***ANALIZA ȘI PREDICȚIA RISCULUI ATACULUI DE CORD***

Cristalin Ciobanu1

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

# ABSTRACT

Atacul de cord rămâne o problemă majoră de sănătate la nivel global, reprezentând o cauză semnificativă de morbiditate și mortalitate. În acest studiu, am abordat analiza și predicția riscului de atac de cord utilizând setul de date "Heart Attack Analysis & Prediction" disponibil pe Kaggle[1]. Scopul nostru a fost de a identifica factorii de risc relevanți și de a dezvolta modele logistice pentru a estima probabilitatea de atac de cord. Am inițiat analiza cu explorarea setului de date, factorizând variabilele categorice și efectuând o analiză descriptivă. Am evidențiat relațiile dintre variabile și am identificat modele semnificative care pot influența riscul de atac de cord. De asemenea, am realizat vizualizări grafice, precum histograma presiunii arteriale medii în funcție de sex și distribuția tipurilor de durere în ambele sexe în funcție de vârstă. Modelarea logistică a constituit o componentă crucială a studiului nostru. Am dezvoltat mai multe modele, explorând diverse combinații de variabile, cum ar fi vârsta, sexul, tipul de durere și nivelurile de colesterol. Analiza VIP (Variable Importance in Projection) ne-a permis să identificăm principalele caracteristici asociate cu riscul de atac de cord, punând în evidență. Rezultatele noastre au demonstrat că anumite variabile, precum vârsta, sexul și nivelurile de colesterol, sunt semnificativ asociate cu probabilitatea de atac de cord. Modelele logistice au oferit o înțelegere mai profundă a interacțiunilor dintre aceste variabile, sprijinind procesul de predicție. În concluzie, această analiză detaliată a setului de date "Heart Attack Analysis & Prediction" a contribuit la înțelegerea factorilor de risc implicați în atacul de cord. Studiul nostru oferă perspective valoroase cum diferite caracteristici ale omului pot influența probabilitatea unui eveniment cardiac și poate servi drept bază pentru dezvoltarea ulterioară a modelelor de predicție și a strategiilor de intervenție.

# INTRODUCERE

Pe parcursul ultimelor decenii, cercetările în domeniul sănătății cardiovasculare au evoluat considerabil, beneficiind de avansurile tehnologice și de accesul sporit la date relevante. Cu toate acestea, există încă lacune semnificative în înțelegerea complexă a interacțiunilor dintre factorii de risc, precum vârsta, sexul, nivelurile de colesterol și presiunea arterială, care contribuie la apariția atacurilor de cord.

Atacul de cord (infarctul miocardic) rămâne una dintre cele mai presante probleme de sănătate la nivel mondial, exercitând o presiune semnificativă asupra sistemelor de sănătate și având consecințe serioase asupra calității vieții și a longevității populației. Conform unui studiu efectuat în Statele Unite ale Americii, bolile de inimă reprezintă principala cauză de deces pentru bărbați, femei și persoane din majoritatea grupurilor rasiale și etnice din Statele Unite. De exemplu, dacă luăm date mai generale legate de boli cardiovasculare, nu neapărat atac de cord, o persoană moare la fiecare 33 de secunde în Statele Unite. Aproximativ 695.000 de persoane din Statele Unite au murit din cauza bolilor de inimă în 2021 - adică 1 din 5 decese.

Bolile de inimă au costat Statele Unite aproximativ 239,9 miliarde de dolari în fiecare an, în perioada 2018-2019. Acest lucru include costul serviciilor de asistență medicală, al medicamentelor și al productivității pierdute din cauza decesului.

Dacă ne axăm pe atacuri de cord, în Statele Unite, cineva suferă un atac de cord la fiecare 40 de secunde. În fiecare an, aproximativ 805.000 de persoane din Statele Unite suferă un atac de cord. Dintre acestea, 605.000 reprezintă un prim atac de cord, iar altele 200.000 de cazuri se întâmplă la persoane care au avut deja un atac de cord. Aproximativ 1 din 5 atacuri de cord sunt silențioase - daunele sunt făcute, dar persoana nu este conștientă de ele[2].

Seturile de date ample și accesibile precum "Heart Attack Analysis & Prediction" oferă oportunități semnificative de a investiga profunzimea și amploarea problemelor de sănătate cardiovasculară[1]. În lumina acestui context, a fost desfășurat un studiu detaliat, utilizând acest set de date, pentru a analiza și a prezice riscul de atac de cord. Scopul principal a fost să se identifice factorii de risc cheie și să se dezvolte un model logistic eficient pentru a estima probabilitatea de a suferi un atac cardiac în viitor.

