[1]

Koncentracija čestica PM 2.5 u Šangaju

Svetlana Baćina, BI 38/2019,
svetlana.bacina@best-eu.org

I. Uvod

Kina nije oduvek bila jedan od glavnih lidera u svetu. Nakon enomskog fijaska Mao Cendunga, Deng Sjaoping 1979. godine najavljuje sveobuhatnu ekonomsku reformu Kine koja dovodi Kinu u redove ekonosmkih velikana. Danas njena ekonomija čini oko 17 odsto svetskog bruto domaćeg proizvoda, više nije društvo siromašnih poljoprivrednika, već sila koja proizvodi I napredne tehnološke inovacije. Stotine miliona ljudi izašlo je iz siromaštvo i danas žive značajno kvalitetniji život. Ovaj napredak, naravno, nije došao besplatno i cena je bila visoka koja se i dan danas se može osetiti. Uporedni napredak BDP-a blisko je pratilo i zagađenje. Baš kao i po pitanju ekonomije, Kina je od države koja ima minimalni udeo u zagađenju planete prerasla u državu koja je ubedljivo najveći zagađivač na celom svetu, što je, otežalo ljudski i životinjski život

Danas jedan od glavnih problema sa kojim se mi kao zemlja suočavamo je zagađenje vazduha. Zagađenje vazduha podrazumeva prisustvo čestica koji nanose štetu ili uzrokuju nelagodnost kod čoveka i drugih živih bića, odnosno koji ugrožavaju prirodnu sredinu u atmosferi. Do zagađenja vazduha dolazi kada se gasovi i mikroskopske čestice čađi i prašine oslobađaju u Zemljinu atmosferu. Ove čestice su veličine 2.5 mikrometra i zbog svoje sitnoće ostaju duže u vazduhu i uzrokuju mnogo veliku štetu. Kako bi se napravila predikcija kvaliteta vazduha u kineskom gradu Šangaju, pet godina su vršena snimanja. Rezultati će biti obrađeni u ovom seminarskom radu.

II. Baza podataka

Baza podataka obrađena u ovom izveštaju sadrži podatke koji su zabeleženi u Šangaju od 1.1.2010. do 31.12.2015. Baza podataka je dimenzija (52584, 17). Podaci koji se nalaze u bazi su beleženi na svakih sat vremena u prethodno navedenom vremenskom periodu. U navedenom vremenskom periodu je zabeleženo 52584 **uzoraka**. Svaki uzorak predstavlja merenje vršeno na svakih sat vremena. Merenjem je beleženo 14 parametara, odnosno 16 **obeležja** baze.

Obeležja koja se nalaze unutar baze možemo podeliti na numerička i kategorička obeležja. Broj numeričkih obeležja je 15, s tim da obeležja *No, year, month, day, hour i season* iako predstavljena brojem pripadaju kategoričkim obeležjima. Jedino kategoričko obeležje je *cbwd* i to je pravac vetra.

Obeležja "PM_Xuhui" i "PM_Nešto" se, po uslovu zadatka, uklanjaju u potpunosti. Nedostatak podataka unutar baze smo otkrili upotrebom funkcije *isnull* kod nekih obeležja. Obeležja kog kojih je primećen nedostatak uzoraka manji od 8% su "PM_US Post", "Precipitation", "Iprec", "PRES", "DEWP", "HUMI", "TEMP", "CBWD" i "Iws". Nakon ovoga iz baze podataka je izuzeto 20704 uzoraka, te baza podata trenutno sadrži 31880 uzoraka i 14 obeležja.

III. Analiza podataka

A. *Statističke veličine parametara*

Korišćenjem funkcije *describe* dobijamo informacije o statističkim parametrima obeležja. Pomoću datih podataka, imali smo uvid o parametrima: srednja vrednost (mean), standardna devijacija (std), minimalna (min) i maksimalna (max) vrednost, medijana (50%), kvartilni (25%) i interkvartilni (75%) opseg.

