Zagađenost vazduha u Pekingu

Marko Vikić, BI 47/2017, vikic@uns.ac.rs

I. Uvod

Zagađenje vazduha postalo je jedan od glavnih i najočiglednijih problema za stanovnike većih gradova, naročito industrijskih centara i oblasti. Veliki uticaj na količinu zagađenja imaju i okolni gradovi čiji zagađivači pomoću strujanja vazduha i kiselih kiša dospijevaju od jednog grada do drugog. Takođe ogroman uticaj na zagađenje imaju industrijska postrojenja i ogroman broj vozila koji je prisutan u većim centrima. Na primjer, u Kini, smatra se da najsitnije čestice imaju ogroman uticaj na zdravlje odraslih i djece, i da dugotrajno izlaganje česticama, naročito $PM_{2.5}$ i PM_{10}, izaziva astmu, bolesti kardiovaskularnog sistema, pa čak može da nastupi i iznenadna smrt. Jedna od prednosti razvijanja modela koji se tiče zagađenja vazduha je predstavljanje hipotetičke situacije prije samog događaja.

II. Baza podataka

Baza podataka koja je analizirana sadrži 35064 uzorka koji predstavljaju vrijednosti koje se tiču kvaliteta vazduha u Pekingu (oblast Tiantan). Vrijednosti su mjerene i bilježene svakih sat vremena, svaki dan u periodu od 01.03.2013. do 01.03.2017. Obilježja koja su posmatrana su: redni broj mjerenja, godina, mjesec, dan u mjesecu, sat u toku dana, konecentracije čestica $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO, O_3 u vazduhu (mikrogram po metru kubnom), temperatura (^OC), pritisak (hPa), tačka rose (^OC), količina padavina (mm), pravac vjetra, brzina vjetra (m/s) i naziv stanice u kojoj je vršeno mjerenje. Ukupan broj obilježja u ovoj bazi je dakle 16, tačnije 15 obilježja i redni broj mjerenja. Bitno je istaći da se analizom obilježja može primijetiti da su koncentracije čestica $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO, O_3 u vazduhu, temperatura, pritisak, tačka rose, količina padavina i brzina vjetra numerička obilježja, dok su ostala obilježja kategorička. Odbačena su obilježja: redni broj mjerenja, pravac vjetra i naziv stanice. Prethodno izbačena obilježja su izbačena pod pretpostavkom da su nebitna za dalju analizu jer za svaki uzorak znamo da broj mjerenja ima jedinstvenu vrijednost i obilježje za naziv stanice u kojoj je vršeno mjerenje ima istu vrijednost i jer su kategorička. Provjerom vrijednosti obilježja za svaki uzorak ustanovljeno je da u obilježjima za koncentraciju čestica $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO i O_3 nedostaje određeni broj podataka.

Procenat nedostajućih podataka je iznad 1% i smatran je suviše velikim da bi se nedostajući podaci mogli totalno odbaciti.

Umjesto odbacivanja obavljeno je popunjavanje nedostajućih podataka sa prvom prethodnom validnom vrijednošću. Popunjavanje medijanom je prvobitno odrađeno ali je ustanovljeno da to nije najadekvatnija metoda jer bi se javile oscilacije u podacima. Pored prethodno navedenih obilježja nedostajao je i određeni broj vrijednosti za temperaturu, pritisak, tačku rose i brzinu vjetra. Procenat nedostajućih podataka u odnosu na ukupan broj podataka je ispod 1% tako da su ona odbačena. Ukupan broj odbačenih uzoraka je 20.

Glavni zadatak ove analize predstavlja predviđanje vrijednosti NO_2 u odnosu na ostala obilježja kvaliteta vazduha.

III. Analiza podataka

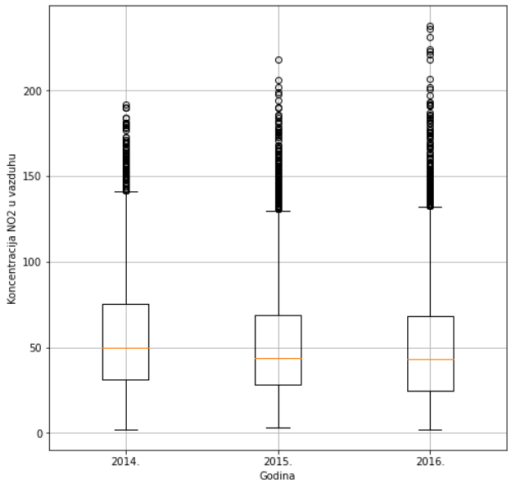
Na početku ovog poglavlja bitno je napomenuti da je od ukupno 35064 uzoraka uklonjeno 20 uzoraka, što je prihvatljivo.