Prin intermediul acestui studiu, ne propunem să aducem o contribuție semnificativă în domeniul analizei și predicției riscului de atac de cord. Vom explora relațiile dintre variabile, vom dezvolta modele logistice relevante și vom evalua impactul acestora asupra probabilității de atac de cord. De asemenea, vom examina modul în care factorii de risc identificați în acest studiu pot influența elaborarea unor strategii personalizate de prevenție și intervenție.

Într-un context mai larg, această cercetare poate contribui la avansarea cunoștințelor noastre despre riscul cardiovascular, oferind fundamentul pentru dezvoltarea unor metode de identificare și intervenție mai eficiente. Prin urmare, această investigație se aliniază eforturilor globale de combatere a bolilor cardiovasculare, reprezentând un pas esențial către îmbunătățirea sănătății publice și a calității vieții.

# MATERIALE ȘI METODE

***Setul de date***

Studiul nostru a utilizat setul de date "Heart Attack Analysis & Prediction" disponibil pe platforma Kaggle[1]. Acesta conține informații variate despre pacienți, aspecte precum vârsta și sexul, măsurători medicale precum presiunea arterială în repaus și nivelurile de colesterol, precum și variabila-cheie de ieșire care indică dacă pacientul a suferit sau nu un atac de cord.

***Preprocesarea datelor***

Înainte de analiză, am efectuat o preprocesare datelor. Variabilele categorice, cum ar fi sexul, tipul de durere toracică, nivelul de colesterol, nivelul de zahăr, tipul de afecțiune a cordului au fost factorizate pentru a permite o analiză eficientă. Nivelul de zahăr conține un singur nivel în urma factorizării. De aceia, nu a fost inclus în crearea modelului logistic, ci doar la analiza exploratorie a datelor. Am utilizat funcții precum `*summary*` și `*glimpse*` pentru a obține o înțelegere detaliată a distribuției și caracteristicilor setului de date.

***Analiza exploratorie***

Analiza exploratorie a fost realizată pentru a evidenția relațiile dintre variabile și pentru a identifica tendințe semnificative. Cu ajutorul pachetelor ggplot2, tidyverse, am generat vizualizări informative, cum ar fi histograma presiunii arteriale în funcție de sex și distribuții ale tipurilor de durere toracică în funcție de vârstă.

***Modelarea logistică***

Setul de date a fost divizat într-un set de antrenare și unul de testare folosind funcția `*initial\_split*` din pachetul rsample. Am construit apoi mai multe modele logistice, considerând combinații diferite de variabile. Modelele au fost dezvoltate folosind funcția `*glm*` pentru regresie logistică, având ca variabilă de ieșire prezența sau absența atacului de cord.

***Analiza VIP (Variable Importance in Projection)***

Pentru a identifica influența variabilelor asupra predicției atacului de cord, am aplicat analiza VIP folosind pachetul `*vip*`. Această metodă ne-a ajutat să evidențiem cele mai semnificative caracteristici asociate cu riscul cardiovascular, contribuind la o înțelegere mai profundă a mecanismelor implicate[3].

Prin această abordare comprehensivă și detaliată, ne-am propus să obținem nu doar rezultate precise, ci și să aducem lumină asupra factorilor cheie care pot influența riscul de atac de cord. Această metodologie robustă și analiza detaliată a setului de date ne-au permis să construim modele informative și să aducem contribuții semnificative la domeniul sănătății cardiovasculare.

***Evaluarea performanței modelelor***

Curba ROC oferă o ilustrare vizuală a performanței unui model de clasificare binară pe întregul spectru al pragurilor de decizie. Adevărata rată pozitivă (TPR) este reprezentată pe axa Y, iar rata fals pozitivă (FPR) este reprezentată pe axa X. Un model perfect ar avea o curba care s-ar deplasa de la colțul din stânga jos la cel din dreapta sus, având o arie sub curba (AUC-ROC) de 1.

AUC-ROC măsoară capacitatea modelului de a distinge între exemplele pozitive și cele negative. Cu cât AUC-ROC este mai mare, cu atât modelul este considerat mai bun la clasificarea corectă a exemplelor pozitive și negative.

