**Predicția Prezenței Afecțiunii Cardiace**

Șevcenco S. IA-212, Universitatea Tehnică a Moldovei, Facultatea Calculatoare, Informatică și Microeletronică

# Abstract

Acest studiu examinează creșterea îngrijorătoare a incidenței bolilor cardiace, o cauză principală a morbidității la nivel global, cu o atenție deosebită acordată tinerilor adulți. Folosind un set de date cu caracteristici clinice diverse, cercetarea noastră aplică prelucrarea complexă a datelor, analiza exploratorie și modele de regresie logistică pentru a identifica factorii determinanți în apariția bolilor cardiace și pentru a dezvolta un model predictiv. Rezultatele descoperă perspective noi asupra importanței variabilelor clinice, punând accentul pe necesitatea unei evaluări detaliate a riscului.

Analiza Importanței Variabilelor în Proiecție evidențiază semnificația tipului durerii în piept (cp), numărului de vase mari colorate la fluoroscopie (ca) și a rezultatelor testului thallium (thal) în predicția bolii cardiace. Analizele comparative confirmă performanța echilibrată a modelului de regresie logistică cu o acuratețe de 81.37%, o sensibilitate de 76.77% și o specificitate de 85.71%, indicând eficiența acestuia în context clinic.

# Introducere

În prezent, cercetarea medicală avansează semnificativ prin utilizarea tehnicilor de analiză a datelor. Potrivit Organizației Mondiale a Sănătății (OMS), bolile de inimă sunt una dintre principalele cauze de deces la nivel global, afectând profund sănătatea oamenilor. Aceste boli cardiovasculare nu numai că provoacă o mortalitate semnificativă, dar adesea lasă sechele pe termen lung, punând o povară semnificativă asupra familiilor și comunităților. Deși este cunoscută pentru prevalența sa la adulții în vârstă, incidența bolilor de inimă la adulții mai tineri a început să crească, indicând o nevoie urgentă de cercetări suplimentare în acest domeniu.

Bolile de inimă, care includ o serie de afecțiuni care afectează inima, pot avea consecințe grave asupra sănătății. Acestea apar atunci când funcția inimii este perturbată, ceea ce poate duce la reducerea fluxului sanguin și, prin urmare, la deteriorarea țesutului cardiac. Cauzele pot varia de la artere înfundate la hipertensiune arterială până la aritmii cardiace.

În studiul nostru, ne-am concentrat pe analiza unui set de date format din 14 caracteristici medicale cu scopul de a identifica și clarifica principalii factori care contribuie la bolile de inimă. Scopul principal este de a dezvolta și antrena un nou model de predicție care poate evalua în mod eficient riscul de boli de inimă. Cercetarea noastră nu a inclus doar o evaluare cuprinzătoare a datelor colectate, ci și crearea și optimizarea unui model de predicție pentru a îmbunătăți predicția riscului de boli de inimă. Acest studiu nu numai că ne îmbogățește înțelegerea factorilor de risc pentru bolile de inimă, dar oferă și o bază solidă pentru viitoarele eforturi de modelare predictivă care pot ghida strategiile de intervenție.Intervenție timpurie și prevenire pentru sănătatea cardiovasculară.

**Materiale și Metode**

Pentru analiza respectivă, s-a accesat un set de date open-source de pe Kaggle;

Sursa: <https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset>

Setul de date inițial furnizat includea 14 variabile, concentrându-se pe caracteristici clinice esențiale pentru studiul bolilor de inimă. Aceste caracteristici cuprindeau o gamă largă de informații relevante, cum ar fi vârsta, sexul, nivelul colesterolului, tensiunea arterială și alți indicatori cheie, oferind o bază solidă pentru analiza aprofundată a factorilor de risc asociați cu bolile cardiace.