A. Dinamički i interkvartilni opseg

Dinamički opseg donija se kao razlika maksimuma i minimuma vrijednosti podataka i on iznosi 239.0. Interkvartilni opseg je izračunat kao opseg u kome se nalazi 50% vrednosti podataka oko srednje vrednosti, odnosno kao razlika 75. i 25. percentila i iznosi 43.0. Donja granica opsega je vrijednost ispod koje se nalazi 25% podataka sa najnižim vrijednostima, dok je gornja granica vrijednost iznad koje se nalazi 25% podataka sa najvišim vrednostima. Vizuelizacija ovih podataka urađena je putem boxplota.

	NO2
count	35044.000000
mean	53.239284
std	32.001414
min	2.000000
25%	28.000000
50%	47.000000
75%	71.000000
max	241.000000

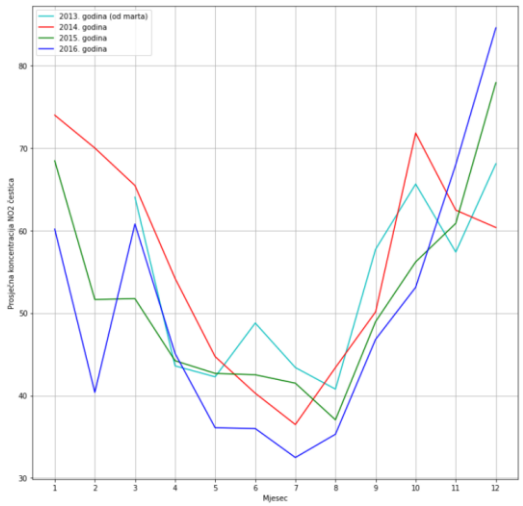
Tabela 1. Prikaz srednje vrijednosti, st. devijacije, min, max, 25. percentila, medijan, 75. percentila



Slika 1. Boxplotovi koji vizuelno prikazuju opsege vrijednosti koncentracije NO_2 u vazduhu po godinama

Sa slike 1. može se zaključiti da se medijani kreću od 40 do 50. Vrijednosti za godine 2013. i 2017. nisu prikazane zbog potpunog izostanka podataka za mjesece januar i februar 2013. godine i od marta do kraja godine za 2017. godinu. Opseg vrijednosti za 2014. je nešto veći od opsega za 2015. i 2016. godinu. Takođe, interkvartilni opseg za vrijednosti koncentracije čestica NO_2 u vazduhu je nešto veći od interkvartilnih opsega ostalih godina. Primijećeno je i postojanje velikog broja autlajera za svaku godinu, odnosno ekstremnih vrijednosti koje ne upadaju u predviđeni opseg.

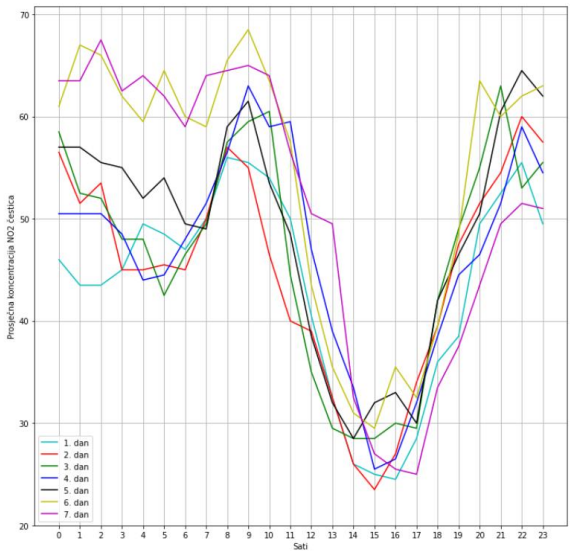
B. Pregled vrijednosti NO_2 u određenim periodima



Slika 2. Grafik prosječnih koncentracija čestica NO_2 po mjesecima

Na slici 2. prikazan je grafik prosječnih koncentracija NO_2 po mjesecima za svaku godinu, izuzev 2017. godine zbog nedostatka podataka od marta do decembra te godine. Sa grafika se može zaključiti da je prosječna koncentracija NO_2 jako osciluje. Uočeno je da od 2014. godine do 2016. godine nivo u januaru znatno opada što može biti posljedica popunjavanja nedostajućih podataka.