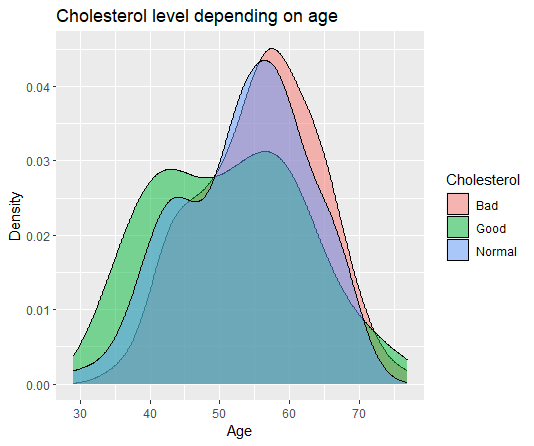
# REZULTATE

În setul de date analizat avem o variabilă dependentă *output,* și mai multe variabile independente precum *sexul, vârsta, tipul de durere prepoderent, numărul maxim de bătăi a cordului înregistrate, nivelul de colesterol, abatareri ale cordului(anomalii) și tensiunea arterială în stare de repaus*. Variabila independentă *age* are un min de 29 de ani și un max de 77 de ani, iar o medie pentru setul de date analizat este 54,37. De exemplu, variabila *trtbps* ce indică numărul maxim de bătai ale inimii are un min de 94 și un max de 200. Dacă ne referim la variabila ce conține observații privitor la tensiunea arterială sistolică, atunci min este de 71 mm Hg, iar max este de 202 mm Hg. Este bine de remarcat că o persoană ce nu prezintă acuze să aibă tensiunea între 120-129 mm Hg.

***Analiza distibuțiilor variabilelor***

În încercarea de a înțelege relația dintre vârstă și nivelul colesterolului din sânge, am analizat setul de date. Una din presupuneri este dacă există o asociere semnificativă între înaintarea în vârstă și nivelurile de colesterol în sânge.

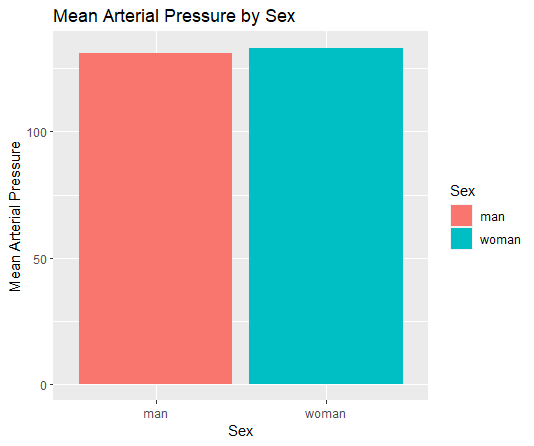
Prin urmare, ne-am concentrat pe variabilele '*age*' și '*chol*' pentru a explora trendurile posibile. Pentru a ilustra această relație și pentru a facilita înțelegerea rezultatelor noastre, am pregătit o figură informativă care sintetizează corelația între vârstă și nivelurile de colesterol în sânge(*figura 1*).



**Figura 1.** *Nivelul de colesterol raportat la vârstă*

Plotul final este un grafic de densitate(*figura 1*) pentru a evidenția distribuția nivelurilor de colesterol în funcție de vârstă. Acesta utilizează curbe de densitate pentru fiecare nivel de colesterol. Putem observa că după vârsta de 40 de ani, nivelul colesterolului are tendința de a crește și a trece din categoria *Normal* în categoria *Bad*.

În efortul de a dezvălui posibile corelații între variabilele tensiunii arteriale și sexului, am analizat atent setul nostru de date. O presupunere pe care am examinat-o este dacă există o asociere semnificativă între sex și nivelurile tensiunii arteriale, cu o posibilă tendință de a observa niveluri mai ridicate la femei în comparație cu bărbații.

