Predicția riscului de boli de inimă folosind date clinice și caracteristici fiziologice

Titica Iana, MI-212

Universitatea Tehnică a Moldovei, Facultatea Calculatoare, Informatică și Microelectronică, Departamentul Informatică și Ingineria Sistemelor

ABSTRACT

În ciuda progreselor în domeniul sănătății, detectarea timpurie și prevenirea BCV rămân provocări majore, conducând adesea la diagnostic târziu și la pierderea oportunităților de intervenție. Această cercetare introduce un model predictiv inovator destinat să evalueze mai precis riscul bolilor de inimă. Modelul integrează date comprehensive ale pacienților, incluzând parametri demografici, clinici și fiziologici. Scopul principal este îmbunătățirea detectării timpurii și facilitarea strategiilor preventive personalizate. Prin identificarea și analiza diverselor factori de risc, modelul își propune să contribuie semnificativ la reducerea morbidității și mortalității asociate BCV. Această abordare predictivă reprezintă un testament al potențialului soluțiilor bazate pe date în transformarea asistenței medicale, în special în gestionarea și prevenirea bolilor cardiovasculare.

INTRODUCERE

În contemporana lume medicală, bolile cardiovasculare constituie o provocare majoră și un domeniu esențial de cercetare, având un impact semnificativ asupra sănătății globale. Aceste afecțiuni persistă în a fi printre cele mai importante cauze de morbiditate și mortalitate la nivel mondial, chiar în contextul progreselor tehnologice și medicale remarcabile. Cu toate acestea, mulți pacienți continuă să fie diagnosticați târziu sau să nu beneficieze de intervenții preventive adecvate. Statisticile actuale subliniază gravitatea situației, arătând că bolile cardiovasculare reprezintă principalul motiv de îngrijorare în sănătate, fiind responsabile pentru aproximativ 17,9 milioane de decese anual, ceea ce reprezintă 31% din totalul deceselor la nivel global. Abordarea acestor probleme necesită o strategie comprehensivă și eficientă.

În fața acestor provocări, dezvoltarea unui model de predicție pentru riscul de boli de inimă devine o inițiativă deosebit de importantă. Acest model, construit pe baza unui set detaliat de date despre pacienți, își propune să ofere o estimare precisă a riscului, luând în considerare variabile precum vârsta, sexul, tipul de dureri toracice, tensiunea arterială, colesterolul, glicemia,

electrocardiograma și altele. Scopul proiectului este să creeze un model solid, capabil să identifice factorii de risc relevanți și să ofere strategii personalizate de intervenție și prevenție.

Obiectivele includ identificarea și pregătirea datelor, crearea unui model de predicție robust și evaluarea performanței acestuia. Cu toate acestea, proiectul nu este lipsit de provocări, precum riscul de overfitting în cazul unui set de date limitat sau necesitatea asigurării calității datelor, evitând confuziile și variabilele confundante care pot afecta corectitudinea relațiilor dintre factorii de risc și rezultate. Prin anticiparea și personalizarea intervențiilor preventive, acest proiect urmărește să contribuie la reducerea semnificativă a mortalității asociate bolilor cardiovasculare. În acest efort, progresul tehnologic devine un aliat esențial în combaterea acestor afecțiuni devastatoare, având potențialul de a aduce beneficii semnificative sănătății globale.

PRESUPUNERI INIȚIALE

1. Persoanele de sex masculin prezintă o predispoziție mai mare către insuficiența cardiacă în comparație cu femeile.
2. Durerea toracică asimptomatică (ASY) se dovedește a fi un indicator mai relevant pentru probleme cardiace decât alte tipuri de dureri toracice.
3. Nivelurile mai ridicate ale glicemiei în post, înregistrate peste 120 mg/dl, sunt asociate mai semnificativ cu riscul de insuficiență cardiacă.
4. Prezența anginei induse de exerciții fizice reprezintă un predictor mai puternic al insuficienței cardiace decât absența acestui simptom.
5. Tensiunea arterială în repaus în intervalul 95-170 mmHg poate avea semnificație în predicția riscului de insuficiență cardiacă.

INFORMAȚII DESPRE SETUL DE DATE

Setul de date furnizează o perspectivă cuprinzătoare asupra pacienților care au fost supuși analizelor cardiace, oferind informații detaliate privind factorii relevanți pentru sănătatea cardiovasculară. Printre acești factori se numără vârsta pacientului, sexul, tipul de durere în piept raportat (cum ar fi angina tipică, angina atipică, durerea non-anginală sau asimptomatică), precum și măsurători obiective, cum ar fi presiunea sanguină în starea de odihnă și nivelul de colesterol seric.

Setul de date include, de asemenea, rezultatele electrocardiogramei în stare de odihnă, furnizând informații despre starea electrică a inimii. Parametrii precum frecvența cardiacă maximă atinsă în timpul exercițiului și prezența anginei induse de efort oferă date semnificative privind răspunsul cardiovascular la stresul fizic. Măsurători specifice, cum ar fi depresia segmentului ST și

inclinația segmentului ST în vârf, completează imaginea clinică, furnizând detalii cruciale pentru evaluarea funcționalității cardiace.

De asemenea, variabila de ieșire indică dacă un pacient suferă sau nu de boală cardiacă, furnizând o bază esențială pentru analiza riscului cardiovascular. Acest set de date, bogat în informații, se pretează analizei exploratorii detaliate și dezvoltării de modele predictive în domeniul sănătății cardiovasculare, având potențialul de a contribui semnificativ la înțelegerea și gestionarea bolilor cardiace.

