ACADEMIA DE STUDII ECONOMICE DIN BUCUREȘTI

FACULTATEA DE CIBERNETICĂ, STATISTICĂ ȘI INFORMATICĂ ECONOMICĂ

DEZVOLTAREA SOFTWARE PENTRU

ANALIZA DATELOR

Analiza bolilor cardiovasculare în rândul populației

Tema 2 – Analiză liniară discriminantă

Andries Victor-Mihai

Arsene Sabin-Marian

Alexandru Robert-Mihai

GRUPA: 1084

Profesor coordonator: Furtuna Felix Titus

București 2024

CUPRINS

Introducere

Proiectul nostru urmărește să exploreze și să analizeze un set de date complex privind bolile de inimă, obținut de la unul dintre spitalele de specialitate din India. Acest set de date reprezintă o resursă valoroasă în domeniul cercetării medicale, oferind o colecție de 14 caracteristici comune, colectate de la 1000 de subiecți. Informațiile fac ca setul de date să fie unul dintre cele mai relevante și cuprinzătoare resurse disponibile pentru studiul bolilor de inimă. Importanţa acestui proiect rezidă în potențialul său de a contribui la dezvoltarea unor modele predictive prin realizarea de tehnici pentru o învățare automată, care să faciliteze detectarea precoce a bolilor de inimă. Un aspect central al analizei noastre va fi aplicarea analizei liniare discriminante (Linear Discriminant Analysis - LDA). Această metodă statistică este utilizată pentru a identifica pattern-uri care separă cel mai eficient clasele în cadrul setului de date - în cazul nostru, prezenta sau absența bolii de inimă.

Prin utilizarea LDA, ne propunem să extragem caracteristici relevante care contribuie la clasificarea stării de sănătate a inimii. Aceasta ne va permite să înțelegem mai bine factorii de risc și să dezvoltăm un model capabil să prevadă riscul de boală cardiacă cu o acuratețe ridicată. Studiul nostru are potențialul de a influența pozitiv practicile medicale prin furnizarea de instrumente avansate pentru diagnosticare și prevenție în domeniul cardiologiei.

- Setul de date utilizat :

https://www.kaggle.com/datasets/jocelyndumlao/cardiovascular-disease-dataset
https://data.mendeley.com/datasets/dzz48mvjht/1

Fisiere utilizate și procesarea datelor

Proiectul conține următoarele fișiere utilizate:

Cardiovascular_Disease.csv

Avem datele anonime despre pacienți cu durerile și bolile cu care au fost găsiți. De exemplu, pacientul cu ID-ul 103368 are 53 de ani, este de gen masculin, a experimentat un anumit tip de durere toracică, are o tensiune arterială ridicată în repaus și un număr de alte caracteristici care sugerează că a fost diagnosticat cu boala de inimă (target = 1).

Cardiovascular_Disease_Description.pdf

Descrierea setului de date pentru boli cardiovasculare

Nr.	Atribut	Cod asignat	Unitate	Tip de Dată
1	Numărul de identificare a pacientului	patientid	Numar	Numeric
2	Vârstă	age	În ani	Numeric
3	Gen	gender	1,0(0=femiinin, 1=masculin)	Binar
4	Tip de durere toracică	chestpain	0,1,2,3 (Value 0: angina tipică Value 1: angină atipică Value 2: durere non-anginoasă Value 3: asimptomatică)	Nominal
5	Tensiunea arterială în repaus	restingBP	94-200 (în mm HG)	Numeric
6	Colesterolul seric	serumcholestrol	126-564 (în mg/dl)	Numeric
7	Glicemia	fastingbloodsugar	0,1 > 120 mg/dl (0 = fals , 1 = adevărat)	Binar
8	Rezultate electrocardiograme de repaus	restingrelectro	0,1,2 (Valoarea 0: normal, Valoarea 1: având o anomalie a undei ST-T (inversări ale undei T și/sau supradenivelare sau scădere a ST > 0,05 mV), Valoarea 2: care arată hipertrofie ventriculară stângă probabilă sau certă de către Estes criterii)	Nominal
9	Ritmul cardiac maxim atins	maxheartrate	71-202	Numeric
10	Angina pectorală indusă de efort	exerciseangia	0,1 (0 = nu, 1 = da)	Binar
11	Segmentul ST	oldpeak	0-6.2	Numeric
12	Panta segmentului ST de vârf de exercițiu	slope	1,2,3 (1-în pantă, 2-plat, 3-în panta descendenta)	Nominal
13	Numărul de vase majore	noofmajorvessels	0,1,2,3	Numeric
14	Clasificare	target	0,1 (0= Absența bolilor cardiace, 1= Prezența inimii boala)	Binar

