Analiza Koncentracije $PM_{2.5}$ Čestica u Šangaju

Đorđe Stanković, IN13-2018,
stankovictab@gmail.com

I. Uvod

Ovaj izveštaj pokriva analizu meteoroloških podataka sakupljenih u Šangaju od 2010. do 2016. radi utvrđivanja uzroka pojave i povećanja koncentracije aerosolnih partikularnih čestica u vazduhu, konkretno PM2.5. Navedene čestice su proizvod "prljave" industrije i imaju jako negativan uticaj na zdravlje ljudi. Čestice nastaju spaljivanjem fosilnih goriva u fabrikama, toplanama i vozilima, ili prilikom požara. Najveće koncentracije su uglavnom beležene oko veoma prometnih saobraćajnih raskrsnica i u industrijskim zonama. Koncentracija se meri u mikrogramima po metru kvadratnom, gde su vrednosti do 35 $\mu g/m^3$ podnošljive, vrednosti od 55 $\mu g/m^3$ do 150 $\mu g/m^3$ nezdrave, a vrednosti do 500 $\mu g/m^3$ opasne po život.

PM2.5 čestice mogu dovesti do srčanih i disajnih problema, najviše zbog njihove veličine od 2.5 mikrometara, i lakoće, što im omogućuje da duže ostanu u vazduhu i lakše uđu u krvotok.

Analizom pojave PM2.5 čestica u odnosu na meteorološke promene možemo utvrditi iz kog predela se čestice javljaju, koliko im treba da nestanu, da li su otporne na različite klimatske promene, i slično. Takođe, analizom možemo napraviti matematički model, odnosno regresor, koji će nam za nove uzorke u budućnosti pokazati da li možemo da očekujemo povećanu koncentraciju čestica ili ne.

II. Baza podataka

Za analizu je korišćena baza podataka od 52584 uzoraka i 17 obeležja, dobijena od konzulata Američke ambasade u Šangaju, i pravljena od 2010. do 2016. godine.

Svaki uzorak baze predstavlja časovno merenje kvaliteta vazduha, i to po sledećim (preimenovanim) obeležjima :

- PM - Koncentracija PM2.5 čestica u $\mu g/m^3$.
- Dew Temp - Temperatura rose u stepenima C.
- Humidity - Vlažnost vazduha u procentima.
- Pressure - Vazdušni pritisak u hPa.
- Temperature - Temperatura u stepenima C.
- Wind Direction - Pravac duvanja vetra.
- Wind Speed - Brzina vetra u m/s.
- Precipitation - Časovna količina padavina u mm.
- TotalPrec - Ukupna količina padavina u mm.

Sva navedena obeležja osim Wind Direction su numeričke prirode. Wind Direction i sva obeležja vezana za datum beleženja uzorka (godina, mesec, dan, sat i godišnje doba) su kategorička. Radi jednostavnosti i lakše upotrebe regresora, obeležje Wind Direction će imati numeričke vrednosti: 1 za severozapadni pravac, 2 za severoistočni, 3 za jugoistočni, 4 za jugozapadni i 0 za miran ili promenljiv vetar.

III. Dopuna Nedostajućih Vrednosti

Baza podataka ima nedostajuće vrednosti za sva obeležja osim za obeležja namenjena za datum merenja. Naime, obeležje PM ima 18545 nedostajućih vrednosti, obeležja vezana za padavine imaju 4009, dok ostala imaju do 28 nedostajućih vrednosti.

Nakon prolaženja kroz podatke možemo da uočimo da se koncentracija PM2.5 čestica merila tek od 28. Decembra 2011. u 18h, odnosno, tek od 17443. uzorka. Odatle, iz baze podataka možemo da izbacimo sve uzorke do navedenog, jer nam ne znače u analizi, odnosno, imali bi veliku rupu u nalazima. To znači da sada imamo 35142 uzoraka u bazi podataka.

Obeležjima Dew Temp, Humidity, Pressure, Temperature, Wind Direction i Precipitation, a i obeležju PM u preostalim uzorcima, možemo da zamenimo nedostajuće vrednosti sa validnim vrednostima iz prethodnog uzorka, odnosno od prethodnog časa, što je bolja praksa nego uzimanje srednje vrednosti obeležja, pogotovo kod obeležja za koncentraciju PM čestica.