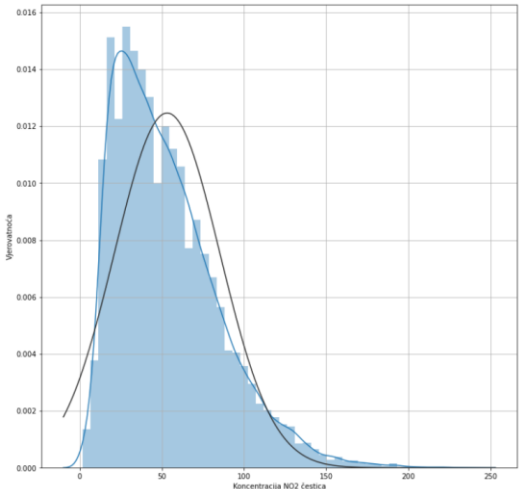
Takođe, prosječne koncentracije NO_2 za svaku godinu znatno opadaju tokom proljeća i ljeta. Veliki porast koncentracija je uočen tokom jeseni i zime, gdje dostižu najviše prosječne vrijednosti.



Slika 3. Grafik prosječnih koncentracija čestica NO_2 po satima u toku dana

Na slici 3. prikazan je grafik prosječnih koncentracija NO_2 u toku dana. Uzete su prosječne vrijednosti za istih 7 uzastopnih dana za svaki mjesec u periodu od 2013. do 2017. godine. Zajedničko za prosječne vrijednosti svih 7 dana je to da su tokom noći nivoi NO_2 viši u odnosu na period poslije podne. Blagi pik javlja se oko 10 časova ujutru. Nakon pada prosječnih vrijednosti u toku poslijepodnevnih časova dolazi do ponovnog porasta koncentracije NO_2 u večernjim časovima. U večernjim časovima su nešto niže vrijednosti od onih u 10 časova ujutru.

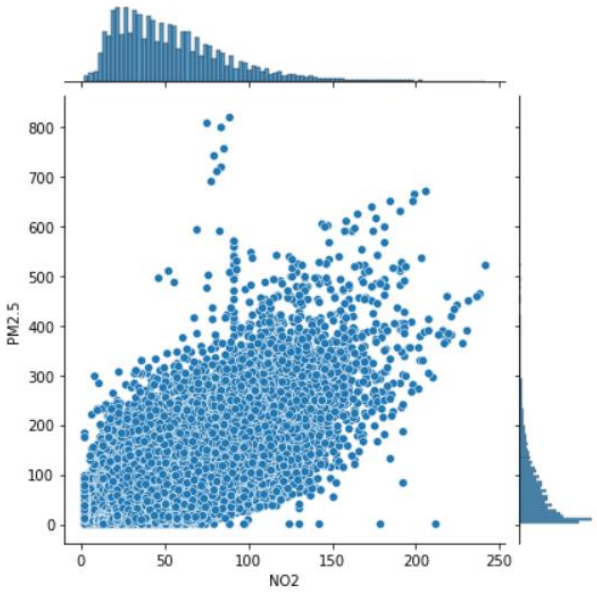
C. Analiza raspodjele u odnosu na normalnu raspodjelu



Slika 4. Raspodjela koncentracija NO_2 u odnosu na normalnu raspodjelu

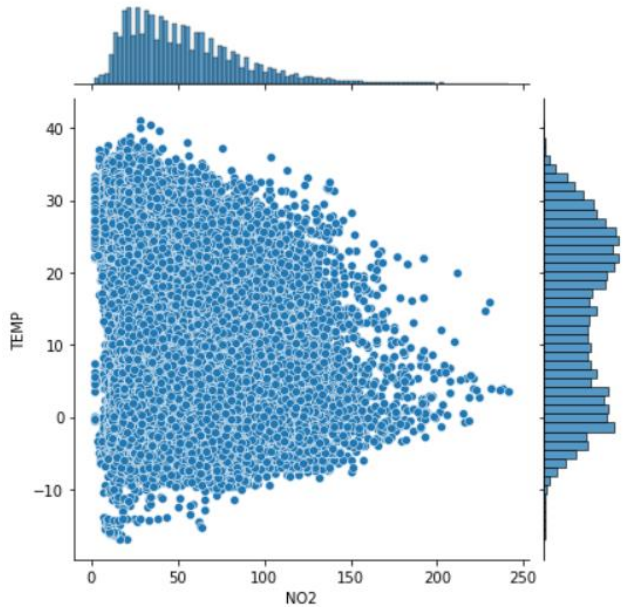
koncentracija čestica NO_2 u vazduhu tokom svih godina i normalna raspodjela približne. Javljaju se odstupanja kod raspodjele NO_2 od normalne na nižim vrijednostima NO_2.

