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

Descriere generată automat

**FACULTATEA DE AUTOMATICĂ ŞI CALCULATOARE DEPARTAMENTUL CALCULATOARE**

**Proiect Sisteme Inteligente**

|  |  |
| --- | --- |
| Student: | **Bota Emanuel-Iosif** |

**1.Introducere**

**1.1 Introducere:**

În era noastra, volumul și complexitatea datelor medicale cresc exponențial iar impreuna cu ritmul alert in care lumea din jurul nostru de misca, ne ofera nenumarate oportunitati de a utiliza sistemele inteligente (machine learning) pentru a îmbunătăți predicțiile și deciziile în domeniul sănătății.

Predicția costului pentru asigurarilor de sanatate este un subiect de interes major în domeniul sănătății, având impact direct asupra economiei și accesului la servicii medicale. Cu ajutorul sistemelor inteligente, se pot construi modele predictive care să estimeze aceste costuri, permițând instituțiilor medicale și companiilor de asigurări să ia decizii informate.

Totodata predictia costurilor pentru asigurarile de sanatate poate veni si in ajutorul oamenilor care ar putea sa isi faca o planificare a veniturilor si a cheltuielilor mult mai exacta si mai usoara prin prisma faptului ca acestia ar putea avea o predictie, destul de apropiata sau chiar identica cu valoarea reala a asigurarii, in functie de datele fiecarei persoane.

**1.2 Motivația alegerii bazei de date:**

Baza de date utilizată în acest proiect, provenind dintr-un set de date public, reprezintă o sursă bogată de informații esențiale despre pacienți, cuprinzând diverse aspecte ale vieții lor și stării de sănătate. Printre aceste informații se numără vârsta, sexul, indicele de masă corporală (BMI), numărul de copii, statutul de fumător și regiunea geografică în care locuiesc.

Alegerea acestei baze de date a fost motivată de diversitatea și relevanța variabilelor incluse, care sunt recunoscute ca fiind factori semnificativi în determinarea costurilor medicale. De exemplu, vârsta poate influența predispoziția către anumite afecțiuni medicale și necesitatea de îngrijiri medicale suplimentare, în timp ce sexul poate avea un impact asupra riscului de anumite boli sau afecțiuni. Indicele de masă corporală (BMI) este un indicator important al stării de sănătate și poate fi asociat cu riscul de obezitate și alte afecțiuni metabolice. Numărul de copii poate influența cheltuielile medicale ale unei familii, iar statutul de fumător poate fi asociat cu riscul crescut de afecțiuni cardiovasculare și respiratorii. În plus, regiunea geografică poate influența accesul la servicii medicale și costurile acestora, datorită diferențelor în infrastructură, resurse și politici de sănătate între diferite regiuni.

Explorarea și înțelegerea mai profundă a relațiilor dintre aceste variabile și costurile medicale pot oferi perspective valoroase asupra modului în care sistemul de sănătate poate fi îmbunătățit și optimizat pentru a furniza servicii mai eficiente și accesibile. Prin utilizarea tehnicilor de analiză și modelare a datelor, putem identifica modele și tendințe care să ne ajute să anticipăm și să gestionăm mai bine costurile medicale, contribuind astfel la îmbunătățirea calității vieții și a sănătății populației

### **2. Contextul bazei de date și al proiectului**

**Contextul bazei de date:**

Baza de date utilizată în acest proiect reprezintă o colecție de informații esențiale despre pacienți, furnizând date relevante pentru analiza și predicția costurilor medicale. Aceasta este structurată în următoarele coloane:

* **age (vârsta):** Această variabilă indică vârsta pacientului, exprimată în ani.
* **sex (sexul):** Variabila indică sexul pacientului și poate avea două valori posibile: masculin sau feminin.
* **bmi (indicele de masă corporală):** Acest indicator reflectă relația dintre greutatea și înălțimea pacientului și este utilizat pentru evaluarea stării de sănătate legate de greutate.
* **children (numărul de copii):** Variabila indică numărul de copii pe care îi are pacientul.
* **smoker (statutul de fumător):** Această variabilă indică dacă pacientul este fumător sau nu, având valorile "da" sau "nu".
* **region (regiunea):** Indică regiunea geografică în care locuiește pacientul.
* **charges (costurile medicale):** Această variabilă reprezintă costurile medicale asociate cu tratamentele și serviciile medicale furnizate pacientului.

