Analiza podataka, predikcija koncentracije PM2.5 čestica u vazduhu i klasifikacija uzoraka

Elena Akik, IN 6/2019, akik.in6.2019@uns.ac.rs

I. Uvod

Tema ovog izveštaja je analiza podataka o vremenskim uslovima koji se odnose na zagađenost vazduha u glavnom gradu provincije Sečuan, gradu Čengdu, u Kini. Jedan od brojnih pokazatelja zagađenosti vazduha je koncentracija PM2.5 čestica, koje se sastoje od finih čestica prečnika manjeg od 2.5 mikrometra. Ove čestice mogu doći iz različitih izvora, uključujući emisije štetnih gasova iz vozila, industrije, ali i šumskih požara. Kako je Čengdu veliki ekonomski, saobraćajni i kulturni centar sa populacijom od preko 14 miliona ljudi, sa visokim nivoom industrijske i saobraćajne delatnosti, ima povišen nivo koncentracije PM2.5 čestica, kao i drugih oblika zagađenja vazduha. Izloženost PM2.5 povezana je sa nizom negativnih efekata na zdravlje, te je za stanovnike ovog grada značajno saznati koliki je zapravo štetan uticaj ovog zagađivača i šta to najviše utiče na isti, kako bi se preduzele adekvatne mere zaštite.

II. Baza podataka

Baza sadrži podatke o 52 584 uzorka, koji su opisani sa 17 obeležja. Kategorička obeležja su: redni broj vrste(*No*), godina(*year*), mesec(*month*), dan u mesecu(*day*), sat u danu(*hour*), godišnje doba(*season*) i pravac vetra(*cbwd*). Numerička obeležja su: koncentracija PM2.5 čestica na nekoliko lokacija(*PM_Caotangsi, PM_Shahepu, PM_US Post*), temperatura rose(*DEWP*), temperatura(*TEMP*), vlažnost vazduha(*HUMI*), vazdušni pritisak(*PRES*), kumulativna brzina vetra(*Iws*), padavine na sat(*precipitation*) i kumulativne padavine(*Iprec*). Jedan uzorak baze odnosi se na izmerene vrednosti za sva pobrojana obeležja u toku jednog sata.

III. ANALIZA PODATAKA

Prilikom analiziranja baze, izostavljene su vrednosti obeležja *PM_Caotangsi i PM_Shaheou*, pa se dalja analiza odnosila samo uzorke vezane za *PM_US Post*. Izbačeno je i obeležje *No*, koji se odnosio na redni broj vrste, jer isto nije relevantno za dalju analizu.

A. Nedostajući podaci

Ispitivanjem dostupnih podataka iz baze ustanovljeno je da nedostajući podaci postoje i to kod sledećih obeležja, u sledećem udelu: *PM_US Post*(45.04%), *precipitation*(5.62%), *Iprec*(5.62%), *HUMI*(1.02%), *Iws*(1.01%), *DEWP*(1.01%), *TEMP*(1%), *PRES*(0.99%), *cbwd*(0.99%). Kako je najveći udeo nedostajućih podataka vezan za obeležje *PM_US Post*, koje je glavno obeležje za dalju analizu, nelogično je i nemoguće izbaciti ga. Nakon analize nedostajućih vrednosti po godinama, uočeno je da kod uzoraka koji se odnose na 2010. i 2011. godinu postoji 100% nedostajućih vrednosti za dato obeležje, te je odluka da se ti uzorci izbace. Tokom 2012. godine, postoji samo polvina dostupnih podataka za dato obeležje, pa će i uzorci za tu godinu biti izbačeni. Tokom 2013, 2014. i 2015. godine, udeo podataka koji nedostaje za dato obeležje je manji od 20%, te ovi podaci neće biti izbačeni, već korigovani metodom prepisivanja prve fizički najbliže poznate vrednosti iz baze za dato obeležje. Nakon izbacivanja i korekcije vrednosti datog obeležja, udeo nedostajućih obeležja je:

Tabela 1: Prikaz obeležja sa udelom nedostajućih vrednosti nakon rešavanja problema nedostajućih vrednosti za *PM_US Post*

Naziv obeležja	Broj uzoraka sa nedostajućim vrednostima
Iws	498
HUMI	493
DEWP	490
TEMP	489
PRES	486
cbwd	1224

Kada je reč o uzorcima kod kojih obeležja *Iws, HUMI, DEWP, TEMP, PRES* i *cbwd* imaju nedostajuće vrednosti, uočava se da vrednosti istovremeno nedostaju, pa će ti uzorci biti izbačeni. Nakon izbacivanja datih uzoraka, opet postoje nedostajuće vrednosti navedenih obeležja za određene uzorke, ali kako ovi podaci nisu vremenski korelisani i ne nedostaju istovremeno, neće biti izbačeni, već će se primeniti metoda prepisivanja prve fizički najbliže poznate vrednosti iz baze za dato obeležje, osim za uzorke kod kojih postoji nedostajuća vrednost za obeležje *Iprec*. Za te uzorke nedostajuće vrednosti korigovaće se sabiranjem količine padavina tog sata(*precipatition*) sa ukupnom količinom padavina prethodnog sata(*Iprec*).

B. Kategorička obeležja

Pre analize obeležja koji se odnose na PM2.5,

kategoričko obeležje koje se odnosi na pravac vetra, *cbwd*, biće predstavljeno numeričkim vrednostima. Dato obeležje je trenutno označeno pomoću kombinacija oznaka strana sveta. Nove numeričke oznake biće formirane na osnovu koncepta kompasa i kretanja po zamišljenoj kružnici u smeru suprotnom od kazaljke na satu, gde će svakom od kategoričkih obeležja ovog atributa biti dodeljen ugao za koji smo se pomerili od nultog stepena kružnice. Smer kretanja može biti pozitivan ili negativan, te će tako postojati pozitivne i negativne vrednosti. Severoistok(NE) će imati vrednost 45, jugoistok(SE) 135, severozapad(NW) -45, jugozapad(SW) -135, dok će vrednosti cv biti dodeljena vrednost 00, koja ukazuje na nepromenjeno stanje ili potpuno promenljivo stanje.

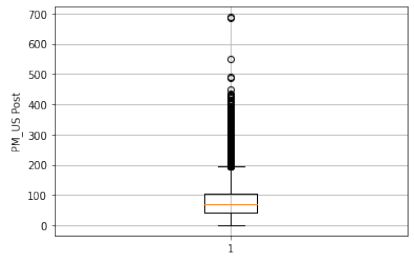
C. *Statističke veličine obeležja*

Tabela 2: Statističke veličine obeležja

	DEWP	HUMI	PRES	TEMP	Iws	precipitation
mean	12.78	72.76	1014.52	18.29	4.30	0.11
min	-16.00	12.78	991.00	-2.00	0.00	1.00
25%	7.00	60.74	1008.00	12.00	1.00	0.00
50%	14.00	76.35	1014.90	19.00	2.00	0.00
75%	19.00	87.75	1021.00	24.00	5.00	0.00
max	28.00	100.00	1041.00	38.00	93.00	51.70
skewness	-0.35	-0.62	0.08	-0.15	4.57	24.29
kurtosis	-0.80	-0.36	-0.76	-0.86	32.74	806.79

Na osnovu podataka datih u tabeli 2 može se uočiti da na osnovu dobijenih statističkih veličina za svako obeležje one su logične i validne, u poređenju sa realnim svetom. Za obeležja *DEWP, HUMI, PRES* i *TEMP* srednja vrednost i medijana su približno jednake što svedoči da ne postoji veliki broj autlajera sa izraženom malom ili velikom vrednošću. Kod obeležja precipitation, koje se odnosi na padavine u toku sata, mogu se uočiti izražene visoke vrednosti kod određenog broja uzoraka. Kako se vrednosti mere iz sata u sat, u ovoj analizi biće pretpostavljeno da su podaci ispravni. U cilju potpune sigurnosti u validnost podataka, korisno je posavetovati se sa nadležnim organom radi utvrđivanja tačnosti datih vrednosti.

