Predikcija nivoa gojaznosti

Sonja Mihajlović, IN33/2020, sonjamihajlovic2001@gmail.com,
Sara Stefanović, IN55/2020, stef.sara01@gmail.com

I. Uvod

Uvod u rešavanje problema gojaznosti putem analize podataka o ishrani i fizičkoj aktivnosti predstavlja ključan korak u suočavanju sa sve većim izazovom povećane gojaznosti na globalnom nivou. Gojaznost, kao globalni zdravstveni problem, zahteva sveobuhvatan pristup, a razumevanje uzroka i faktora koji doprinose ovom zdravstvenom stanju neophodno je za prevenciju i lečenje. Analizom podataka o ishrani i nivou fizičke aktivnosti možemo dublje razumeti veze između ovih faktora i gojaznosti, dok primena algoritama mašinskog učenja za klasifikaciju omogućava identifikaciju rizika. Kreiranje modela klasifikacije pomaže nam da predvidimo rizik od gojaznosti kod pojedinaca i prilagodimo strategije prevencije, što otvara put novim saznanjima i efikasnijem rešavanju ovog globalnog problema.

II. Baza podataka

Baza podataka sadrži podatke o ljudima iz Meksika, Perua i Kolumbije, uzrasta od 14 do 61 godine i karakteristike vezane za ishranu, fizičku aktivnost i druge relevantne faktore za procenu nivoa gojaznosti. Baza sadrži 2111 uzoraka i 17 obeležja.

Obeležja:

- Gender - Pol
- Age - Starost
- Height - Visina
- Weight - Težina
- Family_history_with_overweight - Porodična istorija sa prekomernom težinom (porodična anamneza)
- FAVC - Česta konzumacija visokokalorične hrane
- FCVC - Učestalost konzumiranja povrća
- NCP - Broj obroka
- CAEC - Unos hrane između obroka
- SMOKE - Korišćenje cigareta (pušač)
- CH20 - Dnevni unos vode
- CALC - Konzumiranje alkohola
- SCC - Praćenje potrošnje kalorija
- FAF - Učestalost fizičke aktivnosti
- TUE - Vreme korišćenja tehnoloških uređaja
- MTRANS - Način prevoza
- NObeyesdad - Vrednosti gojaznosti

Baza ima 3 numerička obeležja, to su godine, visina i težina. Ostala obeležja su kategorička. Rešava se klasifikacioni problem sa 7 klasa.

III. Analiza podataka

U bazi podataka postoji obeležje SMOKE, ukazuje na osobe koji su pušači, s obzirom na mali procenat pušača (2,08%), ovo obeležje je izbačeno. Analizom skupa obeležja utvrđeno je da nema nedostajućih i nevalidnih vrednosti. Na osnovu rezultata dobijenih deskriptivnom statistikom numeričkih obeležja mogu se izvesti značajni zaključci:

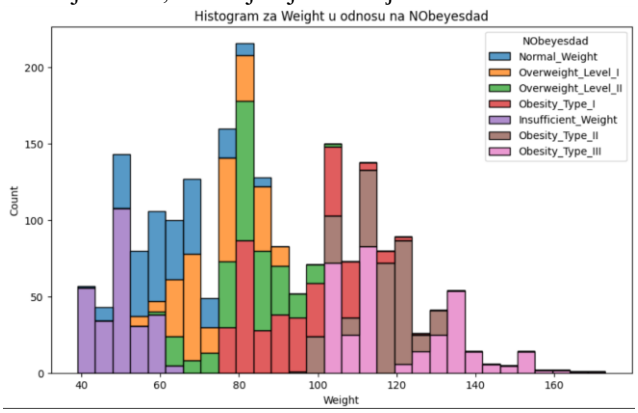
1. Starost (Age): Prosek starosti ispitanika iznosi oko 24 godine, pri čemu se vrednosti kreću od 14 do 61 godine. Standardna devijacija od oko 6,35 ukazuje na relativno veliku varijabilnost u starosnoj strukturi uzorka.
2. Težina (Weight): Prosečna težina je oko 86.6 kilograma. Standardna devijacija ukazuje na raznolikost u težini među ispitanicima.

Prisutan je visok udeo outlier-a među zabeleženim vrednostima obeležja: Starost (Age), Težina (Weight) i Visina (Height). Izbacivanjem ovih podataka model postaje manje sposoban da predvidi adekvatnu klasu, te je bolje ostaviti outlier-e u skupu.