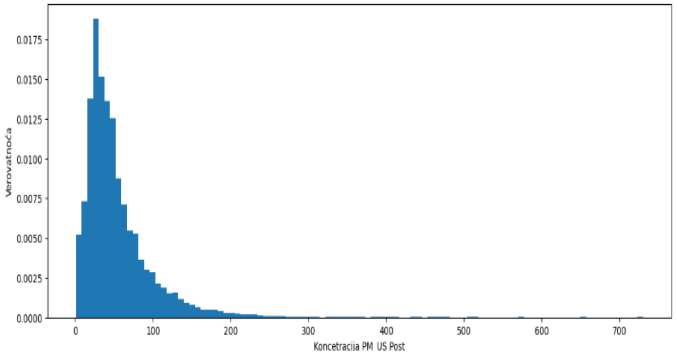
Na osnovu tabele (Tabela 1.) može se uočiti da se vrednosti čestica PM2.5 kreću u opsegu od 1 do 730. Primećujemo da je srednja vrednost veća od medijane, ali kako možemo uočiti da je velika razlika između interkvartalnog opsega i maksimalne vrednosti, možemo zaključiti da je ista očekivana.

Tabela 1: Statistički parametri obeležja PM_US Post

Veličina	Vrednost
Mean	53.41
Std	43.05
Min	1.00
25%	26.00
50%	42.00
75%	67.00
Max	730.00

B. *Raspodela koncentracija PM 2.5*

Na slici 1. je prikazana raspodela verovatnoće koncentracije čestica koja predstavlja izlaznu promenjivu.



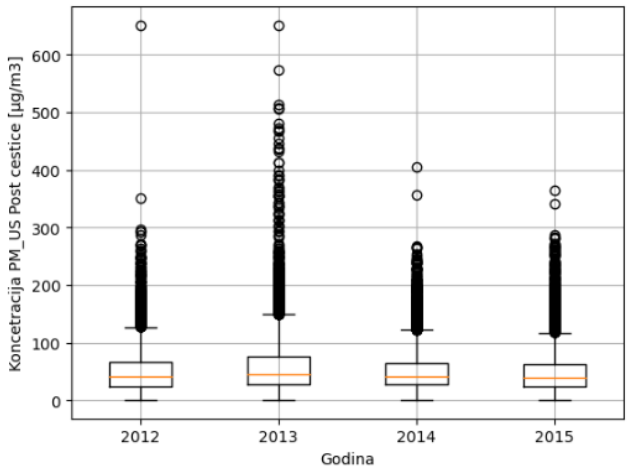
Sl. 1. Verovatnoća za koncentraciju PM 2.5

Na osnovu prikazanog histograma uočavamo da se najčešće koncentracija nalazi u opsegu [20 50] , dok primećujemo da u opsegu [50 300] koncentracija raste, a verovatnoća opada. Vrednosti su nakon 300 mikrometra/m3 beležene do 800 mikrometra/m3, ali su retki slučajevi.

[1]

C. *Vizuelizacija statističkih osobina obeležja*

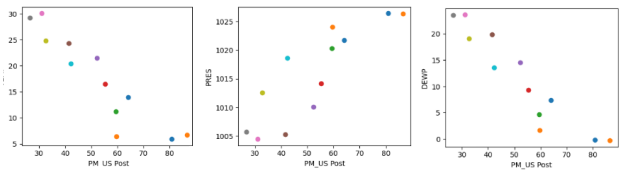
Korišćenjem funkcije *boxplot* iz bibiloteke *matplotlib.pyplot* dobija se grafički prikaz osnovnih statističkih veličina, što je prikazano na slici broj 2. Izostavljena je godina 2011. kako imamo podatke za samo jedan mesec. Na osnovu prikazanog, može se zaključiti da su se najzagađeniji sati dogodili u 2013. godini, ali da postoji mali broj merenja čije su vrednosti premašile 500 mg/m3. Unutar prikaza 2012. godine su prisutni autlajeri kada je zagađenje dostizalo 650 mikrometra/m3.



Sl. 2. Raspodela čestica PM 2.5 u periodu od četiri godine - grafički prikaz

D. *Bivarijantna analiza obeležja*

Pomoću funkcije *scatter* dobijen je grafik rasipanja koji je prikazan na slici 3. Na grafiku rasipanja je prikazan odnos izlaza sa ostalim obeležjima. Iscrtano je 12 tačaka koje predstavljaju svaki mesec u godini i srednju vrednost.