D. *Analiza atributa PM_US Post(PM2.5)*

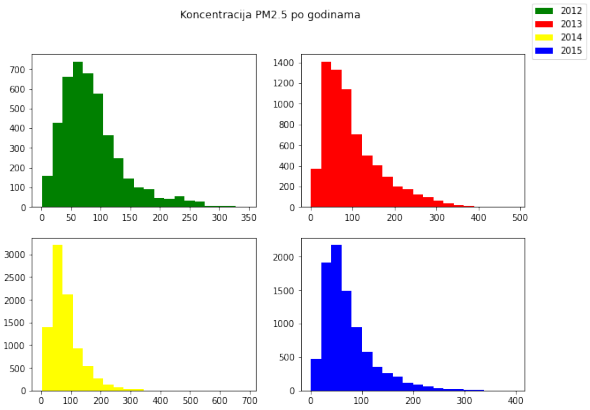


Count	28810
Mean	83.63
Std	57.16
Min	1
25%	44
50%	69
75%	105
max	688

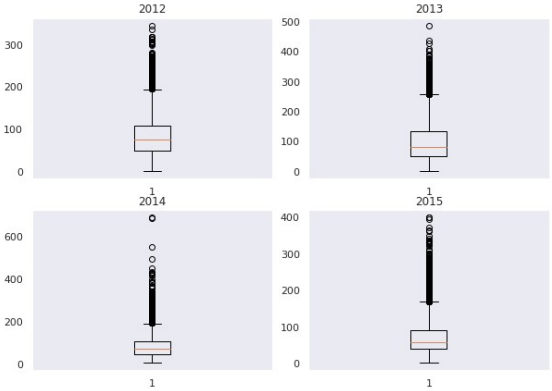
Slika 1: Boxplot i statistički parametri atributa PM_US Post

Na osnovu podataka sa slike 1, može se videti da je interkvartilni opseg na intervalu od 44 ug/m3 do 105 ug/m3, u kome se nalazi 50% uzoraka. Uočava se veliki broj autlajera koji su znatno udaljeni u odnosu na srednju vrednost. Za niske vrednosti ne postoje autlajeri. U bazi se nalazi određeni broj uzastopnih uzoraka čije su vrednosti relativno bliske maksimumu, te je na nadležnima da provere da li je u pitanju validna vrednost ili greška. U daljoj analizi, nije izbačen niti promenjen nijedan uzorak baze. Dobijene vrednosti koeficijenata asimetrije i spljoštenosti ukazuju na postojanje desne asimetrične raspodele analiziranog obeležja, koja je izdignuta u odnosu na normalnu raspodelu, što je posledica velikog dinamičkog opsega od 688 ug/m3, pri čemu se polovina uzoraka pak nalazi na intervalu od 44 ug/m3 do 105 ug/m3, što utiče na iskrivljenost raspodele.

