Koncentracija PM10 čestica u vazduhu

Sara Popov, IN 41/2017, popov.sara@uns.arc.rs

I. UVOD

Zagađenost vazduha je jedan od većih problema današnjice, a najviše zahvata velike industrijske gradove. Kina, sa svojom prestonicom Pekingom, se sa ovim problemom bori već decenijama. Pored jeftinije teške industrije i postavljanje ekonomskog faktora kao bitnijeg prioriteta od očuvanja životne okoline, postoje još razni faktori koji utiču na zagađenost vazduha. Rešavanja problema zagađenosti vazduha je od krucijalne značajnosti, jer je usled povećane koncentracija otrovnih čestica u vazduhu narušeno zdravlje opšte populacije.

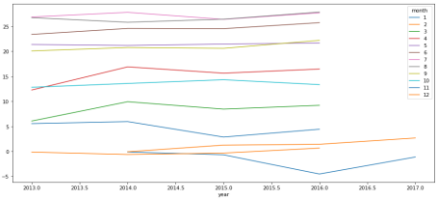
Ovaj rad se bavi analizom podataka vezanih za koncentraciju PM10 čestica u vazduhu i to kako će se ona menjati u skladu sa promenama određenih parametara od kojih ona zavisi.

Analiziranje ovih podataka moguće je uočiti pravilnosti kada je koncentracija PM10 čestica povećana, kao i koji parametri su u korelaciji i samim tim utiču na povećanje ili smanjenje PM10 čestica.

II. BAZA PODATAKA

Baza sadrži 35064 uzoraka, a 18 obeležja, od kojih su 2 kategoričke (pravac duvanja vetra i stanica u kojoj je vršeno ispitivanje) i 16 numeričkih (redni broj merenja, godina, mesec, dan, sat, koncentracija PM2.5, PM10, SO2, NO2, CO, O3 čestica u vazduhu (izraženi u mikrogramu po metru kvadartnom), zatim temperatura, pritisak, tačka rose, količina padavina, brzina vetra). Jedan uzorak predstavlja koncentraciju gore navedenih čestica tačno određenog datuma u određeno vreme, pod određenim vazdušnim i meteorološkim uslovima.

Obeležja koja su izbačena su stanica u kojoj je vršeno ispitivanje i redni broj ispitivanja. Obeležje stanica je izbačena jer je merenje vršeno samo u jednoj stanici i na osnovu te informacije jasno je da to obeležje neće imati uticaja u dalja istraživanja, dok je redni broj merenja predstavlja za svaki uzorak drugačiji broj i takođe ne utiče na dalja istraživanja. Uzorci kod kojih nedostaje 30 ili više posto obeležja se smatraju nepogodnim za dalje ispitavnje i na osnovu toga su izbačeni. Kod ostalih uzoraka koji imaju nedostajuće vrednosti korišćene su različite metode za njihovo popunjavanje. Za koncentraciju čestica i hemijskih elemenata u vazduhu nedostajući podaci su zamenjeni medijanom tog obeležja, dok su za meteorološke parametre, kao što je na primer temperatura zamenjeni medijanom tog meseca kojem pripada nedostajuća vrednost. Ova metoda je primenjena, jer je na osnovu prvobitne intuitivne ideje da se srednja vrednost na primer temperature drastično razlikuje u januaru i u avgustu, tako da se dobijaju približnije vrednosti pretpostavljenim stvarnim vrednostima ako se iskoristi ova metoda. Isto je urađeno za ostale meteorološke parametre. Uzorci koji imaju nedostajuće vrednosti za PM10 obeležje su izbačeni, jer to obeležje će biti korišćeno u regresiji i potrebno je znati tačne vrednosti predviđenog obeležja, kako bi bilo poznata tačnost modela. Zbog ovih korekcija baza se svela na 34635 uzoraka, odnosno broj uzoraka se smanjio za 429, dok je broj obeležja, a samim tim i dimenzionalnost problema smanjena za 2, odnosno sad je 16.



Slika 1. Prikaz promene temperature kroz mesece(y-osa) i godina(x-osa).

