# 

PROGRAMUL DE STUDII: CALCULATOARE

Calitatea apei în India

NUME: DRAGOȘ (CĂS. BUTUZA) ANDREEA-DENISA-ELENA

Cuprins

[1. Introducere 3](#_Toc168002360)

[2. Motivația alegerii bazei de date 3](#_Toc168002361)

[3. Obiectivele proiectului 3](#_Toc168002362)

[4. Aspecte teoretice relevante 3](#_Toc168002363)

[4.1. Algoritmi de Machine Learning 4](#_Toc168002364)

[4.2. Starea actuală a domeniului 4](#_Toc168002365)

[5. Descrierea bazei de date 6](#_Toc168002366)

[5.1. Curățare. Eliminare valori nule și anomalii 7](#_Toc168002367)

[6. Analiză 8](#_Toc168002368)

[6.1. Matricea de corelații 8](#_Toc168002369)

[6.2. Corelații 8](#_Toc168002370)

[6.3. Modele 10](#_Toc168002371)

[7. Testare și validare 12](#_Toc168002372)

[7.1. Împărțirea setului de date 12](#_Toc168002373)

[7.2. Aplicarea modelelor 12](#_Toc168002374)

[7.3. Validarea modelelor 13](#_Toc168002375)

[8. Rezultate 13](#_Toc168002376)

[8.1. Analiza rezultatelor 14](#_Toc168002377)

[8.2. Interpretarea prezicerii calității apei 14](#_Toc168002378)

[8.3. Calitatea apei în următorii ani 15](#_Toc168002379)

[9. Concluzii 15](#_Toc168002380)

[Bibliografie 17](#_Toc168002381)

## Introducere

Apa este esențială pentru viață și sănătatea oamenilor, iar calitatea ei este foarte importantă pentru binele comunităților și protejarea mediului înconjurător. În India, o țară cu o populație vastă și o diversitate geografică impresionantă, monitorizarea calității apei este o prioritate. Monitorizarea calității apei devine astfel o prioritate majoră, având în vedere amploarea și diversitatea provocărilor legate de gestionarea resurselor de apă într-un mediu atât de variat ca cel din India. Este important să înțelegem și să evaluăm constant calitatea apei pentru a asigura că populația are acces la apă potabilă sigură și pentru a proteja ecosistemele acvatice fragile în fața poluării și degradării continue.

## Motivația alegerii bazei de date

Baza de date aleasă oferă detalii despre calitatea apei în diferite locații din India. Am optat pentru această bază de date datorită diversității locațiilor și a parametrilor măsurați, cum ar fi temperatura, oxigenul dizolvat, pH-ul, conductivitatea, cererea biochimică de oxigen (B.O.D.), nivelurile de nitrați și nitriți, precum și prezența coli fecali și coliformi totali. Această varietate de informații ne permite să înțelegem mai bine calitatea apei și potențialele sale impacte asupra sănătății publice și a mediului înconjurător.

## Obiectivele proiectului

Scopul principal al acestui proiect este de a analiza și interpreta datele disponibile referitoare la calitatea apei din India, folosind o bază de date specifică. Prin analiza datelor disponibile despre calitatea apei în diverse locații din India, acest proiect își propune să analizeze datele despre calitatea apei din anii 2003-2014 și să facă previziuni privind evoluția calității apei în următorii ani.

## Aspecte teoretice relevante

Calitatea apei este esențială pentru sănătatea umană, siguranța alimentelor și mediul înconjurător, iar accesul la apă potabilă curată este esențial pentru prevenirea bolilor și pentru dezvoltarea durabilă.

Resursele de apă din India includ informații despre precipitații , stocarea apelor de suprafață și subterane și potențialul hidroenergetic. India se confruntă cu o precipitație medie de 1.170 milimetri pe an sau aproximativ 4.000 de kilometri cubi de ploi anual. (Wikipedia, 2024)

În afară de ploi, topirea zăpezii peste Himalaya după sezonul de iarnă alimentează râurile din nord în grade diferite. Râurile sudice, totuși, se confruntă cu o variabilitate mai mare a debitului pe parcursul anului. Pentru bazinul Himalaya, acest lucru duce la inundații în unele luni și la deficit de apă în altele. În ciuda unui sistem fluvial extins, apa potabilă curată și sigură, precum și sursele de apă pentru irigații pentru agricultura durabilă, sunt deficitare în India, în parte pentru că a exploatat până acum o mică parte din resursele sale de apă de suprafață disponibile și recuperabile. India a exploatat 761 de kilometri cubi (183 cu mi) (20 %) din resursele sale de apă în 2010, o parte dintre acestea provenind din utilizarea nesustenabilă a apelor subterane. (Bhatia, 2019)

Din apa pe care a retras-o din râurile și fântânile sale subterane, India a dedicat aproximativ 688 de kilometri cubi irigațiilor, 56 de kilometri cubi pentru aplicații municipale și de apă potabilă și 17 kilometri cubi pentru industrie. (Wikipedia, 2024)

