select Attributes este un alt pachet WEKA şi permite configurarea şi aplicarea oricărei combinații de atribute din cele ce definesc setul de date pentru a depista care sunt cele mai relevante atribute din set;
- Visualize permite vizualizarea setului curent de date în una sau două dimensiuni, iar dacă atributele au valori continue este utilizat un spectru de nuanțe ale aceleiaşi culori pentru reprezentarea valorilor, pe când pentru atribute discrete fiecare valoare este reprezentată cu altă culoare.

Suplimentar acestor pachete de instrumente pentru lucrul cu seturi de date, WEKA conține şi un clasificator pe bază de arbori de decizie WEKA CLASSIFIERS TREES USERCLASSIFIER şi o interfață grafică pentru realizarea de rețele neuronale WEKA CLASSIFIERS FUNCTIONS NEURAL NEURALNETWORK.

Setul de date utilizat în mediul de programare WEKA trebuie să fie în format ARFF pentru a putea fi prelucrat. Datele provin de cele mai multe ori dintr-o tabelă Excel sau dintr-o bază de date şi trebuiesc convertite în formatul ARFF, cel mai larg răspândit pentru baze de date în fişiere text. Folosirea acestui format în paralel cu suportul direct pentru baze de date este un alt avantaj al WEKA.

Pe lângă aceste elemente favorabile ce caracterizează sistemul WEKA, există şi câteva dezavantaje, şi anume faptul că necesită învățarea utilizării interfeței, înțelegerea algoritmilor şi a modului de interpretare a rezultatelor numerice şi grafice. În plus, WEKA foloseşte termeni statistici în loc să folosească termeni corespunzători datelor de intrare (de exemplu, din aplicațiile economice) aşa cum fac alte produse software specializate pe mediul de afaceri şi mult mai intuitive pentru un manager sau economist.

Studiu de caz: *Data mining* în piața imobiliară

Obiective

Conform unui studiu recent condus de revista financiară „Capital" şi prezentat în ediția cu numărul 18 (24 aprilie 2004) [21], în oraşul Cluj-Napoca există aproximativ 200 de agenții imobiliare, reprezentând cel mai mare număr pe cap de locuitor din țară. O explicație posibilă a acestui fapt este explozia industriei locale în ultimii ani, investitorii români cât şi cei străini considerând Clujul a fi una dintre regiunile cu cea mai rapidă dezvoltare economică.

Fiind date cererea mare de apartamente şi prețurile extrem de ridicate existente în momentul de față pe piața imobiliară clujeană, scopul acestui proiect este de a facilita achiziția unui apartament. Acest obiectiv va fi realizat prin analiza datelor existente pe piața de locuințe utilizând tehnici de *data mining*. În cadrul lucrării se vor determina prețul şi punctajul adecvat fiecărui apartament, care pe o scală de la 1 la 5 va arăta dacă acea locuință merită sau nu să fie cumpărată.

Argumentare

Din cauza numărului foarte mare de agenții imobiliare existente pe piața din Cluj Napoca, găsirea unui apartament poate de-

veni o mare problemă. Mai mult chiar, având atât de multe variante poate deveni chiar imposibilă vizitarea fiecărui apartament disponibil şi luarea unei decizii corecte în funcţie de necesităţile personale ce pot sau nu a fi îndeplinite de locaţiile în cauză.

Alte două motive ce au determinat realizarea acestui proiect au fost dinamica foarte mare a pieţei imobiliare şi cererea în continuă creştere pe această piaţă ce duce la creşteri nejustificate ale preţurilor, ajungându-se astfel la necesitatea reevaluării continue a apartamentelor, lucru ce poate fi realizat mult mai uşor utilizând analiza de tip *data mining*.

Unelte şi metode

În realizarea lucrării s-a utilizat mediul de programare WEKA, iar proiectul a fost implementat cu ajutorul interfeţei grafice a Explorer-ului. Dintre algoritmii de clasificare puşi la dispoziţie de sistem în cadrul lucrării s-au utilizat Regresia Liniară şi algoritmul J.48 de clasificare prin arbori de decizie.