**Analiza exploratorie a datelor (EDA)**

Procesul de Analiză Exploratorie a Datelor (EDA) a implicat folosirea instrumentelor de vizualizare precum R și ggplot2 pentru a analiza setul de date furnizat, care se concentrează pe prezența bolilor cardiace. Analiza a inclus crearea de grafice de tip bară și boxplot pentru a examina distribuția caracteristicilor clinice precum vârsta, sexul, tipul de durere toracică (cp), tensiunea arterială la repaus (trestbps), colesterolul (chol), glicemia a jeun (fbs), rezultatele electrocardiografice în repaus (restecg), frecvența cardiacă maximă atinsă (thalach), angina indusă de exercițiu (exang), depresia ST indusă de exercițiu (oldpeak), panta segmentului ST la exercițiul maxim (slope), numărul de vase majore colorate prin fluoroscopie (ca), și thalasemia (thal). Graficele de tip bară au oferit înțelegere asupra modului în care aceste variabile influențează riscul de boli cardiace. Boxplot-urile au fost utilizate pentru a analiza distribuția și a identifica posibilele asocieri între aceste variabile clinice și ținta studiului, care este prezența bolii cardiace.

**Presupuneri**

Oare există o corelație pozitivă între vârsta unei persoane (age) și riscul de a dezvolta boli cardiace?

Este sexul (sex) un factor determinant în riscul de a dezvolta boli cardiace, cu bărbații fiind mai predispuși decât femeile?

Cum sunt diferite tipurile de dureri toracice (cp) asociate cu riscul de boli cardiace?

Există o relație între nivelurile ridicate de tensiune arterială la repaus (trestbps) și colesterol (chol) și un risc crescut de boli cardiace?

Glicemia în repaus (fbs) peste un anumit prag este un indicator al riscului de boli cardiace?

Există o corelație inversă între frecvența cardiacă maximă atinsă (thalach) și riscul de boli cardiace?

Prezența anginei induse de exercițiu (exang) este un indicator puternic al bolilor cardiace?

Cum sunt asociate variabilele legate de teste cardiace, cum ar fi depresia ST indusă de exercițiu (oldpeak), panta segmentului ST de exercițiu vârf (slope), numărul de vase mari colorate prin fluoroscopie (ca), și defectele de stres (thal), cu prezența bolilor cardiace?

**Regresia logistică**

Un model de regresie logistică a fost dezvoltat pentru a prezice bolile cardiace, utilizând metoda glm cu familia binomială în limbajul R. Modelul a inclus diverse variabile predictor, precum genul, vârsta, colesterolul, tensiunea arterială la repaus și glicemia în repaus. Normalizarea datelor a precedat împărțirea setului în subseturi de antrenament și testare, cu scopul de a asigura o evaluare riguroasă și de a evita potențialul de părtinire în model. Procesul de antrenare nu a implicat validare încrucișată standardizată, ci s-a bazat pe o singură împărțire a setului de date.

**Evaluarea modelelor**

Pentru fiecare model generat, s-a calculat Criteriul Informațional Akaike (AIC) pentru a evalua calitatea ajustării modelului, cu un număr mai mic de AIC indicând o potrivire mai bună a modelului. În final, modelele au fost evaluate prin construirea curbelor ROC și calculul matricelor de confuzie, oferind o perspectivă asupra eficienței fiecărui model în predicția bolii cardiace. Modelul rezultat este gata să ofere înțelegere asupra relațiilor dintre variabilele specificate și probabilitatea de apariție a bolii cardiace.

**Resultate**

1. **Distribuția bolii cardiace în funcție de sex**

Investigarea modului în care se distribuie datele este esențială în numeroase sectoare și domenii de studiu, deoarece furnizează informații esențiale despre structurile, evoluțiile și particularitățile inerente ale datelor. Înainte de toate, să analizăm distribuția bărbaților și femeilor în rândul pacienților afectați și neafectați de boli cardiace:

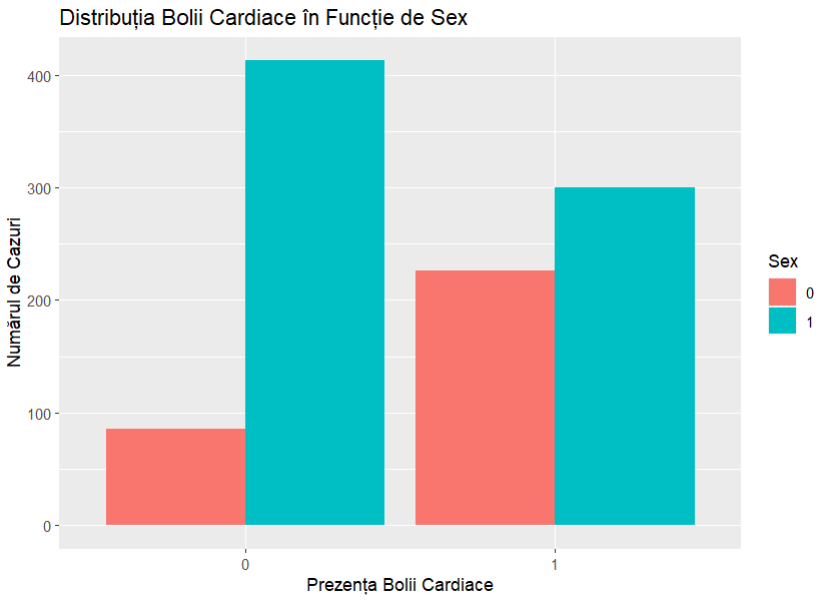
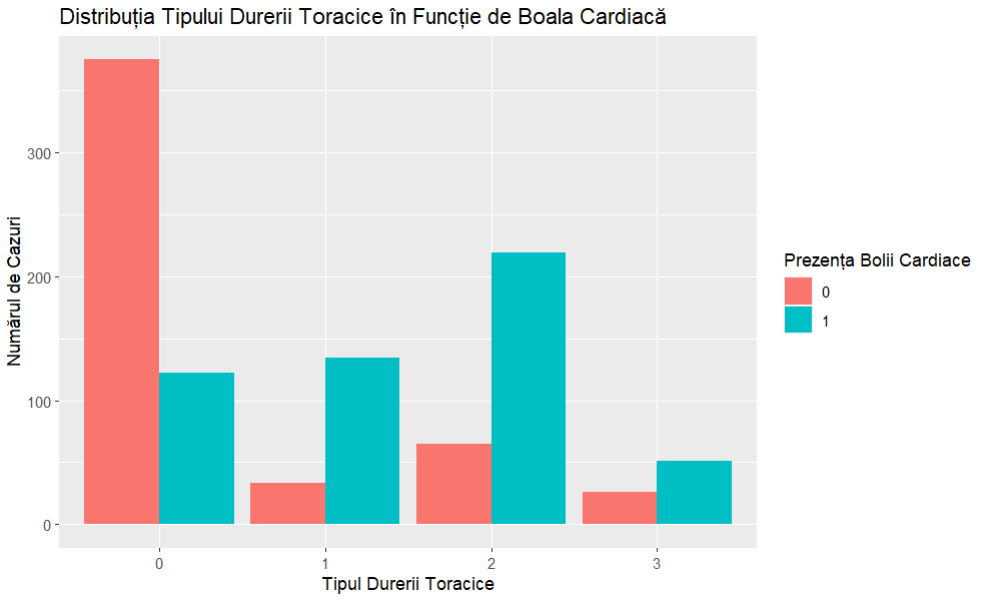


Diagrama ilustrează distribuția bolilor cardiace în funcție de sex, unde sexul este codificat ca 0 și 1. Observăm că există o diferență notabilă între numărul de cazuri de boli cardiace în funcție de sex pentru pacienții care prezintă boala cardiacă (codificat cu 1) comparativ cu cei fără boală cardiacă (codificat cu 0). În particular, sexul codificat cu 1 are o incidență mai mare a bolilor cardiace decât sexul codificat cu 0, indiferent de prezența bolii.

