# Prezicere boli de inima utilizand algoritmi ML

## Intorducere

Am folosit [HEART DISEASE DATASET (COMPREHENSIVE)](https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive) de la IEEE Dataport. Setul de date este organizat prin combinarea a 5 seturi de date populare despre boli de inimă deja disponibile independent, dar necombinate înainte. În acest set de date, 5 seturi de date despre inimă sunt combinate peste 11 caracteristici comune, ceea ce îl face cel mai mare set de date despre boli de inimă disponibil până acum în scopuri de cercetare. Cele cinci seturi de date utilizate sunt:

1. Cleveland
2. Hungarian
3. Switzerland
4. Long Beach VA
5. Statlog (Heart) Data Set.

Setul de date este format din 1190 de interogari cu 11 caracteristici. Aceste seturi de date au fost colectate și combinate într-un singur loc pentru a ajuta la avansarea cercetării privind învățarea automată și algoritmii de extragere a datelor legate de CAD (Coronary artery disease).

## Baza de date

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.No.** | **Attribute** | **Code given** | **Unit** | **Data type** |
| 1 | age | Age | in years | Numeric |
| 2 | sex | Sex | 1, 0 | Binary |
| 3 | chest pain type | chest pain type | 1,2,3,4 | Nominal |
| 4 | resting blood pressure | resting bp s | in mm Hg | Numeric |
| 5 | serum cholesterol | cholesterol | in mg/dl | Numeric |
| 6 | fasting blood sugar | fasting blood sugar | 1,0 > 120 mg/dl | Binary |
| 7 | resting electrocardiogram results | resting ecg | 0,1,2 | Nominal |
| 8 | maximum heart rate achieved | max heart rate | 71–202 | Numeric |
| 9 | exercise induced angina | exercise angina | 0,1 | Binary |
| 10 | oldpeak =ST | oldpeak | depression | Numeric |
| 11 | the slope of the peak exercise ST segment | ST slope | 0,1,2 | Nominal |
| 12 | class | target | 0,1 | Binary |

|  |  |
| --- | --- |
| **Attribute** | **Description** |
| Sex | 1 = male, 0= female; |
| Chest Pain Type | -- Value 1: typical angina  -- Value 2: atypical angina  -- Value 3: non-anginal pain  -- Value 4: asymptomatic |
| Fasting Blood sugar | (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false) |
| Resting electrocardiogram results | -- Value 0: normal  -- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)  -- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria |
| Exercise induced  angina | 1 = yes; 0 = no |
| the slope of the peak exercise ST segment | -- Value 1: upsloping  -- Value 2: flat  -- Value 3: downsloping |
| class | 1 = heart disease, 0 = Normal |

## Aspecte teoretice

Domeniul bolilor de inimă este vast și în continuă evoluție, cu noi descoperiri și tratamente apărute în mod regulat. Iată o privire de ansamblu asupra stării actuale a domeniului, cu 10 referințe științifice cheie:

**1. Factori de risc:**

* **Hipertensiune arterială:** Hipertensiunea arterială rămâne principalul factor de risc pentru bolile cardiovasculare la nivel global.
* **Colesterol ridicat:** LDL ("colesterolul rău") este un alt factor de risc major pentru bolile de inimă.
* **Diabet:** Diabetul de tip 2 este un factor de risc semnificativ pentru bolile de inimă.
* **Fumat:** Fumatul este un factor de risc major pentru bolile cardiovasculare și bolile pulmonare.
* **Obezitate:** Obezitatea este asociată cu un risc crescut de boli de inimă, accident vascular cerebral și alte probleme de sănătate.

**2. Prevenție:**

* **Modificări ale stilului de viață:** Modificările stilului de viață, cum ar fi adoptarea unei diete sănătoase, creșterea activității fizice, renunțarea la fumat și gestionarea stresului, pot reduce semnificativ riscul de boli de inimă.
* **Medicamente:** Există o serie de medicamente care pot fi utilizate pentru a preveni bolile de inimă, cum ar fi statinele pentru a scădea colesterolul, medicamentele pentru tensiunea arterială și aspirina.

**3. Diagnostic:**

* **Electrocardiogramă (EKG):** Un EKG este un test care înregistrează activitatea electrică a inimii. Poate fi utilizat pentru a diagnostica bolile de inimă, cum ar fi aritmiile.
* **Ecocardiogramă:** O ecocardiogramă este o ultrasunete a inimii. Poate fi utilizată pentru a evalua funcția cardiacă și pentru a detecta probleme cardiace, cum ar fi valvele cardiace deteriorate.
* **Test de stres cardiac:** Un test de stres cardiac este un test care evaluează modul în care inima răspunde la efort sau la anumite medicamente. Poate fi utilizat pentru a diagnostica bolile arterelor coronare.