Setul de date

Datele sunt fost furnizate de săptămânalul clujean „Piaţa de la A la *Z*" [22] şi conţin anunţurile de vânzare de apartamente din Cluj-Napoca publicate în luna ianuarie 2001. Datele au fost primite în format FoxPro. Pentru a putea analiza această bază de date folosind WEKA în vederea atingerii obiectivelor propuse, formatul iniţial al datelor a trebuit modificat în paşi succesivi.

În primul rând, anunţurile se regăseau în baza de date în câmpuri de tip memo, câte unul pentru fiecare anunţ. Prin urmare, primul pas a fost de a extrage datele relevante (etajul, prezenţa balconului, televiziune cablu, telefon, centrală termică, garaj etc.) din câmpul memo folosind formule de căutare de text în Excel. După aceasta, pe baza cuvintelor cheie găsite, o nouă bază de date Excel a fost generată conţinând toate cele 18 atribute finale pe care le vom folosi ca set de antrenament în WEKA. Câmpurile libere pentru o instanţă au fost schimbate cu un semn al întrebării, simbolizând informaţie lipsă.

A fost necesară şi efectuarea unor alte modificări înainte de începerea testelor, ţinând cont de faptul că algoritmii aleşi pentru analiză stabilesc o relaţie între atributele de intrare şi cel de ieşire (cauză-efect). Din acest motiv, toate instanţele din baza de date trebuiau să conţină cel puţin două atribute nonvide, unul pentru intrare şi altul pentru ieşire, deoarece este imposibil să se stabilească o relaţie cu un singur atribut. Ca rezultat, din baza de date au fost şterse toate instanţele ce aveau un singur atribut. O altă problemă ce trebuia rezolvată era moneda în care era exprimat preţul, deoarece preţurile din anunţuri erau exprimate atât în lei cât şi în trei valute diferite. Am decis să convertim toate preţurile folosind Euro ca monedă de referinţă, făcând conversia cu ratele de schimb oficiale din 7 ianuarie 2001. În cele din urmă, întrucât WEKA foloseşte ca intrare a datelor formatul ARFF, baza de date din Excel a fost convertită în format CSV (valori separate de virgule), după care i s-a adăugat antetul conţinând descrierea atributelor, a tipurilor şi valorilor acestora.

Nr.	Atribut	Tip
1	camere	nominal: 1, 2, 3, 4, 5
2	decomandat	nominal: decomandat, semidecomandat
3	confort	nominal: sporit, I, II
4	etaj	nominal: intermediar, parter/ultimul
5	balcon	boolean
6	finisare	nominal: superfinisat, finisat, semifinisat, nefinisat
7	parchet	boolean
8	faianţă	boolean
9	gresie	boolean

Nr.	Atribut	Tip
10	termopan	boolean
11	modificări	boolean
12	centrală	boolean
13	contorizat	boolean
14	telefon	boolean
15	cablu	boolean
16	garaj	nominal: garaj, parcare
17	cartierul	nominal: Mănăştur, Gheorgheni, Zorilor, Mărăşti, Grigorescu, Centru
18	preţ	numeric

Fig. 1. Modelul de calcul obţinut în urma algoritmului de regresie liniară

Forma finală a bazei de date conţinea 18 atribute ce descriu informaţiile existente despre

apartamente şi conţine 1981 instanţe. Atributele sunt nominale, booleene şi preţul ca atribut numeric. (Figura 1)

În ceea ce priveşte baza de date, mai trebuie făcută o clarificare: pentru a obţine cele mai bune rezultate folosind algoritmul de clasificare J48, au fost efectuate modificări în baza de date prin selecţia a 7 atribute şi adăugarea unui atribut suplimentar reprezentând o evaluare generală a caracteristicilor fiecărui apartament. Această evaluare a fost efectuată folosind o funcţie expert aplicată tuturor celor 18 atribute iniţiale, obţinând astfel un scor personalizat al fiecărui apartament pe baza preferinţelor medii ale cumpărătorilor. Evaluarea este dată de un număr întreg pe o scală de la 1 la 5, 1 fiind valoarea cea mai mică şi reprezentând o decizie proastă de cumpărare, iar 5 fiind valoarea cea mai mare, corespondentă unei achiziţii oportune.