Sl. 3. Grafički prikaz zavisnosti koncentracije čestica PM2.5 od temperature, pritiska i vlage

Pri posmatranju grafika rasejanja primećena je linearna zavisnost na tri grafika.
Na prikazanim graficima može da se uoči pozitivna korelacija između čestica PM2.5 i vazdušnog pritiska. To znači da je pri većem pritisku, koncentracija čestica veća.
Nasuprot tome, može se primetiti negativna korelacija između PM2.5 čestica i temperature, kao i PM2.5 čestica i temperature rose. Negativna korelacija označava da se pri nižoj temperaturi i nižoj temperaturi rose koncentracija čestica povećava.

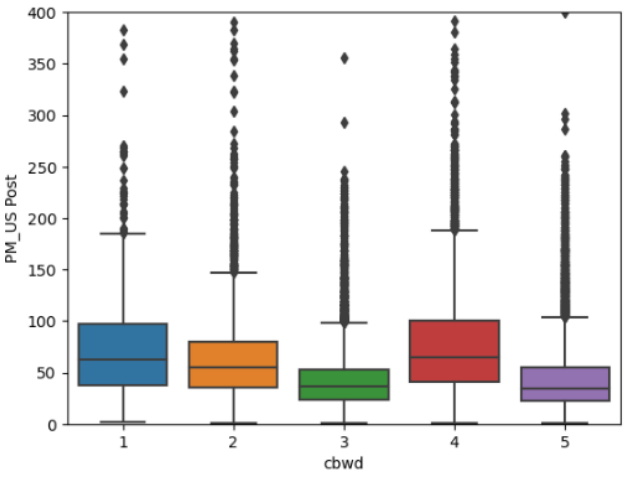
E. *Multivarijantna analiza obeležja*

Korišćenjem toplotne mape je moguće prikazari međusobnu zavisnost svih obeležja. Korelacije prikazane na slici broj 4. dobijamo pomoću funkcije *corr*.
Na pomenutoj slici se mogu uočiti negativne korelacije između obeležja vazdušnog pritiska i obeležja temperature rose, kao i vazdušnog pritiska i temperature. Blaga korelacija se uočava između čestica PM2.5 i temperature.

Sl. 4. Grafički prikaz korelisanosti između obeležja

F. *Vizuelni prikaz zavisnosti koncentracije čestica PM2.5 od pravca vetra*

Kategoričko obeležje pravca vetra (cbwd) je dato u vidu 5 različitih stringova. Radi dalje analize su mu dodeljene numeričke vrednosti 1, 2, 3, 4, 5, za obeležja cv (bez vetra), SW, SE, NW, NE, respektivno. Korišćenjem funkcije *boxplot* prikazan je grafik na slici broj 5.



Sl. 5. Grafički prikaz zavisnosti koncentracije čestica PM2.5 od pravca vetra (cbwd)

Posmatranjem prikazane slike možemo uočiti da je najveća koncentracija na boxplotu pod brojem 4, odnosno onom

koji predstavlja severozapadni pravac vetra. Kako najveći broj čestica nadolazi iz tog pravca, možemo pretpostaviti da se u severozapadnom delu grada nalazi industrijska zona ili neki vid zagađenja koji je izvor ovih čestica.

IV. Linearna regresija

Linearna regresija predstavlja metodu nadgledanog učenja koja koristi se za predviđanje vrednosti kontinualne izlazne promenljive uz pretpostavku da se ta vrednost može dobiti kao linearna kombinacija vrednosti ulaznih obeležja.

A. Podela podataka

Kako bismo testirali podatke na skupu koji nije skup obuke, potrebno je bilo izvršiti podelu na test skup koji čini 15% baze podataka, 15% validacioni skup i ostatak baze koji čine skup za obuku modela. Ukoliko ne bismo prethodno podelili podatke pre samog obučavanja, testirali bi sa već viđenim podacima i procena ne bi bila relevantna. Funkcijom *train_test_split* smo izvršili podelu.