III. ANALIZA OBELEŽJA I NJIHOVIH KORELACIJA

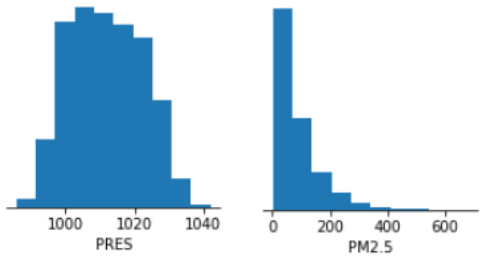
Pomoću statističke analize će vrednosti obeležja,kao i njihova prosečna odstupanja biti prokomentarisana. U bazi je moguće videti da se analiziraju godine od 2013. Do 2017. I za to obeležje, kao i za mesece, dane i sate nije potrebno gledati prosečna odstupanja, jer su unapred poznati opsezi u kojima će se javiti.

TABELA 2 : DINAMIČKI I INTERKVARTALNI OPSEG ATRIBUTA

Atribut	Dinamički	IQR
PM2.5	678	91
PM10	997	109
SO2	292	19
NO2	268	47
CO	9900	1400
O3	415	72
TEMP	57.3	20.1
PRES	56.1	17
DEWP	63.8	23
RAIN	72.5	0
wd	15	7
WSPM	11.2	1.3

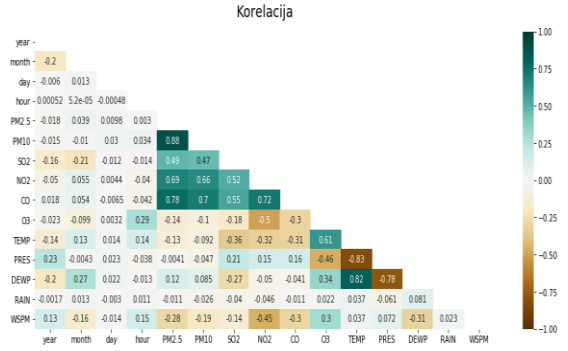
Na osnovu prethodno upoređenog dinamičkog i IQR opsega izveden je zaključak da je merodavniji IQR opseg. Dinamički opseg vrednosti koncentracije PM10 čestica u vazduhu je 997, dok se 50% posto njih nalaze u opsegu 109, slična raspodela se može uočiti i za PM2.5, SO2,

NO2, CO, 03, RAIN. Gde je posebno zanimljivo kod količine padavine(RAIN) da 50% tih vrednosti zapravo ima vrednost 0, dok dinamički opseg daje samo informaciju da se vrednosti kreću u intervalu veličine 72.5. Dok su prethodne raspodele bile asimetrične kod atributa pritisak(PRES) se javlja drugačija situacija, odnosno kod nje su vrednosti ravnomerno raspoređene po celom opsegu.



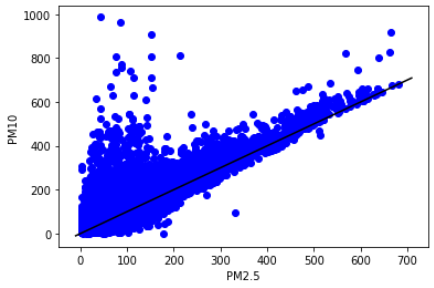
Slika 2. Prikaz raspodele za vazdušni pritisak i koncentraciju PM2.5 čestica, odnosno poređenje dve različite raspodele.

Izlaz linearne regresije je obeležje PM10. PM10 čestice se mogu naći u prašini i dimu, one su izuzetno male i mogu ući u grlo i pluća i tako izazivaju kašalj, povećana vrednost ovih čestica u vazduhu je posebno opasna po ljude koji imaju hroničnih problema sa otežanim disanjem ili koji imaju određene bolesti srca. Kada je vrednost ovih čestica u vazduhu do 40 (mikrograma po metru kubnom), tada ne predstavljaju opasnost i smatra se da je kvalitet vazduha što se tiče ovog parametra dobar, međutim kada vrednosti pređu 240 kvalitet vazduha se smatra veoma opasnim i može značajno uticati na prethodno pomenutu osetljivu grupu. Potrebno je primetiti da se u podacima nad kojim se vrši istraživanje vrednost PM10 obeležja kreće do 999, što predstavlja lošiju vrednost za više od 4 puta već opasne vrednosti. Srednja vrednost ovog obeležja je 109, dok je standardna devijacija 91.57, a 50% uzoraka se nalazi u opsegu 109, dok je dinamički opseg 997, iz toga se vidi da kao i kod obeležja PM2.5 raspodela nije ravnomerna po celom intervalu. Takođe, potrebno je zaključiti da su vrednosti PM10 čestica u vazduhu u dalekoj meri više zastupljene nego što bi bilo poželjno.

