Analiza meteoroloških podataka i procena zagađenosti vazduha u Pekingu 2010-2015

Milan Veljković, IN35-2017, milan.veljkovic@uns.ac.rs

I. Uvod

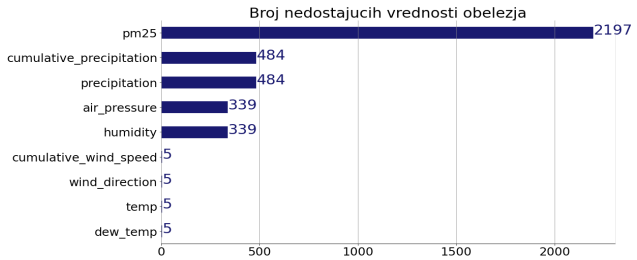
Prema podacima Svetske zdravstvene organizacije (*WHO*), zagađen vazduh je povezan sa više od milion smrtnih slučajeva u Kini svake godine[1]. Među najopasnijm zagađivačima spadaju *Particulate matter (PM)* čestice koje dolaze iz izduvnih gasova vozila, emisije iz sagorevanja uglja i drveta. PM2.5 čestice su se vrlo male čestice, manje od 2.5 mikrometara u prečniku.

Ovaj rad ima za cilj analizu meteoroloških podataka, njihov uticaj na koncentraciju PM2.5 čestica u vazduhu i kreiranje modela za predikciju koncentracije PM2.5 čestica na osnovu meteoroloških podataka.

II. Baza podataka

Skup podataka se sastoji od 52584 uzorka i 18 obeležja. Uzorak predstavlja izmerene vrednosti obeležja u toku jednog sata u Pekingu u periodu od 01-01-2010 00:00 do 31-12-2015 23:00. Postoji 6 kategoričkih i 12 numeričkih obeležja. Kategorička obeležja su: redni broj uzorka(‘*No’*), godina(*‘year’*), mesec(‘*month*’), dan(‘*day*’) i sat merenja(‘*hour*’), godišnje doba(‘*season*’), pravac vetra(‘*wind_direction*’). Numerička obeležja su: temperatura rose(‘dew_temp’), temperatura vazduha(‘temp’), vlažnost vazduha(‘*humidity’*), vazdušni pritisak(‘*air_pressure’*), kumulativna brzina vetra(‘*cumulative_wind_speed’*), padavine na sat(‘*precipitation’*), kumulativne padavine(‘*cumulative_precipitation'*) i koncentracija PM2.5 čestica na stanicama Dongsi, Dongsihuan, Nongzhanguan i US Post.

Značajan broj vrednosti obeležja koncentracije pm2.5 čestica na stanicama Dongsi, Dongsihuan i Nongzhanguan nedostaje, pa su ta obeležja izbačena. Takodje, izbačeno je obeležje redni broj uzorka, jer ne objašnjava ništa o uzorcima. Broj uzoraka sa nedostajućim vrednostima po obeležjma se moze videti na slici 1



Sl. 1. Broj nedostajućih podataka po obeležju

Analizom uzoraka sa nedostajućim vrednostima obeležja primećeno je da uzorcima kojima nedostaje vrednost obeležja temperatura nedostaju i vrednosti za obeležja temperatura rose, pravac vetra, vlaznost vazduha, vazdusni pritisak, kumulativna brzina vetra. Kako tim uzorcima nedostaju vrednosti čak šest obeležja, što predstavlja više od trećine ukupnog broja obeležja, ovi uzorci se izbacuju.

Nedostajuće vrednosti svih ostalih obeležja, sem kumulativne količine padavina, popunjene su poslednjom validnom vrednošću adekvatnog obeležja, dokle god poslednja validna vrednost nije izmerena pre više od 2 sata. Nakon popunjavanja, od 326 nedostajućih vrednosti obeležja vazdušni pritisak i vlažnost vazduha, postoji 293 sukcesivnih redova sa nedostajućim vrednostima, što je vise od 12 dana nedostajućih vrednosti. Popunjavanje nedostajućih vrednosti srednjom vrednošću obeležja grupisanom po mesecu bi bilo neispravno zbog postojanja dugih nizova uzastopnih uzoraka sa nedostajućim vrednostima. Vrednosti koje nedostaju su popunjene srednjom vrednošću dana grupisanih po mesecu. Nedostajuće vrednosti padavina popunjene medijanom meseci grupisanih po godini. Vrednost koje nedostaju za obeležje kumulativne količine padavina izračunate su nakon popunjavanja nedostajućih vrednosti obeležja padavina. Nakon inicijalnog popunjavanja poslednjim validnim vrednostima, ostalo je čak 1828 uzoraka sa nedostajućim vrednostima obeležja koncentracija PM2.5 čestica, kako ovaj broj nije zanemarljiv, podaci su popunjeni medijanom dana grupisanom po mesecu. Grupisanje po mesecima je potrebno jer postoje dani u kojima nedostaju svi podaci, pa ne bi bilo moguće izračunati medijanu tih dana.