ANALIZA PRELIMINARĂ A DATELOR

Setul de date impresionează prin volumul său considerabil de 918 observații, fiecare distribuită pe 12 variabile relevante, capturând aspecte demografice și clinice cruciale pentru cercetarea afecțiunilor cardiace.

Structura setului de date a fost meticulos inspectată în prima fază, confirmând corectitudinea atribuirii tipurilor de date pentru fiecare variabilă. Această verificare a inclus o gamă diversă de tipuri de date, de la numere întregi la șiruri de caractere, pregătind astfel terenul pentru o analiză statistică riguroasă și pertinentă.

Pe parcursul etapei de identificare a duplicatelor, utilizând funcția duplicated(), s-a constatat că setul de date rămâne nealterat de prezența înregistrărilor repetate, consolidând astfel acuratețea și fiabilitatea sa pentru cercetările viitoare. Acest aspect a fost esențial pentru asigurarea integrității datelor și a eliminării potențialelor distorsiuni în rezultatele analizei.

În continuare, verificarea pentru absența valorilor, realizată prin funcția is.na(), a confirmat că setul de date este complet și lipsit de lacune informaționale. Această asigurare a constituat un pas crucial în stabilirea fundamentelor unei analize statistice solide și încrederii, eliminând potențialele interferențe cauzate de absența sau prezența insuficientă a datelor. Astfel, acest proces exhaustiv de pregătire a datelor a stabilit baza pentru o analiză detaliată și în profunzime a fenomenelor legate de sănătatea cardiacă în cadrul acestui set de date complex.

METODE STATISTICE

Intervalului Interquartilic (IQR)

În mediul de programare R, am implementat calculul Intervalului Interquartilic (IQR) pentru variabilele-cheie asociate acestui studiu, utilizând funcția IQR() nativă în R. Alegerea acestei metode de evaluare, rezistentă la valorile extreme, vine în întâmpinarea complexității informațiilor oferite de datele clinice și fiziologice. În urma acestui proces, am prezentat cu claritate rezultatele, furnizând o

perspectivă mai profundă asupra dispersiei și variabilității datelor în cadrul populației studiate în contextul temei noastre de cercetare.

Prin aplicarea acestei abordări în R, am reușit nu doar să obținem cu eficiență valorile IQR pentru variabilele-cheie, dar și să le interpretăm rapid, contribuind astfel la dezvoltarea unei înțelegeri mai cuprinzătoare a factorilor de risc cardiovascular și a modului în care aceștia se reflectă în datele clinice și caracteristicile fiziologice. Această metodologie se aliniază obiectivelor temei noastre de cercetare, oferind un cadru robust pentru analiza și interpretarea datelor în vederea predicției riscului de boli de inimă.

Analiza descriptivă

Analiza statistică descriptivă relevă o serie de tendințe și caracteristici demografice și medicale ale cohortelor studiate. Am adoptat o abordare descriptivă pentru a ilustra caracteristicile cheie ale setului de date, cu accent pe variabilele numerice și categorice relevante pentru predicția riscului de boli de inimă. Am utilizat funcția summary() pentru a obține statistici descriptive, oferind o privire sumară asupra distribuției și tendințelor centrale ale vârstei, presiunii arteriale în repaus, colesterolului, ritmului cardiac maxim și modificărilor segmentului ST. Pentru variabilele categorice, am generat tabele de frecvență folosind funcția table(). Aceste tabele oferă o perspectivă detaliată asupra distribuției genului, tipului de durere toracică, prezenței glicemiei post-absorptive crescute, stării ECG în repaus, exercițiului indus de angină, înclinării segmentului ST și a prezenței bolii de inimă.

ANALIZA EXPLORATORIE A DATELOR

În domeniul cardiologiei, EDA joacă un rol crucial în identificarea factorilor de risc, înțelegerea dinamicii bolilor de inimă și în formularea de ipoteze pentru cercetări ulterioare. Utilizarea pachetului ggplot2 în R oferă unelte puternice pentru vizualizarea și analiza datelor, permițând explorarea relațiilor dintre diverse variabile și detectarea modelor și tendințelor. Astfel, am implementat o serie de grafice pentru a explora diferite aspecte ale datelor, de la distribuții de vârstă și sexe, până la relații între variabile numerice și categorice. Această abordare detaliată ne-a permis să evidențiem variabilitățile semnificative în datele colectate, furnizând o bază solidă pentru interpretarea rezultatelor și pentru orientarea cercetărilor viitoare în domeniul cardiologic.

DEZVOLTAREA MODELELOR

În lumina creșterii alarmante a bolilor cardiovasculare la nivel global, se impune folosirea tehnologiilor avansate, cum ar fi învățarea automată, pentru a anticipa și preveni aceste afecțiuni. Această analiză se axează pe evaluarea a trei modele distincte de învățare automată - Regresia Logistică, Random Forest și Decision Tree - în ceea ce privește predictibilitatea riscului de boli de inimă.

- Regresia Logistică utilizează o abordare statistică pentru predicția binară, examinând relația dintre variabilele independente (vârstă, sex, nivelul colesterolului etc.) și prezența sau absența bolii de inimă.
- Random Forest, un algoritm de învățare supervizată, construiește o colecție de arbori de decizie pentru a realiza predicții precise, fiecare arbore fiind antrenat pe o porțiune aleatoare a datelor de antrenare.
- Decision Tree, un model simplu și intuitiv, este utilizat pentru clasificarea datelor, divizând setul în ramuri pe baza criteriilor specifice pentru a prezice prezența bolilor de inimă. Aceste modele oferă abordări diferite, furnizând soluții eficiente în eforturile de prevenire a bolilor cardiovasculare.