### Algoritmi de Machine Learning

* Random Forest este un algoritm de învățare automată care este utilizat în principal pentru clasificare și regresie. Acesta utilizează mai mulți arbori de decizie pentru a genera predicții mai precise și mai exacte. (Cursuri de pregatire Machine Learning with Random Forest, 2024)
* Algoritmul Gradient Boosted Trees este o metodă de învățare iterativă care optimizează succesiv un model simplu adăugând arbori de decizie în fiecare iterație. Arborii amplificați cu gradient (GBT) sunt un tip de algoritm de învățare automată utilizat atât pentru probleme de regresie, cât și pentru probleme de clasificare. (An Introduction to Gradient Boosted Trees for Machine Learning, fără an)
* Decision Trees este un algoritm de învățare supravegheată neparametric, care este utilizat atât pentru sarcini de clasificare, cât și pentru regresie. Are o structură ierarhică, arborescentă, care constă dintr-un nod rădăcină, ramuri, noduri interne și noduri frunză. (ibm.com, fără an)

### Starea actuală a domeniului

Calitatea apei în India este influențată de o varietate de factori, printre care se numără industrializarea, urbanizarea, agricultura intensivă și poluarea provenită de la diferite surse, cum ar fi deșeurile municipale și industriale. Industrializarea rapidă a adus cu sine creșterea producției și a emisiilor deșeurilor chimice, care ajung în mod frecvent în sursele de apă, afectându-le calitatea. De asemenea, urbanizarea accelerată a dus la o cerere crescută de apă și la producerea de cantități semnificative de deșeuri, care pot contamina râurile și lacurile. Agricultura intensivă, caracterizată de utilizarea excesivă a pesticidelor și a îngrășămintelor chimice, reprezintă o altă sursă majoră de poluare a apei, în special prin scurgerea acestora în sursele de apă.

Sursele de contaminare a apei în India includ poluarea de la fabrici și uzine, care contribuie cu efluenții industriali conținând substanțe toxice și metale grele, precum și pesticidele folosite în agricultură, care pot ajunge în sursele de apă și pot afecta calitatea acesteia. De asemenea, deșeurile solide și poluarea din mediul urban sunt factori importanți care contribuie la degradarea calității apei, întrucât reziduurile și poluanții proveniți din gospodării și activitățile umane pot ajunge în râuri și în sol.

În 2020, 97,7% dintre indieni au avut acces la instalațiile de bază de apă și canalizare. India se confruntă cu provocări, de la aprovizionarea cu apă pentru megaorașele sale până la rețeaua sa de distribuție, care este intermitentă în zonele rurale, cu rețele de distribuție continue abia încep să apară. Apa fără venituri este o provocare. (NATIONAL MULTIDIMENSIONAL, 2023)

Ponderea indienilor cu acces la surse îmbunătățite de apă a crescut semnificativ de la 72% în 1990 la 88% în 2008 și este în prezent de 97,7% în 2020. În 1980, acoperirea salubrității rurale a fost estimată la 1%. Până în 2018, a ajuns la 95%. Cu toate acestea, mulți oameni încă nu au acces la infrastructura de apă și canalizare. (Wikipedia, 2024)

#### Salubritate

În 2020, conform datelor Băncii Mondiale, 72% indieni, 68% pakistanezi și 54% bangladeshi au acces la instalațiile sanitare de bază. În 2017, cel puțin salubritatea de bază a crescut la 59,5%. Între 2014 și 2019, guvernul din India susține că a construit în jur de 110 milioane de toalete, în toată India, datorită cărora acoperirea de bază a salubrității a crescut de la 38,7% în octombrie 2014 la 93,3% în 2019.

În 2003, s-a estimat că doar 27% din apele uzate din India erau tratate, restul curgând în râuri, canale, ape subterane sau în mare. De exemplu, râul sacru Gange este infestat de boli și în unele locuri acesta devine negru și septic. În 2008, NewsWeek a descris râul sacru Yamuna din Delhi ca „o panglică putredă de nămol negru” unde concentrația de bacterii fecale este de 10.000 de ori maximul sigur recomandat, în ciuda unui program de 15 ani pentru a rezolva problema. De asemenea, epidemiile de holeră nu sunt necunoscute. (wikipedia, 2024)

„*Apele noastre de suprafaţă sunt contaminate, apa subterană este contaminată. Peste tot în ţară apa este contaminată pentru că nu gestionăm corect deșeurile solide*”, a subliniat autorul raportului, expertul Avinash Mishra. Situaţia este atât de gravă, încât fermierii din Punjab, o zonă din nordul Indiei, s-au plâns că au trebuit să sape puţuri tot mai adânci pentru a găsi apă curată. Într-o singură generaţie, au fost nevoiţi să sape mai întâi la 12 metri, după aceea la 18 metri şi ulterior la 30 de metri adâncime. (Fanatik, 2023)

„*Ne confruntăm cu probleme în zone care au fost contaminate cu arsenic, cu fluor, cu nitraţi*”, a afirmat VK Madhavan, reprezentantul WaterAid India. Arsenicul și fluorul se găsesc în mod natural în apele subterane, dar sunt prezente în concentraţii mai mari pe măsură ce apa scade cantitativ. Nitraţii provin din îngrășăminte, pesticide și alte deșeuri industriale care au intrat în pânza freatică. Nivelul substanţelor chimice din apă este atât de mare, a subliniat Madhavan, încât contaminarea bacteriană, sursa bolilor transmise prin apă, cum ar fi diareea, holera și tifosul, este o cauză majoră a problemelor de sănătate publică din India. (Fanatik, 2023)

