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

Prof. Dioşan Laura

SANDU ANA-ALEXANDRA

2023-2024

Abstract

Prin lucrarea de față ne propunem ca prin intermediul algoritmilor de clasificare multiplă Decision Tree și Random Forest să construim un model care să prezică cât mai bine Endoscopia Sinusală Perioperatorie (POSE). Deoarece s-au folosit doi algoritmi de învățare automată, am putut să comparăm performanțele și să ne conturăm o idee mai bună asupra rolului acestora în predicția scorului POSE. De asemenea, cei doi algoritmi de învățare automată oferă posibilitatea de a vizualiza factorii cei mai importanți care au dus la predicțiile modelului. Astfel, putem să tragem concluzii importante despre factorii clinici și importanța lor în procesul de tratare a pacienților.

Pentru evaluarea fiecărui model s-au folosit metricile: accuracy, precision, recall și F1 score care ne oferă o mai bună înțelegere a performanței în ceea ce privește predicțiile făcute. Pentru a avea o perspectivă obiectivă asupra rezultatelor ne-am folosit de tehnica Cross-Validation pentru a vedea capacitățile modelului de generalizare.

Setul de date utilizat conține 86 de înregistrări, iar în urma curățării acestuia am folosit 79 de înregistrări în crearea modelului. Factorii folosiți au fost Age, Astm, Eo%/val, Fumător, Gen, Initial SNOT, Preop HPQ-9, Lund-Mackay, miR-125, miR-203, Tratament postop, iar factorii care au fost targetați au fost POSE 6 luni și POSE 1 an.

În ceea ce privește rezultatele obținute în această lucrare, putem spune că a existat o evoluție în cele trei iterații prezentate, unde iterația a treia a avut cele mai bune rezultate. De asemenea, modelul Random Forest a depășit modelul Decision Tree, obținând o acuratețe mai bună. Rezultatele obținute demonstrează potențialul pe care învățarea automată îl are în ceea ce privește sprijinirea proceselor de luare a deciziilor medicale pentru pacienții cu rinosinuzită cronică cu polipi nazali. Random Forest a obținut pentru predicția scorului POSE 6 luni o acuratețe de 77% și pentru POSE 1 an o acuratețe de 82%. Factorii care au avut o mare importanță în determinarea predicțiilor au fost miR-125, miR-203, vârsta și Initial SNOT.

În concluzie, putem spune faptul că lucrarea de față oferă informații și concluzii importante în ceea ce privește predicția scorului de Endoscopie Sinusală Perioperatorie (POSE), ajutând astfel la înțelegerea mai bună a factorilor determinanți și a modului în care un pacient poate fi tratat.

Cuprins

Lista de tabele

Lista de figuri

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 predicția probabilităților unor astfel de urgențe medicale după intervențiile chirurgicale. Această prognoză 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 post-operatie.

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 cu o precizie cât mai mare 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 are o structură de șase capitole, fiecare capitol având următorul rol:

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.

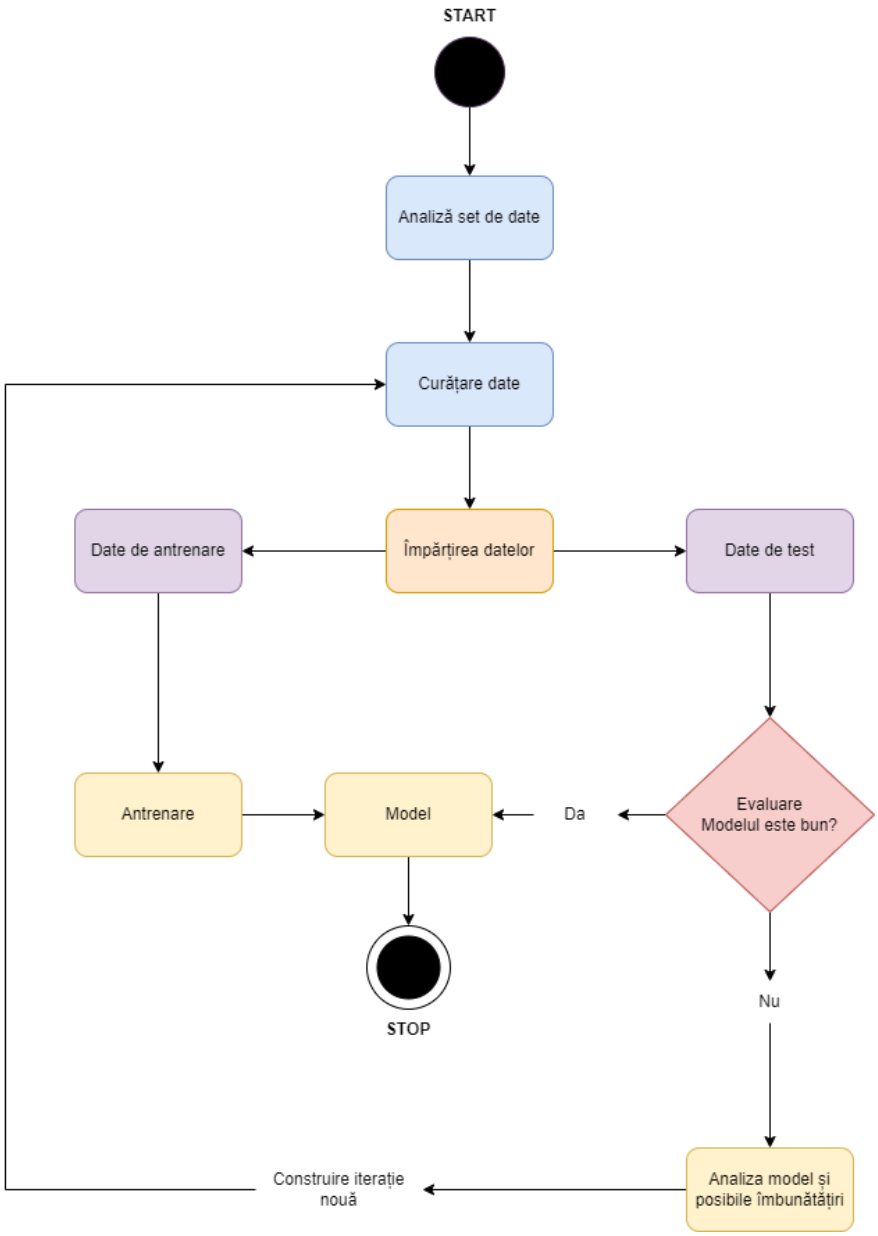
Prin această lucrare se dorește crearea unui model care să prezică starea unei persoane în urma operației de rinosinuzită cronică cu polipi nazali. S-a ales folosirea algoritmilor de învățare automată, datorită complexității subiectului și a factorilor implicați care interacționează în stabilirea unor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



Figură 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 va fi utilizat pentru a determina 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.

