Analiza podataka i predikcija koncentracije PM2.5 čestica u vazduhu

Natalija Krsmanović, IN21/2018, krsmanovic.natalija99@gmail.com

I. Uvod

Tema izvještaja jeste analiza podataka o zagađenosti vazduha u gradu Čengdu u Kini. Glavni pokazatelj zagađenosti jeste koncentracija PM2.5 čestica koje imaju prečnik manji od 2.5 mikrometra i mogu se vidjeti samo uz pomoć mikroskopa. Nazivaju se još i finim česticama i nastaju sagorijevanjem fosilnih goriva. Posjeduju sposobnost da prodru duboko u pluća i da izazovu ili pogoršaju hronična oboljenja,kao i dovedu do prevremene smrti od kardiovaskularnih i pulmonalnih bolesti. Cilj analize jeste uočavanje uticaja pojedinih meteoroloških elemenata na količinu PM2.5 čestica u vazduhu i kreiranje modela za predikciju istih u budućnosti zarad smanjenja njihove koncentracije.

II. Opis baze podataka

Baza sadrži podatke o 52584 uzorka i 17 obilježja. U kategorička obilježja spadaju: redni broj mjerenja (No), godina (year), mjesec (month), dan u mjesecu (day), sat u danu(hour), godišnje doba (season) i pravac vjetra (cbwd). Numeričkim obilježjima pripadaju: koncentracija PM2.5 čestica na tri lokacije (PM_Caotangsi, PM_Shahepu, PM_US Post), temeprtura rose/kondenzacije-C^o (DEWP), vlažnost vazduha-% (HUMI), vazdušni ppritisak–hPa (PRES) ,temperatura-C^o (TEMP), kumulativna brzina vjetra-m/s (Iws), padavine na sat-mm (precipitation), kumulativne padavine-mm (Iprec). Jedan uzorak predstavlja izmjerene vrijednosti za koncentraciju PM2.5 čestica kao i ostalih navedenih obilježja u toku jednog sata.

III. Analiza podataka

Prilikom analiziranja baze izostavljeni su podaci o obilježjima ‘PM_Caotangsi’ i ‘PM_Shaheou’ i analizirana je samo količina PM2.5 čestica u okviru ‘PM_US Post’. Također izbačeno je i obilježje ‘No’ jer ne predstavlja obilježje od značaja za datu analizu.

A. *Nedostajući podaci*

Ispitivanjem je utvrđeno da se nedostajući podaci javljaju kod obilježja: ‘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%). S obzirom da koncentracije PM2.5 čestica nisu poznate za skoro polovinu uzoraka, njihovom dopunom bi se unijela velika količina grešaka u bazu. Nakon analize nedostajućih vrijednosti po godinama izbačeni su oni uzorci koji su imali nedostajuće vrijednosti za PM2.5 obilježje za 2010.,2011,2012 i 2013.-tu godinu. Podaci za 2014. i 2015.-tu godinu su dopunjeni sa prvom poznatom PM2.5 vrijednošću. Također izbačeni su i uzorci sa nedostajućim vrijednostima za sva obilježja [DEWP, HUMI, PRES, TEMP, cbwd, Iws] istovremeno. U tabeli 1 su prikazana obilježja koja i nakon spomenutih izmjena imaju nedostajuće vrijednosti.

Tabela 1: Prikaz obilježja sa nedostajućim vrijednostima

Naziv obilježja	**Broj uzoraka sa nedostajućim podacima**
Precipation	1223
Iprec	1223
Iws	12
HUMI	7
DEWP	4
TEMP	3

Sa pretpostavkom da se metereološki podaci ne mijenjaju drastično iz sata u sat tj. očekuju se slične vrjienosti obilježja susjedih uzoraka, nepoznate vrijednosti obilježja ‘HUMI’, ‘DEWP’, ‘TEMP’, ‘precipitation’, ‘Iws’ su dopunjena sa prvom prethodnom poznatom vrijednošću. Vrijednosti obilježja ‘Iprec’ su rezultat sabiranja količine padavine tog sata(precipation) sa ukupnom količinom padavina prethodnog sata (Iprec). Ukoliko nije bilo padavina tog sata, Iprec dobija nula vrijednost. Konačna modifikovana baza sadrži 28810 uzoraka i 14 obilježja.