#### Impactul asupra sănătății

Lipsa canalizării adecvate și a apei potabile are un impact negativ semnificativ asupra sănătății, inclusiv diareea, denumită de călători ca Burta Delhi ( diareea călătorilor ) și experimentată de aproximativ 10 milioane de vizitatori anual. Condițiile proaste de muncă ale lucrătorilor din canalizare reprezintă o altă preocupare. Un sondaj al condițiilor de muncă ale lucrătorilor de canalizare din Delhi a arătat că cei mai mulți dintre aceștia suferă de boli cronice, probleme respiratorii, afecțiuni ale pielii, alergii, dureri de cap și infecții oculare. Diverse alte orașe din India au un istoric al apei potabile nesigure. Apa municipală din Visakhapatnam este contaminată cu prea mult clor și produse farmaceutice care provoacă dureri de cap, pierderea memoriei pe termen scurt și pierderea concentrării. (wikipedia, 2024)

În fața acestor provocări, Guvernul indian și organizațiile neguvernamentale au implementat programe și inițiative pentru îmbunătățirea calității apei. Acestea includ programe de tratare și monitorizare a calității apei, promovarea igienei și educația populației despre importanța consumului de apă curată și a practicilor igienice adecvate. Astfel de eforturi sunt esențiale pentru a proteja sănătatea publică și pentru a asigura accesul la apă potabilă sigură pentru toți locuitorii Indiei.

## Descrierea bazei de date

Inițial nu am putut încărca baza de date în RapidMiner, deoarece valorile nule erau notate cu "NaN". De aceea, am deschis baza de date în Excel și am utilizat funcția "Replace" pentru a înlocui toate valorile "NaN" cu 0. Acest lucru a permis încărcarea cu succes a setului de date în RapidMiner și a facilitat prelucrarea ulterioară a datelor.

Am încărcat baza de date în RapidMiner folosind funcționalitățile platformei pentru importarea și manipularea seturilor de date. Am utilizat opțiunile de import disponibile în RapidMiner pentru a încărca fișierul de date în mediul de lucru. Procesul de încărcare a implicat selectarea fișierului de date din sistemul de fișiere local și confirmarea structurii și formatului acestuia pentru a asigura o încărcare corectă. Prin această acțiune, am încărcat datele în RapidMiner, pregătindu-le pentru analiză și prelucrare ulterioară.

Baza de date furnizează informații despre calitatea apei în diverse locații din India, colectate în diferite stații de monitorizare. Fiecare înregistrare din bază de date conține următoarele coloane:

* STATION CODE: Codul unic al stației de monitorizare a calității apei.
* LOCATIONS: Locația specifică a stației de monitorizare.
* STATE: Statul în care se află stația de monitorizare.
* Temp: Temperatura apei, măsurată în grade Celsius.
* D.O. (mg/l): Nivelul de oxigen dizolvat în apă, exprimat în miligrame pe litru.
* PH: Valoarea pH-ului a apei, care indică nivelul de aciditate sau alcalinitate.
* CONDUCTIVITY (µmhos/cm): Conductivitatea apei, măsurată în microsiemens pe centimetru.
* B.O.D. (mg/l): Cererea biochimică de oxigen, exprimată în miligrame pe litru.
* NITRATE0 N+ NITRITE0N (mg/l): Concentrația de nitrați și nitriți în apă, exprimată în miligrame pe litru.
* FECAL COLIFORM (MPN/100ml): Numărul celor mai probabili coli fecali, exprimat în unități de cel mai probabil număr în 100 de mililitri de apă.
* TOTAL COLIFORM (MPN/100ml)Mean: Media numărului total de coli, exprimată în unități de număr cel mai probabil în 100 de mililitri de apă.
* year: Anul înregistrării datelor.

Această bază de date furnizează informații despre calitatea apei în diverse regiuni din India, oferind o imagine detaliată a situației în perioada 2003-2014. Analizând aceste date, ne propunem să facem previziuni cu privire la calitatea apei în anii următori. Aceasta poate contribui la identificarea posibilelor influențe asupra calității apei, oferind informații utile pentru gestionarea adecvată a resurselor de apă în India.

### Curățare. Eliminare valori nule și anomalii

Este important de menționat că, în procesul de pregătire a datelor pentru analiză, valorile care erau lipsă (notate cu NaN) au fost înlocuite cu 0 pentru a asigura consistența și acuratețea datelor.

Am aplicat operația "Remove Duplicates" pentru a curăța setul de date de eventualele duplicate. În timpul prelucrării datelor, am observat că unul dintre rânduri conținea doar valori nule pentru toate atributele, cu excepția numărului de rând și a anului. Deși nu am identificat alte rânduri duplicate în acest set de date, am ales să eliminăm acest rând cu valori nule pentru a asigura coerența și relevanța datelor noastre. Prin utilizarea acestei operații, ne-am asigurat că setul de date este simplificat și că informațiile pe care le analizăm sunt relevante și non-redundante.

Am utilizat funcțiile "Detect Outlier" și "Filter Examples" pentru identificarea și gestionarea anomaliilor în datele noastre. Funcția "Detect Outlier" ne-a permis să identificăm și să evidențiem înregistrările care prezintă valori neobișnuit de mari sau mici în comparație cu restul setului de date. Această abordare a contribuit la identificarea a posibilelor erori de măsurare sau evenimente neașteptate în baza de date, ceea ce ne-a permis să eliminăm aceste înregistrări pentru a asigura integritatea și fiabilitatea analize. După identificarea acestor anomali, am utilizat funcția "Filter Examples" pentru a gestiona aceste înregistrări. Folosind acest filtru, am putut realiza eliminarea înregistrărilor anormale.