D. Zavisnoti obilježja



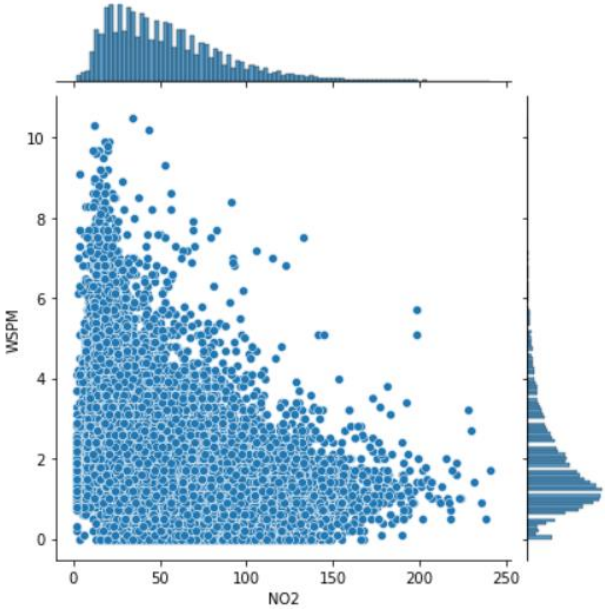
Slika 5. Vizuelizovana zavisnost koncentracija NO_2 i $PM_{2.5}$ čestica i raspodjele vrijednosti

Sa slike 5. može se uočiti da se vrijednosti čestica $PM_{2.5}$ i NO_2 ponašaju slično na nižim vrijednostima koncentracija. Na višim koncentracijama NO_2 postoji određeni broj uzoraka u kojima su koncentracije $PM_{2.5}$ takođe visoke, ali su ta poklapanja mala u odnosu na niže vrijednosti. Takođe se sa raspodjele za obilježje $PM_{2.5}$ jasno može utvrditi da su dominantno zastupljene niže vrijednosti.



Slika 6. Vizuelizovana zavisnost koncentracija NO2 čestica i temperature i raspodjele vrijednosti

Sa
NO_2 javljaju na nižim temperaturama, što je takođe rečeno u ovom poglavlju kada je vršen pregled vrijednosti NO_2 u određenim periodima. Tada je rečeno da se sa grafika može zaključiti da su prosječne koncentracije čestica NO_2 veće u jesenjem i zimskom periodu, kada su i temperature znatno manje od onih u ljetnjem periodu.



Slika 7. Vizuelizovana zavisnost koncentracija NO_2 i brzine vjetra i raspodjele vrijednosti

Sa slike 7. može se uočiti da su veće vrijednosti čestica NO_2 u vazduhu primjetnije u periodima kada je brzina vjetra jako mala, dok su tokom većih brzina vjetrova koncentracije NO_2 znatno manje.

IV. KREIRANJE I ODABIR MODELA

Kreiranje modela započeto je podjelom podataka na trening i test skupove podataka. Podaci koji su uzeti za test skup čine 10% ukupnih podataka u bazi. Ostalih 90% podataka uzeto je za treinranje modela. Prvi model koji je kreiran je kreiran na osnovu osnovnog oblika linearne regresije i odrađena je evaluacija modela da bi se prikazale vrijednosti srednje kvadratne i apsolutne greške (MSE i MAE), kao i vrijednosti R2 skora. Vrijednosti su date u tabeli 2.

	Vrijednosti
MSE	296.26
MAE	12.70
RMSE	17.21
R2	0.707
R2 adjusted	0.707

Tabela 2. Vrijednosti za prvi model

Vrijednost za MSE nakon kreiranja prvog modela su visoke. Nakon te konstatacije urađena je selekcija obilježja svih preostalih obilježja. Za granicu *p* vrijednosti uzet je broj 0.001. Vrijednosti svih *p* vrijednosti su date u tabeli 3.