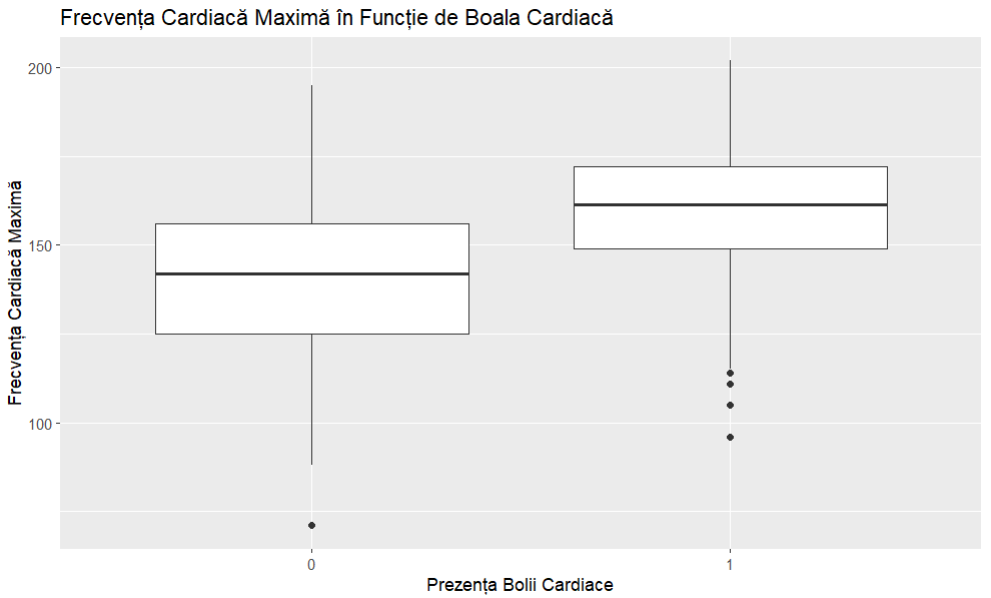
1. **Distribuția Tipului Durerii Toracice în Funcție de Boala Cardiacă**

Diagrama prezentată analizează relația dintre diferite tipuri de durere toracică și prezența bolii cardiace. Se observă că tipul 0 de durere toracică este cel mai frecvent întâlnit printre pacienții fără boală cardiacă, sugerând o corelație slabă cu afecțiunile cardiace. Pe de altă parte, tipul 2 se remarcă prin asocierea sa puternică cu bolile cardiace, având cel mai mare număr de cazuri confirmate. Tipul 1 arată o distribuție mai egală între cele două stări, indicând o posibilă legătură moderată. În sfârșit, tipul 3 este cel mai puțin întâlnit în ambele categorii, ceea ce ar putea sugera o incidență general mai scăzută a acestui tip de durere sau o relevanță limitată în diagnosticarea bolilor cardiace. Aceste observații subliniază importanța tipului durerii toracice ca factor potențial predictiv în evaluarea riscului de boală cardiacă.