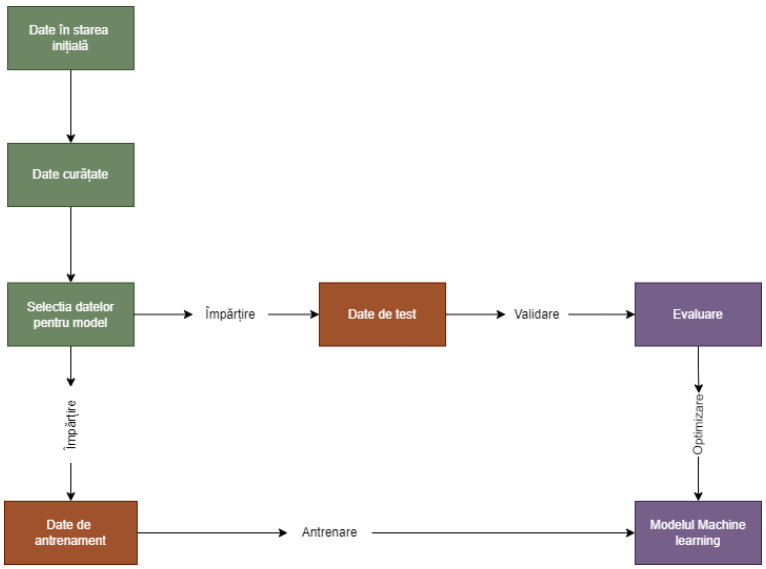
Lucrarea de față se folosește de parametrii asemănători, cum ar fi dacă pacientul are sau nu astm, folosirea unui tratament după operație. De asemenea, avem evaluări similare cum ar fi Lund-McKay și POSE, care s-au dovedit a fi importante în înțelegerea mai buna a stării pacienților. Setul de date din 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.



Figură 4.1: Fluxul de creare al modelului de învățare automată

Setul de date pe care îl vom folosi are un număr de 86 de pacienți. Atributele setului de date sunt următoarele: Vârstă, Astm, Alergii/Intoleranțe, Eo%/val, Fumător, Gen, Scor inițial SNOT, Preop HPQ-9, Lund-Mackay, Scor endoscopie, mir 125, mir 203, SNOT 6 luni, HPQ-9 6 luni, POSE 6 luni, SNOT 1 an, HPQ-9 1 an, POSE 1 an, Tratament postoperator. Câmpurile Fumător, Gen și Tratament postoperator sunt valori de tip text, restul atributelor sunt valori numerice.

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

În Tabelul 4.1 sunt prezenți patru pacienți și valorile pentru fiecare parametru în parte, pentru a ne contura o imagine de ansamblu mai clară asupra datelor.

	Age	Astm	Alergii /Intoler 1/2	Eo\n%/val	Fum	Sex	Initial SNOT	Preop HPQ-9	Lund-Mackay	Endoscopy score	mir 125	mir 203	SNOT 6 luni	HPQ-9 6 luni	POSE 6 luni	SNOT 1 an	HPQ-9 1 an	POSE 1 an	Tratament postop
0	74.0	0.0	0.0	1.70	0.0	0.0	44.0	7.0	14.0	4.0	1.80	0.19	2.0	4.0	4.0	2.0	2.0	0.0	1.0
1	56.0	0.0	0.0	0.00	0.0	0.0	16.0	12.0	22.0	6.0	2.70	0.32	13.0	9.0	5.0	46.0	10.0	8.0	1.0
2	61.0	1.0	1.0	0.00	0.0	1.0	11.0	2.0	12.0	3.0	1.11	0.53	0.0	0.0	1.0	0.0	0.0	0.0	1.0
3	53.0	1.0	2.0	0.00	0.0	1.0	67.0	5.0	22.0	6.0	2.39	0.50	19.0	3.0	5.0	29.0	3.0	6.0	1.0
4	52.0	0.0	1.0	0.64	0.0	1.0	24.0	3.0	19.0	5.0	0.97	0.83	2.0	2.0	6.0	5.0	3.0	13.0	0.0

Tabel 4.1: Tabel cu datele a patru pacienți

Pentru atributele 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.

Pentru a evalua performanța algoritmilor Decision Tree și Random Forest în cazul clasificării multiple ne vom folosi de următoarele metrici [12]:

Accuracy (Acuratețe):

Acuratețea reprezintă proporția de predicții corecte din totalul predicțiilor.

$$\textbf{Acuratețe} = \frac{\text{Numărul de predicții corecte}}{\text{Numarul total de predicții}}$$

Precision (Precizie):

Precizia reprezintă proporția de predicții corecte din totalul predicțiilor pozitive identificate de model.

$$\textbf{Precizia} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

True Positives reprezintă numărul de instanțe pozitive care au fost corect identificate de model. False Positives reprezintă numărul de instanțe negative incorect identificate de model.

Recall:

Recall reprezintă proporția de predicții pozitive corecte din totalul instanțelor pozitive reale.

$$\textbf{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

True Positives reprezinta numarul de instanțe pozitive corect identificate de model. False Negatives este numărul de instanțe pozitive incorect identificate ca negative de model.

F1 Score:

Acest scor combinând atât precizia cât și recall-ul. Se realizează media armonică dintre precizie și recall.

$$\textbf{Scorul F1} = 2 \times \frac{\text{Precizie} \times \text{Recall}}{\text{Precizie} + \text{Recall}}$$

4.1 Decision Tree (Arbore Decizional)

Primul algoritm pe care îl vom folosi este algoritmul de clasificare Arbore de decizie [5]. 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.

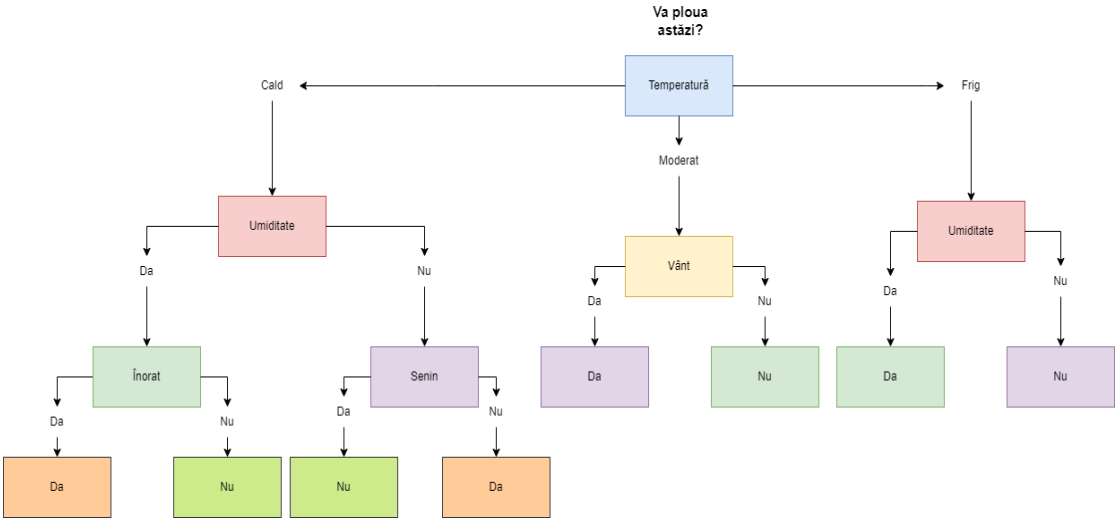
Predicție:

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:

După obținerea predicțiilor, este posibil să se evalueze performanța modelului pe baza metricilor, cum ar fi acuratețea, precizia, recall-ul pentru a obține o înțelegere mai detaliată a performanței modelului.