Procentualna raspodela uzoraka po klasama:

Obesity_Type_I	16.63%
Obesity_Type_III	15.35%
Obesity_Type_II	14.07%
Overweight_Level_I	13.74%
Overweight_Level_II	13.74%
Normal_Weight	13.6%
Insufficient_Weight	12.88%

Obeležje koje se izdvaja kao najdiskriminatornije je obeležje težine, ono najbolje razdvaja klase.



Sl. 3.1: Grafički prikaz histograma za težinu u odnosu na klase

Ne postoje parovi obeležja koja su visoko korelisana (korelacija veća od 0.7).

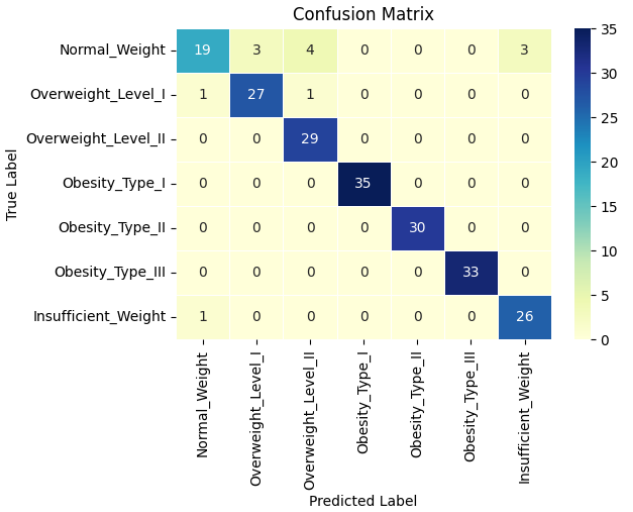
IV. KNN KLASIFIKACIJA

A. *KNN klasifikacija*

Istraživanje započinjemo KNN (*K Nearest Neighbours*) klasifikacijom, gde prvenstveno izdvajamo ciljnu promenljivu *NObeyesdad*. Nad preostalim obeležjima primenjen je *one-hot encoding* kako bi se nadalje koristile *dummy* promenljive. Za potrebe finalnog testiranja modela izdvojeno je 10% podataka, dok je preostalih 90% iskorišćeno za unakrsnu validaciju sa podelom na 5 podskupova. Kao metrika distance za KNN klasifikator korišćena je *Manhattan* distanca, koja je određena *Grid Search-om*, kao i optimalan broj suseda za klasifikator. Pretraga optimalnog broja suseda vršena je nad opsegom [1, 5, 10, 20] i unakrsnom validacijom za svaki od brojeva iz opsega zaključeno je da je optimalan broj suseda 1.

Nakon identifikacije optimalnih hiperparametara, treniramo KNN model nad celim trening skupom podataka. Potom sledi validacija performansi prethodno obučenog modela. Rezultat klasifikacije KNN algoritmom nam daje 13 loše klasifikovana uzorka.

Matricom konfuzije (Slika 4.1) prikazan je broj pogrešno klasifikovanih uzoraka za svaku klasu.



Sl. 4.1: Matrica konfuzije za KNN klasifikator

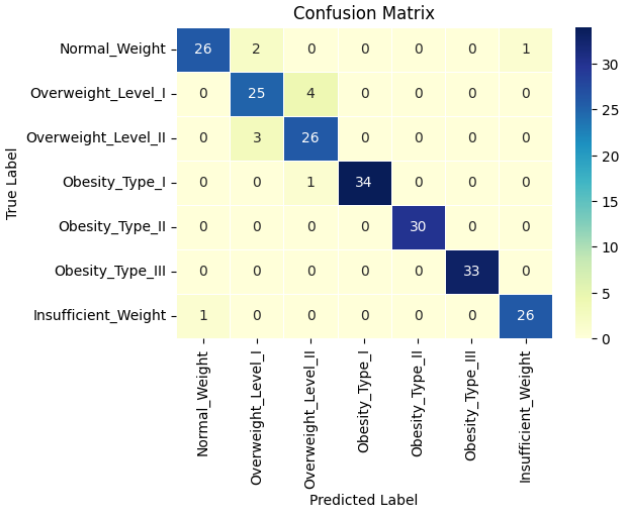
B. *KNN klasifikacija sa smanjenjem dimenzionalnosti*

Podaci koji se nadalje koriste u projektu su standardizovani tako da svako obeležje ima srednju vrednost 0 i standardnu devijaciju 1. PCA (*Principal Component Analysis*) i LDA (*Linear Discriminant Analysis*) algoritmi su primenjeni kako bi se smanjila dimenzionalnost podataka na 2 dimenzije radi lakše vizualizacije. Došle smo do zaključka da LDA algoritam za smanjenje dimenzionalnosti bolje odgovara našim podacima jer ih bolje razdvaja po klasama.