Obeležje Wind Speed predstavlja kumulativnu vrednost za brzinu vetra do sledeće promene pravca duvanja vetra, slično kao i obeležje TotalPrec, koje predstavlja kumulativnu količinu padavina, do prestanka padavina. Nedostajuće vrednosti ovih obeležja takođe možemo da zamenimo vrednošću od prethodnog uzorka, uzimajući u obzir da nedostajuće vrednosti za, na primer ukupne padavine, ne znače da u tim trenucima nije bilo padavina.

IV. Analiza podataka

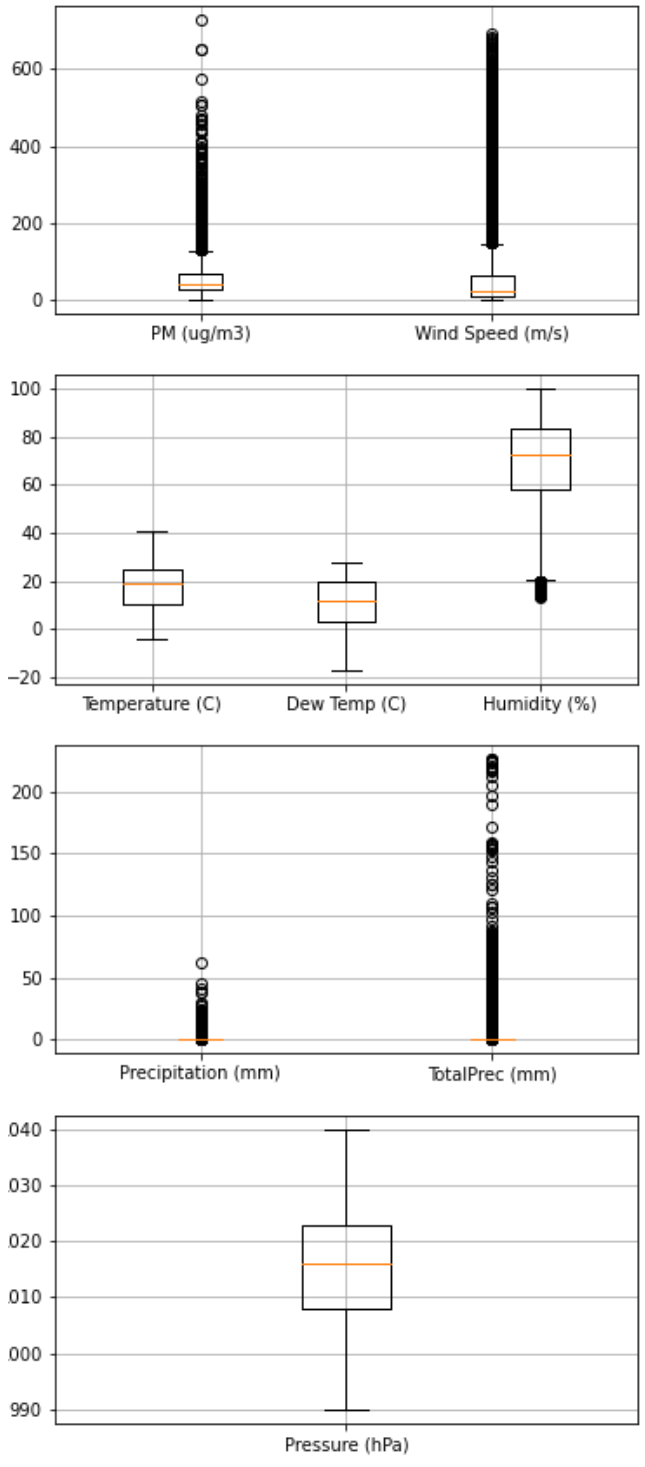
Obeležja Year, Month, Day, Hour i Season imaju očekivane statistike za obeležja vezana za datum. Obeležje PM ima maksimum od 730 $\mu g/m^3$ što ukazuje na postojanje outlier-a, koje ne izbacamo, već ostavljamo za analizu. Dew Temp, Humidity, Pressure, Temperature i Precipitation imaju očekivane statistike.

