Predviđanje metaboličkog sindroma

Marina Bijelović, BI12/2020, marinabijelovic@gmail.com

Višnja Stojšin, BI14/2020, visnja.stojsin@gmail.com

I. Uvod

Metabolički sindrom predstavlja skup faktora rizika koji povećavaju verovatnoću za razvoj srčanih bolesti, dijabetesa tipa 2 i drugih zdravstvenih problema. Ova kompleksna bolest obuhvata faktore poput gojaznosti, visokog krvnog pritiska, visokog nivoa šećera u krvi, povećanja nivoa lošeg holesterola i insulinske rezistencije. Analiza podataka i istraživanja o metaboličkom sindromu imaju ključnu ulogu u razumevanju njenih mehanizama i razvoju efikasnih terapija. Kreiranje modela klasifikacije može pomoći u predviđanju rizika kod pacijenata, što je od vitalnog značaja za prevenciju i pravovremeno lečenje ovog sindroma.

II. Analiza baze podataka

U pitanju je klasifikacioni problem kog kojeg se analizom podataka predviđa da li pacijent ima ili nema metabolički sindrom. Predviđanje se vrši na osnovu demografskih, kliničkih i laboratorijskih merenja. Baza podataka sadrži 2041 uzorak, od kojih svaki predstavlja jednog pacijenta, 14 obeležja, koja opisuju uzorke, i jednu klasnu labelu – metabolički sindrom. Procenat ispitanika koji boluju od ovog sindroma je 34.24%, dok je 65.76% ispitanika negativno. U bazi postoji 11 numeričkih obeležja: sekvencijalni identifikacioni broj, starost, prihod, obim struka, indeks telesne mase, odnos albumina i kreatinina u mokraći , nivo mokraćne kiseline u krvi, nivo glukoze u krvi, HDL – nivo lošeg holesterola u krvi i nivo triglicerida u krvi. U kategorička obeležja se ubrajaju pol, rasa i bračni status pacijenta.

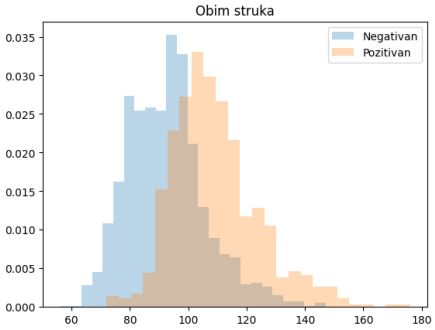
III. Obrada baze podataka

Prilikom analize baze, utvrdile smo da je sekvencijalni identifikacioni broj jedino obeležje koje je nerelevantno, s obzirom na to da ima jedinstvenu vrednost za svaki uzorak i stoga neupotrebljivo za predviđanje klase. Zbog navedenog, u nastavku nećemo razmatrati ovo obeležje i izbacujemo ga iz baze.

Nakon provere postojanja nedostajućih vrednosti, utvrdile smo da su ona prisutna kod obeležja: prihod (4.87%), bračni status (8.66%), obim struka (3.54%) i indeks telesne mase (1.08%). Nedostajuće vrednosti za prihod smo popunile medijanom kako bismo očuvale raspodelu uzoraka koja je asimetrična. U okviru obeležja obim struka i indeks telesne mase smo nedostajuće vrednosti popunile medijanom zbog prisustva autlajera, zasebno za svaki pol. Analizom obeležja bračni status, došle smo do zaključka koje kategorije bračnog statusa su najzastupljenije u određenim starosnim grupama, prema polu. Na osnovu toga smo popunile preostale nedostajuće vrednosti.

Obeležja koja sadrže autlajere su: obim struka, indeks telesne mase, odnos albumina i kreatinina u mokraći , nivo mokraćne kiseline u krvi, nivo glukoze u krvi, HDL, nivo triglicerida u krvi. Autlajeri koji su prisutni ne predstavljaju nevalidne vrednosti. Oni su karakteristični za pacijente koji su jako lošeg zdravstvenog stanja, među kojima su i oni sa metaboličkim sindromom.

Analizom histograma zaključeno je da nijedno obeležje nije dovoljno diskriminatorno kako bi podelilo uzorke na dve klase, što se i očekivalo. Od svih obeležja, na osnovu obika struka, vrši se najbolje razdvajanje klasa (Sl. 1.).



