# Dezvoltarea unui model de predicție a supraviețuirii

**Deleu Iulius**

Facultatea Calculatoare, Informatică și Microelectronică

Universitatea Tehnică a Moldovei, Chișinău

Email: iulius.deleu@gmail.com

# ABSTRACT

Acest articol prezintă o investigație cuprinzătoare privind dezvoltarea unui model predictiv conceput pentru a spori acuratețea predicțiilor de supraviețuire. Studiul cuprinde două componente primare: Materiale și metode și Rezultate. În secțiunea Materiale și metode, este realizată o prezentare detaliată a setului divers de modele de învățare automată utilizate în cadrul acestui studiu. Regresia logistică, clasificatorul K-Nearest Neighbors, Gaussian Naive Bayes, Bernoulli Naive Bayes, clasificatorul Decision Tree, clasificatorul RandomForest și clasificatorul XGBoost au fost evaluate cu și utilizate în cadrul acestui studiu. Setul de date utilizat în acest studiu a fost prelucrat cu atenție, constând în caracteristici importante și etichete de supraviețuire corespunzătoare, creând o bază solidă pentru formarea și evaluarea modelului. Trecând la secțiunea Rezultate, este prezentată o analiză detaliată a performanței fiecărui model. Parametrii de evaluare, cum ar fi acuratețea, precizia, reamintirea și scorul F1, sunt examinați în detaliu, oferind o înțelegere nuanțată a punctelor forte și a limitărilor fiecărui model. Ulterior, se prezintă o explorare aprofundată a procesului de optimizare și de reglare. Acesta implică o strategie de rafinare sistematică și iterativă care vizează îmbunătățirea capacităților de predicție ale modelelor luate în considerare. Demn de remarcat este punctul interesant al acestui proces iterativ, în care, după o optimizare riguroasă, s-a luat decizia de a utiliza exclusiv clasificatorul XGBoost în modelul final de predicție a supraviețuirii. Ne adâncim în factorii care au influențat această alegere, precizând caracteristicile specifice și parametrii de reglare care au contribuit la performanța sa superioară. Eficacitatea modelului ales este subliniată, accentuând robustețea acestuia în predicția rezultatelor de supraviețuire. Acest studiu contribuie cu informații valoroase la domeniul modelării predictive, în special în domeniul sănătății și în domenii conexe.

# INTRODUCERE

Analiza de supraviețuire este analiza datelor de timp până la eveniment. Astfel de date descriu durata de timp de la o origine temporală până la un punct final de interes. De exemplu, ar putea fi studiate persoane care ar putea fi urmărite de la naștere până la debutul unei anumite boli. Metodele de analiză a supraviețuirii sunt utilizate de obicei pentru a analiza datele colectate prospectiv în timp, cum ar fi datele dintr-un studiu de cohortă prospectiv sau datele colectate pentru un studiu clinic.

Originea temporală trebuie să fie specificată astfel încât indivizii să fie cât mai mult posibil pe picior de egalitate. De exemplu, în cazul în care se studiază durata de supraviețuire a pacienților cu un anumit tip de boală cardiacă, se poate alege ca origine temporală momentul diagnosticării acelui tip de boală cardiacă. La fel de important, punctul final sau evenimentul de interes ar trebui să fie specificat în mod corespunzător, astfel încât timpii luați în considerare să fie bine definiți. În exemplul de mai sus, ar putea fi vorba de decesul datorat bolii cardiace studiate. Apoi, ar putea fi calculată durata de timp de la momentul de origine până la punctul final.

Unul dintre motivele pentru care analiza supraviețuirii necesită tehnici "speciale" este posibilitatea de a nu observa evenimentul de interes pentru anumite persoane. De exemplu, indivizii pot renunța la un studiu sau pot avea un eveniment diferit, cum ar fi, în exemplul de mai sus, decesul din cauza unui accident, care nu face parte din punctul final de interes. O altă posibilitate este aceea că ar putea exista un moment la care studiul se încheie și, astfel, dacă unii indivizi nu au avut încă evenimentul lor, momentul evenimentului nu va fi observat. Aceste observații incomplete nu pot fi ignorate, dar trebuie să fie tratate diferit. Acest lucru se numește cenzurare. O altă caracteristică a datelor de supraviețuire este faptul că distribuțiile sunt adesea asimetrice și, prin urmare, tehnicile simple bazate pe distribuția normală nu pot fi utilizate direct.

Obiectivele analizei de supraviețuire includ analiza modelelor de evenimente în timp, compararea distribuțiilor timpilor de supraviețuire în diferite grupuri de indivizi și examinarea dacă și în ce măsură anumiți factori afectează riscul unui eveniment de interes.

***Cenzurarea***

Cel mai frecvent întâlnit tip de cenzură și cel mai ușor de gestionat în analiză este cenzura dreaptă. Cenzura dreaptă apare atunci când un individ este urmărit de la un moment de origine până la un moment ulterior și nu a avut evenimentul de interes, astfel încât tot ceea ce știm este că evenimentul nu a avut loc până la momentul de cenzură. Acest lucru se poate întâmpla, de exemplu, în cazul în care o persoană se retrage dintr-un studiu înainte de producerea evenimentului de interes. În mod obișnuit, studiile se încheie la un anumit moment specificat, iar la sfârșitul studiului unele persoane nu au avut încă evenimentul lor. Acest lucru este denumit uneori cenzura administrativă. În unele studii, majoritatea participanților sunt cenzurați.

Un alt tip de cenzură este cenzura de stânga. Cenzura la stânga este situația în care se știe că un individ a avut evenimentul înainte de un anumit moment, dar acesta poate fi orice moment înainte de cenzură.

Un alt concept este cel de trunchiere. Trunchierea este ceva ce se întâmplă în mod deliberat. Trunchierea stângă este cel mai des întâlnit tip de trunchiere, în care indivizii intră în studiu după ce au avut evenimentul lor de trunchiere (care nu este același cu evenimentul studiat). Intrarea întârziată, în care, de exemplu, un set de adulți sunt recrutați într-un studiu, dar cei care au avut evenimentul înainte de vârsta adultă nu sunt incluși deloc, este foarte frecventă. Trunchierea corectă apare atunci când întreaga populație studiată a experimentat deja evenimentul de interes.