Tabelul prezinta datele specifice problemelor de sanatate cu care pacienții au fost diagnosticați și unitatea de masura a acestora prezente în setul de date de mai sus.Fiecare analiza a fost realizata prin limbajul de programare Python, după ce au fost importate datele.

Pentru a începe analiza, am citit inițial datele din fișierul "Cardiovascular_Disease.csv". Aceste date au fost apoi împărțite în două subseturi distincte: un set de date pentru antrenare și unul pentru testare. Această separare este esențială pentru a evalua performanța modelelor de învățare automată.

Un aspect important al acestei etape a fost utilizarea testului Fisher pentru a evalua semnificația fiecărui predictor în discriminarea bolilor cardiovasculare. Acest test compară variația dintre grupurile de pacienți cu și fără boli cardiovasculare și generează rezultate importante despre importanța fiecărui predictor în ceea ce privește identificarea bolilor. Rezultatele testului Fisher au fost stocate în fișierul "Predictori.csv", permițându-ne să identificăm predictorii critici. Rezultatele testului Fisher indică faptul că nu toți predictorii sunt la fel de semnificativi în distingerea între prezența și absența bolilor cardiovasculare. De exemplu, vârsta (age) și genul (gender) au avut o putere de discriminare scăzută și valori p ridicate, sugerând că acestea nu sunt indicatori semnificativi în acest set de date pentru bolile cardiovasculare.

$$F_j = \frac{MSB_{jj}}{MSW_{jj}}.$$

Figura 1 - Statisca testului F

Pe de altă parte, variabile precum tipul durerii în piept (chestpain), tensiunea arterială în repaus (restingBP), colesterolul seric (serumcholestrol), zahărul din sânge pe nemâncate (fastingbloodsugar), rezultatele electrocardiogramei în repaus (restingrelectro), rata maximă a inimii (maxheartrate), panta segmentului ST al exercițiului ECG (slope) și numărul de vase majore vizibile la fluoroscopie (noofmajorvessels) au demonstrat o putere de discriminare însemnată cu valori p extrem de mici, semnalând o semnificație statistică robustă. Aceasta înseamnă că acești predictorii au un rol puternic în modelarea riscului de boli cardiovasculare și pot fi considerați variabile-cheie în predicția acestora.

Analiza Modelului LDA

În cadrul analizei noastre a datelor cardiovasculare, un pas crucial a fost aplicarea Analizei Discriminante Liniare (LDA) pentru a extrage variabilele discriminatorii — componente cheie care ne ajută să clasificăm eficient cazurile studiate. Acest proces transformă complexitatea multidimensională a datelor originale într-un număr redus de dimensiuni, numite discriminatori sau axe discriminante, care oferă cea mai mare separare între grupurile predefinite ale setului de date. Prin aplicarea LDA pe setul nostru de date, am identificat un discriminator dominant, denumit z1. Acesta reprezintă o nouă variabilă sintetică, calculată ca o combinație liniară a variabilelor originale, care captează esența variației între clasele de pacienți în studiul nostru — în acest caz, cei cu și fără boli cardiovasculare.

