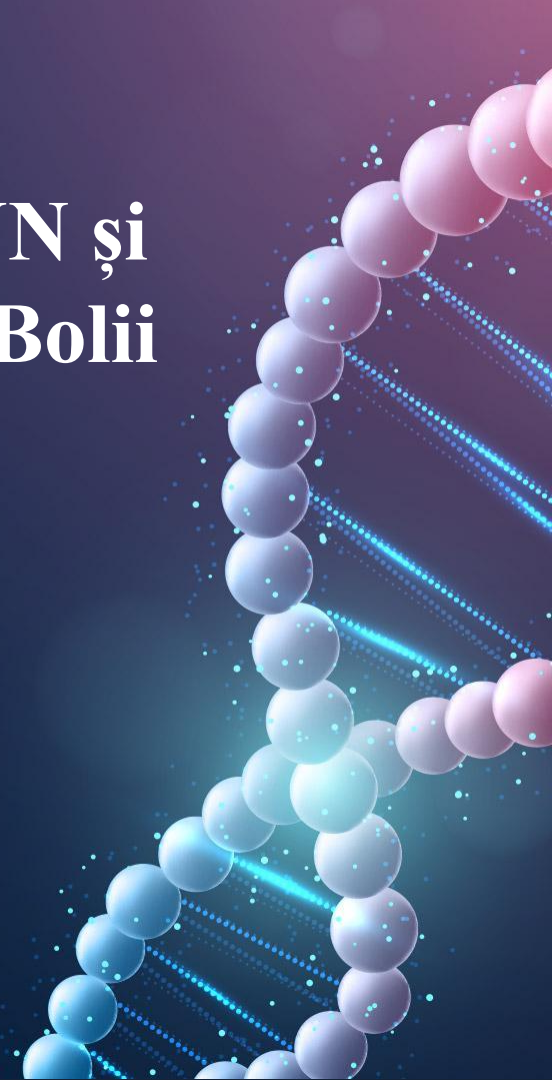
Compararea Metodelor K-NN și Random Forest în Predicția Bolii Cardiace