Sl. 4.2: Raspodela uzoraka po klasama, primenom LDA algoritma

Kao metrika distance za KNN klasifikator sa smanjenom dimenzionalnošću (na 3 dimenzije, primenom LDA algoritma) korišćena je *Manhattan* distanca i broj suseda 5 (oba hiperparametra dobijena su primenom *Grid Search-a*). Broj pogrešno klasifikovanih uzoraka je 12.

Matricom konfuzije (Slika 4.3) prikazan je broj pogrešno klasifikovanih uzoraka za svaku klasu.



Sl. 4.3: Matrica konfuzije za KNN klasifikator primenom LDA algoritma za smanjenje dimenzionalnosti.

C. *Mere uspešnosti*

Kao mere uspešnosti klasifikatora korišćene su tačnost, preciznost, osetljivost i F1 skor. Iz tabele se može primetiti da je vrednost svih mera uspešnosti približno jednaka. Ovo je posledica korišćenja makroprosečnih mera, koje su karakteristične za dobro balansirane klase.

Measure	Value for KNN	Value for KNN with LDA
Accuracy	93.9%	94.3%
Precision	93.6%	94.3%
Sensitivity	93.6%	94.1%
F1-score	93.2%	94.2%

Tabela 4.1: Mere uspešnosti za oba algoritma primenom makro prosečnih mera.

Mere uspešnosti za pojedinačne klase date su u tabeli 4.2. Gornju vrednost u ćeliji predstavlja rezultat dobijen KNN klasifikatorom, a donju vrednost KNN klasifikator sa smanjenom dimenzionalnošću.

Class	Precision	Accuracy	Sensitivity	Specificity	F-score
Normal_Weight	90.48 96.30	94.34 98.11	65.52 89.66	98.91 99.45	76.00 92.86
Overweight_Level_I	90.00 83.33	97.64 95.75	93.10 86.21	98.36 97.27	91.53 84.75
Overweight_Level_II	85.29 83.87	97.64 96.23	100.00 89.66	97.27 97.27	92.06 86.67
Obesity_Type_I	100.00 100.00	100.00 99.53	100.00 97.14	100.00 100.00	100.00 98.55
Obesity_Type_II	100.00 100.00	100.00 100.00	100.00 100.00	100.00 100.00	100.00 100.00
Obesity_Type_III	100.00 100.00	100.00 100.00	100.00 100.00	100.00 100.00	100.00 100.00
Insufficient_Weight	89.66 96.30	98.11 99.06	96.30 96.30	98.38 99.46	92.86 96.30

Tabela 4.2: Prikaz mera uspešnosti, izraženih u procentima

V. SVM KLASIFIKATOR

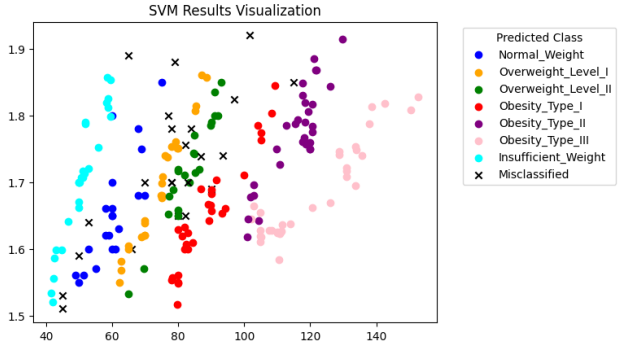
A. *SVM klasifikacija*

U ovom poglavlju predstavljamo analizu rezultata primenom SVM (*Support Vector Machine*) algoritma za problem klasifikacije gojaznosti. Ovaj deo istraživačkog rada fokusira se na evaluaciju performansi modela, odabir optimalnih hiperparametara i interpretaciju rezultata. Za optimizaciju performansi SVM modela *Grid Search-om* testiramo kernel funkcije, uključujući linearnu, polinomijalnu, RBF (Gaussian) i sigmoidalnu funkciju.