Obeležja Wind Speed i TotalPrec imaju kumulativne vrednosti, pa im maksimumi ne predstavljaju pogrešna merenja (na primer 1110 m/s za brzinu vetra ili 226 mm za padavine), odnosno statistike su očekivane.

	PM (ug/m3)	Dew Temp (C)	Humidity (%)	Pressure (hPa)	Temperature (C)	Wind Speed (m/s)	Precipitation (mm)	TotalPrec (mm)
count	35142.00	35142.00	35142.00	35142.00	35142.00	35142.00	35142.00	35142.00
mean	52.71	11.45	69.90	1015.87	17.56	49.45	0.15	1.00
std	42.34	9.67	17.80	9.03	9.18	71.64	1.06	7.31
min	1.00	-17.00	13.09	990.00	-4.00	0.00	0.00	0.00
25%	26.00	3.00	58.17	1008.00	10.00	6.00	0.00	0.00
50%	41.00	12.00	72.96	1016.00	19.00	20.00	0.00	0.00
75%	67.00	20.00	83.60	1023.00	25.00	62.00	0.00	0.00
max	730.00	28.00	100.00	1040.00	41.00	691.00	61.60	226.40

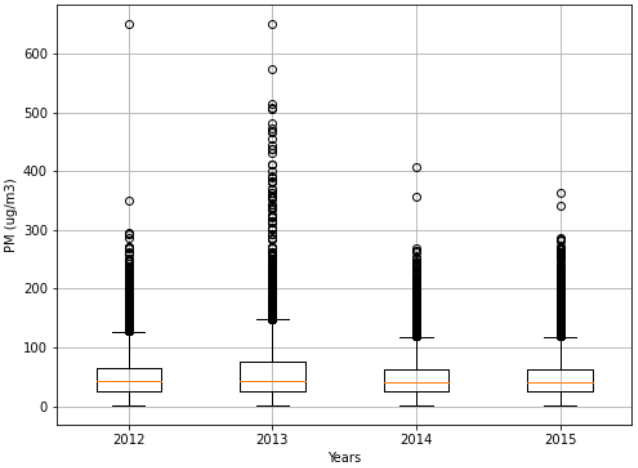
Slika 1 - Osnovne statistike obeležja.

Statistike obeležja možemo bolje predstaviti koristeći boxplot dijagrame (Slika 2). Obeležjima PM i Wind Speed su jasno izraženi outlier-i zbog čestih naglih skokova u njihovim vrednostima - outlier-i obuhvataju do šest puta veći opseg vrednosti nego što obuhvata interkvartilni opseg boxplot-ova. Kod obeležja vezana za padavine (Precipitation i TotalPrec), interkvartilni opseg je jako mali, ali i očekivan, zbog velikog broja dana bez padavina. Outlier-i za obeležja Wind Speed i TotalPrec su očekivana, jer je reč o kumulativnim obeležjima.



Slika 2 - Boxplot reprezentacija obeležja.

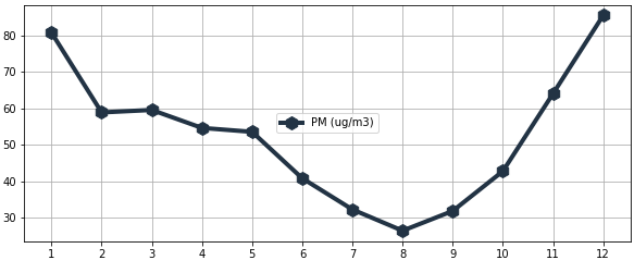
Pošto imamo podatke za četiri kontinualne godine, možemo za njih uporediti koncentracije PM čestica.



Slika 3 - Upoređivanje koncentracije PM čestica po godinama.

Koncentracija PM čestica varira uglavnom u istim opsezima za 2012., 2014. i 2015. godinu, dok je u 2013. zabeležena najveća varijacija, odnosno najveća koncentracija, pokazano outlier-ima.