III. Analiza obeležja

Nakon preobrade podataka, ostalo je 52579 uzorka i 14 obeležja. Deskriptivnom statističkom analizom utvrdjeno je da dinamički opseg(prikazan na tabeli 1) obeležja padavina na sat i kumulativnih količina padavina iznosi 999990.00 i da je u bazi zabeležen samo jedan uzorak sa vrednošću 999990.00mm. Medjutim, prema Svetskoj meteorološkoj organizaciji *(WMO)*, najveća zabeležena količina padavina po satu iznosi 305mm [2]. Iz navedenog razloga, moze se pretpostaviti da je došlo do greške pri unosu vredosti ovih obeležja, pa se ovaj uzorak izbacuje. Polovina uzoraka obeležja kumulativna brzina vetra se

[1] https://apps.who.int/iris/bitstream/handle/10665/250141/9789241511353-eng.pdf?sequence=1&isAllowed=y

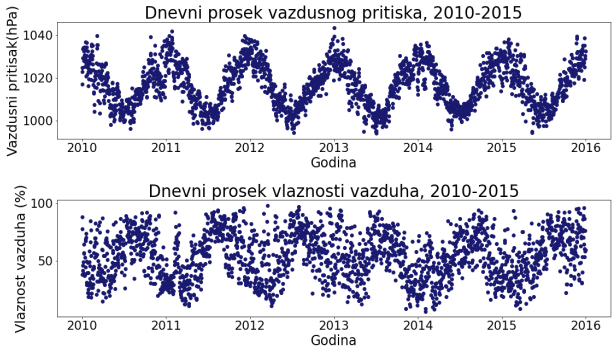
[2] https://wmo.asu.edu/content/world-greatest-sixty-minute-one-hour-rainfall

nalazi u opsegu od 19.23, dok mu je dinamički opseg čak 558.15. Raspodele obeležja kumulativna brzina vetra, količina padavina i kumulativna količina padavina imaju koeficijente zakrivljenosti od 4.399, 38.110, 30.597, respektivno, što znaci da su raspodele ovih obeležja u velikoj meri asimetrične. Takvi koeficijenti zakrivljenosti obeležja padavina i kum.kol. padavina mogu biti objašnjene činjenicom da oko 96% uzoraka ima vrednost 0 ovih obeležja. Ista ova obeležja imaju i najveće vrednosti koeficijenta spljoštenosti, i to 24.489, 2211.874 i 1379.411, tim redom.

Tabela 1: Dinamički, interkvartilni opseg i medijana obeležja nakon popunjavanja nedostajućih vrednosti

Obelezje	Dinamicki opseg	IQR	Medijana
pm25	993.00	100.00	69.00
dew_temp	68.00	25.00	2.00
humidity	98.00	46.00	55.00
air_pressure	55.00	17.00	1016.00
temp	61.00	21.00	14.00
cumulative_wind_speed	585.15	19.23	4.92
precipitation	999990.00	0.00	0.00
cumulative_precipitation	999990.00	0.00	0.00

Sa slike 2 moze se zakljuciti da prosečne dnevne vrednosti vazdušnog pritiska i vlažnosti vazduha imaju sezonsku komponentu.

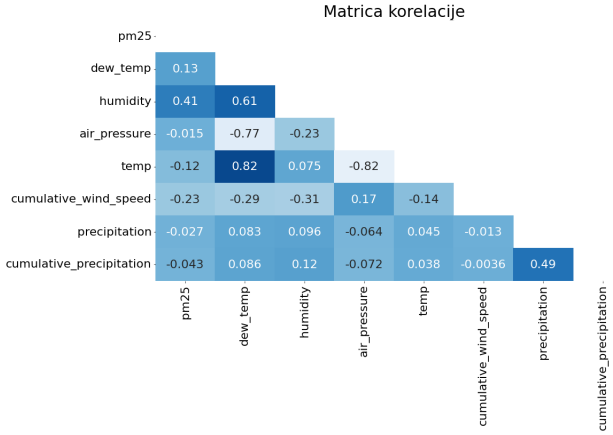


Sl. 2. Dnevne srednje vrednosti vlaznosti vazduha i vazdusnog pritiska od 2010. do 2015. godine.