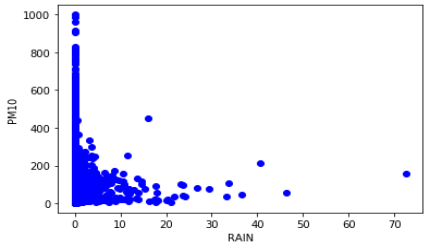


Slika 3. Matrica korelacije

U matrici korelacije se jasno vidi da je obeležje koje je izlaz regresije u najvećoj korelaciji, 0.88, sa PM2.5 obeležjem, sledeće po redu obeležje sa kojim se javlja relativno velik stepen korelacije je CO, 0.77, a odmah posle njega NO2, 0,69. Obeležje PM10 je u najmanjoj korelaciji sa obeležjem koje predstavlja količinu padavine (rain), zatim vazdušnim pritiskom (PRES) i tačkom rose (DEWP).



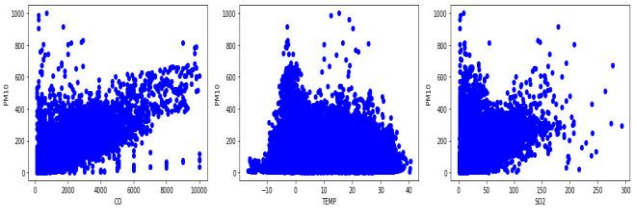
Slika 4. Prikaz korelisanosti obeležja PM2.5 i PM10



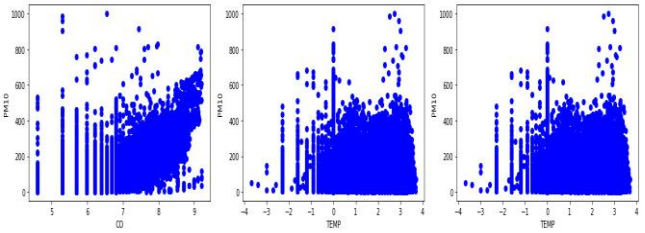
Slika 5. Prikaz korelisanosti obeležja PM2.5 i RAIN

Korelacija ostalih obeležja su takođe prikazana kroz matricu korelacije i već spomenuta korelacija između PM2.5 i PM10 je najveća, takođe su visoko korelisane i temperatura i tačka rose, kao i količina čestica PM2.5 u vazduhu i CO. Dok je najmanji stepen korelacije između temperature i vazdušnog pritiska, kao i vazdušnog pritiska i stepena.

U obeležjima CO, TEMP i SO2 se javljaju uzorci kod kojih dolazi do velikih odstupanja, odnosno vrednosti tih uzoraka se smatraju netipičnim za datu raspodelu obeležja. Ove netipične vrednosti mogu da naruše model linearne regresije, koji će biti objašnjen u narednim poglavljima. Iz datog razloga će se vrednosti za ova obeležja logaritmovanjem preskalirati na manje vrednosti, gde se logaritam koristi da bi se razmera između obeležja sačuvala, a da bi se velike vrednosti eksponencijalno smanjile.



Slika 6. Prikaz postojanja outliera kod obeležja, redom, CO, TEMP i SO2



Slika 7. Prikaz obeležja CO, TEMP i SO2 nakon logaritmovanja

Nakon logaritmovanja opseg na kojim se nalaze uzorci se smanjio za svako obeležje. Gde na primer kod SO2 opseg je bio od 0 do 300, a nakon logaritmovanja vrednosti je od 0 do 6. Što će se kasnije u procesu linearne regresije ispostaviti da je pogodno za model.

Obeležje wd predstavlja kategoričko obeležje i kao takvo ne može se koristiti za potrebe linearne regresije, nego su njegove vrednosti pretvorene u numeričke. Ovaj postupak je urađen tako što je kategoriji obeležja sa najmanjom vrednošću za PM10 dodaljena vrednost 0 i tako redom kako se nastavljaju i kateogorije po sortiranim srednjim vrednostima za PM10.