În figura 4.1.1 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.



Figură 4.1.1: Diagrama pentru algoritmul Arbore de decizie pentru a determina dacă azi va ploua

4.2 Random Forest (Pădure de arbori decizionali)

Random Forest este un algoritm puternic de învățare în machine learning [6]. 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:

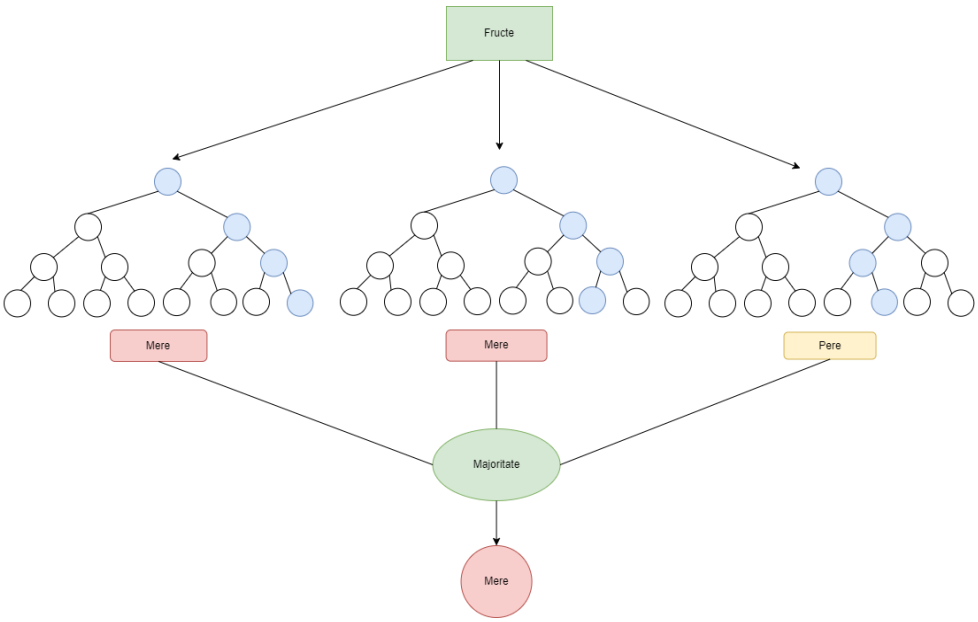
Performanța modelului de Random Forest este evaluată pe baza predicțiilor făcute pe datele de testare sau prin Cross-Validation. Metrici comune utilizate pentru evaluarea performanței includ acuratețea, precizia și scorul F1.

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.



Figură 4.2.1: Diagrama pentru algoritmul de Random Forest pentru a determina tipul de fruct

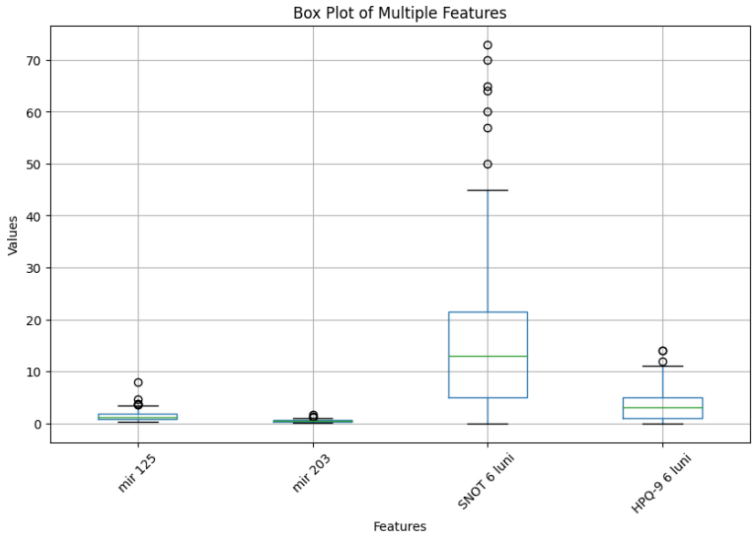
În Figura 4.2.1 de mai sus este reprezentat modul în care deciziile sunt luate în algoritmul Random Forest. Vedem cum în timpul fazei de antrenare, fiecare arbore de decizie generează un rezultat și apoi prezice decizia finală pe baza celor mai frecvente rezultate.

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.

S-au construit trei iterații pentru a ajunge la un model cât mai bun. S-a pornit de la o iterație în care setul de date a fost modelat după criteriile prezentate mai sus, toate înregistrările fiind transformate în valori numerice, iar pentru evaluările ambelor fose nazale s-a făcut media.

Pentru iterația a doua s-au modificat valorile din câmpurile POSE 6 luni și POSE 1 an, reprezentând câmpurile pe care dorim să le prezicem. Intervalul valorilor era unul între 0 și 16, astfel am clasificat aceste valori în patru categorii. S-a dovedit faptul că această clasificare a ajutat la îmbunătățirea modelului.

În a treia iterație ne-am folosit de box plot pentru a vizualiza dacă datele noastre conțin valori outlier. Se poate observa în Figura 4.2.2 de mai jos, faptul că există valori care nu se încadrează. Astfel, s-a folosit algoritmul Isolation Forest pentru a elimina anomaliile, iar din cele 79 de înregistrări au rămas 72. Rezultatele din această iterație nu au fost cu mult mai bune decât cele din a doua iterație.



Figură 4.2.2: Box Plot pentru mir 125, mir 203, SNOT 6 luni si HPQ-9 6 luni

4.3 Isolation Forest

Isolation Forest este un algoritm de învățare automată care este folosit pentru a detecta anomaliile. [7] Nu presupune nicio distribuție specifică a datelor, spre deosebire de alte metode de detecție a anomaliilor.

Isolation Forest creează un număr mare de arbori de izolare (iTrees) și pentru fiecare arbore se selectează aleatoriu un factor și un punct de tăiere. Setul de date este divizat în funcție de acest punct de tăiere, procesul continuă până în momentul în care fiecare punct se află într-un nod terminal. Se calculează astfel un scor de anomalie pentru fiecare punct de date. Un scor mare va indica faptul că acel punct este o anomalie, în timp ce un scor mic va indica faptul că punctul de date este normal.

Capitolul 5

Aplicare (Studiu de caz)

5.1 Descrierea aplicației și principalele funcționalități

Scopul acestei lucrări este acela de a modela, prin intermediul algoritmilor de învățare automată, un model cu o acuratețe cât mai mare. Vom putea compara rezultatele celor doi algoritmi aplicați pe setul nostru de date și, de asemenea, vom putea determina care factori sunt determinanți în luarea unei decizii. Procesul prin care s-a trecut pentru realizarea modelului este:

Analiza datelor: Se analizează câmpurile pe care le avem, semnificația lor pentru a înțelege problema abordată. De asemenea, ne ajută să înțelegem distribuția datelor și să identificăm tiparele.

Procesarea datelor: Curățarea și pregătirea datelor pentru antrenarea modelelor.