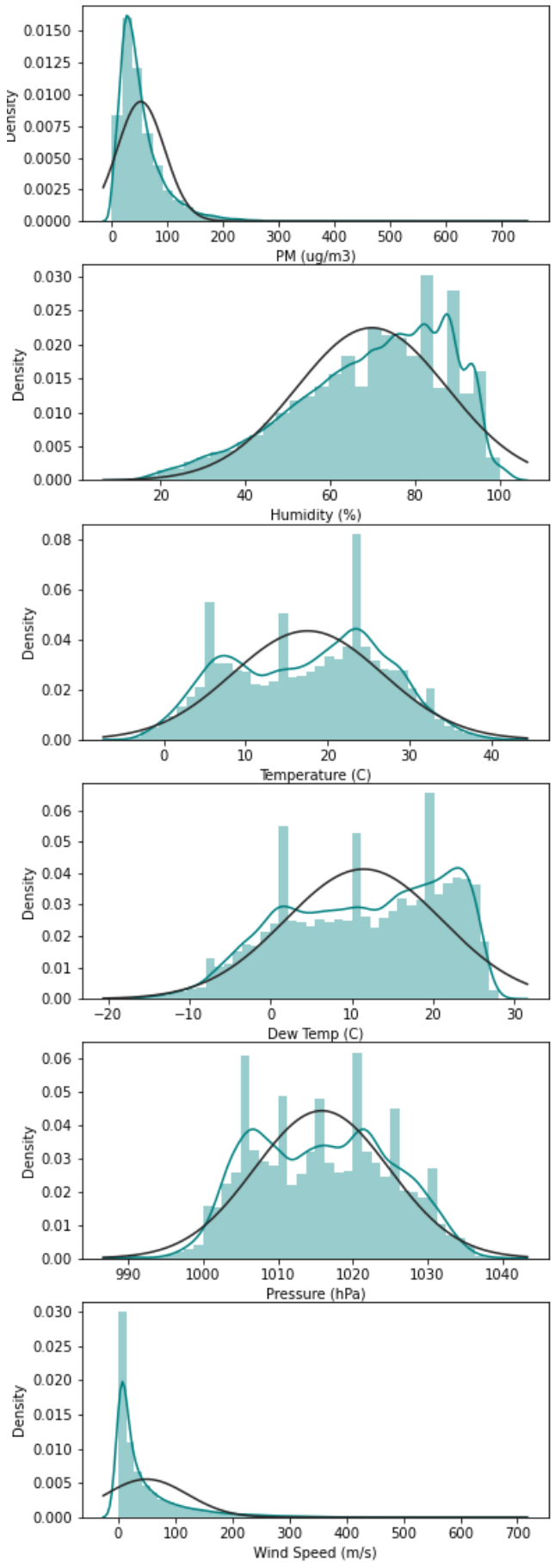
Na sledećem grafiku (Slika 4) je prikazana prosečna mesečna koncentracija PM čestica, na osnovu kojeg možemo da zaključimo da su PM čestice najzastupljenije zimi, odnosno u periodima kada se najviše koriste toplane i grejna tela.



Slika 4 - Upoređivanje koncentracije PM čestica po mesecima.

Na osnovu grafika koji prikazuju koncentraciju PM čestica možemo jasno videti da se koncentracija kreće u rizičnim opsezima - ogroman broj outlier-a i veliki interkvartilni opsezi na boxplot-ovima, kao i činjenica da je prosečna koncentracija u skoro svakom mesecu veća od 30 i sa srednjom vrednošću od 53 $\mu g/m^3$, ukazuju na to da stanovnici Šangaja, barem za godine za koje imamo pristup podacima, žive u uslovima koji su opasni po zdravlje, uzimajući u obzir da je granica za nezdrav nivo koncentracije 55 $\mu g/m^3$.

V. Raspodele Obeležja

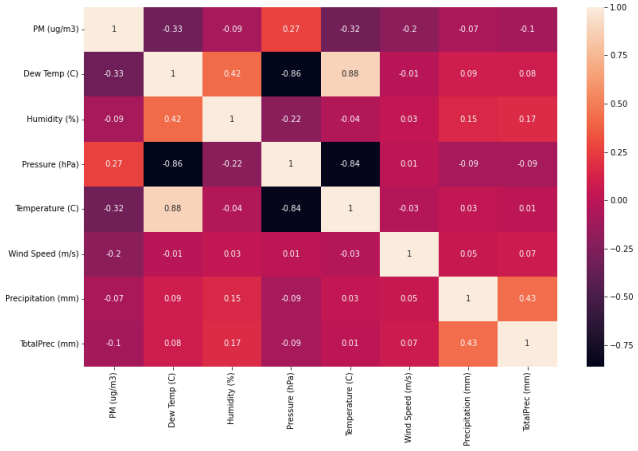


Slika 5 - Raspodele glavnih obeležja.

