Predicția diabetului de tip 2 utilizând regresia logistică

Cerescu Marius
Universitatea Tehnică a Moldovei
marius.cerescu@gmail.com

Abstract

Diabetul zaharat este una dintre cele mai frecvente boli umane la nivel mondial și poate provoca mai multe complicații legate de sănătate. Acesta este responsabil pentru o morbiditate, mortalitate și pierderi economice. Un diagnostic și o predicție în timp util a acestei boli ar putea oferi pacienților o oportunitatea de a adopta strategii preventive și de tratament adecvate. Pentru a îmbunătăți înțelegerea factorilor de risc, am prezis diabetul de tip 2 pentru femeile indiene Pima utilizând un model de regresie logistică. Analiza noastră găsește cinci predictori principali pentru diabetul de tip 2: glucoza, sarcina, indicele de masă corporală (IMC), funcția pedigree a diabetului și vârsta. Specificația noastră preferată produce o precizie de predicție de 78,26 % și o rată de eroare de validare încrucișată de 21.74%. Susținem că modelul nostru poate fi aplicat pentru a face o predicție rezonabilă a diabetului de tip 2, și ar putea fi utilizat pentru a completa măsurile preventive existente pentru a reduce incidența diabetului zaharat și reducerea costurilor asociate.

1. Introducere

Diabetul este una dintre cele mai frecvente boli umane și a devenit o problemă semnificativă de sănătate publică la nivel mondial. Au fost diagnosticate aproximativ 450 de milioane de persoane cu diabet, ceea ce a dus la aproximativ 1,37 milioane de decese la nivel global în 2017 [1]. Mai mult de 100 de milioane de adulți din SUA trăiesc cu diabet, iar aceasta a fost a șaptea cauză principală de deces în SUA în 2020 [2]. Unul din zece adulți americani suferă de diabet în prezent, iar dacă actuala tendință continuă, se estimează că până la unul din trei adulți americani ar putea avea diabet până în 2050 [2].

Pacienții cu diabet prezintă un risc ridicat de a dezvolta complicații de sănătate, cum ar fi insuficiența renală, pierderea vederii, boli de inimă, accident vascular cerebral, deces prematur și amputarea membrelor, ceea ce poate duce la disfuncționalități și deteriorarea cronică a țesuturilor [3]. În plus, există costuri economice substanțiale asociate cu boala. Prețul total estimat al diabetului diagnosticat în SUA a crescut la 237 miliarde USD în 2017, de la 188 miliarde USD în 2012. Costurile medicale în exces pe persoană asociate cu diabetul au crescut la 9601 USD de la 8417 USD în aceeași perioadă [2]. Există, de asemenea, pierderi de productivitate în cadrul forței de muncă din cauza pacienților cu diabet.

Este posibil ca o persoană cu risc ridicat de diabet să nu fie conștientă de factorii de risc asociați cu această boală. Având în vedere prevalența și gravitatea ridicată a diabetului, cercetătorii sunt interesați de identificarea celor mai frecvenți factori de risc, deoarece acesta poate fi cauzat de o combinație de mai multe motive. Determinarea factorilor de risc și predicția precoce a diabetului sunt vitale în reducerea complicațiilor [4,5] și a costurilor economice [6], beneficiind atât practicii clinice, cât și sănătății publice [7]. Studiile au constatat că identificarea persoanelor cu risc ridicat este esențială pentru implementarea eficientă a măsurilor de prevenire [8]. Intervenția timpurie poate contribui la prevenirea complicațiilor și îmbunătățirea calității vieții [9,10].

Există dovezi tot mai mari că modificarea stilului de viață previne sau întârzie diabetul de tip 2 [11]. Principalii factori de risc ai diabetului sunt o dietă nesănătoasă, îmbătrânirea, istoricul familial, grupurile etnice, obezitatea, sedentarismul și istoricul anterior de diabet gestațional [6,7,12]. Studiile anterioare au arătat, de asemenea, că sexul, indicele de masă corporală (IMC), sarcina și starea metabolică sunt asociate cu diabetul [13,14].

Modelele de predicție pot identifica persoanele cu prediabet sau cu risc crescut de a dezvolta diabet, ajutând la luarea deciziilor în

managementul clinic al pacienților. Numeroase modele și ecuații predictive au fost propuse pentru a modela factorii de risc ai diabetului incident [15-17]. De exemplu, Heikes și colaboratorii săi [18] au studiat un instrument de predicție a riscului de diabet în SUA folosind date nediagnosticate și date privind prediabetul, în timp ce Razavian și colaboratorii săi [19] au dezvoltat modele de predicție bazate pe regresia logistică pentru apariția diabetului de tip 2. Aceste modele ajută, de asemenea, la identificarea persoanelor cu risc ridicat de a dezvolta diabet. Zou și colaboratorii săi [20] au utilizat metode de învățare automată pentru a prezice diabetul în Luzhou, China, validând modelele prin cinci validizări încrucișate. Nguyen și colaboratorii săi [5] au prezis apariția diabetului utilizând algoritmi de învățare profundă, sugerând că metodele sofisticate pot îmbunătăți performanța modelelor. În schimb, alte studii au arătat că regresia logistică are performanțe cel puțin la fel de bune ca și tehnicile de învățare automată pentru predicția riscului de boală ([21,22], de exemplu). Anderson și colaboratorii săi au utilizat regresia logistică împreună cu algoritmi de învățare automată și au constatat o acuratețe mai mare cu modelul de regresie logistică. Aceste modele se bazează în principal pe evaluarea factorilor de risc ai diabetului, cum ar fi caracteristicile gospodăriei și individuale; cu toate acestea, lipsa unui instrument obiectiv și imparțial de evaluare rămâne o problemă. Există, de asemenea, îngrijorări tot mai mari cu privire la dezvoltarea slabă a acestor modele predictive din cauza selecției necorespunzătoare a covarianței, datelor lipsă, mărimii mici a eșantionului și modelelor statistice greșit specificate [25,26]. În acest sens, doar câteva modele de predicție a riscului sunt utilizate în mod curent în practica clinică. Fiabilitatea și calitatea acestor instrumente și ecuații de predicție variază semnificativ în funcție de geografie, datele disponibile și etnie [5]. Este posibil ca factorii de risc pentru un grup etnic să nu fie generalizați la altele; de exemplu, prevalența diabetului este raportată ca fiind mai mare în rândul comunității indiene Pima. Prin urmare, acest studiu utilizează setul de date Pima Indian pentru a prezice dacă o persoană este expusă riscului de a dezvolta diabet pe baza unor diagnostice specifice factori specifici [27,28].