**4. Tratament:**

* **Angioplastie și stent:** Angioplastia este o procedură care utilizează un balon pentru a deschide arterele coronare blocate. Un stent poate fi apoi implantat pentru a menține artera deschisă. [11]
* **Bypass coronarian:** Bypass-ul coronarian este o operație care creează o cale nouă pentru ca sângele să curgă către inimă, ocolind arterele coronare blocate.
* **Medicamente:** Există o serie de medicamente care pot fi utilizate pentru a trata bolile de inimă, cum ar fi medicamente pentru a scădea tensiunea arterială, medicamente pentru a reduce colesterolul și medicamente pentru a trata insuficiența cardiacă.

**5. Prognoză:**

* Prognosticul pentru bolile de inimă depinde de o serie de factori, inclusiv tipul de boală cardiacă, severitatea bolii și starea generală de sănătate a persoanei.
* Cu tratament, mulți oameni cu boli de inimă pot trăi o viață lungă și sănătoasă.

**Referinte:**

1. <https://www.heart.org/en/health-topics/high-blood-pressure>
2. <https://www.nhlbi.nih.gov/health/blood-cholesterol>
3. <https://www.cdc.gov/diabetes/diabetes-complications/diabetes-and-your-heart.html>
4. <https://www.cdc.gov/tobacco/about/cigarettes-and-cardiovascular-disease.html>
5. <https://www.cdc.gov/nccdphp/index.html>
6. <https://www.cdc.gov/heart-disease/about/index.html>
7. <https://www.mayoclinic.org/diseases-conditions/heart-disease/in-depth/heart-disease-prevention/art-20046502>
8. <https://www.mayoclinic.org/tests-procedures/ekg/about/pac-20384983>

Logistic Regression (Regresia logistică) este un algoritm statistic utilizat pe scară largă în domeniul învățării automate pentru clasificare. Spre deosebire de regresia liniară care prezice valori continue, regresia logistică se concentrează pe predicția probabilității ca un anumit eveniment să aparțină uneia dintre două categorii.

Iată o descompunere a modului în care funcționează:

Scenariu: Imaginați-vă că încercați să preziceți dacă un pacient are sau nu boli de inimă pe baza factorilor de risc precum vârsta, tensiunea arterială și colesterolul.

Date de intrare: Regresia logistică ia în considerare valorile variabilelor independente (vârsta, tensiunea arterială, colesterolul) pentru fiecare pacient.

Model: Algoritmul învață un model matematic care ia în considerare aceste variabile și produce o valoare cuprinsă între 0 și 1. Această valoare reprezintă probabilitatea ca pacientul să aibă boala de inimă.

Decision tree (arbore de decizie) este un instrument de învățare automată utilizat pe scară largă pentru **atât sarcini de clasificare, cât și de regresie**. Este o structură ierarhică asemănătoare unui copac, care modelează o serie de decizii și consecințele lor posibile.

Arborii de decizie pot fi utilizați ca instrument valoros în domeniul bolilor de inimă pentru a:

1. Clasificarea pacienților cu risc:

Factori de risc: Un arbore de decizie poate fi antrenat pe baza datelor medicale ale pacienților, inclusiv factori de risc cunoscuți precum vârsta, tensiunea arterială, colesterolul, diabetul și fumatul.

Predicția probabilității: Arborele poate fi utilizat pentru a prezice probabilitatea ca un pacient individual să dezvolte boli de inimă, oferind o estimare personalizată a riscului.

Identificarea grupurilor de risc: Arborele poate ajuta la identificarea grupurilor de pacienți cu caracteristici similare care prezintă un risc mai mare de boli de inimă, permițând o intervenție preventivă țintită.

2. Asistarea în luarea deciziilor medicale:

Selectarea testelor diagnostice: Un arbore de decizie poate fi utilizat pentru a sugera teste diagnostice specifice pe baza factorilor de risc și a simptomelor pacientului, optimizând eficiența diagnosticului.

Recomandări de tratament: Arborele poate fi instruit să ia în considerare riscul individual al pacientului, comorbiditățile și severitatea bolii pentru a recomanda opțiuni de tratament adecvate, cum ar fi medicamente, modificări ale stilului de viață sau proceduri intervenționale.

3. Îmbunătățirea cercetării bolilor de inimă:

Identificarea factorilor de risc noi: Arborii de decizie pot fi utilizați pentru a analiza seturi mari de date medicale pentru a descoperi noi factori de risc asociați cu bolile de inimă.

Înțelegerea relațiilor complexe: Arborele poate ajuta la dezvăluirea relațiilor complexe dintre factorii de risc, oferind o perspectivă mai profundă asupra modului în care aceștia contribuie la dezvoltarea bolilor de inimă.