Selecția factorilor: Pe baza matricei de corelație vom alege care factori îi vom folosi în construirea modelului. Dacă corelația dintre doi factori este mare, atunci vom analiza care dintre cei doi factori are o corelație mai mare cu factorul țintă și se păstrează factorul cu corelația mai mare. De asemenea, scorurile care au fost refăcute după operație la șase luni respectiv un an (SNOT 6 luni, HPQ-9 6 luni, SNOT 1 an, HPQ-9 1 an) nu au fost luate în calcul în construirea modelului, deoarece dorim să folosim date pre-operatorii și intra-operatorii. Factorul Tratament postoperator a fost luat în considerare în construirea modelului, deoarece dorim să vedem dacă tratarea pacientului post-operatoriu poate avea un impact asupra stării pacientului.

Antrenarea modelelor: Antrenăm modelele pe baza algoritmilor Decision Tree și Random Forest.

Evaluarea performanței: Evaluăm modelele și analizăm dacă se pot aduce îmbunătățiri printr-o nouă iterație.

În fiecare dintre cele trei iterații se urmărește prezicerea scorului POSE după șase luni și după un an. Dorim să folosim ambele coloane pentru a vedea care dintre cei doi factori poate fi prezis mai precis. Astfel, după vizualizarea datelor și eliminarea coloanelor cu ajutorul matricei de corelație, se folosesc algoritmii Decision Tree și Random Forest pentru a crea un model pentru prezicerea scorului POSE.

5.2 Designul aplicației

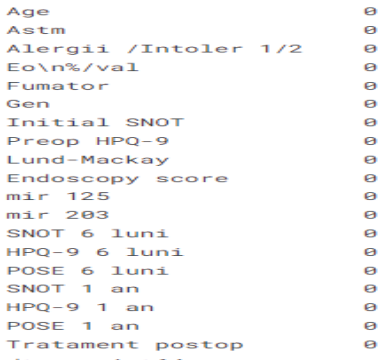
Setul nostru inițial de date conținea 86 de pacienți, iar în urma eliminării celor care nu aveau date complete au rămas 79. Datele pacienților sunt următoarele:

- **Prezența unui astm bronșic:** Este notat cu 1 pentru da și cu 0 pentru nu. După cum am văzut în studiul prezentat mai sus, pacientul prezintă posibilele agravări ale stării dacă acesta are astm și rinosinuzită cronică cu polipi nazali.

- **Vârsta**
- **Prezența unei alergii de mediu sau a alergiei/intoleranței la Aspirină:** (sunt notate cu 0, 1, 2) : alergiile de mediu au fost notate cu 1, alergia la aspirină a fost notată cu 2 și cazul în care pacientul nu are alergii s-a notat cu 0.
- **Valoarea eozinofilelor din sângele periferic:** valorile normale a eozinofilelor sunt între 0-5%, respectiv 0-35x109/l.
- **Status de fumător:** valoarea 1 dacă este fumător și 0 dacă nu este.
- **Genul:** valoarea 1 pentru feminin și 0 pentru masculin.
- **SNOT inițial:** acest scor se referă la calitatea vieții pacientului cu polipoză nazală. Acest scor a fost făcut înainte de operație. Scorul are următoarele încadrări:
    - 0-10 reprezintă lipsa unei probleme/problemă ușoară
    - 11-40 problemă moderată
    - 41-69 moderat spre sever
    - 70-100 sever spre foarte grav
- **Preop HPQ9:** această evaluare ne ajută să înțelegem starea psihică a pacientului și gradul de depresie. Scorul are următoarele încadrări:
    - 0-4 fără depresie
    - 5-9 depresie ușoară
    - 10-14 depresie moderată
    - 15-19 depresie moderat spre severă
    - 20-27 depresie severă
- **Lund-Mackay:** acest scor determină existenta polipozei nazale printr-un examen CT. Scorul maxim este de 12 și acest scor este efectuat separat pentru fiecare fosa nazală.
- **Endoscopy score:** prin intermediul unei camere video se evaluează dimensiunea polipilor. Este evaluată fiecare fosa nazală. Scorul are următoarele încadrări:
    - 0- fără polipi nazali
    - 1- polipi mici
    - 2- polipi de dimensiuni medii
    - 3- polipi care ocupă toată fosa nazală
- **Tratamentul postoperator**: toți pacienții primesc recomandare de tratament postoperator cu un spray antiinflamator. S-a notat cu 1 dacă pacientul a luat tratamentul și cu 0 dacă pacientul nu a luat tratamentul.

Pentru fiecare pacient s-a făcut o evaluare după 6 luni și un an. Scorurile pentru care se face reevaluarea sunt: SNOT, HPQ9. După 6 luni și după un an a fost efectuată evaluarea Endoscopiei Sinusale Perioperatorie (POSE), iar rezultatul este minim 0- fără inflamație, maxim 16. Acest scor evaluează fiecare fosa nazală în parte. MiR 125 și 203 reprezintă microARN-uri determinate din polipii obținuți de la fiecare pacient în timpul intervenției chirurgicale. Știm pe baza studiului prezentat mai sus faptul că între micro-ARN-uri și rinosinuzita cronică cu polipi nazali există o posibilă legătură.

Putem vedea în Figura 5.2.1 faptul că toate înregistrările folosite nu au valori nule.



Figură 5.2.1: Ilustrare câmpuri cu valori nule

5.3 Implementare

Pentru realizarea modelului s-a folosit limbajul de programare Python. Am ales acest limbaj de programare, deoarece Python are o sintaxă simplă și ușor de înțeles, ceea ce îl face accesibil. [8] De asemenea, Python are multe librării special realizate pentru învățarea automată și data science. Python are o comunitate mare și activă de cercetători și oameni de știință.

Împreună cu Python, s-a utilizat Jupyter Notebook, o aplicație open-source ce permite crearea și partajarea de documente care includ cod și vizualizări de date. Jupyter Notebook permite rularea de cod în blocuri separate și permite vizualizarea imediată a rezultatelor.[9] Acest lucru facilitează experimentarea pe diferite segmente de date.

Pentru scrierea codului s-a folosit editorul Visual Studio Code [10]. Acest editor a fost dezvoltat de Microsoft. S-a ales folosirea acestui editor, deoarece interfața este intuitivă și ușor de utilizat. De asemenea, este un editor puternic și versatil care este folosit pe scară largă.

5.4 Testare

Testarea fiecărui model s-a făcut prin împărțirea setului de date inițial în două seturi, unul de antrenare și unul de test. Se evaluează acuratețea și performanța modelului pe setul de test pentru a decide dacă este un model bun sau unul care necesită modificări.

De asemenea, pentru fiecare model se folosește tehnica Cross-Validation [11], prin care setul de date este împărțit în mai multe subseturi, iar modelul este antrenat și testat de mai multe ori, folosind diferite combinații. La final, se face o medie a rezultatelor obținute din iterații. Această tehnică ne ajută să obținem o evaluare mai obiectivă a performanței modelului și putem evita astfel overfitting-ul sau underfitting-ul.

