Analiza kvaliteta vazduha u Pekingu

Jovana Bogojević, IN32/2017, bogojevic.in32.2017@uns.ac.rs

I. UVOD

U ovom izveštaju predstavljena je analiza podataka koji se odnose na kvalitet vazduha u Pekingu za period od marta 2013. godine do februara 2017. godine. Kao glavni grad Narodne Republike Kine, Peking je poznat i kao grad sa velikom zagađenosti vazduha koja se menjala tokom godina.

Analizom dostupnih podataka iz baze mogu se oučiti pravilnosti i zavisnosti između obeležja na osnovu kojih se primenom različitih tehnika može predvideti vrednost nekog od obeležja u budućnosti.

II. OPIS BAZE PODATAKA

Baza sadrži 35.064 uzorka i 18 obeležja. Jedan uzorak u bazi predstavlja izmereni kvalitet vazduha u jednom satu tokom određenog dana u mesecu. Numerička obeležja su: redni broj merenja, godina, koncentracija čestica PM2.5, PM10, SO_2, NO_2, CO i O_3, temperatura, pritisak, tačka rose, količina padavina i brzina vetra. Preostala obeležja - mesec, dan u mesecu, sat u toku dana, pravac vetra i stanica gde je izvršeno merenje, su kategorička. Tipovi obeležja u bazi uzimaju vrednosti iz tri skupa: int64, float64 i object.

KOREKCIJE U BAZI PODATAKA

S obzirom na to da je obeležje koje se odnosi na redni broj merenja jedinstveno za svaki uzorak i činjenicu da su sva merenja izvršena u stanici Tiantan, iz baze će biti izostavljena ova dva obeležja tokom dalje analize. U bazi su takođe sadržane i nedostajuće vrednosti.

Naziv atributa	Broj nedostajućih vrednosti	Procenat
PM2.5	677	1.93%
PM10	597	1.70%
SO2	1118	3.19%
NO2	744	2.12%
CO	1126	3.21%
O3	843	2.40%
TEMP	20	0.06%
PRES	20	0.06%
DEWP	20	0.06%
wd	78	0.22%
WSPM	14	0.04%

TABELA 1: Nedostajuće vrednosti po obeležjima

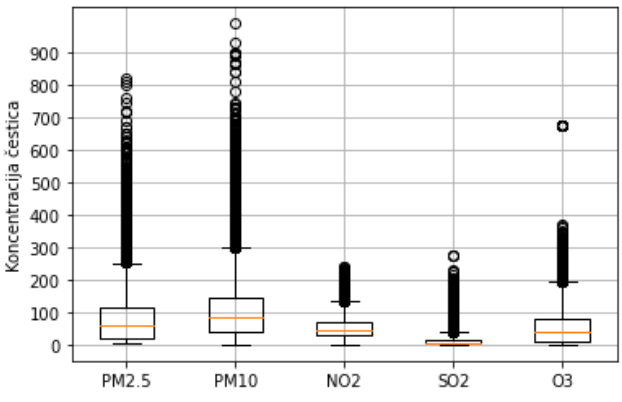
U ovoj tabeli prikazana su sva obeležja koja imaju nedostajuće vrednosti kao i njihov procenat. Obeležja koja imaju manje od 1% nedostajućih vrednosti su: temperatura, pritisak, tačka rose, brzina i pravac vetra. Iz razloga što je procenat vrednosti koje nedostaju vrlo nizak za data obeležja, svi uzorci koji imaju nedostajuće vrednosti za neko od navedenih obeležja će biti izostavljeni iz baze.

Nedostajuće vrednosti ostalih obeležja (čiji je procenat veći od 1%) biće zamenjene poslednjim validnim vrednostima. *Primer*: ukoliko nedostaje vrednost za koncentraciju O_3 za neki sat tokom dana, ta vrednost će biti zamenjena vrednošću O_3 izmerene za prvi prethodni sat za koji postoji izmerena vrednost. Vrednosti se dopunjavaju na opisani način jer je mala verovatnoća da će se koncentracija O_3 naglo promeniti u toku jednog sata u odnosu na prethodnu izmerenu vrednost.

Nakon primene prethodnih korekcija, dimenzije baze su: 34.980 uzoraka i 16 obeležja.

III. ANALIZA BAZE PODATAKA

Analizom statističkih parametara dobijeni su sledeći rezultati – merenja u bazi započinju u martu 2013. godine, a završena su u februaru 2017. godine. Merenja su vršena svakog dana u mesecu u razmaku od jednog sata. Najniža izmerena temperatura za dati period iznosi -16.8 °C, najviša 41.1 °C, a najčešće izmerena vrednost temperature iznosila je približno 14 °C, što je u skladu sa klimom. Na osnovu analize prosečne količine padavina po mesecima dolazi se do rezultata da najveća količina padavina se izruči u letnjim mesecima kao i početkom jeseni, dok su početkom i krajem godine padavine minimalne, približne 0 *mm*.



