***Predicția prezenței afecțiunilor cronice cardiovasculare***

Realizat de Onofrei Anastasia, gr. MI-211

Verificat de Munteanu Viorel, lect. univ.

**Universitatea Tehincă a Moldovei**

**Chișinău 2023**

**ABSTRACT**

INTRODUCERE. Bolile cardiovasculare (BC) reprezintă una dintre principalele cauze de deces la nivel mondial. Acest studiu a urmărit să investigheze cauzele apariției bolilor cardiovasculare. Setul de date este colectat în baza unor indicatori ale pacienților care au fost înregistrați în cadrul unor instituții medicale, și anume, vârsta, sexul, colesterolul, tensiunea arterială în stare de repaus, glicemia, prezența unei boli cronice de inimă și altele. Iar în urma unei analize minuțioase a acestor date se poate de extras informații valoroase și semnificative.

În această lucrare, se va explora importanța și beneficiile analizei exploratorii a datelor, evidențiind metodele și instrumentele cheie utilizate în acest proces. De asemenea, se va ilustra rolul său esențial în construirea unui model de regresie pentru a prezice valoarea țintă.

MATERIALE ȘI METODE. Pentru această lucrare s-a utilizat ca mediu de dezvoltare și limbaj de programare R-ul, care este un instrument puternic în analiza datelor, împreună cu unele pachete ce permit crearea modelelor și vizualizarea rezultatelor. Iar ca metodă fundamentală de analiză a datelor s-a utilizat EDA (Exploratory Data Analysis), după care s-au creat câteva modele de predicție și anume, regresia logistică, arborele de decizie și random forest.

REZULTATE. În urma celor efectuate conform EDA s-au obținut diverse vizualizări ce descriu setul de date, distribuțiile acestora și alte elemente care oferă informații aplicabile în cercetare. S-a constatat faptul că cei mai mulți dintre cei ce manifestă oarecare semne de boli cardiovasculare sunt bărbații, constituind un aport de aproximativ 70% în comparație cu femeile care constituie 30%, iar vârsta care predomină la aceștia este cuprinsă între 50 – 60 de ani.

CONCLUZII. S-a observat faptul că variabilele corelează între ele diferit, ceea ce înseamnă că a fost nevoie de a elimina unele caracteristici sau de a normaliza prin eliminarea outlierilor, ulterior obținând un rezultat mai bun în urma modelelor de predicție. Iar conform rezultatului final de predicție a prezenței afecțiunilor cardiovasculare la noi pacienți se atestă o precizie destul de bună.

**INTRODUCERE**

Bolile cardiovasculare sunt printre principalele cauze de deces în întreaga lume. Un sistem de predicție poate ajuta la identificarea persoanelor cu risc crescut și la intervenții mai timpurii, reducând astfel șansele de a dezvolta evenimente cardiace majore precum infarctul miocardic sau accidentul vascular cerebral.

De asemenea, identificarea și gestionarea precoce a riscului de afecțiuni cardiace poate îmbunătăți semnificativ calitatea vieții pacienților.

Scopul aceste lucrări este de a oferi informații utile pentru diagnosticarea precoce și prevenirea acestor afecțiuni. Aceste predicții se bazează pe analiza datelor relevante despre pacienți, cum ar fi istoricul medical, factorii de risc. Prin utilizarea algoritmilor de învățare automată sau alte tehnici de analiză a datelor, se pot identifica modele și corelații între factorii de risc și probabilitatea apariției bolilor cardiovasculare. Iar pentru aceasta s-au propus următoarele obiective:

* analiza domeniului de studiu, și anume, afecțiuni cardiovasculare, indicatorii corpului uman care indică spre afecțiuni cardiovasculare;
* colectarea unui set de date în conformitate cu tema proiectului;
* aplicarea instrumentelor de analiză a datelor - R;
* crearea modelelor de prezicere a prezenței unei afecțiuni cardiovasculare.

Impactul rezolvării acestei probleme asupra publicului țintă sau a părților interesate va fi unul semnificativ de important în domeniul medical. Acesta se referă la prioritizarea pacienților cu un risc mai mare de afecțiuni cardiovasculare, oferindu-le un tratament adecvat de timpuriu. Respectiv și speranța la viață și calitatea vieții pacienților se va îmbunătăți.

Pentru problema propusă sunt necesari mai mulți indicatori ai diferitor pacienți ca sex și vârstă, ca exemplu: tensiune arterială în repaus, electrocardiograma, glicemia, angina endusă de effort, colesterol, altele și un target (prezența sau neprezența bolilor cardiovasculare). Sursa setului de date este repozitoriul Kaggle. Aceste date au fost colectate în cadrul unor instituții medicale și aparțin unor pacienți reali selectați aleatoriu. Datele sunt în mare parte numerice și categoriale, ele reprezintă niște caracteristici ale pacienților (variabilele independente – x1, x2, …, xn) și variabila care spune dacă este sau nu prezentă o boală cardiovasculară (variabila dependentă - y). Dimensiunea setului de date este de 1025 de înregistrări, 1025 de rânduri și 14 coloane.