Aceste date sunt esențiale pentru analiza costurilor medicale și pot oferi informatii valoroase cu privire la factorii care influențează aceste costuri.

**Cerințe:**

Proiectul are ca scop principal dezvoltarea unui model predictiv precis pentru costurile medicale, folosind datele disponibile în baza de date. Cerințele specifice ale proiectului includ:

1. **Încărcarea și curățarea datelor:** Procesul de pregătire a datelor implică încărcarea datelor din sursa lor și efectuarea operațiilor de curățare pentru a elimina datele lipsă, a corecta erorile și a standardiza formatele.
2. **Explorarea și analiza datelor:** Această etapă implică explorarea datelor pentru a înțelege distribuția, corelațiile și tendințele din setul de date. Vizualizările grafice, analiza statistică și identificarea outlier-ilor sunt aspecte cheie în această etapă.
3. **Construirea și antrenarea mai multor modele de regresie:** Se vor construi și antrena mai multe modele de regresie folosind diferite algoritmi și tehnici pentru a prezice costurile medicale.
4. **Optimizarea hiperparametrilor pentru fiecare model:** Pentru fiecare model, se vor ajusta hiperparametrii pentru a maximiza performanța și generalizarea.
5. **Compararea performanțelor modelelor:** Performanța modelelor va fi evaluată folosind metodele relevante, cum ar fi root mean squared error (RMSE), mean absolute error (MAE) și coeficientul determinării (R-squared).
6. **Selectarea și interpretarea modelului final:** Bazându-se pe performanța și interpretabilitatea modelului, se va selecta un model final care să fie utilizat pentru predicții și interpretări.

**Obiectiv:**

Scopul final al proiectului este de a dezvolta un model de regresie robust și precis care să prezică costurile medicale în funcție de variabilele demografice și legate de stilul de viață al pacienților. Prin înțelegerea factorilor cheie care influențează costurile, se poate oferi recomandări pentru optimizarea cheltuielilor și îmbunătățirea politicilor de asigurare, contribuind astfel la îmbunătățirea accesului și eficienței serviciilor medicale.

**3. Aspecte teoretice relevante:**

**Regresia liniară:** Regresia liniară este una dintre cele mai fundamentale și larg utilizate metode de modelare în analiza regresiei. În esență, regresia liniară încearcă să găsească o relație liniară între variabila dependentă (în acest caz, costurile medicale) și una sau mai multe variabile independente (age, sex, bmi, children, smoker, region). Această relație este exprimată printr-o ecuație liniară, unde coeficienții corespunzători variabilelor independente sunt estimați pentru a minimiza eroarea dintre valorile prezise și cele observate.

**Regresia polinomială:** Regresia polinomială reprezintă o extensie a regresiei liniare, care permite modelarea relațiilor non-liniare dintre variabile. În loc să presupună o relație liniară, acest model poate ajusta o curbă polinomială la datele de intrare. Astfel, este capabil să capteze nuanțele și complexitățile din date care nu ar putea fi reprezentate de o relație liniară simplă.

**Support Vector Regression (SVR):** Support Vector Regression (SVR) este o tehnică de regresie care se bazează pe principiile mașinilor de vector suport (SVM). Scopul SVR este de a găsi o funcție care să se încadreze într-un set de toleranță predefinit în jurul datelor de antrenare, în timp ce minimizează eroarea de predicție. SVR este eficient în tratarea datelor non-liniare și în gestionarea problemelor de regresie cu un număr mare de dimensiuni sau cu seturi de date mici.

**Gradient Boosted Trees (GBT):** Gradient Boosted Trees (GBT) este un model de ansamblu care combină mai mulți arbori de decizie slabi pentru a forma un model puternic. În fiecare etapă, GBT adaugă un nou arbore pentru a corecta erorile modelului existent. Această abordare iterativă duce la o îmbunătățire continuă a performanței modelului. GBT este eficient în tratarea datelor non-liniare și în capturarea relațiilor complexe dintre variabile.