Utilizând o abordare precum regresia logistică, ne propunem să prezicem factorii de risc ai diabetului zaharat de tip 2. Regresia logistică compară mai multe modele de predicție pentru prezicerea diabetului. Am utilizat diverse criterii de selecție populare în literatura de specialitate, cum ar fi AIC, BIC, Cp al lui Mallows, R ajustat 2, și selecția înainte și înapoi pentru a identifica predictorii semnificativi. Lucrarea noastră contribuie, de asemenea, la literatura extinsă despre factorii de risc ai diabetului. Extindem cercetările anterioare prin explorarea unor metode tradiționale și a algoritmilor simpli de învățare automată pentru a valorifica literatura de specialitate privind predicția diabetului.

Analiza noastră identifică cinci predictori principali ai diabetului: glucoza, sarcina, indicele de masă corporală, vârsta și funcția pedigreu a diabetului. Un model de regresie care include acești cinci factori de predicție produce o precizie de predicție de 77,73% și o rată de eroare de validare încrucișată de 22,65%. Acest studiu contribuie la informarea factorilor de decizie și la proiectarea politicii de sănătate pentru a reduce prevalența diabetului. Metodele care prezintă cele mai bune performanțe de clasificare variază în funcție de structura datelor; prin urmare, studiile viitoare ar trebui să fie prudenți în utilizarea unui singur model sau a unei singure abordări pentru predicția riscului de diabet. Restul acestui articol este structurat în următoarele secțiuni: "Date și statistici sumare", "Metode", "Rezultate" și "Concluzii".

2. Date și statistici sumare

Mai multe variabile pot fi asociate cu diabetul, inclusiv tensiunea arterială, sarcina la femei, vârsta și indicele de masă corporală. Pentru a gestiona diabetul, este crucial să înțelegem ce factori sunt legați de această boală. Am utilizat setul de date Pima Indian de la Universitatea Johns Hopkins, care oferă informații pentru a prezice riscurile asociate cu diabetul. Indienii Pima sunt nativi americani stabiliți în sudul Arizonei, de-a lungul râurilor Gila și Salt. Fiecare intrare din setul de date reprezintă un pacient și conține informații medicale precum numărul de sarcini, concentrația plasmatică de glucoză, grosimea pliului cutanat triceps, indicele de masă corporală (IMC), tensiunea arterială diastolică, nivelul seric al insulinei după

2 ore, vârsta și funcția pedigree pentru diabet. Variabila noastră de răspuns, diabetul, este marcată cu 1 pentru un diagnostic de tip 2 și cu 0 în caz contrar. Din eșantion, 268 de pacienți (34,9%) au fost diagnosticați cu diabet. Pentru cinci dintre variabile - insulină, glucoză, IMC, grosimea pielii și tensiunea arterială - existau valori lipsă, marcate cu zero, care nu erau semnificative. Prin urmare, am înlocuit aceste zero-uri cu valorile lor mediane corespunzătoare. Datele au fost analizate folosind programul statistic R versiunea 4.0.5.

Tabelul 1 prezintă statisticile descriptive pentru toți predictorii după imputarea medianelor pentru valorile lipsă. Observăm că tensiunea arterială, IMC-ul și grosimea pielii au medii și mediane apropiate. Variabila "pedigree" prezintă cea mai mică deviație standard, în timp ce insulina are cea mai mare deviație standard, indicând o variație mai mică pentru "pedigree" și o variație mai mare pentru insulina prezentă în distribuțiile lor. Scopul nostru este să identificăm un subset potrivit de covariante pentru a fi incluse într-un model predictiv pentru diabet. Excluderea unor predictori poate distorsiona variabilele, în timp ce prea mulți predictorii pot scădea precizia. Există diverse metode pentru a dezvolta un model predictiv și ecuații adecvate. Printre acestea, am aplicat regresia logistică.

Table 1. Descriptive statistics.

