UNIVERSITATEA BABEŞ-BOLYAI CLUJ-NAPOCA

FACULTATEA DE MATEMATICĂ ŞI INFORMATICĂ

SPECIALIZAREA BAZE DE DATE

LUCRARE DE DISERTAŢIE

Prezicerea scorului de Endoscopie Sinusală Perioperatorie prin intermediul algoritmilor de învățare automată

Conducător ştiinţific

Laura Silvia Dioșan Profesor Universitar

SANDU ANA-ALEXANDRA

2023-2024

Abstract

Cuprins

Capitolul 1

Introducere

1.1 Ce? De ce? Cum?

Rinosinuzita cronică cu polipi nazali este o boală inflamatorie cronică a cavităților nazale și sinusurilor, care se caracterizează prin prezența polipilor, care sunt creșteri benigne ale țesutului ce apar din mucoasa nasului și a sinusurilor. Printre cei mai importanți factori de risc pentru rinosinuzita cronică cu polipi nazali se numără alergiile, astmul, sensibilitatea la aspirină și antecedentele de sinuzită cronică.

Simptomele rinosinuzitei cronice cu polipi nazali includ congestie nazală, scurgeri nazale, pierderea mirosului sau a gustului, dureri faciale sau presiune și dureri de cap. Diagnosticul se face, de obicei, pe baza unei combinații de simptome, constatări ale examinării fizice și studii imagistice, cum ar fi tomografiile. În ceea ce privește opțiunile de tratare a rinosinuzitei cronice cu polipi nazali, acestea variază în funcție de severitatea bolii și de circumstanțele fiecărui pacient. Printre opțiunile valabile se numără corticosteroizii intranazali, corticosteroizii orali, antibioticele și intervențiile chirurgicale.

Una dintre modalitățile cele mai eficiente de a evalua starea pacientului postoperatorie este scorul de Endoscopie Sinusală Perioperatorie (POSE). Acest scor servește drept busolă și ajută la evaluarea riscurilor și optimizarea intervențiilor pentru a îmbunătăți rezultatele pacienților. Lucrarea de față se axează pe datele unor pacienți care au avut parte de o intervenție chirurgicală pentru a trata rinosinuzita cronică cu polipi nazali.

Astfel, prin această lucrare se dorește prognozarea probabilităților unor astfel de urgențe medicale după intervențiile chirurgicale. Această prognozare va ajuta la o mai bună înțelegere a complicațiilor posibile și oferă posibilitatea prevenirii acestora. De asemenea, se dorește să se înțeleagă mai bine factorii care pot avea un rol esențial în evoluția pacientului postoperațiu.

Abordarea din această lucrare este aceea ca, pe baza datelor pe care le avem și prin utilizarea a doi algoritmi de învățare automată, să obținem un model capabil să estimeze probabilitatea ca, în urma intervenției chirurgicale, persoana să întâmpine sau nu complicații. Integrarea algoritmilor de învățare automată în gestionarea prezicerilor medicale reprezintă un potențial mare pentru îmbunătățirea diagnosticului, tratamentului și monitorizării pacienților. Această colaborare interdisciplinară facilitează o abordare mai precisă și personalizată a îngrijirii pacienților, contribuind astfel la optimizarea procesului medical.

Astfel, putem concluziona faptul că modelele create prin intermediul algoritmilor de învățare automată eficientizează procesele din domeniul sănătății, contribuind la îmbunătățirea rezultatelor clinice.

1.2 Structura lucrării si contribuția originală

Prin această lucrare dorim să prezentăm atât aspectele teoretice ale modelelor de învățare automată, cât și implementarea acestora pe datele medicale, astfel încât să putem contura concluzii cât mai exacte.

Principala contribuție a acestei lucrări este aceea de a prezenta și compara rezultatele obținute prin algoritmii de învățare automată folosiți. Cei doi algoritmi pe care îi vom prezenta în această lucrare sunt Random Forest și Decision Tree.