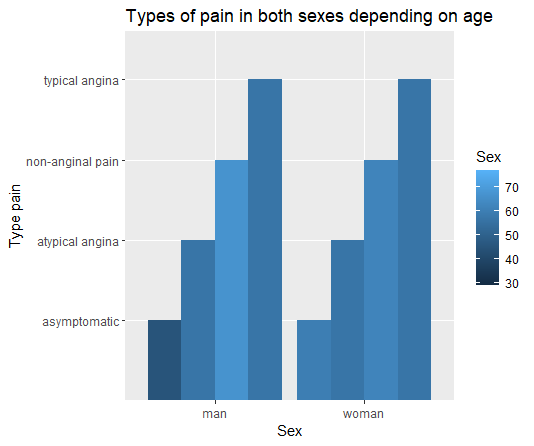


**Figura 2.** *Media tensiunii arteriale pentru ambele sexe*

În urma analiza acestei presupuneri, am optat pentru vizualizarea cu ajutorul unui grafic de tip bar(*figura 2*) pentru a compara distribuțiile tensiunii arteriale în funcție de sex. Dar, spre surprindere presupunerea nu s-a adeverit, fiind depistată o diferență minoră în setul de date analizat.

O altă presupunere pe care am cercetat-o este dacă există vreo asociere semnificativă între vârstă și prevalența anumitor tipuri de durere, sugerând posibile schimbări în raportarea sau frecvența acestora odată cu înaintarea în vârstă.

Pentru a evidenția aceste potențiale corelații, am focalizat atenția asupra variabilelor relevante, respectiv '*age*' și *'cp'*.