Regresia liniară

Relaţia dintre caracteristicile şi preţul apartamentelor poate fi estimată intuitiv ca fiind liniară. Aceasta înseamnă că modificarea unui atribut are ca rezultat o modificare proporţională în preţ. Algoritmul corespunzător unei legături liniare între atributele de intrare şi cel de ieşire este modelul regresiei liniare, ce face parte din setul de algoritmi funcţionali din WEKA. Folosind clasificarea statistică, regresia liniară determină coeficientul numeric al fiecărui atribut prin analiza unui set de date de antrenament şi raportează eroarea statistică a algoritmului prin calculul coeficientului de corelaţie. Ca o particularitate, algoritmul de regresie liniară implementat în WEKA nu este limitat la atribute numerice, fiind astfel o metodă excelentă de analiză a bazei de date de antrenament în care 17 atribute (caracteristici) determină în mod direct cel de-al 18-lea atribut (preţul).

Cel mai important parametru pentru algoritmul de regresie liniară este metoda de selecţie a atributelor, cu trei opţiuni posibile. La extreme se află metoda „Fără selecţie" ce conduce la obţinerea rapidă a unor rezultate, însă fiind mai puţin selectivă, respectiv metoda „Greedy" ce este considerabil mai lentă însă cu rezultate mai precise. Între ele se află metoda „M5" ce face un compromis între viteză şi acurateţe. Totuşi, trebuie luată în considerare o restricţie: metoda „Greedy" determină formula cea mai precisă, însă are nevoie ca relaţia dintre atributele de intrare şi cel de ieşire să fie cât mai apropiată de una liniară pentru a determina corect valorile maxime locale. Dacă relaţia nu este liniară, metoda „Greedy" va da o formulă eronată.

```
Preţ =
      4.537,74 € pentru 2, 3, 4 sau 5 camere +
      3.851,08 € pentru 3, 4 sau 5 camere +
      2.897,56 € pentru 4 sau 5 camere +
     28.548,94 € pentru 5 camere +
      2.015,49 € pentru decomandat +
      5.328,60 € pentru confort I sau sporit +
      4.467,44 € pentru confort sporit +
      1.201,83 € pentru finisat sau superfinisat +
      3.754,58 € pentru superfinisat +
      1.682,11 € pentru garaj +
      2.520,75 € pentru cartierele Gheorgheni, Grigorescu, Zorilor sau Centru +
       -595,08 € pentru cartierele Grigorescu, Zorilor sau Centru +
      6.855,38 € pentru cartierul Centru +
      1.265,03 €.
```

Fig. 2. Modelul de calcul obţinut în urma algoritmului de regresie liniară

Prin aplicarea algoritmului de clasificare prin regresie liniară pe baza de date completă (1981 înregistrări), atât metoda „Greedy" cât şi metoda „M5" au găsit exact aceeaşi formulă de calcul a preţului apartamentului pe baza caracteristicilor acestuia. (Figura 2)

Coeficientul de corelaţie a fost de 0,702, ceea ce înseamnă că legătura dintre caracteristicile apartamentelor şi preţ se poate aproxima bine cu o legătură liniară. Un coeficient de corelaţie cu valoarea 1 ar indica o legătură liniară perfectă, în timp ce un coeficient de corelaţie cu valoarea 0 ar indica lipsa unei legături între atributele de intrare şi cel de ieşire. Algo-

ritmul de regresie liniară a considerat relevante doar 6 atribute de intrare din 17.
Folosind formula găsită, putem calcula uşor preţul unui apartament în funcţie de caracteristicile acestuia. De exemplu, un apartament cu 3 camere, decomandat, confort sporit, finisat, fără garaj, aflat în cartierul Zorilor, preţul estimat este calculat prin simpla adăugare a coeficienţilor corespunzători valorii fiecărui atribut: 4.537,74 + 3.851,08 + 2.015,49 + 5.328,60 + 4.467,44 + 1.201,83 + 2.520,75 - 595,08 + 1.265,03 = 24.592,88 €.