A doua contribuție constă în conturarea unui model care să prezică cât mai precis scorul de Endoscopie Sinusală Perioperatorie, pentru a oferi o mai bună înțelegere a gravității situației pacientului înainte de operație.

A treia contribuție constă în găsirea celor mai importanți factori care ajută la determinarea scorului POSE. Acești factori ajută la conturarea unei imagini de ansamblu a fiecărui caz în parte cu mult înainte de operație.

Lucrarea de față este structurată în șase capitole, după cum urmează.

Primul capitol conține o scurtă descriere a lucrării și a motivației pentru care s-a ales realizarea acesteia. De asemenea, se descrie problema medicală care se dorește să fie analizată.

Al doilea capitol oferă o descriere mai detaliată a subiectului pe care lucrarea îl prezintă și motivația pentru care s-a ales folosirea algoritmilor de învățare automată.

Al treilea capitol prezintă lucrările științifice care abordează subiectul rinosinuzitei cronice cu polipi nazali și le compară cu lucrarea prezentă, fiind descrise atât diferențele cât și elementele comune dintre acestea.

Al patrulea capitol oferă informații detaliate despre cei doi algoritmi folosiți, Decision Tree și Random Forest, prezentând motivele pentru care aceștia au fost folosiți în această lucrare.

Al cincilea capitol descrie aplicarea algoritmilor și testarea modelelor implementate, cât și detalii despre datele folosite. Se vor prezenta cele trei iterații prin care s-a trecut până la determinarea celui mai bun model. De asemenea, sunt prezentate rezultatele și sunt comparate pentru a putea determina care model și algoritm a oferit cea mai bună performanță.

Al șaselea capitol conține concluziile lucrării și cele mai importante aspecte care au fost prezentate pe parcurs. De asemenea, se prezintă posibile îmbunătățiri care se pot aduce în viitor.

Capitolul 2

Problema științifică

2.1 Definirea problemei

Problema abordată în această lucrare se referă la predicția scorurilor de endoscopie sinusală perioperatorie (POSE) la pacienții cu rinosinuzită cronică cu polipi nazali folosind algoritmi de învățare automată. Importanța acestei probleme constă în potențialul ei de a îmbunătăți evaluarea preoperatorie și procesele de luare a deciziilor în practica medicală. Scorurile POSE ajută la evaluarea riscurilor asociate cu intervențiile chirurgicale.

Prin această lucrare se dorește crearea unui model care să prezică starea unei persoane în urma operației de rinosinuzită cronică cu polipi nazali. S-a ales folosirea algoritmilor de învățare automată, datorită complexității subiectului și a factorilor implicați care interacționează în stabilirea unor concluzii mai precise.

Se vor folosi algoritmi de învățare supervizată, deoarece aceștia au avantaje care ne sunt utile în acest caz. Metodele tradiționale de evaluare a pacienților se pot baza pe evaluări subiective și experiența clinică a medicului, în timp ce algoritmii de învățare automată oferă o abordare obiectivă și bazată pe date, reducând astfel potențiale erori.

Un alt motiv pentru care s-a ales folosirea acestor algoritmi este faptul că avem date etichetate și clare, ceea ce ne ajută să definim un obiectiv clar în formarea modelului. De asemenea, o parte din algoritmii de învățare supervizată, cum ar fi arborii de decizie sunt ușor de interpretat și de înțeles, oferind astfel o transparență în ceea ce privește rezultatele modelului creat. Datele pe care le vom folosi sunt atât date personale ale pacientului: vârsta, gen, fumător, cât și date medicale: astm, alergii și evaluări ale stării persoanei înainte și după operație.

Pentru construirea modelului vom folosi algoritmi de clasificare cu partiționare multiplă, deoarece valorile scorului POSE sunt între 0 și 16 pentru fiecare fosa nazală. Avantajele utilizării algoritmilor de clasificare includ obiective clare, interpretabilitate, versatilitate, analiza importanței caracteristicilor, rezultate probabilistice, metrici de evaluare bine stabilite. Algoritmii de clasificare pot identifica cele mai relevante caracteristici sau variabile. Analizând scorurile de importanță a caracteristicilor, putem obține o perspectivă asupra factorilor de bază care conduc deciziile de clasificare, ajutând la selectarea caracteristicilor și la înțelegerea mult mai bună a problemei.