Pentru metodele standard de analiză, cenzura ar trebui să fie neinformativă, adică momentul cenzurii ar trebui să fie independent de momentul evenimentului care ar fi fost altfel observat, având în vedere orice variabile explicative incluse în analiză, altfel inferența va fi distorsionată.

Un exemplu de cenzură informativă care nu trebuie ignorat este următorul: într-un studiu privind supraviețuirea după diagnosticarea unei boli, pacienții ar putea fi pierduți la urmărire deoarece starea lor s-a agravat și nu mai pot participa la programări. Sau, într-un studiu privind tratamentele pentru o afecțiune care nu pune în pericol viața, unii pacienți ar putea renunța la studiu deoarece starea lor s-a îmbunătățit și aleg să întrerupă tratamentul. De obicei, nu este posibil să se știe dacă cenzura într-un studiu este într-adevăr neinformativă.

***Modele de regresie***

Unul dintre obiectivele analizei datelor de supraviețuire ar putea fi acela de a examina dacă timpii de supraviețuire sunt legați de alte caracteristici. Modelele de regresie pot fi utilizate pentru a evalua efectul covariatelor asupra rezultatului. Acestea sunt similare cu analizele de regresie pentru alte tipuri de rezultate, cum ar fi regresia liniară pentru un rezultat numeric continuu sau regresia logistică pentru rezultate binare. Două tipuri obișnuite de modele de regresie pentru datele de supraviețuire, clasificate în funcție de modul în care se presupune că covariatele afectează timpii de supraviețuire, sunt modelul riscurilor proporționale Cox și mdoelul timpului de eșec accelerat.

***Alegerea scării temporale***

În unele exemple, scara de timp poate fi specificată în mai multe moduri. De exemplu, într-un studiu în care indivizii intră în studiu ca adulți și sunt urmăriți până la apariția unei anumite boli, este posibil să se utilizeze vârsta ca scală de timp, adică nașterea unui individ ca origine temporală și apariția bolii ca punct final, permițând intrarea întârziată, astfel încât indivizii să nu fie "la risc" până când nu intră în studiu. Alternativ, se poate alege ca origine temporală să fie momentul intrării în studiu. Alegerea se bazează, de obicei, pe considerente legate de subiect, cea mai bună opțiune fiind cea în cadrul căreia indivizii sunt cât mai asemănători posibil în ceea ce privește rostul subiacent al evenimentului la momentul de origine.

***Variabile explicative dependente de timp***

Până acum am presupus că o variabilă explicativă este măsurată o singură dată și reprezintă fie o caracteristică a unui individ la un moment fix în raport cu originea timpului, fie o caracteristică care rămâne neschimbată pentru momentul observației, de exemplu sexul unui pacient. Cu toate acestea, valoarea unei variabile explicative se poate modifica în timp. Noi numim o astfel de variabilă o variabilă explicativă dependentă de timp. Modelul Cox al pericolelor proporționale poate găzdui astfel de variabile. În practică, acest lucru ar necesita divizarea timpului în unități de timp discrete, de exemplu ani, și atribuirea unei valori pentru fiecare variabilă la fiecare astfel de unitate de timp în care o persoană este expusă riscului. Prin urmare, datele de la timp la eveniment cu variabile explicative dependente de timp sunt reprezentate de observații multiple per individ, fiecare reprezentând o unitate de timp în care individul a fost expus riscului.

Procesul complex al analizei și dezvoltării modelelor de predicție a supraviețuirii reprezintă o componentă esențială în evoluția domeniilor medicinii, asigurărilor și explorării spațiale. Acest studiu se adâncește în aspectele cheie ale acestui proces, oferind o perspectivă detaliată asupra etapelor implicate și evidențiind modalitățile prin care se identifică și se adaptează modelele pentru a fi cât mai eficiente într-un context specific.

# MATERIALE ȘI METODE

**Colectarea Datelor**

Prima etapă a procesului de analiză și dezvoltare a modelelor de predicție a supraviețuirii, este colectarea datelor relevante. Informațiile privind vârsta, prezența bolilor de inimă, nivelul de glicoză, genul și alți parametri relevanți sunt esențiale pentru construirea unui set de date robust. Această colectare se poate realiza prin intermediul studiilor clinice, a bazelor de date medicale sau a altor surse specializate, asigurându-se astfel că informațiile sunt cuprinzătoare și de încredere, pentru aceasta se respectă pașii ce urmează.

***1. Selectarea Publicului Țintă***

Selectarea publicului țintă constă în definirea grupului de subiecți sau obiecte asupra cărora se va aplica modelul (de exemplu, pacienți cu o anumită afecțiune medicală).

***2. Identifficarea Surselor de Date***

Identificarea surselor potențiale de date, cum ar fi baze de date medicale, registre, studii clinice, sau alte surse relevante pentru domeniul de interes.

***3. Evaluarea Calității Datelor***

Evaluarea calității datelor se realizează cu scopul de a asigura integritatea și relevanța acestora, verificarea dacă acestea sunt complete, consistente și actializate, identificarea și tratarea eventualelor erori sau lipsuri în setul de date.

***4. Selectarea Metodelor de Colectare a Datelor***

Selectarea metodelor de colectare a datelor, cum ar fi chestionare, interviuri, analiza documentelor, sau utilizarea datelor preexistente din registre electronice de sănătate, implementarea tehnicilor adecvate pentru colectarea datelor, în funcție de natura informațiilor și a obiectivelor.

***5. Validarea și Verificarea Datelor***

Validarea datelor pentru a confirma corectitudinea și acuratețea acestora, implementarea verificărilor incroșate și a procedurilor de reconciliere a datelor pentru a evita erorile sistemice.

Colectarea datelor pentru un model de predicție a supraviețuirii este un proces riguros, care necesită atenție la detalii și respectarea standardelor etice. Asigurarea calității și reprezentativității datelor este esențială pentru obținerea unor rezultate precise și relevante în cadrul modelului.