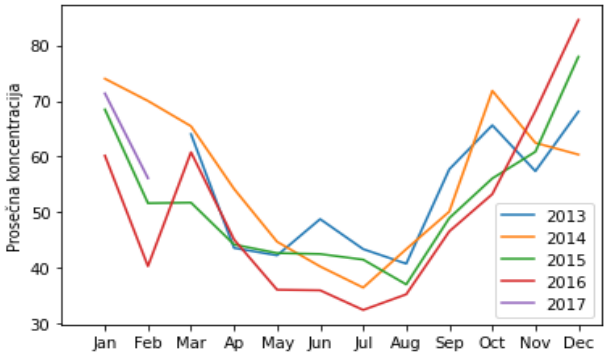
SLIKA 1: Koncentracija čestica PM2.5, PM10, NO2, SO2 i O3 za period od 2013. do 2017. prikazana putem boxplot-ova

Na osnovu analize dobijenih raspodela može se primetiti da sva obeležja imaju vrednosti koje su netipične, odnosno *outlier-e*. Na primer, najveća izmerena koncentracija čestice PM2.5 iznosila je 821 μg/m³, najniža 3 μg/m³, međutim 50% vrednosti nalazi se u opsegu od 22-114 μg/m³. Kada su u pitanju *outlier-i* obeležja O3, odnosno koncentracija ozona u vazduhu koji je u nižim slojevima atmosfere sastavni deo gradskog smoga , svi se javljaju u 2013. godini u mesecu aprilu. Može se uočiti da najčešće vrednosti svih čestica ne prelaze 150 μg/m³.

Čestica čije prosečne vrednosti odstupaju od vrednosti ostalih jeste ugljen-monoksid čiji se interkvartilni opseg kreće od 500 do 1600 μg/m³. Najveća zabeležena vrednost obeležja CO iznosila je 10.000 μg/m³ , izmerena sredinom decembra 2016. godine i početkom 2017. godine.

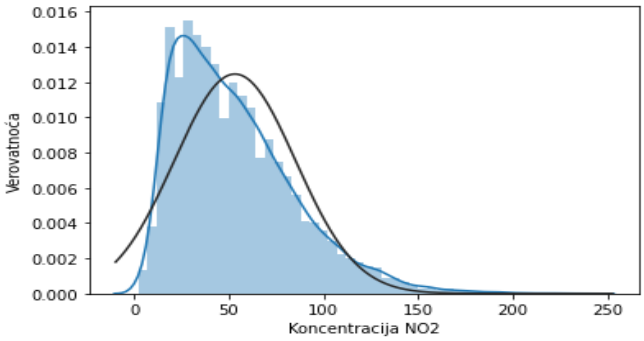
KONCENTRACIJA AZOT-DIOKSIDA

Najveća količina azot-dioksida nastaje pri radu elektrana i motornih vozila koje za svoj rad koriste tečno gorivo. Analizom podataka koji se odnose na koncentraciju ove čestice u vazduhu može se uočiti da su najčešće vrednosti između 28 i 71 μg/m³, dok medijana iznosi oko 53 μg/m³.



SLIKA 2: Prosečna koncentracija NO_2 po mesecima svake godine

Najveće vrednosti azot-dioksida zabeležene su početkom i krajem godine, a u letnjim mesecima, konkretno u julu, koncentracija azot-dioksida je značajno niža i dostiže svoj minimum. Ovakav trend je verovatno posledica sistema grejanja u Pekingu zimi koji uključuje i sagorevanje uglja, što utiče na zagađenje vazduha.

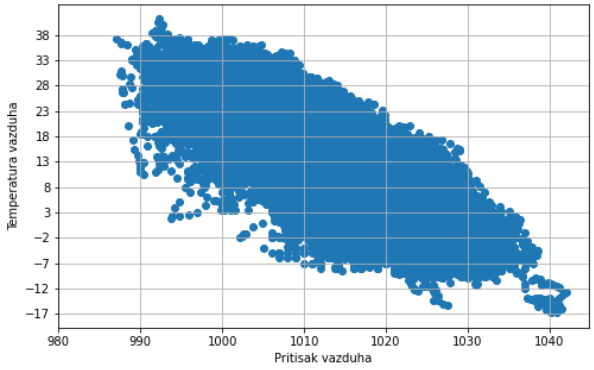


SLIKA 3: Raspodela vrednosti NO_2

Na osnovu prikazanog grafikona, dolazi se do zaključka da je mala verovatnoća da je koncentracija čestice NO_2 preko 150 μg/m³. Takođe se može primetiti odstupanje od normalne raspodele, što potvrđuju i koeficijenti spljoštenosti i asimetrije, čije vrednosti redom iznose 1.40 i 1.07. Budući da su oba koeficijenta pozitivna, raspodela je viša i pomerena u levu stranu u odnosu na normalnu raspodelu obeležja NO_2.