Putem concluziona astfel că, problema descrisă în această lucrare poate aduce un beneficiu zonei medicale prin înțelegerea mai detaliată a legăturilor dintre caracteristicile fiecărei persoane și starea de sănătate a acestuia după operație. Utilizarea algoritmilor inteligenți poate îmbunătăți semnificativ precizia și eficiența acestor procese, contribuind la diagnosticarea precoce a afecțiunilor, la identificarea celor mai eficiente strategii de tratament și la îmbunătățirea predicțiilor pentru pacienți.

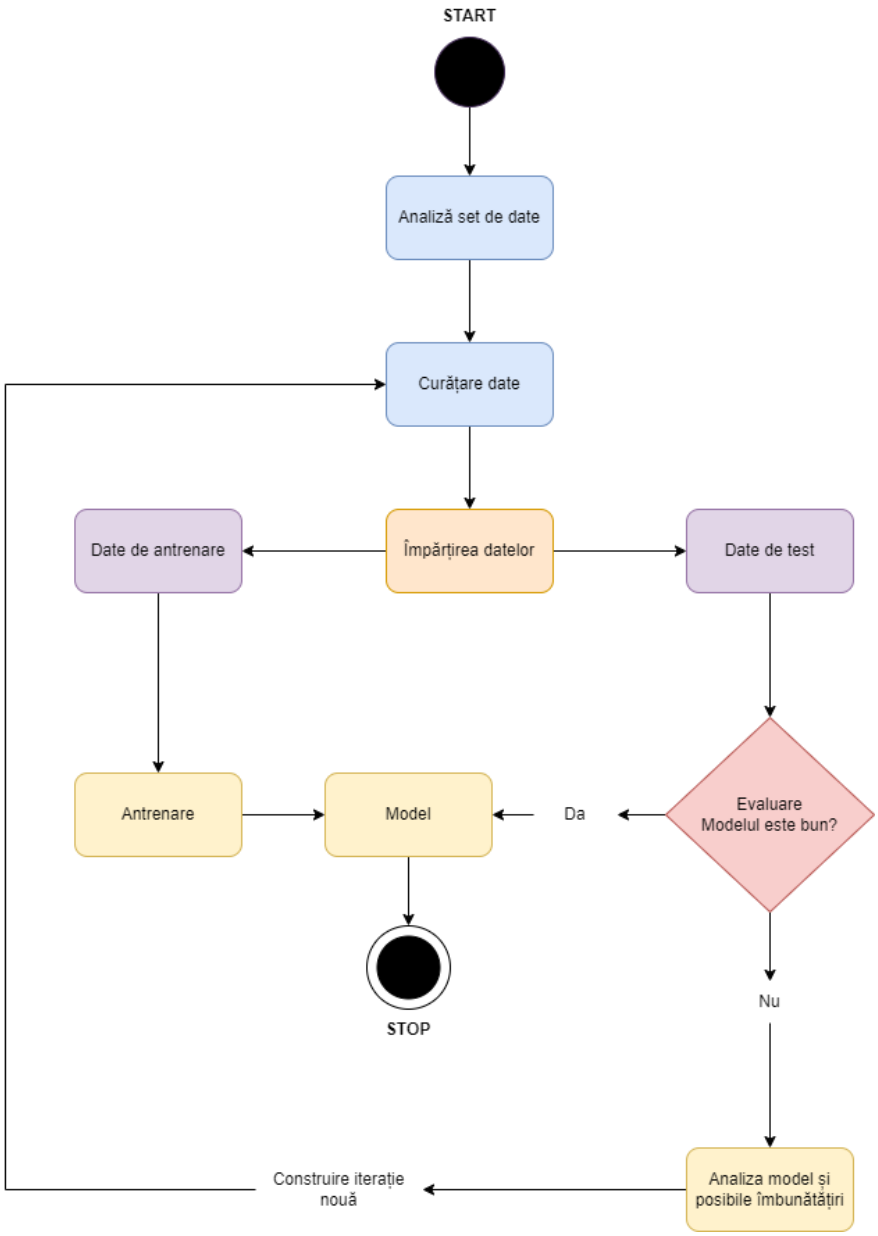

Figura 2.1.1: Fluxul de realizare al modelului pe baza algoritmului de învățarea automată