Na slici 5 su prikazane raspodele glavnih obeležja, kao i grafici normalne raspodele koji odgovaraju datim podacima. Raspodele obeležja PM i Wind Speed su jako iskošene ulevo i oštre, što ukazuje da su očekivane vrednosti za ova obeležja jako male. Obeležja Humidity, Temperature i Pressure dobro prate normalnu raspodelu, dok Dew Temp ima zakošenje udesno, što pokazuje da temperatura rose može često dostići više vrednosti.

VI. Korelacija Podataka

Korelacije između podataka možemo da vidimo na grafičkom prikazu korelacione matrice (Slika 6).

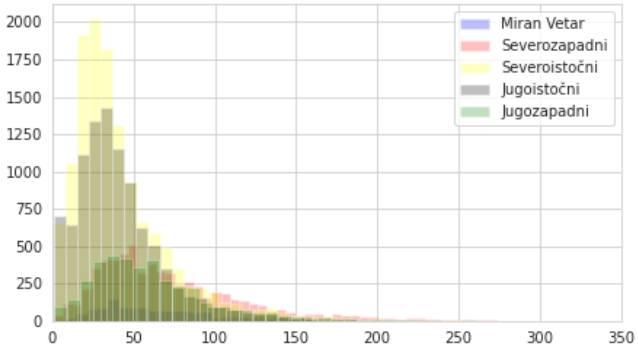


Slika 6 - Grafički prikaz korelacione matrice.

Što se tiče obeležja PM možemo videti dosta slabe negativne korelacije sa obeležjima Dew Temp, Temperature i Wind Speed. Kao što smo pokazali, koncentracija PM čestica je znatno veća zimi nego leti, pa je korelacija sa obeležjima vezanim za temperaturu opravdana. Negativna korelacija od -0.2 sa obeležjem Wind Speed može ukazati na to da je koncentracija PM čestica veća kada je vetar slabiji, što ćemo i pokazati daljim istraživanjima.

Pored toga, možemo da vidimo da je Dew Temp u pozitivnoj korelaciji sa Humidity i Temperature, i u negativnoj korelaciji sa Pressure obeležjem, što je i logično. Slično, Humidity ima slabu negativnu korelaciju sa Pressure obeležjem, i Pressure ima negativnu korelaciju sa Temperature obeležjem. Takođe je logična i pozitivna korelacija između obeležja Precipitation i TotalPrec, zbog prirode kumulativnog obeležja.

Obeležje Wind Direction nije uključeno u korelacionu matricu jer je kategoričko, pa njegov uticaj na koncentraciju PM čestica moramo posebno razmatrati. Ako bi obeležja grupisali po Wind Direction, videli bi da imamo 1330 uzoraka gde nije duvao vetar, 6530 gde je duvao severozapadni, 13273 gde je duvao severoistočni, 9984 gde je duvao jugoistočni i 4025 gde je duvao jugozapadni vetar. Raspodelu PM čestica po smeru duvanja vetra možemo videti na sledećem grafiku (Slika 7). Takođe, srednje vrednosti koncentracije PM čestica za određene smerove duvanja vetra su 72.98, 76.02, 43.28, 42.61 i 64.36 $\mu g/m^3$ respektivno.



Slika 7 - Raspodela PM čestica po Wind Direction obeležju (broj uzoraka po koncentraciji PM čestica).

Na osnovu ovoga možemo da primetimo da je najveća koncentracija PM čestica zabeležena kada je duvao severozapadni vetar, nakon čega su najveće koncentracije kada vetar nije duvao. Ovo je očekivano zato što se industrijska zona Šangaja nalazi na jugozapadnom delu grada, i zato što sa istočne strane grad izlazi na more, pa vetar nema nikakvih prepreka da čestice odnese.