B. Normalizacija obeležja

Izvršena je standardizacija obeležja pomoću funkcije *StandardScaler*. Normalizacijom se obeležja svode na iste ili slične opsege i obuka obeležja bude brža.

C. Funkcija za procenu tačnosti

Nakon implementacije modela koji predviđa vrednosti određene kontinualne promenljive potrebno je proceniti tačnost obučenog modela. Kako i dalja analiza bila olakšana, napravljena je funkcija mere_uspesnosti koja računa različite mere uspešnosti regresora, a koja će biti korišćena nakon svake obuke i testiranja. Unutar funkcije se izračunavaju: srednja kvadratna greška (mse), srednja apsolutna greška (mae), koren srednje kvadratne greške (rmse), R kvadrat skor (r2) i prilagođeni R kvadrat skor (r2_adj). Ova funkcija će takođe prikazati prvih osam pravih i predviđenih vrednosti izlazih obeležja.

D. Prva hipoteza

Prva hipoteza predstavlja predviđanje izlaza na osnovu parametara linearnog modela. Rezultati su prikazani unutar Tabele 2.

Tabela 2: Mere uspešnosti prvog modela

Veličina	Vrednost
mse	1585.8236
mae	27.1891
rmse	39.8224
r2	0.1707
r2_adj	0.1704

E. Druga hipoteza

Druga hipoteza predstavlja predviđanje izlaza na osnovu parametara linearnog modela o njihove međusobne interakcije. Rezultati su prikazani unutar Tabele 3.

Tabela 3: Mere uspešnosti drugog modela

Veličina	Vrednost
mse	1480.0639
mae	26.2397
rmse	38.4715
r2	0.2260
r2_adj	0.2244

F. Treća hipoteza

Treća hipoteza predstavlja predviđanje izlaza na osnovu parametara linearnog modela, njihove međusobne interakcije i njihovih drugih stepena. Veći broj parametara znači složeniji model. što znači da postoji opasnost od natprilagođenja. Rezultati su prikazani unutar Tabele 4.

Tabela 4: Mere uspešnosti trećeg modela

Veličina	Vrednost
mse	1454.9062
mae	25.9238
rmse	38.1432
r2	0.2391
r2_adj	0.2373

G. Selekcija obeležja

Korelisanost obeležja i visoka dimenzionalnost otežavaju proces obuke kod linearne regresije, te se preporučuje smanjenje dimenzionalnosti selekcijom ili redukcijom obeležja. Moguće je klasifikovati obeležja odabirom manje značajnih i uklanjanjem.

H. Regularizacione tehnike

Kako ni se pronašao kompromis između pristrasnosti i varijanse, koriste se Ridge i Lasso metode. Kod navedenih metoda je moguće podesiti nenegativni parametar alfa - regularizacioni parametar i broj iteracija. Ove metode takođe služe kako bi se ograničile procene koeficijenata. Najbolji rezultati su se pokazali kod modela Ridge regresije kada je alfa parametar bio postavljen na 1. Rezultati su prikazani u tabeli broj 4.

Tabela 4: Mere uspešnosti modela nakon selekcije

Veličina	Vrednost
mse	1456.1507
mae	25.9323
rmse	38.1595
r2	0.2385
r2_adj	0.2366

V. KNN KLASIFIKATOR

Metoda k najbližih suseda (engl. *k nearest neghbors - k*NN) predstavlja neparametarsku metodu klasifikacije. Ova metoda spada u grupu metoda kasnog učenja (engl: lazy learning), koje podrazumevaju odlaganje obrade uzoraka za obuku do trenutka kada treba klasifikovati neobeleženi uzorak.

A. *Priprema*

Dodeljene su labele vrednostima PM_US Post koje su numerički predstavljene kao 1, 2, 3 za bezbedno, nebezbedno i *opasno*, respektivno. Gde za koncentracije čestica do 55.4 važi da su bezbedne, za koncentracije između 55.5 i 150.5 važi da su nebezbedne, dok za koncentracije preko 150.5 važi da su opasne.
Kako se sada navedene klase nalaze u poslednjoj koloni, u *X* su pomoću funkcije *iloc* smeštena sva obeležja osim poslednje kolone, a u promenjivu *Y* je smeštena poslednja kolona.
Pomoću funkcije *train_test_split* izvršena je podela uzoraka na 15% za test finalnog klasifikatora i 85% ostalih uzoraka na kojima će se vršiti metoda unakrsne validacije sa 10 podskupova.