**Procesarea Datelor**

Odată ce datele au fost adunate, pasul următor în procesul analizei datelor constă în prelucrarea acestora. În această fază, tehnici avansate de analiză, precum machine learning și algoritmi de învățare automată, devin instrumente fundamentale pentru identificarea tiparelor și corelațiilor semnificative. Această etapă revine la centrul investigației, fiind crucială pentru a dezvălui relațiile complexe dintre diverși parametri și rezultatele privind supraviețuirea într-un set de date. În cadrul acestui proces de prelucrare, recurgem la utilizarea funcțiilor de vizualizare a informației despre datele obținute. Aspecte precum numărul total de înregistrări, tipul de date și volumul eșantionului de date sunt esențiale pentru obținerea unei perspective complete. După gestionarea valorilor lipsă, urmează realizarea și vizualizarea corelației dintre diversele variabile. Diferite tipuri de grafice și ploturi sunt utilizate pentru a oferi o viziune clară asupra distribuției datelor în funcție de diverse caracteristici, cum ar fi genul, vârsta, mediul de trai, prezența hipertensiunii și rezultatul privind supraviețuirea. În acest context, se explorează distribuția datelor în raport cu variabile numerice, precum nivelul de glucoză și indicatorul de indice de masă corporală. Analiza acestor distribuții ajută la obținerea unei înțelegeri mai profunde a relațiilor dintre aceste variabile și evenimentul de supraviețuire. Prin utilizarea ploturilor comparative, putem observa distribuția datelor în funcție de diverși factori, punând în lumină modul în care aceștia influențează probabilitatea de supraviețuire. Această analiză comparativă aduce în prim plan variațiile semnificative și oferă o perspectivă detaliată asupra datelor prezentate de eșantionul de date investigat. În final, această etapă de prelucrare și analiză a datelor este primordială pentru dezvăluirea insight-urilor semnificative și luarea deciziilor informate în cadrul domeniilor variate care beneficiază de învățarea automată și analiza datelor.

**Separarea Setului de Date**

Procesul de separare a datelor reprezintă o etapă fundamentală în domeniul învățării automate, având un rol crucial în crearea și evaluarea modelelor de predicție. Această procedură implică împărțirea inițială a setului de date în subseturi distincte, fiecare având scopul său bine definit în procesul de instruire, validare și testare a eficacității modelului. Primul pas esențial în acest proces constă în divizarea setului de date în trei componente principale: setul de instruire, setul de validare și setul de testare. Această împărțire strategică asigură o abordare sistematică și eficientă în antrenarea și evaluarea modelului. Proportiile în care se face această împărțire sunt determinate de dimensiunea totală a setului de date și de particularitățile problemei specifice. Un aspect important în procesul de separare constă în amestecarea datelor înainte de împărțire, evitând astfel distorsiunile care pot apărea în seturile rezultate. Aceasta contribuie la eliminarea posibilelor modele sau corelații nedorite care pot influența negativ performanța modelului. În cazul în care există o distorsiune în distribuția claselor din setul de date, se impune o divizare stratificată pentru a asigura o proporție echitabilă de date din fiecare clasă în fiecare subset. Există și metode avansate de separare a datelor, precum validarea încrucișată, care implică împărțirea setului de date în mai multe pliuri sau fold-uri. Aceasta permite antrenarea și evaluarea repetată a modelului pe porțiuni diferite ale datelor, furnizând astfel o evaluare mai robustă și mai generalizată a performanței modelului. Separarea corespunzătoare a datelor devine esențială pentru asigurarea învățării corecte a modelului și pentru capacitatea acestuia de a se generaliza eficient la date noi. Prin adoptarea unei abordări riguroase în procesul de separare a datelor, se creează premisele unei învățări automate de calitate și a dezvoltării unor modele de predicție cu performanțe superioare și capacitate de adaptare la variabilitatea datelor de intrare.

**Supraeșantionarea Sintetică a Minorităților**

Supraeșantionarea sintetică a minorităților, cunoscută și sub denumirea de tehnica SMOTE (Synthetic Minority Over-sampling Technique), reprezintă o strategie eficientă în domeniul învățării automate, în special atunci când se lucrează cu seturi de date în care există un dezechilibru semnificativ între clase, cum ar fi în cazul modelelor de predicție a supraviețuirii. Această tehnică are scopul de a compensa disparitatea dintre clasele majore și minore prin generarea sintetică a unor exemple suplimentare din clasa minoritară. Principiul fundamental al SMOTE constă în crearea de exemple noi pentru clasa minoritară prin interpolarea între exemplele deja existente. Acest proces presupune selectarea unui exemplu din clasa minoritară și identificarea vecinilor săi în spațiul caracteristicilor. Apoi, pentru fiecare vecin identificat, se generează un exemplu sintetic situat pe segmentul de linie dintre exemplul inițial și vecinul respectiv. Prin această metodă, se extinde setul de date al clasei minoritare, contribuind la îmbunătățirea performanței modelului în identificarea și predicția instanțelor din această clasă. Procesul de supraeșantionare sintetică a minorităților are avantajul de a evita supraeșantionarea simplă, care constă în duplicarea aleatoare a exemplelor din clasa minoritară. SMOTE furnizează exemple sintetice care prezintă variații și caracteristici mai semnificative, contribuind astfel la îmbunătățirea generalizării modelului. Aplicarea SMOTE în contextul unui model de predicție a supraviețuirii ajută la echilibrarea setului de date, ceea ce poate conduce la o învățare mai precisă și la obținerea unor rezultate mai fiabile. Această tehnică reprezintă o soluție eficientă pentru a gestiona provocările create de dezechilibrul claselor în seturile de date, contribuind la creșterea acurateței și robusteții modelului în prezența datelor cu caracteristică minoritară.