În era tehnologică actuală, volumul imens de date generat zilnic constituie o resursă inestimabilă ce ascunde potențialul de a oferi perspective semnificative și informații cruciale pentru luarea deciziilor. Cu toate acestea, abordarea acestor seturi de date necesită instrumente și tehnici adecvate pentru a extrage informații relevante. Una dintre metodele fundamentale în acest proces este analiza exploratorie a datelor (EDA). EDA reprezintă o etapă esențială în procesul de analiză a datelor, prin care se explorează și se înțelege structura, relațiile și distribuția variabilelor din setul de date. Acest proces nu se limitează doar la calculul de statistici descriptive, ci își propune să dezvăluie posibile corelații, tendințe și modele ascunse, oferind astfel o perspectivă detaliată asupra informațiilor conținute în datele analizate.

Un aspect foarte important al EDA îl are de asemenea și agregarea sau segmentarea datelor pentru analiză și vizualizare. Luarea în considerare a modului de grupare a datelor poate avea un impact asupra informațiilor derivate din EDA.

De asemenea este necesară o inginerie sau o transformare a caracteristicilor pentru o analiză mai eficientă. Uneori, crearea de noi variabile sau transformarea variabilelor poate îmbunătăți analiza semnificativ. Este necesară normalizarea sau standardizarea datelor atunci când avem outliers (excepții). Crearea valorilor binare (0,1) simplifică analiza semnificativ. Iar clusterizarea este foarte binevenită pentru o vizualizare mai clară a datelor.

EDA este un proces iterativ și flexibil, este o parte esențială a procesului de analiză a datelor și oferă o fundație solidă pentru dezvoltarea de modele și luarea de decizii în domeniul analizei datelor.

**MATERIALE ȘI METODE**

Pentru această lucrare s-a utilizat ca mediu de dezvoltare și limbaj de programare **R**-ul, care este un instrument puternic în analiza datelor, crearea modelelor și vizualizarea rezultatelor. Iar pentru a efectua toate acestea au fost utilizate câteva pachete, care reprezintă unelte puternice și utile în analiza datelor, vizualizarea lor și construirea și evaluarea modelelor statistice și de învățare automată în R. **Tidyverse** este un set de pachete din R care lucrează împreună pentru a simplifica și eficientiza procesul de lucru cu datele. Include pachete precum ggplot2, dplyr, tidyr și altele, promovând un stil de lucru coerent și asociat cu datele. **Dplyr** este un pachet esențial pentru manipularea datelor în R. Furnizează funcții concise și eficiente pentru filtrare, selectare, rearanjare și sintetizare a datelor. **Ggplot2** este un pachet pentru crearea de grafice în R, bazat pe filozofia "grammar of graphics". Oferă o abordare puternică și flexibilă pentru realizarea diverselor tipuri de grafice. Iar pentru vizualizarea matricilor de corelație într-un mod simplu și atractiv este utilizat **corplot**. Ajută la înțelegerea relațiilor dintre variabilele dintr-un set de date. **pROC** este utilizat pentru analiza și vizualizarea performanței modelelor de diagnostic sau de clasificare, oferind unelte pentru curbele ROC (Receiver Operating Characteristic). Pentru construirea și evaluarea modelelor de învățare automată s-a utilizat **caret**. Acesta simplifică procesul de ajustare a modelelor și de selecție a parametrilor.Top of Form

Iar ca metodă fundamentală de analiză a datelor s-a utilizat EDA, după care s-au creat câteva modele de predicție și anume, regresia logistică, arborele de decizie și random forest.

Analiza exploratorie a datelor se referă la procesul de investigare a unui set de date pentru a extrage informații utile, recunoaște tipare, identifica relații și detecta anomaliile. Scopul principal al EDA este de a înțelege mai bine datele, pentru a face apoi analize și luări de decizie mai informate.

Procesul de EDA implică de obicei următoarele etape:

* **Sumarizarea datelor**

Acest pas implică înțelegerea cantității de date, identificarea coloanelor (variabilelor) și observarea primelor câteva rânduri din setul de date. Acest studiu a inclus 1025 de pacienți care au fost înregistrați în cadrul unor instituții medicale și selectați aleatoriu. Datele sunt în mare parte numerice și categoriale, ele reprezintă niște caracteristici ale pacienților (variabilele independente – x1, x2, …, xn) și variabila care spune dacă este sau nu prezentă o boală cardiovasculară (variabila dependentă - y). Dimensiunea setului de date este de 1025 de înregistrări, 1025 de rânduri și 14 coloane. Aceste 14 coloane include caracteristici precum vârsta, sexul, colesterolul, glicemia, ritmul cardiac maxim atins, alte caractieristici ce țin de identificarea unei afecțiuni cardiovasculare precum și targetul, care definește dacă pacientul care o afecțiune cardiacă sau nu.