Sl. 1. Prikaz diskriminantnog svojstva obeležja obim struka pomoću histograma.

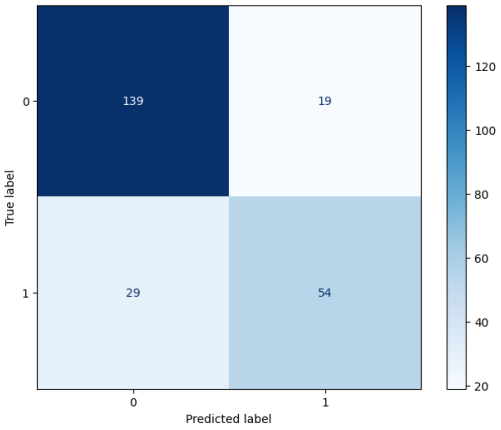
IV. Klasifikacija

Klasifikacija je problem u nadgledanom učenju gde, za dat skup uzoraka, izlazne promenljive zapravo predstavljaju oznake određenih kategorija. Cilj obuke je kreiranje modela klasifikacije koji može da predvidi klasnu pripadnost za uzorke čije oznake nisu poznate. Pre nego što se klasifikator formira, neophodno je da uzorci u bazi imaju odgovarajuće oznake, što je već urađeno u ovom slučaju. Za skraćivanje vremena obuke, izdvojen je trening skup koji sadrži po 10% uzoraka iz svake klase. Takođe, radi ubrzanja procesa treniranja modela, obeležja su standardizovana, što znači da su normalizovana tako da imaju srednju vrednost 0 i standardnu devijaciju 1. Nakon što su odabrani optimalni parametri, model je treniran na celom trening skupu i testiran na testnom skupu.

A. *Mere uspešnosti klasifikatora*

Za evaluaciju performansi klasifikatora koristili smo matricu konfuzije, koja omogućava poređenje stvarnih i predviđenih labela. Matrica konfuzije jasno pokazuje kako klasifikator kombinuje različite klase, pri čemu ispravno klasifikovani uzorci zauzimaju glavnu dijagonalu. Na osnovu matrice konfuzije, izračunavamo različite mere uspešnosti klasifikatora, uključujući tačnost, preciznost, osetljivost, specifičnost i F-meru. Ove mere se računaju na osnovu broja TP (true positive - tačno pozitivnih), TN (true negative - tačno negativnih), FP (false positive - lažno pozitivnih) i FN (false negative - lažno negativnih) slučajeva, kako je prikazano u tabeli.

Nakon primene algoritma k-najbližih suseda izvršena je klasifikacija uzoraka. Tačno pozitivni (TP) uzorci su oni koji stvarno pripadaju klasi 1, već klasi 0 (imaju metabolički sindrom - bolesni) i pravilno su klasifikovani u tu klasu. Lažno pozitivni (FP) uzorci su oni koji ne pripadaju klasi 1 (nemaju metabolički sindrom - zdravi), ali su greškom klasifikovani kao da joj pripadaju. Lažno negativni (FN) uzorci su oni koji stvarno pripadaju klasi 1, ali su greškom klasifikovani u klasu 0. Tačno negativni (TN) uzorci su svi oni koji ne pripadaju klasi 1 i pravilno nisu klasifikovani u klasu 1. Primer matrice konfuzije koja je dobijena nakon izvršene KNN klasifikacije prikazan je na slici Sl. 2. U tabeli gore levo nalazi se TP, gore desno FP, dole levo FN, dole desno TN.



Sl. 2. Matrica konfuzije nakon izvršene KNN klasifikacije sa prikazanim vrednostima za TP, FP, FN i TN.

Pri evaluaciji, za glavnu meru uspešnosti izabrana je senzitivnost. Visoka senzitivnost znači da je klasifikator sposoban da među svim osobama koje zaista imaju bolest, većinu identifikuje kao stvarne pozitive, iz razloga što želimo da minimizujemo broj propuštenih slučajeva bolesti kako bi bolesni ispitanici dobili pravovremeno lečenje.