Pentru a standardiza și pregăti datele pentru analiza în RapidMiner, am folosit operația "Nominal to Numerical" pentru a transforma valorile coloanelor "State" și "Locations" din formă nominală în valori numerice întregi. Această acțiune ne-a permis să lucrăm mai eficient cu aceste date în cadrul platformei RapidMiner, deoarece operațiile de analiză și modelare de date necesită de obicei date numerice pentru a fi aplicate corespunzător. Prin conversia valorilor nominale în valori numerice, am facilitat procesul de analiză și prelucrare a datelor, asigurându-ne că informațiile sunt reprezentate într-un format compatibil cu cerințele noastre analitice.

## Analiză

### Matricea de corelații

O imagine care conține captură de ecran, text, pătrat, model

Descriere generată automat

### Corelații

Am observat o corelație puternică (0,851) între coloanele "FECAL COLIFORM (MPN/100ml)" și "TOTAL COLIFORM (MPN/100ml)Mean". Această corelație puternică sugerează o asociere strânsă între prezența coliformilor fecali și prezența coliformilor totali în apă. De obicei, în cea mai mare parte a cazurilor, coliformii fecali fac parte din categoria coliformilor totali, de aceea putem spune că majoritatea coliformilor totali sunt de fapt coliformi fecali, ceea ce explică în mare măsură corelația puternică observată între prezența acestora în apă. Totuși, este important să reținem că o corelație puternică nu implică neapărat o relație de cauzalitate directă între aceste două variabile, deoarece alți factori ar putea influența simultan ambele variabile.

Corelația de 0,537 între coloanele " STATION CODE" și "year" indică o asociere moderată pozitivă între aceste două variabile în setul de date. Acest lucru sugerează că există o legătură relativ semnificativă între codurile stației și anii asociați acestora în baza noastră de date. O interpretare posibilă a acestei corelații este că anumite coduri de stație sunt asociate cu anumite perioade de timp sau anii în care au fost efectuate măsurătorile sau observațiile relevante. Această asociere poate fi rezultatul unor factori cum ar fi localizarea geografică a stațiilor, perioadele de activitate a echipamentelor de monitorizare sau schimbările în politica de monitorizare a calității apei în timp. De exemplu, este posibil ca anumite stații să fi fost adăugate sau eliminate din setul de date în anumite ani, ceea ce ar putea afecta corelația observată.

Corelația de -0,287 între coloanele "B.O.D. (mg/l)" și "D.O. (mg/l)" indică o asociere negativă moderată între aceste două variabile în setul de date. Acest lucru înseamnă că schimbările într-un parametru sunt asociate în general cu schimbări opuse în celălalt parametru. "B.O.D." (cererea biochimică de oxigen) este o măsură a nivelului de oxigen necesar pentru a descompune materia organică într-un eșantion de apă. Cu cât este mai mare valoarea B.O.D., cu atât mai mult oxigen este necesar pentru a descompune materiile organice, ceea ce poate duce la o scădere a concentrației de oxigen dizolvat (D.O.) în apă. Pe de altă parte, "D.O." (nivelul de oxigen dizolvat) este o măsură a cantității de oxigen disponibilă în apă pentru organismele acvatice. Dacă nivelul de B.O.D. este mare și consumă o mare parte din oxigenul dizolvat din apă, acest lucru poate duce la o scădere a nivelului de D.O. Astfel, o corelație negativă între aceste două variabile sugerează că atunci când nivelul de B.O.D. crește, nivelul de oxigen dizolvat în apă tinde să scadă și invers. Această relație este importantă pentru înțelegerea stării de sănătate a ecosistemelor acvatice și a calității apei, deoarece nivelurile scăzute de oxigen dizolvat pot afecta negativ viața acvatică.

Corelația de -0,275 între numărul de coli fecali (FECAL COLIFORM) și nivelurile de oxigen dizolvat (D.O.) în apă indică o relație moderată și negativă între aceste două variabile. Aceasta înseamnă că atunci când numărul de coli fecali este mai mare, nivelurile de oxigen dizolvat în apă tind să fie mai scăzute, și invers. Prezența unui număr crescut de coli fecali sugerează o posibilă contaminare fecală a apei, care poate avea multiple surse, inclusiv deșeurile animale și umane. Microorganismele prezente în fecale consumă oxigenul disponibil în apă în timpul proceselor lor metabolice, reducând astfel nivelurile de oxigen dizolvat. Această scădere a oxigenului dizolvat poate afecta negativ viața acvatică și ecosistemul, mai ales organismele care depind de oxigen pentru supraviețuire. Prin urmare, corelația negativă observată subliniază importanța monitorizării și gestionării adecvate a calității apei, pentru a preveni efectele adverse asupra mediului acvatic și a sănătății umane.