* **Curățarea datelor**

Acest pas implică tratamentul datelor lipsă, gestionarea valorilor aberante (outliers) și corectarea altor probleme legate de calitatea datelor. În acest set de date valori lipsă nu s-au depistat, însă s-au găsit câteva înregistrări cu valori outliers pentru colesterol, repectiv aceste valori s-au înlocuit cu valoarea medianei. Pentru a vizualiza valorile oaberente s-a utilizat biblioteca ggplot2, și anume, scatterplot și în special boxplot, unde se văd clar valorile excepții, iar în cazul în care micșorăm eșantionul de valori în plot pot dispărea unele excepții și apărea altele.

* **Vizualizarea datelor**

Acesta este un pas crucial în EDA. Include crearea de grafice, diagrame și alte vizualizări pentru a reprezenta datele într-un mod ușor de înțeles și interpretabil. Aceste vizualizări pot ajuta la identificarea tiparelor și relațiilor în date. În această lucrare s-a utilizat pachetul ggplot2 pentru a vizualiza variabilele numerice sub formă de barplot, boxplot, histogramă, scatterplot, diagramă de dispersie, iar pentru cele categoriale s-a utilizat mai mult diagramele cu bare și boxploturi comparative.

* **Analiza statistică**

La această etapă s-au calculat unele măsuri statistice de bază, cum ar fi medii, mediane, deviații standard etc. Aceste măsuri pot oferi o înțelegere mai profundă a distribuției datelor. Ca rezumate statistice s-au utilizat mărimi ale tendinței centrale (media, mediana), măsuri ale dispersiei pentru valori numerice. Iar pentru cele categorice s-au utilizat ploturi care ilustrează frecvența, s-au creat tabele de cotingență și de asemenea diagrame de bare. Iar cele mai potrivite metode și modele statistice care au fost luate în considerare și utilizate pentru datele și obiectivele proiectului au fost analiza clusterizată pentru a categoriza unele variabile și regresia logistică care se va aplica pentru a face predicția bolilor de inimă pentru o nouă înregistrare.

* **Explorarea relațiilor între variabile**

Aici se examinează cum variabilele interacționează între ele. Acest lucru s-a cercetat prin calcularea de corelații între variabile și construirea de grafice care evidențiază relațiile.

* **Formularea ipotezelor**

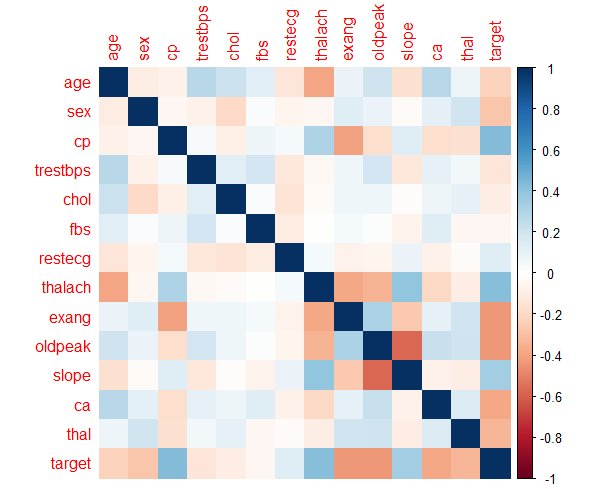
Pe baza observațiilor făcute în timpul EDA, se pot formula ipoteze care pot fi ulterior testate folosind tehnici statistice mai avansate. Iar aceasta este o ipoteză care pe parcursul cercetării se demonstrează că alternativa ei este corectă.

*Ipoteza nulă (H0):*Nu există o corelație semnificativă între ritmul cardiac maxim atins și riscul de afecțiuni cardiace (target).

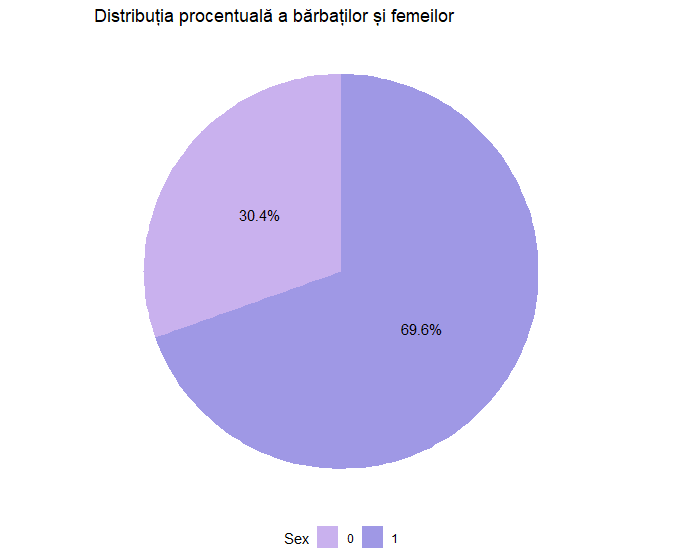
*Ipoteza alternativă (H1):*Există o corelație semnificativă între ritmul cardiac maxim atins și riscul de afecțiuni cardiace (target).