B. *Kategorička obilježja*

Jedino kategoričko obilježje koje nije predstvaljeno numeričkim vrijednostima je pravac vjetra (cbwd). Pravac je označen putem kombinacija strana svijeta i zbog toga su

one preslikane na vrijednsoti uglova. Sjeveru je dodjeljena vrijednost od 0^o, na osnovu koje su dobijene ostale kombinacije. Oznaci pravca 'cv' koja označava promjenjivo ili mirno stanje dodijeljeno je 360^o.

C. *Statistička analiza obilježja*

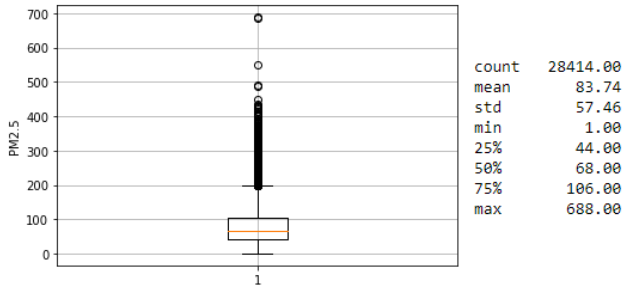
Tabela 2: Statističke mjere obilježja

	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

Iz Tabele 2 može se uočiti da su za obilježja 'DEWP', 'HUMI', 'PRES' i 'TEMP' medijana i srednja vrijednost približno jednake što ukazuje na to da ne postoji veliki broj autlajera sa izraženo velikom ili malom vrijednošću. Također ako se podaci posmatraju u kontekstu da li su ispravni i logični u realnom svijetu, za pomenute se može zaključiti da ispunjavaju taj uslov. Na osnovu koeficijenata spljoštenosti i asimetrije važi da sva obilježja osim 'PRES','Iws' i 'precipitation' imaju pozitivnu tj.desnu asimetričnu raspodjelu što znači da imaju autlajere za visoke vrijednosti.

Za padavine u toku sata(precipitation) su izražene visoke vrijednosti kod pojedinih uzoraka. S obzirom da su ove vrijednosti izmjerene u uzastopnim satima, može se pretpostaviti da su podaci ispravni, ali bi se radi sigurnosti trebalo posavjetovati sa nadležnim ukoliko je došlo do greške u procesu mjerenja ili unosa u bazu.

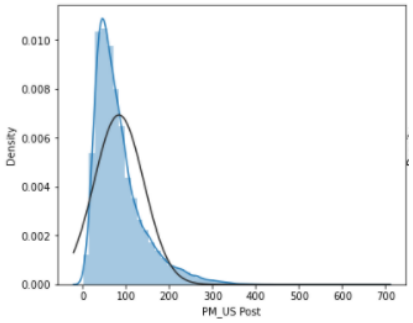
D. *Analiza obilježja PM2.5*



Slika 1: Boxplot obilježja PM2.5 i statistički parametri

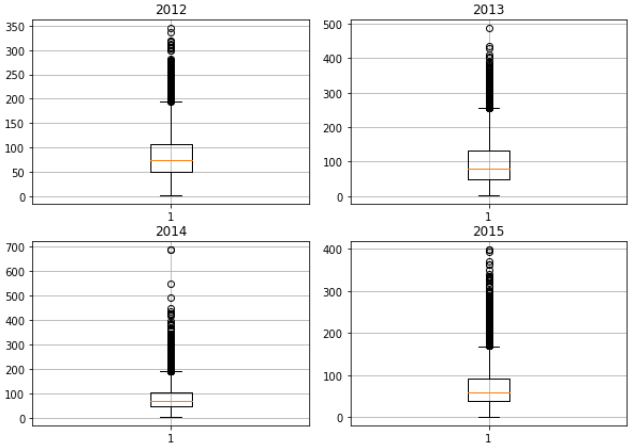
Iz prikazanih podataka se vidi da je interkvartilni opseg između 44ug/m3 i 106ug/m3 i tu se nalazi 50% uzoraka. Međutim postoji veliki broj autlajera koji su mnogo udaljeni od srednjih vrijednsoti ,gdje se pojavljuje čak i maksimalna vrijednost od 688 ug/m3, dok za niske vrijednosti ne postoje autlajeri. Baza sadrži par uzastopnih uzoraka sa vrijednostima približne maksimumu I zbog toga se također ne može sa sigurnošću zaključiti da li je u pitanju greška ili stvarna vrijednost. Odlučeno je da se ni jedan uzorak ne izbacuje niti mijenja nakon date analize.