B. *Klasifikator zasnovan na k - najbližih suseda (KNN)*

K najbližih suseda (KNN) je jednostavan algoritam nadgledanog učenja koji se koristi za klasifikaciju i regresiju. Osnovna ideja ovog algoritma je da novi uzorak klasifikujemo na osnovu klase najčešćih K suseda iz trening skupa. Za svaki novi uzorak, KNN traži K najbližih suseda iz trening skupa i dodeljuje mu klasu koja je najzastupljenija među tim susedima. Jedna od važnih karakteristika KNN algoritma je parametar K, koji predstavlja broj najbližih suseda koji se uzimaju u obzir prilikom klasifikacije. Veće vrednosti K mogu dovesti do glađih granica odluke, dok manje vrednosti K mogu dovesti do preobučavanja. Važno je odabrati odgovarajuću vrednost K kako bi se postigao balans između tačnosti modela i sposobnosti generalizacije.

Prilikom ispitivanja najboljih parametara ovog klasifikatora, u obzir su uzete sledeće vrednosti*: 1, 3, 5, 6, 7, 8, 9, 11, 13* i *15* za broj suseda, odnosno parametar K, *distance* i *uniform* za težine kojima se ponderišu susedni uzorci i *euclidean*, *manhattan* i *chebyshev* za metriku. Evaluacijom modela nad izdvojenim skupom, utvrđeno je da je senzitivnost najbolja i ima vrednost 65.06% kada se za klasifikaciju koristi 6 najbližih suseda, euklidska metrika kao funkcija rastojanja, a za parametar weights uzme vrednost distance što znači da je značaj susednih uzoraka obrnuto proporcionalan njihovoj udaljenosti od neobeleženog uzorka. Matrica konfuzije je prikazana na slici Sl. 2, dok su prosečne mere uspešnosti klasifikatora prikazane u tabeli (Tabela 1), obe dobijene klasifikacijom sa konačno odabranim parametrima na celokupnom trening skupu i testiranjem na skupu za test.

Tabela 1: Prikaz mera uspešnosti sa konačno odabranim parametrima za KNN klasifikator.

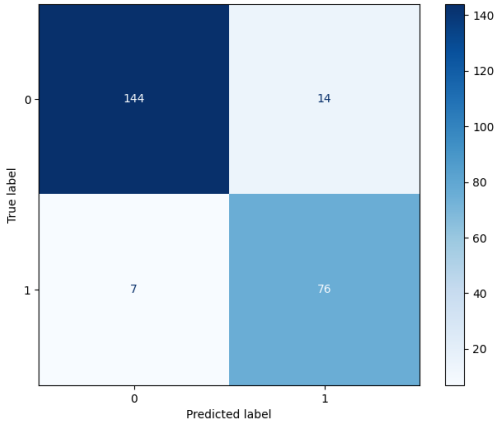
Mera uspešnosti	Vrednost
Preciznost	73.97%
Tačnost	80.08%
Senzitivnost	65.06%
F-mera	69.23%

C. *Klasifikator zasnovan na stablima odluke (Decision Tree)*

Stabla odluke su metod ranog nadgledanog učenja koji se može koristiti za klasifikaciju i regresiju, što znači da podržava rad sa numeričkim i kategoričkim obeležjima. Koren stabla, od kojeg se stablo grana, sadrži sve uzorke. Proces obuke obuhvata odabir pitanja koja na najbolji način dele uzorke. Svako pitanje deli uzorke na dva podskupa (čvora), zavisno od odgovora. Čistoća čvora se koristi kako bi se opisala uspešnost podele uzoraka i definiše se kao dominacija jedne od klasa. Kriterijumi za zaustavljanje grananja mogu biti potpuno čist čvor, ograničenje broja uzoraka u čvoru ili dubina stabla. Klasifikacija se vrši tako što novi uzorak prolazi kroz stablo sa početkom u korenu stabla a zatim se na osnovu pitanja određuje kom će čvor i na kraju klasi pripasti [1].