Obeležje pm25 se kreće u opsegu 1-994 mikrometara po kubnom metru, dok se 50% uzoraka nalazi u opsegu od 100. Uočeno je da postoji veliki broj autlajera obeležja pm25 u januaru. Analizom obeležja primećuje se da su tri najviše vrednosti obeležja pm25 izmerene na dan Lunarne nove godine, jednog od najvećih praznika u Kini. Po tradiciji, za proslavu Lunarne nove godine ispaljuje se velika količina vatrometa, a emisije iz vatrometa utiču na povećanje koncentracije PM2.5 čestica u vazduhu. Narednih vrednosti sa najvecom koncentracijom PM2.5 čestica su zabeležene dana 12.01.2013., koji važi za dan sa najvišim zabeleženim vrednostima PM2.5 čestica u Pekingu do tada[3]. Konstatovano je da su ovi autlajeri ispravno uneti i ne treba ih ukloniti.

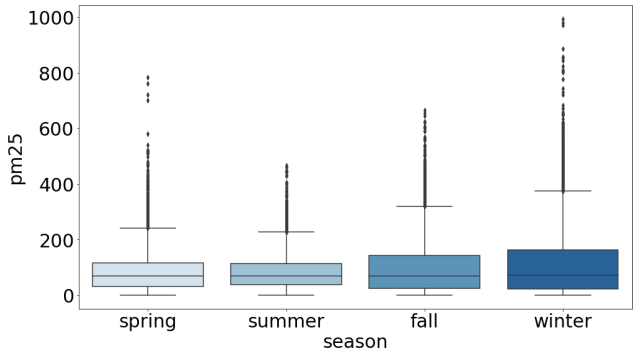
[3] https://www.cnbc.com/id/100375537

Sa matrice korelacije moze se pročitati da je izlazna promenljiva u najvećoj korelaciji sa obeležjima vlažnost vazduha, temperatura rose i temperatura vazduha. Obeležje pm25 je u najnižoj korelaciji sa obeležjima vazdušni pritisak, količina padavina i kumulativna količina padavina. Među nezavisnim obeležjiama, najveća korelacija postoji između obeležja temperatura i temperatura rose, temperatura i vazdušni pritisak i obeležja temperatura rose i vazdušni pritisak.



Sl. 3. Matrica korelacije između numeričkih obeležja

Primećuje se razlika između interkvartilnih opsega koncentracije pm25 čestica po godišnjim dobima. IQR obeležja PM2.5 u zimu iznosi 141, dok je leti IQR samo 76. Takodje, zimi postoji značajnije vise autlajera nego drugim godišnjim, što je moguće prouzrokovano spaljivanjem uglja za potrebe grejanja



Sl. 4. Boxplot obeležja PM2.5 po godisnjim dobima

Obeležje pravac vetra ima vrednosti tipa *string*. Kao takvo, nepogodno je za obučavanje modela linearnog regresora. Da bi bilo moguće upotrebiti ga u obuci, njegove vrednosti su prebačene u numeričke na taj način što su grupisani podaci o koncentraciji PM2.5 čestica po pravcu duvanja vetra i onoj grupi sa najvećom srednjom vrednošću je dodeljenja najveća vrednost pravca duvanja vetra i tako po grupama sve do najmanje srednje vrednosti. Uzorci koji su odbirkovani u mesecu januar imaju 12 puta veću vrednost obeležja *month* od uzoraka koji su odbirkovani u mesecu decembar. Ovakav odnos vrednosti kategoričkog obeležja nije prikladan za treniranje modela linearne regresije. Obeležja godina, mesec merenja i godišnje doba su *one-hot* kodovana kako bi veze između različitih vrednosti navedenih obeležja bile održane.

One-hot kodovanjem obeležja sat i dan dimenzionalnost problema bi se povecala za 46 obeležja, sto onemogućava obuku kompleksnijih modela na dostupnim računarima. Načinjen je pokusaj da se dani kodiranju jedinicom ako predstavljaju subotu ili nedelju, a nulom za druge vrednosti i da se sati kodiraju jedinicom ako imaju vrednost vecu od 16, a nulom za druge vrednosti, međutim, pri selekciji obeležja unazad, ova obeležja imaju p-vrednosti veće od zadatog praga(1%) i trebalo bi ih izbaciti. Izbacivanjem nekodovanih obeležja dobijaju se modeli sa lošijim performansama, pa su, iz navedenih razloga, ova obeležja ostavljena u formi u kakvoj jesu.