J48 - Arborele de decizie

J.48 este de fapt implementarea algoritmului C4.5 creat de către J.Ross Quinlan [23], care foloseşte metoda inductivă top-down de construire a arborilor de decizie. Aceştia se construiesc pe baza testării fiecărui nod al arborelui, începând cu nodul rădăcină, pentru fiecare înregistrare. Fiecare nod reprezintă numele unui atribut. Se încearcă introducerea instanţei într-o clasă existentă, pe baza caracteristicilor comune, evaluându-se atributul corespunzător nodului la care s-a ajuns. În funcţie de valoarea sa, instanţa va urma o ramură. Când nu mai există noduri de evaluat, instanţa este clasificată. Dacă o anumită clasă nu mai diferă într-un mod evident de alta în urma introducerii a tot mai multe înregistrări, cele două se vor uni, proces numit „pruning”.

Locul	Atributul	Intensitatea legăturii		Locul	Atributul	Intensitatea legăturii
1	preţ	2.398,843		10	balcon	0
2	camere	1.760,281		11	faianţă	0
3	cartierul	120,16		12	contorizat	0
4	confort	44,849		13	telefon	0
5	finisare	10,946		14	cablu	0
6	etaj	8,646		15	centrală	0
7	garaj	0,475		16	gresie	0
8	decomandat	0,296		17	termopan	0
9	parchet	0		18	modificări	0

Fig. 3. Ordonarea atributelor folosind algoritmul „ChiSquaredAttributeEval”

Deoarece algoritmul J.48 realizează o clasificare cu rezultat discret, nu continuu, şi întrucât preţul apartamentelor este un atribut de tip continuu, s-a impus evaluarea fiecărui apartament pe baza tuturor atributelor, mai puţin numărul de camere, atribut care am considerat că nu trebuie sa influenţeze punctajul. Rezultatul evaluării este un scor nominal pe o scală de la 1 la 5, unde 5 este punctajul cel mai bun şi 1 cel mai slab. Scorul a fost introdus în baza de date pe ultima poziţie, rezultând 19 atribute.
Pentru a realiza o clasificare de calitate a fost necesară evaluarea atributelor. Instrumentul WEKA ales pentru a realiza acest lucru a fost „Attribute evaluator”, folosind „ChiSquaredAttributeEval” prin metoda „Ranker”.
„ChiSquaredAttributeEval” calculează intensitatea legăturii dintre variabile folosind testul HI pătrat (χ^2). Metoda „Ranker” notează atributele în funcţie de evaluare, ordonându-le. Am aplicat evaluarea în funcţie de tipul apartamentelor, adică numărul de camere. (Figura 3) Din rezultatul obţinut reiese că atributele clasate pe locurile 9-18 au obţinut punctaj 0, ceea ce ne-a determinat să renunţăm la folosirea lor în clasificare, relevanţa lor fiind nulă în acest caz însă adăugând în mod inutil complexitate procesului de clasificare. Excluderea lor din baza de date folosită pentru aplicarea algoritmului J.48 duce la creşterea vitezei de creare a modelului şi la o mai bună acurateţe a acestuia.
Algoritmul J.48 a fost aplicat întregului set de date ca set de antrenament, urmând ca testarea să se facă pe un set de test format din alte date. Atributul în funcţie de care se face clasificarea este scorul. Vom obţine astfel un arbore de decizie ale cărui frunze reprezintă scorul categoriei respective. În urma rulării algoritmului s-a obţinut o acurateţe de 87,2%, ceea ce înseamnă că 1728 de instanţe din 1981 au fost clasificate corect în cadrul