1) *Analiza obeležja PM_US Post po godinama*

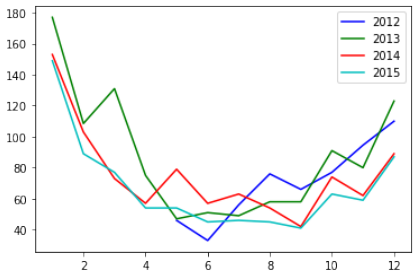


Slika 2: Prikaz koncentracije PM2.5 čestica po dostupnim godinama



Slika 3: Boxplot-ovi atributa PM_US Post po dostupnim godinama

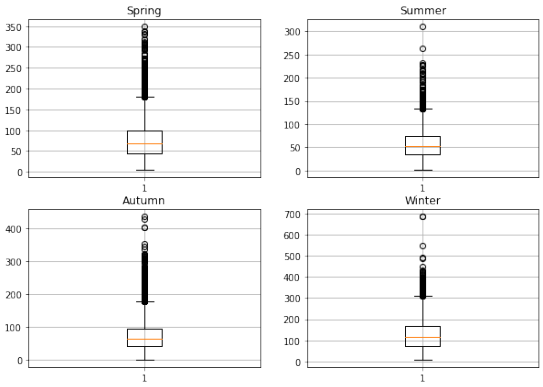
Na osnovu slika 2 i 3 vidi se da se koncentracija PM2.5 čestica ne menja značajno tokom dostupnih godina, te da je najviša koncentracija zabeležena tokom 2014, kad se uočava i prisustvo većeg broja autlajera sa višim izraženim vrednostima, u odnosu na preostale godine.



Slika 4: Srednja mesečna koncentracija PM2.5 čestica za sve dostupne godine

Na osnovu grafikona sa slike 4, može se videti da se tokom svih dostupnih godina, posmatrajući svaki mesec pojedinačno, za iste mesece beleže relativno slične vrednosti. Odstupanje u smislu većih vrednosti javlja se u periodu marta 2013. godine. Vrednosti PM2.5 su najniže tokom aprila, maja, juna i jula, dok se najviše vrednosti beleže tokom januara i decembra. Kako je uočena „pravilnost“ po sezonama, biće izvršena i analiza koncentracije PM2.5 čestica i po sezonama, u cilju preciznije analize.

2) Analiza obeležja PM_US Post po sezonama

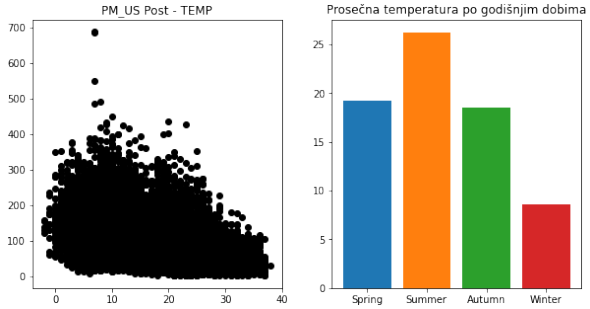


Slika 5: Boxplot-ovi za koncentraciju PM2.5 tokom sezona za sve dostupne godine

Na osnovu slika 4 i 5 vidi se da su vrednosti PM2.5 najniže tokom leta, kada je zagađenje vazduha najmanje i da se kreću na intervalu od 36 ug/m3 do 75 ug/m3. Najveće vrednosti beleže se tokom zime, gde se vrednosti kreću na intervalu od 74 ug/m3 do 168 ug/m3. Ovakav rezultat očekivan je zbog sledećih faktora:

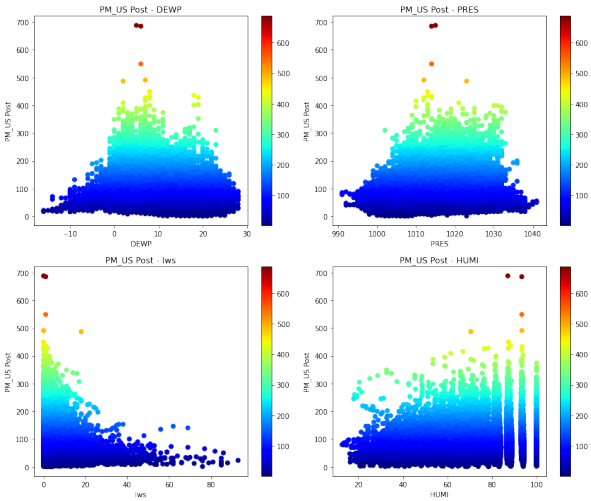
- Tokom zimskih meseci, ljudi imaju tendenciju da koriste grejanje u svojim domovima, što dovodi do povećane upotrebe uglja, drveta i drugih fosilnih goriva, Sgorevanjem ovih goriva, oslobađaju se štetne čestice u vazduh, uključujući i PM2.5.
- Smanjeno je mešanje vazduha, odnosno, hladan vazduh gušći je od toplog, pa tako ima tendenciju da ostane blizu tla. Ovo dovodi do temperaturne inverzije, gde sloj hladnog vazduha zadržava zagađivače blizu površine, što ima za posledicu nakupljanje PM2.5 čestica.
- Smanjena je sunčeva svetlost, tokom zime su dani kraći, što može dovesti do smanjenja količine fotohemijskih reakcija koje se dešavaju u atmosferi. Ove reakcije mogu pomoći da se razbiju zagađivači poput PM2.5, pa samnjenje sunčeve svetlosti može dovesti po povećane koncentracije PM2.5.