Isprobane su vrednosti *Đinijevog indeksa (gini)* i *entropije (entropy)* kao kriterijumi podele, odnosno mere nečistoće. Za maksimalnu dubinu stabla testirane su vrednosti *4, 5, 6, 7, 8, 9, 10*, dok su za broj obeležja koja će stablo razmatrati pri izboru najboljeg pitanja za podelu čvora *None, sqrt, log2*. Prilikom evaluacije modela nad izdvojenim skupom, najbolja senzitivnost u vrednosti od 91.56% dobijena je za entropiju kao kriterijum podele,

maksimalnu dubinu 5 i kada algoritam odlučivanja koristi sva dostupna obeležja pri grananju svakog čvora, odnosno None. Nakon izbora ovih optimalnih parametara, model je obučen na kompletnom skupu za obuku i testiran na skupu za testiranje. Rezultati su prikazani u matrici konfuzije i merama uspešnosti klasifikatora, koje se mogu videti na slici Sl. 3. i u tabeli (Tabela 1).



Sl. 3. Matrica konfuzije nakon izvršene klasifikacije pomoću stabla odluke sa konačno odabranim parametrima.

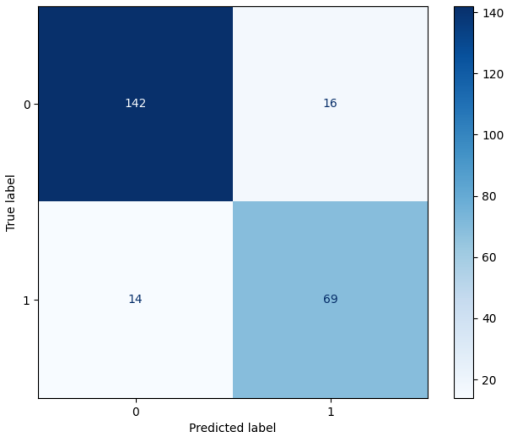
Tabela 2: Prikaz mera uspešnosti sa konačno odabranim parametrima za klasifikator zasnovan na stablima odluke.

Mera uspešnosti	Vrednost
Preciznost	84.44%
Tačnost	91.29%
Senzitivnost	91.57%
F-mera	87.86%

D. Klasifikator„mašina na bazi vektora nosača“ (SVM)

Mašina na bazi vektora nosača (SVM) se zasniva na principu klasifikatora maksimalne margine. Ova metoda se uglavnom primenjuje za rešavanje binarnih klasifikacionih problema, gde se identifikuje hiperpovrš koja treba da razdvoji uzorke u prostoru obeležja tako da se uzorci jedne klase nalaze u jednom delu, a uzorci druge klase u drugom. Uzorci koji su najbliži ovoj hiperravni nazivaju se vektori nosači. Optimalna hipperravan mora da obezbedi najveću moguću marginu, a da bi ovo bilo ostvarivo, klase moraju da budu linearno separabilne.Da bi ovo bilo ostvarivo, klase moraju biti linearno separabilne, a optimalna hiperravan razdvajanja mora obezbediti najveću moguću marginu. To znači da je potrebno maksimizovati rastojanje između hiperravni i najbližeg uzorka iz svake klase. Ova definicija hiperravni omogućava bolju generalizaciju, što znači da se očekuje tačnija klasifikacija novih uzoraka u poređenju sa bilo kojom drugom hiperravni koja bi takođe mogla da razdvoji uzorke iz skupa za obuku[1].

Prilikom ispitavanja parametara, isprobane su vrednosti *poly* i *linear* za kernel, *0.2, 0.5, 2, 5* i *10* za regularizacioni parametar C. Kao poseban slučaj ispitane su i performanse klasifikatora sa parametrima: *rbf* za kernel, vrednostima *0.2, 0.5, 2, 5* i *10* za regularizacioni parametar C i vrednostima *scale* i *auto* za *gamma*. Analizom rezultata utvrđeno je da je za datu bazu najefikasnije koristiti radijalni kernel, a vrednosti parametra C i gamma postavljene na *10* i *scale* respektivno. Matrica konfuzije je prikazana na slici Sl. 4, dok su prosečne mere uspešnosti klasifikatora prikazane u tabeli (Tabela 3), obe dobijene klasifikacijom sa konačno odabranim parametrima na celokupnom trening skupu i testiranjem na skupu za test.



Sl. 4. Matrica konfuzije nakon izvršene klasifikacije pomoću mašine na bazi vektora nosača sa konačno odabranim parametrima.

Tabela 3: Prikaz mera uspešnosti sa konačno odabranim parametrima za klasifikator zasnovan na mašini na bazi vektora nosača sa konačno odabranim parametrima.