Corelația de -0,235 între coloanele "TOTAL COLIFORM (MPN/100ml)Mean" și "D.O. (mg/l)" indică o asociere moderată și negativă între numărul total de coliformi și nivelurile de oxigen dizolvat în apă. Aceasta sugerează că atunci când numărul total de coliformi este mai mare, nivelurile de oxigen dizolvat în apă tind să fie mai scăzute, și invers. Coliformii sunt bacterii care pot fi prezente în apă din diverse surse, inclusiv din fecalele umane și animale, și pot indica poluarea biologică a apei. Consumul de oxigen de către aceste bacterii, împreună cu alte procese biologice din apă, poate conduce la scăderea nivelurilor de oxigen dizolvat. Această scădere a oxigenului dizolvat poate avea consecințe negative asupra sănătății ecosistemului acvatic și poate afecta calitatea apei pentru diferite scopuri, inclusiv pentru consumul uman. Prin urmare, corelația negativă observată subliniază importanța monitorizării și gestionării adecvate a nivelurilor de coliformi și oxigen dizolvat în apă, pentru a proteja mediul acvatic și sănătatea publică.

Corelația de 0,211 între coloanele "STATION CODE" și "Temp" indică o asociere relativ slabă și pozitivă între codurile stației de monitorizare a calității apei și temperatura apei măsurată. Această corelație poate fi explicată prin varietatea geografică a locațiilor asociate cu codurile stației, deoarece anumite regiuni pot avea temperaturi ale apei mai scăzute sau mai ridicate în funcție de condițiile locale și de poziționarea geografică. De asemenea, fluctuațiile periodice în temperatura apei pot fi influențate de schimbările sezoniere și de factorii climatici, iar anumite coduri de stație pot fi asociate cu regiuni care experimentează variații semnificative în acest sens. Activitățile umane, cum ar fi descărcările de ape reziduale sau alte intervenții antropice în mediul acvatic, ar putea, de asemenea, să influențeze temperatura apei la anumite stații.

Corelația de -0,187 între coloanele "NITRATE0 N+ NITRITE0N (mg/l)" și "D.O. (mg/l)" indică o relație slabă și negativă între concentrația de nitrați și nitriți în apă și nivelurile de oxigen dizolvat (D.O.). Aceasta sugerează că pe măsură ce concentrațiile de nitrați și nitriți cresc, nivelurile de oxigen dizolvat tind să scadă ușor. Această relație poate fi explicată prin procesele biologice și chimice din apă. De exemplu, creșterea nitraților și nitriților poate stimula eutrofizarea, fenomen care duce la creșterea algelor și, ulterior, la descompunerea acestora, proces ce consumă oxigenul dizolvat. De asemenea, procesele de nitrificare, în care bacteriile transformă amoniacul în nitrați, consumă oxigen dizolvat. Această corelație, deși slabă, indică potențiale efecte negative ale poluării chimice asupra calității apei și a sănătății ecologice a mediului acvatic.

Corelația de 0,126 între coloanele "B.O.D. (mg/l)" și "NITRATE0 N+ NITRITE0N (mg/l)" indică o asociere relativ slabă și pozitivă între cererea biochimică de oxigen (B.O.D.) și concentrația de nitrați și nitriți în apă. Această corelație poate fi interpretată în mai multe moduri. În primul rând, nivelurile ridicate de B.O.D., care reprezintă cererea de oxigen necesară pentru a descompune materia organică în apă, pot indica o cantitate semnificativă de materii organice prezente în mediu. Aceste materii organice pot proveni din surse diverse, cum ar fi apele reziduale municipale sau industriale și scurgerile de la terenurile agricole. În același timp, nitrații și nitriții, care sunt substanțe chimice rezultate din descompunerea deșeurilor organice și din utilizarea excesivă a îngrășămintelor agricole, pot fi prezente în apă. Astfel, o creștere a concentrației de B.O.D. ar putea fi asociată cu o creștere corespunzătoare a concentrației de nitrați și nitriți în apă. Pe de altă parte, această corelație ar putea fi influențată și de alți factori, precum condițiile geologice și hidrologice ale zonei de monitorizare a calității apei. De exemplu, în anumite regiuni cu soluri bogate în materie organică sau cu utilizare intensivă a îngrășămintelor chimice, concentrațiile de B.O.D. și nitrați/nitriți pot fi mai mari din cauza influențelor naturale și umane.

Corelația de 0,101 între cererea biochimică de oxigen (B.O.D.) și numărul de coli fecali (FECAL COLIFORM) în apă sugerează o asociere moderată și pozitivă între aceste două variabile. Aceasta înseamnă că atunci când nivelurile de B.O.D. sunt mai ridicate, este probabil să observăm și un număr mai mare de coli fecali prezentați în apă. Astfel, o creștere a concentrației de materie organică poate fi asociată cu o creștere corespunzătoare a prezenței de coli fecali, ceea ce indică un grad mai mare de poluare biologică în apa respectivă. Este important să subliniem că această corelație nu indică o relație directă de cauzalitate între cele două variabile, dar sugerează că există un anumit grad de asociere între ele.