Puterea de discriminare a z1 este extraordinar de înaltă, cu o valoare de 1868.08, sugerând că acest discriminator este extrem de eficient în a diferenția între stările de sănătate cardiacă ale subiecților. Mai mult, p-value-ul asociat este mai mic decât pragul standard de semnificație de 0.05, fiind de 1.11e-16 — practic zero. Acest lucru indică o semnificație statistică extremă și ne asigură că diferențele pe care le observăm nu sunt datorate fluctuațiilor aleatorii, ci reflectă diferențe reale și convingătoare între grupele de pacienți.

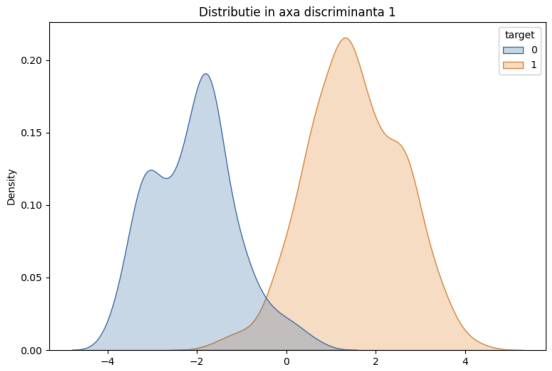

Figura 2 - Distribuție în axa discriminanta

Validitatea Modelului LDA

Una dintre cele mai importante etape ale analizei a fost construirea unui model de învățare automată folosind algoritmul Linear Discriminant Analysis (LDA). Acest model a fost instruit pe datele de antrenare, învățând să facă predicții precise pe baza predictorilor și a informațiilor despre prezența sau absența bolilor cardiovasculare.

Modelul LDA instruit a fost evaluat pe datele de testare pentru a măsura performanța sa. Acuratețea modelului a fost calculată pentru a determina cât de precis poate face predicții. De asemenea, o matrice de confuzie a fost creată pentru a evalua modul în care modelul a clasificat corect și incorect pacienții în funcție de starea lor de sănătate.

Rezultatele acestor măsurători au fost stocate în fișiere separate: "Acuratete_lda.csv" și "MatriceaDeConfuzie_lda.csv".

Datele din "Acuratete_lda.csv" și "MatriceaDeConfuzie_lda.csv" furnizează informații esențiale despre performanța modelului Linear Discriminant Analysis (LDA) în predicția bolilor cardiovasculare. Iată concluziile pe care le putem trage pe baza acestor date:

1. **Acuratețe globală:** Modelul LDA prezintă o acuratețe globală de 96.0%. Acest lucru înseamnă că modelul a clasificat corect aproximativ 96% din cazurile din setul de date de testare.

2. **Acuratețe medie:** Acuratețea medie este de 95.82%. Aceasta reprezintă media acurateții pentru clasificarea pacienților cu și fără boli cardiovasculare. Rezultatul de 95.82% arată că modelul are o performanță bună în identificarea ambelor categorii de pacienți.
3. **Index Cohen-Kappa:** Indexul Cohen-Kappa este de 0.9165. Acest indice măsoară acordul între predicțiile modelului și realitatea observată, corectându-se pentru acordul aleator. Valoarea mare a acestui indice (aproape de 1) indică o concordanță foarte bună între predicții și realitatea observată.
4. **Matricea de confuzie:** Matricea de confuzie arată numărul de cazuri corect și incorect clasificate pentru fiecare categorie (0 și 1). Avem următoarele concluzii:
    a. Pentru clasa 0 (pacienți fără boli cardiovasculare), 94.97% dintre cazuri au fost clasificate corect, iar 5.03% au fost clasificate greșit ca având boala cardiovasculară.
    b. Pentru clasa 1 (pacienți cu boli cardiovasculare), 96.68% dintre cazuri au fost clasificate corect, iar 3.32% au fost clasificate greșit ca fiind sănătoși.