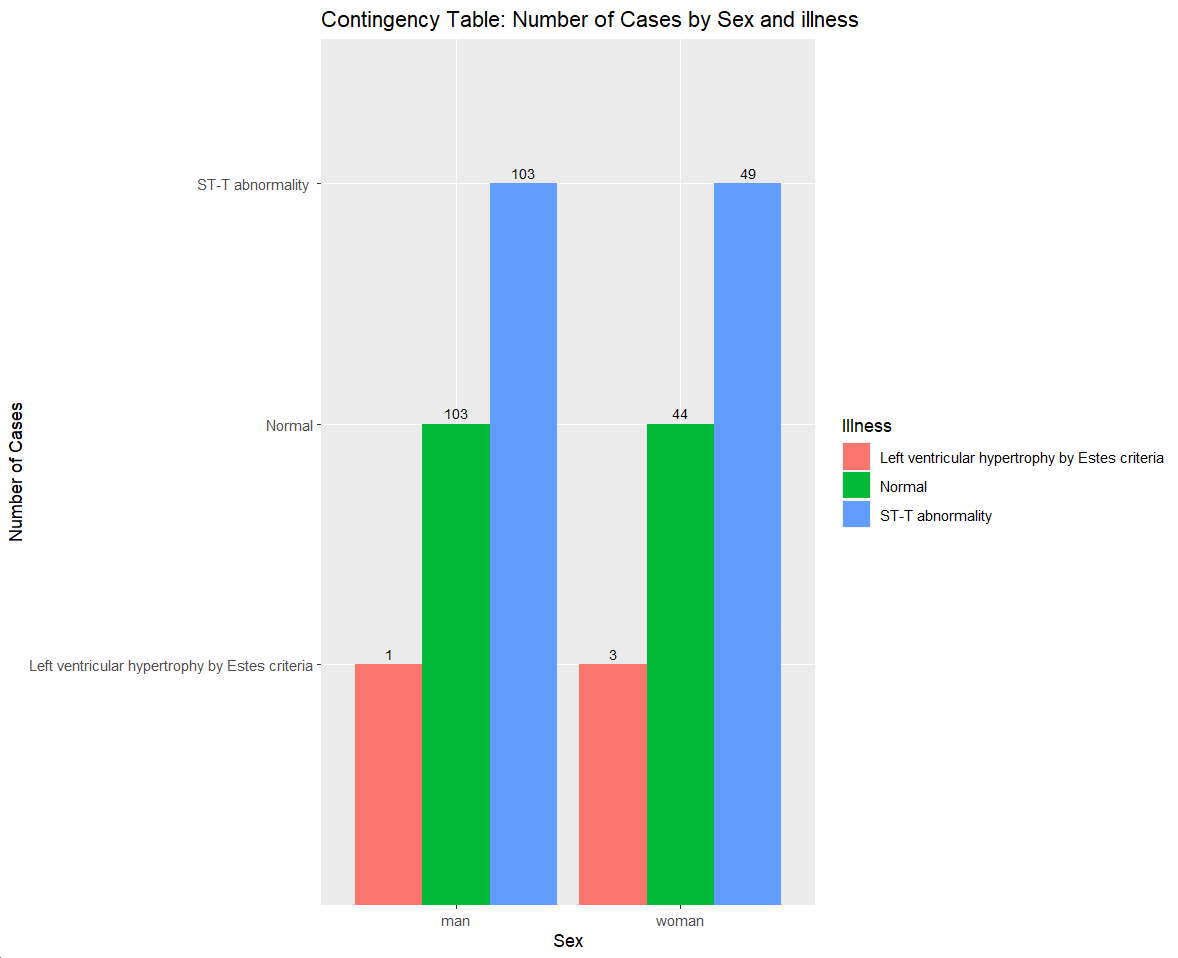


**Figura 3.** *Tipul de durere raportat la sex*

Folosind un grafic de tip bar, am intenționat să evidențiem frecvența diferitelor tipuri de durere în grupuri de vârstă distincte. Astfel, putem observa(*figura 3*) că bărbații cu vârsta cuprinsă între 30 și 40 de ani nu prezintă dureri, iar la sexul opus diapazonul fiind mai mare, ajungând și la vârsta de 70 de ani în unele cazuri. Totuși, vârsta își ia a ei, de aceia, o dată cu înaintarea în vârstă sunt tot mai prezente diferite tipuri de dureri, făcându-și apariția unele doar la efort fizic continu, altele fiind permanente, mai ales în rândul bărbaților ce trec de vârsta de 50 de ani.

În procesul de cercetare a datelor, am analizat atent variabilele legate de sexul pacienților și numărul de cazuri de diferite boli pentru a evalua posibile diferențe semnificative în incidența acestora în funcție de sex. O presupunere pe care am investigat-o este dacă există vreo asociere semnificativă între sex și frecvența anumitor boli, sugerând că acestea pot avea o incidență mai ridicată la un anumit sex în comparație cu celălalt.

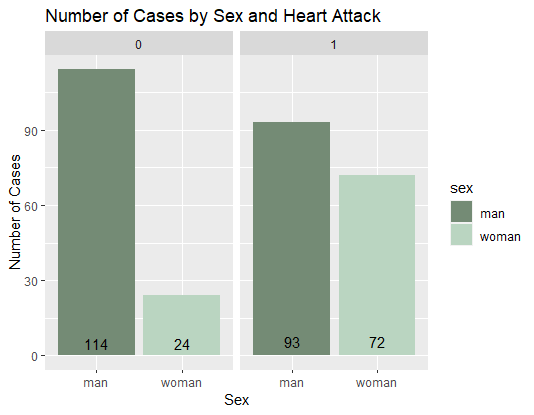
În vederea evidențierii acestor posibile diferențe, am direcționat atenția către variabilele relevante, cum ar fi '*sex*' și '*restcg*'.



**Figura 4.** *Afecțiuni ale cordului în dependență de sex*

Cu ajutorul unui grafic de tip bar, am dorit să comparăm numărul de cazuri ale diferitelor boli între sexe. Plotul din *figura 4* ne oferă claritate. În cazul hipertrofiei ventriculare stângă în rândul bărbaților este doar un singur caz, comparativ cu femeile unde sunt 3 cazuri. De asemenea, în rândul bărbaților au fost înregistrate 103 persoane la care nu au fost detectate afectiuni ale cordului, comparativ cu femeile la care această cifră este mult mai mică, fiind doar 44 de cazuri. Ce ține de a 3-a afecțiune, care poate fi depistată cu efectuarea doar a unei electrocardiograme, rezultatul este opus, fiind înregistrate 103 cazuri între rândurile bărbaților și 49 la femei. În urma analizei, putem afirma cu siguranță că unele afecțiuni ale cordului sunt prepoderente doar la un sex, celălalt sex fiind mai imun unor asemenea afecțiuni.

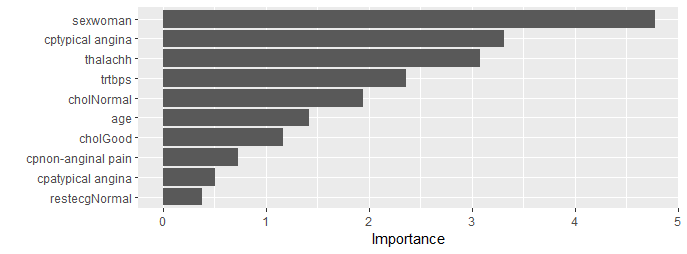
De asemenea, datele medicale arată că, în general, bărbații au o incidență mai mare de atacuri de cord decât femeile. Acest lucru contravine presupunerii inițiale, care sugera că femeile ar fi mai predispuse la atacuri de cord. Factori precum diferențele biologice, hormonale și comportamentale pot influența acest fenomen. Este esențial să subliniem că aceste concluzii se bazează pe tendințe generale și că fiecare individ poate prezenta riscuri specifice în funcție de istoricul său medical, stilul de viață și alți factori.