Kao rezultat *Grid Search-a* dobijamo linearnu funkciju kernela kao optimalnu. Linearna funkcija za kernel pruža jednostavno linearno razdvajanje između klasa u prostoru atributa. Ova odluka omogućava jasno tumačenje rezultata modela i pruža korisne uvide za dalje istraživanje i kliničku praksu.

Nakon identifikacije najboljeg kernel-a, treniramo SVM model koristeći optimalne hiperparametre nad celim trening skupom podataka. U ovom istraživanju koristimo vrednost C=1 radi kontrolisanja regularizacije modela i izbegavanja preobučenosti. Potom sledi validacija performansi modela koja je ključna za razumevanje njegove sposobnosti klasifikacije.

Rezultat klasifikacije SVM algoritmom nam daje 24 loše klasifikovana uzorka.



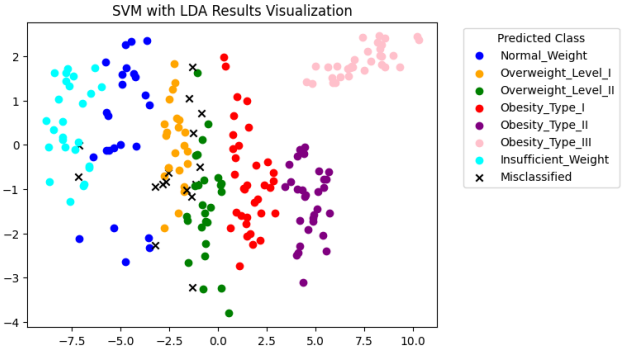
Sl. 5.1: Vizualizacija rezultata SVM klasifikacije

B. *SVM klasifikacija sa smanjenjem dimenzionalnosti*

Ovaj segment istraživanja usredsređen je na analizu rezultata primenom SVM klasifikatora i LDA algoritma za smanjenje dimenzionalnosti. Optimizacija performansi modela je rađena istom metodologijom kao u poglavlju *A. SVM klasifijacija*. Kao rezultat *Grid Search-a* dobijamo RBF (*Radial Basis Functions*) funkciju kernela kao optimalnu.

Nakon identifikacije najboljeg kernel-a, treniramo SVM model koristeći optimalne hiperparametre na celom trening skupu podataka. C parametar je postavljen na vrednost 1 radi kontrolisanja regularizacije modela i izbegavanja preobučenosti. Nakon toga sledi evaluacija performansi modela koja je ključna je za razumevanje njegove sposobnosti klasifikacije.

Rezultati klasifikacije SVM algoritmom uz primenu LDA algoritma pokazuju da je od ukupnog broja uzoraka njih 16 loše klasifikovano.



Sl. 5.2: Vizualizacija rezultata SVM klasifikacije uz primenu LDA algoritma

C. *Mere uspešnosti*

Koristimo mere uspešnosti: tačnost (accuracy), preciznost (precision), osetljivost (sensitivity) i F1-mera (F1-score) kako bi se kvantifikovale performanse SVM modela, bez primene LDA algoritma za smanjenje dimenzionalnosti i sa primenom, na test skupu podataka. Ove metrike omogućavaju nam detaljan uvid u to kako model klasifikuje različite kategorije gojaznosti.

Measure	Value for SVM	Value for SVM with LDA
Accuracy	88.7%	92.5%
Precision	88.5%	92.5%
Sensitivity	88.5%	92.2%
F1-score	88.3%	92.2%

Tabela 5.1: Mere uspešnosti za oba algoritma primenom makro prosečnih mera.

Iz analize tabele 5.1 možemo zaključiti da su vrednosti svih mera uspešnosti približno jednake, što je posledica korišćenja makroprosečnih mera. Makroprosečne mere uzimaju u obzir ravnotežu između preciznosti, osetljivosti i F1-skora za svaku klasu pojedinačno, što može rezultirati sličnim vrednostima mera uspešnosti.

VI. Random Forest Klasifikacija

A. Random Forest klasifikacija

Pristupamo problemu klasifikacije koristeći Random Forest algoritam na skupu podataka koji sadrži kategorička obeležja, za razliku od prethodna dva gde su korišćene *dummy* promenljive. Korišćenje kategoričkih obeležja može doprineti raznolikosti stabala u Random Forest algoritmu, što može uticati na smanjenje varijanse modela, čime se povećava stabilnost i često dovodi do boljih performansi modela. Nakon toga, podatke smo ponovo podelile na trening i test skupove, zbog kategoričkih obeležja. Zatim, primenom *Label Encoding-a* kategoričkim vrednostima dodeljujemo labele, tj. jedinstvene vrednosti svakoj kategoriji.