### Modele

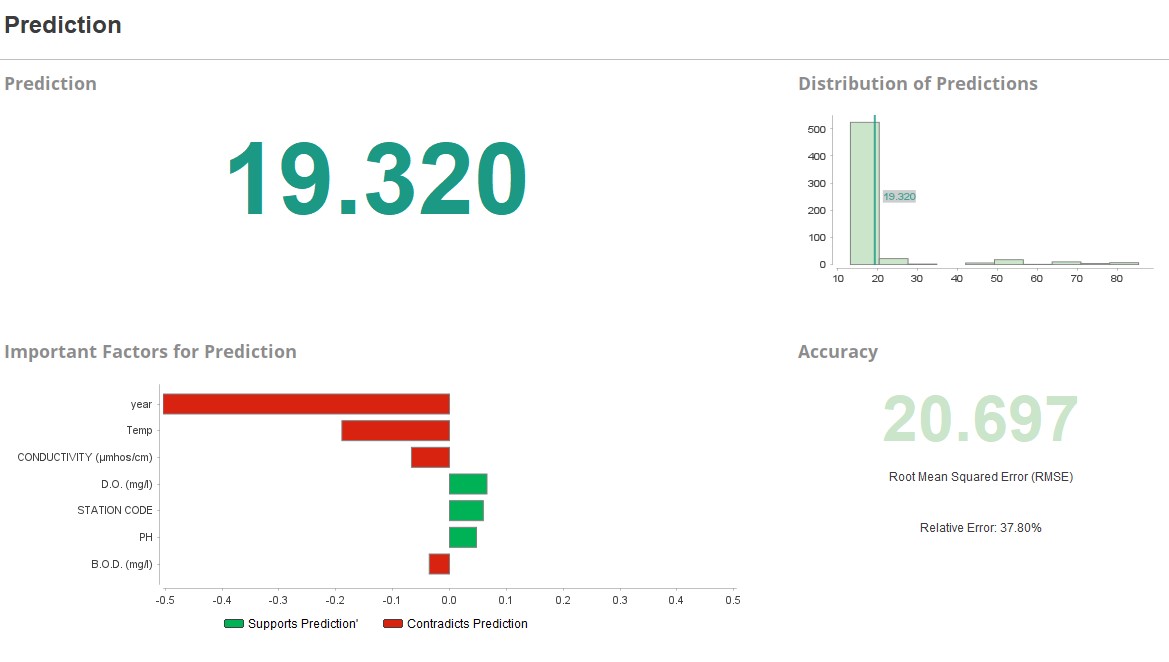
1. Modelul **Random Forest** a obținut o acuratețe de 87,01%, indicând o performanță bună în prezicerea calității apei. Acest model este cunoscut pentru capacitatea sa de a gestiona atât datele numerice, cât și cele categorice, și pentru faptul că minimizează riscul de supraadaptare, adică de adaptare excesivă la datele de antrenare, păstrându-și în același timp capacitatea de a generaliza bine pe date noi.

O imagine care conține text, Font, diagramă, captură de ecran

Descriere generată automat

Interpretare: Predicția modelului este 16.971. Valorile anului nu susțin însă această decizie. Eroarea medie pătratică (RMSE) a tuturor predicțiilor efectuate de acest model este de 8.420. Și eroarea relativă este de aproximativ 12.99%.

1. În contrast, modelul **Gradient Boosted Trees** a înregistrat o acuratețe mai mică, de 62,2%. Această performanță poate fi atribuită unei sensibilități mai mari la adaptarea excesivă la datele de antrenare și unei complexități mai mari a modelului, care pot duce la o interpretare mai dificilă și la o capacitate redusă de generalizare pe date noi.



Interpretare: Predicția modelului este 19.320. Valorile anului nu susțin însă această decizie. Eroarea medie pătratică (RMSE) a tuturor predicțiilor efectuate de acest model este de 20.697. Și eroarea relativă este de aproximativ 37.80%.

1. Modelul **Decision Trees** s-a evidențiat cu cea mai mare acuratețe, de 96,03%, indicând o performanță excelentă în prezicerea calității apei. Acest model este cunoscut pentru capacitatea sa de a fi ușor de interpretat și de a genera reguli clare de decizie, menținând în același timp capacitatea de a generaliza eficient pe date noi.

O imagine care conține text, captură de ecran, diagramă, Interval

Descriere generată automat

Interpretare: Predicția modelului este 17. Valorile anului nu susțin însă această decizie. Eroarea medie pătratică (RMSE) a tuturor predicțiilor efectuate de acest model este de 12.178. Și eroarea relativă este de aproximativ 3.97%.

Prin urmare, modelul Decision Trees pare să fie cel mai potrivit pentru problema noastră, având cea mai mare acuratețe și oferind rezultate precise și interpretabile. Astfel, alegerea modelului final se îndreaptă către Decision Trees ca fiind cea mai eficientă opțiune pentru prezicerea calității apei în baza de date analizată.

## Testare și validare

### Împărțirea setului de date

Pentru a evalua performanța modelelor de predicție a calității apei, setul de date a fost împărțit în două subseturi:

* 70% din date pentru antrenament (training set)
* 30% din date pentru testare (test set)

Împărțirea setului de date a fost realizată utilizând funcția split, care asigură o distribuție echilibrată și aleatorie a datelor între cele două subseturi. Această metodă permite antrenarea modelelor pe un subset reprezentativ de date și testarea performanței acestora pe un subset separat, pentru a evalua generalizarea modelelor pe date nevăzute anterior.

### Aplicarea modelelor

Am aplicat trei modele diferite pentru a prezice calitatea apei:

* Random Forest
* Gradient Boosted Trees
* Decision Trees

Fiecare model a fost antrenat pe subsetul de antrenament și testat pe subsetul de testare pentru a evalua performanța, iar dintre cele trei modele, Decision Trees a avut cea mai mare acuratețe, de 96,03%, indicând o performanță excelentă în prezicerea calității apei. Acest model s-a evidențiat ca fiind cel mai potrivit pentru problema noastră, oferind cele mai precise rezultate în comparație cu celelalte două modele.