**Figura 5.** *Numărul de atac de cord pe sex*

Similar cu presupunerea precedentă, vom utiliza un grafic de tip bar pentru a compara numărul de cazuri de atacuri de cord între sexe. Astfel, putem observa(*figura 5*) că din 207 bărbați și 96 femei, 114 bărbați și 24 de femei nu au suferit un atac de cord, pe când 93 de bărbați și 72 de femei au suferit un atac de cord. Putem concluziona că doar 25% din femei nu au suferit atact de cord, spre deosebire de bărbați unde raportul este 55%.

Odată ce am terminat analiza exploratorie a datelor, pentru setul de date ales am creat modele de regresie logistică. Am selectat regresia logistică deoarece este o metodă de analiză statistică utilizată pentru a modela relația dintre o variabilă dependentă binară și una sau mai multe variabile independente. Acest model este numit logistic deoarece utilizează funcția logistică pentru a transforma o combinație liniară a variabilelor independente într-o probabilitate cuprinsă între 0 și 1.

**

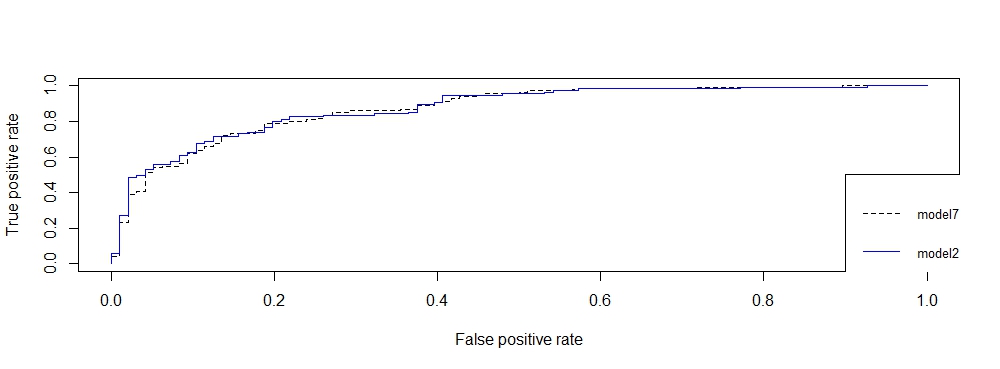
**Figura 6.** *Importanța variabilelor pentru model*

Pentru a putea creia un model este necesar și de a ști care variabile sunt importante și care sunt redundante și antrenează modelul greșit. Pentru a putea spune cu precizie ce este util, am utilizat funcția *vip* care ne indică importanța unei variabile pentru model(*figura 6*).

**Tabelul 1.** *Selectarea modelelor în baza AICc*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model name** | **K** | **AICc** | **Delta\_AICc** | **AICcWt** | **Cum.Wt** | **LL** |
| heart\_atack\_model\_2 | 10 | 196,74 | 0,00 | 0,33 | 0,33 | -87,82 |
| heart\_atack\_model\_7 | 9 | 196,79 | 0,05 | 0,32 | 0,66 | -88,95 |
| heart\_atack\_model\_4 | 8 | 197,56 | 0,83 | 0,22 | 0,88 | -90,42 |
| heart\_atack\_model\_5 | 10 | 199,98 | 3,25 | 0,07 | 0,94 | -89,44 |
| heart\_atack\_model\_1 | 12 | 200,25 | 3,52 | 0,06 | 1,00 | -87,34 |
| heart\_atack\_model\_3 | 9 | 223,28 | 26,54 | 0,00 | 1,00 | -102,19 |
| heart\_atack\_model\_6 | 5 | 233,59 | 36,86 | 0,00 | 1,00 | -111,65 |

În *tabelul 1* putem observa compararea tuturor modelelor după scorul AIC, disponibil în biblioteca *AICcmodavg.* Cele mai bune rezultate s-au obținut la modul 2 care conține variabilele *age, sex, cp, trtbps, chol, thalachh* și modelul 7 care este compus din următoarele variabile: *sex, cp, trtbps, chol, thalachh.*



**Figura 7.** *Compararea celor două mai bune modele*

În plotul de curbe ROC(*figura 7*), avem două linii reprezentând modelele `*model2*` și `*model7*`. Aceste curbe oferă o imagine vizuală a performanței modelelor în clasificarea datelor în funcție de pragurile de decizie.