IV. Linearna regresija

U ovom odeljku su opisani kreirani modeli linearne regresije. Nakon popunjavanja nedostajućih vrednosti i kodovanja kategoričkih obeležja, baza se sastoji od 29 obeležja i sadrži 52578 uzoraka. Baza je podeljena na tri dela: 70% je ostavljeno za trening modela, 15% uzoraka za testiranje modela i 15% uzoraka za validaciju.

Modeli sa različitim hipotezama i različitim vrednostima hiperparametara su trenirani nad trening skupom, testirani na validacionom skupu, nakon čega su modeli sa najboljim rezultatima ponovo istrenirani nad unijom trening i validacionog skupa i testirani nad test skupom. Nad podacima je urađena z-normalizacija, tj. srednje vrednosti obeležja su svedene na 0, a standardna devijacija na 1. Normalizacija je urađena modeli bili manje osetljivi na dinamičke opsege obeležja i kako bi se obuka ubrzala. Kod selekcije obeležja unazad redom se odbacuju obeležja sa velikom p-vrednošću. Međutim, sva obeležja imaju vrednosti manje od praga 0.01. To znači da su sva obeležja značajna za predikciju modela. Obučeni model linearne regresije sa hipotezom $h_1(x) = \theta_0 + \theta_1 * x_1 + \theta_2 * x_2 + ... + \theta * x$ je model bez regularizacionog člana. Ovakav model daje rezultate MSE = 4565.131, MAE = 49.075, RMSE = 67.566, R^2 = 0.412 i R^2 prilagodjen = 0.411. Ovaj prost model se pokazao kao neperformantan i u proseku greši za 49.075 µg/m3, što je nešto manje od polovine interkvartilnog opsega obeležja koncentracije pm2.5 čestica. Kada se u hipotezu uključe i interakcije između obeleža, hipoteza dobija oblik

$h_2(x) = \theta_0 + \theta_1 * x_1 + \theta_2 * x_2 + ... + \theta * x + \theta_{+1} * x_1x_2 + \theta_{+2} * x_1x_3 + ... + \theta_2 {}_{-1} * x {}_{-2}x {}_{-1} + \theta_2 * x {}_{-1}x$.

Ovakvi modeli daju bolje rezultate kada postoje jake korelacije izmedju nezavisnih obeležja. Mere uspešnosti ovakvih modela, sa i bez uključenim regularizacionim parametrima, prikazane su u tabeli 2.

Tabela 2: Mere uspešnosti modela kada hipoteza uključuje interakcije među obeležjima

Mere uspesnosti	Lasso $\alpha = 0.005$	Ridge $\alpha = 0.5$	Bez regularizacije
MSE	3445.734	3442.760	3442.652
MAE	41.589	41.571	41.570
RMSE	58.700	58.675	58.674
R^2	0.556	0.556	0.556
R^2 prilagođen	0.552	0.552	0.552

Modeli sa interakcijama imaju značajno bolje rezultate u odnosu na model bez interakcija, što se može videti i po vrednosti R^2 prilagođenog skora, koji uzima u u obzir i broj obeležja modela i ima vrednost 0.552.

Sledeći modeli čije su vrednosti razmotrene su modeli sa hipotezom $h_3(x) = \theta_0 + \theta_1 * x_1 + \theta_2 * x_2 + ... + \theta * x + \theta_{+1} * x_1^2 + \theta_{+2} * x_1x_2 + ... + \theta_2 {}_{-1} * x {}_{-1}x + \theta_2 * x {}^2$. U ovim modelima su uključeni i monomi na kvadrat. Mere uspešnosti testiranih modela su opisane u tabeli 3.

Tabela 3: Mere uspešnosti modela kada hipoteza uključuje interakcije među obeležjima i monome do stepena 2

Mere uspesnosti	Lasso $\alpha = 0.005$	Ridge $\alpha = 0.5$	Bez regularizacije
MSE	3405.714	3405.202	3405.237
MAE	41.415	41.422	41.422
RMSE	58.358	58.354	58.354
R^2	0.561	0.561	0.561
R^2 prilagođen	0.556	0.557	0.556

Sa ovom hipotezom, *Lasso* regularizacijom odbačeno je 108 obeležja, a selektovano je preostalih 356. Rezultati se nisu značajno poboljšali u odnosu na prethodno opisani model.