**REZULTATE**

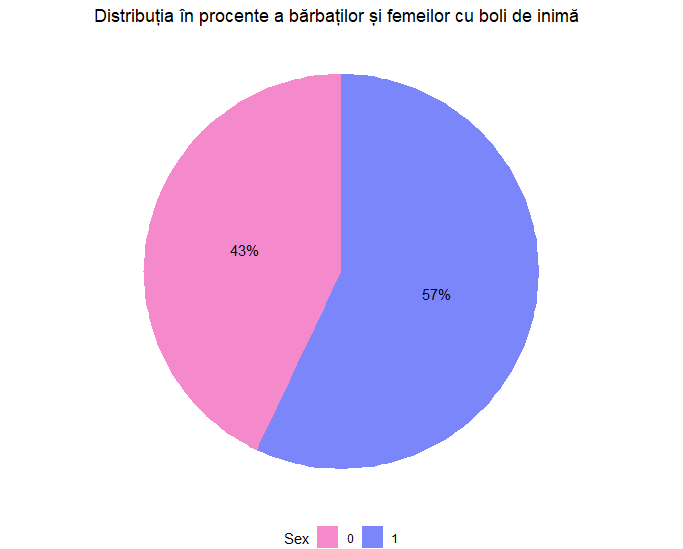
O parte esențială a EDA este identificarea corelațiilor potențiale sau a relațiilor de cauzalitate. Respectiv, s-a utilizat correlation heatmap pentru a vizualiza relațiile între variabile (în ce măsură depinde o variabilă de altă variabilă).

 **Figura 1.1.** Correlation heatmap

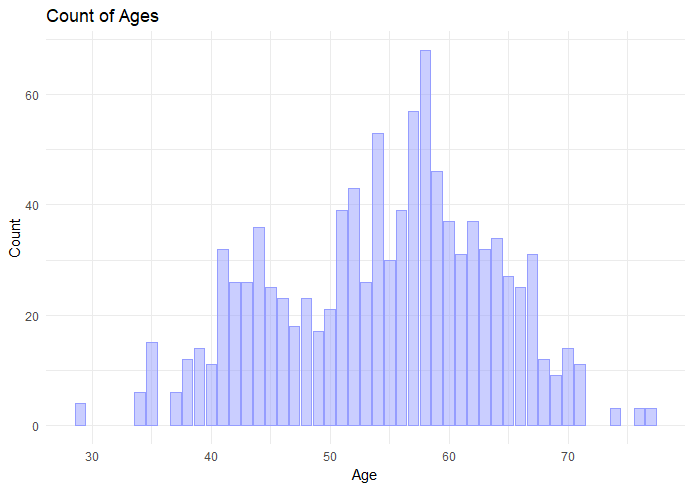
S-a observat în figura 1.1 că este o corelație bună între următoarele variabile: prezența bolilor de inimă – tip de durere în piept, vârstă – tensiune arterială, ritm cardiac maxim atins - prezența bolilor de inimă, tip de durere în piept - prezența bolilor de inimă.

**  
Figura 1.2.** Distribuția pe sexe a pacienților

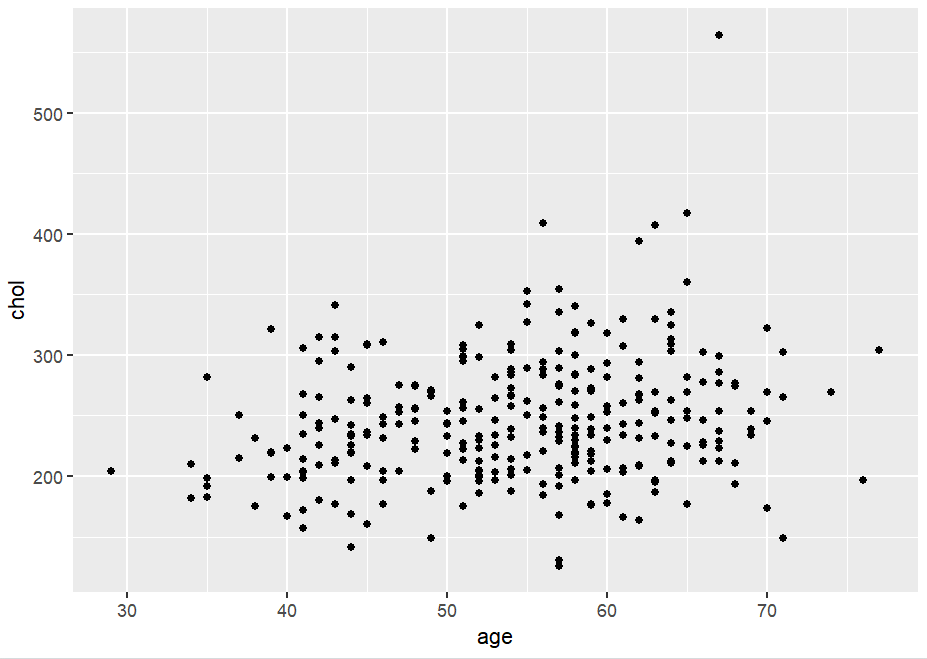
În figura 1.2 se observă faptul că majoritatea pacienților care manifestă oarecare semne de boli cardiovasculare sunt bărbați, având un aport de 70% comparativ cu femeile care constitue 30% doar.