În Figura 2.1.1 se poate vedea fluxul prin care se trece pentru construirea modelului. Se începe cu analiza și înțelegerea datelor pe care le avem pentru ca apoi să putem curăța setul de date. Curățarea constă atât în eliminarea înregistrărilor care nu sunt complete, cât și în modificarea tuturor câmpurilor în valori numerice.

Următorul pas constă în împărțirea datelor în date de test și date de antrenare. Setul de date pentru testare este utilizat pentru a evalua performanța modelului de învățare automată pe date noi după ce acesta a fost antrenat.

După evaluarea capacității modelului de a prezice corect scorul POSE se decide dacă modelul este unul bun sau dacă avem nevoie de o nouă iterație. Dacă se dorește să se construiască o nouă iterație, atunci procesul se reia începând cu pasul de curățare a datelor.

Capitolul 3

Metode existente de rezolvare a problemei

În procesul de cercetare și analiză a informațiilor despre rinosinuzita cronică cu polipi nazali s-au luat în considerare lucrări și studii deja existente pe acest subiect. S-au selectat patru studii care conțin informații relevante pentru lucrarea de față. Prin prezentarea lor putem înțelege mai bine contextul lucrării prezente și putem aduce argumente suplimentare în privința abordărilor luate.

Primul articol pe care îl vom discuta este "Impact of Perioperative Systemic Steroids on Surgical Outcomes in Patients With Chronic Rhinosinusitis With Polyposis: Evaluation With the Novel Perioperative Sinus Endoscopy (POSE) Scoring System" [1]. Obiectivul principal al acestui studiu este acela de a evalua impactul pe care steroizii sistemici preoperatori îl pot avea asupra rezultatelor chirurgicale ale pacienților cu rinosinuzită cronică cu polipoză. Studiul își propune să evalueze eficacitatea terapiei cu steroizi în îmbunătățirea rezultatelor postoperatorii, iar pentru a evalua pacienții se folosește scorul POSE.

Studiul evidențiază faptul că rezultatele chirurgicale ale pacienților se îmbunătățesc semnificativ dacă aceștia au un tratament cu steroizi sistemici înainte de operație. De asemenea, se evidențiază faptul că cei care nu au primit un tratament au o evoluție mai slabă.

În articol este discutat și scorul POSE ca fiind un scor care îmbunătățește evaluarea stării pacienților după operație. De asemenea, în articol se discută faptul că scorul POSE pare să fie mult mai sensibil față de alte scoruri folosite, cum este Lund-McKay, astfel folosirea scorului POSE aduce un avantaj atunci când un pacient este evaluat. Între scorul Lund-McKay și POSE s-a descoperit o corelare ridicată, ceea ce validează faptul că acest scor este important în evaluarea stării unui pacient.

Al doilea studiu este "Subepithelial neutrophil infiltration as a predictor of the surgical outcome of chronic rhinosinusitis with nasal polyps" [2], unde, spre deosebire de articolul precedent, sunt folosiți algoritmi de învățare automată. Obiectivul principal al acestui studiu este înțelegerea modului în care densitatea neutrofilelor din stratul subepitelial al țesuturilor polipului nazal se corelează cu rezultatele postoperatorii.

Studiul s-a folosit de algoritmul Decision Tree pentru a prezice rezultatele chirurgicale la pacienții cu rinosinuzită cronică cu polipi nazali. Algoritmul Decision Tree a demonstrat în acest studiu faptul că celulele HNE-pozitive subepiteliale, scorul Lund-McKay și endotipul au fost factori critici pentru rezultatele chirurgicale ale pacienților. De asemenea, pentru pacienții cu mai mult de 45 de celule HNE-pozitive subepiteliale exista o șansă de 75% ca rezultatele chirurgicale să fie slabe.