Na osnovu koeficijenta spljoštenosti i asimetrije može se uočiti da ovo obilježje ima desnu asimetričnu raspodjelu i izdignuta je u odnosu na normalnu raspodjelu što je prikazano na slici 2. To je rezultat velikog dinamičkog opsega od 677 ug/m3, gdje se zapravo 50% uzoraka nalazi između 44 i 106 ug/m3 čime izazivaju iskrivljenost raspodjele.

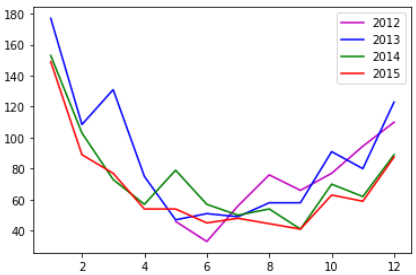


Slika 2: Raspodjela PM2.5 čestica naspram normalne raspodjele

1) *Analiza obilježja PM2.5 po godinama*



Slika 3: Boxplotovi obilježja PM2.5 po godinama

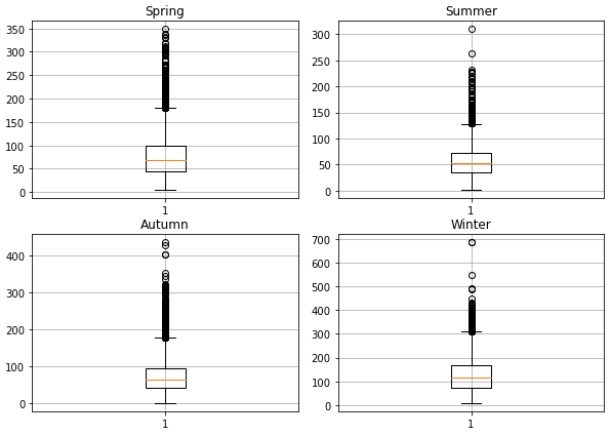


Slika 4: Srednje mjesečne vrijednosti PM2.5 čestica za svaku godinu

Na osnovu slike 3 uočava se da su interkvartilni opsezi za svaku godinu približno jednaki i da se koncentracija PM2.5 čestica iz godine u godinu ne mijenja drastično ali

blago opada. Najviša koncentrtacija je bila zabilježena 2014. godine kada su generalno prisutni autlajeri sa većim vrijednostima u odnosu na ostale godine. Analiziranjem slike 4 došlo se do zaključka da pored činjenice da su koncentracije slično raspoređene i blago opadaju sa godinama, najveće vrijednosti su tokom decembra i januara,a najniže tokom juna i jula. Iz tog razloga je izvršena analiza po godišnjim dobima.

2) Analiza obilježja PM2.5 po godišnjim dobima

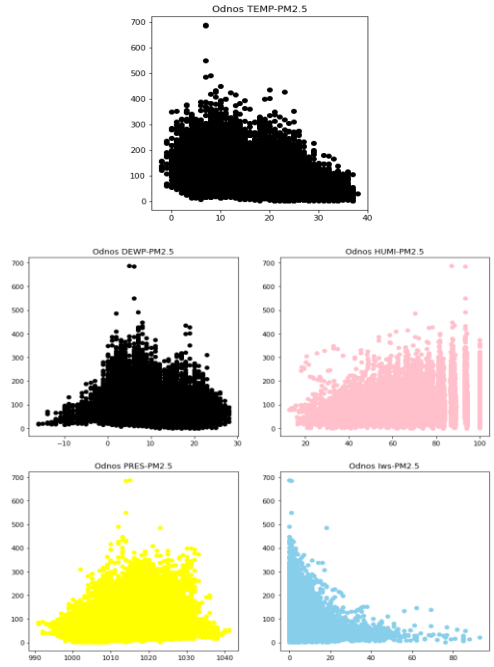


Slika 5: Boxplotovi obilježja PM2.5 po godišnjim dobima

Analizom slike 5 uočava se zavisnot koncenrtracije PM2.5 od godišnjih doba. 75% uzoraka ljeti ima vrijednost PM2.5 manju ili jednaku od 73 ug/m3, dok za 75% uzoraka zimi taj prag je 168 ug/m3. Također kada se porede maksimalne izmjerene koncenttracije PM2.5 tokom godišnjih doba, najviše su zapažene zimi a najniže ljeti. Razlog tome su toplane i grijanje na drva ili ugalj tokom zime kao dodatni izvor PM2.5 čestica.