Performanța fiecărui model poate fi evaluată prin mai multe aspecte. Curba ROC pentru `*model7*` se apropie mai mult de colțul din stânga sus al graficului decât curba pentru `*model2*`. Aceasta indică faptul că `*model7*` are o performanță globală mai bună în a distinge între clasele pozitive și negative în comparație cu `*model2*`.

În teorie, un model perfect ar avea o curba ROC care să meargă direct la colțul din stânga sus al graficului, având un AUC (Area Under the Curve) de 1. Cu cât curba se apropie mai mult de această perfecțiune, cu atât performanța modelului este mai bună.

În concluzie, bazându-ne pe aceste metrici, `*model7*` pare să fie mai eficient în clasificarea corectă a datelor decât `*model2*`, sugerând o performanță mai robustă în identificarea claselor pozitive și negative. De asemenea, rezultatele și toate materialele sunt plasate pe platforma de control a versiunilor GitHub[4].

# DISCUȚII

Studiul efectuat aduce lumină asupra relațiilor complexe dintre diferite variabile asociate sănătății cardiovasculare și furnizează o înțelegere detaliată a impactului acestora. Analiza vârstei și nivelului de colesterol dezvăluie o variabilitate semnificativă, sugestivă pentru posibile tendințe asociate îmbătrânirii. Această observație are implicații importante pentru implementarea unor strategii personalizate de gestionare a colesterolului, orientate către grupurile de vârstă cu risc crescut.

O altă dimensiune investigată în studiu se referă la relația dintre tensiunea arterială și gen. Cercetările efectuate nu indică diferențe semnificative între sexe.

Analiza tipurilor de durere raportate în funcție de vârstă aduce în discuție necesitatea adaptării abordărilor de gestionare a durerii în diferitele categorii de vârstă, oferind o perspectivă utilă pentru îmbunătățirea calității îngrijirii medicale.

Un aspect deosebit de relevant este identificarea disparităților de gen în ceea ce privește incidența bolilor cardiovasculare. Aceste constatări subliniază importanța unei abordări diferențiate în prevenție și tratament în funcție de sex.

Vulnerabilitățile specifice genului în cazurile de atac de cord, dezvăluite în studiu, accentuează necesitatea unor măsuri preventive personalizate pentru a reduce incidența și gravitatea acestor evenimente în rândul populației.

Cu toate acestea, trebuie menționat că studiul are limite, inclusiv sursa de date și variabilele selectate, ceea ce deschide noi orizonturi în interpretarea rezultatelor. Necesitatea unor cercetări ulterioare este evidentă pentru validarea și aprofundarea constatărilor în acest domeniu vital al cercetării medicale.

Stilul de viață a unei persoane poate multe spune despre el și despre sănătatea lui. Un aspect important sunt factori precum fumatul sau consumarea alcolului în exces. Aceste date nu sunt disponibile în setul de date analizat. De asemenea, activitatea fizică zilnică este un factor care tot nu este în set de date, dar el poate micșora apariției unei boli cardiovasculare. Alimentația tot este importantă. O persoană ce consumă mâncări grase, îngreunează lucrul inimii, astfel fiind mai predispus unui atac de cord.

# BIBLIOGRAFIE

1. Kaggle, platformă online pentru seturi de date, ‘Heart Attack Analysis & Prediction Dataset’ [Accesat:04.12.23], disponibil: <https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset?resource=download>
2. CDC, resursă web a Centrului pentru Prevenirea și Controlul Bolilor, Departamentului de Sănătate Publică al SUA, ‘Heart Disease Facts’ [Accesat:04.12.23], disponibil: <https://www.cdc.gov/heartdisease/facts.htm>
3. RDocumentation, resursă web pentru documentație privitor la RStudio , ‘RDocumentation’ [Accesat:04.12.23], disponibil: <https://www.rdocumentation.org/>
4. GitHub, platformă de control a versiunilor, ‘Heart\_Attack\_Analysis’ [Accesat:18.12.23], disponibil: <https://github.com/Cristalin8/Heart_Attack_Analysis>