Algoritmul Random Forest a avut o acuratețe de 84,04% în prezicerea rezultatelor chirurgicale. Scorul Lund-McKay, vârsta și numărul de celule HNE-pozitive subepiteliale au fost în acest caz cei mai importanți factori pentru rezultatele chirurgicale.

Un alt studiu care merită să fie menționat este "Chronic Rhinosinusitis with Nasal Polyps and Asthma" [3], care se concentrează pe asocierea dintre rinosinuzita cronică cu

polipi nazali și astm, cu scopul de a oferi perspective asupra caracteristicilor clinice. Rinosinuzita cronică cu polipi nazali și astmul bronșic sunt asociate cu o severitate crescută a bolii, rezultate mai slabe ale tratamentului și calitatea vieții afectată.

În articolul „Research advances in roles of microRNAs in nasal polyp”[4] se discută despre corelația dintre miARN-uri și rinosinuzita cronică cu polipi nazali. Studiul evidențiază faptul că modificările în conținutul endogen al miARN-urilor au efecte atât asupra producției de citokine inflamatorii, cât și asupra remodelării căilor respiratorii în rinosinuzita cronică cu polipi nazali. Articolul subliniază importanța miARN-urilor în tratarea rinosinuzitei cronice cu polipi nazali.

Lucrarea de față se folosește de parametrii asemănători, cum ar fi dacă pacientul are sau nu astm, folosirea unui tratament după operație. De asemenea, avem evaluări similare cum ar fi Lund-McKay și POSE, care s-au dovedit a fi importante în înțelegerea mai buna a stării pacienților. Setul de date din această lucrare conține două coloane pentru miARN-uri, miR-125 și miR-203, care ne vor ajuta în trasarea unor concluzii în legătură cu importanța miARN-urilor.

Putem concluziona astfel că, prin utilizarea informațiilor deja existente și a algoritmilor de învățare automată, putem aduce noi informații legate de rinosinuzita cronică cu polipi nazali, dar putem și să validăm încă o dată informațiile deja prezentate în articole.

Capitolul 4

Abordarea investigată

În Figura 4.1 s-a ilustrat fluxul prin care setul de date trece din starea inițială până când este gata să devină set de antrenament sau de test. Crearea și selecția datelor reprezintă un proces esențial în realizarea unui model cu o acuratețe cât mai mare.

În ceea ce privește setul nostru de date, s-au efectuat următoarele modificări:

Fumător:

- "Da" a fost înlocuit cu 1
- "Nu" a fost înlocuit cu 0

Gen:

- Valorile "M" au fost înlocuite cu 0
- Valorile "F" au fost înlocuite cu 1

Tratament postoperator:

- "Da" a fost înlocuit cu 1
- "Nu" a fost înlocuit cu 0

Pentru câmpurile Lund-Mackay, Endoscopy score, POSE 6 luni, POSE 1 an, care conțin scorul pentru fiecare nară, s-a făcut media celor două valori. După realizarea acestor modificări, s-au eliminat o parte din rândurile unde existau valori lipsă. Câmpurile unde nu s-au găsit valori au fost Fumător, Preop HPQ-9, POSE 6 luni. S-a ales eliminarea acestor rânduri pentru a nu impacta rezultatele modelului. În continuare, datele au fost împărțite într-un set de antrenament care să construiască modelul și un set de test prin care să putem evalua performanțele modelului.