**Selectarea Modelului Optim**

Selectarea modelului optim reprezintă un proces în cadrul dezvoltării unui sistem de învățare automată sau a unui model de predicție eficient. Pentru a face alegerea potrivită, este esențial să ne stabilim cu claritate obiectivele și metricile de evaluare. Această etapă inițială constituie fundația deciziei și ghidează în direcția potrivită. Înainte de a alege un anumit model, este necesar să explorăm o gamă diversificată de algoritmi de învățare automată. Comprendererea modului în care aceștia se potrivesc cu setul nostru de date și cu cerințele specifice ale problemei este crucială pentru a face o alegere informată. Validarea încrucișată reprezintă un instrument esențial pentru evaluarea performanței modelului pe date diferite și pentru prevenirea suprainstruirii. Această abordare ne asigură că modelul poate generaliza bine pe datele noi, evitând astfel interpretarea greșită a modelelor. Reglarea parametrilor modelului este o altă componentă vitală a procesului. Aceasta poate avea un impact semnificativ asupra performanței generale a modelului, de aceea, ajustările adecvate sunt necesare pentru a atinge rezultate optime. Compararea performanțelor diferitelor modele este, de asemenea, esențială pentru a identifica cel mai potrivit și eficient model pentru problema noastră specifică. În cadrul acestui proces, evitarea suprainstruirii și utilizarea tehnicilor de regularizare devine esențială. Acestea contribuie la menținerea generalizării modelului și împiedică adaptarea excesivă la datele de antrenament. Interpretarea modelului reprezintă o etapă suplimentară, dar utilă în anumite contexte. Înțelegerea modului în care modelul ia decizii poate adăuga transparență și încredere în rezultatele obținute. Nu în ultimul rând, resursele și scalabilitatea trebuie luate în considerare. Asigurarea faptului că modelul poate funcționa eficient pe seturi mari de date și poate scala în conformitate cu cerințele noastre este esențială pentru implementarea cu succes în mediul dorit. Parcurgerea atentă a acestor pași permite să luăm decizii informate și să asigurăm că modelul ales se aliniază perfect cu obiectivele și cerințele noastre specifice. Este un proces complex, dar esențial în construirea unui sistem de învățare automată sau a unui model de predicție performant și adaptat nevoilor.

**Adaptarea Modelului**

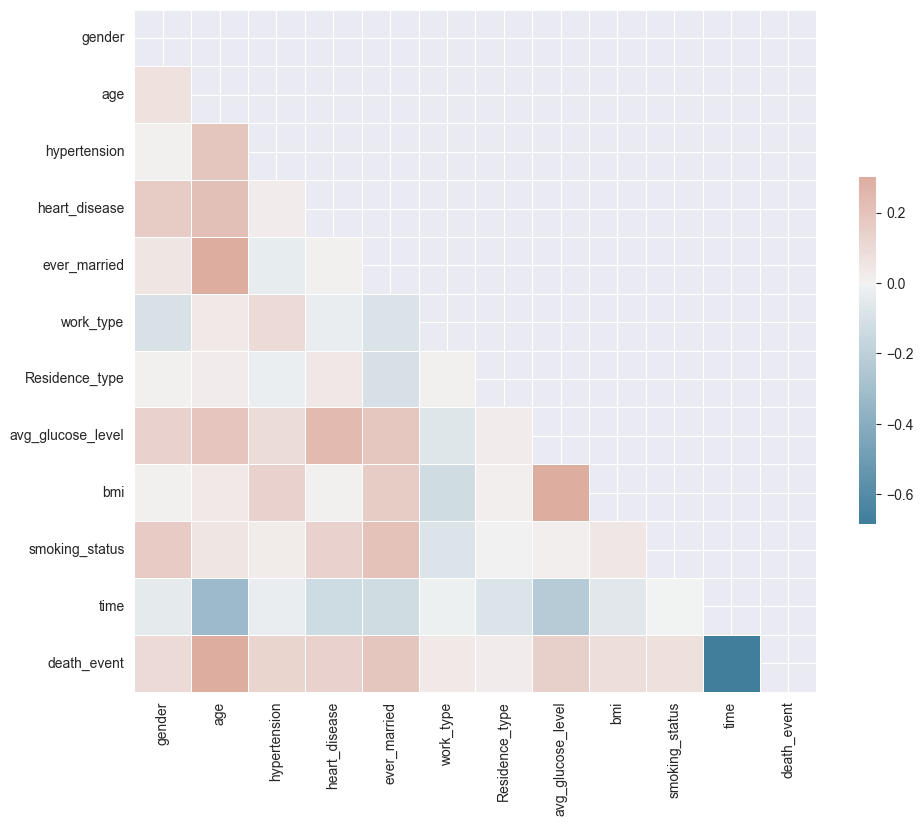
Adaptarea modelului reprezintă un aspect esențial în domeniul învățării automate, constând în procesul de optimizare a performanței unui model existent, fie în cadrul datelor de testare, fie în mediul de producție. Această etapă complexă implică mai multe strategii și factori esențiali, care necesită atenție detaliată. Unul dintre pilonii adaptării modelelor constă în reantrenarea acestora cu date actualizate. Actualizarea constantă a setului de date utilizat pentru antrenare contribuie la menținerea relevanței modelului în fața schimbărilor din mediul înconjurător. În același timp, ajustarea fină a parametrilor și optimizarea hiperparametrilor sunt practici esențiale pentru a asigura o adaptare corespunzătoare la caracteristicile noilor date. O altă componentă importantă a adaptării modelelor este reevaluarea și validarea acestora. Procesul de testare și validare periodică confirmă sau ajustează performanța modelului în funcție de schimbările intervenite. Totodată, gestionarea derivării conceptelor este vitală pentru a gestiona și integra modificările semnificative în conceptele sau caracteristicile cheie ale datelor. Optimizarea resurselor și monitorizarea continuă a performanței modelului completează această panoramă complexă. Eficientizarea utilizării resurselor și evaluarea constantă a performanței sunt esențiale pentru a asigura o funcționalitate optimă și pentru a anticipa eventualele probleme ce pot apărea în timp. Este important de subliniat faptul că adaptarea modelului nu este o activitate punctuală, ci o necesitate continuă. Pentru a menține relevanța și eficiența modelului pe întreaga sa durată de viață, adaptarea continuă este imperativă, astfel încât acesta să rămână în pas cu schimbările din mediul înconjurător și să ofere performanțe superioare în orice context.

# REZULTATE

Această secțiune dezvăluie rezultatele explorării modelului de predicție a supraviețuirii. După evaluarea a diverși algoritmi, este prezentată o analiză detaliată a performanțelor acestora și a procesului de optimizare ulterior. Accentul este pus pe parametrii cheie și pe raționamentul care a stat la baza selectării clasificatorului XGBoost ca model exclusiv. Această prezentare succintă oferă informații despre procesul de identificare a celui mai eficient model de predicție pentru rezultatele de supraviețuire.

După procesul de colectare a datelor, a urmat procesarea datelor, anume înlăturarea dalor lipsă, normalizarea datelor, ca urmare am obșinut un eșantion de date a circa 300 de înregistrări structurate în 12 coloane, dintre care, în urma transformărilor, am obținut 9 coloane ce conțin date de tip categorial, inclusiv valoarea țintă (death\_event).