B. *Evaluacija klasifikatora*

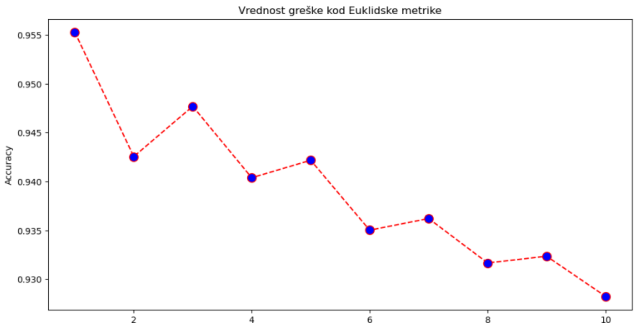
U cilju određivanja mera uspešnosti za svaku klasu, napisana je funkcija *Evaluation_Classifier.* Kao ulaz u funkciju unosi se matrica konfuzije. U matrici konfuzije pojedine vrste odgovaraju stvarnim vrednostima labela, dok pojedine kolone odgovaraju predviđenim vrednostima (oznaka klase koju je uzorku dodelio klasifikator). Na osnovu matrice konfuzije računaju se mere uspešnosti klasifikatora: tačnost, preciznost, osetljivost , specifičnost i F-mera.

C. *Određivanje optimalnih parametara metodom unakrsne validacije*

Metoda unakrsne validacije skup podataka za obuku se deli na približno 10 jednakih podskupova. Ono po čemu se odlikuje ova metoda je to što se svaki od uzoraka jednom javi kao test uzorak. Skup podataka bude bolje iskorišćen. Prosek ponašanja ovih 10 klasifikatora daje očekivano ponašanje na neočekivanim uzorcima.
Podelu na 10 podskupova smo izvršili pomoću funkcije *StratifiedKFold.* Vrednost parametra *k* označava koliko će suseda biti uzeto u obzir, te kroz petlje isprobavamo vrednosti od 1 do 10, kako bismo konačan model mogli da obučimo pomoću klasifikatora koji daje najbolje rezultate. Za metriku su izabrane *hamming* i *euclidean*.

D. *Analiza rezultata iteracije unakrsne validacije*

Nakon izvršenih iteracija najveću tačnost je pokazala metrika *euclidean*, za k=5, gde je uspešnost 95%. (slika 6.)



Sl. 6 Grafik uspešnosti kNN klasifikacije

Matrica konfuzije je prikazana na slici 7. Vrste prikazuju prave vrednosti, dok se u kolonama nalaze vrednosti koje su predviđene.

	Predviđene vrednosti		
Stvarne vrednosti	20426	577	0
	670	8959	85
	0	94	1069

Slika 7. Matrica konfuzije za parametar k=5

Na glavnoj dijagonali matrice konfuzije nalaze se ispravno klasifikovani uzorci. Primećujemo da se 670 puta dogodilo da se bezbedna koncentracija čestica klasifikuje kao nebezbedna, dok se 577 puta nebezbedna koncentracija čestica klasifikovala kao bezbedna. To može da ukazuje na su faktori koji utiču na koncentraciju čestica graničnog karaktera. Takođe, 94 puta se opasna koncentracija čestica klasifikovala kao nebezbedna, a 85 puta se nebezbedna koncenracija klasifikovala kao opasna. Ono što je odlično je što se nijednom nije dogodilo da se opasan vazduh klasifikuje kao bezbedan.

E. *Testiranje odabranog modela na test skupu*

Konačna faza je testiranje modela koji je nastao pomoću metrike Euklidske i klasifikatora 5 na test skupu koji je izvojen. Testiranje na kom bismo videli ponašanje našeg modela na test skupu kao i matricu konfuzije nije izvršeno ovaj put.