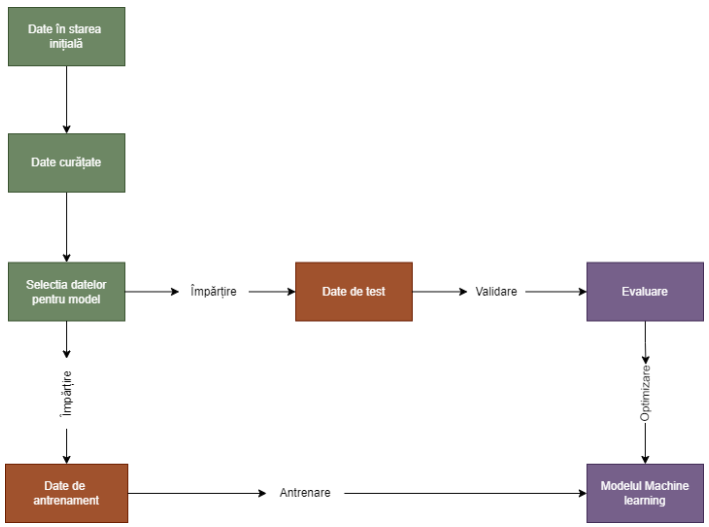

Figura 4.1: Fluxul de creare al modelului de învățare automată

Decision Tree (Arbore Decizional)

Primul algoritm pe care îl vom folosi este algoritmul de clasificare Arbore de decizie. Este un algoritm cunoscut care se folosește atât pentru sarcini de regresie cât și pentru sarcini de clasificare. Un motiv pentru care s-a ales folosirea lui este faptul că procesul prin care se trece pentru luarea unei decizii este ușor de înțeles.

Inițializare:

Algoritmul începe prin inițializarea arborelui de decizie cu setul de date de antrenare și caracteristicile disponibile.

Selectarea Diviziei Optime:

Se selectează criteriul de divizare optim pentru a separa setul de date în funcție de impuritățile prezente în diferitele noduri ale arborelui.

Divizarea Datelor:

Setul de date este divizat în funcție de caracteristica selectată și pragul asociat. Se creează două subseturi: unul în care valorile sunt mai mici sau egale cu pragul, iar celălalt în care valorile sunt mai mari decât pragul.

Construirea Arborelui:

Procesul de divizare este recursiv și continuă până când se îndeplinesc criteriile de oprire, cum ar fi atingerea adâncimii maxime a arborelui sau când nu mai există caracteristici disponibile pentru divizare.

Evaluarea Performanței Modelului:

După construirea arborelui, acesta este evaluat folosind setul de date de testare pentru a determina precizia și performanța sa generală.

Pruning (Opțional):

În unele cazuri, se poate aplica tăierea arborelui pentru a preveni supraînvățarea și pentru a îmbunătăți generalizarea modelului. Acest lucru implică eliminarea unor noduri sau ramuri care nu contribuie semnificativ la îmbunătățirea performanței.

Predicție:

Arborele de decizie antrenat este utilizat pentru a face predicții pe datele noi, folosind caracteristicile acestora pentru a naviga prin arbore și pentru a determina clasa corectă.

Evaluare Suplimentară:

După obținerea predicțiilor, este posibil să se evalueze performanța modelului pe baza altor metrici, cum ar fi matricea de confuzie, acuratețea, precizia, pentru a obține o înțelegere mai detaliată a performanței modelului.

În figura 4.2 de mai jos prezentăm un posibil rezultat al aplicării algoritmului de Decision Tree pe un set de date legat de vreme pentru a prezice dacă astăzi va ploua sau nu.