Aceste rezultate indică faptul că modelul LDA are o performanță excelentă în identificarea bolilor cardiovasculare. Cu o acuratețe globală de 96.0%, un indice Cohen-Kappa ridicat și o matrice de confuzie care arată o proporție foarte mare de clasificări corecte, putem concluziona că modelul LDA este un instrument puternic pentru diagnosticarea bolilor cardiovasculare pe baza predictorilor din setul de date.

Discriminarea Bayesiană

$$P(A_k / A) = \frac{P(A_k)P(A / A_k)}{\sum_{i=1}^{n} P(A_i)P(A / A_i)}$$

Figura 3 - Relația lui Bayes

1. **Acuratețea Globală**: Modelul Naive Bayes a atins o acuratețe globală de 94.25%, ceea ce indică un nivel înalt de performanță generală. Acest procent ne spune că, din totalul predicțiilor făcute de model pe setul de testare, 94.25% au fost corecte.
2. **Acuratețea Medie**: Acuratețea medie, calculată ca media aritmetică a acurateților pentru fiecare clasă, este de 93.86%. Această valoare reflectă capacitatea modelului de a clasifica corect atât pacienții sănătoși, cât și pe cei cu boli cardiovasculare, arătând că modelul oferă o performanță echilibrată pentru ambele clase.
3. **Indexul Cohen-Kappa**: Cu un indice Cohen-Kappa de 0.881, modelul demonstrează o concordanță foarte bună între predicțiile sale și realitatea observată, depășind acuratețea ce ar putea fi atribuită șansei. Acest indice apropiat de 1 sugerează că modelul este foarte fiabil în predicțiile sale.

Analiza Matricei de Confuzie

- Clasa 0 (Pacienți Fără Boli Cardiovasculare): Din totalul pacienților fără boli cardiovasculare, modelul a identificat corect 91.57% dintre ei ca fiind sănătoși. Acest lucru indică o sensibilitate bună a modelului în a recunoaște pacienții fără condiții de sănătate afectate.
- Clasa 1 (Pacienți Cu Boli Cardiovasculare): În cazul pacienților cu boli cardiovasculare, modelul a avut o acuratețe și mai mare, reușind să clasifice corect 96.15% dintre cazuri. Aceasta evidențiază o specificitate ridicată și o abilitate deosebită a modelului de a detecta prezența bolilor cardiovasculare.

Cuprins

Proiectul nostru a abordat provocarea complexă a detectării și diagnosticării precoce a bolilor cardiovasculare folosind metode de învățare automată. Printr-o analiză detaliată a unui set de date extensiv, care a inclus 14 variabile clinice esențiale colectate de la 1000 de subiecți, am demonstrat că modelele de învățare automată pot juca un rol semnificativ în îmbunătățirea diagnosticării bolilor de inimă.Am aplicat metoda LDA pentru a extrage caracteristici relevante și pentru a construi un model predictiv cu o acuratețe globală de 96.0%, o acuratețe medie de 95.82%, și un indice Cohen-Kappa de 0.9165, indicând o concordanță excepțională între predicțiile modelului și realitatea clinică. Matricea de confuzie a confirmat capacitatea înaltă a modelului de a distinge între pacienții sănătoși și cei cu boli cardiovasculare, o descoperire care subliniază utilitatea modelului LDA în contextul medical.

În comparație, modelul Naive Bayes a prezentat, de asemenea, rezultate admirabile, cu o acuratețe globală de 94.25%, o acuratețe medie de 93.86%, și un indice Cohen-Kappa de 0.881. Deși puțin mai scăzută decât LDA, performanța modelului Naive Bayes rămâne impresionantă, mai ales în clasificarea corectă a pacienților cu boli cardiovasculare, unde a atins o acuratețe de 96.15%.Aceste rezultate ne conduc la concluzia că, în timp ce LDA a avut o performanță generală ușor superioară, ambele modele sunt instrumente valoroase pentru diagnosticarea bolilor cardiovasculare, fiecare cu punctele sale forte.