Variable	Definition	Mean	Std. dev.	Median
Pregnancy	Frequency of pregnancy	3.85	3.37	3.00
Glucose	Concentration of plasma glucose (mg/dL)	121.66	30.44	117.00
BP	Diastolic blood pressure (mm Hg)	72.39	12.10	72.00
Skin	Tricep skinfold thickness (mm)	29.11	8.79	29.00
Insulin	Two-hour serum insulin (mu U/mL)	140.67	86.38	125.00
BMI	Body mass index (kg/m^2)	32.46	6.88	32.30
Pedigree	A pedigree function for diabetes	0.47	0.33	0.37
Age	Age (log (years))	33.24	11.76	29.00

3. Metode

Regresiile logistice explorează legătura dintre variabila de răspuns categorică și covariante. Ele utilizează o combinație liniară a variabilelor independente pentru a estima log-odds-ul probabilității unui eveniment într-un model logistic. În cazul regresiilor logistice binare, se evaluează probabilitatea ca o caracteristică a unei variabile binare să fie prezentă, având în vedere valorile covariantelor. Presupunem că avem o variabilă de răspuns binară, unde Yi = 1 dacă caracteristica este prezentă și Yi = 0 dacă caracteristica este absentă, iar datele [Y1, Y2, ..., Yn] sunt independente. πi reprezintă probabilitatea de succes. De asemenea, avem x = (x1, x2, ..., xpS) ca un set de variabile explicative, care pot fi discrete, continue sau o combinație a ambelor tipuri. Funcția logistică pentru πi este definită de relația:

$$logit(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{i,p};$$

unde

$$\pi_i = \frac{exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{i,p})}{1 + exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{i,p})} = \frac{exp(x_i'\beta)}{1 + exp(x_i'\beta)} = \Lambda(x_i'\beta)$$

În această ecuație, πi reprezintă probabilitatea ca un eșantion să aparțină unei anumite categorii a variabilei de răspuns binare, adesea denumită "probabilitatea de succes", având o valoare între 0 și 1. Funcția Λ(..) este funcția de distribuție cumulată logistică, cu λ(z) = e^z / (1 + e^(-z)) = 1 / (1 + e^(-z)), iar β reprezintă un vector de parametri care urmează să fie estimat.

3.1 Estimarea și testul raportului de verosimilitate

Metoda preferată pentru estimarea lui β este cea a probabilității maxime, deoarece are proprietăți statistice mai bune, deși putem folosi și metoda celor mai mici pătrate. Să considerăm, modelul logistic cu o singură variabilă predictivă X dată de funcția logistică de

$$\pi(X) = \frac{\exp(X_i\beta)}{1 + \exp(X_i\beta)}$$

Dorim să găsim estimările astfel încât, dacă introducem βˆ în modelul pentru π(X), să obținem un număr apropiat de unu pentru toți subiecții care au diabet și aproape de zero în caz contrar. Din punct de vedere matematic, funcția de verosimilitate este dată de formula

$$L(\beta_0, \beta_1) = \prod_{i:y_i=1} \pi(x_i) \prod_{i':y_{i'}=0} (1 - \pi(x_{i'}))$$

Estimările βˆ sunt alese pentru a maximiza această funcție de verosimilitate. Se ia logaritmul pe ambele părți pentru a calcula și a utiliza funcția logaritmică de verosimilitate în scopul estimării. Am utilizat raportul de verosimilitate pentru a testa dacă vreun subset de estimări β este zero. Să presupunem că p și r reprezintă numărul de β în modelul complet și, respectiv, în modelul redus. Statistica de testare a raportului de verosimilitate este dată de

$$\Lambda^* = -2[l(\hat{\beta}^{(0)}) - l(\hat{\beta})],$$

unde l(βˆ) și l(βˆ(0)) sunt probabilitățile logaritmice ale modelului complet și, respectiv, ale modelului redus, evaluate la estimarea de maximă verosimilitate (MLE) a modelului redus și Λ∗ ∼ χ 2 p-r ; p și r fiind numărul de parametri

din modelul complet și, respectiv, din modelul redus.

3.2 Metoda de validare și validare încrucișată

Putem estima eroarea de testare folosind metodele de validare a setului de validare și de eroare de validare încrucișată ca alternative la abordările descrise mai sus. Am calculat eroarea de validare încrucișată și eroarea setului de validare pentru fiecare model luat în considerare, apoi am ales modelul cu cea mai mică eroare de testare ca fiind specificația noastră preferată. Această metodă poate fi aplicată și în alte situații de selecție a modelelor, mai ales atunci când numărul de covariante din model este dificil de estimat sau când varianța erorii σ^2 este complexă.

Abordarea setului de validare: Pentru a evalua rata de eroare de testare asociată cu o anumită metodă pe un set de eșantioane, am folosit abordarea setului de validare. Aceasta constă în împărțirea aliatoare a eșantioanelor disponibile într-un set de instruire și un set de validare. Modelul este antrenat folosind setul de instruire, iar apoi este utilizat pentru a prezice răspunsurile din setul de validare. Rata de eroare a setului de validare este evaluată în mod obișnuit utilizând eroarea medie pătratică (MSE).

Validarea încrucișată k-fold: Această abordare implică împărțirea aliatoare a eșantionului în k grupuri sau pliuri de dimensiuni egale. Fiecare grup este utilizat pe rând ca set de validare, în timp ce modelul este antrenat folosind celelalte k-1 grupuri. Eroarea medie pătratică (MSE) este calculată folosind datele din grupul reținut. Acest proces este repetat de k ori, cu fiecare grup servind ca set de validare cel puțin o dată. Avantajul major al acestei metode constă în faptul că întregul set de date este atât antrenat, cât și testat, contribuind la reducerea variației. Estimarea CV k-fold este determinată prin calculul mediei acestor valori MSE1, MSE2, - - - - , MSEk.