E. Analiza zavisnosti PM2.5 od ostalih obilježja

Na osnovu prethodne analize po godišnjim dobima zaključuje se da postoji zavisnost između obilježja 'TEMP' i PM2.5.

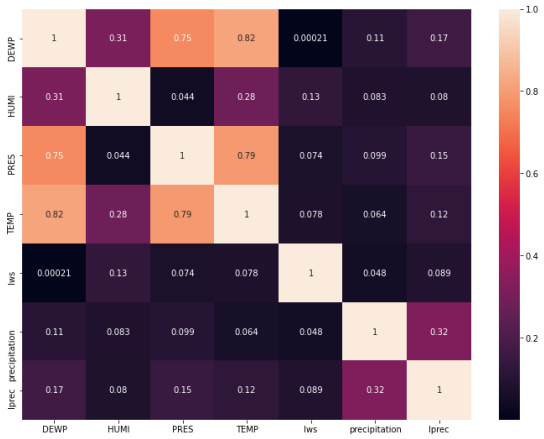


Slika 6: Prikaz odnosa PM.5 sa drugim obilježjima

Vrijednost korelacije između temepreature i PM2.5 čestica je 0.41. Sa slike 6 se vidi da je u pitanju negativna korelacija tj. zabilježene su najveće vrijednosti koncentracije zagađujućih čestica na nižim temepraturama(do 10 C^o) . Kako se temepratura povećava koncentracija PM2.5 opada i dostiže minimum za temeprature između 35C^o i 40C^o.

Također sa slike 6 se zapaža da obilježje PM2.5 naglo opada za temperature rose ispod 0C^o kao i za temperature više od 20 C^o. Slično ponašanje se uočava i u odnosu na obilježje 'PRES' gdje je količina PM2.5 čestica mala za granične vrijednosti vazdušnog pritiska. PM2.5 je pozitivno korelisano sa obilježjem 'HUMI'. Vidi se da i što su brzine vjetrova veće, PM2.5 čestice su slabije postojane.

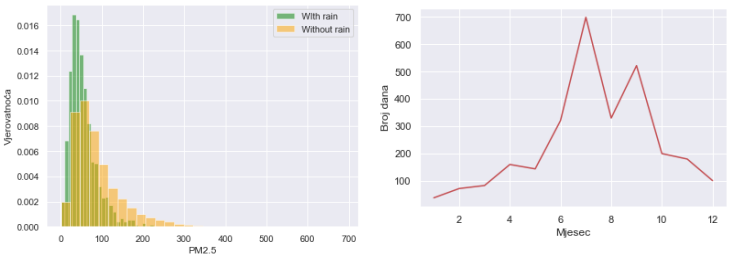
F. Analiza međukorelacije obilježja



Slika 7: Međukorelacija obilježja

Slika 7 pokazuje da su najviše korelisana obilježja 'TEMP' i 'DEWP' (0.82), 'TEMP' i 'PRES' (0.79),kao i 'PRES' i 'DEWP' (0.75). Kako vrijednost temeprature raste to dovodi do povećanja temeprature rose(pozitivna korelacija) ali i smanjenja vazdušnog pritiska (negativna korelacija).Posljedično korelacija između pritiska i temeprature rose je negativna. Kulumativna brzina vjetra(Iws) i padavine(precipitation) imaju malu korelaciju sa svim oblježjima.

G. Analiza uticaja padavina



Slika 8: Raspodjela kišnih i sušnih dana i proječne padavine po mjesecima

Radi lakše analize obilježje 'percipitation' tj. padavine su predstavljena kao binarno obilježje(1 kada pada kiša i 0 kada ne pada). Na slici 8 je prikazano da su tokom kišnih dana najčešće vrijednsoti za PM2.5 između 25 ug/m3 i 75

ug/m3, dok su te vrijednosti mnogo veće za dane bez padavina. Tada su izmjerene i ekstremne koncentracije čestica u vazduhu. Ovu interpretaciju potvrđuje slika sa padavina po mjesecima koja pokazuje da je najviše kišnih dana tokom ljetnih mjeseci a već je analizirano da je to period sa najnižim vrijednostima PM2.5 čestica.

IV. Linearna regresija

Početni skup podatka je radi obučavanja modela linerane regresije podijeljen na dva podskupa. To su podskup za obuku iz koga je izbačeno obilježje PM2.5 i podskup za testiranje koji sadrži samo vrijednosti za PM2.5 obilježje. 90% uzoraka pripada skupu za obučavanje a preostalih 10% testnom skupu. Radi što boljih rezultata prilikom obučavanja izvršena je standardizacija kojom se obilježja normalizuju tako da imaju srednju vrijednost 0 i standardnu devijaciju 1 čime se postiže ubrzanje obuke. Nakon toga izvršena je selekcija obilježja čiji je rezultat bio da se izbaci obilježje temepratura iz trening skupa što je i odrađeno.