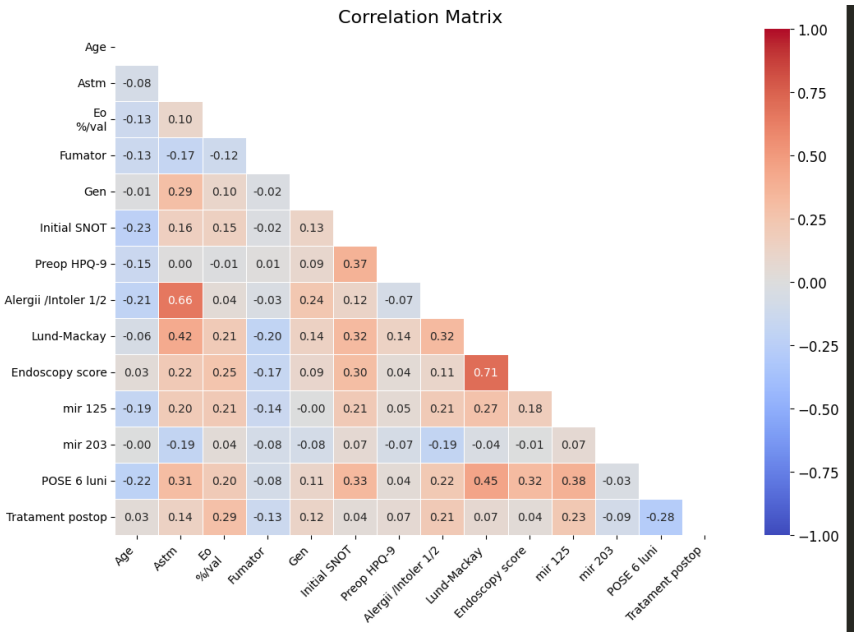
5.5 Validare numerică

5.5.1 Rezultate

Setul de date pe care îl folosim în această lucrare a fost împărțit pentru fiecare iterație și algoritm de învățare automată în:

- Set de antrenare - 70%
- Set de testare - 30%

De asemenea, câmpurile care s-au folosit în fiecare model sunt: Age, Astm, Eo%/val, Fumător, Gen, Initial SNOT, Preop HPQ-9, Lund-Mackay, miR-125, miR-203, Tratament postop. Câmpurile care au fost eliminate pe baza matricei de corelație sunt Alergii/Intoleranțe, care avea o corelație mare cu Astm-ul, și Endoscopy score, care avea o corelație mare cu Lund-Mackay, după cum se poate vedea în Figura 5.5.1.1.



Figură 5.5.1.1: Matricea de corelație

Pentru a evalua modelele ne-am folosit de următoarele metrici:

- Accuracy, precision, recall și F1 score.
- După, se folosește Cross-Validation pentru a se face media scorurilor calculate pe model.

Iterația 1

În această iterație s-au utilizat 79 de înregistrări. Modificările care s-au făcut în această iterație pe date au constat în transformarea tuturor valorilor în valori numerice (Fumător, Gen, Tratament postop) și pentru scorurile unde se evaluează ambele fose nazale s-a făcut media valorilor (Lund-Mackay, Endoscopy score, POSE).

POSE 6 luni

Model	Accuracy	Precision	Recall	F1 score
Decision Tree	25%	22%	25%	23%
Random Forest	29%	41%	29%	29%

Tabel 5.5.1.2: Valori model pentru iterația 1 cu target POSE 6 luni

Model	Accuracy Corss-Valiadtion	Precision Corss-Valiadtion	Recall Corss-Valiadtion	F1 score Corss-Valiadtion
Decision Tree	22%	15%	16%	12%
Random Forest	23%	18%	23%	20%

Tabel 5.5.1.3: Valori evaluare Cross-Validation pentru iterația 1 cu target POSE 6 luni

În această iterație se poate observa faptul că valorile obținute sunt destul de slabe, atât pentru Random Forest cât și pentru Decision Tree. Random Forest are o acuratețe mai mare (29%) comparativ cu Decision Tree (25%) pe setul de testare, ceea ce sugerează că Random Forest are o performanță generală mai bună. Valorile pentru acuratețe din Cross-Validation sunt mai mici pentru ambele modele (23% pentru Random Forest și 22% pentru Decision Tree) comparativ cu acuratețea pe setul de testare.

De asemenea, se poate observa faptul că Random Forest face mai puține predicții false bazat pe valorile preciziei. Același lucru se poate spune și despre recall, Random Forest are capacitatea să indice mai corect instanțele pozitive. Pentru scorul F1 observăm faptul că Random Forest are o valoare mai mare în comparație cu Decision Tree, ceea ce sugerează că există un echilibru mai bun între precizie și recall.

POSE 1 an

Model	Accuracy	Precision	Recall	F1 score
Decision Tree	21%	17%	21%	18%
Random Forest	21%	18%	21%	18%

Tabel 5.5.1.4: Valori model pentru iterația 1 cu target POSE 1 an

Model	Accuracy Corss-Valiadtion	Precision Corss-Valiadtion	Recall Corss-Valiadtion	F1 score Corss-Valiadtion
Decision Tree	21%	23%	20%	23%
Random Forest	20%	14%	22%	15%

Tabel 5.5.1.5: Valori evaluare Cross-Validation pentru iterația 1 cu target POSE 1 an

Ambele modele, Decision Tree și Random Forest, au performanțe similare, cu o acuratețe de 21% și scorul F1 de 18%. De asemenea, rezultatele oferite de Cross-Validation nu sunt cu mult diferite de cele pe setul de testare, indicând o generalizare relativ bună a modelelor pe date noi. Cu toate acestea, performanța ambelor modele rămâne slabă.

Iterația 2

Pentru iterația a doua s-a decis să se facă o clasificare a valorilor din câmpul POSE 6 luni și POSE 1 an. Valorile au fost încadrate astfel:

- **0-3** - valori scăzute, s-au notat cu **1**
- **4-7** - valori medii, s-au notat cu **2**

- **8-11**- valori mari, s-au notat cu **3**
- **12-16** - valori foarte mari, s-au notat cu **4**

POSE 6 luni

Model	Accuracy	Precision	Recall	F1 score
Decision Tree	71%	79%	71%	74%
Random Forest	79%	78%	79%	79%

Tabel 5.5.1.6: Valori model pentru iterația 2 cu target POSE 6 luni

Model	Accuracy Corss-Valiadtion	Precision Corss-Valiadtion	Recall Corss-Valiadtion	F1 score Corss-Valiadtion
Decision Tree	60%	52%	53%	51%
Random Forest	64%	62%	61%	62%

Tabel 5.5.1.7: Valori evaluare Cross-Validation pentru iterația 2 cu target POSE 6 luni

Se poate observa o diferență semnificativă în rezultatele din această iterație. Random Forest are o performanță mai bună față de Decision Tree pe majoritatea metricilor. Performanța modelelor pare să scadă la Cross-Validation, ceea ce poate indica o capacitate mai mică de generalizare pe date noi.