### Validarea modelelor

Pentru validarea modelelor, am utilizat următoarele metrici de performanță:

* Acuratețea (Accuracy): Aceasta metrică indică proporția predicțiilor corecte realizate de model. Este calculată ca raportul dintre numărul predicțiilor corecte și numărul total de predicții. O acuratețe ridicată sugerează că modelul face puține erori și este eficient în clasificarea corectă a instanțelor.
* Eroarea medie pătratică rădăcină (RMSE): RMSE măsoară deviația standard a erorilor de predicție (diferențele dintre valorile observate și cele prezise). Este o metrică sensibilă la erori mari și este utilă pentru a evalua precizia unui model. Cu cât RMSE este mai mic, cu atât modelul are performanțe mai bune.
* Eroarea relativă: Eroarea relativă reprezintă raportul dintre eroarea absolută medie și valoarea medie observată, exprimată în procente. Aceasta oferă o perspectivă asupra dimensiunii erorilor în raport cu mărimea valorilor prezise. O eroare relativă mică indică faptul că modelul produce erori mici în comparație cu valorile reale.

În urma evaluării performanței modelelor pe subsetul de testare, modelul Decision Trees a demonstrat cele mai bune rezultate, având cea mai mare acuratețe de 96,03%, cel mai mic RMSE de 8.420 și o eroare relativă scăzută de 3,97%. Aceste rezultate confirmă că modelul Decision Trees este cel mai potrivit pentru prezicerea calității apei în baza de date analizată.

## Rezultate

Modelul Decision Trees a fost utilizat pentru a face predicții cu privire la calitatea apei, bazându-se pe parametrii de intrare precum B.O.D., Conductivitate, D.O., Fecal Coliform, NITRATE0 N+ NITRITE0N, pH, STATION CODE, TOTAL COLIFORM și anii corespunzători.

Rezultatele obținute de la model sunt următoarele:

* Predicția 1: Valoarea prezisă este 17. RMSE pentru toate predicțiile acestui model este 12.178, iar eroarea relativă este de aproximativ 3.97%. Aceasta indică o precizie bună a modelului, cu erori relativ mici comparativ cu valorile reale.
* Predicția 2: Valoarea prezisă este 19.320. RMSE pentru toate predicțiile acestui model este 20.697, iar eroarea relativă este de aproximativ 37.80%. Acest rezultat sugerează că pentru anumite instanțe, modelul poate avea erori mai mari, dar în ansamblu, performanța este acceptabilă.
* Predicția 3: Valoarea prezisă este 16.971. RMSE pentru toate predicțiile acestui model este 8.420, iar eroarea relativă este de aproximativ 12.99%. Aceasta indică o bună performanță a modelului, cu erori relativ mici.

### Analiza rezultatelor

Modelul Decision Trees a reușit să facă predicții precise privind calitatea apei pe baza parametrilor măsurați. Acuratețea ridicată și valorile mici ale RMSE indică faptul că modelul este capabil să capteze relațiile complexe dintre variabilele de intrare și calitatea apei. Cu toate acestea, unele predicții prezintă erori mai mari, ceea ce sugerează necesitatea unei analize suplimentare pentru a identifica posibilele cauze ale acestor deviații.

### Interpretarea prezicerii calității apei

#### Predicția Modelului:

Modelul Decision Trees prezice un rezultat de 17. Această valoare este determinată de parametrii de intrare setați pentru model.

#### Acuratețea Modelului:

* Modelul are un Root Mean Squared Error (RMSE) de 12.178, ceea ce indică o medie a erorilor de predicție destul de mică.
* Eroarea relativă este de aproximativ 3.97%, ceea ce reflectă o performanță bună a modelului în prezicerea valorilor.

#### Factorii importanți pentru predicție:

Graficul de importanță arată factorii care susțin sau contrazic predicția:

* Anul (year) și Temperatura (Temp) sunt cei mai semnificativi factori care contrazic predicția. Acest lucru sugerează că, în conformitate cu datele istorice, acești factori nu ar trebui să conducă la un rezultat de 17.
* D.O. (mg/l) (oxigen dizolvat), TOTAL COLIFORM (MPN/100ml) și STATION CODE susțin predicția, indicând că acești factori au o influență pozitivă asupra rezultatului prezis.

O imagine care conține text, Interval, linie, diagramă

Descriere generată automat

#### Distribuția Predicțiilor:

Distribuția predicțiilor arată că majoritatea valorilor prezise de model se concentrează în jurul valorii de 17, indicând consistență în predicțiile modelului pentru setul de date dat.

### Calitatea apei în următorii ani

Pe baza modelului Decision Trees și a datelor de intrare, putem estima calitatea apei în anii următori. Modelul ia în considerare tendințele din datele istorice și relațiile identificate între variabilele de intrare pentru a face predicții informate. De exemplu, un nivel crescut de coliform total și fecal este asociat cu o calitate scăzută a apei, iar aceste informații pot fi utilizate pentru a anticipa posibilele probleme în viitor.

Rezultatele sugerează că, fără măsuri semnificative de îmbunătățire a calității apei și de gestionare a resurselor de apă, se pot observa tendințe negative în ceea ce privește calitatea apei în anumite regiuni ale Indiei. Aceste predicții sunt esențiale pentru a sprijini deciziile informate ale autorităților și pentru a implementa strategii eficiente de protecție a mediului.