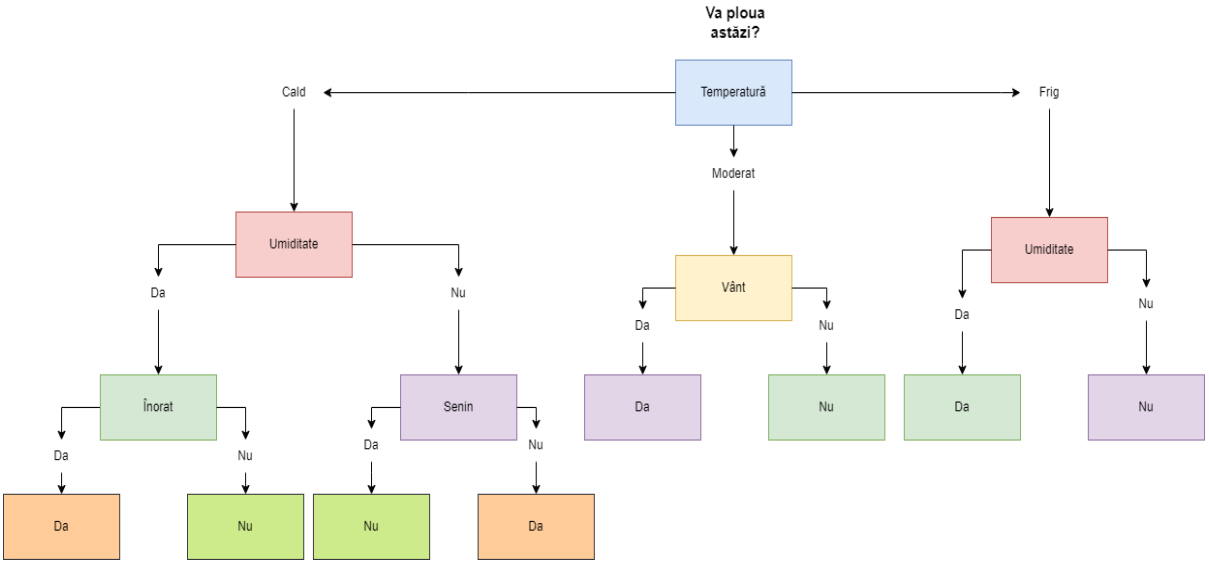

Figura 4.2: Diagrama pentru algoritmul Arbore de decizie pentru a determina dacă azi va ploua

Random Forest (Pădure de arbori decizionali)

Random Forest este un algoritm puternic de învățare în machine learning. Funcționează prin crearea mai multor arbori de decizie în timpul fazei de instruire. Fiecare arbore este construit folosind un subset aleatoriu al setului de date pentru a măsura un subset aleatoriu de factori în fiecare partiție.

Procesul prin care se iau decizii prin algoritmul Random Forest este:

Inițializare:

Algoritmul începe prin inițializarea unui număr specificat de arbori de decizie, fiecare cu un set aleatoriu de date de antrenare.

Antrenare Arbori de Decizie:

Pentru fiecare arbore din Random Forest, se selectează un subset aleatoriu de date de antrenare din setul complet. Fiecare arbore este antrenat folosind acest subset de date și o submulțime de factori, ceea ce ajută la diversificarea arborilor și la reducerea corelațiilor între aceștia.

Predicție:

După antrenare, fiecare arbore din Random Forest poate fi utilizat pentru a face predicții pe datele de testare sau pe datele noi. Pentru clasificarea cu clase multiple, fiecare arbore oferă o predicție pentru clasa fiecărui exemplu de testare.

Votare Majoritară:

Clasa finală este determinată printr-un vot majoritar, unde clasa care primește cel mai mare număr de voturi din toți arborii este considerată clasa finală pentru acel exemplu.

Evaluarea Performanței Modelului:

Performanța modelului de Random Forest este evaluată pe baza predicțiilor făcute pe datele de testare sau prin Cross-Validation. Metrici comune utilizate pentru evaluarea performanței includ acuratețea, precizia și scorul F1.

Importanța Factorilor:

Random Forest poate furniza, de asemenea, informații despre importanța fiecărui factor în cadrul modelului. Această importanță este calculată pe baza contribuției fiecărui factor la îmbunătățirea performanței modelului.

Optimizarea Parametrilor (Opțional):

În unele cazuri, parametrii Random Forest, cum ar fi numărul de arbori, adâncimea maximă a arborilor și numărul de factori utilizați în fiecare split, pot fi optimizați.