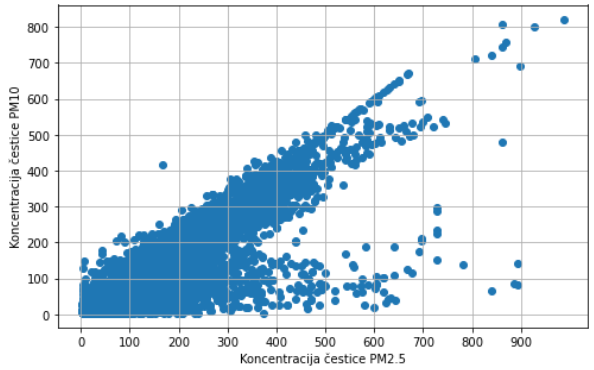
ZAVISNOSTI IZMEĐU ATRIBUTA

Na osnovu *SLIKE 4.* može se konstatovati visoka negativna korelacija (u iznosu od -0.84) između temperature i pritiska vazduha. To bi značilo da visoka temperatura podrazumeva nizak pritisak vazduha, stoga smanjenjem temperature, pritisak vazduha se povećava.



SLIKA 4: Zavisnost između pritiska i temperature vazduha

Najveća pozitivna korelacija je između atributa PM2.5 i PM10. Visoka koncentracija čestice PM2.5 ujedno označava i visoku prisutnost čestice PM10 u vazduhu.



SLIKA 5: Zavisnost između PM2.5 i PM10

IV. PREDVIĐANJE VREDNOSTI NO_2

Za predviđanje vrednosti obeležja NO_2 biće iskorišćen algoritam linearne regresije. Pre kreiranja modela linearne regresije, koji će predviđati koncentraciju NO_2 u vazduhu na osnovu dostupnih podataka o drugim atributima, neophodno je prvo pretvoriti vrednosti kategoričkog obeležja koje se

odnosi na pravac vetra u numeričke vrednosti, a nakon toga podeliti skup podataka, na osnovu kojih će se vršiti obuka modela, na trening i test podatke.

Nakon inicijalizacije i obuke modela linearne regresije sa osnovnom hipotezom, za mere uspešnosti modela, odnosno za srednju kvadratnu grešku, srednju apsolutnu grešku, koren srednje kvadratne greške, R^2 score i prilagođeni R^2 score, dobijeni su sledeće vrednosti:

```
Mean squared error:  288.8657972725798
Mean absolute error:  12.622501702574565
Root mean squared error:  16.996052402619256
R2 score:  0.7191364677603969
R2 adjusted score:  0.7190025787060654
```

SLIKA 6: Vrednosti mera uspešnosti

U cilju daljeg poboljšanja modela, primenjuje se metoda selekcije koja podrazumeva odabir bitnijih obeležja iz skupa svih raspoloživih obeležja. Nakon selekcije unazad, iz početnog skupa obeležja za obuku modela izostavljena su obeležja koja se odnose na količinu padavina i pritisak vazduha na osnovu p-vrednosti, odnosno značaja obeležja, koja je bila veća od 0.001 i potom je izvršena obuka modela linearne regresije sa osnovnom hipotezom. S obzirom na to da nije došlo do poboljšanja mera uspešnosti, za inicijalizaciju sledećeg modela iskoristiće se i prethodno izbačena obeležja.

NAJUSPEŠNIJI MODEL

Na osnovu matrice korelacije uočene su zavisnosti između obeležja, što je dobra osnova da se primeni interakcija između njih. Nakon primene linearne regresije sa hipotezom koja uključuje interakciju među obeležjima, odnosno kreiranja obeležja prvog, drugog i trećeg stepena, dobija se novi skup podataka za obuku. Obukom modela nad novodobijenim podacima primećuje se poboljšanje mera uspešnosti, posebno srednje apsolutne greške koja je iznosila oko 9,11.

U cilju daljeg poboljšanja modela upotrebljena je regularizaciona tehnika – ridge regresija. Kao ulazni parametri za ridge regresiju korišćeni su podaci dobijeni primenom interakcije među obeležjima, a kao vrednost regulariozacionog parametra uzeta je vrednost 10. Rezultati mera uspešnosti nakon primene navedenih tehnika prikazani su na *SLICI 7*.

```
Mean squared error:  186.76953269847078
Mean absolute error:  9.09128968651952
Root mean squared error:  13.66636501409467
R2 score:  0.8184044246022893
R2 adjusted score:  0.8135782198821062
```

SLIKA 7: Rezultati za mere uspešnosti nakon primene interakcije među obeležja, potom regulariozacione tehnike

Na osnovu dobijenih rezultata za mere uspešnosti, kao i novodobijenih koeficijenata uz obeležja, prethodno opisani model je izabran kao najbolji za predviđanje vrednosti obeležja NO_2 na osnovu vrednosti ostalih obeležja.