**  
Figura 1.3.** Distribuția pe sexe a pacienților cu boli cardiovasculare

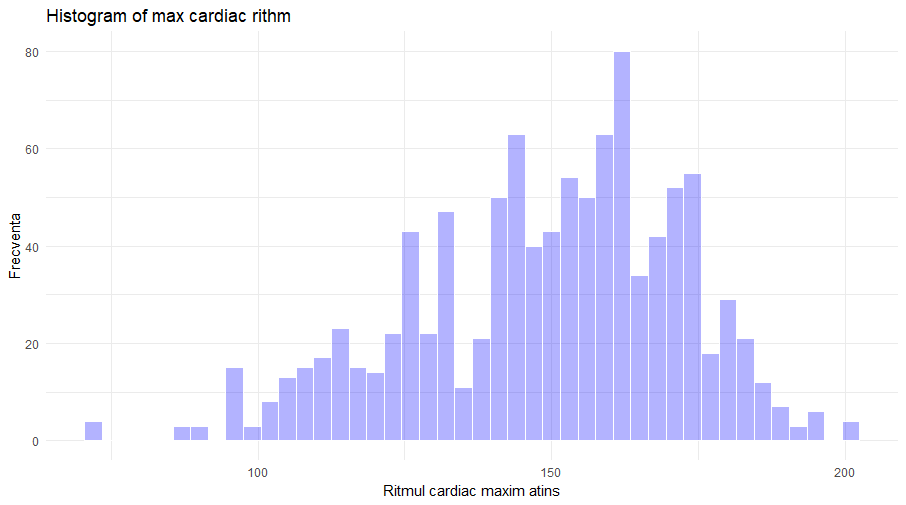
Iar în figura 1.3 se observă un aport aproape egal în ce privește distribuția pe sexe a pacienților cu boli cardiovasculare, partea masculină depășind doar cu 14 % femeile. Aceasta denotă faptul că atât bărbații cât și femeile sunt dispuși la afecțiuni cardiovasculare.

**  
Figura 1.4.** Vârstele pacienților

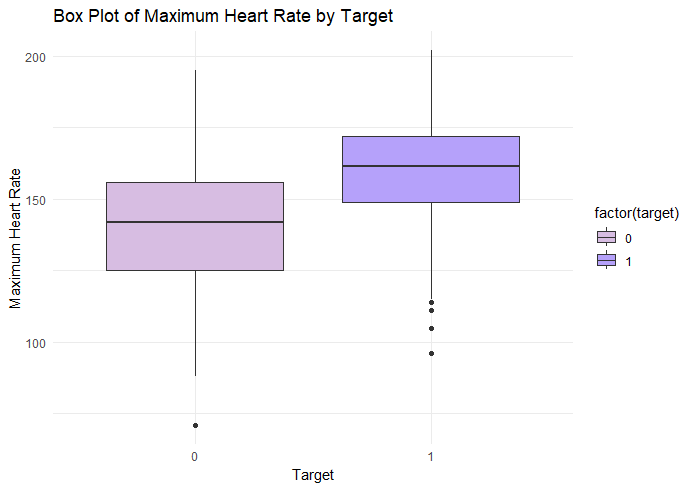
În figura 1.4 este prezentată distribuția vârstelor pacienților, iar cei mai mulți se centrează în vârstele cuprinse între 50 și 60 de ani.

**  
Figura 1.5.** Prezența colesterolului pentru toate vârstele

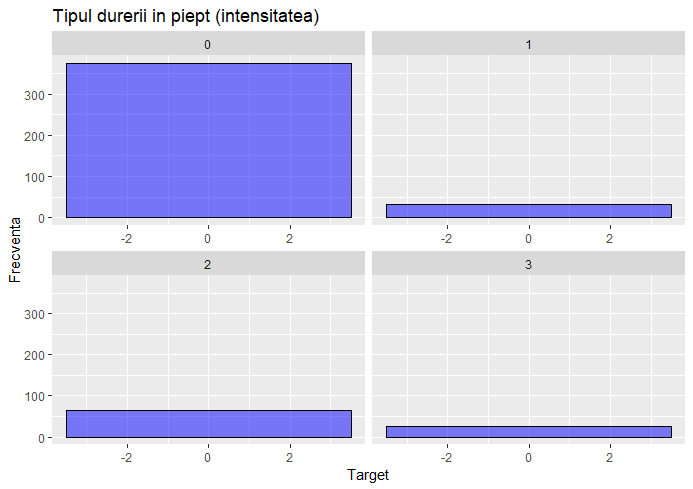
În figura 1.5 este prezentată distribuția colesterolului pentru toate vârstele și se observă că avem valori outlieri cu valori de peste 550 de mg/dl, care ulterior s-au normalizat.