În contextul elaborării unui model de predicție a supraviețuirii, crearea unei matrice de corelații implică calcularea sistematică a coeficienților de corelație între diverse caracteristici din setul de date. Utilizând metode precum coeficientul de corelație Pearson, fiecare celulă din matrice semnifică intensitatea corelației dintre două variabile. Acest coeficient măsoară relația liniară dintre factori, variind de la -1 (indicând o corelație negativă perfectă) la 1 (indicând o corelație pozitivă perfectă), 0 indicând absența corelației. Matricea de corelație servește drept instrument vizual pentru a rezuma în mod cuprinzător interdependențele dintre aceste caracteristici. Prin explorarea matricei, pot fi identificate potențiale modele sau asociații relevante pentru predicția supraviețuirii, iar drept rezultat este prezentată matricea de corelație din figura 1.1.



**Figura 1.1 Matricea de corelație**

Divizarea datelor este o etapă ce urmează în dezvoltarea modelelor, în care un set de date este împărțit în subseturi distincte pentru instruire, validare și testare. Setul de instruire este utilizat pentru a instrui modelul, setul de validare ajută la optimizarea și reglarea fină a parametrilor, iar setul de testare evaluează performanța modelului pe date nevăzute. Acest proces garantează că modelul se generalizează bine la date noi, sporindu-i fiabilitatea și eficiența. Împărțirile obișnuite includ proporții de 70-30 sau 80-20 pentru instruire și testare, cu subseturi suplimentare pentru validare, dacă este necesar.

SMOTE, sau Tehnica de supraeșantionare sintetică a minorităților, abordează dezechilibrul clasei prin crearea de instanțe sintetice ale clasei minoritare. Aceasta selectează o instanță minoritară, identifică cei mai apropiați vecini ai acesteia și generează noi instanțe prin interpolare. Acest proces se repetă până când se obține echilibrul dorit, sporind capacitatea modelului de a învăța din clasa minoritară în seturi de date dezechilibrate. Echilibrul obținut este prezentat în tabelul 1.2.

**Tabel 1.2 Supraeșantionarea datelor**

|  |  |  |
| --- | --- | --- |
| Phase | Label | Count |
| Before OverSampling | 1 | 198 |
| Before OverSampling | 0 | 42 |
| After OverSampling | Train\_x Shape | (396, 11) |
| After OverSampling | Train\_y Shape | (396, ) |
| After OverSampling | 1 | 198 |
| After OverSampling | 0 | 198 |

În evaluarea diferitelor modele de clasificare pentru predicția supraviețuirii, au fost evaluați mai mulți parametri de performanță, prezentați în continuare:

1. Regresie logistică:

Utilizând regresia logistică am obținut o acuratețe de 87%, cu o precizie (0,96) și recall (0,88), rezultând un scor F1 de 0,92. Modelul a demonstrat o precizie medie de validare K-Fold de 91,40% și un scor ROC AUC de 0,83.

2. SVM:

Modelul SVM a atins o acuratețe de 85%, cu valori ale preciziei, rechemării și scorului F1 de 0,94, 0,88 și, respectiv, 0,91. Modelul a prezentat o consistență ridicată în validarea K-Fold (precizie medie: 95,96%) și a obținut un scor ROC AUC de 0,77.

3. KNeighbors:

KNeighbors oferă o acuratețe de 80%, cu o precizie (0,93) și recall (0,82) echilibrate, rezultând un scor F1 de 0,87. Validarea K-Fold a indicat stabilitate (Precizia medie: 92,91%), iar scorul ROC AUC a fost de 0,75.

4. GaussianNB:

A afișat o acuratețe de 85%, cu o precizie, recall și valori ale scorului F1 de 0,94, 0,88 și 0,91. Validarea K-Fold a relevat coerență (Precizia medie: 89,38%), iar scorul ROC AUC a fost de 0,77.

5. BernoulliNB:

Modelul dat a prezentat o acuratețe ridicată de 88%, cu o precizie notabilă (0,98), recall (0,88) și un scor F1 (0,93). Validarea K-Fold a raportat stabilitate (Precizia medie: 87,38%), iar scorul ROC AUC a fost de 0,89.

6. Arbore de decizie:

Arborele de decizie a demonstrat o acuratețe de 88%, cu o precizie (0,92) și recall (0,94) echilibrate, rezultând un scor F1 de 0,93. Modelul a demonstrat robustețe în validarea K-Fold (acuratețe medie: 93,70%) și a produs un scor ROC AUC de 0,75.

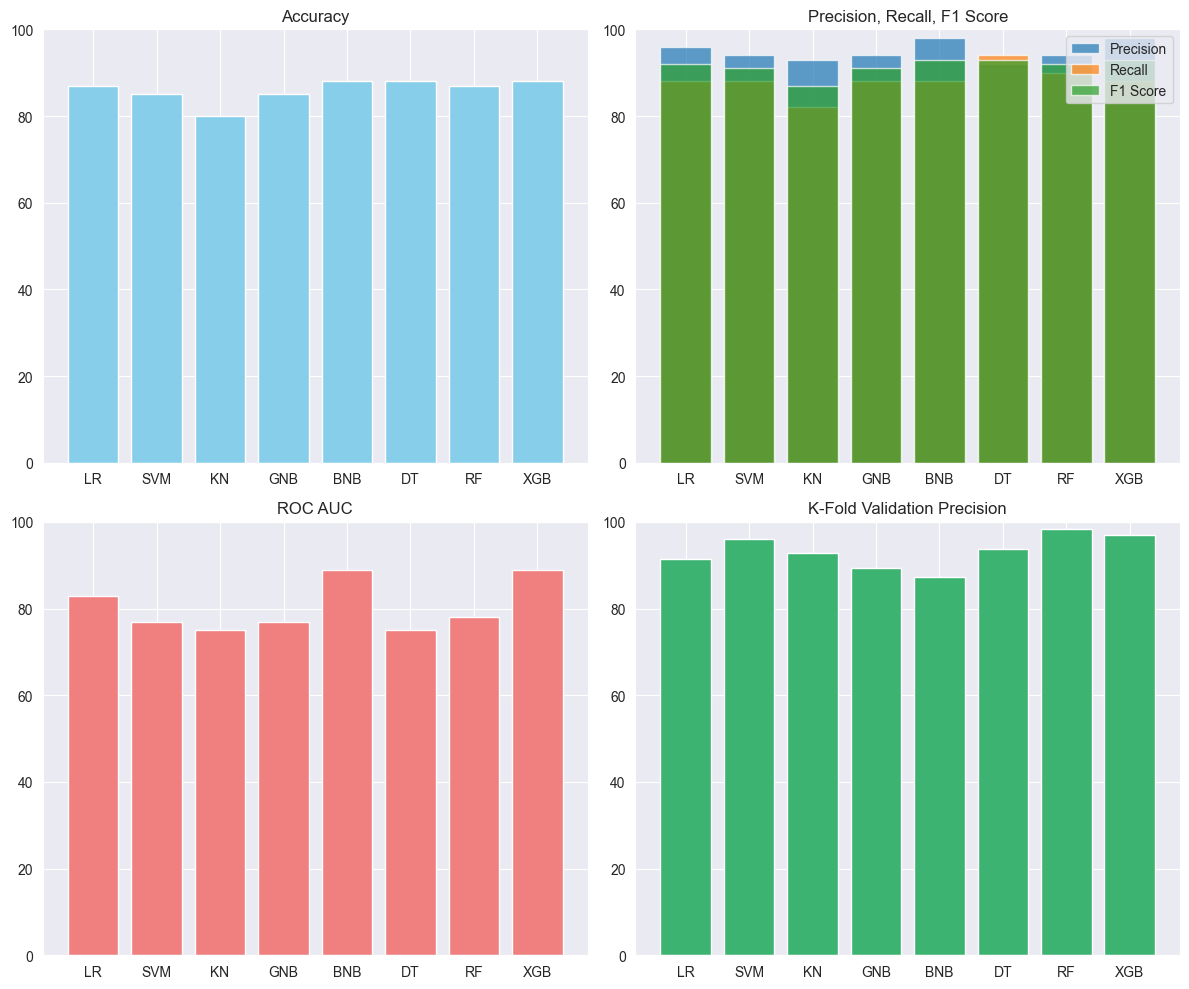
7. Random Forest:

Random Forest prezintă o acuratețe de 87%, cu valori ale preciziei, recall și scorului F1 de 0,94, 0,90 și, respectiv, 0,92. Validarea K-Fold a demonstrat o consistență ridicată (Precizia medie: 98,24%), iar scorul ROC AUC a fost de 0,78.

8. XGBoost:

Modelul dat a oferit o acuratețe de 88%, cu valori ale preciziei, recall și scorului F1 de 0,98, 0,88 și 0,93. Validarea K-Fold a indicat o coerență puternică (precizie medie: 96,98%), iar scorul ROC AUC a fost de 0,89.

Aceste rezultate oferă o înțelegere cuprinzătoare a performanțelor fiecărui model, aceestea fiind prezentate în figura 1.3, permițând luarea unor decizii informate în ceea ce privește selectarea celui mai potrivit model pentru predicția supraviețuirii.



**Figura 1.3 Performanța modelelor**

Pe scurt, au fost evaluate diferite modele de clasificare pentru predicția supraviețuirii, fiecare model fiind evaluat pe baza acurateței și a altor parametri cheie. Modelele s-au clasificat după cum urmează cele mai potrivite, adică cei mai înalți indicatori de performanță:

1. XGBoost (Precizie: 88,33%).

- Notabil pentru obținerea celei mai mari precizii.

- A demonstrat o performanță cu o precizie (0,98) și recall (0,88) ridicate.

- Scor general F1 puternic (0,93).

- Precizia medie de validare K-Fold (96,98%) și scorul ROC AUC (0,89) sunt constant ridicate.

2. Random Forest (Precizie: 86,67%).

- S-a lăudat cu un scor de acuratețe puternic.

- A prezentat o precizie ridicată (0,94) și recall echilibrat (0,90), ceea ce a dus la un scor F1 solid (0,92).

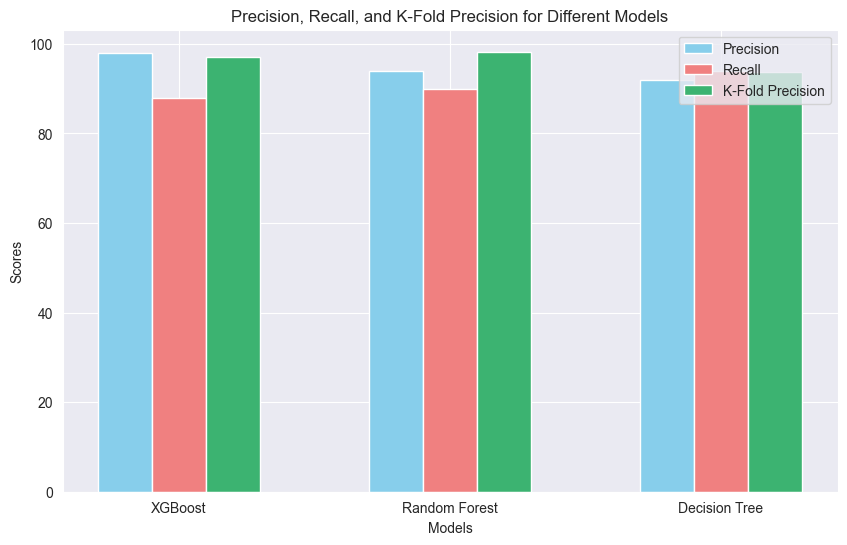
- Precizie medie excepțională a validării K-Fold (98,24%) și un scor ROC AUC (0,78).

3. Arbore de decizie (acuratețe: 88,33%).

- La egalitate cu Random Forest pentru cea mai mare acuratețe.

- A prezentat o precizie (0,92) și recall (0,94), ceea ce a condus la un scor F1 bun (0,93).

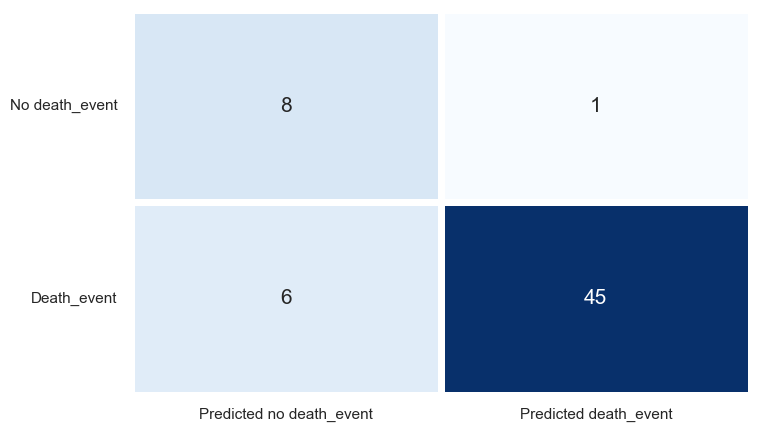
- Precizie medie puternică de validare K-Fold (93,70%) și un scor ROC AUC competitiv (0,75).