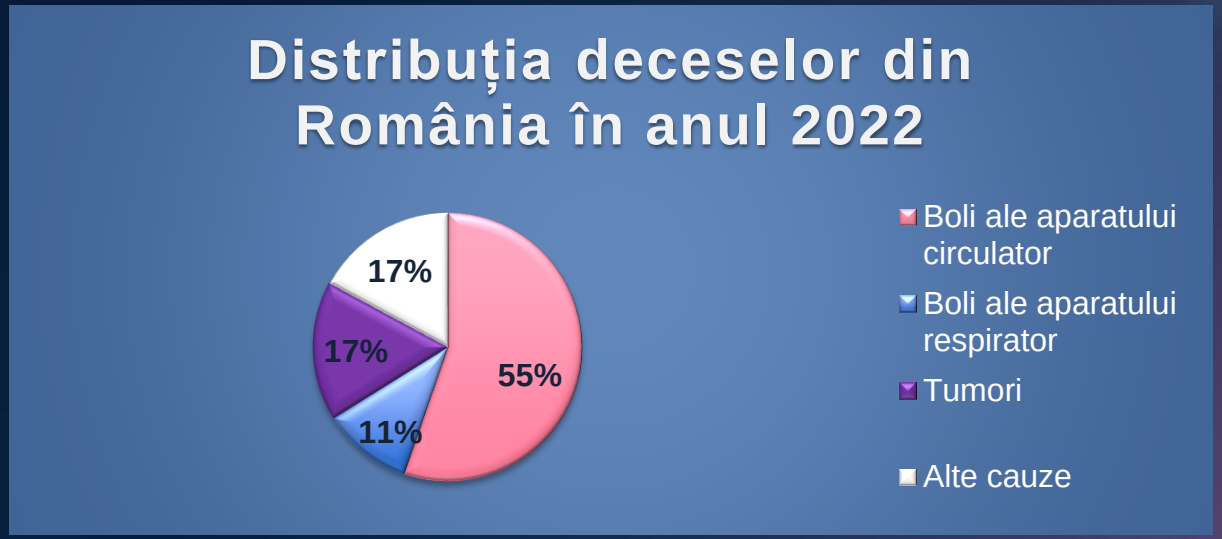


Studenta: Munteanu Maria-Daniela
Grupa: SCI 1B

Cuprins

I. Descrierea problemei propuse si a setului de date

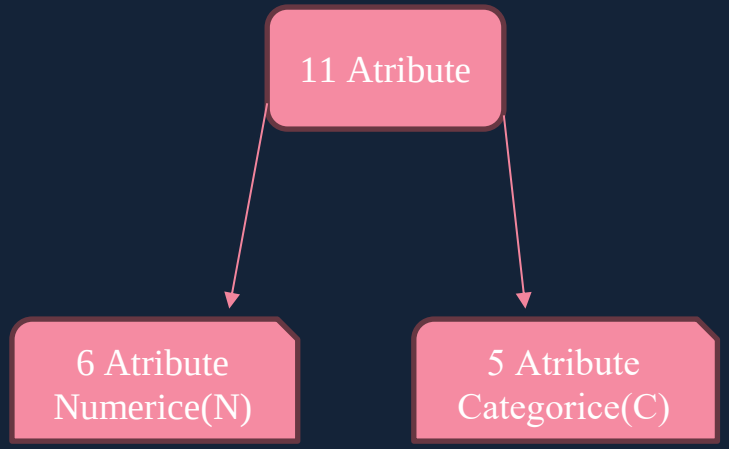
Bolile cardiovasculare reprezintă **una dintre principalele cauze de mortalitate** la nivel global. Un diagnostic **corect și precoce** poate îmbunătăți tratamentul și reduce riscurile. Scopul proiectului este să folosim **învățarea automată** pentru a **predice prezența bolii cardiace** pe baza unor variabile medicale.



I. Descrierea problemei propuse si a setului de date

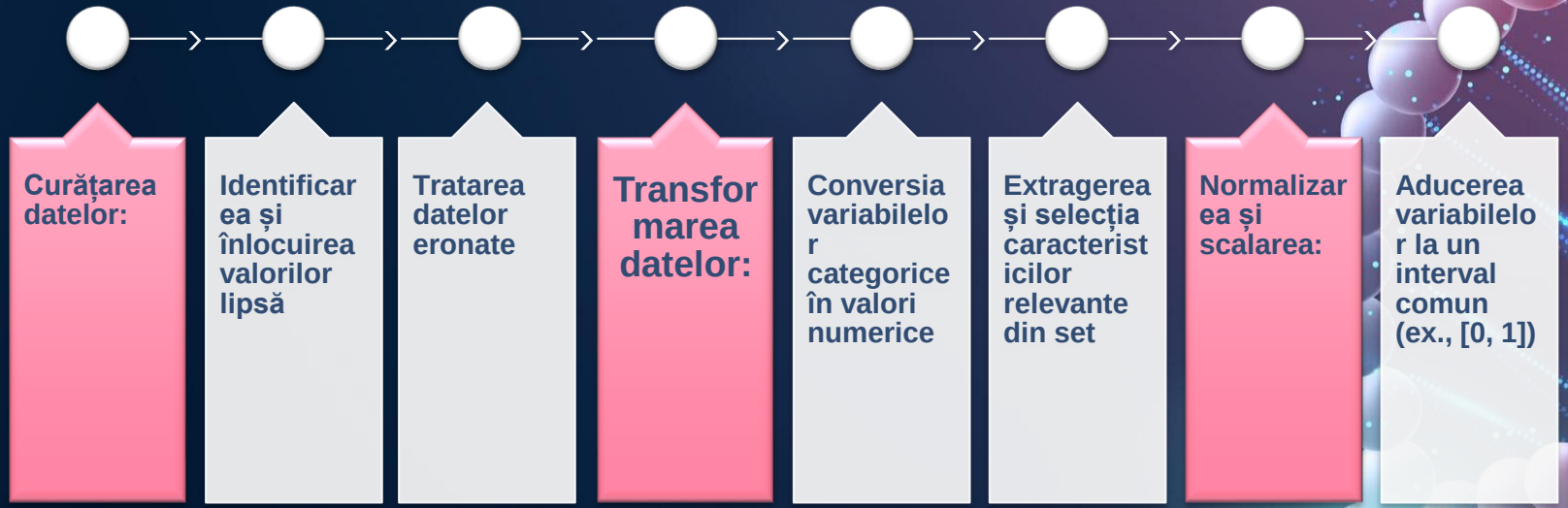
Setul de date conține informații medicale ale pacienților, fiecare fiind etichetat cu **prezența sau absența bolii cardiace**. Folosind aceste date, vrem să **identificăm pacienții cu risc de boală cardiacă**.