**Referințe științifice (minim 10):**

1. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232.
2. Vapnik, V. N. (1995). The Nature of Statistical Learning Theory. Springer-Verlag.
3. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.
4. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297.
5. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.
6. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.
7. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301-320.
8. Schölkopf, B., & Smola, A. J. (2002). Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press.
9. Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81-106.
10. Witten, I. H., Frank, E., & Hall, M. A. (2011). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.

**4. State-of-the-Art:**

**Evoluția tehnologică continuă:**

În lumea analizei datelor și a machine learning-ului, inovația este la ordinea zilei. De la începuturile sale, domeniul a cunoscut o creștere exponentială, alimentată de avansurile tehnologice și de disponibilitatea datelor. Noile descoperiri și tehnici apar constant, oferind oportunități nelimitate pentru cercetare și aplicații practice.

**Rețelele neurale profunde:**

Printre cele mai de seamă realizări recente se numără progresele în domeniul rețelelor neurale profunde. Aceste arhitecturi complexe, inspirate de modul în care funcționează creierul uman, sunt capabile să înțeleagă și să reprezinte datele în moduri fascinante și, uneori, surprinzătoare. Utilizate într-o gamă largă de aplicații, de la recunoașterea facială până la analiza textului, rețelele neurale profunde au redefinit limitele cunoștințelor noastre despre învățarea automată.

**Abordări inovatoare în învățarea automată:**

Pe lângă rețelele neurale profunde, există numeroase alte tehnici inovatoare care își găsesc locul în arsenalul practicienilor de machine learning. De exemplu, învățarea cu transfer a câștigat popularitate datorită capacității sale de a utiliza cunoștințele dobândite într-un domeniu pentru a rezolva probleme în alt domeniu. Învățarea semi-supervizată, în care modelele sunt antrenate folosind atât date etichetate, cât și neetichetate, reprezintă o altă zonă de interes, având în vedere dificultatea obținerii unor seturi mari de date etichetate în multe domenii.

**Importanța rămânerii la curent cu noile tehnologii:**

Într-o lume în continuă schimbare, adaptarea la noile tehnologii este esențială pentru succes. Specialiștii în domeniul analizei datelor și al machine learning-ului trebuie să fie la curent cu cele mai recente avansuri pentru a putea profita de oportunitățile care se ivește și pentru a obține cele mai bune rezultate în proiectele lor. Rămânerea la curent cu "state of the art"-ul nu este doar un obiectiv, ci o necesitate pentru cei care doresc să rămână relevanți și eficienți într-un domeniu în continuă schimbare.

**5.Aspecte teoretice implementate:**

**Încărcarea și preprocesarea datelor:**

În procesul de încărcare și preprocesare a datelor, am aplicat mai multe operații pentru a asigura că datele noastre sunt pregătite corespunzător pentru analiză:

* **Încărcarea datelor:** Am folosit operatorul Retrive din RapidMiner pentru a încărca setul nostru de date dintr-un fișier CSV importat in RapidMiner. Acesta ne-a permis să citim datele din fisierul CSV, fisier pe care il avem sub forma unei baze de date.
* **Gestionarea valorilor lipsă:** După încărcarea datelor, am identificat și gestionat valorile lipsă folosind operația Replace Missing Values. Pentru a asigura coerența și integritatea datelor noastre, am optat pentru strategii precum înlocuirea valorilor lipsă cu mediana sau media caracteristicilor corespunzătoare.
* **Curățarea datelor și eliminarea anomaliilor:** Am aplicat operația Detect Outliers pentru a elimina potențialele anomalii și erori din setul nostru de date. Acest lucru ne-a permis să ne asigurăm că datele folosite în analiză sunt valide și relevante pentru obiectivele noastre.

**Construirea și antrenarea modelelor:**

Am implementat și antrenat mai multe modele de regresie, fiecare cu caracteristicile și particularitățile sale specifice:

* **Random Forest și Decision Tree:** Am folosit operatorii corespunzători din RapidMiner pentru a construi și antrena aceste modele. Am definit parametrii relevanți, cum ar fi numărul de arbori sau adâncimea maximă a acestora, pentru a obține modelele optime.
* **Gradient Boosted Trees (GBT):** Pentru GBT, am utilizat un operator dedicat care ne-a permis să configurăm parametrii algoritmului, cum ar fi rata de învățare și numărul de arbori, pentru a obține un model bine ajustat și performant.