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$

4. Rezultate

Figura 1 reprezintă graficul de corelație, care indică intensitatea corelațiilor dintre diferite perechi de predictori. Corelațiile pe perechi (r) dintre sarcină și vârstă (0,59) și dintre IMC și grosimea pielii (0,54) sunt (r > 0,5) ridicate în comparație cu alte perechi, ceea ce indică faptul că aceste două perechi de predictori sunt corelate în mod semnificativ.

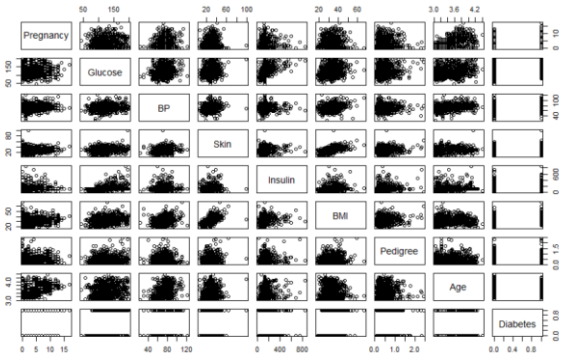


Figure 1. Correlation plot of the variables used in the study. Note: BP and BMI represent blood pressure and body mass index, respectively.

4.1 Evaluarea variațiilor nivelurilor de glucoză din sânge între subiecții diagnosticați cu diabet și cei sănătoși

Prima ipoteză care a fost testată este că nivelurile de glucoză din sânge variază semnificativ între indivizii diagnosticați cu diabet și cei fără această.

Primul lucru pe care l-am făcut a fost să vizualizez cu ajutorul unui box plot nivelurile de glucoză pentru persoanele care au și care nu au diabet.

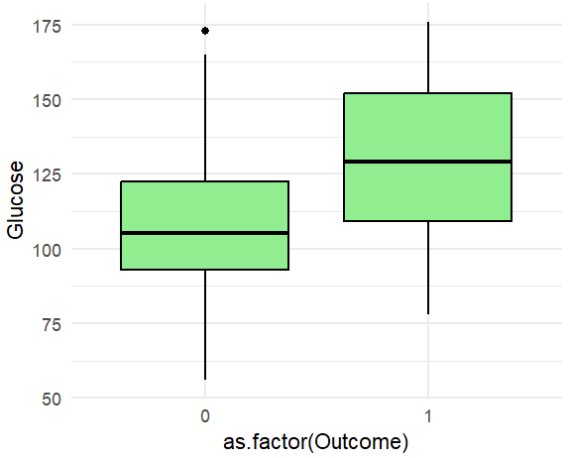


Figura 2. Analiza comparativă a nivelurilor de glucoză

Cu ajutorul box plot-ului am observat că mediana este mai mare pentru grupul de persoane care au diabet, însă trebuia să testăm ipoteza cu ajutorul unui test statistic, de aceea am decis să utilizăm testul statistic t.

```
        Welch Two Sample t-test

data:  Glucose by Outcome
t = -6.7198, df = 82.684, p-value = 2.164e-09
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -31.34895 -17.02902
sample estimates:
mean in group 0 mean in group 1
       108.2277        132.4167
```

În urma testului efectuat am obținut valoare p = 2.165e-9 care este o valoare extrem de mică, iar deoarece valoarea obținută este mai mică decât 0.05 putem afirma că într-adevăr concentrația de glucoză în sânge dintre persoanele care au și care nu au diabet este diferită.

4.2 Examinarea variațiilor incidenței diabetului în funcție de grupul de vârstă

Următoarea ipoteză testată a fost că incidența diabetului variază în funcție de grupul de vârstă.

Primul lucru pe care l-am făcut a fost să împărțim în 3 grupe persoanele în dependență de vârstă. Young – persoane până la 25 de ani. Middle-Aged – persoane între 25 – 40 de ani. Old – persoane mai în vârstă de 40 de ani.

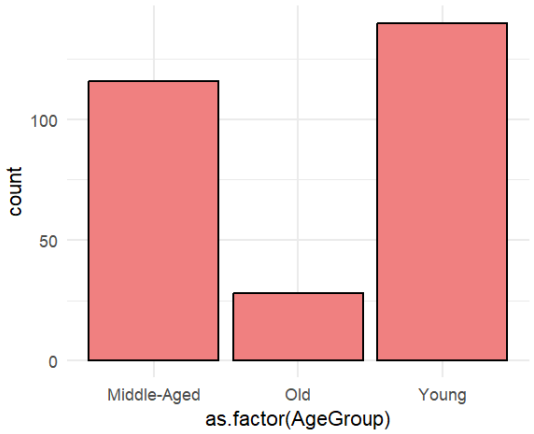


Figura 3. Distribuția grupelor de vârstă în studiul populației

În continuare am luat 10 persoane la întâmplare din fiecare categorie și am numărat câte persoane au diabet din fiecare grup.

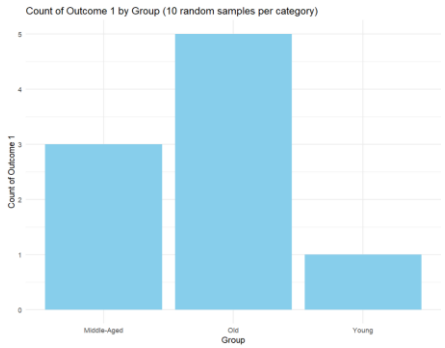


Figura 4. Analiza frecvenței prezenței diabetului pe grupe de vârstă

Am observat că în grupurile Middle-Aged și Old sunt mult mai multe persoane cu diabet.