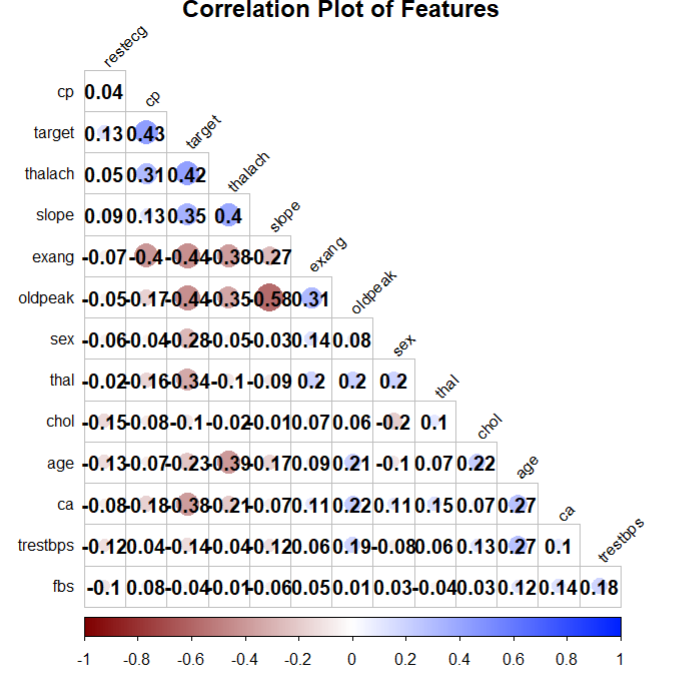


1. **Distribuția Anginei Induse de Exercițiu în Funcție de Boala Cardiacă**

Boxplot-ul evidențiază o frecvență cardiacă maximă general mai redusă pentru pacienții diagnosticați cu boli cardiace comparativ cu cei fără, sugerând o posibilă corelație între o frecvență scăzută și afecțiunile inimii. Variabilitatea mai mare în grupul cu boli cardiace ar putea indica diverse grade de severitate a afecțiunilor.



**Diagramă de corelație a caracteristicilor**



Analiza coeficienților de corelație din setul de date axat pe predicția bolilor cardiace oferă o perspectivă asupra relațiilor dintre diferite variabile și riscul de boală cardiacă. Se observă că vârsta are o corelație pozitivă moderată cu prezența bolii cardiace, ceea ce sugerează că probabilitatea de a dezvolta o boală cardiacă crește pe măsură ce pacienții îmbătrânesc. Variabile precum tensiunea arterială la repaus și nivelul colesterolului au arătat, de asemenea, corelații pozitive, indicând o asociere cu riscul crescut de boli cardiace.

În cadrul analizei exploratorii a datelor, s-a constatat o corelație puternică între vârstă și prezența bolii cardiace, iar variabilele continue precum tensiunea arterială și colesterolul au fost identificate ca având distribuții care le pot face predictori relevanți ai bolii. Frecvența cardiacă maximă și tipul durerii toracice s-au evidențiat de asemenea ca având legături semnificative cu starea cardiacă a pacienților.

