# INTELIGENȚĂ ARTIFICIALĂ - TEMA 2
# MODELE DE ÎNVĂȚARE AUTOMATĂ




**Cătălin-Alexandru Rîpanu, 341C3**
Facultatea de Automatică și Calculatoare
Universitatea Națională de Știință și Tehnologie Politehnica București
catalin.ripanu@stud.acs.upb.ro

14 Ianuarie 2024

**ABSTRACT**

Această temă propune implementarea, folosind diverse tehnici, a unor modele de învățare automată ce folosesc 2 algoritmi des întâlniți în practică, și anume algoritmul de **Regresie Logistică**, respectiv algoritmul de **Arbore de Decizie**. Ambii algoritmi au fost evaluați pe baza acestor măsurători: **Precizie**, **Recall**, **Acuratețe** și **F1**. Mai mult, pentru variantele alternative de implementare, în cazul celor 2 algoritmi, s-a folosit pachetul de Python **sklearn** ce expune numeroase funcții specializate în domeniul Învățării Automate.

## 1 Cerința 1: Analiza Datelor

Prima parte se referă la Analiza Datelor, mai exact se stabilește, prima dată, dacă seturile de antrenare, respectiv de testare sunt echilibrate, după se formează grafice / histograme elocvente pentru vizualizarea distribuțiilor valorilor atributelor în percentile cu granularitate de 10%, și, nu în ultimul rând, se trasează plot-uri în cazul fiecarui atribut pentru a se observa relevanța acestuia în raport cu predicția variabilei țintă (cu denumirea de **"Revenue"**).

Se studiază **Corelația Point-Biserial** și **Valoarea Statistică Chi-Squared** pentru a stabili, numeric, gradul de corelație cu ținta predicției (se folosesc 2 analize diferite întrucât atributele din setul de date sunt fie categorice, fie pur numerice).

S-au făcut 10 împărțiri într-un mod aleatoriu în cazul setului inițial de date (toate împărțirile au rezultat în seturi dezechilibrate), de asemenea, toate aceste diviziuni au fost date spre modelele de interes pentru a studia capacitatea lor de predicție în raport cu informațiile din fișier (prin aceste teste putem observa dacă sunt suficiente date și dacă sunt relevante în raport cu variabila ce se dorește a fi prezisă).

În general, se consideră faptul că atributele care au p-value mai mic decat 0.05 în raport cu un alt atribut sunt dependente de acel atribut. În cazul acesta, există câteva atribute care sunt corelate cu ținta predicției, prin urmare, dacă aceste variabile au valori extreme (outliers) și nu se scalează corespunzător, atunci modelul (fie el de regresie logistică sau de arbore de decizie) riscă să nu producă rezultate satisfăcătoare.

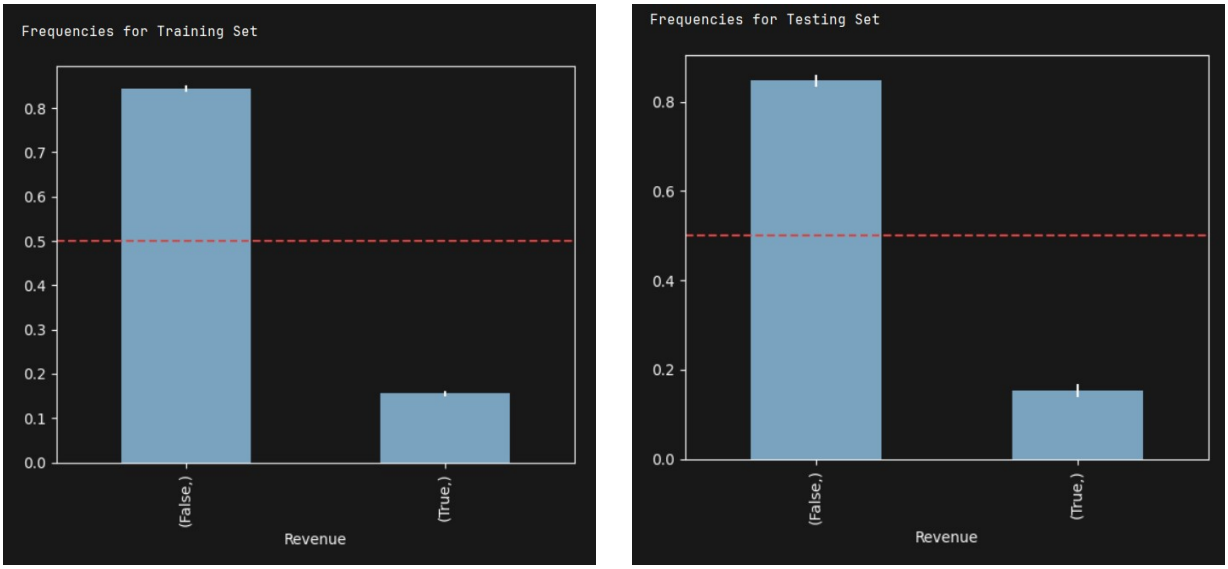


Figure 1: Distribuția claselor în ambele seturi de date

Tabel pentru raportarea valorii Point-Biserial în cazul tuturor atributelor numerice:

| Atribut | Valoarea Point-Biserial | Valoarea P-Value |
|---|---|---|
| Administrative | 0.13891709415067247 | 3.5197598377762157e-54 |
| Administrative_Duration | 0.09358671905704398 | 2.1465136347391026e-25 |
| Informational | 0.09520034257205935 | 3.174034311253924e-26 |
| Informational_Duration | 0.0703445023459854 | 5.282870860711972e-15 |
| ProductRelated | 0.15853798428087626 | 3.2411873289126146e-70 |
| ProductRelated_Duration | 0.15237261055701173 | 6.1153382534504874e-65 |
| BounceRates | -0.15067291192605678 | 1.594197538507772e-63 |
| ExitRates | -0.20707108205527436 | 1.662653625062138e-119 |
| PageValues | 0.49256929525120163 | 0.0 |
| SpecialDay | -0.08230459817953507 | 5.498934260216157e-20 |
| OperatingSystems | -0.01466755959420953 | 0.10339431071061646 |
| Browser | 0.023984289404092515 | 0.007736888294955789 |
| Region | -0.01159506777780058 | 0.19794262499410128 |
| TrafficType | -0.0051129705027557195 | 0.5702433635793045 |

Se observă că, de exemplu, atributul "PageValues" are o corelație mai mică decât 0.05 (este chiar aprox. 0), deci spunem despre acesta că este dependent de țintă.

Tabel pentru raportarea valorii statistice de Chi-Squared în cazul tuturor atributelor categorice:

| Atribut | Valoarea Statistică Chi-Squared | Valoarea P-Value |
|---|---|---|
| Weekend | 10.58183514829754 | 0.005037136192599038 |
| VisitorType | 135.2519228192047 | 3.991223433643256e-29 |
| TrafficType | 373.14556468814857 | 7.442024647470126e-67 |
| Region | 9.252751430579846 | 0.4142778663586834 |
| Browser | 27.71529940138156 | 0.00991396435211927 |
| OperatingSystems | 75.02705620298461 | 4.871504098402714e-13 |
| Month | 384.93476153599426 | 1.5092752396293818e-76 |

## 2 Cerința 2: Antrenarea modelelor

Aici se vor pune statisticile relevante pentru cele 2 modele menționate anterior, folosind configurațiile aferente lor (scalări, adâncimi, etc).

Măsurătorile implementărilor manuale:

| Model | Scalare | Adâncime | Medie Precision | Medie Recall | Medie F1 | Varianță Precision | Varianță Recall | Varianță F1 |
|---|---|---|---|---|---|---|---|---|
| Regresie Logistică | MinMax | - | 0.753 | 0.319 | 0.339 | 0.0334 | 0.0896 | 0.0512 |
| Regresie Logistică | Standard | - | 0.565 | 0.441 | 0.462 | 0.0362 | 0.02003 | 0.0088 |
| **Regresie Logistică** | Robust | - | 0.545 | 0.487 | 0.483 | 0.0323 | 0.0507 | 0.029 |
| Arbore de Decizie | None | 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| Arbore de Decizie | None | 4 | 0 | 0 | 0 | 0 | 0 | 0 |
| Arbore de Decizie | None | 5 | 0 | 0 | 0 | 0 | 0 | 0 |
| Arbore de Decizie | None | 6 | 0 | 0 | 0 | 0 | 0 | 0 |
| Arbore de Decizie | MinMax | 3 | 0.289 | 0.056 | 0.087 | 0.017 | 0.001 | 0.0021 |
| Arbore de Decizie | MinMax | 4 | 0.267 | 0.078 | 0.116 | 0.00204 | 0.00164 | 0.00180 |
| Arbore de Decizie | MinMax | 5 | 0.256 | 0.121 | 0.155 | 0.00153 | 0.00401 | 0.00247 |
| Arbore de Decizie | MinMax | 6 | 0.23 | 0.175 | 0.193 | 0.0005 | 0.0068 | 0.0032 |
| Arbore de Decizie | Standard | 3 | 0.237 | 0.116 | 0.141 | 0.0055 | 0.0089 | 0.0067 |
| Arbore de Decizie | Standard | 4 | 0.271 | 0.134 | 0.159 | 0.0101 | 0.0102 | 0.0113 |
| **Arbore de Decizie** | Standard | 5 | 0.27 | 0.263 | 0.252 | 0.0133 | 0.0449 | 0.0215 |
| Arbore de Decizie | Standard | 6 | 0.26 | 0.228 | 0.228 | 0.0015 | 0.01301 | 0.0045 |
| Arbore de Decizie | Robust | 3 | 0.183 | 0.068 | 0.091 | 0.0055 | 0.00204 | 0.0024 |
| Arbore de Decizie | Robust | 4 | 0.211 | 0.136 | 0.128 | 0.00276 | 0.0285 | 0.0089 |
| Arbore de Decizie | Robust | 5 | 0.189 | 0.128 | 0.138 | 0.00238 | 0.0099 | 0.0072 |
| Arbore de Decizie | Robust | 6 | 0.23 | 0.152 | 0.165 | 0.00208 | 0.01392 | 0.00785 |

S-a constatat faptul că arborele de decizie fără ajutorul unei normalizări nu poate prezice clasa etichetată cu True (clasifică toate exemplele de test ca fiind False, acest lucru petrecându-se din cauza faptului că există valori extreme și că setul de date este dezechilibrat).

În schimb, Regresia Logistică cu scalarea RobustScale obține cele mai bune rezultate datorită faptului că gestionează eficient variabilele continue cu ajutorul normalizării atașate.