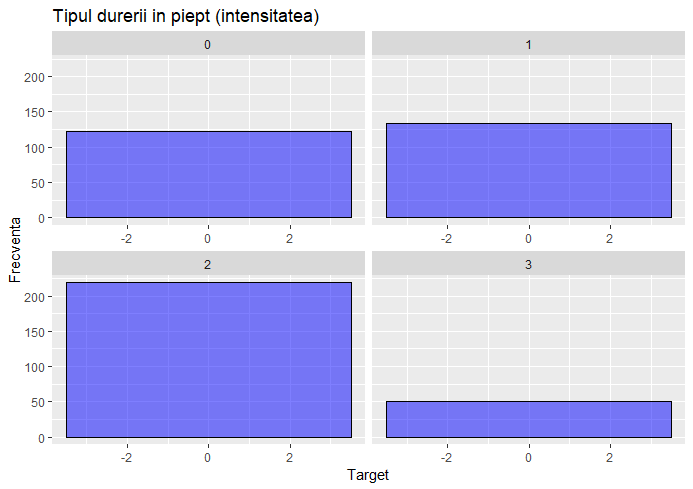
**  
Figura 1.6.** Frecvența ritmului cardiac maxim atins

În această figură se observă faptul că frecvența ritmului cardiac maxim atins se centrează pe valoarea de 160 de bătăi pe minut, iar majoritatea valorilor maxime cele mai frecvente sunt în diapazonul 140 – 160 de bătăi pe minut.

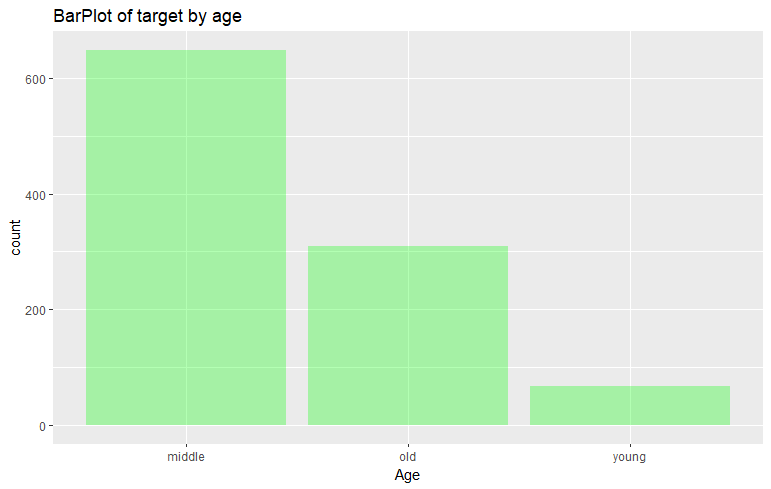
  
**Figura 1.7.** Ritmul cardiac maxim atins în dependență de prezența afecțiunilor cardiovasculare

În diagrama de mai sus (figura 1.7) sunt reprezentate comparativ valorile ritmului cardiac maxim atins pentru cei ce nu suferă de o boală cardiovasculară și cei ce suferă de una, iar în rezultat se vede faptul că cei ce suferă au un ritm cardiac mai mărit, desigur cu unele excepții.

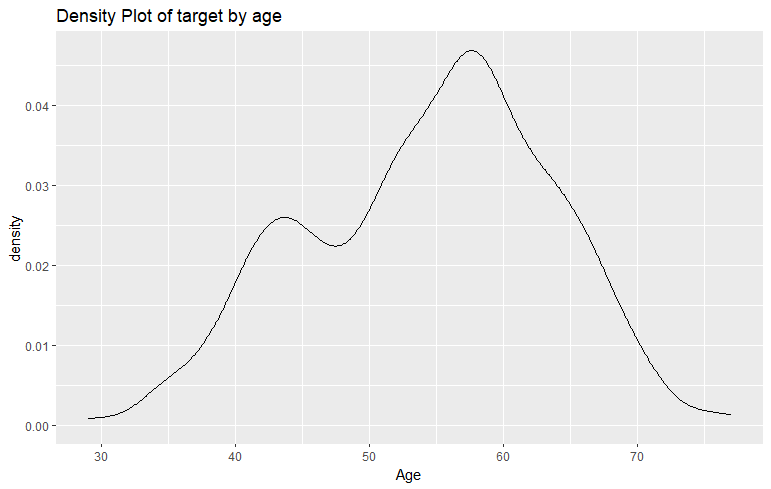
  
**Figura 1.8.** Frecvența neprezenței bolilor de inimă în dependență de categorii de dureri în piept

  
**Figura 1.9.** Frecvența prezenței bolilor de inimă în dependență de categorii de dureri în piept

Aici avem două diagrame cu bare în care, de asemenea, s-a făcut comparația în dependență de prezența bolilor cardiovasculare. În figura 1.8 este reprezentată frecvența neprezenței bolilor cardiovasculare în dependență de tipul durerii în piept care se împarte în 4 categorii (0,1,2,3) și se observă faptul că cei care nu suferă de o boală de inimă practic nici nu manifestă vreo durere în piept. Iar în figura 1.9 se observă faptul că acei pacienți care suferă de o boală de inimă manifestă și dureri în piept de tipul 1 și 2, mai puțin ultima categorie și de asemenea au rămas pacienți care nu manifestă dureri.