Dezvoltarea de strategii preventive mai eficiente: Informațiile obținute din arborii de decizie pot fi utilizate pentru a dezvolta strategii preventive mai eficiente, adaptate nevoilor specifice ale diferitelor grupuri de pacienți.

Exemplu:

Un arbore de decizie ar putea fi utilizat pentru a evalua riscul individual al unui pacient de a dezvolta boli de inimă coronariene (CHD). Arborele ar lua în considerare factori de risc precum vârsta, sexul, tensiunea arterială, colesterolul, diabetul și fumatul. Pe baza valorilor acestor factori, arborele ar prezice probabilitatea ca pacientul să dezvolte CHD și ar putea recomanda teste diagnostice sau intervenții preventive adecvate.

Random Forest (Pădurea aleatorie) este un algoritm de învățare automată de tip **ansamblu** utilizat pe scară largă pentru **atât sarcini de clasificare, cât și de regresie**. Este formată din numeroși **arbori de decizie** antrenați independent, combinându-i predicțiile pentru a obține o performanță mai robustă și mai precisă.

Pădurea Aleatorie poate fi utilizată ca instrument valoros în domeniul bolilor de inimă pentru a:

1. Clasificarea pacienților cu risc:

Factori de risc: O Pădure Aleatorie poate fi antrenată pe baza datelor medicale ale pacienților, incluzând factori de risc cunoscuți precum vârsta, tensiunea arterială, colesterolul, diabetul și fumatul.

Predicția probabilității: Modelul poate fi utilizat pentru a prezice probabilitatea ca un pacient individual să dezvolte boli de inimă, oferind o estimare personalizată a riscului.

Identificarea grupurilor de risc: Pădurea Aleatorie poate ajuta la identificarea grupurilor de pacienți cu caracteristici similare care prezintă un risc mai mare de boli de inimă, permițând o intervenție preventivă țintită.

2. Asistarea în luarea deciziilor medicale:

Selectarea testelor diagnostice: O Pădure Aleatorie poate fi utilizată pentru a sugera teste diagnostice specifice pe baza factorilor de risc și a simptomelor pacientului, optimizând eficiența diagnosticului.

Recomandări de tratament: Modelul poate fi instruit să ia în considerare riscul individual al pacientului, comorbiditățile și severitatea bolii pentru a recomanda opțiuni de tratament adecvate, cum ar fi medicamente, modificări ale stilului de viață sau proceduri intervenționale.

3. Îmbunătățirea cercetării bolilor de inimă:

Identificarea factorilor de risc noi: Pădurea Aleatorie poate fi utilizată pentru a analiza seturi mari de date medicale pentru a descoperi noi factori de risc asociați cu bolile de inimă.

Înțelegerea relațiilor complexe: Modelul poate ajuta la dezvăluirea relațiilor complexe dintre factorii de risc, oferind o perspectivă mai profundă asupra modului în care aceștia contribuie la dezvoltarea bolilor de inimă.

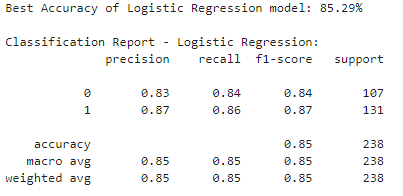
Dezvoltarea de strategii preventive mai eficiente: Informațiile obținute din Pădurea Aleatorie pot fi utilizate pentru a dezvolta strategii preventive mai eficiente, adaptate nevoilor specifice ale diferitelor grupuri de pacienți.

Exemplu:

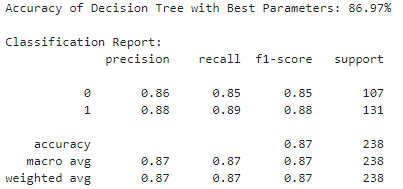
O Pădure Aleatorie ar putea fi utilizată pentru a evalua riscul individual al unui pacient de a dezvolta boli de inimă coronariene (CHD). Modelul ar lua în considerare factori de risc precum vârsta, sexul, tensiunea arterială, colesterolul, diabetul și fumatul. Pe baza valorilor acestor factori, Pădurea Aleatorie ar prezice probabilitatea ca pacientul să dezvolte CHD și ar putea recomanda teste diagnostice sau intervenții preventive adecvate.

## Rezultate:

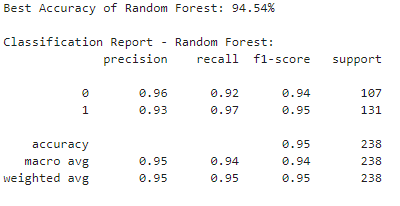
### Logistic regression



### Decision Tree



### Random forest



## Comparații:

