Predikcija rezultata medicinskog testa na dijabedes

Nenad Joldić
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad
joldicnenad13@gmail.com

Bogdan Blagojević
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad
blagojevicbogdan1@gmail.com

***Apstrakt*—Kako zdravstvo poseduje ograničene resurse i kapacitete neophodno je napraviti model koji će im olakšati testiranja i ubrzati rad. Kreiranjem modela koji može ispravno da izvrši predikciju ranih faza nastanka dijabetesa olakšao bi eventualno lečenje pacijenta kao i prevenciju pogoršanja njegovog stanja. Podsticanjem istraživanja i inovacija bi se smanjila potreba za fizičkim testiranjem što bi dovelo do smanjenja troškova. U ovom radu predstavljen je model za predikciju rezultata testa na dijabetes gde su moguće sledeće vrednosti: negativan, preddijabetično stanje i pozitivan. Kako su klase u osnovnom skupu podataka nebalansirane, problem je rešen pre same obuke modela. Skup podataka je podeljen na obučavajući skup, na kom je vršena obuka modela kao i na test skup na kom su evaluirani modeli. Za potrebe ovog rada su obučeni sledeći modeli: random forest, logistička regresija, k-najbližih suseda (knn), xgboost klasifikator kao i klasterovanje *kmeans* metodom kako bi se identifikovali potencionalni obrazci unutar podataka. Pokušana je i redukcija dimenzionalnosti koja nije donela nikakve benefite, dok se rezultati metoda pre kao i posle redukcije dimenzionalnosti korišćenjem PCA značajno ne razlikuju. Rezultati koje je ostvario algoritam logističke regresije su za nijanu bolji od ostalih gore pomenutih algoritama mašinskog učenja. Takođe, rezultati se značajno ne razlikuju od rezultata ostvarenih u radu [1]. Na samom kraju vršeno je I tumačenje modela LIME I SHAP tehnikama.**

Ključne reči—dijabetes; nebalansiran skup podataka; redukcija dimenzionalnosti; klasterovanje;

I. UVOD

Dijabetes, hronični metabolički poremećaj koji karakteriše povišen nivo šećera u krvi, predstavlja značajan i rastući globalni zdravstveni problem [8]. Kako prevalencija dijabetesa nastavlja da raste, efikasno i tačno predviđanje medicinskih rezultata postaje sve važnije za pravovremenu intervenciju i personalizovanu zdravstvenu negu. Rešavanje ovog izazova je od najveće važnosti u poboljšanju ishoda pacijenata, smanjenju troškova zdravstvene zaštite i poboljšanju ukupnog kvaliteta upravljanja dijabetesom.

Zamršena priroda dijabetesa zahteva napredne analitičke tehnike za efikasno predviđanje medicinskih rezultata [9]. U ovom izveštaju se bavimo istraživanjem prediktivnih metodologija za predikciju dijabetesa. Cilj je da se razviju modeli koji mogu predvideti medicinske ishode, koji mogu biti negativan, preddijabetično stanje i pozitivan. Predviđanje se vrši na osnovu sveobuhvatnog skupa karakteristika, koji obuhvataju demografiju pacijenata, faktore životnog stila i kliničke indikatore.

U ovom radu korišćen je skup podataka koji je javno dostupan na popularnoj platformi *Kaggle* [4]. Sačinjavaju ga podaci o ispitanicima kao i rezultat testa koji je ostvaren. Detaljan opis skupa podataka predstavljen je u poglavnju III.

Naš pristup prvobitno uključuje sprovođenje eksplorativne analize, normalizacije obeležja odnostno vršenje *OneHotEncoding* algoritma. Nakon toga sledi implementacija algoritama mašinskog učenja kao što su *random forest*, logistička regresija, K-najbliže suseda (KNN) i XGBoost klasifikator. Da bi se rešio problem nebalansiranosti klasa u skupu podataka, ugrađena je tehnika prekomernog uzorkovanja sintetičke manjine (SMOTE), koji se pokazao bolji i brži u odnosu na isprobani SMOTE-ENN metod. SMOTE tehnika je obezbedila robusniji i reprezentativniji skup podataka za obuku modela.

Tokom naše analize, procenjujemo prediktivne modele koristeći ključne metrike učinka kao što su tačnost, preciznost, osetljivost i F1 mera [3]. Posebno, naši preliminarni rezultati pokazuju da različite metodologije, čak i sa različitim algoritmima i tehnikama preprocesiranja, daju slične rezultate. Ovo ukazuje na robusnost i svestranost odabranih algoritama u predviđanju medicinskih ishoda povezanih sa dijabetesom. Najbolje rezultate ostvarila je logistička regresija sa sledećim metrikama: tačnost – 0.83, preciznost – 0.43, osetljivost – 0.36 i f1 mera – 0.37. U pogledu ostvarenih rezultata na istom skupu podataka [1] – tačnosti - 0.86, preciznost - 0.49, osetljivost - 0.57 i f1 mera - 0.53 primećuje se da se rezultati značajno ne razlikuju.

U potrazi za poboljšanjem interpretabilnosti modela i smanjenjem dimenzionalnosti, koristi se analiza glavnih komponenti (PCA). Istražuju se dva različita pristupa PCA – jedan koji koristi 95% uzoraka da zadrži većinu informacija, a drugi sa samo tri obeležja koji imaju najznačajniji uticaj na podatke. Zanimljivo je da naši preliminarni rezultati ukazuju na to da uključivanje PCA ne mora uvek dovesti do poboljšanih prediktivnih performansi. Rezultati dobijeni nakon primene PCA pokazali su se za nijansu lošiji nego pre primene samog algoritma za smanjenje dimenzionalnosti.

Pored toga, istražuje se primena grupisanja K-means [7] kao sredstvo za identifikaciju potencijalnih obrazaca unutar podataka koji mogu pomoći u prečišćavanju procesa predviđanja. Utvrđeno je 4 klastera između kojih se vidi jasna separacija kao na slici broj 6. Kao najznaćajnija obeležja na osnovu kojih su identifikovani klasteri su: pol, starost, BMI i opšte zdravlje. Takođe, tumačenje modela vršeno je pomoću tehnika kao što su LIME i SHAP.

Kako budemo napredovali kroz ovaj izveštaj prvenstveno ćemo govoriti o srodnim istraživanjima te koji su njihovi zaključci i kako su doprineli našem radu. Nakon toga biće reči o skupu podataka – odakle je prikupljen, koji je odnos broja uzoraka i obeležja kao i opisu obeležja, detalji o ciljnom obeležj i eksplorativnoj analizi kao i kako se vršila podela na trening i test skup. U poglavnju IV. stavljen je akcenat na korišćenim algoritnima mašinsgo učenja koji su prethodno pomenuti u ovom poglavlju. Odgovore na pitanja koja su najbitnija obeležja, šta su najbitnije PCA komponente, zašto se klase 0 i 1 mešaju i slično biće obrazloženi u V. poglavlju. Odnosno, biće jasno predstavljena interpretacija dobijenih rezultata. U pretposlednjem poglavlju osvrnućemo se na kompletan rad i o najbitnijim detaljima. Na samom kraju biće predstavnjena relevatna literature korišćena u ovom radu.

II. Srodna istraživanja

U radu [1] predstavljeno je rešenje u predikciji rezultata medicinskog testa na dijabetes gde je korišćen isti skup podataka kao i u ovom radu. Cilj rada je postavljen na poboljšanje tačnosti prilikom predikcije dijabetesa preko uzorkovanja podataka. Ova studija istražuje efektivnost različitih strategija za balansiranje skupa podataka čime se poboljšava senzitivnost i tačnost modela za najmanje 52% i 15%. Poboljšanja su konstantna za sva četiri testirana skupa podataka(2015, 2017, 2019 i 2021. godine). Kako bi se rešio problem nebalansiranosti koristi se kombinacija nasumičnog *undersampling-a i oversampling-a,* čime se smanjuje pristrasnost i povećava tačnost modela. Korišćen skup podataka sastoji se od 4 disjunktna skupa podataka iz sledećih godina: 2015, 2017, 2019, 2021, nad kojima se zasebno primenjuju metode mašinskog učenja. Prisutan je problem nebalansiranosti jer sadrži značajno manje uzoraka manjinske klase – osobe sa dijabetesom, naspram većinske klase – osobe bez dijabetesa. Kako bi se rešio problem nebalansiranosti korišćeni su tehnike *SMOTE i ADASYN,* kao i sledeći algoritmi mašinskog učenja: *The Mulitilayer Perceptron, Random Forest i Logistic Regression.* Ostvareni rezultati pomenuti su u poglavlju V.

Chowdhury, Ayon, Hossain u radu [2], stavljaju akcenat na augmentaciju podataka u cilju poboljšanja efikasnosti algoritama mašinskog učenja. Autori ovog rada pokazali važnost za primenom tehnika za povećanje podataka u slučaju nebalansiranog skupa podataka. Korišćeni skup podataka podeljen je na trening i test skup u odnosu 80:20 koji se sastoji od sledećih obeležja: Indeks telesne težine, Starost, Prihodi, Pušenje, Krvni pritisak, Holesterol, Srčano oboljenje, Asma, Oboljenje bubrega, Bračni status, Edukacija, Opšte zdravlje, Vežbanje, Artritis, Depresija, Konzumacija hrane i voća, Pol i Dijabetes kao zavisna promenljiva. Korišćeni algoritmi su *Gradient Boosting*, Logistička Regresija, *Random Forest i AdaBoost.* Kao najbolja metoda za uzorkovanje podataka pokazala se ENN metoda, dok je model sa najboljim rezultatima u kombinaciji sa prethodno navedenom metodom *Gradient Boosting* sa preciznosti od 0.635, osetljivošću 0.717 i tačnošću 0.703.

U radu [3], Mujumdar, Vaidehi unapredili su postojeći skupa podataka, dodavanjem novih obeležja koja su relevantna u predikciji rezultata medicinskog testa na dijabetes, dobavljenih iz drugih podataka koji su javno dostupni. Za razliku od skupa podataka u ovom radu, Mujumdar i Vaidehi su rešavali problem binarne klasifikacije. Skup podataka koji su koristili sastoji se od 800 uzoraka i sledećih obeležja: broj trudnoća, nivo glukoze, krvni pritisak, debljina kože izražena u milimetrima, insulin, index telesne težine, starost i tip posla (kancelarijski posao, terenski posao, rad u fabrici). Korišćeni algoritmi su Logistička Regresija, *Gradient Boos Classifier, LDA, AdaBoost Classifier, Extra Trees Classifier, Gaussian NB, Bagging, Random Forest, KNN, Decision Tree, Perceptron, SVC.* Razvijeni model se pokazao kao vrlo efikasan u otkrivanju dijabetesa kod pacijenata, kao i u detekciji mogućnosti nastanka dijabetesa u narednih nekoliko godina. Kao najbolji model se pokazala Logistička Regresija. Rezultati svakog pojedinačnog modela su evaluirani korišćenjem klasifikacionih metrika kao što su f1 mera, preciznost i tačnost. Najbolja se pokazala Logistička Regresija koja ima tačnost od 96%, dok ostali navedeni algoritmi za tačnost imaju između 86 i 93 procenta.

III. Opis skupa podataka

Javno dostupan skup podataka[1] [1] preuzet je sa popularne platforme *Kaggle* [3]. Skup podataka kreiran je uz pomoć sistema za nadzor bihevioralnih faktora rizika, koji pomoću telefonskih poziva prikuplja podatke vezane za zdravstvene rizike, hronična stanja i upotrebu preventivnih sredstava za odrasle osobe starije od 18 godina. Ovaj sistem pruža uvid u zdravstvena stanja odraslih osoba u Sjedinjenim Američkim Državama jos od 1984. godine. Dok je naš dobijen prikupljanjem informacija iz 2021. godine.

Skup podataka sastoji se od 236378 jedinstvenih uzoraka i 21 obeležja. Ciiljno obeležje (*Diabetes*) nad kojim se vrši klasifikacija ima moguće vrednosti 0 za osobe koje nemaju dijabetes ili žene koje su pozitivne u stanju trudnoće, 1 za preddijabetično stanje i 2 za osobe pozitivne na dijabetes.

Atributi uz pomoć kojih se vrši predikcija rezultata medicinskog testa na dijabetes su sledeći:

- visok krvni pritisak (*high blood pressure*),
- visok holesterol (*high cholesterol*),
- provera holesterola (*cholesterol check*),

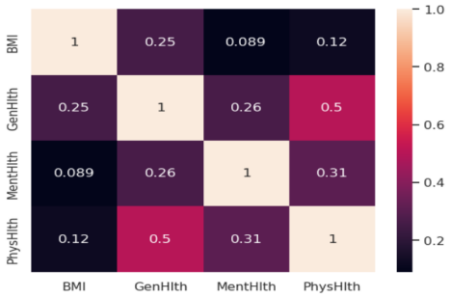
[1] https://www.kaggle.com/datasets/julnazz/diabetes-health-indicators-dataset/

- indeks telesne težine (*body mass index*),
- pušenje (*smoking*),
- udar (*stroke*),
- srčano oboljenje ili srčani napad (*heart disease or heart attack*),
- fizička aktivnost (*physical activity*),
- konzumacija voća (*fruit consumption*),
- konzumacija povrća (*vegetable consumption*),
- visoka konzumacija alkohola (*high alcohol consumption*),
- zdravstvena nega (*health care*),
- odlaganje posete doktoru zbog nedostatka novca (*postponing a visit to the doctor due to lack of money*),
- opšte zdravlje (*general health*),
- mentalno zdravlje (*mental health*),
- fizičko zdravlje (*physical health*),
- poteškoće prilikom šetanja ili penjanja uz stepenice (*difficulty walking or climbing stairs*),
- pol (*gender*),
- starost (*age*),
- obrazovanje (*education*),
- prihodi (*income*).

.

Indeks telesne težine, opšte zdravlje, mentalno zdravlje i fizičko zdravlje prestavljaju numerička obeležja dok su sva ostala obeležja kategorička. Za dalje procesiranje bilo je neophodno pretvoriti kategoričke podatke u numeričke. Pretvaranje *one hot encoding* tehnikom vršeno je na svim obeležjima sem *Education, Age, Income i General Health.* Ona su transformisana na sledeći način: obeležje *Education* sa vrednostima *Never attended high school or kindergarden, Elementary, Some high school, High school graduate, Some college or technical school i College graduate* transformisani su redom u numeričke vrednosti na skali od 1 do 6 u zavisnosti od stepena obrazovanja. Slično je vršeno i kod obeležja *Age* i *Income* gde su ispitanici grupisani u neki od intervala gde je razmak između intervala 10 godina odnosno 20 000 dolara. Kod obeležja *General Health* čije vrednosti mogu biti *Poor, Fair, Good, Very good i Excelent* transforminane su redom u numeričke vrednosti od 1 do 5.

Matrica korelacije na slici 1. prikazuje da je najveća korelacija između opšteg i fizičkog zdravlja, dok je korelacija između ostalih obeležja neprimetna.



Slika 1. Matrica korelacije numeričkih obeležja

Eksplorativnom analizom, uz pomoć chi-kvadrat testa, uočeno je da su obeležja koja imaju najmanji uticaj na ciljno obeležje:

- pol,
- odlaganje posete doktoru zbog nedostatka novca,
- zdravstvena nega,
- konzumacija voća,
- konzumaija povrća.

te su ona odbaćena iz skupa podataka nakon utvđivanja da metode daju za nijansu bolje rezultate bez gore pomenutih obeležja.

Normalizacija obeležja vršena je određivanjem procečne vrednosti (*x_avg_train*) i standardne devijacije (*x_std_train*) trening skupa, te se svaka vrednost zamenila sa

$$(x - x\_avg\_train) / x\_std\_train \quad (1)$$

Normalizacija na test skupu vršena je korišćenjem prosečne vrednosti i standardne devijacije određeme nad test skupom na sledeći način

$$(x - x\_avg\_train) / x\_std\_train \quad (2)$$

Korišćenjem nasumično stratifikovane podele dobijen je skup podataka podeljen je na 80% trening, dok je ostatak ostavljen za testiranje model.

Daljom analizom kategoričkih varijabli uz pomoć histograma, box plot-ova i drugih grafikova nije utvrđeno obeležje koje je u jakoj korelaciji sa ciljnim obeležjem te ni jedno obeležje nije dodatno odbačeno.

IV. METODOLOGIJA

Za rešavanje multiklasnog klasifikacionog problema korišćeni su algoritmi kao što su *Random-Forest, K-Nearest Neighbors (KNN), XGBoost* i *Logistic Regression*, koji će biti detaljnije opisani u nastavku ovog poglavlja.

U cilju obučavanja klasifikatora, pronalazka optimalnih hiperparametara i sprečavanja natprilagođenja, korišćena je unakrsna validacija [10] sa 10 fold-ova.

Kako skup podataka ima problem sa nebalansiranom raspodelom klasa gde je klasa 0 prisutna u 83% uzoraka, klasa 1 u 2% i klasa 2 u 15%, primenjena je tehnika *SMOTE (Synthetic Minority Over-sampling Technique)*. Takođe isprobana je varijacija ove metode (*SMOTEENN)* koja je dala identične rezultate ali je vreme izvršavanja bilo znatno duže.

SMOTE je korišćen u kombinaciji sa unakrsnom validacijom, tako što je primenjen na 9 trening fold-ova u svakoj iteraciji. To je podešeno prilikom kreiranja *pipeline*-a.

A. *Logistic Regression*

U cilju obučavanja klasifikatora evaluirane su vrednosti l1 i l2 regularizacije, za regularizacioni parametar C evaluirane su vrednosti od 0.001 do 1000, dok su za parameter solver, koji se odnosi na pronalaženje optimalnih koeficijenata, evaluirane vrednosti newton-cg, lbfgs i liblinear.

penalty	I1	I2	
solver	newton-cg	lbfgs	liblinear
C	np.logspace(-3,7)		

Tabela 1. Logistička regresija hiperparametri

B. *Random Forest*

U cilju obučavanja klasifikatora za parametar koji se odnosi na ukupan broj stabala u šumi razmatrane su vrednosti 25, 50 i 100, a za parametar koji se odnosi na maksimalnu dubinu razmatrane su vrednosti *None*, 8, 10 i 20.

max_depth	none	8	10	20
n_estimators	25	50	100	

Tabela 2. Slučajne šume hiperparametri

C. *K-Nearest Neighbors*

U cilju obučavanja klasifikatora za parametar koji se odnosi na broj komšija razmatrane su vrednosti 1, …, 25, za parametar težine razmatrane su *uniform* i *distance*, a za metrike razmatrane su Euklidska, Minkowski i Čebiševa.

n_neighbors	1…25	
weights	uniform	distance
metric	minkowski	chebyshev

Tabela 3. KNN hiperparametri

D. *XGBoost*

U cilju obučavanja klasifikatora za parametar koji se odnosi na maksimalnu dubinu razmatrane su vrednosti 3, 4, i 5, za parametar *learning rate* razmatrane su vrednosti, 0.1, 0.01 i 0.001, a za broj estimatora 50, 100 i 150.

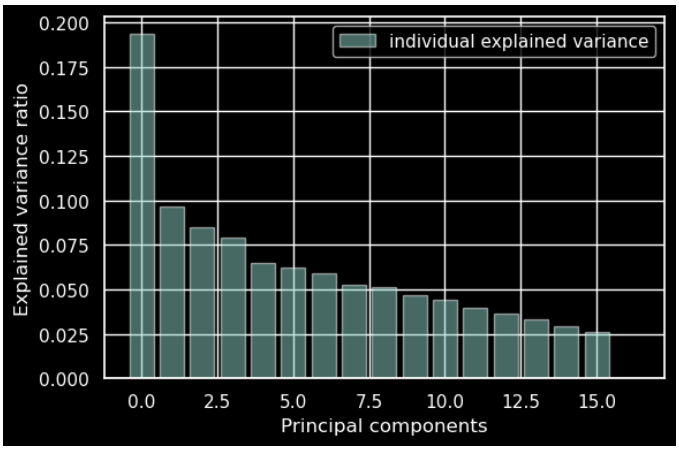
max_depth	3	4	5
learning_rate	0.1	0.01	0.001
n_estimators	50	100	150

Tabela 4. XGBoost hiperparametri

E. *Principal Component Analysis*

PCA je korišćen u cilju redukcije dimenzionalnosti, zbog većeg broja obeležja koja dovode do dužeg obučavanja modela i potencijalnog natprilagođenja. PCA omogućava redukciju broja obeležja a pritom sačuvava varijansu u skupu podataka. Takođe je korišćena i za vizualizaciju klastera.

Primenom PCA broj obeležja koji očuvava 95% varijanse je 15, što se može videti na sledećoj slici.



Slika 2. Odnos varijanse i broja obeležja

Inicijalno model ima ukupno 22 obeležja. Redukcijom broja obeležja na 15 dobijaju se identični rezultati kao u modelima bez primene PCA. Jedina prednost primene PCA je brzina izvršavanja.

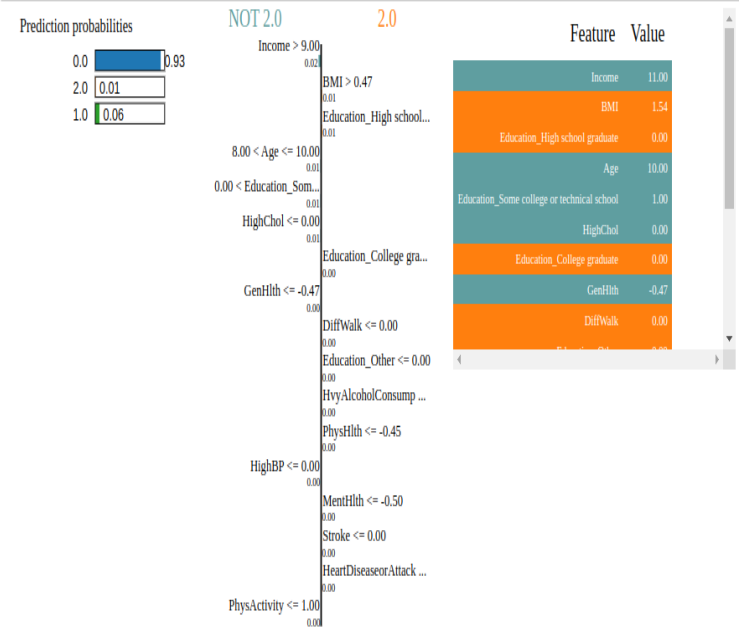
Takođe je isprobana i redukcija na samo 3 obeležja(opšte zdravlje, starost i prihod) koja imaju najveći uticaj na očuvanje varijanse ali ovo je dalo znatno lošije rezultate.

F. *Local Interpretable Model-Agnostic Explanations*

Tumačenje modela je izuzetno važno, posebno kada su u pitanju rezultati medicinskih testova. Pruža uvid u to kako se model može poboljšati, i jasnije prikazuje na osnovu kojih parametara je određenom uzorku dodelio neku klasnu labelu.

LIME je tehnika koja se koristi u mašinskom učenju da objasni predikcije kompleksnih modela. Korisna je za interpretaciju *black-box* modela, odnosno modela koji nisu lako razumljivi za čoveka. LIME ima za cilj da pruži lokalne interpretacije za pojedinačne predikcije od strane datih modela.

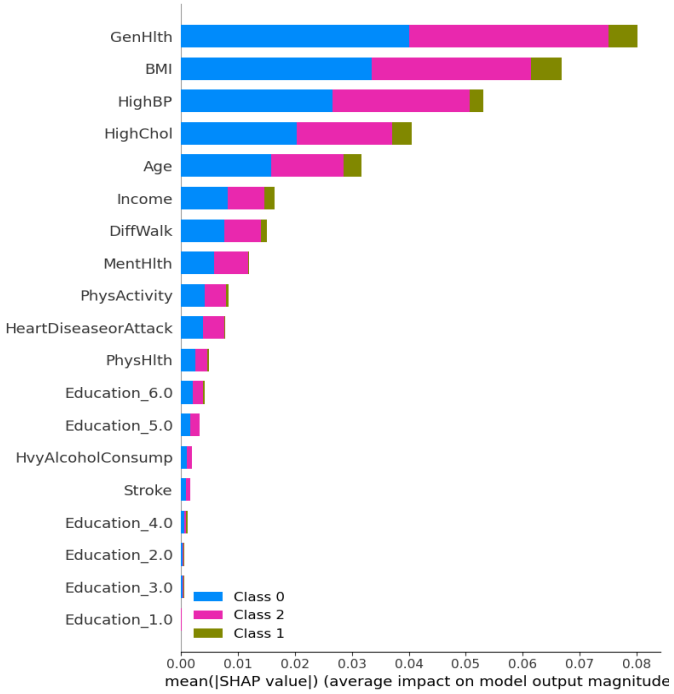
Ovaj algoritam funkcioniše tako što za željenu instancu čiju predikciju želimo da objasnimo, LIME generiše veštačke instance ali se trudi da sačuva većinu originalnih obeležja. Zatim koristi neki jednostavan model koji je najčešće linearna regresija u pitanju i predviđa izlaze tih veštački kreiranih instanci. Instancama dodeljuje težine na osnovu njihove distance od originalne instance. Nakon toga dobijaju se koeficijenti uz svako obeležje, koji predstavljaju koliko je dato obeležje uticalo na finalni rezultat modela.



Slika 3. LIME

Na slici 3. dat je prikaz interpretacije odluke modela Logističke regresije, gde se vide verovatnoće koje označavaju kojoj klasi dati uzorak pripada. Sivom bojom prikazana su obeležja koja su uticala da dati uzorak ne bude klasifikovan kao klasa 2, dok su narandžastom bojom prikazana obeležja koja su uticala da dati uzorak bude klasifikovan kao klasa 2.

Obeležja koja najviše utiču na odluku modela su prihod, BMI, edukacija i broj godina, dok obeležja kao što su fizička aktivnost, srčano oboljenje, srčani udar i mentalno zdravlje, imaju najmanji uticaj na odluku modela

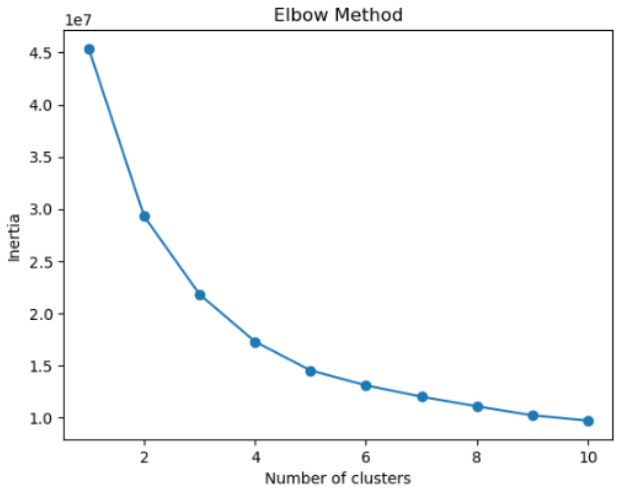


Slika 4. Interpretacija odluka modela korišćenjem tehnike SHAP

Na slici 4. dat je prikaz svih obeležja koja su učestvovala u obuci modela, plavom bojom je označena klasa 0, rozom bojom klasa 2 i zelenom bojom klasa 1. Obeležja koja su najviše uticala na odluku modela su opšte zdravlje, BMI, visok krvni pritisak i visok holesterol. Sva obeležja su normalizovana i na X osi su prikazane granice na osnovu kojih je model razvrstavao uzorke u date klase. Za konkretan primer uzet je model *Random Forest* [6].

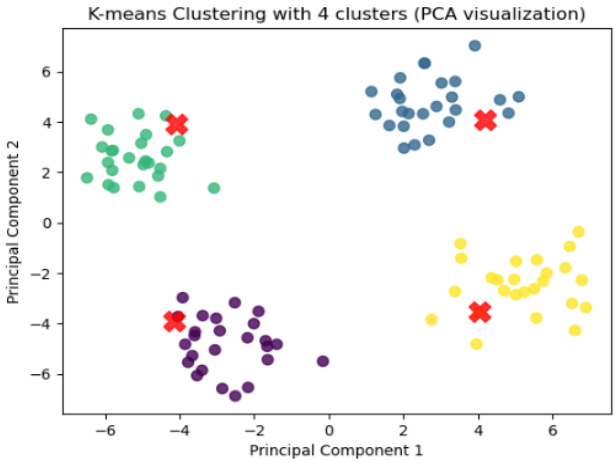
G. *K-means*

K-means je nenadgledani algoritam mašinskog učenja i koristi se za podelu skupa podataka u klastera. Ova podela pomaže u identifikovanju sličnosti između uzoraka koji se nalaze u istom klasteru kao i uočavanju nekih trendova. Problem je odrediti optimalan broj klastera, u ovom radu je to određeno korišćenjem lakat metode.

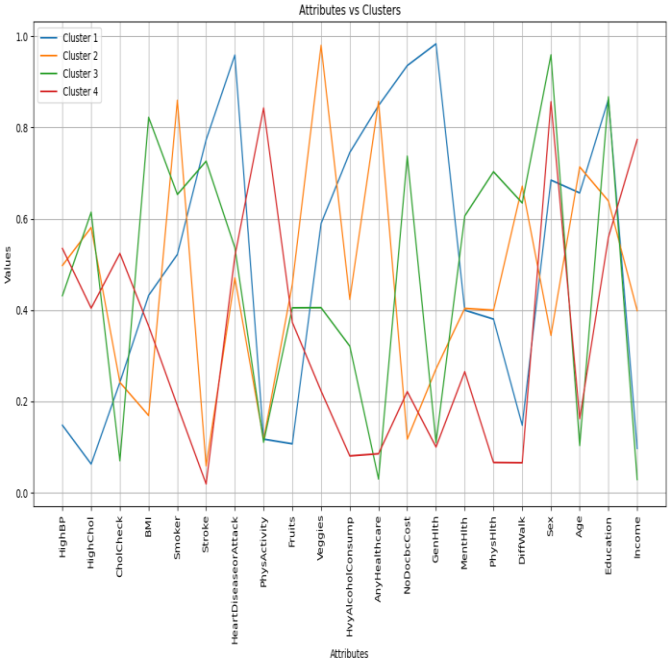


Slika 5. Lakat metoda

Na slici 5. je prikazana vizualizacija lakat metode, gde se vidi da je optimalan broj klastera 4.



Slika 6. Raspored uzoraka po klasterima



Slika 7. Analiza odluka pojedinačnik klastera

U klasteru 1 vidimo da se nalaze ispitanici koji nemaju visok holesterol, imaju dobro opšte zdravlje i nemaju poteškoća prilikom hodanja. Za klaster 2 vidimo da se tu nalaze ispitanici sa malim vrednostima BMI, da su pušači i konzumiraju povrće. Za klaster 3 vidimo da se nalaze ispitanici koji su uglavnom muškarci, ne proveravaju često holesterol i imaju više obrazovanje. U klasteru 4 se nalaze ispitanici koji nemaju rizik od srčanih oboljenja, ne konzumiraju alkohol i imaju manji broj godina.

V. REZULTATI

Zbog nezadovoljavajućih rezultata na evaluacionim metrikama ni jedan model nije moguće koristiti u praksi. Model ne raspoznaje klasu 0 i 1, zbog toga što samo 2% od ukupnog broja uzoraka pripada klasi 1, a 85% klasi 0 pa se model prilagodio većinskoj klasi.

Prilikom poređenja sa drugim radovima zaključujemo da je problem u samom skupu podataka, gde nije pronađena neka značajna korelacija između obeležja. Rezultati iz radova [2] i [5] su slični kao i u ovom radu, što ukazuje na to da se nalazi mali broj podataka za pacijente koji boluju od dijabetesa ili su u preddijabetičnom stanju, kako bi model mogao da uoči obrasce.

Kao najbitnija obeležja pokazala su se BMI, prihod, starost i edukacija, što je i utvrđeno eksplorativnom analizom podataka.

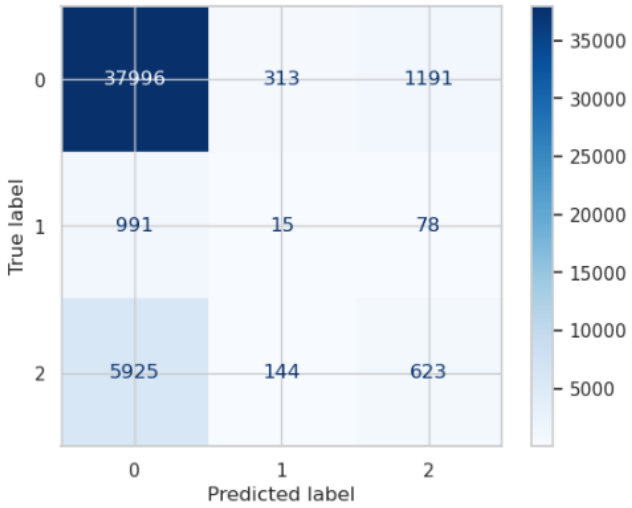
Najbitnije PCA komponente ne uključuju obeležja pol, konzumacija voća, konzumacija povrća, odlaganje posete zbog nedostatka novca i zdravstvena nega.

	accuracy	precision	recall	f1-score
Random Forest	0.816672	0.400819	0.367816	0.369759
Logistic Regression	0.834461	0.439854	0.368698	0.370449
KNN	0.817201	0.402357	0.356286	0.354895
XGBoost	0.836556	0.451554	0.352323	0.343004

Slika 8. Rezultati

Na slici 8. su prikazani rezultati klasifikatora na test skupu. Vidimo da su svi klasifikatori dali slične rezultate, gde logistička regresija za nijansu daje bolje rezultate od ostalih za metriku f1-mera i *recall,* koji su najrelevantniji.

Rezultati za *accuracy* su znatno bolji jer se algoritam prilagodio većinskoj klasi, pa samim tim ova metrika nam ne govori puno o performansama naših modela. Svi modeli su imali problem da razlikuju klasu 0 i klasu 1, što označava problem u samom skupu podataka.



Slika 9. Matrica konfuzije za model Logističke regresije

Na slici 9. prikazana je matrica konfuzije gde vidimo da model klasu 1 vidi kao klasu 0. Za konkretan primer uzet je model logističke regresije, a ostali modeli daju gotovo identičnu matricu konfuzije.

	accuracy	precision	recall	f1-score
Random Forest	0.826804	0.430408	0.383153	0.389835
Logistic Regression	0.83975	0.461637	0.378255	0.384326
KNN	0.821051	0.425258	0.36608	0.369162
XGBoost	0.840553	0.470663	0.369598	0.371703

Slika 10. Rezultati PCA

Na slici 10. prikazani su rezultati nakon redukcije dimenzionalnosti korišćenjem PCA, gde je očuvano 95% varijanse korišćenjem 15 obeležja. Rezultati su gotovo

identični, odnosno ne postoji neko poboljšanje. Takođe svi modeli imaju skoro iste performanse.

VI. Zaključak

U ovom radu je vršena predikcija dijabetesa na osnovu podataka o pacijentima. Kako je dijabetes oboljenje koje se sve češće javlja kod ljudi, pravljenje modela koji bi uspešno mogao da detektuje ovu bolest u ranom stadijumu bio bi od velike koristi. Za predikciju su korišćena četiri algoritma mašinskog učenja, a to su logistička regresija, KNN, slučajne šume i *XGBoost.*

Pre primene samih algoritama bilo je potrebno obraditi podatke, jer je to neophodan korak za pravljenje dobrih modela. Skup podataka nije imao nedostajućih vrednosti, a za transformaciju kategoričkih promenljivih korišćen je *one-hot-encoding.* U skupu se javlja nebalansiranost klasa gde dominira klasa 0, koja predstavlja ljude koji nemaju dijabetes. Za rešavanje problema nebalansiranosti korišćena je tehnika *SMOTE.* Pronalaženje optimalnih hiperparametara vršeno je koristeći kros-validaciju sa 10 foldova. Za evaluaciju modela korišćene su evaluacione metrike, a detaljniji prikaz rezultata modela je predstavljen matricom konfuzije. Tumačenje ponašanja modela je izvršeno korišćenjem LIME i SHAP tehnike, gde se dobio jasniji prikaz koja to obeležja utiču na date klasifikacije.

Svi klasifikatori su dali slične rezultate što ukazuje na to da sam skup podataka ima veliki uticaj. Nisu postignuti željeni rezultati koji bi sa većom preciznosti mogli da odrede da li osoba boluje od dijabetese ili da li je u pred-dijabetičnom stanju. Takođe u radu nije uočena neka pravilnost koja je karakteristična za određenu klasu.

Predviđanje dijabetesa je složen zadatak i uvek postoji mesta za napredak. Moguće je dalje nastaviti sa razvojem projekta, prikupljanjem još podataka o pacijentima koji boluju od dijabetesa ili su u pred-dijabetičnom stanju. Dodavanjem još nekih obeležja koja bi mogla da jasnije diferenciraju date klase bi bila od velikog značaja. Konsultovanje sa domenskim stručnjakom koji bi mogao da pruži bolji uvid u same podatke, kao i analiziranje grešaka modela bi mogli da doprinesu daljem razvoju projekta

Literatura

[1] Tanmoy Sarkar Pias, Yiqi Su, Xuxin Tang, Haohui Wang, Shahriar Faghani, Danfeng (Daphne) Yao, Virginia Tech, USA(2023.) *Enhancing Fairness and Accuracy in Type 2 Diabetes Prediction through Data Resampling*

[2] Mohammad Mihrab Chowdhury, Ragib Shahariar Ayon, Md Sakhawat Hossain (2023.) *Diabetes Diagnosis through Machine Learning: Investigating Algorithms and Data Augmentation for Class Imbalanced BRFSS Dataset*

[3] Aishwarya Mujumdar, Dr. Vaidehi V.(2019.) Diabetes Prediction using Machine Learning Algorithms

[4] *Data Science* platforma "*Kaggle*" - https://www.kaggle.com/

[5] Abigail R Cartus, Lisa M Bodnar, and Ashley I Naimi. The impact of undersampling on the predictive performance of logistic regression and machine learning algorithms: a simulation study. Epidemiology (Cambridge, Mass.), 31(5):e42, 2020.

[6] Mani Butwall and Shraddha Kumar," A Data Mining Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier", International Journal of Computer Applications, Volume 120 - Number 8,2015

[7] Dost Muhammad Khan1, Nawaz Mohamudally2, "An Integration of K-means and Decision Tree (ID3) towards a more Efficient Data Mining Algorithm ", Journal Of Computing, Volume 3, Issue 12, December 2011

[8] Ullah, F. Saleem, M. Jamjoom, B. Fakieh, F. Kateb, A. M. Ali, B. Shah, et al. Detecting high-risk factors and early diagnosis of diabetes using machine learning methods. Computational Intelligence and Neuroscience, 2022, 2022

[9] Diabetic mellitus prediction with BRFSS datasets, Marwa Husein Mohamed, Mohamed Helmy Khafagy, Nesma Mohamed Mahmoud Kamel and Wael Said, University Cairo Egypt

[10] Diabetes type 2 classifcation using machine learning algorithms with up-sampling technique, Mariwan Ahmed Hama Saeed