**Figura 1.4 Performanța modelelor selectate**

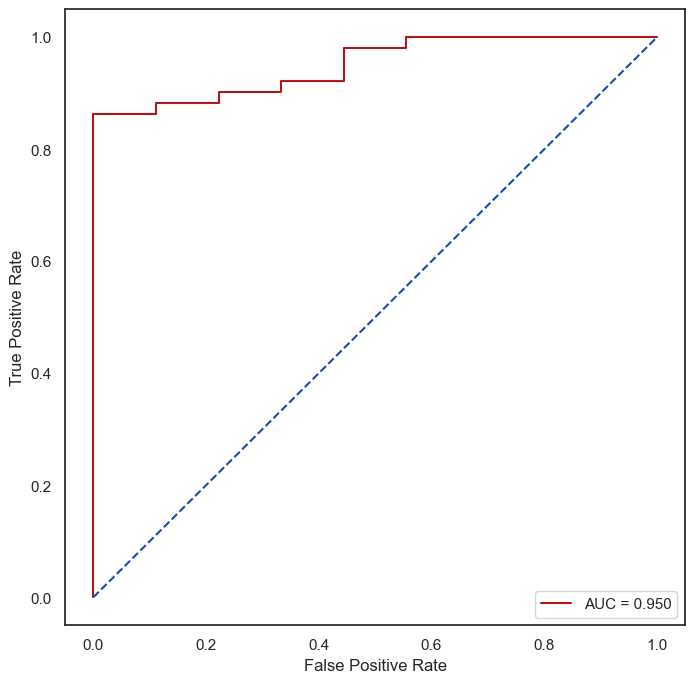
Având în vedere evaluarea cuprinzătoare a acurateței, preciziei, recall, scorului F1, validării K-Fold și scorului ROC AUC, XGBoost apare ca fiind cel mai promițător model pentru predicția supraviețuirii. Performanța sa constant ridicată în cadrul diferitelor măsurători, ceea ce se poate observa în figura 1.4, sugerează că este potrivit pentru a prezice cu precizie rezultatele supraviețuirii în setul de date dat,.

Clasificatorul XGBoost utilizat pentru clasificarea binară este concentrat pe prezicerea evenimentelor de deces. Clasificatorul este antrenat pe date reeșantionate și este evaluat folosind parametri precum matricea de confuzie, raportul de clasificare, scorul ROC AUC și acuratețea. Vizualizările, inclusiv o hartă termică a matricei de confuzie, figura 1.5.



**Figura 1.5 Matricea de confuzie**

Curba ROC din figura 1.6, este prezentată pentru a îmbunătăți înțelegerea performanței modelului. Această abordare cuprinzătoare asigură o evaluare completă a capacității de generalizare și a acurateței predictive a modelului XGBoost într-un context de clasificare binară.



**Figura 1.6 Curba ROC**

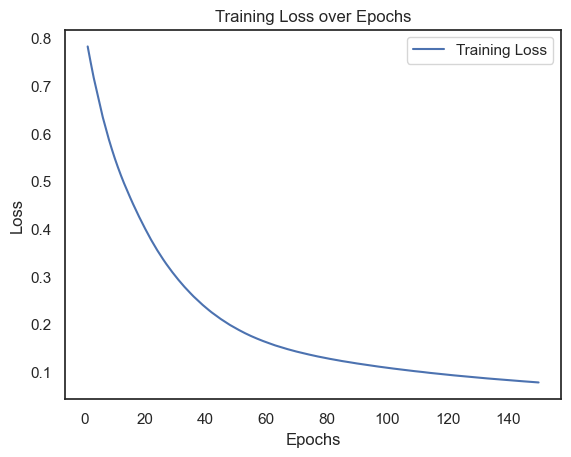
După toate procesele descrise anterior, este realizată construirea și optimizarea unei rețele neuronale MLP (Multi-Layer Perceptron) pentru sarcina de clasificare. Procesul implică definirea unei funcții pentru a crea clasificatorul MLP cu parametri personalizabili, specificarea unei grile de hiperparametri, aceștia fiind prezentați în continuare, pentru o căutare exhaustivă utilizând GridSearchCV și evaluarea performanței modelului prin validare încrucișată.

Best Parameters: {‘alpha’: 1, ‘hidden\_layer\_sizes’: (8, 8), ‘max\_iter’: 150, ‘solver’: ‘adam’}

Best Accuracy: 91.92%

Rezultatul final include cei mai buni hiperparametri și scorul de precizie corespunzător, oferind informații despre configurația optimă pentru rețeaua neuronală MLP în contextul de clasificare dat. Acest proces general poate fi aplicat în diverse sarcini de clasificare pentru a regla cu precizie modelele de rețele neuronale pe baza unor caracteristici specifice ale datelor.

Graficul din figura 1.7 arată evoluția pierderii de antrenament în funcție de numărul de epoci în timpul instruirii unui model de învățare automată. Pe măsură ce numărul de epoci crește, pierderea de antrenament scade, indicând că modelul devine mai precis. Cea mai mică valoare a pierderii este de aproximativ 0,1, suggerând o înaltă precizie în predicțiile modelului. Observațiile includ o scădere rapidă a pierderii la începutul instruirii, o scădere mai lentă pe măsură ce epocile avansează și posibilitatea unei creșteri a pierderii în cazul unui antrenament excesiv, semn al supraînvățării.



**Figura 1.7 Evoluția pierderii de antrenament**

În ansamblu, graficul indică faptul că modelul se antrenează eficient și devine mai precis pe parcursul timpului. Detaliile graficului includ axa X, reprezentând numărul de epoci, axa Y, reflectând pierderea de antrenament, și linia albastră, care ilustrează evoluția acesteia în timpul antrenamentului. Este o modalitate utilă de evaluare a performanței modelului pe durata procesului de învățare automată.

