

**Stelian STANCU**

# **Analiza Datelor în mediul R**

## **Teorie și aplicații**

Colecția  
Cibernetică

**Editura ASE**  
**București**  
**2022**



**ACADEMIA DE STUDII ECONOMICE DIN BUCUREȘTI**

**Copyright © 2022, Editura ASE**

Toate drepturile asupra acestei ediții sunt rezervate editurii.

**Editura ASE**

Piața Romană nr. 6, sector 1, București, România

cod 010374

[www.ase.ro](http://www.ase.ro)

[www.editura.ase.ro](http://www.editura.ase.ro)

[editura@ase.ro](mailto:editura@ase.ro)

**Descrierea CIP a Bibliotecii Naționale a României**

**STANCU, STELIAN**

**Analiza Datelor în mediul R : teorie și aplicații** / Stelian Stancu. –  
București : Editura ASE, 2022

Conține bibliografie

ISBN 978-606-34-0417-7

005

**Editura ASE**

**Redactor, tehnoredactor și copertă:** Claudia-Marinela Dumitru

Autorul își asumă întreaga responsabilitate pentru: ideile exprimate, corectitudinea științifică, originalitatea materialului și sursele bibliografice menționate.

# Cuprins

---

<b>Despre autor .....</b>	<b>15</b>
<b>Cuvânt-înainte.....</b>	<b>17</b>
<b>Capitolul 1</b>	
<b>Introducere în Analiza Datelor în contextul Data Science</b>	
<b>Aspecte teoretice generale privind analiza datelor .....</b>	<b>19</b>
1.1 Știința datelor (Data Science) .....	19
1.2 Analiza datelor (Data Mining) .....	19
1.3 Analiza de text (Text Mining).....	21
1.4 Prelucrarea limbajului natural (NLP – Natural Language Processing).....	22
1.5 Tipuri de analiză / Analytcs .....	22
1.6 Clasificarea metodelor de Analiză a Datelor .....	22
1.7 Preprocesarea și standardizarea datelor, utilizând mediul R.....	24
1.7.1 Normalizarea datelor utilizând pachetul reshape, funcția rescaler .....	24
1.7.2 Normalizarea datelor utilizând pachetul clusterSim, funcția data.Normalization .....	26
<b>Capitolul 2</b>	
<b>Reprezentări, descriptori și metrici la nivelul datelor multidimensionale .....</b>	<b>28</b>
2.1 Reprezentarea datelor multidimensionale primare.....	28
2.1.1 Matricea observațiilor (obiectelor, indivizilor).....	28
2.1.2 Matricea sau tabelul de contingență.....	28
2.1.3 Matricea sau tabelul de proximitate .....	29
2.2 Descriptori și metrici la nivelul datelor multidimensionale.....	29
2.2.1 Descriptori și metrici în spațiul indivizilor .....	29
2.2.2 Descriptori și metrici în spațiul variabilelor .....	30
2.2.2.1 Matrice centrată (matricea observațiilor centrate).....	30
2.2.2.2 Matricea datelor (observațiilor) standardizate.....	30
2.2.2.3 Matricea de covarianță .....	30
2.2.2.4 Matrice a corelațiilor (matrice de corelație).....	30
2.3 Schimbarea de variabilă în analiza datelor.....	31
2.3.1 Schimbarea de variabilă prin schimbarea structurii.....	31
2.3.2 Schimbarea de variabilă prin codificare .....	31
2.4 Măsuri de similaritate/disimilaritate. Metrici și ultrametrici .....	32
2.4.1 Definirea noțiunilor de bază .....	32
2.4.2 Măsuri de similaritate/disimilaritate în cazul variabilelor cantitative.....	32
2.4.3 Reprezentări grafice și metrici suplimentare .....	35
2.4.4 Măsuri de similaritate/disimilaritate între grupe de indivizi.....	36

**Capitolul 3****Metode de învățare nesupervizată și supervizată**

<b>(Analiza în componente principale – ACP) .....</b>	<b>39</b>
3.1 Metode aplicate în analiza datelor.....	39
3.2 Metode de învățare nesupervizată.....	41
3.2.1 Analiza în componente principale (ACP). Prezentarea teoretică a metodei ACP.....	41
3.2.2 Aplicație privind analiza în componente principale (ACP) folosind mediul de programare R.....	50
3.2.2.1 Încărcarea pachetelor necesare analizei, în mediul R.....	50
3.2.2.2 Pregătirea datelor .....	51
3.2.2.3 Valori proprii/vectori proprii/varianțe.....	52
3.2.2.4 Prezentarea metodei ACP/PCA.....	53
3.2.2.5 Vizualizarea și interpretarea datelor și rezultatelor obținute.....	57
3.2.2.6 Funcții PCA încorporate în R.....	57
3.2.2.7 Proporția de variație explicată la nivel de ACP/PCA.....	60
3.2.2.8 Stabilirea numărului optim de componente principale într-un model.....	61

