Aplicație de clasificare statistică implementată cu SVM

Trandafir Victor-Gabriel

January 5, 2024

Introducere

Clasificarea este problema (de învățare automată) de a identifica din ce categorie face parte o observație (sau mai multe observații, "etichetate" corespunzător, clasificarea având nevoie de metode de tip "supervised learning") dintr-un set de categorii. Un exemplu este atribuirea unui diagnostic unui anumit pacient pe baza caracteristicilor observate ale acestuia. Adesea, observațiile individuale sunt analizate într-un set de proprietăți cuantificabile, cunoscute în mod diferit ca variabile explicative sau caracteristici. Aceste proprietăți pot fi, în mod variat, categorice (de exemplu, "A", "B", "AB" sau "O", pentru grupa sanguină), ordinale (de exemplu, "mare", "mediu" sau "mic"), cu valori întregi sau cu valori reale.

Support vector machine

SVM este un model cu marjă maximă (calculează parametrii unui hiperplan pentru a separa obiecte în 2 clase,chiar și în cazul în care nu toate obiectele pot fi separate) cu algoritmi de învățare care analizează datele pentru clasificare. În plus față de efectuarea clasificării liniare, SVM-urile pot efectua eficient o clasificare neliniară folosind ceea ce se numește "kernel trick", cartografiind (proiectând) implicit intrările lor în spații de caracteristici de dimensiuni înalte. Popularitatea SVM-urilor se datorează probabil ușurinței cu care se pretează la analize teoretice, flexibilității cu care se aplică la o mare varietate de sarcini, inclusiv la probleme de predicție structurate.

Formulare Matematică pentru SVM

Pentru modelul SVM, luăm în considerare următoarele elemente:

- Setul de date de antrenare: $(\mathbf{X}_i, y_i)$, unde $\mathbf{X}_i$ reprezintă vectorul de caracteristici, iar y_i este eticheta de clasă asociată.
- Hyperplanul definit de $\mathbf{w}^T\mathbf{X} + b = 0$, unde $\mathbf{w}$ este vectorul de greutăți, iar b este termenul de decizie.
- Termenul de penalizare pentru clasificare greșită: C, controlând costul erorilor de clasificare.
- Parametrii modelului: $\mathbf{w}$ și b sunt optimizați pentru a maximiza marginea dintre clase și a minimiza erorile de clasificare.

$$\text{Minimizare} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i \tag{1}$$

Formulare Matematică pentru SVM (continuare)

Sub restricția:

$$y_i(\mathbf{w}^T\mathbf{X}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \tag{2}$$

unde ξ_i sunt variabilele de slack, iar N reprezintă numărul de instanțe în setul de date.

SVM Dual cu Kernel Trick:

Problema dual̆ă implică introducerea unei matrice de kernel K, unde $K_{ij} = y_i y_j K(\mathbf{X}_i, \mathbf{X}_j)$. Clasificatorul este apoi definit ca:

$$\text{Clasificare}(\mathbf{a}) = \sum_i y_i \alpha_i K(\mathbf{X}_i, \mathbf{a}) - b \tag{3}$$

Kernel trick-ul permite gestionarea eficientă a intrărilor non-separabile.

Setul de date

- Aplicația aleasă: Clasifică persoanele descrise de un set de atribute ca fiind bune sau rele ca risc de credit.
- Informații din Dataset: Date privind creditele în Germania (https://archive.ics.uci.edu/dataset/144/statlog+german+credit+da
- 20 Atribute (multivariate ca tip de date: categorice, numerice întregi și reale): soldul contului curent, Durata (în luni), Istoricul de credit, Scopul, Valoarea creditului, Economii/obligațiuni, Informații privind angajarea actuală, Rata în procent din venitul disponibil, Stare civilă, Alți debitori/garantori, Reședința actuală, Proprietăți, Vârsta, Alte planuri de rate, Alte informații privind locuința, Numărul de credite existente la aceeași bancă, Ocupația, Numărul de persoane responsabile pentru întreținere, Telefonul, Dacă este lucrător străin.
- Variabilă Țintă: clasă (1 = Bun, 2 = Rău)
- Instanțe: 1000.