**Constatările VIP**

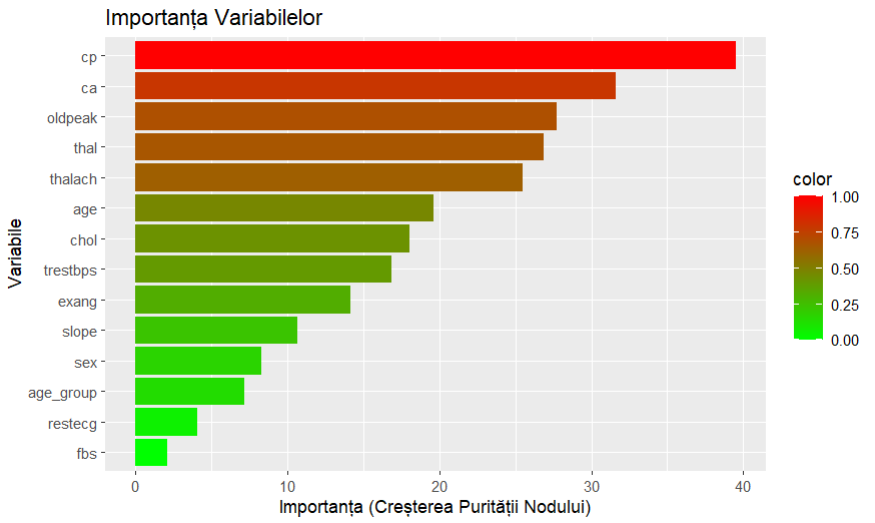
****

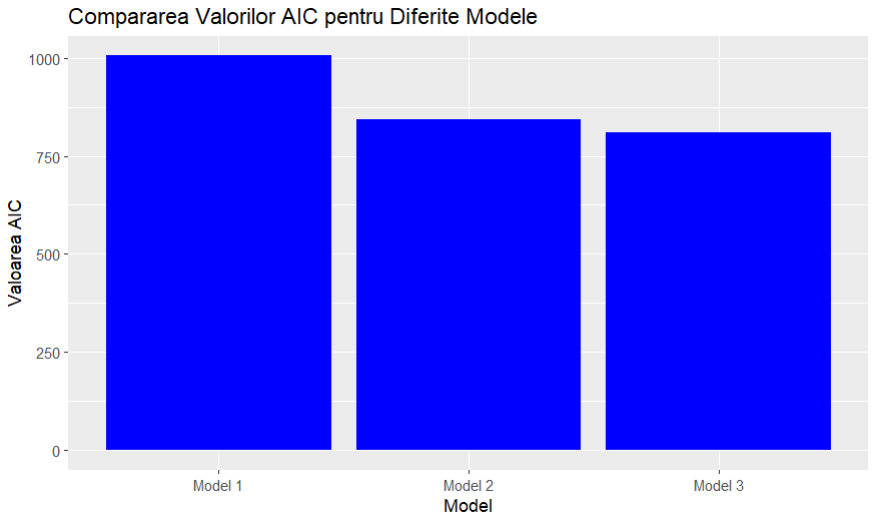
Diagrama privind importanța variabilelor în contextul unui model predictiv pentru bolile cardiace pune în evidență contribuțiile diferitelor caracteristici în predicția riscului de boală. Este remarcabilă proeminența tipului durerii toracice (cp), care domină graficul cu o importanță semnificativă, indicând o influență puternică în model. Alte variabile, cum ar fi numărul de vase majore colorate prin fluoroscopie (ca) și depresia segmentului ST (oldpeak), urmează în ordinea importanței, sugerând că acestea dețin și ele un rol cheie în determinarea riscului.

Pe măsură ce ne aprofundăm analiza, observăm că frecvența cardiacă maximă (thalach) și vârsta sunt, de asemenea, factori cu greutate, subliniind legătura dintre capacitatea cardiacă fizică și vârsta avansată cu probabilitatea crescută de boli cardiace. Chiar dacă nivelul de colesterol (chol) și tensiunea arterială la repaus (trestbps) prezintă o importanță moderată, acestea nu trebuie neglijate în evaluarea riscului cardiac.

Aceste descoperiri ne direcționează către o înțelegere mai detaliată a complexității bolilor cardiace și subliniază importanța unei abordări multifactoare în modelarea riscului, care să ia în considerare nu doar vârsta sau factorii fiziologici, dar și aspecte precum caracteristicile clinice specifice ale inimii și impactul lor cumulat asupra sănătății cardiace.

**Compararea Valorilor AIC pentru Diferite Modele**

Diagrama ilustrează o evaluare a trei modele diferite prin intermediul Criteriului Informațional Akaike (AIC), o metrică ce echilibrează bunătatea ajustării modelului cu complexitatea sa. Modelul 1 arată cea mai mare valoare AIC, sugerând o ajustare mai puțin optimă comparativ cu celelalte două modele. Atât Modelul 2 cât și Modelul 3 prezintă valori AIC mai mici, indicând o ajustare mai bună la date, cu Modelul 2 având un ușor avantaj. Astfel, în contextul selecției modelului pentru predicția bolilor cardiace, Modelul 1 pare să fie mai puțin adecvat, în timp ce Modelul 2 și Modelul 3 ar putea oferi o acuratețe predictivă mai mare și o interpretare mai fiabilă.