IV.LINEARNA REGRESIJA

Nakon analize i obrade podataka je određeno predviđanje PM10 atributa pomoću univarijantne linearne regresije. Nad prethodno opisanim podacima je urađena standardizacija, kako bi se ubrzala konvergencija metode opadanja gradijenta. Skaliranje je izvršeno na isti opseg vrednosti, kako bi bile sprečene da se potencijalno odluči o izlaznoj vrednosti na osnovu manjeg broja obeležja, koje variraju u većem opsegu vrednosti. Takođe, priprema podataka koja je dodatno uređena je selekcija obeležja. Kako je pokazano u prethodnom poglavlju ne utiču sva obeležja u podjednakoj meri jedna na druge, pa tako ni na obeležje PM10. Selekcijom obeležja unazad gde je p vrednost su izbačena obeležja CO, hour, month, wd. Vrednost za p je izabrana prvo empirijski, a nagon toga je posmatrana bila RSS greška i izbacivanjem jednog po jednog obeležja kao i podešavanjem parametra p na druge vrednosti utvrđeno je da sa ovim paramtrom i ovim izbačenim obeležjima se dobija najmanja RSS greška. Izbacivanjem obeležja došlo je do smanjenja dimenzionalnosti problema, samim tim i bržeg izvršavanja. Ostala obeležje su ostavljena, jer je u interesu ostaviti obeležja koja u određenoj meri utiču na obeležje koje je izlaz linearne regresije, jer se tako dobija tačnije predviđanje.

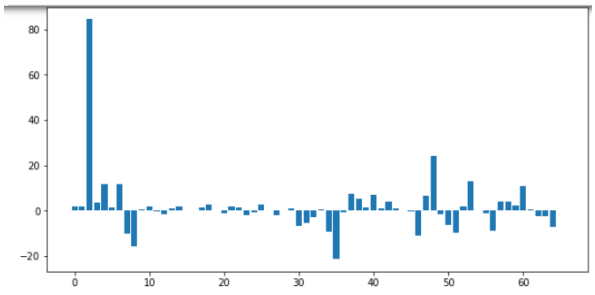
Uzorci su podaljeni u dva dela, od kojih je 90% korišćeno za obuku linearnog modela, a ostalih 10% za testiranje uspešnosti.

Linearnom regresijom sa hipotezom y=b0+b1x1+...+bnxn, odnosno kada na izlaz modela utiču samo ostala obeležja koja su uzeta za predviđanje i prethodno pripremljeni podaci.

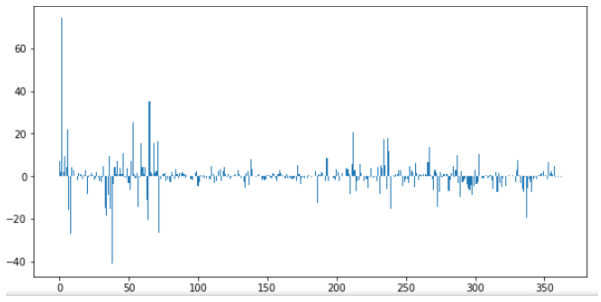
Primenom ove hipoteze dobija se model kod kojeg je MSE 1596, ono što nama više govori o grešci MAE 23.5 i R^2 0.81. Na osnovu MAE znamo da će u proseku ovaj model pogrešiti u predviđanju za 23.5, kako znamo da je opseg vrednosti za PM10 997, zaključujemo da model ne odstupa previše od stvarne vrednosti.

U narednom modelu su u obzir uzete i interakcije među obeležjima, odnosno korišćena je sledeća hipoteza y=b0+b1x1+b2x2+...+bnxn+c1x1x2+c2x1x3+.... Sa tim da se koristi prethodno redukovana baza, koja je i korišćena u prethodnom modelu regresije, odnosno baza u kojoj smo prethodno izbacili obeležja selekcijom obeležja unazad. Sa ovakvim modelom regresije dobijaju se sledeće greške: MSE 1376 i MAE 22.3, dok je R^2 sada 0.83. Potrebno je primetiti da je ovakav oblik linearne regresije uspesniji odnosno, da kada se uzmu u razmatranje I interakcije među obeležjima popravi se razlika između stvarne i predviđene vrednosti izlazne promenljive. Što se može objasniti time da su neka obeležja visoko korelisana i dosta utiču jedna na drugu, a i dodavanjem interakcija povećava se broj obeležja, samim tim i količina informacija, koje utiču na izlaznu vrednost i samim tim na uspešnost modela, međutim proces predviđanja se usložava i usporava.