**Capitolul 4**

<b>(Analiza factorială – AF) .....</b>	<b>63</b>
4.1 Introducere.....	63
4.2 Definiții și aplicații majore ale analizei factoriale (AF).....	64
4.3 Modele de analiză factorială .....	66
4.4 Probleme fundamentale ale analizei factoriale (AF).....	68
4.5 Condiții preliminare .....	71
4.6 Structura generală a modelului factorial .....	72
4.7 Descompunerea variabilității spațiului inițial .....	75
4.7.1 Spațiul factor și exprimarea conținutului său informațional.....	75
4.7.2 Componentele varianței .....	76
4.8 Configurația factor și structura factor .....	78
4.8.1 Definierea configurației factor.....	78
4.8.2 Definierea structurii factor .....	78
4.9 Calculul scorurilor factoriale .....	79
4.10 Criterii de alegere a numărului de factori .....	80
4.10.1 Criteriul procentului de acoperire .....	80
4.10.2 Criteriul lui Kaiser .....	81
4.10.3 Criteriul "granulozității" .....	81
4.11 Aspecte problematice ale analizei factoriale (AF).....	81
4.12 Aspecte analitice ale analizei factoriale .....	81
4.13 Aplicație privind Analiza factorială (AF) .....	85
4.13.1 Introducere.....	85
4.13.2 Încărcarea pachetelor necesare .....	86
4.13.3 Datele utilizate .....	86

4.13.4 Descrierea datelor .....	87
4.13.5 Evaluarea factoriabilității datelor.....	87
4.13.5.1 Testul Bartlett, de sfericitate .....	88
4.13.5.2 Testul KMO .....	88
4.13.6 Determinarea numărului de factori de extras/comuni.....	89
4.13.6.1 Scree plot-ul .....	89
4.13.6.2 Analiza paralelă.....	90
4.13.7 Realizarea analizei factoriale .....	91
4.13.7.1 Factorizarea axei principale (analiza factorilor comuni).....	91
4.13.7.2 Analiza în componente principale.....	95
4.13.8 Concluzii finale.....	97

## Capitolul 5

<b>Analiza factorială a corespondențelor – AC .....</b>	<b>98</b>
5.1 Introducere .....	98
5.2 Analiza factorială a corespondențelor simple .....	99
5.3 Matricea sau tabelul de contingență.....	100
5.4 Analiza bivariată a corespondențelor .....	101
5.4.1 Aspecte generale.....	101
5.4.2 Analiza în componente principale a matricei de inerție.....	102
5.5 Analiza factorială a corespondențelor multiple (Analiza Omogenității – HOMALS) .....	104
5.6 Analiza factorială a corespondențelor multiple în mediul R. Aplicație .....	105
5.6.1 Aspecte generale .....	105
5.6.2 Rezumatul datelor .....	107
5.6.3 Vizualizarea și interpretarea ieșirilor ACM .....	111
5.6.4 Valori proprii/variații și scree plot .....	112
5.6.5 Calitatea reprezentării variabilelor categoriale .....	118
5.6.6 Graficul indivizilor .....	124
5.6.7 Descrierea dimensiunii .....	129
5.6.8 Utilizarea funcției plotellipses(), din pachetul FactoMineR .....	133

## Capitolul 6

<b>Învățarea automată nesupervizată, supervizată și semisupervizată în mediul R .....</b>	<b>137</b>
6.1 Introducere .....	137
6.2 Învățarea automată versus programarea .....	137
6.3 Aspecte teoretice generale .....	138
6.4 Tipuri de învățare automată .....	140
6.4.1 Învățarea automată supervizată/supravegheată (Supervised Machine Learning) .....	140
6.4.2 Învățarea automată nesupervizată/nesupravegheată (Unsupervised Machine Learning) .....	141

6.4.3 Învățarea automată semisupervizată/semisupravegheată sau prin reîntărire (de consolidare).....	143
6.5 Învățarea automată (ML) – o abordare pentru realizarea inteligenței artificiale.....	144
6.5.1 Aspecte generale.....	144
6.5.2 Învățarea automată (ML) – rol de clasificator de obiecte.....	145
6.6 Învățarea profundă/complexă (DL) – o tehnică pentru implementarea învățării automate (ML).....	147
6.7 Extragerea automată a caracteristicilor relevante, utilizând învățarea profundă/complexă (DL).....	148
6.8 Exemple de aplicare a învățării profunde/complexe (DL).....	149
6.8.1 Detectarea obiectelor.....	149
6.8.2 Analiza de text.....	149
6.8.3 Traducerea automată.....	150
6.8.4 Generarea imaginilor pentru legendă (subtitrarea imaginilor).....	150
6.8.5 Recunoașterea entităților definite.....	150
6.9 Algoritmi de învățare nesupervizată/nesupravegheată.....	150