**Optimizarea hiperparametrilor:**

Pentru a găsi cei mai buni hiperparametri pentru fiecare model, am folosit operatorul Optimize Parameters (Grid) din RapidMiner:

* Am definit un set de valori posibile pentru fiecare hiperparametru, acoperind un spectru larg de valori relevante.
* Am utilizat o metodă de căutare în grilă pentru a testa toate combinațiile posibile de hiperparametri și pentru a identifica acele combinații care maximizează performanța modelului pe datele noastre de antrenare.

**Evaluarea performanței modelelor:**

Am evaluat performanța fiecărui model utilizând metrici specifice de regresie, cum ar fi RMSE, MAE și R-squared:

* Am utilizat operatorii Performance (Regression) pentru a calcula aceste metrici pentru fiecare model antrenat pe datele noastre de testare.
* Am comparat rezultatele obținute pentru fiecare metrică pentru a evalua și a alege modelul final, care prezintă cea mai bună performanță în predicția costurilor medicale pe baza datelor noastre disponibile.

**6.Testare si Validare:**

**Împărțirea setului de date:**

* În etapa inițială a procesului, setul de date este împărțit în două subseturi distincte: setul de date pentru antrenament și setul de date pentru testare. Această diviziune este crucială pentru a evalua performanța modelului pe date "nevăzute" anterior, simulant condițiile reale de utilizare a modelului.
* Alegerea unei proporții adecvate pentru împărțirea setului de date este importantă pentru a evita supraînvățarea (overfitting) sau subînvățarea (underfitting) modelului. În acest caz, am alocat 70% din datele pentru antrenament, iar restul de 30% pentru testare.

**Testarea modelelor:**

* După ce setul de date a fost împărțit, fiecare model a fost antrenat pe subsetul de antrenament. Acest proces implică ajustarea parametrilor modelului pentru a găsi relațiile optime între caracteristicile de intrare și variabila de ieșire (costurile medicale în acest caz).
* După antrenament, modelele au fost aplicate pe setul de date de testare utilizând operatorul Apply Model. Acest lucru a generat predicții pentru costurile medicale pe baza caracteristicilor din setul de date de testare.
* Prin evaluarea performanței fiecărui model pe setul de date de testare, am putut determina eficacitatea și acuratețea fiecărui model în predicția costurilor medicale.

**Validarea modelelor:**

* Pentru validarea modelelor, am comparat performanța fiecărui model pe setul de date de testare utilizând diferite metrici de evaluare, cum ar fi Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) și R-squared.
* Analizând rezultatele obținute pentru fiecare metrică de performanță, am identificat modelul care prezintă cea mai mică eroare și cea mai mare acuratețe în predicția costurilor medicale pe baza datelor de testare.
* Modelul care a obținut cele mai bune rezultate în validare a fost selectat ca model final pentru a fi utilizat în predicțiile ulterioare ale costurilor medicale.

Prin aplicarea riguroasă a acestor etape de estimare și validare, am asigurat că modelul nostru este bine pregătit și poate oferi predicții precise pentru costurile medicale, ceea ce ne permite să luăm decizii informate și să obținem rezultate fiabile în cadrul proiectului nostru de analiză a datelor.

**7.Rezultate**

**Performanța modelelor:**

1. **Decision Tree:**
   * RMSE (Root Mean Squared Error): 2313.618
   * MAE (Mean Absolute Error): 960.195
   * R-squared: N/A
2. **Random Forest:**
   * RMSE: 1833.837
   * MAE: 928.000
   * R-squared: N/A
3. **Gradient Boosted Trees:**
   * RMSE: 8164.494
   * MAE: 6282.925
   * R-squared: N/A