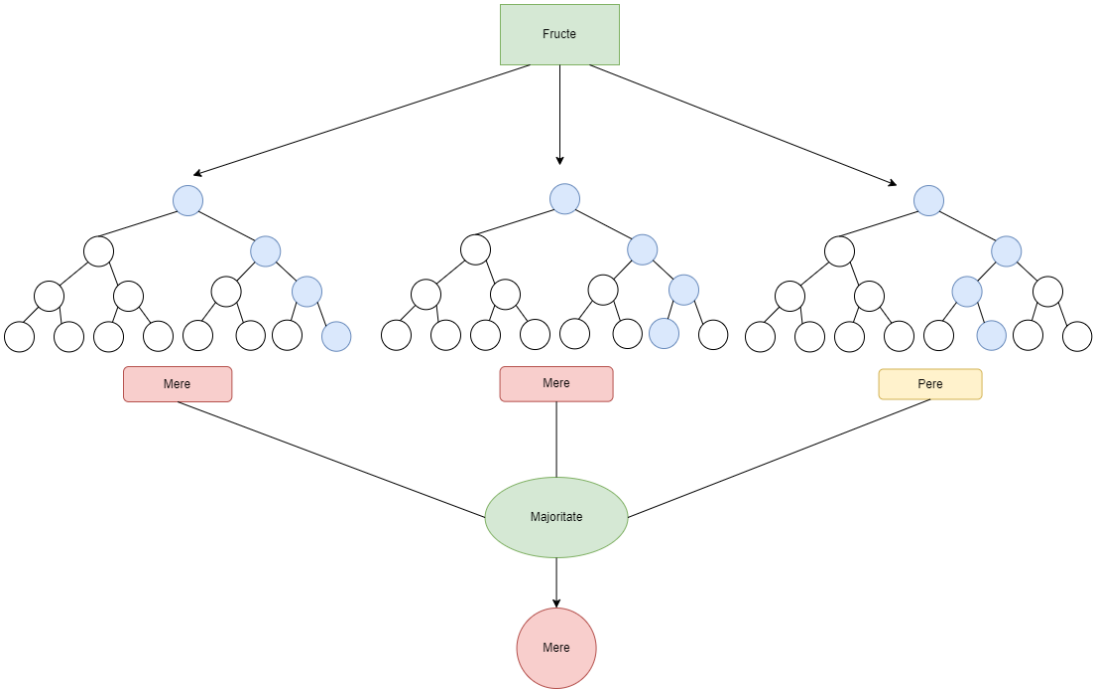

Figura 4.3: Diagrama pentru algoritmul de Random Forest pentru a determina tipul de fruct

În Figura 4.3 de mai sus este reprezentat modul în care deciziile sunt luate în algoritmul Random Forest. Vedem cum în timpul fazei de antrenare, fiecare arbore de decizie generează un rezultat și apoi prezice decizia finală pe baza celor mai frecvente rezultate.

Deși Random Forest este o colecție de Decision Tree, există diferențe semnificative între cele două abordări. Random Forest are tendința de a fi mai rezistent la overfitting decât un singur Decision Tree, datorită faptului că rezultatele multiple ale arborilor sunt mediate pentru a lua decizia finală. Un Decision Tree este ușor de interpretat și de înțeles, deoarece poate fi reprezentat grafic. Random Forest, din cauza complexității sale mai mari, poate fi mai dificil de interpretat, dar poate oferi o performanță mai bună în multe.

Bibliografie

[1] Wright, E. D., & Agrawal, S. (2007). Impact of perioperative systemic steroids on surgical outcomes in patients with chronic rhinosinusitis with polyposis: evaluation with the novel Perioperative Sinus Endoscopy (POSE) scoring system. The Laryngoscope, 117(S115), 1-28. https://onlinelibrary.wiley.com/doi/10.1097/MLG.0b013e31814842f8

[2] Subepithelial neutrophil infiltration as a predictor of the surgical outcome of chronic rhinosinusitis with nasal polyps https://www.rhinologyjournal.com/Rhinology_issues/manuscript_2701.pdf

[3] Chronic Rhinosinusitis with Nasal Polyps and Asthma, Tanya M. Laidlaw, Joaquim Mullol, Katharine M. Woessner, Nikhil Amin, Leda P. Mannent,

https://www.sciencedirect.com/science/article/pii/S2213219820311132

[4] Research advances in roles of microRNAs in nasal polyp, Niu Zhipu, Huo Zitao, Sha Jichao, and Meng Cuida