Preprocesarea (prelucrearea) Datelor

- Obținerea și examinarea setului de date: presupunem că datele culese dunt suficiente, reprezentative și de calitate (fără erori și zgomot), deci curată.
- Verificarea existenței valorilor nule și duplicatelor și scoaterea lor în caz că există
- Transformarea datelor categorice în date numerice

Preprocesarea datelor: Reducerea dimensionalității 1

- Analiza de corelație evaluează relațiile liniare dintre caracteristici, ajutând la identificarea și abordarea multicoliniarității. O "hartă termică" (heatmap) evidențiază caracteristicile puternic corelate, permițând selectarea celor relevante și ghidând crearea de variabile compozite pentru a păstra informațiile esențiale:

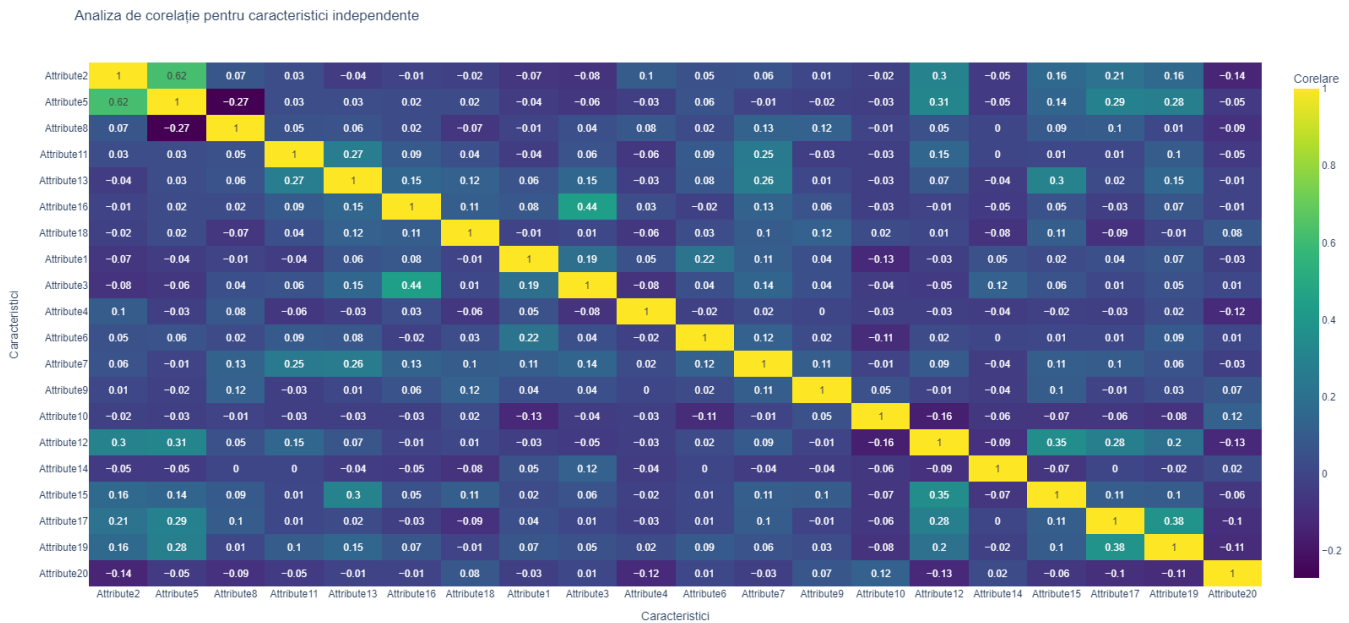


Figure: Heatmap inițial

Preprocesarea datelor: Reducerea dimensionalității 2

- Din rezultatele de mai sus, observăm corelarea dintre durată și valoarea creditului (> 0.6), deci aplicăm PCA pentru a reține informațiile importante ale acestor 2 caracteristici și punerea rezultatului într-o coloană nouă și scoatem coloanele cu atributele corelate:

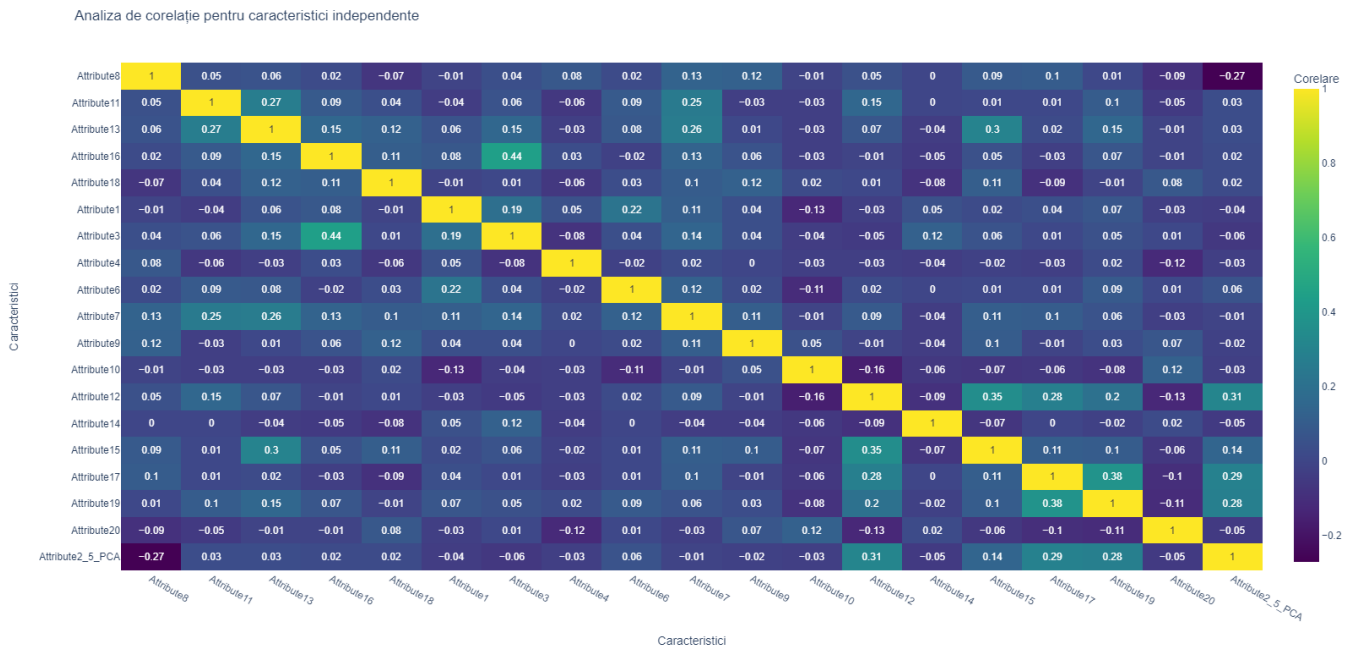


Figure: Heatmap după reducerea dimensională

Pregătirea și aplicarea Modelului

- Împărțirea datelor în seturi de antrenare și testare
- Normalizarea prin scalare min-max a datelor
- Calcularea acurateței de referință (care trebuie depășită): 70.5% (pentru ambele seturi de date).
- Aplicarea SVM

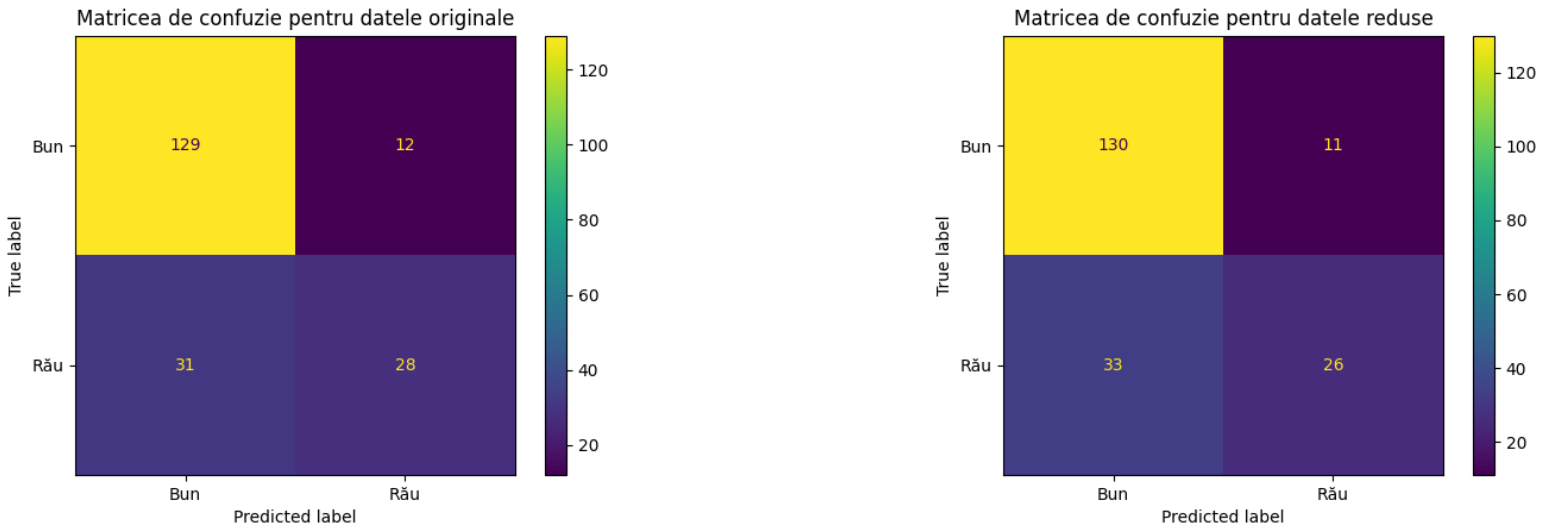
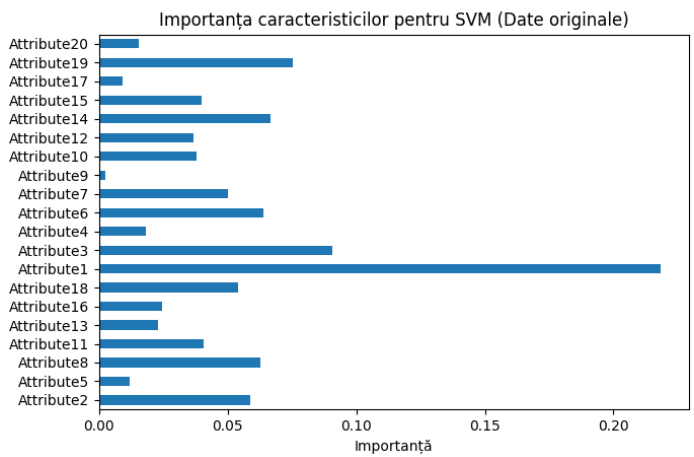