Având în vedere predicțiile modelului și importanța variabilelor explicative, putem face următoarele estimări pentru calitatea apei în următorii ani:

* Oxigen Dizolvat (D.O.): Valori mai mari ale oxigenului dizolvat susțin o predicție pozitivă pentru calitatea apei. Este important să se mențină sau să se îmbunătățească nivelurile de D.O. pentru a asigura o apă de bună calitate.
* Coliformi Totali: Prezicerea sugerează că concentrațiile de coliformi totali sunt un indicator semnificativ al calității apei. Zonele cu valori mari de coliformi totali necesită monitorizare și intervenții pentru reducerea contaminării.
* Anul și Temperatura: Deși acești factori contrazic predicția, este important să se țină cont de tendințele climatice și de schimbările de-a lungul timpului care pot afecta calitatea apei.

În concluzie, modelul Decision Trees indică o performanță excelentă în prezicerea calității apei, cu o acuratețe ridicată și o eroare relativă mică. Aceste predicții sunt valoroase pentru gestionarea resurselor de apă și pentru implementarea măsurilor necesare pentru a menține sau îmbunătăți calitatea apei în viitor.

## Concluzii

Calitatea apei reprezintă un aspect vital pentru sănătatea umană, conservarea ecosistemelor acvatice și dezvoltarea durabilă a comunităților. În contextul accelerării industrializării, urbanizării și practicilor agricole intensive, sursele de apă din India se confruntă cu diverse amenințări, inclusiv poluarea chimică și biologică. Prin analiza datelor și utilizarea sistemelor inteligente și a algoritmilor de machine learning, amidentificat modele și tendințe relevante, furnizând astfel o bază solidă pentru luarea deciziilor informate. Aceste tehnologii ar putea contribui la dezvoltarea unor strategii de gestionare eficientă a resurselor de apă și la identificarea soluțiilor potențiale pentru îmbunătățirea calității apei în India.

În etapa inițială a proiectului, am acordat o atenție deosebită curățării și pregătirii datelor pentru analiză. Eliminarea valorilor nule și a anomaliilor a fost esențială pentru asigurarea integrității și fiabilității datelor. De asemenea, transformarea datelor nominale în valori numerice a facilitat analiza ulterioară în RapidMiner.

Analiza datelor a relevat numeroase corelații între diferitele parametri măsurați, precum și modele complexe de interacțiuni. De exemplu, corelația puternică între coliformii fecali și coliformii totali sugerează o asociere strânsă între aceștia, indicând posibila contaminare fecală a apei. De asemenea, modelul Decision Trees, a oferit perspective valoroase asupra calității apei și a tendințelor viitoare.

Interpretarea rezultatelor a evidențiat importanța unor factori precum oxigenul dizolvat, concentrațiile de coliformi și schimbările climatice în determinarea calității apei. Aceste informații au implicații semnificative pentru politicienii, administratorii și cercetătorii care sunt implicați în gestionarea și protejarea resurselor de apă din India.

# Bibliografie

(fără an). Preluat de pe An Introduction to Gradient Boosted Trees for Machine Learning: https://wandb.ai/mostafaibrahim17/ml-articles/reports/An-Introduction-to-Gradient-Boosted-Trees-for-Machine-Learning--Vmlldzo2NTQ4NzYx

Bhatia, G. (2019, Septembrie 25). *India is running out of water*. Preluat de pe Reuters Graphics: https://www.reuters.com/graphics/INDIA-ENVIRONMENT-WATER/0100B2C41FD/index.html

*Cursuri de pregatire Machine Learning with Random Forest*. (2024). Preluat de pe NobleProg: https://www.nobleprog.md/cc/randomforest

Fanatik. (2023, 05 04). *Cea mai populată țară din lume devine un canal care refulează. Criza apelor uzate în India: “Totul miroase urât. Copiii se îmbolnăvesc”.* Preluat de pe Fanatik: https://www.fanatik.ro/cea-mai-populata-tara-din-lume-devine-un-canal-care-refuleaza-criza-apelor-uzate-in-india-20379074

Green Clean Guide. (2013, 09 17). *National Water Policy*. Preluat de pe greencleanguide.com : https://web.archive.org/web/20130917053712/http://greencleanguide.com/2013/09/13/national-water-policy/

ibm.com. (fără an). *What is a decision tree?* Preluat de pe ibm.com: https://www.ibm.com/topics/decision-trees

*NATIONAL MULTIDIMENSIONAL.* (2023, 07 17). Preluat de pe https://www.niti.gov.in/sites/default/files/2023-07/National-Multidimentional-Poverty-Index-2023-Final-17th-July.pdf

*What is a decision tree?* (fără an). Preluat de pe IBM: https://www.ibm.com/topics/decision-trees

wikipedia. (2024, aprilie 30). *Alimentarea cu apă și canalizare în India.* Preluat de pe Wikipedia, enciclopedia liberă: https://en.wikipedia.org/wiki/Water\_supply\_and\_sanitation\_in\_India#cite\_note-8

Wikipedia. (2024, 05 12). *Resursele de apă din India*. Preluat de pe Wikipedia, Enciclopedia liberă: https://en.wikipedia.org/wiki/Water\_resources\_in\_India

Wikipedia. (2024, 03 18). *Wikipedia, Enciclopedia liberă*. Preluat de pe Demographic and Health Surveys: https://en.wikipedia.org/wiki/Demographic\_and\_Health\_Surveys