Evaluarea performanțelor modelelor

Pentru a evalua eficacitatea modelelor noastre în predicția riscului de boli de inimă, am adoptat un set de metrici standard. Acuratețea furnizează o perspectivă asupra proporției de predicții corecte, în timp ce precizia și recall-ul oferă informații despre exactitatea predicțiilor pozitive și capacitatea modelului de a identifica corect cazurile reale pozitive.

În paralel cu aceste metrici, am integrat Curba ROC pentru a evalua performanța modelelor într-un context mai larg, explorând raportul dintre rata de adevărate pozitive și rata de false pozitive. Matricea de Confuzie a fost utilizată pentru a oferi o imagine detaliată a rezultatelor, evidențiind numărul de adevărate pozitive, adevărate negative, false pozitive și false negative.

Implementarea acestor evaluări a fost realizată eficient în Python, prin intermediul bibliotecilor specializate precum scikit-learn. Această abordare asigură o înțelegere cuprinzătoare a performanțelor modelelor noastre, esențială în contextul cercetării noastre asupra riscului de boli de inimă.

REZULTATE

Analiza distribuției datelor este crucială în diverse domenii, deoarece furnizează perspective valoroase asupra modelelor, tendințelor și caracteristicilor subiacente ale datelor. Să explorăm mai întâi pacienții cu și fără afecțiuni cardiovasculare în contextul bolilor de inimă:

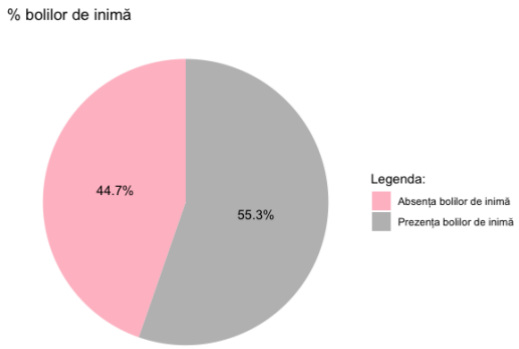

Figura 1. % bolilor de inimă

Diagrama circulară evidențiază procentajele relative, cu peste jumătate din populație (55.3%) prezentând boli de inimă, în timp ce o proporție ușor mai mică (44.7%) nu prezintă astfel de afecțiuni.

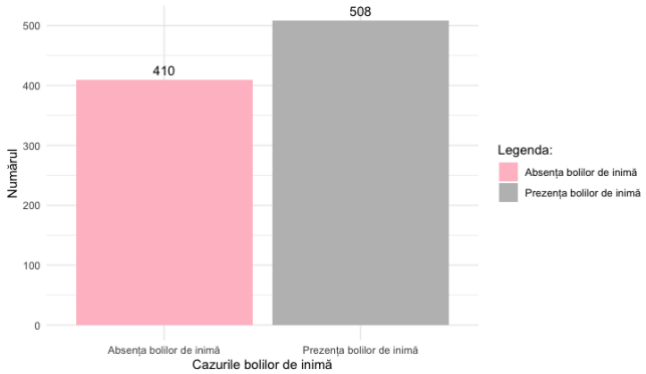

Figura 2. Numărul absolut al cazurilor de boli e inimă

În complementaritate, diagrama cu bare demonstrează numărul absolut al cazurilor, oferind o ilustrare directă a impactului bolilor de inimă: 410 indivizi fără boală cardiacă stau în contrast cu 508 indivizi care suferă de această condiție.

Analiza univariată

În contextul analizei univariate a datelor, se evidențiază discrepanțe semnificative între populația de sex masculin și cea de sex feminin în ceea ce privește incidența bolilor cardiace. Conform constatărilor, numărul de pacienți cu afecțiuni cardiace este mai mare în rândul bărbaților în

comparație cu cei fără astfel de probleme, în timp ce la femei, observăm o proporție inversă, cu un număr mai mic de pacienți cu boli de inimă în comparație cu cei sănătoși.

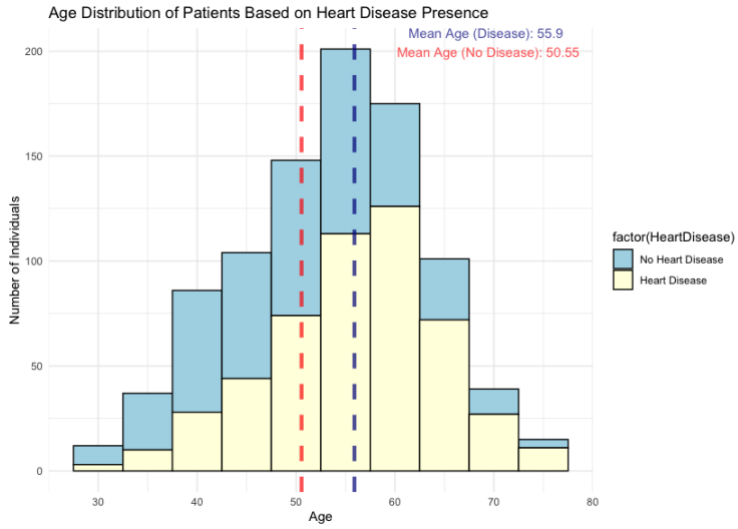

Figura 3. Distribuția vârstei pacienților în funcție de prezența bolilor de inimă

O concluzie solidă poate fi trasă din asocierea fermă dintre durerea toracică de tip ASY și probabilitatea semnificativă de a dezvolta boli cardiace.