Variabilă	Descriere
Age	Vârsta pacientului (N)
Sex	Genul pacientului (C)
ChestPainType	Tipul durerii toracice (C)
RestingBP	Tensiunea arterială în repaus (N)
Cholesterol	Nivelul colesterolului (N)
FastingBS	Glicemia pe nemâncate (N)
RestingECG	Rezultatele ECG (C)
MaxHR	Frecvența cardiacă maximă atinsă (N)
ExerciseAngina	Angina indusă de efort (Da/Nu) (C)
Oldpeak	Depresiunea segmentului ST(N)
ST_Slope	Forma pantei ST (C)



Variabila țintă (HeartDisease) →
0 = sănătos
1 = pacient cu boală cardiacă

II.Preprocesarea și Împărțirea Setului de Date

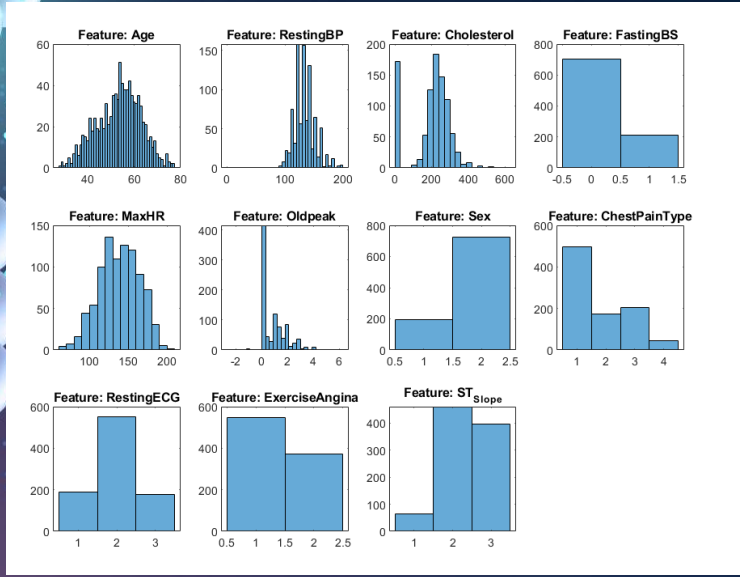


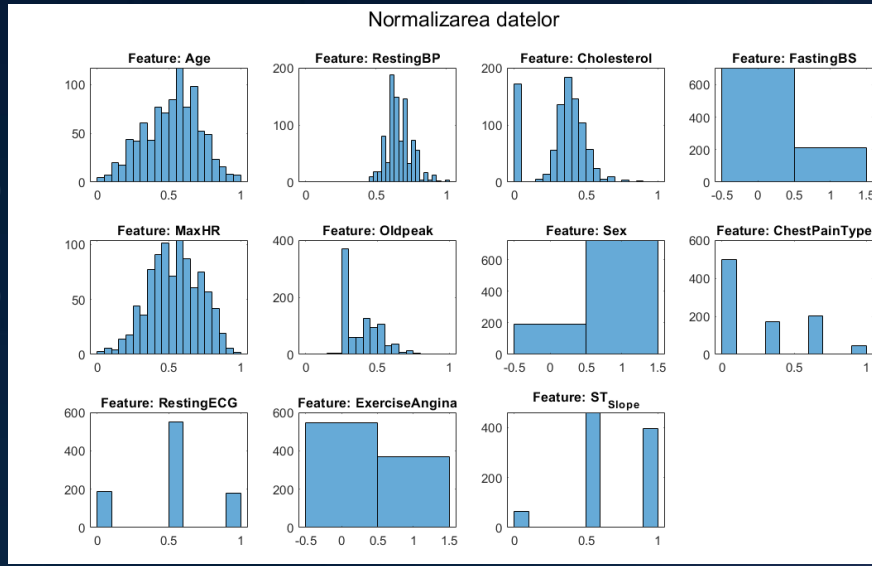
Setul de date a fost împărțit astfel:

🔹 **Set de antrenare** (80%) → Utilizat pentru a învăța modelele.

🔹 **Set de testare** (20%) → Folosit pentru evaluarea performanței.

II.Preprocesarea și Împărțirea Setului de Date





III. Modelele utilizate
K-NN & Random Forest

K-NN

✓ Este un **algoritm de clasificare bazat pe similaritate**.

✓ Clasificarea unui pacient se face **prin analiza celor mai apropiați K vecini** din setul de date.

🔷 **Cum funcționează?**

1. Se calculează **distanța** dintre pacientul nou și toți pacienții din setul de antrenare.
2. Se selectează **cei mai apropiați K vecini**.
3. Se face **clasificarea** bazată pe majoritatea etichetelor vecinilor.