## Capitolul 7

### Învățarea nesupervizată

<b>Analiza de tip cluster și algoritmi de clusterizare în mediul R.....</b>	<b>152</b>
7.1 Analiza cluster – aspecte generale.....	152
7.1.1 Introducere.....	152
7.1.2 Scopul analizei cluster.....	152
7.1.3 Încărcarea pachetelor de bază în mediul R.....	152
7.2 Metoda <i>k-mean clustering</i> sau algoritmul <i>centrului mobil</i> .....	153
7.2.1 Aspecte generale.....	153
7.2.2 Pregătirea datelor.....	154
7.2.3 Metoda propriu-zisă.....	156
7.2.4 Măsurarea calității unei partiții <i>k-mean</i> .....	161
7.2.5 Pașii pentru implementarea <i>k-mean clustering</i> pe datele iris.....	162
7.3 Metode de grupare ierarhică (Hierarchical Clustering Algorithms).....	166
7.3.1 Aspecte generale.....	166
7.3.2 Pregătirea datelor.....	166
7.3.3 Metoda propriu-zisă (alomerativă).....	168
7.3.4 Lucrul cu dendrograme.....	176
7.4 Determinarea numărului optim de clustere în cazul metodei <i>k-mean clustering</i> .....	179
7.4.1 Metoda cotului.....	179
7.4.2 Metoda siluetei medii.....	182
7.4.3 Metoda statisticii gap.....	184
7.4.4 Extragerea rezultatelor.....	186

7.5 Determinarea numărului optim de clustere în cazul metodelor de grupare ierarhică (Hierarchical Clustering Algorithms).....	189
7.5.1 Metoda cotului .....	189
7.5.2 Metoda siluetei medii .....	190
7.5.3 Metoda statisticii gap .....	191
7.6 Alegerea celui mai bun dintre algoritmi de clustering .....	191
7.7 Analiza cluster Top 50 melodii Spotify – 2019 .....	194
7.7.1 Analiza explorativă a bazei de date .....	194
7.7.2 Selectarea variabilelor .....	197
7.7.3 K-means clustering .....	198
7.7.4 Analiza în componente principale .....	199
7.7.5 Combinarea PCA cu K-Means .....	203

## Capitolul 8

### Învățarea supervizată

<b>Algoritmi de clasificare a datelor în mediul R.....</b>	<b>206</b>
8.1 Introducere .....	206
8.2 Regresia logistică .....	206
8.2.1 Regresia logistică binomială.....	206
8.2.2 Regresie logistică multinomială .....	209
8.2.3 Regresie logistică ordinală.....	209
8.2.4 Regresie logistică în R cu glm .....	209
8.3 Clasificatorul naiv Bayesian (Naive Bayes) .....	217
8.4 Algoritmii K-vecini apropiati (K-Nearest Neighbors Classification-KNN) .....	221
8.4.1 Funcționarea algoritmului KNN .....	221
8.4.2 Aplicații .....	223
8.5 Random Forest – ca algoritm de clasificare.....	236

## Capitolul 9

<b>Suport vectori mașină (SVM) în mediul R .....</b>	<b>241</b>
9.1 SVM – aspecte generale.....	241
9.1.1 Introducere.....	241
9.1.2 Încărcarea pachetelor suplimentare, necesare pe lângă pachetul de bază din mediul R .....	241
9.2 Metoda SVM folosită în clasificarea simplă (booleană).....	242
9.2.1 SVM – clasificator liniar. Construirea modelului și prezentarea clasificatorului de marjă maximă .....	242
9.2.2 SVM – clasificator neliniar.....	245
9.3 Metoda SVM folosită în clasificarea multiplă (cel puțin 3 variabile).....	248
9.4 Metoda SVM aplicată setului de date producția .....	251
9.4.1 Pregătirea datelor pentru aplicarea SVM.....	251
9.4.2 Prelucrarea propriu-zisă a datelor utilizând SVM .....	253

## Capitolul 10

<b>LDA (Linear Discriminant Analysis) versus PCA (Principal Component Analysis), rol în clasificarea și reducerea dimensionalității datelor .....</b>	<b>257</b>
10.1 LDA ca tehnică de reducere a dimensionalității și totodată	
un algoritm de clasificare.....	257
10.1.1 Introducere în LDA.....	257
10.1.2 Analiza discriminant liniară (LDA) – utilizarea funcției lda() .....	258
10.1.3 Aplicarea funcției lda() pe setul de date "iris" .....	258
10.1.4 LDA ca algoritm de clasificare.....	265
10.1.5 LDA ca tehnică de reducere a dimensionalității .....	266
10.1.6 Rezultate comparative obținute din cele două abordări.....	268
10.2 PCA ca tehnică de reducere a dimensionalității și totodată	
un algoritm de clasificare.....	268
10.2.1 Introducere în PCA .....	268
10.2.2 PCA ca algoritm de clasificare .....	268
10.2.3 Structura spațiului componentelor principale .....	269
10.2.4 Coduri R în spațiul componentelor principale.....	270