modelului creat. (Figura 4)
Arborele de decizie generat are ca nod rădăcină numărul de camere, acest atribut fiind principalul mijloc de a diferenția apartamentele. În cadrul modelului au mai fost găsite ca fiind relevante pentru diferențiere atributele preț, prezent în fiecare ramură principală a arborelui, gradul de finisare, confortul şi cartierul.
În cazul apartamentelor de 5 camere, modelul nu poate fi aplicat cu succes deoarece în setul de date nu există decât 6 înregistrări pentru apartamentele de 5 camere. Lipsa unui număr mai mare de înregistrări a dus la imposibilitatea creării unui model viabil pentru această clasă.
Pentru testarea modelului am utilizat un set de date de test format din 200 de înregistrări, rezultatul având o acuratețe de 85% , ceea ce demonstrează viabilitatea modelului decizional creat pe setul de antrenament.
Exemplu de interpretare a unei ramuri din arborele decizional: „Dacă apartamentul are patru camere, prețul său este între 31.864 şi 44.722 Euro şi este situat în Centru, atunci scorul său este 4.”

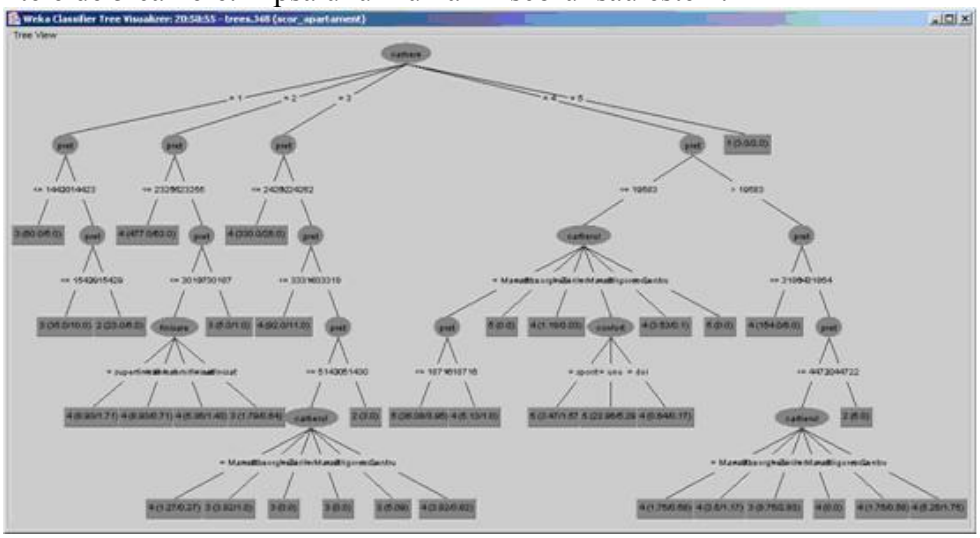

Fig. 4. Arborele de decizie rezultat în urma aplicării J.48 pe setul de antrenament

Concluzii
Rezultatele acestui proiect dovedesc faptul că WEKA este o unealtă potrivită ce oferă metodele fundamentate ştiințific de a extrage cunoştințe noi privind prețurile apartamentelor în funcție de caracteristicile lor şi de a folosi aceste cunoştințe în luarea deciziei de cumpărare.
Luând în considerare schimbările ce au loc în fiecare zi în activitatea economică şi care pot conduce la o piață şi concurență imperfectă, erorile nu pot fi evitate. Deoarece această aplicație se bazează pe informații de pe piață, şi erorile legate de datele propriu-zise sunt prezente. Numărul instanțelor din baza de date nu este foarte mare şi baza de date este rară, lipsind multe valori ale atributelor. Prețurile sunt cele solicitate de către vânzători în anunțurile din ziar, nereflectând cu precizie valorile apartamentelor ci aproximări, de multe ori prețurile fiind psihologice, oportunistice, care nu sunt atinse niciodată în urma negocierilor între vânzători şi cumpărători. O altă posibilă sursă de eroare este convertirea prețurilor exprimate în patru monede diferite la un numitor comun, deoarece nu toți vânzătorii au luat în considerare acelaşi curs de schimb valutar.
Perspective
O aplicație practică a acestui proiect este implementarea lui pe un server web pentru a estima instantaneu prețul unui apartament şi a asista cumpărătorii, vânzătorii şi agențiile imobiliare în evaluarea şi luarea unei decizii privind o tranzacție de vânzare-cumpărare, de asemenea colectând automat date privind noile tranzacții ce au loc şi generând noi modele de calcul în funcție de evoluția pieței. O asemenea bază de date poate oferi rezultate excelente în alte studii statistice şi *data mining*, inclusiv prezicerea schimbărilor în cererea de apartamente în diferite cartiere, in-