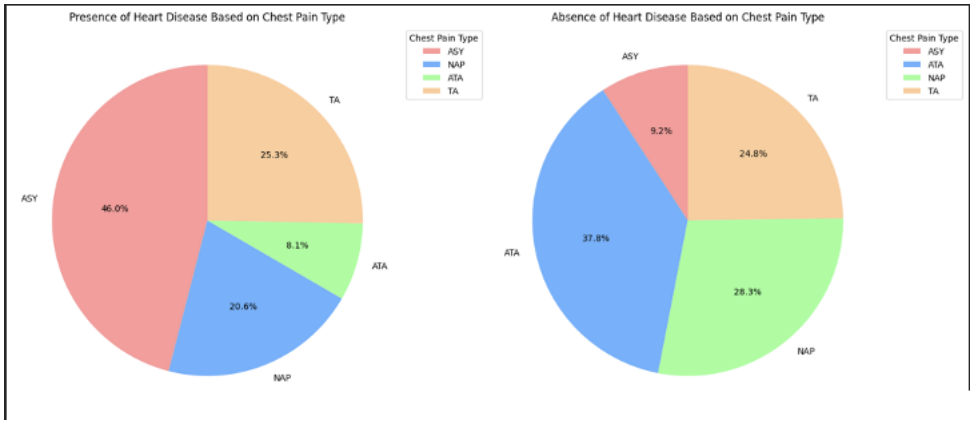

Figura 4. Diagramă "Prezența bolii de inimă în funcție de tipul durerii în piept"

De asemenea, nivelul glicemiei în post a fost identificat ca un factor delicat, deoarece atât pacienții diagnosticați, cât și cei nedepistați cu glicemie în post prezintă un număr semnificativ de cazuri de boli cardiace. Cu toate acestea, analiza ECG-ului în repaus nu furnizează o categorie distinctă care să delimiteze pacienții cu boli de inimă, deoarece toate cele trei valori conțin un număr considerabil de astfel de cazuri. În schimb, angina indusă de exerciții fizice a fost identificată ca un factor semnificativ care crește probabilitatea diagnosticării bolilor cardiace.

Analiza detaliată a valorilor pantelor ST subliniază diferențele semnificative între tipurile de pantă și probabilitatea de a diagnostica bolile de inimă. Pantele ST plate prezintă o probabilitate ridicată de a indica afecțiuni cardiace, în timp ce cele descendente sau ascendente oferă rezultate similare, dar într-un număr mai mic de puncte de date.

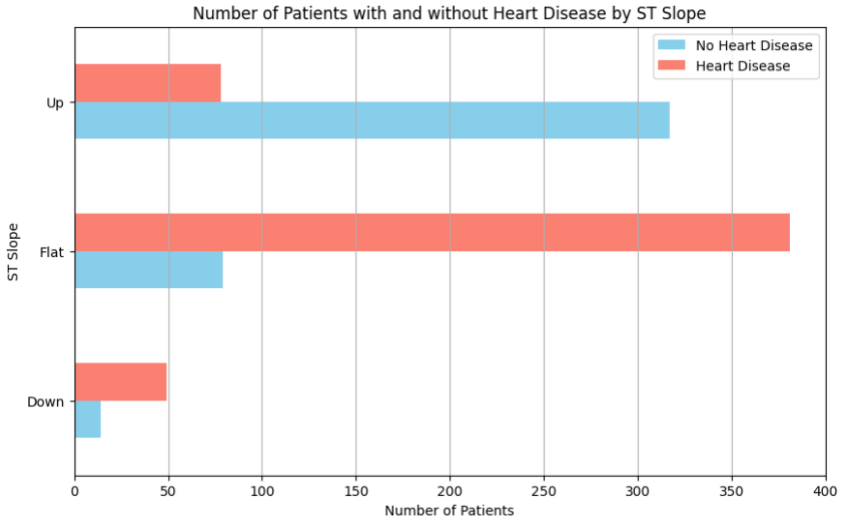

Figura 5. Numărul de pacienți cu și fără boală de inimă în funcție de panta segmentului ST

Aceste constatări oferă o perspectivă detaliată asupra semnificației clinice a diferitelor parametri în ceea ce privește diagnosticul bolilor cardiace, furnizând informații esențiale pentru îmbunătățirea preciziei diagnosticului și optimizarea strategiilor de intervenție medicală.

Analiza bivariată

Analiza bivariată relevă diferențe semnificative între populațiile masculine și feminine în asocierea cu bolile cardiace. Pentru bărbați, bolile cardiace sunt prezente la majoritatea valorilor caracteristicilor numerice, cu depresia segmentului ST și frecvența cardiacă maximă sub 140 de ani devenind semnificative după vârsta de 50 de ani. La femei, datele sunt mai limitate.

Durerea toracică de tip ASY este predominantă, iar după vârsta de 50 de ani, bolile cardiace apar indiferent de diagnosticul de glicemie în post. Pacienții cu glicemie în post și tensiune arterială peste 100 prezintă o prevalență crescută a bolilor cardiace. Combinația colesterol-glicemie nu pare să influențeze semnificativ cauzele bolilor cardiace.

Pacienții fără glicemie în post și frecvența cardiacă maximă sub 130 sunt mai susceptibili la boli cardiace. Valorile ECG în repaus de tip Normal, ST și LVH indică bolile cardiace începând cu vârstele de 30, 40 și 40 de ani, cu o creștere notabilă după vârsta de 50 de ani. Colesterolul între 200 și 300, împreună cu valori ST la ECG în repaus, evidențiază un grup de pacienți cu boli cardiace.

Pentru frecvența cardiacă maximă sub 140 și ECG în repaus normal, cazurile de boli cardiace sunt dense.