  
**Figura 1.10.** Prezența bolilor de inimă în dependență de categorie de vârstă

În figura 1.10 este reprezentată distribuția pacienților ce suferă de o boală de inimă în dependență de categorie de vârstă sub forma unui barplot, care arată că cei mai mulți pacienți sunt de vârstă medie. Iar mai jos, pentru o vizualizarea mai clară, este reprezentată aceiași informație în dependență de vârstă sub forma unui density plot, care denotă faptul că cei mai mulți dintre aceștia au vârsta cuprinsă între 50 și 60 de ani.

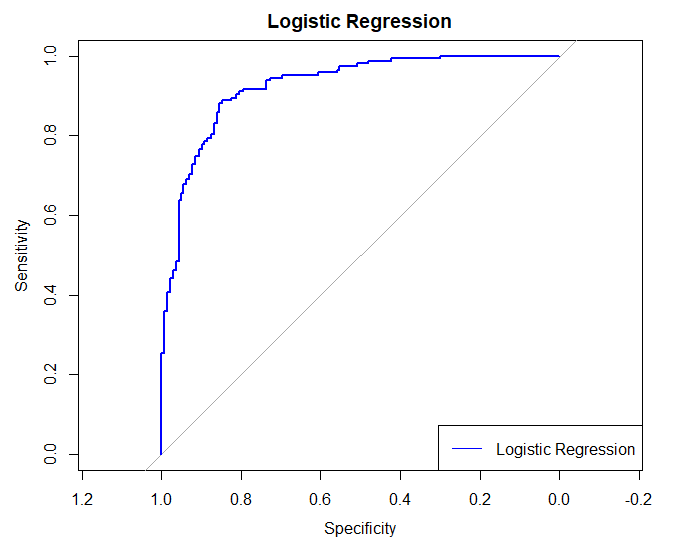
 **Figura 1.11.** Prezența bolilor de inimă în dependență de vârstă

REGRESIA LOGISTICĂ

Regresia logistică este o tehnică statistică utilizată pentru modelarea relației între variabilele independente și o variabilă dependentă binară.



În acest caz, modelul nostru a furnizat o acuratețe de aproape 86%, ceea ce indică că acesta a identificat și capturat asocierea între variabilele independente și probabilitatea de apartenență la una dintre cele două categorii ale variabilei dependente, cu o precizie semnificativă. Acest rezultat indică faptul că modelul are o bună capacitate de predictivitate pentru datele folosite în evaluarea sa.

  
**Figura 1.12.** Curba ROC pentru regresia logistică

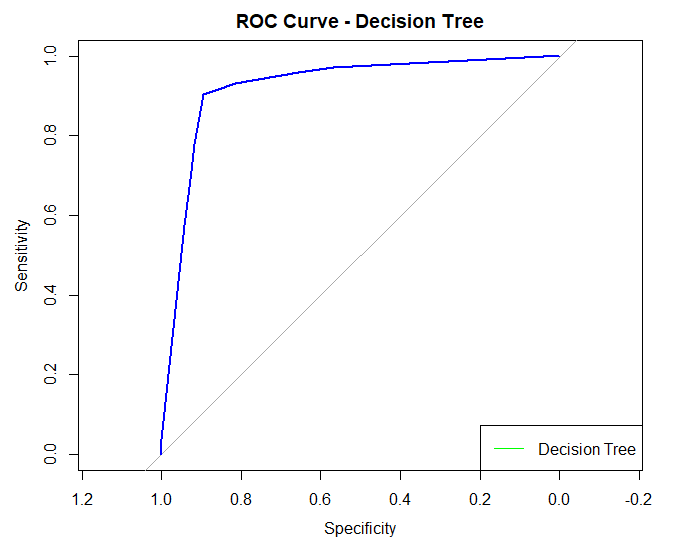
Acuratețea de 86% sugerează faptul că modelul logistic a reușit să distingă și să prezică corect în marea majoritate a cazurilor apartenența la fiecare categorie a variabilei dependente, pe baza informațiilor furnizate de variabilele independente selectate pentru model. Este important să se continue evaluarea și îmbunătățirea modelului pentru a se asigura că acesta se comportă la fel de bine pe oricare date noi, necunoscute anterior.

ARBORELE DE DECIZIE

Arborele de decizie este o metodă de învățare automată utilizată pentru clasificare și regresie. În cazul acestei lucrări, modelul de arbore de decizie a obținut o acuratețe de aproape 90%, ceea ce indică o precizie remarcabilă în clasificarea datelor. Acest rezultat sugerează că modelul a putut captura și utiliza cu succes informațiile din variabilele independente pentru a face predicții precise pentru variabila dependentă.



Cu o acuratețe de 90%, arborele de decizie a demonstrat o bună capacitate de înțelegere a relațiilor dintre variabile și de generare a unor reguli clare de decizie, permitând clasificarea corectă a majorității datelor. Continuarea evaluării și validării acestui model poate contribui la asigurarea generalizării bune a rezultatelor sale pe date noi.