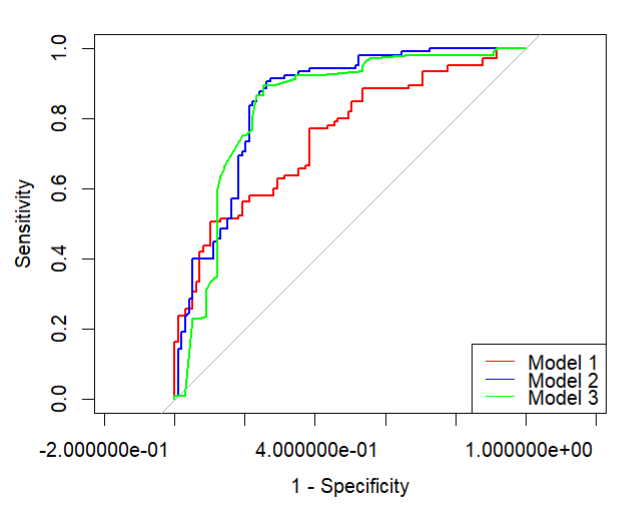


**Curba ROC**

În cadrul evaluării comparative a modelelor destinate predicției bolilor cardiace, curbele Receiver Operating Characteristic (ROC) ofereau o perspectivă profundă. În cercetarea lor, autorii studiului construiau și comparau trei modele predictive distincte, cu scopul de a determina care dintre acestea demonstra cea mai mare precizie în clasificarea pacienților în categorii de risc pentru boli cardiace.

Reprezentarea grafică a curbelor ROC descoperea o diferență notabilă între modele în ceea ce privește capacitatea lor de a face distincția între clasificările pozitive și negative. Modelul 1, indicat de linia albastră, arăta o performanță superioară, apropiindu-se considerabil de idealul reprezentat de colțul din stânga sus al graficului. Acest lucru indica faptul că Modelul 1 realiza un echilibru optim între sensibilitate și specificitate, fiind esențial pentru un diagnostic precis.

În contrast, curbele pentru Modelele 2 și 3, deși treceau peste nivelul de performanță aleatorie indicat de linia diagonală, prezentau o acuratețe inferioară în comparație cu Modelul 1. Această variație în eficacitatea modelelor sublinia importanța unei selecții meticuloase a caracteristicilor și a metodologiei de modelare pentru a îmbunătăți rezultatele predictive.

****

**Rezultatele matricei de confuzie**

În cadrul analizei noastre, am examinat performanța a trei modele statistice pentru clasificarea datelor. Mai întâi, Modelul 1 a obținut o acuratețe de 65.69%, cu o sensibilitate (rata de detecție corectă pentru clasa 1) de 71.72% și o specificitate (rata de detecție corectă pentru clasa 0) de 60.00%. Statistica Kappa, care măsoară acordul peste așteptările întâmplării, a fost de 0.3159, indicând un nivel moderat de acord. De asemenea, Modelul 1 a prezentat o precizie pozitivă (PPV) de 62.83% și o precizie negativă (NPV) de 69.23%. Acuratețea echilibrată a modelului, calculată la 65.86%, sugerează un echilibru moderat între sensibilitate și specificitate.

În ceea ce privește Modelul 2, acesta a obținut o acuratețe de 81.37%, cu o sensibilitate de 76.77% și o specificitate de 85.71%. Statistica Kappa a fost de 0.6263, indicând un nivel semnificativ de acord dincolo de întâmplare. De asemenea, Modelul 2 a înregistrat o precizie pozitivă (PPV) de 83.52% și o precizie negativă (NPV) de 79.65%. Acuratețea echilibrată a modelului, înregistrată la 81.24%, sugerează un echilibru excelent între sensibilitate și specificitate.

În fine, Modelul 3 a prezentat o acuratețe de 78.92%, cu o sensibilitate de 77.78% și o specificitate de 80.00%. Statistica Kappa pentru Modelul 3 a fost de 0.5779, indicând un nivel moderat de acord. De asemenea, Modelul 3 a înregistrat o precizie pozitivă (PPV) de 78.57% și o precizie negativă (NPV) de 79.25%. Acuratețea echilibrată a modelului, calculată la 78.89%, sugerează un echilibru corect între sensibilitate și specificitate.

Modelul 2 s-a evidențiat ca având cea mai bună performanță în majoritatea metricilor, în timp ce Modelul 3 a prezentat, de asemenea, rezultate solide. Modelul 1, în schimb, a înregistrat o performanță relativ mai scăzută în comparație cu celelalte două modele. Aceste constatări reflectă abilitatea fiecărui model de a clasifica corect instanțele în sarcina dată.