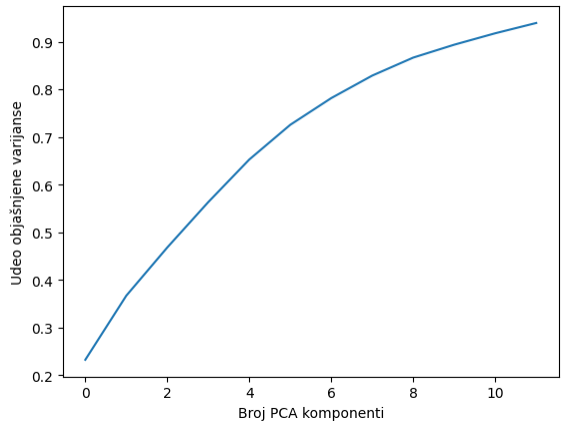
Mera uspešnosti	Vrednost
Preciznost	81.18%
Tačnost	87.55%
Senzitivnost	83.13%
F-mera	82.14%

V. RAZLAGANJE NA GLAVNE KOMPONENTE – PCA

Metoda razlaganja na glavne komponente (PCA) je tehnika za smanjenje dimenzionalnosti koja ima za cilj da uzorke iz visokodimenzionalnog prostora predstavi u prostoru sa manjim brojem dimenzija što vernije. Osnovni cilj PCA je da, polazeći od pretpostavke da su raspoloživa obeležja korelisana i da postoji redundantnost u višedimenzionalnoj reprezentaciji podataka, identifikuje nove, nekorelisane obeležja (PCA komponente) kao linearnu kombinaciju postojećih obeležja, pri čemu se očuvava varijansa podataka. Odabir novih obeležja rezultira formiranjem novog prostora sa manjim brojem dimenzija, u koji se uzorci projektuju. [1].

Sprovedena je nova faza klasifikacije, pri čemu je model obučen na smanjenom skupu za obuku i testiran na smanjenom testnom skupu. Glavni parametar klase je *n_components*, a za vrednost ovog parametra postavili smo 0.92, koja je davala najbolje rezultate. Grafik koji nam pomaže da odredimo optimalan broj PCA komponenti koje treba zadržati kako bismo zadržali što veći deo informacija u podacima, ali izbegli prenaglašavanje i

gubitak generalizacije modela prikazan je na slici Sl.5.



Sl. 5. Grafik zavisnosti objašnjenog dela varijanse od broja PCA komponenata uzetih u obzir.

Rezultati su pokazali da redukcija dimenzionalnosti primenom PCA metode nije unapredila performanse algoritma u poređenju sa performansama postignutim nad originalnim skupom obeležja, osim u slučaju klasifikatora zasnovanom na k – najbližih suseda, gde se senzitivnost krajnjeg modela povećala za 6.9%. Suprotno, klasifikator koji je postigao najbolje rezultate nad originalnim skupom obeležja, baziran na stablu odluke, pokazao je pogoršanje performansi nakon redukcije dimenzionalnosti, bez obzira na broj zadržanih PCA komponenti i to za 14.5%. Poređenje vrednosti senzitivnosti dobijene nakon klasifikacije sa i bez PCA, prikazano je u Tabeli 4.

Tabela 4: poređenje vrednosti senzitivnosti klasifikatora sa i bez primene PCA algoritma.

Klasifikator	Sa PCA	Bez PCA
K-najbližih suseda	72.29%	65.3%
Stablo odluke	77.11%	91.57%
Vektori nosači	80.72%	83.13%

VI. Zaključak

Za binarnu klasifikaciju pacijenata najbolje se pokazao klasifikator na bazi stabla odluke, i to sa sledećim parametrima: algoritam odlučivanja koristi sva dostupna obeležja pri grananju svakog čvora, entropijom kao kriterijumom podele i maksimalnom dubinom grananja 5. Ovaj klasifikator radi sa prosečnom tačnošću od 91.57% Iako je izvršena i redukcija dimenzionalnosti, ona nije izazvala poboljšanje performansi klasifikatora.

VII. Literatura

[1] Tijana Nosek, Branko Brkljač, Danica Despotović, Milan Sečujski i Tatjana Lončar-Turukalo. Praktikum iz mašinskog učenja. Fakultet tehničkih nauka, Univerzitet u Novom Sadu, 2020.