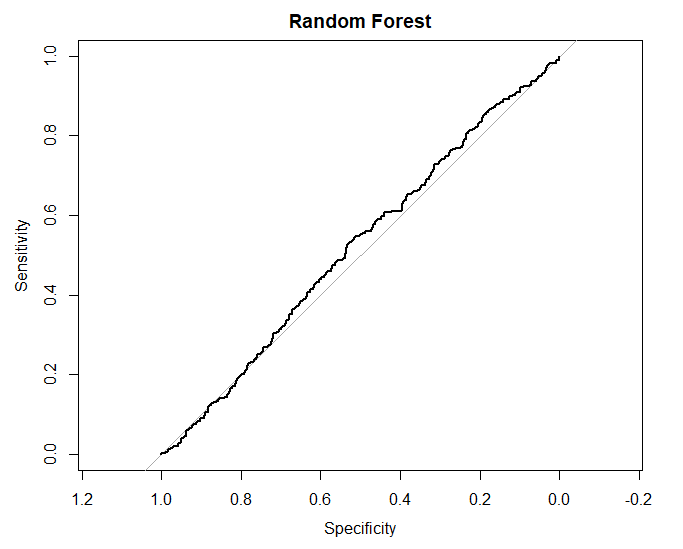
  
**Figura 1.13.** Curba ROC pentru arborele de decizie

RANDOM FOREST

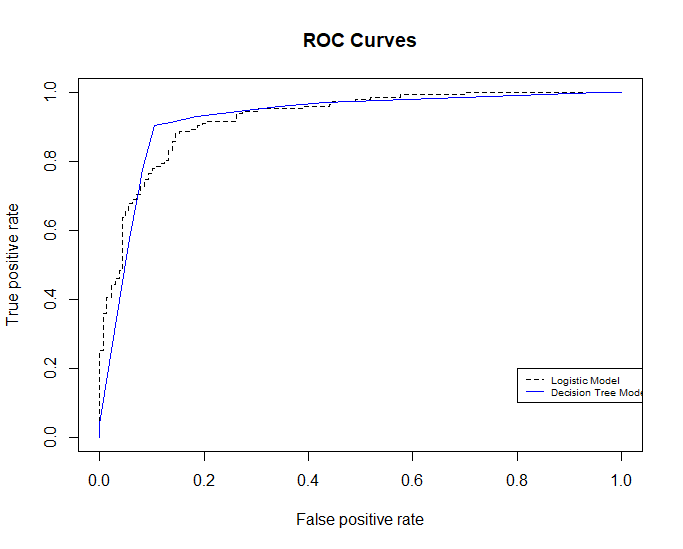
Metoda Random Forest este o tehnică de învățare automată ce utilizează un ansamblu de arbori de decizie.



În cazul acesta, modelul bazat pe Random Forest a obținut o acuratețe de 56%, ceea ce indică o performanță nu foarte bună în clasificarea datelor, sugerând că acesta nu a reușit să prezică cu precizie mare categoriile variabilei dependente pe baza informațiilor din variabilele independente selectate.

  
**Figura 1.14.** Curba ROC pentru Random Forest

Acuratețea de 56% indică o performanță limitată a modelului Random Forest în ceea ce privește clasificarea corectă a datelor. Sunt necesare ajustări suplimentare ale modelului, însă e deja evident faptul că regresia logistică și arborele de decizie rămân a fi cele mai precise.

  
**Figura 1.15.** Curbele ROC comparative

În figura 1.15 este reprezentat grafic comparația între Regresia logistica și Arborele de decizie și se observă faptul că arborele de decizie are acuratețea și precizia ce mai mare în cazul acestei cercetări.

**CONCLUZII**

În urma acestei lucrări s-au utilizat mai multe biblioteci a limbajului R, cu ajutorul cărora s-au analizat datele, s-au verificat dependeța unei variabile de alta, s-au vizualizat rezultatele obținute, ajustând setul de date unde a fost nevoie. S-a observat că variabilele corelează între ele, ceea ce denotă faptul că este o legătură între anumite variabilele și că este posibil de făcut o predicție a riscului afecțiunilor cardiovasculare la noi pacienți. Iar în urma cercetării s-a depistat modelul Decision Tree ca fiind modelul cu precizia ce mai înaltă pentru datele din această lucrare. Ulterior modelul s-a testat și pe date noi, acesta demonstrând o înțelegere corectă a datelor.

Crearea unui model este un proces iterativ, iar o concluzie importantă poate fi că acesta reprezintă doar un punct de plecare. Înțelegerea datelor, interpretarea rezultatelor și continuarea optimizării sunt etapele cheie în evoluția și îmbunătățirea performanței modelului.

**BIBLIOGRAFIE**

1. Sursa setului de date: <https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset/versions/2?resource=download>

2. Laurie Alexander, U.S. EPA, ORD, NCEA, Washington, DC, 10.08.23 - Exploratory Data Analysis - <https://www.epa.gov/caddis-vol4/exploratory-data-analysis>