POSE 1 an

Model	Accuracy	Precision	Recall	F1 score
Decision Tree	75%	85%	75%	80%
Random Forest	71%	68%	71%	69%

Tabel 5.5.8: Valori model pentru iterația 2 cu target POSE 1 an

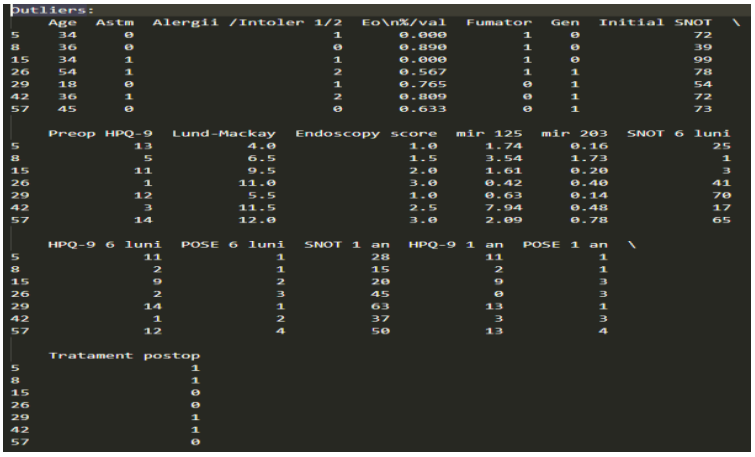
Model	Accuracy Corss-Valiadtion	Precision Corss-Valiadtion	Recall Corss-Valiadtion	F1 score Corss-Valiadtion
Decision Tree	66%	61%	63%	59%
Random Forest	64%	64%	62%	61%

Tabel 5.5.1.9: Valori evaluare Cross-Validation pentru iterația 2 cu target POSE 1 an

Pentru POSE 1 an, se observă faptul că acuratețea modelului Decision Tree de 75% este mai mare decât cea a Random Forest, care este de 71%. De asemenea, putem observa că valorile pentru precizie, recall și scorul F1 sunt mai mari pentru Decision Tree.

Iterația 3

În iterația a treia ne-am folosit de algoritmul Isolation Forest pentru a elimina anomaliile din setul de date. Astfel, din cele 79 de valori au mai rămas 72. Nu au fost eliminate multe valori, dar această eliminare poate contribui la îmbunătățirea performanței modelelor. Se pot vedea în Figura 5.5.1.2 de mai jos înregistrările care au fost eliminate.



Figură 5.5.1.2: Valorile eliminate prin Isolation Tree

POSE 6 luni

Model	Accuracy	Precision	Recall	F1 score
Decision Tree	77%	79%	77%	78%
Random Forest	77%	75%	77%	76%

Tabel 5.5.1.10: Valori model pentru iterația 3 cu target POSE 6 luni

Model	Accuracy Corss-Valiadtion	Precision Corss-Valiadtion	Recall Corss-Valiadtion	F1 score Corss-Valiadtion
Decision Tree	72%	74%	74%	75%
Random Forest	66%	67%	63%	64%

Tabel 5.5.1.11: Valori evaluare Cross-Validation pentru iterația 3 cu target POSE 6 luni

Observăm faptul că pe setul de test, ambele modele au o acuratețe de 77%, cu toate acestea, metricile pentru Decision Tree sunt ușor mai mari. Se observă o îmbunătățire a modelului față de iterația 2. În Cross-Validation, ambele modele au înregistrat scăderi ale performanței în comparație cu setul de date de testare. Decision Tree a obținut o acuratețe de 72%, în timp ce Random Forest a obținut o acuratețe mai mică, de 66%. Totuși, Random Forest a înregistrat o precizie și un scor F1 ușor mai mari decât Decision Tree în acest caz.

POSE 1 an

Model	Accuracy	Precision	Recall	F1 score
Decision Tree	73%	81%	73%	74%
Random Forest	82%	82%	82%	81%

Tabel 5.5.1.12: Valori model pentru iterația 3 cu target POSE 1 an

Model	Accuracy Corss-Valiadtion	Precision Corss-Valiadtion	Recall Corss-Valiadtion	F1 score Corss-Valiadtion
Decision Tree	72%	78%	72%	76%
Random Forest	74%	79%	72%	73%

Tabel 5.5.1.13: Valori evaluare Cross-Validation pentru iterația 3 cu target POSE 1 an

Random Forest are o performanță mai bună atât pentru valorile modelului, cât și pentru cele de la Cross-Validation. Valorile scad pentru ambele modele atunci când se evaluează modelul cu Cross-Validation. Scăderea este mai mică pentru Decision Tree, ceea ce indică o capacitate mai bună de generalizare pentru Decision Tree.

Putem concluziona faptul că modelele au evoluat pe parcursul iterațiilor, atât pentru setul de date de test, cât și pentru valorile din Cross-Validation. Iterația 3 pare să fie cea care are cele mai bune performanțe. În Iterația 3, atât modelul de Decision Tree cât și cel de Random Forest au obținut rezultate mai bune în ceea ce privește acuratețea, precizia, recall-ul și scorul F1, comparativ cu celelalte iterații. În special, pentru datele POSE 1 an, atât Decision Tree cât și Random Forest obțin o îmbunătățire semnificativă a performanței în Iterația 3, comparativ cu Iterațiile 1 și 2.

Performanța modelului Random Forest pare să fie mai stabilă și mai consistentă în timp, comparativ cu modelul Decision Tree, deoarece valorile metricilor de evaluare pentru Random Forest sunt mai puțin afectate de schimbările iterațiilor. De asemenea, predicția scorului POSE la 1 an pare să fie mai dificilă decât la 6 luni, acest fapt poate să indice faptul că evaluarea pe termen lung poate fi mai complexă și necesită o atenție sporită în dezvoltarea modelelor.

Trebuie precizat faptul că valorile fluctuează în funcție de rularea modelelor, atât pentru valorile modelului, cât și pentru cele generate de Cross-Validation. Diferența este de aproximativ 3-5% pentru toate modelele. O astfel de fluctuație a valorilor modelului și ale celor din Cross-Validation în funcție de rulare este destul de comună. O variație de aproximativ 3-5% poate fi considerată acceptabilă, mai ales în cazul seturilor de date mici.

5.5.2 Importanța factorilor

Pentru toate modelele din iterații s-a calculat și afișat importanța factorilor în construirea modelului. Măsura care s-a folosit a fost Gini index. Este o măsură a gradului de impuritate a unui set de date în contextul algoritmilor de clasificare. Indexul Gini măsoară probabilitatea de clasificare greșită a unui factor ales în mod aleatoriu din setul de date. Cu cât valoarea indexului Gini este mai mică, cu atât split-ul este considerat mai bun.

$$Gini(p) = 1 - \sum_{i=1}^{J} p_i^2$$

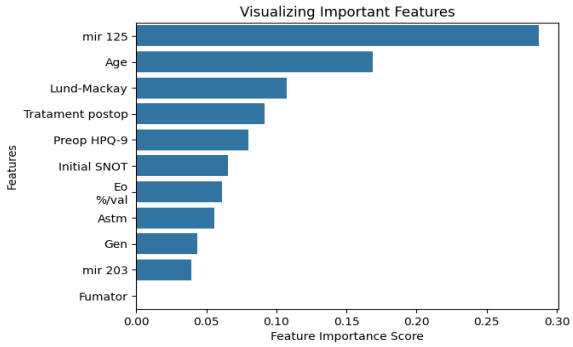
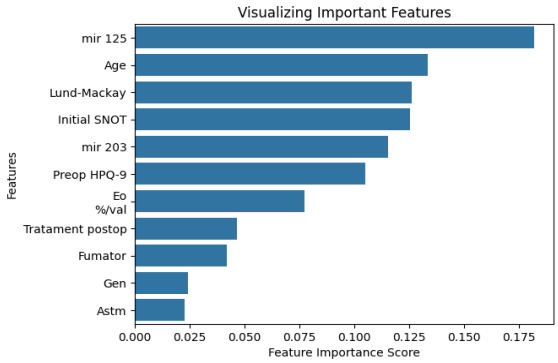
- **J** reprezintă numărul de clase
- $\boldsymbol{p_i}$ reprezintă probabilitatea unei instanțe de a fi clasificată în clasa i

Pentru fiecare nod din Decision Tree sau din Random Forest, câștigul Gini este calculat pentru fiecare factor atunci când se face împărțirea datelor în nodurile copil. Câștigul Gini este definit ca diferența dintre impuritatea Gini a nodului părinte și suma ponderată a impurităților Gini ale nodurilor copil.