**Discuții**

Studiul de față reprezintă o contribuție valoroasă în domeniul cardiologiei, oferind o bază pentru îmbunătățirea continuă a metodelor de predicție a bolilor cardiace. Prin analiza datelor și interpretarea rezultatelor vizuale, am identificat variabile cu impact semnificativ în predicția afecțiunilor cardiace. Interesant este că, deși Modelul 1 indică o valoare AIC mai mare, sugerând o ajustare mai puțin optimă a datelor, el rămâne fundamental în înțelegerea dinamicii bolilor cardiace.

Conform graficului de importanță a variabilelor, tipul durerii în piept (cp), numărul de vase mari vizibile la fluoroscopie (ca), și rezultatele testului thallium (thal) se disting ca fiind predictorii principali în modelul nostru. Acești indicatori, alături de depresia ST indusă de exercițiu (oldpeak) și ritmul cardiac maxim atins (thalach), subliniază zonele cheie pentru evaluarea riscului cardiac.

Curbele ROC pentru toate cele trei modele dezvăluie variații în capacitatea lor de a clasifica corect cazurile, cu Modelul 2 afișând o performanță deosebită, apropiindu-se de idealul unei rate maxime de adevărat pozitiv și minime de fals pozitiv. Acest lucru ne sugerează că Modelul 2 ar putea fi mai eficient în practica clinică pentru identificarea pacienților cu risc de boli cardiace.

Analiza distribuției bolii cardiace în funcție de sex a relevat de asemenea diferențe semnificative, punând în lumină potențialul de gen ca factor de risc. În plus, corelația observată între cp și target întărește ipoteza că simptomatologia durerii în piept este un marker important pentru prezența bolii cardiace.

Prin integrarea acestor descoperiri, putem afirma că modelarea predicției bolilor cardiace beneficiază de pe urma unui model complex care încorporează o varietate de variabile clinice. Totodată, recunoaștem necesitatea îmbunătățirii continue a acestor modele prin extinderea și diversificarea seturilor de date, ceea ce ne-ar permite o înțelegere mai aprofundată a interacțiunilor complexe dintre caracteristicile clinice și riscul de boală cardiacă.

**Bibliografie**

1. Heart Disease Dataset **-** <https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset>

[accesat pe 20.11.2023]

2.Heart Diseases **-** <https://medlineplus.gov/heartdiseases.html> [accesat pe 19.10.2023]

3.Heart Disease and Stroke **-**

<https://www.cdc.gov/chronicdisease/resources/publications/factsheets/heart-disease-stroke.htm#:~:text=Leading%20risk%20factors%20for%20heart,unhealthy%20diet%2C%20and%20physical%20inactivity> [accesat pe 11.11.2023]

**Anexă**

|  |  |  |
| --- | --- | --- |
| Variabile | Tip var. | Descrierea |
| age | Numeric | Vârsta individului. |
| sex | Numeric | Sexul individului (1: Masculin, 0: Feminin). |
| cp | Numeric | Tipul durerii toracice (Valori 0: angină tipică; 1: angină atipică; 2: durere non-anginală; 3: asimptomatic). |
| trestbps | Numeric | Tensiunea arterială în repaus (în mm Hg la internare în spital). |
| chol | Numeric | Colesterolul seric în mg/dl. |
| fbs | Numeric | Glicemia a jeun > 120 mg/dl (1: Da; 0: Nu). |
| restecg | Numeric | Rezultatele electrocardiografice în repaus (Valori 0,1,2). |
| thalach | Numeric | Frecvența cardiacă maximă atinsă. |
| exang | Numeric | Angina indusă de exercițiu (1: Da; 0: Nu). |
| oldpeak | Numeric | Depresia ST indusă de exercițiu relativ la repaus. |
| slope | Numeric | Panta segmentului ST la exercițiul maxim (Valori 0: ascendentă; 1: plată; 2: descendentă). |
| ca | Numeric | Numărul de vase majore (0-3) colorate prin fluoroscopie. |
| thal | Numeric | Talasemia (3 = normal; 6 = defect fix; 7 = defect reversibil). |
| target | Numeric | Prezența bolii de inimă (1: Da; 0: Nu). |