E. Analiza zavisnosti obeležja PM2.5 od ostalih obeležžja



Slika 6: Zavisnost obeležja PM2.5 i TEMP

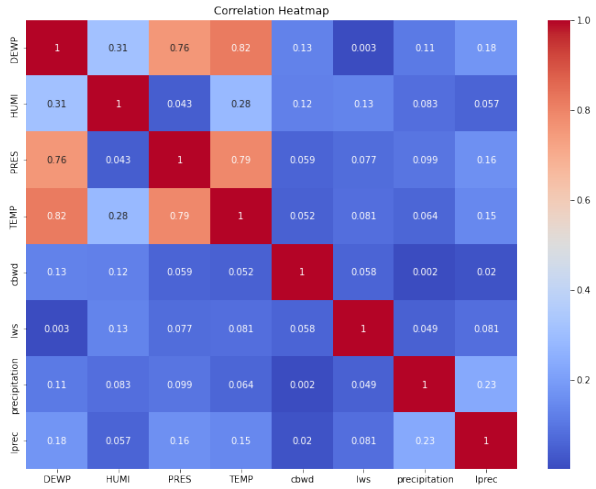
Kako je vrednost korelacije između koncentracije PM2.5 i temperature 0.41, reč je o srednje jakoj korelaciji između obeležja. Na osnovu grafikona, uočava se da su najveće vrednosti koncentracije PM2.5 uočene kada su temperature bile niže, što ukazuje na negativnu korelaciju.



Slika 7: Zavisnost obeležja PM2.5 od obeležja DEWP, PRES, Iws i HUMI

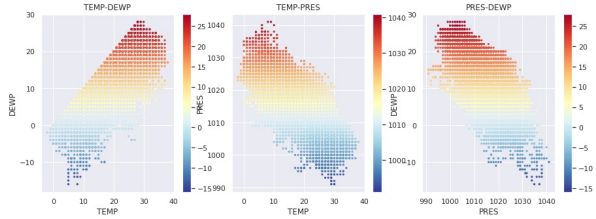
Na osnovu slike 7, može se videti da je koncentracija PM2.5 prilično mala za vrednosti kada je temperatura rose imala vrednost sličnu medijani. Isti zaključak može biti donet i za obeležje *PRES*, koje se odnosi na vazdušni pritisak. Vrednosti za kumulativnu brzinu vetra bile su niže istovremeno kada je i koncentracija PM2.5 bila manja, dok je sa porastom vlažnosti vazduha rasla i koncentracija PM2.5.

F. *Međusobna korelacija obeležja*



Slika 8: Prikaz korelacije među obeležjima bez obeležja PM2.5

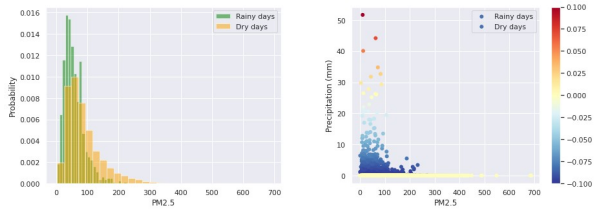
Na osnovu slike 8, može se videti da su najviše korelisani sledeći parovi obeležja: *DEWP – TEMP*(0.82%), *PRES – TEMP* i *PRES*(0.79%) – *DEWP*(0.75%).