Povećavanjem parametra degree, odnosno podešavanja do kog stepena će biti formirana polinomijalna obeležja. Inicijalna vrednost je 2 i za taj slučaj su u prethodnom primeru pokazane performanse. Povećanjem ovog parametra se povećava broj obeležja, a to prouzrokuje bolje rezultate modela, ali se povećava i kompleksnost modela. Ako je parametar degree inicijalizovan na vrednost 3, onda je MSE 1163, MAE 20.7, dok je R^2 0.86. Kada je degree postavljen na 4, odnosno kada je određeno da će obeležja ići do četvrtog stepena dolazi do lošijih rezultata, R^2 je u tom slučaju 0.75, dok je MSE 2033 i MAE 21. Kako se dalje povećava stepen dolazi do sve lošijih rezultata. Razlog zašto se ovo dešava je zato što je došlo do preobučavanja modela i model ne samo što je uočio pravilnosti u podacima nego je uzeo u obzir i njihov šum i tako izvršena linearna regresija bi davala odlične rezultate kada bi ih merili na testnom skupu. Međutim, kako to nije poenta, nego predviđanje na neviđenim podacima, ova mogućnost se odbacuje i smatra se da je model davao najbolje rezultate kada je vrednost stepena do kog se kreiraju polinomijalna obeležja je 3.



Slika 8. Prikaz koeficijenata kada je stepen do kog se formiraju polinomijalna obeležja postavljen na 2



Slika 9. Prikaz koeficijenata kada je stepen do kog se formiraju polinomijalna obeležja postavljen na 3

Kao što je prikazano na prethodne dve slike, kada je stepen postavlje na 2 opseg koeficijenata je manji nego kada je postavljen na 3, takođe ovde je vizualno prikazano da je broj obeležja drastično povećan.

Naredne dve metode linearne regresije koje će biti objašnjene su modeli kod kojih se vrši regularizacija. Regulrizacija je korišćena sa ciljem da se u isto vreme i isprati trend u podacima i spreči variranje vrednosti koje se predviđaju na testnom skupu, odnosno skupu na kom nije vršena obuka.

Ridge regresija je korišćena sa hiperaparametrom alpha,

odnosno f-ja cene, na 6 dok je degree 3. Ove vrednosti hiperparametra su usledile nakon prvobitnog postavljanja empirijski zadanih vrednosti i onda su isprobavane različite kombinacije i upoređivanje rezultata ta kombinacija je dala najbolji R^2 i najmanje greške. Rezultati Ridge regresije su poprilično slični kao i rezultati sa regresijom koja uvažava interakcije. Razlog zašto je korišćena Ridge regresija je jer ona uvodi alpha, odnosno vrednost koja otežava izbor modela sa velikim koeficijentima, kako bi šanse da će doći do natprilagođavanja bile umanjene.

Lasso regresija je korišćena sa hiperparametrom alpha 0.001 i stepenom 3, što dovodi do približno istih rezultata kao Ridge regresiju i sa interakcijama. Razlika je što Lasso u odnosno na Ridge vrši direktnu realizaciju selekcije obeležja.

V. ZAKLJUČAK

Zagađenost vazduha je problem koji direktno utiče na ljudske živote time što narušava zdravlje celokupne populacije. Na osnovu toga, neophodno je posmatrati ovaj problem kao ozbiljan problem današnjice i tretirati ga kao takvog. Povećan broj PM10 čestica u vazduhu utiče kako na zdrave tako i na hronične bolesnike. Pokazano je da je PM10 u najvećoj pozitivnoj korelaciji sa PM2.5 česticama, CO i NO2 i da kako se ova obeležja menjaju tako će se i srazmerno menjati PM10. A da ne zavisi od dana u mesecu, sata kao ni pravca duvanja vetra, zato ova obeležja nije potrebno uzeti u razmatranje prilikom predviđanja linearnom regresijom. Baza u kojo su upisani podaci merenja kao i samo obeležje PM10 daje relativno dobre rezultate, R^2 0.86, predviđanjem linearnom regresijom, što ne znači da se model ne može unaprediti i dobiti precizniji rezultati.