În continuare pentru a testa ipoteza am decis să folosesc testul statistic Chi squared.

```
        Pearson's Chi-squared test

data:  cont_table[, 2:3]
X-squared = 20.775, df = 2, p-value = 3.081e-05
```

Deoarece am obținut p-value 3.08e-5 < 0.05 se poate conclude că există o diferență de proporții dintre aceste 3 categorii.

4.3 Investigarea variațiilor mediei nivelului de insulină în funcție de categoria de greutate a individului

Următoarea ipoteză testată a fost că media nivelului de insulină variază în funcție de categoria de greutate a individului.

Inițial am decis să vizualizez categoriile de BMI și numărul de persoane din fiecare categorie.

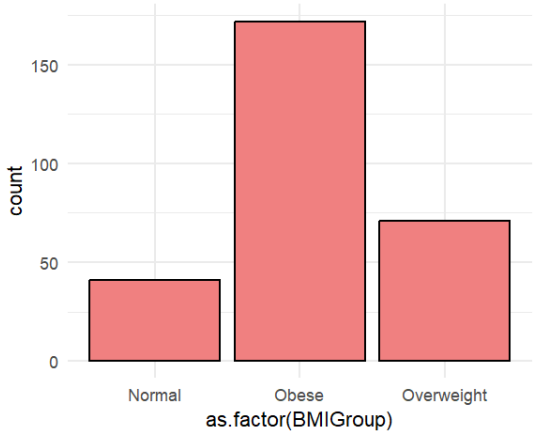


Figura 5. Analiza comparativă a prevalenței BMI

În continuare am decis să vizualizez distribuția de insulină din fiecare categorie.

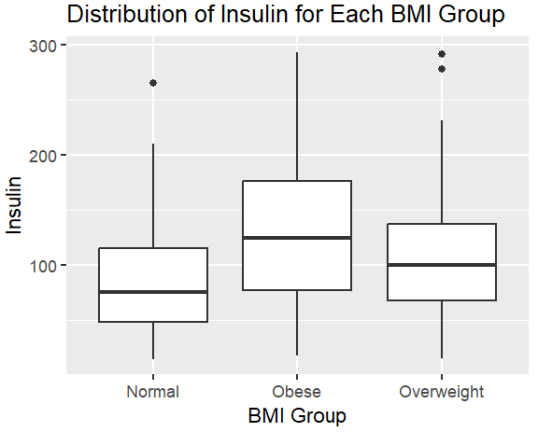


Figura 6. Distribuția nivelurilor de insulină pe grupe BMI

Cu ajutorul boxplot-urilor am observant că există o diferență în media nivelului de insulină. Am decis să realizez testul statistic ANOVA pentru a testa ipoteza.

```
            Df  Sum Sq Mean Sq F value   Pr(>F)
BMIGroup     2   65580   32790   8.958 0.000169 ***
Residuals  281 1028541    3660
```

Deoarece p-value este 0.000169 < 0.05 putem conclude că ipoteza este adevărată.

4.4 Explorarea corelației dintre grosimea pielii și indicele de masă corporală.

Următoarea ipoteză testată a fost că grosimea pielii este corelată cu indicele de masă corporală.

Inițial am realizat un scatter plot cu grosimea pielii și indicele de masă corporală pentru a observa dacă există o corelație.

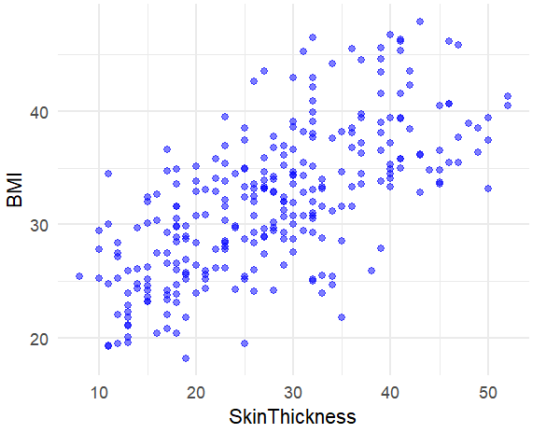


Figura 7. Vizualizarea relației dintre BMI și grosimea pielii

În urma analizei plot-ului am observat că există o corelație între nivelul indicelui de masă corporală și grosimea pielii.

În continuare pentru a testa ipoteza am decis să realizez testul de corelație.

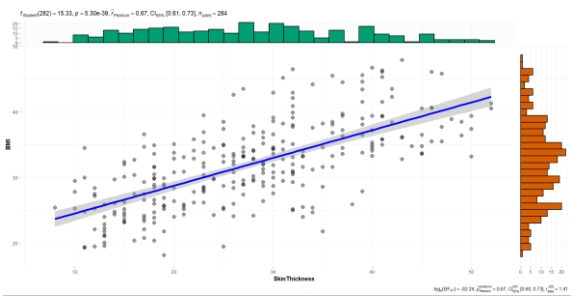


Figura 8. Corelația între grosimea pielii și BMI

În urma testului realizat am obținut valoarea corelației 0.67 și p-value = 5.30e-39 ceea ce conclude că există o corelație foarte puternică între BMI și SkinThickness.