Slika 9: Odnos parova obeležja, redom DEWP - TEMP, PRES - TEMP, PRES – DEWP

Na osnovu slika 8 i 9 vidi se da kako vrednost izmerene temperature raste, tako se povećava temperatura rose, te su ta dva obeležja u pozitivnoj korelaciji. Za razliku od temperature rose, vrednost vazdušnog pritiska opada, pa su temperatura i vazdušni pritisak u negativnoj korelaciji. Temperatura rose i vazdušni pritisak su u negativnoj korelaciji. Obeležja koja se odnose na padavine(*precipitation*) i kumulativnu brzinu vetra(*Iws*) imaju malu korelaciju sa drugim obeležjima.

G. *Analiza uticaja padavina*



Slika 10: Odnos kišnih i sušnih dana

Obeležje *precipitation* koje se odnosi na padavine predstavljeno je binarno, gde ima vrednost 0 u slučaju da ne pada kiša, odnosno 1, gde označava da pada kiša. Sa slike 10 uočava se da je mnogo veća verovatnoća pojave viših vrednosti PM2.5 tokom kišnih, nego tokom sušnih dana, kada su i izmerene ekstremne koncentracije PM2.5 čestica u vazuhu.

IV. LINEARNA REGRESIJA

Polazni skup podataka podeljen je na 2 disjunktna podskupa, gde se u jednom nalaze vrednosti svih atributa osim atributa *PM_US Post*, dok drugi podskup čine isključivo vrednosti atributa *PM_US Post*. Izvršena je i trostruka podela podataka, gde se 70% koristi za obuku modela, 15% za validaciju, dok je 15% namenjeno za testiranje. Kako bi obučavanje imalo što bolje rezultate, prethodno su izvršene standardizacija i normalizacija, koje su za cilj imale skaliranje obeležja, kako bi sva obeležja imala slične opsege i dristribucije. Takođe, srednja vrednost postavljena je na 0, dok je standardna devijacija postavljena na 1. Izvršena je i selekcija obeležja, nakon koje je izbačeno obeležje *TEMP*, koje se pokazalo kao obeležje koje nije statistički značajno za dalju analizu.

A. *Hipoteza: y=b0+b1x1+b2x2+...+bnxn*

Tabela 3: Rezultati predikcije modela A

Mera uspešnosti	Model (bez izvršene regularizacije)
MSE	2444.52
RMSE	49.44
MAE	36.29
R2 score	0.27
R2 score adjusted	0.27

Na osnovu podataka prikazanih u tabeli 3, dobijenih nakon predikcije na osnovu modela A, nakon koga su dobijene velike greške, uočava se da model A ne pokazuje bolje rezultate u odnosu na model koji na osnovu određene vrednosti zavisnog obeležja predviđa srednju vrednost istog. Postupak regularizacije nije izvršen jer su kod ovog modela težine koeficijenata ravnomerno raspoređene.

B. *Hipoteza: y=b0+b1x1+b2x2+..+bnxn+c1x1x2+c2x1x3+...*

Tabela 4: Rezultati predikcije modela B

Mera uspešnosti	Model bez regularizacije	Lasso	Ridge
MSE	2188.13	2189.52	2189.64
RMSE	46.77	46.79	46.79
MAE	34.38	34.38	34.34
R2 score	0.34	0.34	0.34
R2 score adjusted	0.34	0.34	0.34

Nakon predikcije na osnovu modela B, u tabeli 4 prikazani su novodobijeni rezultati, na osnovu kojih se može videti da su bolji u poređenju sa rezultatima dobijenim nakon predikcije korišćenjem modela A. RMSE vrednost ukazuje na to da dobijene predviđene vrednosti odstupaju od stvarnih vrednosti za oko 46.77 ug/m3, dok R2 skor pokazuje da model B objašnjava 34% varijanse u zavisnoj promenljivoj, to jest, da je model B u stanju da u

uzme u obzir 34% varijanse u ciljnoj promenljivoj korišćenjem prediktorskih varijabli. Preostala varijansa nije objašnjena modelom i posledica je faktora koji nisu prediktori u modelu ili su slučajne greške. U cilju smanjenja grešaka i poboljšanja rezultata, uvode se novi modeli.