**Interpretarea rezultatelor:**

1. **Decision Tree:** Modelul Decision Tree a obținut un RMSE de 2313.618 și un MAE de 960.195. Aceste valori reflectă nivelul de eroare al modelului în predicția costurilor medicale. Cu cât aceste valori sunt mai mici, cu atât modelul este considerat mai precis în estimarea costurilor. Totuși, este important să reținem că R-squared nu este disponibil pentru modelul Decision Tree în RapidMiner, ceea ce limitează capacitatea noastră de a evalua ajustarea modelului la datele disponibile.
2. **Random Forest:** Modelul Random Forest a prezentat un RMSE de 1833.837 și un MAE de 928.000. Aceste valori sunt mai mici decât cele ale modelului Decision Tree, indicând o performanță mai bună în predicția costurilor medicale. Random Forest este un model de ansamblu care combină mai mulți arbori de decizie, ceea ce îi conferă o capacitate mai mare de generalizare și o rezistență mai bună la suprapunerea datelor.
3. **Gradient Boosted Trees:** Gradient Boosted Trees a obținut un RMSE de 8164.494 și un MAE de 6282.925. În comparație cu celelalte modele, acesta prezintă cea mai mare eroare în predicția costurilor medicale. Acest rezultat ar putea fi atribuit unei sensibilități mai mari a modelului la variabilitatea datelor și unei tendințe mai mari de a se supraantrena la datele de antrenament.

**Concluzii:**

Analizând rezultatele obținute pentru fiecare model, putem trage următoarele concluzii:

* Random Forest demonstrează cea mai bună performanță în predicția costurilor medicale, având cele mai mici valori pentru RMSE și MAE. Acest lucru sugerează că este cel mai potrivit model pentru a fi utilizat în predicțiile ulterioare ale costurilor medicale pe baza caracteristicilor disponibile în setul de date.
* Decision Tree ocupă locul doi în ceea ce privește performanța, în timp ce Gradient Boosted Trees prezintă cea mai mare eroare în predicție. Alegerea modelului potrivit depinde de prioritățile și necesitățile specifice ale proiectului și poate necesita o investigație mai detaliată a caracteristicilor și a comportamentului fiecărui model.

**8.Concluzii si Discutii:**

**Rezumat:**

După analiza atentă a performanței modelelor de regresie în cadrul proiectului nostru, constatăm că modelul Random Forest s-a evidențiat ca fiind cel mai potrivit pentru predicția costurilor medicale. Acest lucru se reflectă în cele mai mici valori obținute pentru metricile de evaluare, RMSE și MAE, indicând o performanță superioară în comparație cu modelele Decision Tree și Gradient Boosted Trees.

**Contribuții:**

1. **Importanța preprocesării datelor și selecției modelului:**
   * Am subliniat rolul crucial al preprocesării adecvate a datelor și a selecției modelului potrivit pentru obținerea unor predicții precise în domeniul costurilor medicale. Alegerea unui model adecvat și manipularea datelor într-un mod corespunzător sunt aspecte esențiale pentru succesul proiectului.
2. **Identificarea factorilor cheie:**
   * Prin analiza datelor, am identificat factorii cheie care influențează costurile medicale, furnizând astfel informații valoroase pentru optimizarea cheltuielilor în domeniul sănătății. Această înțelegere mai profundă a relațiilor între variabile ne permite să luăm decizii mai informate și să elaborăm strategii mai eficiente în managementul costurilor.

**Perspective viitoare:**

1. **Extinderea setului de date:**
   * O direcție viitoare importantă ar fi extinderea setului de date pentru a include variabile suplimentare relevante, cum ar fi istoricul medical al pacienților. Integrarea acestor informații suplimentare ar putea îmbunătăți precizia și generalizabilitatea modelului nostru, oferind o perspectivă mai cuprinzătoare asupra determinanților costurilor medicale.
2. **Investigarea altor tehnici:**
   * În viitor, este esențial să explorăm și să investigăm alte tehnici de machine learning și deep learning pentru a îmbunătăți performanța predicțiilor și a crește acuratețea modelelor noastre. Această explorare continuă ne poate ajuta să identificăm și să implementăm soluții inovatoare care să răspundă nevoilor și cerințelor din ce în ce mai complexe ale domeniului sănătății. Prin adăugarea de noi abordări și tehnici în arsenalul nostru de analiză, putem îmbunătăți continuu procesul nostru de luare a deciziilor și să contribuim la avansarea cunoștințelor în domeniul sănătății.