4.5 Crearea unui model de prezicere a diabetului zaharat

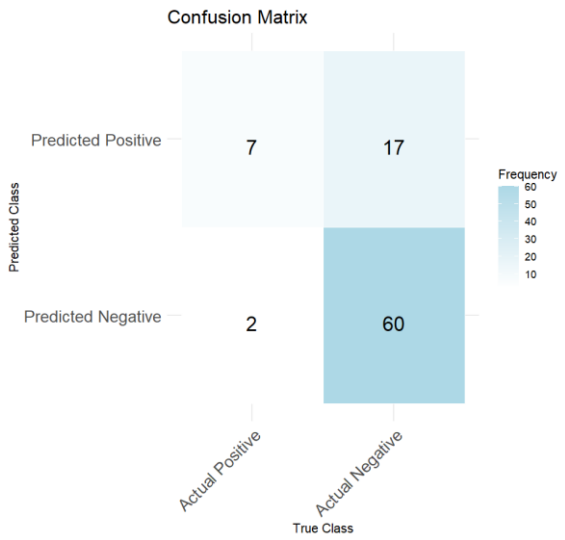
Deoarece setul de date conține atât variabile numerice cât și categoriale, iar clasa de prezicere este o variabilă binară pentru crearea acestui model se potrivește cel mai bine regresia logistică.

Inițial am împărțit setul de date într-un set de antrenare și unul pentru testare, oferind pentru antrenare 70% din date și pentru testare 30%.

testing_data	86 obs. of 12 variables
training_data	198 obs. of 12 variables

În continuare am creat modelul utilizând setul de date pentru antrenare, am creat matricea de confuzie și am calculat acuratețea modelului.

Matricea de confuzie:



Acuratețea modelului:

```
"Accuracy: 0.779069767441861"
```

Conform matricei de confuzie, modelul a prezis corect clasa “0” de 60 de ori și a greșit de 2 ori. De asemenea, a prezis corect clasa “1” de 7 ori și a greșit de 17 ori.

În urma calculării acurateței am obținut valoarea 0.78, ceea ce este o valoare bună, dar care mai trebuie îmbunătățită.

În continuare am folosit librăria caret și ggplot2 pentru a afla și vizualiza importanța fiecărei variabile din model.

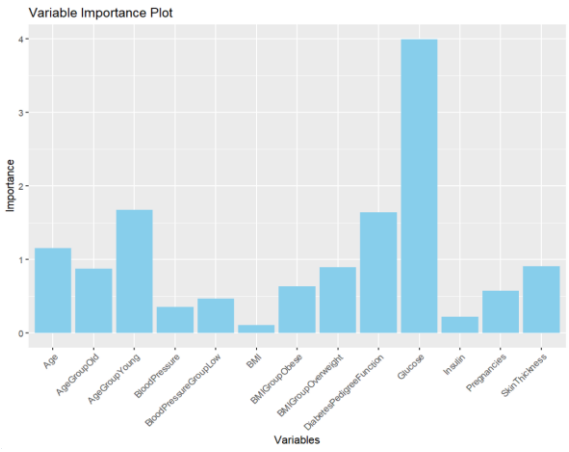


Figura 9. Analiza importanței variabilelor

În diagrama prezentată se observă că cea mai importantă variabilă este glucoza, urmată de funcția pedigree și AgeGroupYoung.

Este foarte surprinzător că astfel de variabile precum insulina și BMI sunt cele mai nesemnificative variabile.

În urma analizei rezultatelor am ajuns la concluzia că modelul este compus din multe variabile care au o influență foarte mică și că glucoza este cu mult mai importantă decât restul variabilelor.

4.6 Crearea unor modele noi

În continuare am decis că creez două modele noi. Unul care folosește doar glucoza și unul care folosește glucoza și funcția pedigree.

Pentru compararea modelelor am utilizat diagrama ROC și am calculat aria de sub curbă.

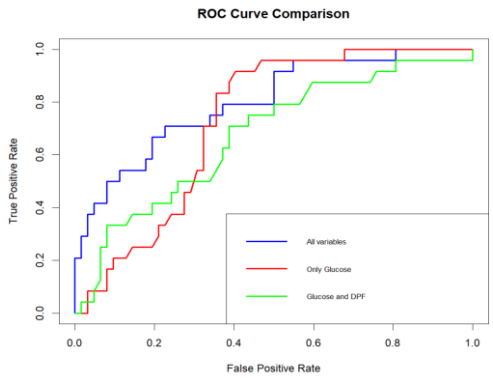


Figura 10. Analiza curbelor ROC pentru diferite seturi de variabile

```
Area under the ROC curve (AUC) - All variables: 0.8010753

Area under the ROC curve (AUC) - Only Glucose: 0.7281586

Area under the ROC curve (AUC) - Glucose and DPF: 0.6743952
```

Conform valorilor AUC obținute pentru fiecare model în parte, se poate observa că primul model, care a fost antrenat pe toate variabilele a obținut cea mai mare valoare, prin urmare îl putem considera cel mai bun.

De asemenea se poate observa că modelul antrenat doar pe glucoză a obținut o valoare AUC mai mare decât modelul antrenat atât pe glucoză cât și pe funcția pedigree. Putem face concluzia că funcția pedigree a contribuit la înrăutățirea modelului în acest caz.