U sledećoj tabeli razmotreni su modeli sa hipotezom $h_4(x) = \theta_0 + \theta_1 * x_1 + \theta_2 * x_2 + ... + \theta * x + \theta_{+1} * x_1^2 + \theta_{+2} * x_1x_2 + ... + \theta_2 {}_{-1} * x {}_{-1}x + \theta_2 * x {}^2 + \theta_2 {}_{+1} * x_1^3 + \theta_2 {}_{+2} * x_1^2x_2 + ... + \theta_3 {}_{-1} * x {}_{-1}x {}^2 + \theta_3 * x {}^3$

Tabela 4: Mere uspešnosti modela kada hipoteza uključuje interakcije među obeležjima i monome do stepena 3

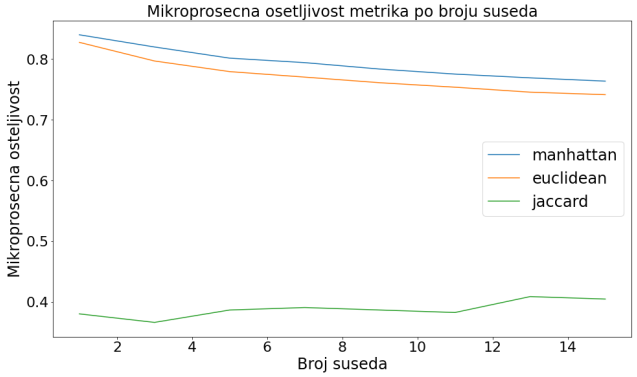
Mere uspesnosti	Lasso $\alpha = 0.05$	Ridge $\alpha = 500$	Bez regularizacije
MSE	2725.109	2785.994	3135.248
MAE	36.280	36.454	36.833
RMSE	52.202	52.782	55.993
R^2	0.649	0.641	0.596
R^2 prilagođen	0.605	0.596	0.546

Broj obeležja se povećao na 4959 obeležja. *Lasso* regresijom je selektovano 1241, a odbačeno preostalih 3718 obeležja. U modelima sa ovom hipotezom se vrednost srednje apsolutne greške smanjila na oko 36.0 mikrometara po metru kubnom, što je za oko 10% niža vrednost u odnosu na modele sa prethodne dve hipoteze. Takodje, R^2 prilagodjeni skor je primetno bolji.

Modeli sa hipotezom $h_5(x) = \theta_0 + \theta_1 * x_1 + \theta_2 * x_2 + ... + \theta * x + \theta_{+1} * x_1^2 + \theta_{+2} * x_1x_2 + ... + \theta_2 {}_{-1} * x {}_{-1}x + \theta_2 * x {}^2 + \theta_2 {}_{+1} * x_1^3 + \theta_2 {}_{+2} * x_1^2x_2 + ... + \theta_3 {}_{-1} * x {}_{-1}x {}^2 + \theta_3 * x {}^3 + \theta_3 {}_{+1} * x_1^4 + ... + \theta_4 * x {}^4$ imaju 27840 obeležja i zbog svoje kompleksnosti, i postojanja tehničkih ograničenja, nije bilo moguće obučiti ih i testirati. Od svih testiranih modela, model sa hipotezom $h_4(x)$, koji minimizuje L1 normu, ima najbolje mere uspešnosti.