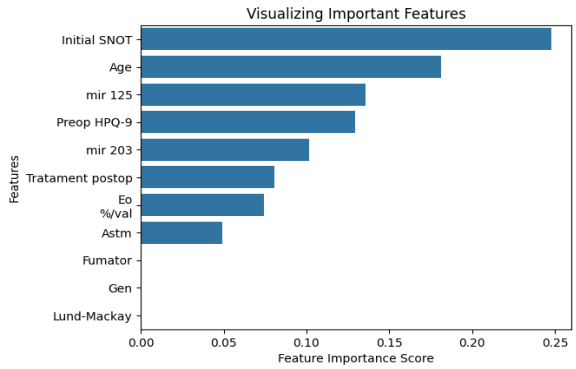
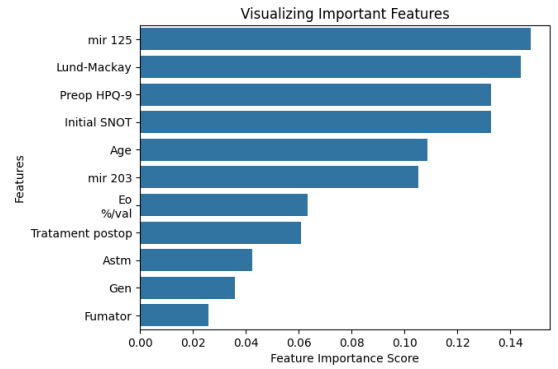
$$\boldsymbol{Gini\ Gain = Gini(Parinte) - \sum(\frac{Ni}{N} \times Gini(Fiul_i))}$$

- **Gini(Părinte)** reprezintă impuritatea Gini a nodului părinte
- **N** este numărul total de instanțe din nodul părinte
- **Ni** este numărul de instanțe din nodul copil i
- **Gini($\mathbf{Fiul_i}$)** reprezintă impuritatea Gini a nodului copil i

Iterația 1

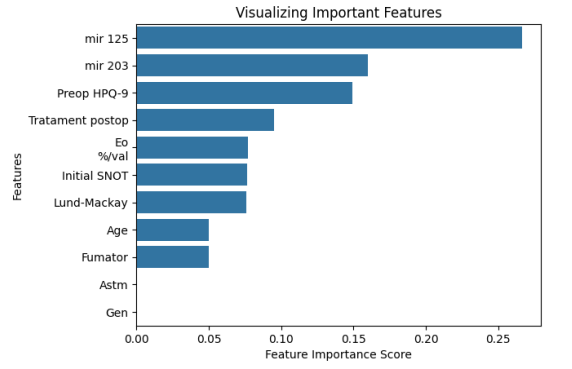
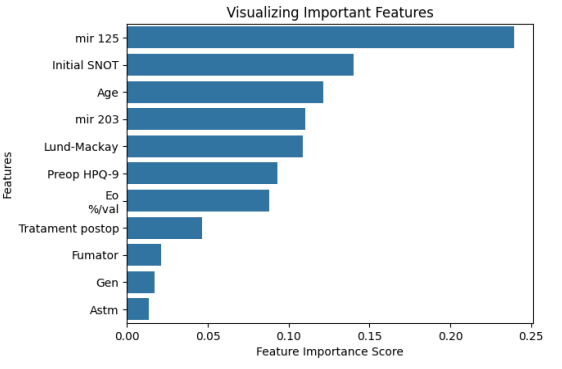
POSE 6 luni	
Decision Tree	**Random Forest**
	

Tabel 5.5.2.14: Scorul de importanță al factorilor iterația 1 POSE 6 luni

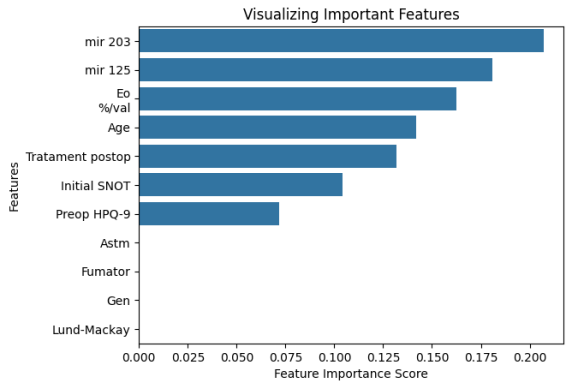
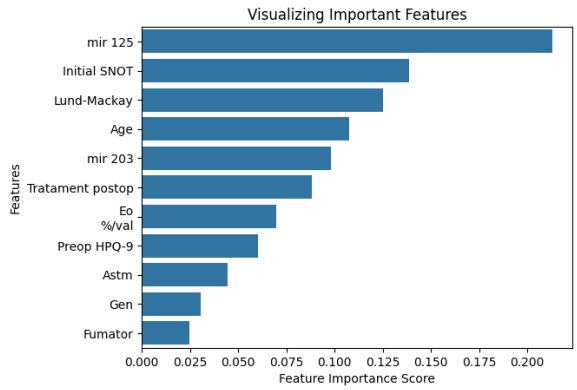
POSE 1 an	
Decision Tree	**Random Forest**
	

Tabel 5.5.2.15: Scorul de importanță al factorilor iterația 1 POSE 1 an

Iterația 2

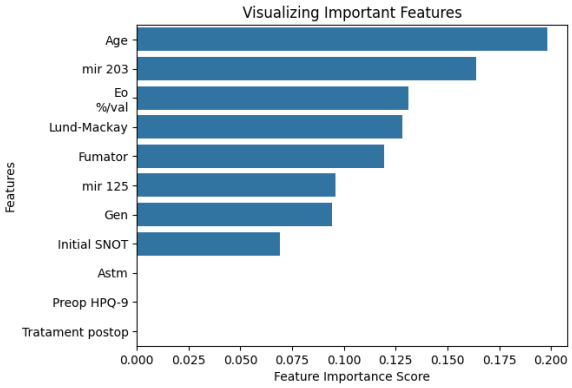
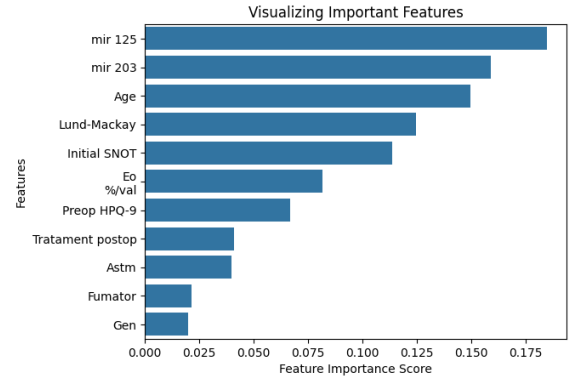
POSE 6 luni	
Decision Tree	**Random Forest**
	