C. Hipoteza:
y=b0+b1x1+b2x2+..+c1x1x2+c2x1x3..+d1x1^2+d2x2^2+...+ dnxn^2

Tabela 5: Rezultati predikcije modela C

Mera uspešnosti	Model bez regularizacije	Lasso	Ridge
MSE	1814.26	1816.33	1844.83
RMSE	42.59	42.61	42.95
MAE	31.27	31.26	31.43
R2 score	0.46	0.45	0.45
R2 score adjusted	0.45	0.45	0.44

Korišćenjem modela C dobijeni su bolji rezultati u odnosu na rezultate dobijene korišćenjem modela B – RMSE sada ima vrednost 42.59 u odnosu na prethodnih 46.77, što znači da je smanjeno odstupanje predviđenih od stvarnih vrednosti, dok model sada objašnjava 46% varijanse, u odnosu na prethodnih 34%, što je relativno značajan pomak.

D. Hipoteza:
y=b0+b1x1+b2x2+..+c1x1x2+c2x1x3..+d1x1^2+d2x2^2+...+ e1x1^3+..enxn^3

Tabela 6: Rezultati predikcije modela D

Mera uspešnosti	Model bez regularizacije	Lasso	Ridge
MSE	1724.87	1557.52	1586.77
RMSE	41.53	39.46	39.83
MAE	28.68	28.51	28.71
R2 score	0.48	0.53	0.52
R2 score adjusted	0.47	0.52	0.51

Nakon predikcije korišćenjem modela D, koji je model 3. stepena, dobijaju se do sada najbolji rezultati, koji pokazuju da predviđene vrednosti odstupaju od stvarnih za 41.53 ug/m3, dok model sada objašnjava 48% varijanse. Predikcija je izvršena i modelima višeg reda, no, kako su korišćenjem istih dobijeni lošiji rezultati, kao finalni model uzima se model D.

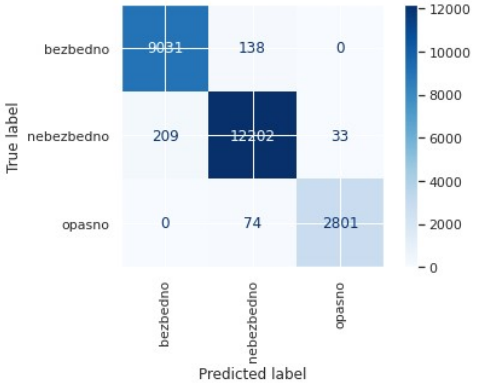
V. KNN KLASIFIKATOR

Za klasifikaciju uzoraka iz baze biće korišćen klasifikator metodom k najbližih suseda. Uzorke je potrebno klasifikovati u jednu od tri grupe zagađenja, na osnovu koncentracije PM2.5 čestica, te je prvo izvršeno labeliranje na osnovu date koncentracije, gde su labele: bezbedno, nebezbedno i opasno. Prvobitno će baza biti podeljena na dva disjunktna podskupa, gde prvi predstavlja skup za metodu unakrsne validacije, koji čini 85% uzoraka, dok će se ostalih 15% uzoraka koristiti za testiranje finalnog klasifikatora.