## Capitolul 11

<b>Rolul rețelelor neuronale în clasificare folosind programarea R .....</b>	<b>275</b>
11.1 Aspecte generale .....	275
11.2 Construirea, antrenarea/testarea, predicția și validarea unei RNA	
utilizând mediul R .....	276
11.3 Implementarea în R a clasificării (învățării supervizate) .....	284
11.4 Construirea unui clasificator, folosind biblioteca neuralnet în mediul R.....	294
11.4.1 Despre pachetul neuralnet.....	294
11.4.2 Pașii de construire a unei rețele neuronale artificiale (RNA)	
utilizând pachetul neuralnet.....	295
11.4.3 Implementarea unei rețele neuronale în mediul R	
utilizând pachetul neuralnet fără normalizarea datelor.....	297
11.4.3.1 Instalarea pachetului necesar.....	297
11.4.3.2 Crearea unui set de date de antrenare/instruire .....	298
11.4.3.3 Construirea unui clasificator, folosind biblioteca neuralnet .....	298
11.4.3.4 Reprezentarea grafică a unei rețele neuronale .....	298
11.4.3.5 Crearea unui set de date de testare .....	299
11.4.3.6 Previzionarea rezultatelor pentru setul de testare .....	299
11.4.3.7 Convertirea probabilităților în clase binare .....	299
11.5 Învățarea automată nesupravegheată în RNA .....	299



## Capitolul 12

### Învățarea automată (ML)

<b>și algoritmi de tip Machine Learning în mediul R .....</b>	<b>302</b>
12.1 Pașii de pregătire și aplicare a învățării automate (ML) pe un set de date....	302
12.1.1 Instalarea pachetelor de lucru și a seturilor de date .....	302
12.1.1.1 Instalarea pachetelor de lucru și a seturilor de date inițiale.....	302
12.1.1.2 Crearea setului de date de validare.....	302
12.1.2 Rezumatul statistic al setului de date, summary() .....	302
12.1.3 Vizualizarea setului de date .....	303
12.1.4 Evaluarea algoritmilor .....	304
12.1.4.1 Cadrul automat de testare și evaluare a modelelor.....	304
12.1.4.2 Construirea de modele.....	304
12.1.5 Selecția celui mai bun model.....	305
12.1.6 Realizarea de predicții .....	306
12.2 Gruparea primară a algoritmilor de tip Machine Learning .....	306
12.2.1 Algoritmi de învățare supervizată/supravegheată.....	306
12.2.2 Algoritmi de învățare nesupervizată/nesupravegheată .....	307
12.2.3 Algoritmi de învățare semisupervizată/semisupravegheată.....	307
12.3 Prezentare generală a algoritmilor de tip Machine Learning (ML) .....	308
12.3.1 Algoritmi bazați pe similitudini .....	308
12.3.1.1 Algoritmi de tip ML pentru regresie .....	309
12.3.1.2 Algoritmi bazați pe instanțe .....	309
12.3.1.3 Algoritmi de regularizare .....	310
12.3.1.4 Algoritmi bazați pe arbori de decizie .....	311
12.3.1.5 Algoritmi Bayes-ieni.....	311
12.3.1.6 Algoritmi de clustering.....	312
12.3.1.7 Algoritmi de învățare bazați pe reguli de asociere .....	312
12.3.1.8 Algoritmi bazați pe rețele neuronale artificiale.....	313
12.3.1.9 Algoritmi de tip Deep Learning-DL (de învățare profundă).....	314
12.3.1.10 Algoritmi de reducere a dimensionalității.....	314
12.3.1.11 Algoritmi de asamblare .....	315
12.3.1.12 Alți algoritmi de tip Machine Learning.....	316
12.4 Pachetele de top din mediul R utilizate în sfera Machine Learning.....	316

## Capitolul 13

<b>Predicția unui fenomen socioeconomic .....</b>	<b>319</b>
13.1 Formularea problemei .....	319
13.2 Predicția utilizând modele stohastice.....	319
13.3 Predicția utilizând modele bazate pe funcții de transfer .....	320
13.4 Alegerea unei metode de predicție.....	320

<b>Bibliografie .....</b>	<b>327</b>
<b>Anexe .....</b>	<b>333</b>
A1. Eliminarea outlierilor în Excel.....	333
A2. Eliminarea outlierilor în R.....	333