5. Discuții

Diabetul a devenit una dintre principalele cauze ale decesului uman în ultimele decenii, iar incidența sa a crescut continuu, din motive precum obiceiurile alimentare, stilul de viață sedentar și prevalența alimentelor nesănătoase. Dezvoltarea unui model de predicție pentru diabet poate fi esențială în gestionarea clinică, ajutând la identificarea potențialilor factori de risc și la identificarea persoanelor cu risc crescut pentru prevenirea diabetului. Regresia logistică [21] și arborii de clasificare bazată pe învățarea automată [20] sunt printre cele mai populare metode utilizate pentru acest scop. Unele studii, cum ar fi cel realizat de Habibi et al. [6], sugerează că un arbore de clasificare simplu ar putea fi utilizat pentru detectarea diabetului fără a necesita un laborator. Cu toate acestea, valabilitatea acestor modele pentru diferite locații, populații cu diete diferite, stiluri de viață, rase și patrimoniu genetic rămâne încă necunoscută. În plus, există un număr limitat de ecuații predictibile fiabile pentru femeile indiene Pima. Performanța și valabilitatea acestor predicții variază considerabil. Pentru a corecta această lacună în literatură, acest studiu a utilizat regresia logistică pe setul de date al indienilor Pima pentru a identifica factorii importanți pentru diabetul de tip 2. Am selectat variabilele bazate pe testul de potrivire a datelor și pe criteriile de selecție a modelului, cum ar fi AIC, BIC și Mallows' Cp.

Un IMC mai mare, secreția și acțiunea redusă a insulinei, antecedentele familiale de diabet, tensiunea arterială, fumatul și starea de sarcină sunt considerate factori de risc comuni pentru diabetul de tip 2 [33,34]. Conform studiului nostru, cei cinci predictorii principali pentru diabet sunt frecvența sarcinilor la femei, nivelul de glucoză, pedigreul, IMC-ul și vârsta.

Aceste variabile au fost utilizate și în studii anterioare pentru a prezice diabetul. De exemplu, Bays et al. au constatat că creșterea IMC-ului este asociată cu un risc crescut de diabet zaharat. Un studiu din Finlanda a dezvoltat un scor de risc pentru diabet și a identificat că vârsta, istoricul parental de diabet, IMC-ul, nivelul crescut de zahăr din sânge și activitatea fizică sunt printre predictori. Persoanele cu niveluri mai mari de glucoză au un risc crescut de a dezvolta diabet, deoarece glucoza este legată de răspunsul la insulină. Lyssenko et al. au raportat că antecedentele familiale de diabet ar putea dubla riscul de boală. Pedigreul furnizează informații despre istoricul familial al diabetului pentru a prezice probabilitatea unei persoane de a dezvolta diabet. Un IMC mai mare poate duce la obezitate, afectând funcția celulelor pancreatice și conducând la rezistență la insulină. Vârsta este un alt factor de risc pentru debutul diabetului, fiind asociată cu scăderea sensibilității la glucoză și a secreției de insulină odată cu înaintarea în vârstă. Validarea modelului nostru arată o performanță predictivă relativ bună, cu o acuratețe de 78%, apropiată de studiile anterioare privind factorii de risc pentru diabet. De exemplu, Lyssenko et al. au raportat rate de acuratețe de 74-77% pentru două locații diferite în studiul factorilor de risc pentru diabet. Zou et al. [20] au prezis diabetul cu o acuratețe de 77% și 81% pentru seturile de date Pima Indian și Luzhou, respectiv. Așa cum indică și Wilson et al., am constatat că modelele complexe nu sunt întotdeauna necesare pentru a prezice boli; regresia logistică și tehnici precum arborii de clasificare pot fi la fel de utile în prezicerea diabetului. Totuși, validarea modelului propus între diferite grupuri de populație ar trebui să fie efectuată.

Recunoaștem limitele analizei noastre. În primul rând, doar câțiva predictorii au fost luați în considerare din cauza limitărilor datelor disponibile, ceea ce ar putea să nu permită generalizarea concluziilor la seturi de date mai ample cu mai mulți predictori. În al doilea rând, chiar și cele mai bune modele predictive și procesele de selecție a variabilelor pot genera rezultate diferite în funcție de locație, tipul setului de date și algoritmii utilizați. În fine, înlocuirea valorilor lipsă cu medianele variabilelor respective, deși o practică comună, ar putea influența rezultatele. Studiile viitoare ar trebui să includă o varietate mai largă de factori de risc, cum ar fi caracteristicile genetice, genul, statutul socio-economic, activitatea fizică, fumatul, comportamentul și atitudinile legate de sănătate, consumul alimentar și cheltuielile pentru a prezice diabetul într-o populație mai diversă.

6. Concluzii

A identifica indivizii cu un risc crescut de a dezvolta diabet reprezintă un pas crucial în prevenirea și gestionarea acestei boli. Studiul propune o ecuație de predicție a diabetului pentru a ilustra factorii de risc esențiali, facilitând astfel clasificarea persoanelor cu risc ridicat, diagnosticul și strategiile de prevenție. Cele cinci variabile critice identificate pentru predicția diabetului de tip 2 sunt vârsta, IMC-ul, pedigreul, nivelul glucozei și frecvența sarcinilor. Concluzionăm că modelul nostru propus prezintă o acuratețe de predicție de 78.26%, cu o rată de eroare în validare încrucișată de 22.86%. Aceste rezultate sugerează că gestionând acești cinci predictori prin intermediul măsurilor adecvate, putem reduce prevalența diabetului de tip 2. De asemenea, o prezicere precisă a diabetului ar putea servi ca fundament pentru elaborarea intervențiilor și implementarea politicilor de sănătate care să contribuie la prevenirea acestei boli.

Referințe

1. Cho, N.; Shaw, J.; Karuranga, S.; Huang, Y.; da Rocha Fernandes, J.; Ohlrogge, A.; Malanda, B. IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes Res. Clin. Pract. 2018, 138, 271–281. [CrossRef] [PubMed]
2. CDC. Centers for Disease Control and Prevention and Others; National Diabetes Statistics Report; Centers for Disease Control and Prevention, US Department of Health and Human Services: Atlanta, GA, USA, 2020; pp. 12–15.
3. Krasteva, A.; Panov, V.; Krasteva, A.; Kisselova, A.; Krastev, Z. Oral cavity and systemic diseases—Diabetes mellitus. Biotechnol. Biotechnol. Equip. 2011, 25, 2183–2186. [CrossRef]
4. Alghamdi, M.; Al-Mallah, M.; Keteyian, S.; Brawner, C.; Ehrman, J.; Sakr, S. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford Exercise Testing (FIT) project. PLoS ONE 2017, 12, e0179805. [CrossRef]
5. Nguyen, B.P.; Pham, H.N.; Tran, H.; Nghiem, N.; Nguyen, Q.H.; Do, T.T.; Tran, C.T.; Simpson, C.R. Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records. Comput. Methods Programs Biomed. 2019, 182, 105055. [CrossRef] [PubMed]
6. Habibi, S.; Ahmadi, M.; Alizadeh, S. Type 2 diabetes mellitus screening and risk factors using decision tree: Results of data mining. Glob. J. Health Sci. 2015, 7, 304. [CrossRef]
7. Ryden, L.; Standl, E.; Bartnik, M.; Van den Berghe, G.; Betteridge, J.; De Boer, M.J.; Cosentino, F.; Jönsson, B.; Laakso, M.; Malmberg, K.; et al. Guidelines on diabetes, pre-diabetes, and cardiovascular diseases: Executive summary: The Task Force on Diabetes and Cardiovascular Diseases of the European Society of Cardiology (ESC) and of the European Association for the Study of Diabetes (EASD). Eur. Heart J. 2007, 28, 88–136. [PubMed]
8. Tuso, P. Prediabetes and lifestyle modification: Time to prevent a preventable disease. Perm. J. 2014, 18, 88. [CrossRef] [PubMed]
9. IDF Clinical Guidelines Task Force. Global Guideline for Type 2 Diabetes: Recommendations for standard, comprehensive, and minimal care. Diabet. Med. 2006, 23, 579–593. [CrossRef]
10. Gregg, E.W.; Geiss, L.S.; Saaddine, J.; Fagot-Campagna, A.; Beckles, G.; Parker, C.; Visscher, W.; Hartwell, T.; Liburd, L.; Narayan, K.V.; et al. Use of diabetes preventive care and complications risk in two African-American communities. Am. J. Prev. Med. 2001, 21, 197–202. [CrossRef]
11. Knowler, W.C.; Barrett-Connor, E.; Fowler, S.E.; Hamman, R.F.; Lachin, J.M.; Walker, E.A.; Nathan, D.M.; Diabetes Prevention Program Research Group . Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. N. Engl. J. Med. 2002, 346, 393–403.
12. Wild, S.; Roglic, G.; Green, A.; Sicree, R.; King, H. Global prevalence of diabetes: Estimates for the year 2000 and projections for 2030. Diabetes Care 2004, 27, 1047–1053. [CrossRef] [PubMed]
13. Engelgau, M.M.; Narayan, K.; Herman, W.H. Screening for type 2 diabetes. Diabetes Care 2000, 23, 1563–1580. [CrossRef] [PubMed]
14. Rolka, D.B.; Narayan, K.V.; Thompson, T.J.; Goldman, D.; Lindenmayer, J.; Alich, K.; Bacall, D.; Benjamin, E.M.; Lamb, B.; Stuart, D.O.; et al. Performance of recommended screening tests for undiagnosed diabetes and dysglycemia. Diabetes Care 2001, 24, 1899–1903. [CrossRef] [PubMed]
15. Schwarz, P.E.; Li, J.; Lindstrom, J.; Tuomilehto, J. Tools for predicting the risk of type 2 diabetes in daily practice. Horm. Metab. Res. 2009, 41, 86–97. [CrossRef]
16. Yu, W.; Liu, T.; Valdez, R.; Gwinn, M.; Khoury, M.J. Application of support vector machine modeling for prediction of common diseases: The case of diabetes and pre-diabetes. BMC Med. Inform. Decis. Mak. 2010, 10, 1–7. [CrossRef]
17. Naz, H.; Ahuja, S. Deep learning approach for diabetes prediction using PIMA Indian dataset. J. Diabetes Metab. Disord. 2020, 19, 391–403. [CrossRef]
18. Heikes, K.E.; Eddy, D.M.; Arondekar, B.; Schlessinger, L. Diabetes Risk Calculator: A simple tool for detecting undiagnosed diabetes and pre-diabetes. Diabetes Care 2008, 31, 1040–1045. [CrossRef]
19. Razavian, N.; Blecker, S.; Schmidt, A.M.; Smith-McLallen, A.; Nigam, S.; Sontag, D. Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. Big Data 2015, 3, 277–287. [CrossRef] [PubMed]
20. Zou, Q.; Qu, K.; Luo, Y.; Yin, D.; Ju, Y.; Tang, H. Predicting diabetes mellitus with machine learning techniques. Front. Genet. 2018, 9, 515. [CrossRef]
21. Christodoulou, E.; Ma, J.; Collins, G.S.; Steyerberg, E.W.; Verbakel, J.Y.; Van Calster, B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J. Clin. Epidemiol. 2019, 110, 12–22. [CrossRef] [PubMed]