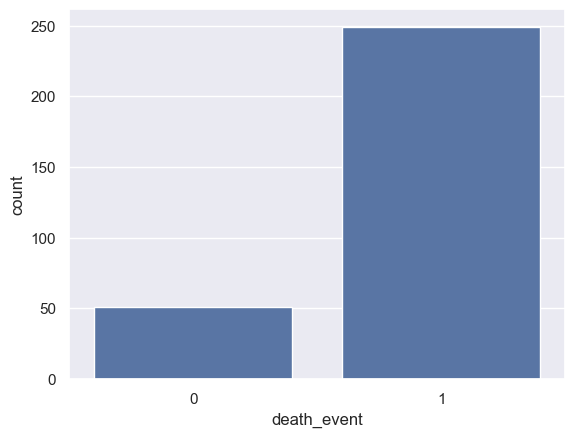
În concluzie, acest articol a detaliat un proces cuprinzător de dezvoltare a unui model de predicție a supraviețuirii, acoperind fiecare etapă crucială, de la colectarea datelor până la construirea modelului. Călătoria a început cu achiziționarea meticuloasă a datelor relevante din lumea reală, urmată de o fază de preprocesare atentă pentru a asigura calitatea și coerența datelor. S-au aplicat tehnici de inginerie a caracteristicilor pentru a transforma datele brute în predictori semnificativi, îmbunătățind capacitatea modelului de a capta modelele de supraviețuire.

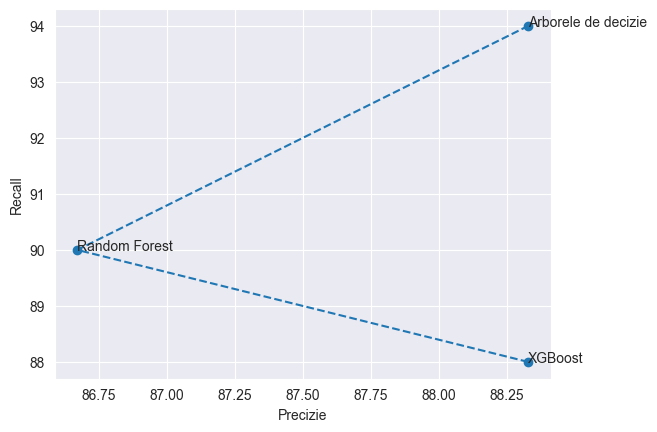
Articolul a elucidat procesul de construire a modelului, utilizând un clasificator perceptron multistrat (MLP) ca algoritm ales. Faza de reglare a hiperparametrilor, facilitată de căutarea în grilă, a avut ca scop optimizarea performanței modelului prin explorarea sistematică a diferitelor combinații de parametri. Modelul cu cele mai bune performanțe a fost selectat și antrenat pe datele de formare reeșantionate pentru a spori robustețea și generalizabilitatea.

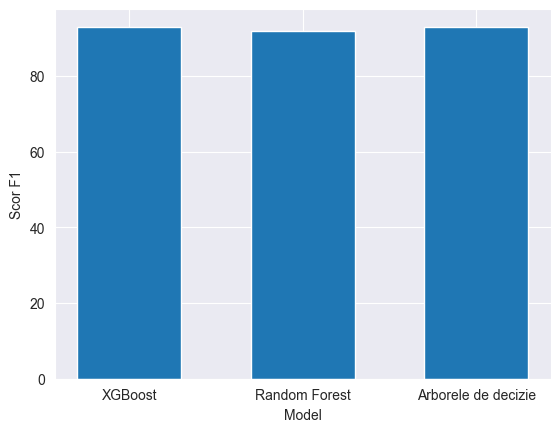
Rezultatele au fost prezentate folosind modelul selectat pentru a face predicții pe un set de date de testare separat. S-au utilizat parametri de evaluare, cum ar fi matricile de confuzie, pentru a evalua acuratețea modelului în ceea ce privește predicția rezultatelor supraviețuirii. Vizualizările, inclusiv reprezentările de hărți termice ale matricelor de confuzie, au oferit o imagine de ansamblu clară a performanței modelului.

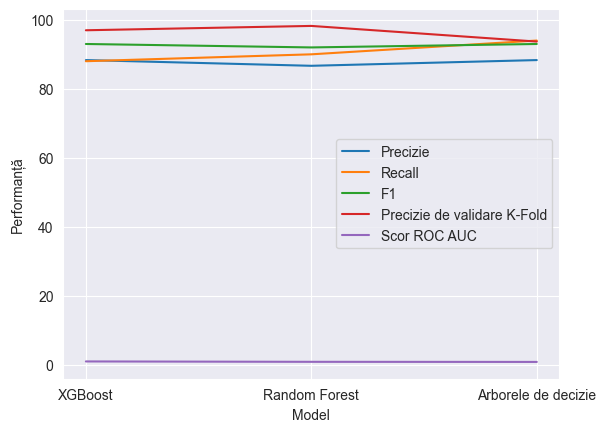
Această abordare cuprinzătoare nu numai că a prezentat complexitatea tehnică a modelării predicției supraviețuirii, dar a subliniat, de asemenea, importanța calității datelor, a preprocesării și a selecției modelului în obținerea unor rezultate precise și fiabile. Metodologia prezentată servește drept un ghid valoros pentru dezvoltarea și să evaluarea modelelor de predicție a supraviețuirii în diverse domenii.

# ANEXE









# BIBLIOGRAFIE

1. Repozitoriu web: https://github.com/iuliusd/data-analysis/tree/main
2. Sursă set de date: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset
3. UCI Machine Learning Repository: https://archive.ics.uci.edu/
4. Kalbfleisch, J. D., & Prentice, J. R. (2002). Survival analysis: A self-learning approach.
5. P.Armitage MA, PhD,, G.Berry MA, PhD,, J.N.S. Matthews MA. : Statistical Methods in Medical Research: https://onlinelibrary.wiley.com/doi/book/10.1002/9780470773666
6. Ripley BD, Ripley RM. Neural networks as statistical methods in survival analysis. Clin Appl Artif Neural Netw. 2001.
7. Wang P, Li Y, Reddy CK. Machine learning for survival analysis: A survey. ACM Comput Surv (CSUR). 2019.
8. Brown SF, Branford AJ, Moran W. On the use of artificial neural networks for the analysis of survival data. IEEE Trans Neural Netw. 1997.