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1305.2703	179.712	-7.263	0.000	-1657.513	-953.028
year	0.6709	0.090	7.451	0.000	0.494	0.847
month	0.1936	0.033	5.805	0.000	0.128	0.259
day	0.0489	0.011	4.402	0.000	0.027	0.071
hour	0.3479	0.015	23.016	0.000	0.318	0.378
PM2.5	0.0631	0.003	19.253	0.000	0.057	0.070
PM10	0.0617	0.002	24.771	0.000	0.057	0.067
SO2	0.2333	0.006	39.132	0.000	0.222	0.245
CO	0.0078	0.000	51.663	0.000	0.008	0.008
O3	-0.2270	0.002	-102.568	0.000	-0.231	-0.223
TEMP	0.7484	0.025	29.783	0.000	0.699	0.798
PRES	-0.0126	0.019	-0.649	0.516	-0.050	0.025
DEWP	-0.5138	0.018	-28.515	0.000	-0.549	-0.479
RAIN	-0.3354	0.125	-2.676	0.007	-0.581	-0.090
WSPM	-4.2707	0.095	-44.845	0.000	-4.457	-4.084

Tabela 3. Vrijednosti za prvi model

Uočene su vrijednosti *p* koje su iznad postavljenje granice od 0.001. U narednom koraku uklonjena su obilježja za pritisak vazduha i količinu padavina iz test i trening skupa. Nakon toga urađena je standardizacija obilježja da bi se dobile nove vrijednosti obilježja za sve uzorke i te vrijednosti su u normalnoj raspodjeli oko 0.
Ponovo je urađena selekcija obilježja da bi se potvrdile vrijednosti *p*. Utvrđeno je da su sve *p* vrijednosti 0.000.

Ponovljena je ista obuka modela sa osnovnim oblikom linearne regresije ali ovaj put sa standardizovanim obilježjima. Dobijene vrijednosti za MSE, MAE i R2 su identične vrijednostima iz prethodnog modela. Potom je urađena Ridge i Lasso regresija da bi se dobili manji koeficijenti. Nakon Lasso regresije vrijednosti za MSE, MAE i R2 su ostale iste ali su koeficijenti manje oscilovali i imali su maksimalnu vrijednost od -4.29, za razliku od koeficijenata prije Ridge i Lasso regresije koji su išli i do -13.87. Model nakon Lasso regresije uzet je kao trenutno najbolji model. Nakon toga određena je matrica korelacija svih dostupnih obilježja i grafički prikazana na toplotnoj mapi. Toplotna mapa prikazana je na slici 8. Uočene su visoke vrijednosti korelacije između obilježja $PM_{2.5}$, PM_{10} i CO te je odlučeno da je potrebno uraditi obuku modela uzimajući u obzir parove obilježja. Nakon obuke modela putem linearne regresije sa drugačijom hipotezom zaključeno je da su dobijene povoljnije vrijednosti za MSE, MAE i R2 skor. Vrijednosti su prikazane u tabeli 4.

Slika 8. Toplotna mapa korelacija svih dostupnih obilježja

	Vrijednosti
MSE	221.80
MAE	11.01
RMSE	14.89
R2	0.780
R2 adjusted	0.780

Tabela 4. Vrijednosti za treći model

Nakon toga su urađene Ridge i Lasso regresije ali su rezultati ostali isti. Nakon toga obučen je model putem linearne regresije ali ovaj put dodavanjem kuba obilježja. Vrijednosti koje su dobijene su date u tebeli 5.

	Vrijednosti
MSE	152.67
MAE	8.96
RMSE	12.36
R2	0.849
R2 adjusted	0.847

Tabela 5. Vrijednosti za četvrti model

Vrijednosti koje su dobijene su do sada najbolje i ovaj model je uzet kao najbolji. Koeficijenata je bilo znatno više jer je urađeno kubiranje obilježja. Vrijednosti tih koeficijenata su takođe bile veće. Urađena je Ridge i Lasso regresija sa vrijednostima *alpha*=5 za Ridge i *aplha*=0.001 za Lasso. Model nakon Lasso regresije smatra se najboljim modelom.

V. Literatura

Pisani i video materijali sa vježbi iz predmeta Prepoznavanje oblika, Fakultet tehničkih nauka, Novi Sad

Air Quality of Beijing and Impacts of the New Ambient Air Quality Standard, Wei Chen , Fusheng Wang, Guofeng Xiao , Kai Wu and Shixuan Zhang