A. Hipoteza: $y=b_0+b_1x_1+b_2x_2+...+b_nx_n$

Tabela 3: Rezultati predviđanja modela A

Mjera uspješnosti testa	Model bez regularizacije
Srednja kvadratna greška(MSE)	2372.91
Korijen srednja kvadratne greške(RMSE)	48.71
Srednja apsolutna greška(MAE)	35.48
R^2 skor	0.29
R^2 prilagođen	0.28

Za ovaj model nije rađen postupak regularizacije jer su težine koeficijenata relativno ravnomjerno raspoređene i nije došlo do natprilagođenja što pokazuju i dobijene greške koje su dosta velike i R2 skor koji je 0.29 i ukazuje na to da model nije puno bolji od modela koji bi za svaku vrijednost predvidjeo srednju vrijednost zavisnog obilježja.

B. Hipoteza:
$y=b_0+b_1x_1+b_2x_2+..+b_nx_n+c_1x_1x_2+c_2x_1x_3+...$

Tabela 4: Rezultati predviđanja modela B

Mjera uspješnosti testa	Model bez regularizacije	Lasso regularizacija	Ridge regularizacija
Srednja kvadratna greška(MSE)	2119.14	2126.09	2124.77
Korijen srednja kvadratne greške(RMSE)	46.03	46.10	46.09
Srednja apsolutna greška(MAE)	33.75	33.74	33.68
R^2 skor	0.36	0.36	0.36
R^2 prilagođen	0.36	0.36	0.36

Rezultati datog modela sa interakcijama između obilježja su bolji u odnosu na prethodni gdje sada obučen model pokriva oko 36% ukupne varijanse. Srednja apsolutna greška govori da dobijene predviđene vrijednosti u prosjeku odstupaju za oko 46 ug/m3 od stvarnih vrijednsoti. Odrađena je regularizacija radi postizanja kompromisa između pristrasnosti i varijanse. Sa ciljem snižavanja vrijednosti grešaka uvedeni su naredni modeli.

C. Hipoteza:
$y=b_0+b_1x_1+b_2x_2+..+c_1x_1x_2+c_2x_1x_3..+d_1x_1\text{^}2+d_2x_2\text{^}2+...+d_nx_n\text{^}2$

Tabela 5: Rezultati predviđanja modela C

Mjera uspješnosti testa	Model bez regularizacije	Lasso regularizacija	Ridge regularizacija
Srednja kvadratna greška(MSE)	1767.90	1754.92	1773.17
Korijen srednja kvadratne greške(RMSE)	42.04	41.89	42.10
Srednja apsolutna greška(MAE)	31.09	30.90	30.96
R^2 skor	0.47	0.47	0.46
R^2 prilagođen	0.47	0.47	0.46

Rezultati dobijeni sa datom hipotezom pokazuju da se srednja kvadratna greška dodatno smanjila na oko 42 ug/m3 i R^2 se približava jedinici koja bi bila optimalna vrijednost te mjere uspješnosti.

D. Hipoteza:
$y=b_0+b_1x_1+b_2x_2+..+c_1x_1x_2+c_2x_1x_3..+d_1x_1\text{^}2+d_2x_2\text{^}2+...+e_1x_1\text{^}3+..e_nx_n\text{^}3$

Tabela 6: Rezultati predviđanja modela D

Mjera uspješnosti testa	Model bez regularizacije	Lasso regularizacija	Ridge regularizacija
Srednja kvadratna greška(MSE)	1469.39	1466.41	1484.32
Korijen srednja kvadratne greške(RMSE)	38.33	38.29	38.52
Srednja apsolutna greška(MAE)	28.57	28.58	28.76
R^2 skor	0.56	0.56	0.56
R^2 prilagođen	0.55	0.55	0.54

Dobijeni rezultati pokazuju da je datim modelom obezbjeđena pokrivenost varijanse od 56% sa srednjom apsolutnom greškom od oko 38 ug/m3. R^2 skor je doveden do vrijednosti 0.55. Daljim povećvanjem stepena modela dobijaju se lošiji rezultati što implicira da se za konačan model izabere posljednji sa stepenom 3.