A. Određivanje optimalnih parametara

Da bi se primenio kNN metod, pored labeliranja uzoraka neophodno je znati broj najbližih suseda koji se razmatraju u procedu odlučivanja - k, kao i metriku kojom se izračunava rastojanje - m. U ovoj analizi, korišćeni su samostalno implementirana unakrsna validacija za kombinacije vrednosti k i n, kao i GridSearchCV funkcija. Kao moguće vrednosti za parametar k date su vrednosti: 1, 3, 5, 10, 15, 17,19, dok su moguće vrednosti za parametar m: „Manhattan“, „Euclidean“ i „Minkowski“. Za unakrsnu validaciju korišćena je StratifiedKFold funkcija, sa 10 particija. Oba načina dala su iste rezultate, to jest, prvi način rađen je samo kao provera i potvrda rezultata drugog. Dobijeni rezultati su: k = 19, m = „Euclidean“. Mikroprosečna osetljivost je 98,15 %.

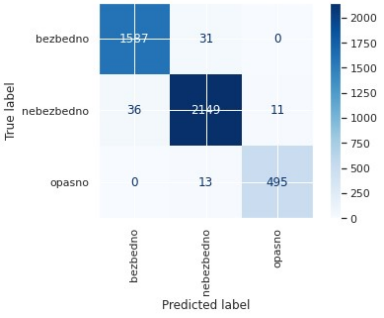
B. Analiza matrice konfuzije nakon unakrsne validacije



Slika 11: Matrica konfuzije nakon unakrsne validacije

Na osnovu matrice konfuzije nakon unakrsne validacije sa optimalnim parametrima, prikazane na slici 11, uočava se da su uzorci koji pripadaju svakoj od navedenih kategorija veoma dobro klasifikovani. Na osnovu vrednosti osetljivosti, zaključuje se da od 100 uzoraka koji su kategorisani kao pripadnici grupe bezbedno, njih 98.49% je tačno kategorizovano. Kod uzoraka koji su u grupi *nebezbedno*, njih 98.05 % je korektno kategorisano, dok je kod kategorije *opasno* tačno kategorisano 97.42 % uzoraka.

C. Analiza matrice konfuzije na test skupu



Slika 12: Matrica konfuzije nad test skupom

Nakon izvršene unakrsne validacije, sledi obuka modela nad trening skupom, koji nije bio prethodno korišćen. Parametri k i m nisu menjani u odnosu na prethodne korake. Nova matrica konfuzije prikazana je na slici 12, a novi rezultati za osetljivost dati su u tabeli 7. Sada, za uzorke klasifikovane kao *bezbedne*, 98.08% klasifikovano je korektno, za uzorke iz klase *nebezbedno* 97.85% uzoraka je tačno klasifikovano, dok je celih 97.44% uzoraka iz klase *opasno* tačno klasifikovano.

Tabela 7: Rezultati osetljivosti nakon unakrsne validacije i obuke finalnog modela

Labela	**Osetljivost nakon unakrsne validacije**	**Osetljivost – finalni model**
Bezbedno	0.9849	0.9808
Nebezbedno	0.9805	0.9785
Opasno	0.9742	0.9744
UKUPNO	**Mikroprosečna osetljivost**	
	0.9799	0.9779

VI. ZAKLJUČAK

Nakon analize podataka, dolazi se do zaključka da je koncentracija PM2.5 čestica izražena tokom zimskih meseci, usled prethodno pobrojanih razloga, te se stanovnicima grada savetuje oprez pri izboru mogućnosti i sredstava za grejanje. Posebnu pažnju treba obratiti i na gasove koje ispuštaju vozila, kao i gasove koji dolaze iz industrija.

Kada je reč o kNN klasifikatoru, nakon poređenja dobijenih rezultata mikroprosečne osetljivosti dobijenih nakon unakrsne validacije, te nakon testiranja na test skupu, uočava se pad iste sa 97.99% na 97.79%, što ukazuje na to da nije došlo do preobučavanja skupa. Ni kod jedne klase uzoraka nema značajnog odstupanja po pitanju osetljivosti, vrednosti za osetljivost svake od njih se u proseku razlikuju oko 1%, te je i dalje većina uzoraka korektno klasifikovana.

Savetuje se primena još nekog od mogućih klasifikatora, poput SVM klasifikatora, koji ima određene prednosti u odnosu na kNN, te poređenja dobijenih rezultata ova dva navedena klasifikatora.