Tabel 5.5.2.16: Scorul de importanță al factorilor iterația 2 POSE 6 luni

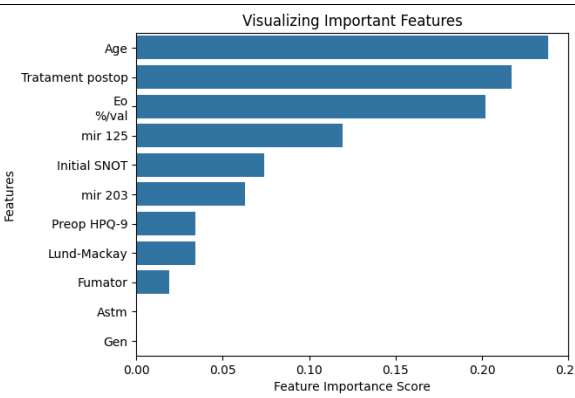
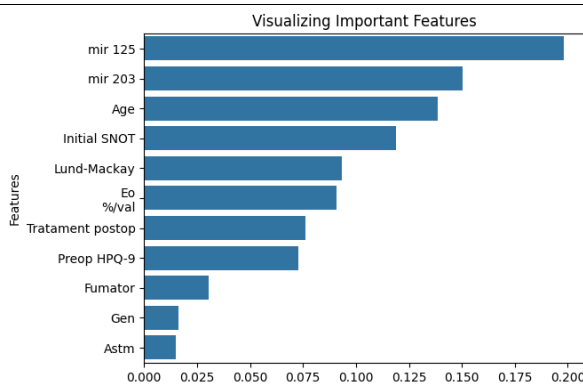
POSE 1 an	
Decision Tree	**Random Forest**
	

Tabel 5.5.2.17: Scorul de importanță al factorilor iterația 2 POSE 1 an

Iterația 3

POSE 6 luni	
Decision Tree	**Random Forest**
	

Tabel 5.5.2.18: Scorul de importanță al factorilor iterația 3 POSE 6 luni

POSE 1 an	
Decision Tree	**Random Forest**
	

Tabel 5.5.2.19: Scorul de importanță al factorilor iterația 3 POSE 1 an

Deși există o variație a importanței factorilor, se poate observa faptul că anumiți factori rămân constant pe primele locuri în ceea ce privește importanța. Mir 125 este unul din factorii care apare cel mai frecvent atât la POSE 6 luni cât și la POSE 1 an. Age este un factor care apare adesea în partea superioară a listei de importanță, indicând o corelație mare cu starea pacientului după operație. De asemenea, mir 203 și Initial SNOT apar destul de des printre primele valori. Putem concluziona faptul că ipoteza privind impactul anumitor factori asupra scorului POSE este susținută.

În ceea ce privește Tratamentul postoperator, putem observa faptul că acesta pare să influențeze mai mult rezultatele pentru POSE 1 an. Ceea ce poate să indice faptul că Tratamentul postoperator are rezultate după o perioadă mai lungă de timp. Factorii care par să fie cei mai puțin importanți sunt Fumător și Gen, aceștia apar adesea cu valori apropiate sau egale cu zero.

Capitolul 6

Concluzii și posibile îmbunătățiri

În urma celor prezentate pe parcursul acestei lucrări, putem concluziona faptul că s-a reușit obținerea unui model de învățare automată cu o performanță bună în determinarea scorului de Endoscopie Sinusală Perioperatorie. S-a putut observa o îmbunătățire a rezultatelor pornind de la iterația unu până la iterația trei, care s-a dovedit a avea cele mai bune valori. În ceea ce privește algoritmii folosiți, Random Forest s-a dovedit a avea rezultate mai bune, depășind constant Decision Tree. Prin rafinarea modelului iterativ și încorporarea feedback-ului din fiecare iterație, studiul a asigurat o îmbunătățire constantă a performanței predictive.

În ceea ce privește limitările acestei lucrări, putem spune că dimensiunea setului de date, de 79 de pacienți, poate să reprezinte un impediment în testarea și validarea datelor prezise. De asemenea, nu există o validare externă a modelului pe date independente de setul nostru.

Astfel, în viitor se dorește să se ia în considerare colectarea mai multor date pentru antrenarea și testarea modelului. Acest aspect se poate realiza prin construirea unei aplicații care să permită colectarea de date noi. De asemenea, în această aplicație se poate integra și modelul construit pentru a oferi cadrelor medicale posibilitatea de a utiliza predicțiile modelului. Acest lucru ar facilita luarea de decizii în ceea ce privește personalizarea tratamentul pentru pacienții cu rinosinuzită cronică cu polipi nazali.

Factori precum miR-125, miR-203, vârsta și Initial SNOT au fost identificați ca predictori importanti pentru scorul POSE, evidențiind relevanța clinică potențială a acestora. S-a confirmat faptul că miARN-ul joacă un rol important în starea pacientului cu rinosinuzită cronică cu polipi nazali, așa cum am văzut și în studiul „Research advances in roles of microRNAs in nasal polyp” [4].

Algoritmul Random Forest a demonstrat o performanță superioară în comparație cu Arborele de Decizie, sugerând potențialul său pentru prezicerea scorurilor POSE. Acest aspect este în concordanță cu constatările din studiul "Subepithelial neutrophil infiltration as a predictor of the surgical outcome of chronic rhinosinusitis with nasal polyps" [2] unde, algoritmul Random Forest a avut o performanță mai bună de 84,04% față de Decision Tree.

Astfel, putem spune că lucrarea de față oferă o bază bună de unde se poate continua studierea rinosinuzitei cronice cu polipi nazali, oferind informații utile atât în domeniul analizei predictive, cât și în cel al medicinei, în ceea ce privește scorul Endoscopie Sinusală Perioperatorie (POSE).

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 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9732428/

[5] Decision Tree https://www.geeksforgeeks.org/decision-tree/

[6] Random Forest https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/?ref=header_search

[7] Isolation Forest https://www.geeksforgeeks.org/what-is-isolation-forest/

[8] Machine Learning with Python https://www.geeksforgeeks.org/machine-learning-with-python/

[9] The Jupyter Notebook https://jupyter-notebook.readthedocs.io/en/stable/notebook.html

[10] Visual Studio Code https://code.visualstudio.com/docs

[11] Cross Validation in Machine Learning https://www.geeksforgeeks.org/cross-validation-machine-learning/

[12] Different Metrices in Machine Learning for Measuring performance of Classification Algorithms https://medium.com/@sachinsoni600517/different-metrices-in-machine-learning-for-measuring-performance-of-classification-algorithms-509e55c0a451

[13] What is Gini Impurity ? https://hidir-yesiltepe.medium.com/what-is-gini-impurity-b821dfb63b6e