Random Forest

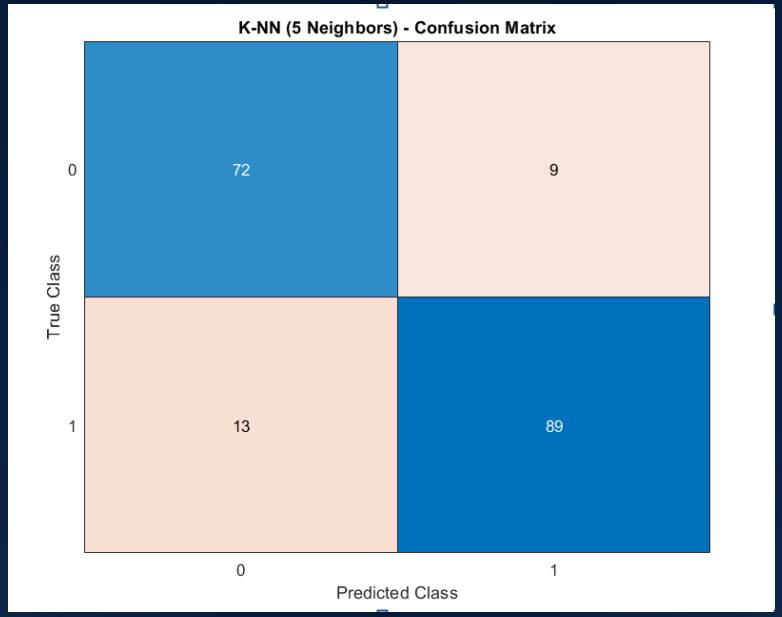
✓ Este un **model bazat pe arbori de decizie**, unde mai mulți arbori sunt antrenați pe subseturi de date.

✓ Clasificarea finală se face **prin votul majorității arborilor**.

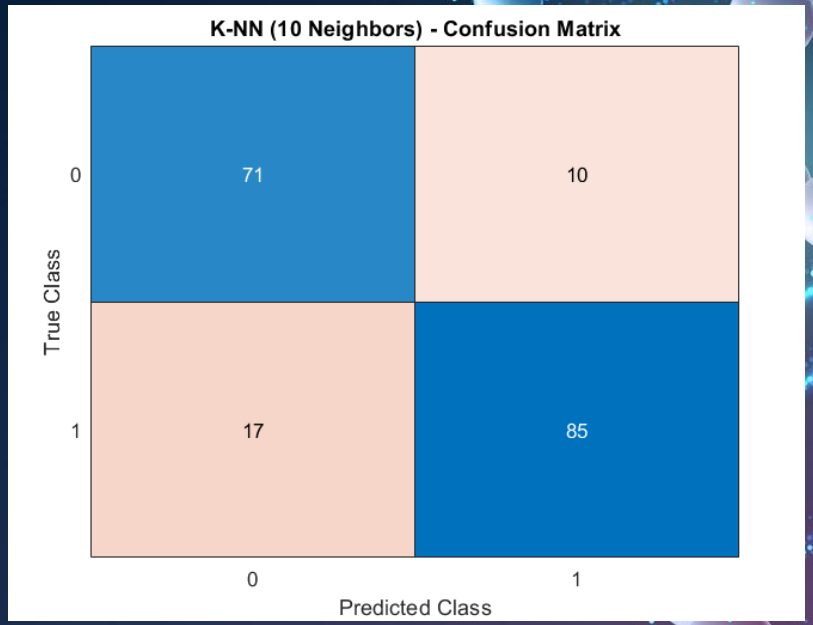
🔷 **Cum funcționează?**

1. Setul de date este **împărțit în subseturi** pentru antrenarea mai multor arbori.
2. Fiecare arbore învață **reguli diferite** pentru clasificare.
3. **Toți arborii votează**, iar rezultatul final este determinat de **majoritatea voturilor**