Figure: Matricile de confuzie

Rezultate și evaluarea modelului

- Kernel optim pentru SVM:am parcurs diferite tipuri de kernel pentru un clasificator vectorial de suport (SVC) pentru a determina care kernel oferă cea mai bună performanță - polinomial pentru datele orginale și gaussian (RBF) pentru datele reduse.
- Acuratețea Clasificatorului de Suport Vectorial (SVC) (cu cel mai bun kernel): 78.5% pentru datele originale și 78% pentru datele reduse.



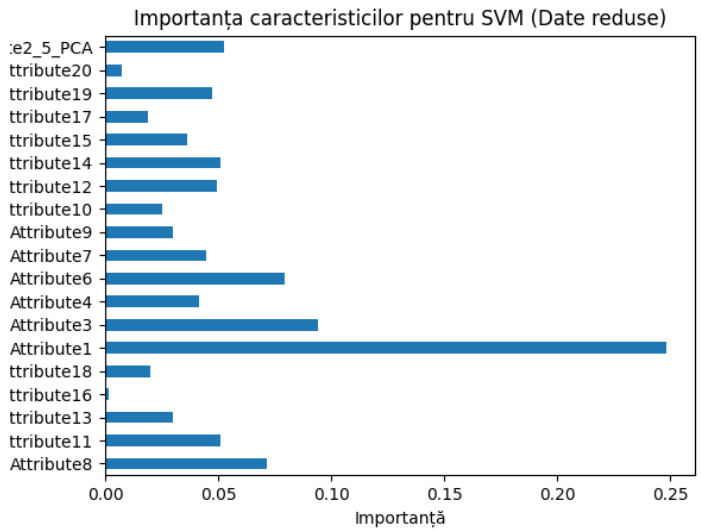


Figure: Diagramele importanței caracteristicilor

Matricea de cost

Acest set de date vine cu o matrice de cost:

	1	2
1	0	1
2	5	0

(1 = Bun, 2 = Rău) Rândurile reprezintă clasificarea reală, iar coloanele reprezintă clasificarea prezisă. Deci, este mai rău să clasifici un client drept bun când este rău (5), decât să îl clasifici drept rău când este bun (1). Dacă aplicăm această matrice, kernelul cel mai bun pentru datele originale și cele reduse este cel liniar, iar precizia, în ambele cazuri, este egală cu precizia de bază (70.5%).

Matricea de confuzie cu aplicarea matricei cost

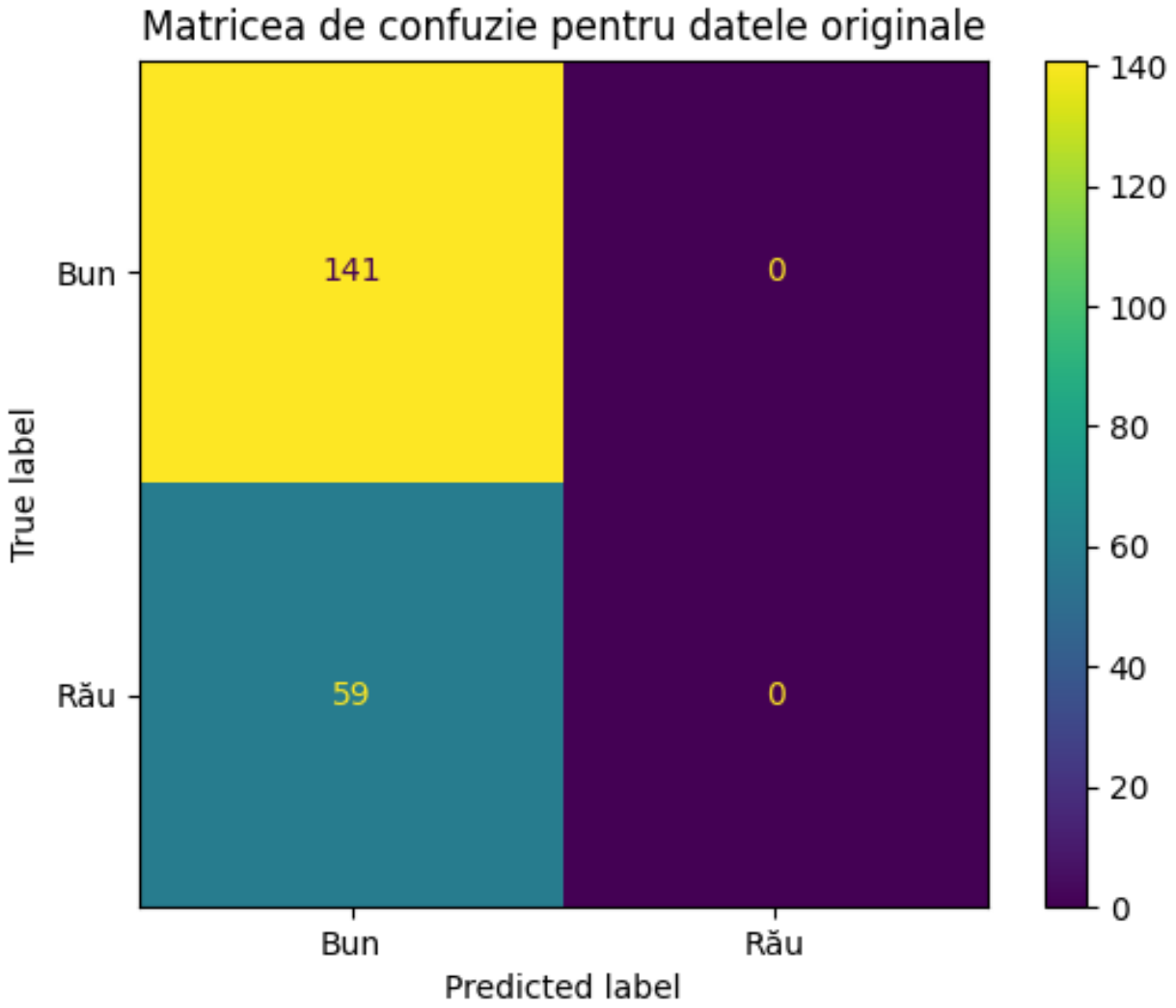


Figure: Matricea de confuzie la aplicarea matricei cost

Principalele constatări și implicațiile rezulatatelor

Caracteristica cea mai importantă a reieșit a fi soldul contului curent (care poate fi în una din 4 categorii: mai puțin de 0 mărci, între 0 și 200 mărci, mai mult de 200 mărci sau încasări salariale de minim un an sau că nu există vreun cont curent), iar persoanele sunt mai degrabă apte să ia un credit, în cazul aplicării matricei de cost chiar total. Performanțele SVM sunt acceptabile iar probabil reducerea preciziei odată cu reducerea dimensiunii datelor este rezonabilă rezonabilă (0.5% pentru 5% - un atribut scos din 20).

Considerente viitoare și îmbunătățiri

- Explorarea și utilizarea altor modele de învățare automată, precum rețele neuronale sau alte tehnici de clasificare, pentru a compara performanța și a identifica soluții mai eficiente.
- Extinderea setului de date sau adăugarea de caracteristici noi pentru a obține o reprezentare mai cuprinzătoare și complexă a comportamentului oamenilor relativ la creditele lor.
- "Refine" pentru hiperparametri și selecția kernelului pentru SVM pentru a îmbunătăți potențial acuratețea clasificării.
- Investigarea unor metode avansate de preprocesare a datelor și reducere a dimensionalității pentru a optimiza performanța modelelor.
- Integrarea unei evaluări de impact social și etic în procesul de clasificare, pentru a asigura o decizie corectă și imparțială.

Vă mulțumesc!