Nakon pripreme podataka, primenile smo *Grid Search Cross Validation* funkciju kako bismo pronašle optimalne hiperparametre za naš model. Prvo smo istraživale optimalan broj estimatora (broj stabala) u rasponu od 10 do 100. Nakon toga, istraživale smo optimalnu dubinu stabala u rasponu od 1 do 20. Ovi parametri su ključni za performanse Random Forest algoritma. Rezultati *Grid Search-a* pokazali su da je optimalan broj estimatora 99, dok je optimalna dubina stabala 14.

Nakon identifikacije optimalnih hiperparametara, treniramo Random Forest model nad celim trening skupom podataka. Potom sledi validacija performansi modela koja je ključna je za razumevanje njegove sposobnosti klasifikacije.

Rezultat klasifikacije SVM algoritmom nam daje 25 loše klasifikovana uzorka.

B. Random Forest klasifikacija sa smanjenjem dimenzionalnosti

Sprovedena je primena LDA algoritma za smanjenje dimenzionalnosti u Random Forest algoritmu. Prvo, podaci su pripremljeni kroz ponovnu standardizaciju koristeći *StandardScaler* funkciju. Zatim je primenjen Random Forest algoritam nad transformisanim podacima kako bi se izvršila klasifikacija. Kroz postupak optimizacije hiperparametara na prethodno opisan način, pronađen je optimalan broj estimatora 78 i optimalna dubina stabala 14. Model je treniran na trening skupu podataka i evaluiran na testnom skupu koristeći tačnost kao metriku evaluacije.

C. Mere uspešnosti

Koristimo prethodno opisane mere uspešnosti kako bi se kvantifikovale performanse Random Forest modela, bez primene LDA algoritma za smanjenje dimenzionalnosti i sa primenom, na test skupu podataka.

Measure	Value for Random Forest	Value for Random Forest with LDA
Accuracy	88.2%	91.0%
Precision	88.6%	91.1%
Sensitivity	88.2%	90.7%
F1-score	88.1%	90.8%

Tabela 6.1: Mere uspešnosti za oba algoritma primenom makro prosečnih mera.

VII. Zaključak

Analizom performansi tri različita algoritma klasifikacije - K najbližih suseda (KNN), Mašina na bazi vektora nosača (SVM) i Slučajne šume (Random Forest), kako u osnovnoj implementaciji tako i uz primenu LDA algoritma za smanjenje dimenzionalnosti, došle smo do nekoliko važnih zaključaka.

Kada je u pitanju KNN algoritam, uočavamo minimalno poboljšanje performansi nakon primene LDA, što ukazuje na to da KNN već dobro generalizuje u osnovnoj formi. S druge strane, SVM pokazuje značajno poboljšanje mera uspešnosti nakon primene LDA, što sugeriše da je LDA pomogao algoritmu da bolje razdvoji klase u prostoru manje dimenzionalnosti. Random Forest, iako već efikasan algoritam u osnovnoj formi, takođe pokazuje neznatno poboljšanje u performansama nakon primene LDA, što ukazuje na to da je LDA koristan u smanjenju dimenzionalnosti podataka i boljoj separaciji klasa.

Čisti podaci (bez nevalidnih, nekonzistentnih ili nedostajućih vrednosti) osiguravaju da algoritmi ne budu ometeni šumom ili nekonzistentnošću u podacima, što bi moglo dovesti do nepouzdanih rezultata. Kvalitet podataka igra ključnu ulogu u postizanju visoke tačnosti u analizi podataka, upravo to je razlog zbog čega sva tri algoritma, sa i bez smanjenja dimenzionalnosti, postižu visoku tačnost. Uprkos tome, KNN algoritam sa smanjenjem dimenzionalnosti možemo izdvojiti kao najbolji od tri navedena u problemu klasifikacije, što implicira da ovaj algoritam pruža najpouzdanije predikcije za ciljnu varijablu.

VIII. Literatura

Mueller, A., & Guido, S. (2016). *Introduction to Machine Learning with Python: A Guide for Data Scientists* (1st ed.). O'Reilly Media