IV. Aplicarea si compararea metodelor

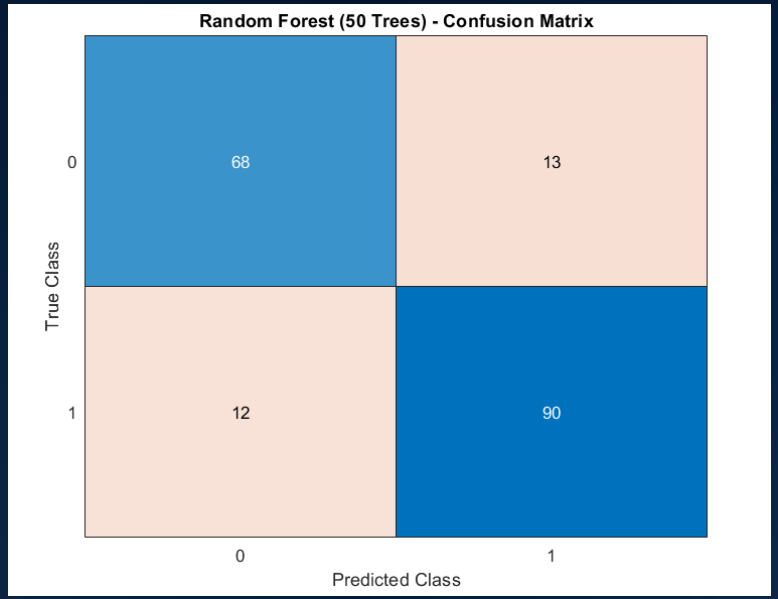


Precision: 0.86
Recall: 0.8866
F1 Score: 0.8731
Accuracy: 0.86339

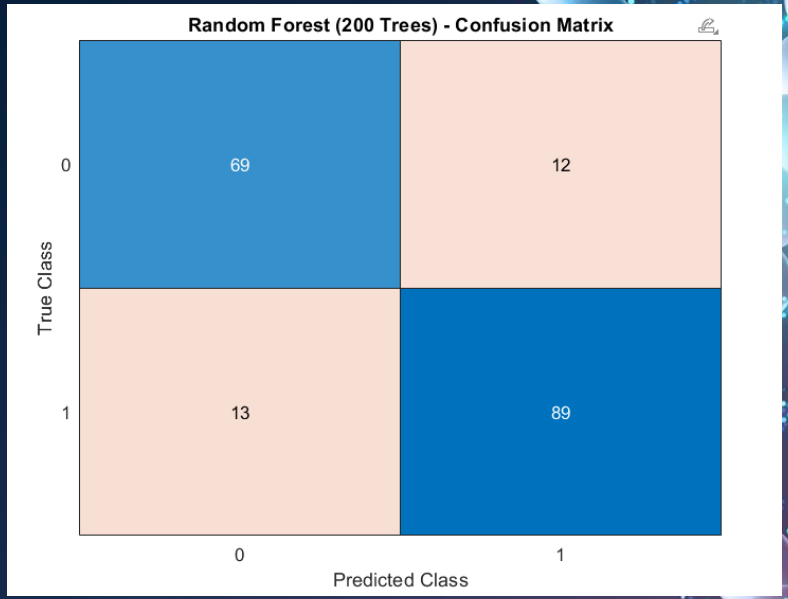


Precision: 0.87629
Recall: 0.87629
F1 Score: 0.87629
Accuracy: 0.86885

IV. Aplicarea si compararea metodelor



Precision: 0.87129
Recall: 0.90722
F1 Score: 0.88889
Accuracy: 0.87978
Prediction time (50 Trees): 0.1189 sec
Training time (50 Trees): 0.2779 sec



Precision: 0.85437
Recall: 0.90722
F1 Score: 0.88
Accuracy: 0.86885
Prediction time (200 Trees): 0.20479 sec
Training time (200 Trees): 0.34742 sec

IV.Avantaje și Dezavantaje ale K-NN și Random Forest

K-Nearest Neighbors (K-NN)

Avantaje

- ✓ Ușor de implementat și intuitiv
- ✓ Nu necesită antrenare (doar caută vecinii în timpul predicției)
- ✓ Performanță bună pe date cu relații locale
- ✓ Flexibilitate în alegerea lui k pentru optimizare

Dezavantaje

- ✓ Lent pentru seturi mari de date deoarece caută în întregul set la fiecare predicție.
- ✓ Sensibil la outlieri și la alegerea lui k
- ✓ Nu funcționează bine dacă datele nu au modele clare de proximitate

IV.Avantaje și Dezavantaje ale K-NN și Random Forest

Random Forest

Avantaje

- ✓ Rezistent la overfitting datorită combinației de arbori
- ✓ Funcționează bine cu date complexe și zgomotoase
- ✓ Poate calcula importanța caracteristicilor
- ✓ Rapid în inferență (odată antrenat)

Dezavantaje

- ✓ Necesită mai multă putere de calcul în faza de antrenare
- ✓ Mai puțin interpretabil decât un singur arbore de decizie
- ✓ Poate deveni redundant cu prea mulți arbori, crescând inutil complexitatea

Concluzii și Posibile Îmbunătățiri ale Modelului

- Prin această analiză, am demonstrat importanța învățării automate în clasificarea și predicția medicală. Alegerea algoritmilor și prelucrarea datelor influențează major performanța, iar optimizarea continuă ne permite să construim sisteme mai precise și mai robuste.

- În viitor, îmbunătățiri prin stacking, selecția predictorilor și interpretabilitatea modelului vor aduce soluții predictive mai eficiente pentru aplicații reale

Q&A Time!