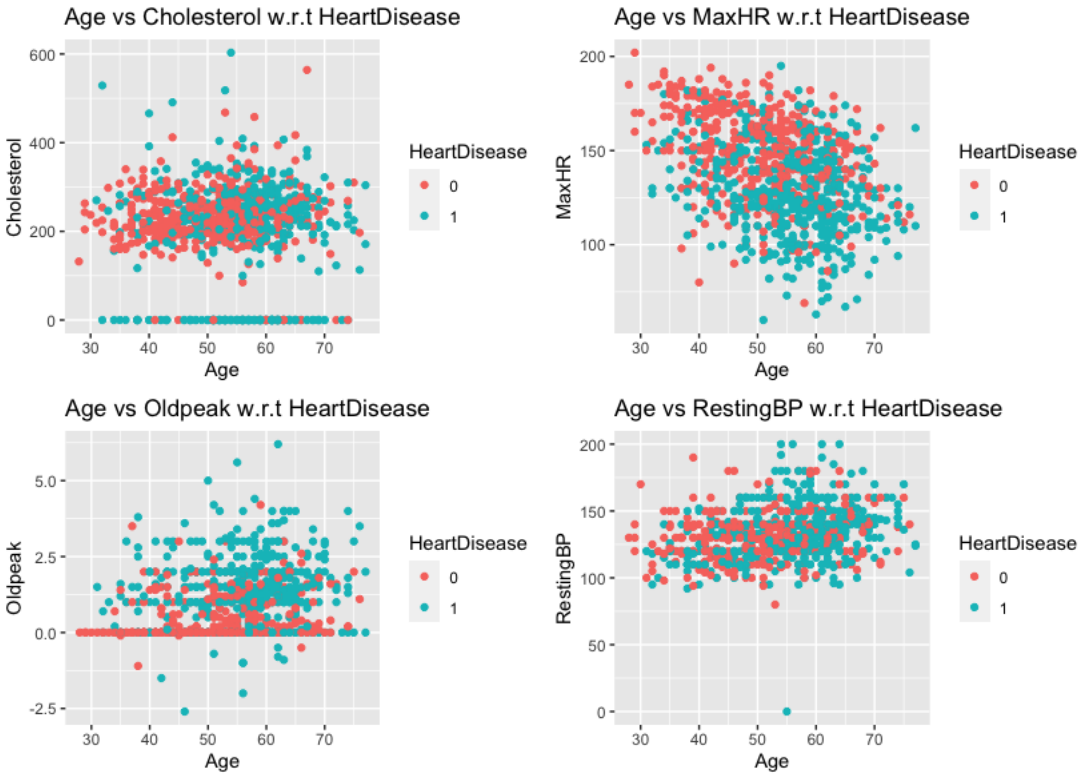

Figura 6. Analiza corelației între vârstă și factori de risc cardiovasculari

Angina indusă de exerciții fizice și valorile pantelor ST, în special cele plate, prezintă o corelație semnificativă cu bolile cardiace, indicând influențe diferite asupra diagnosticului. Aceste constatări contribuie la înțelegerea complexității asocierilor dintre factorii analizați și bolile cardiace, oferind informații utile pentru strategii de diagnostic și intervenție medicală.

Intervalul Interquartilic pentru variabilele numerice

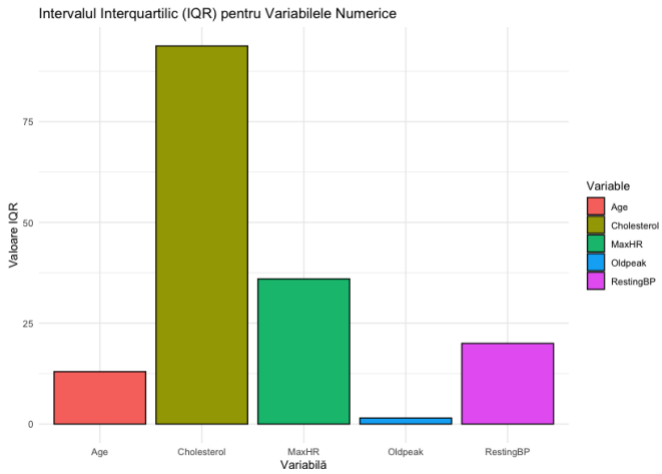

Figura 7. Varianța variabilelor numerice IQR

Se observă că variabila cu cel mai mare IQR este 'MaxHR', indicând o variație mai mare în datele pentru această variabilă comparativ cu celelalte, cum ar fi 'Age' și 'RestingBP', care prezintă intervale interquartilice mult mai mici.

REZULTATELE MODELELOR

Analiza performanței modelelor în diagnosticarea bolilor cardiace prin regresie logistică, Random Forest și Decision Tree oferă perspective valoroase. Regresia logistică a obținut o acuratețe de aproximativ 85%, indicând o bună capacitate de clasificare a cazurilor pozitive și negative. Precizia, recall-ul și scorul F1 au atins valori de 89.22%, 85.05%, respectiv 87.08%, evidențiind o echilibrare între exactitatea predicțiilor pozitive și capacitatea de a identifica corect cazurile pozitive.

În comparație, modelul Random Forest a obținut o acuratețe superioară de 87.50%, excelând în detectarea cazurilor de boală a inimii, cu 95 de adevărat pozitive și doar 12 fals negative. Metricile preciziei, recall-ului, scorului F1 și AUC-ROC au atins valori înalte (89.62%, 88.79%, 89.20%, 92.88%, respectiv), evidențiind eficiența îmbunătățită comparativ cu regresia logistică.

În ceea ce privește Decision Tree, acuratețea a fost similară cu cea a regresiei logistice, de 85.33%. Cu toate acestea, performanța acestui model a variat în celelalte metrici, cu precizia de 86.36%, recall-ul de 88.79%, scorul F1 de 87.56%, și AUC-ROC de 84.65%. Desigur, Decision Tree a prezentat o performanță comparabilă cu Random Forest în ceea ce privește adevărat pozitive, dar cu o ușoară creștere a falselor pozitive la 15, indicând o tendință ușor mai mare de a clasifica greșit cazurile negative ca pozitive.

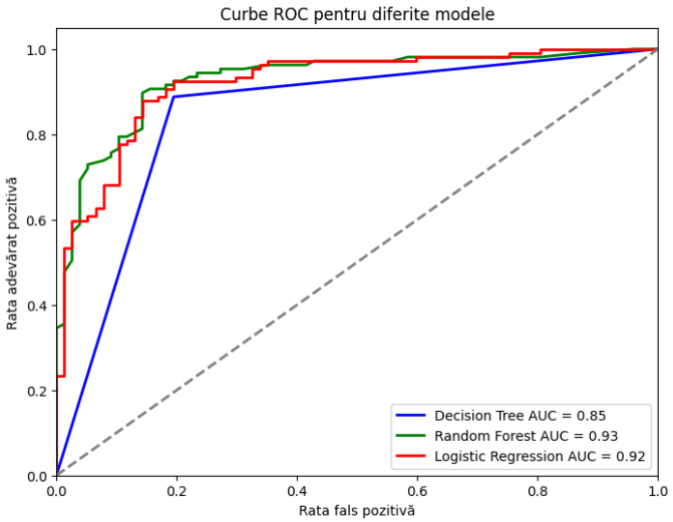

Figura 8. Compararea performanței modelelor de clasificare prin curbele ROC și aria de sub curbe (AUC)

DISCUȚII

Această cercetare reprezintă o inițiativă semnificativă în abordarea provocărilor asociate detecției și prevenirii bolilor cardiovasculare (BCV). Prin introducerea unui model predictiv inovator, axat pe date comprehensive ale pacienților, inclusiv parametri demografici, clinici și fiziologici, studiul urmărește să îmbunătățească detectarea timpurie și să faciliteze strategii preventive personalizate. Cu toate că modelele de regresie logistică, Random Forest și Decision Tree au demonstrat performanțe promițătoare în predicția riscului de BCV, există direcții viitoare esențiale pentru consolidarea și extinderea cercetării. Astfel, integrarea unor informații suplimentare în setul de date, cum ar fi istoricul familial și markeri genetici, ar adăuga o perspectivă mai detaliată asupra riscului cardiovascular. Validarea externă a modelelor pe seturi de date independente și explorarea tehnologiilor emergente, precum învățarea adâncă, reprezintă pași importanți pentru confirmarea generalizabilității și îmbunătățirea performanțelor acestora. În paralel, optimizarea modelelor și gestionarea dezechilibrului în distribuția datelor ar contribui la maximizarea eficacității predicțiilor. De asemenea, este esențială o interpretare clinică îmbunătățită a rezultatelor și o colaborare continuă cu profesioniștii din domeniul sănătății pentru a adapta modelele la nevoile reale ale practicii medicale. Prin aceste direcții și îmbunătățiri, cercetarea reprezintă o contribuție semnificativă la transformarea asistenței medicale în gestionarea și prevenirea bolilor cardiovasculare.

MATERIALE ADIȚIONALE

Anexa 1. Detalii suplimentare privind variabilele clinice pentru evaluarea bolilor cardiace

Tabel 1. Date Clinice și Diagnostice pentru Pacienți cu Boli Cardiace

Variabila	Tipul de date	Descriere
Age	Numeric	Vârsta pacientului
Sex	Text	Sexul pacientului (M: Masculin, F: Feminin)
ChestPainType	Text	Tipul durerii de piept (ATA, NAP, ASY, TA)
RestingBP	Numeric	Tensiunea arterială în repaus (mm Hg)
Cholesterol	Numeric	Colesterolul seric (mm/dl)
FastingBS	Numeric	Glicemia în repaus (1: dacă Glicemie în repaus > 120 mg/dl, 0: altfel)
RestingECG	Text	Rezultatele electrocardiografice în repaus (Normal, ST, LVH)
MaxHR	Numeric	Frecvența cardiacă maximă atinsă
ExerciseAngina	Text	Angina indusă de efort (Y: Da, N: Nu)
Oldpeak	Numeric	Depresia segmentului ST indusă de exercițiu în raport cu repausul
ST_Slope	Text	Inclinația segmentului ST de vârf al exercițiului maxim (Sus, Plat, Jos)
HeartDisease	Numeric	Prezența bolii de inimă (1: Da, 0: Nu)

Anexa 2. Figuri adăugătoare folosite pentru verificarea ipotezelor

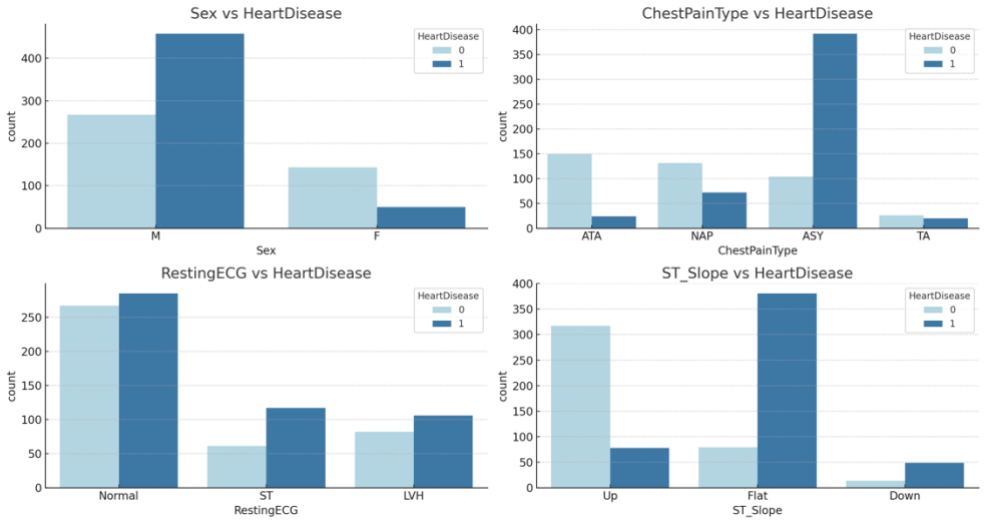

Figura 9. Distribuția bolii de inimă în funcție de variabilele categorice

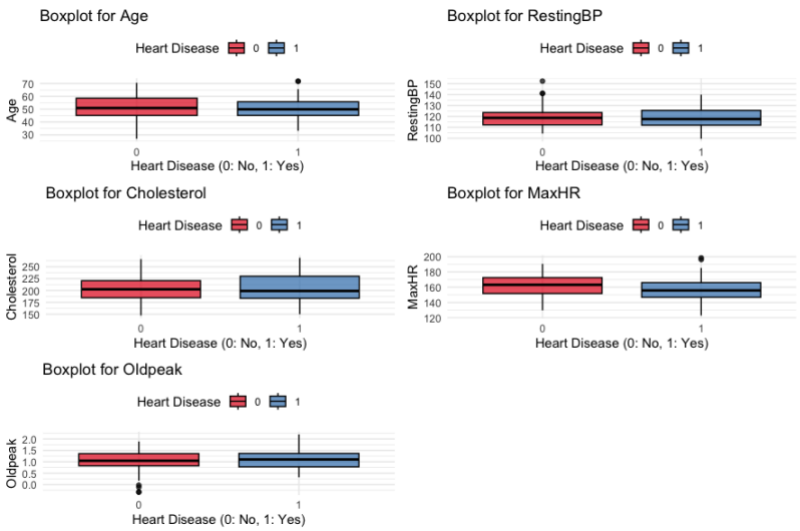

Figura 10. Distribuția bolii de inimă în funcție de variabilele numerice

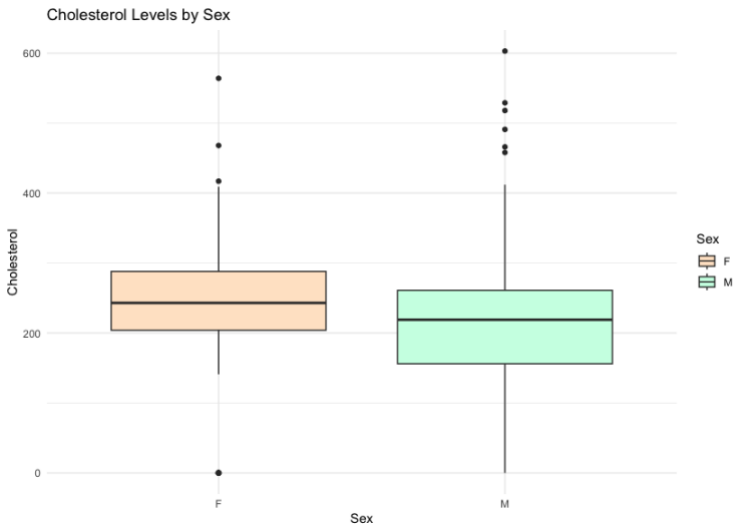

Figura 11. Boxplot cu nivelurile de colesterol după sex

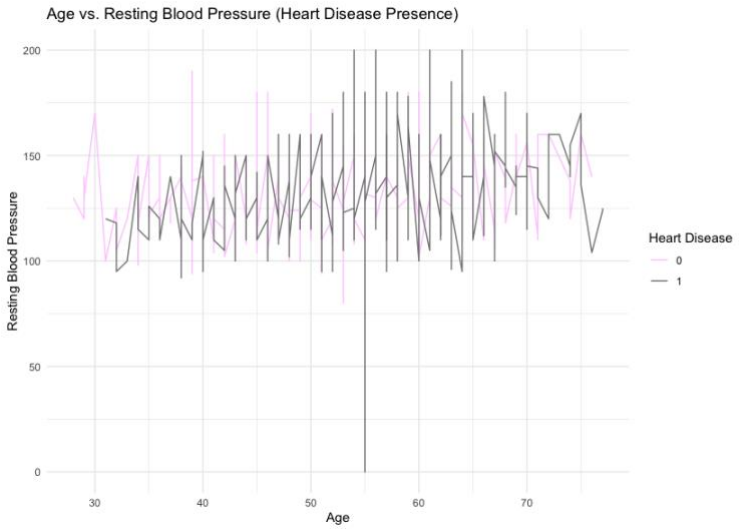

Figura 12. Relația dintre vârstă și tensiunea arterială

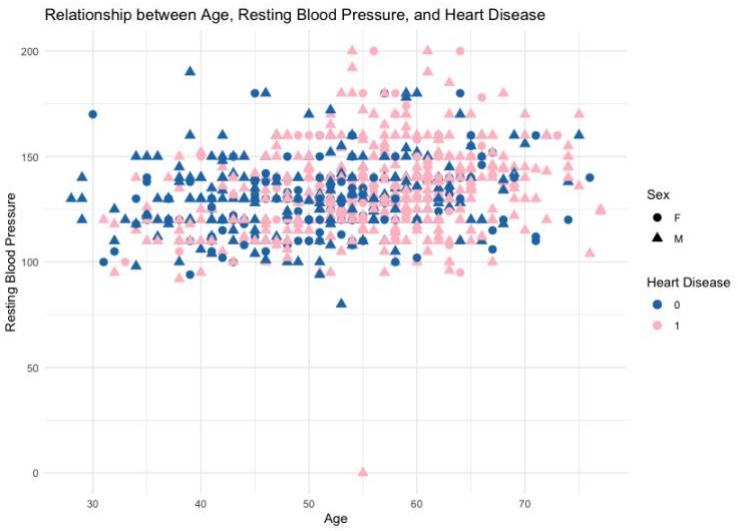

Figura 13. Relația dintre vârstă, tensiunea arterială în repaus și prevalența bolii de inimă

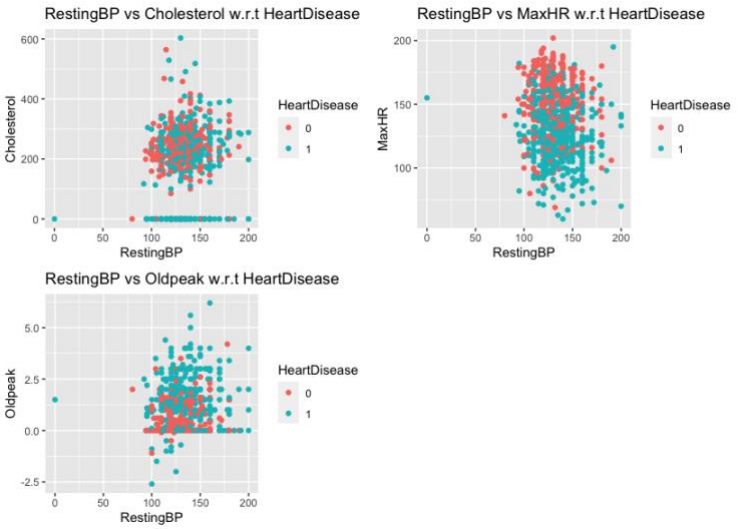

Figura 14. Analiza relațiilor dintre parametrii fiziologici și boala de inimă

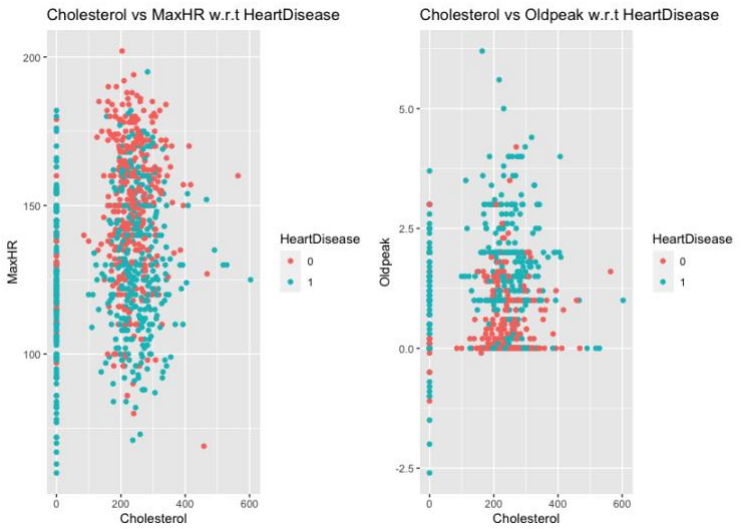

Figura 15. Relația dintre nivelurile de colesterol și alți parametri fiziologici în contextul bolii de inimă

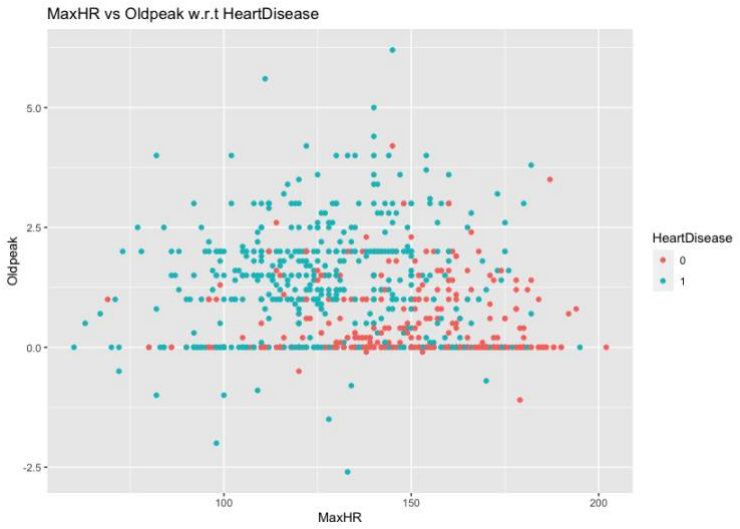

Figura 16. Compararea ratei maxime a inimii și oldpeak în cazurile de boală cardiacă

BIBLIOGRAFIE

1. R: The R Project for Statistical Computing [Internet]. [citat 12 decembrie 2023]. Disponibil la: https://www.r-project.org/
2. Tidyverse [Internet]. [citat 12 decembrie 2023]. Disponibil la: https://www.tidyverse.org/
3. A Grammar of Data Manipulation [Internet]. [citat 12 decembrie 2023]. Disponibil la: https://dplyr.tidyverse.org/
4. Tidy Messy Data [Internet]. [citat 12 decembrie 2023]. Disponibil la: https://tidyr.tidyverse.org/
5. https://github.com/YanaT26/BCV