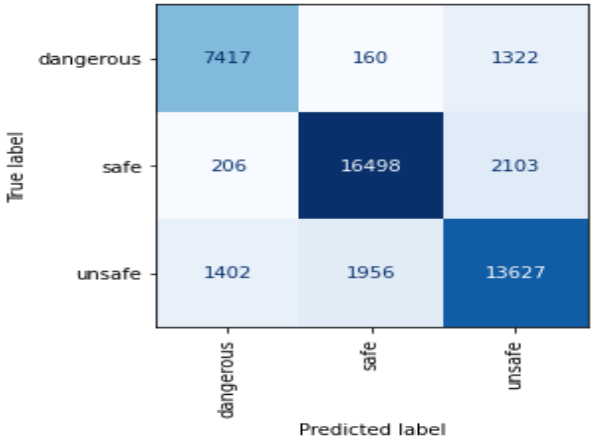
V. KNN KLASIFIKATOR

Za potrebe KNN klasifikatora, baza podataka je podeljena na trening skup, koji čini 85% baze i test skup koji predstavlja 15% baze. Za pronalazak optimalnih hiperparametara iz skupa ponuđenih, korišćena je unakrsna validacije sa 10 podskupova uz pretragu optimalne kombinacije parametara. Prosleđene vrednosti parametara su 1,3,5,7,9,11,13,15 za parametar k-broj suseda koji se posmatraju pri odluci, a za parametar metrike ponuđene su Hamingova, Euklidska i Menhetn metrika. Kako broj klasa koje se predviđaju u bazi nije jednak, za meru uspešnosti koristi se mikroprosečna osetljivost. Rezultati pretrage su prikazani na slici 5



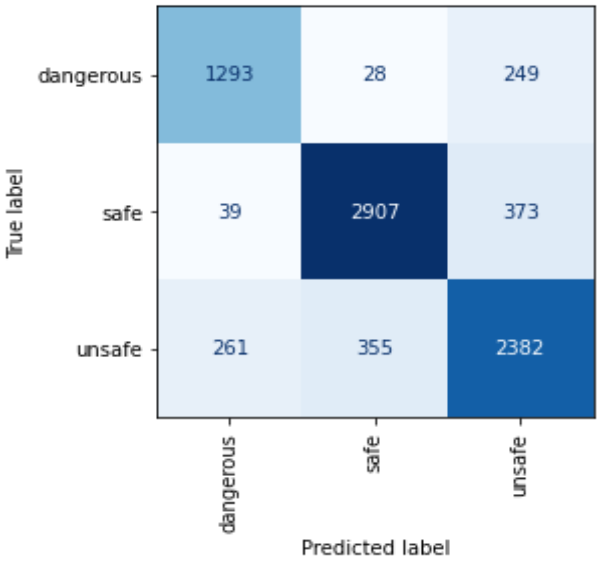
Slika 5: Mikro preciznost po parametru i metrici

Kao optimalni parametri izabrani su Menhetn metrika i 1 za broj suseda. Nakon pronalaska parametara, nad trening skupom izvršena je unakrsna validacija, sa tim da podskupovi imaju očuvane zastupljenosti svake klase. Sa matrice konfuzije, prikazane na slici 6. može se primetiti da model u velikom broju slučajeva pravi ispravne predikcije. Najveći broj grešaka model je napravio pri klasifikaciji uzoraka iz klase ‘nebezbedno’. Najtačnije je procene je pravio nad uzorcima iz klase ‘opasno’



Slika 6. Matrica konfuzije modela dobijenog unakrsnom validacijom sa optimalnim parametrima

Klasifikator sa optimalnim vrednostima parametara je ‘obucen’ nad celim trening skupom i testiran nad test skupom. Matrica konfuzije treniranog klasifikatora je na slici 7.



Slika 6. Matrica konfuzije modela dobijenog unakrsnom validacijom sa optimalnim parametrima

Tabela 5: Mere uspešnosti KNN klasifikatora nad test skupom po određenim klasama

Mere uspesnosti	Bezbedno	Nebezbedno	Opasno
Tačnost	0.899	0.843	0.927
Preciznost	0.883	0.793	0.811
Senzitivnost	0.875	0.794	0.823
Specifičnost	0.916	0.873	0.952
F-mera	0.879	0.793	0.817

Mere uspešnosti klasifikatora testiranog na test skupu se nisu značajno promenile u odnosu na model dobijen unakrsnom validacijom na trening skupom. Klasifikator je ponovo najveći broj grešaka napravio pri proceni uzoraka koji pripadaju klasi ‘nebezbedno’. Mikroprosečna osetljivost novog klasifikatora, koja predstavlja količnik sume stvarnih pozitiva po svim klasama i sume stvarnih pozitiva i lažnih negativa po klasama, iznosi 0.834.

VI. ZAKLJUČAK

Može se zaključiti da su oba pristupa predikciji zagađenosti vazduha validni i da tip modela treba birati u zavisnosti od karakteristika podataka koji su dostupni. KNN klasifikator se pokazao kao pouzdan model za predikciju zagadjenosti u određenim opsezima. Kako kompleksnost modela linearne regresije raste, tako imaju bolje rezultate u predviđanju koncentracije PM2.5 čestica. Ostaje prostora za ispitivanje kompleksnijih modela linearne regresije, i generalno, za unapređenje opisanih modela.