formaţii de mare valoare pentru agenţiile imobiliare, oficialităţi locale responsabile cu dezvoltarea oraşului, arhitecţi şi firme de construcţii.
Folosind acest proiect putem elabora şi determina modele şi grafice privind fluctuaţiile sezoniere şi anuale ale preţurilor în funcţie de mai mulţi factori, cum ar fi rotaţia populaţiei studenţeşti a oraşului. O evaluare comparativă a industriei şi pieţei imobiliare din diferite oraşe ale ţării poate ajuta persoanele fizice şi juridice să aleagă un oraş potrivit intereselor lor. O altă perspectivă demnă de luat în seamă este cea a unui studiu comparativ în piaţa imobiliară comunitară pentru a înţelege evoluţia preţurilor apartamentelor înainte şi după aderarea la Uniunea Europeană, sau posibilităţile cetăţenilor europeni de a cumpăra apartamente.

Bibliografie

[1] Enciclopedia Wikipedia: www.wikipedia.org/wiki/Artificial_intelligence
[2] Kasparov versus Deep Blue: www.research.ibm.com/deepblue
[3] Enciclopedia Wikipedia: www.wikipedia.org/wiki/Machine_learning
[4] Enciclopedia Wikipedia, www.wikipedia.org/wiki/Data_mining
[5] PATEL, Nitin, *„Data Mining”*, Cursul 15.062, Massachusetts Institute of Technology, MIT OpenCourseWare, Sloan School of Management, 2003: ocw.mit.edu
[6] Knowledge Discovery Nuggets, 2004: www.kdnuggets.com/polls/2004/data_mining_applications_industries.htm
[7] Oracle9i Data Mining pentru Oracle9i Database Enterprise Edition, şi Oracle Data Mining Suite: www.oracle.com
[8] Teradata Warehouse Mining: www.ncr.com
[9] SAS Enterprise Miner: www.sas.com
[10] Crystal Analysis and Decisions: www.businessobjects.com
[11] Clementine: www.spss.com
[12] SmartDiscovery: www.inxight.com
[13] Monarch: www.datawatch.com
[14] Statistica: www.statsoftinc.com
[15] InfoBase: www.acxiom.com
[16] PolyAnalyst: www.megaputer.com
[17] IBM Intelligent Miner: www.developer.ibm.com/university/scholars
[18] WEKA (Waikato Environment for Knowledge Analysis): www.cs.waikato.ac.nz/ml/weka
[19] ADaM (Algorithm Development and Mining system): datamining.itsc.uah.edu/adam
[20] YALE (Yet Another Learning Environment): www-ai.cs.uni-dortmund.de/SOFTWARE/YALE
[21] OBAE, Petrişor, *„Clujul tace şi le face”*, Capital nr. 18/2004: www.capital.ro/index.jsp?page=archive&magazine_id=279&article_id=14048
[22] Celina Prodcom SRL, *„Piaţa de la A la Z”*, Cluj Napoca: www.piata-az.ro
[23] QUINLAN, Ross, *„C4.5: Programs for Machine Learning”*, Morgan Kaufmann Publishers, San Mateo, CA, 1993