VII. Linearna Regresija

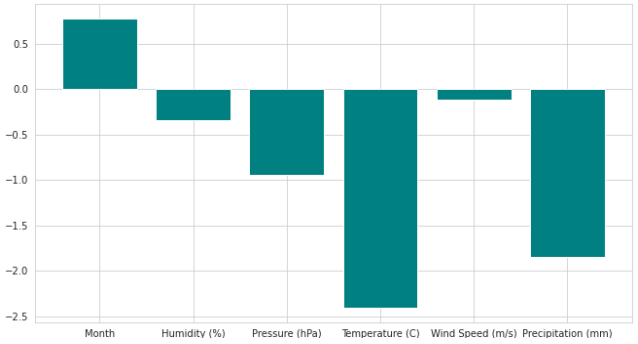
Možemo napraviti matematički model linearne regresije da bi izvršili predviđanje vrednosti koncentracije PM čestica u novim uzorcima, odnosno u budućnosti. Model možemo implementirati koristeći metod gradijentnog silaska. Za linearnu regresiju možemo uzeti u obzir standardnu linearnu regresiju prvog stepena, i Ridge regresiju sa standardizovanim uzorcima i kvadriranim parametrima. Takođe, koristeći Ridge regresiju možemo da napravimo više modela sa različitim parametrima, i da uporedimo rezultate, kako bi izabrali model sa najmanjom izabranom greškom. Na kraju, možemo uporediti rezultate najboljih modela iz obe metode linearne regresije, i videti kako će se oni ponašati u slučaju da izbacimo određena obeležja iz baze podataka, odnosno, da uradimo selekciju obeležja unazad.

Regresioni model se u početku trenira nad 70% uzoraka, gde se ostalih 30% podataka polovi na validacioni i test skup. Odatle, trening skup sadrži 24599 uzoraka, dok validacioni i test skup sadrže 5271 uzorak.

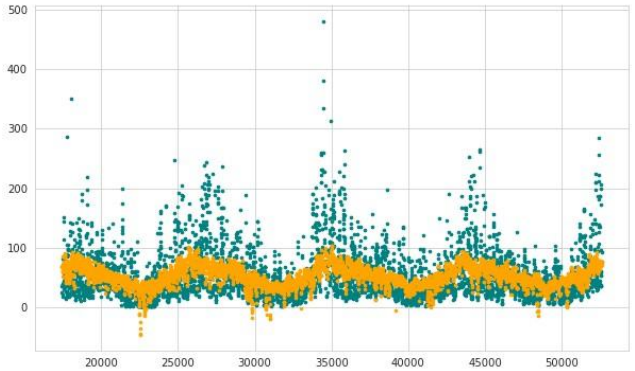
Regresiju u početku radimo nad obeležjima Month, Humidity, Pressure, Temperature, Wind Speed i Precipitation. Month je jedino obeležje vezano za datum koje nas zanima, Dew Temp je dosta korelisan sa Temperature pa ne mora ući u analizu, TotalPrec je kumulativno obeležje, i Wind Direction je kategoričko sa numeričkim vrednostima, što nije pogodno za regresor u slučaju velikih koeficijenata (iako je vrednost obeležja zapravo bitna za regresiju).

Prva faza predstavlja kreiranje modela linearne regresije prvog stepena. Nakon treniranja modela, predikcije nad validacionim skupom, ponovnim treniranjem nad spojenim trening i validacionim skupovima i ponovnom predikcijom nad test skupom, dobijeni su sledeći rezultati :

- MSE = 1395.48,
- MAE = 26.4
- RMSE = 37.35
- R2 / R2 Adjusted = 0.181



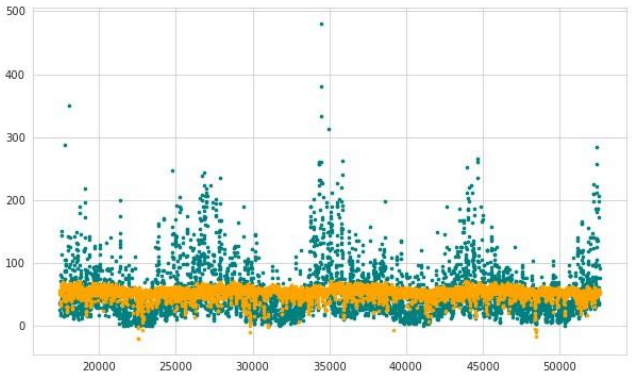
Slika 8 - Koeficijenti modela linearne regresije prvog stepena.



Slika 9 - Poređenje pravih vrednosti (plavo) sa predviđenim vrednostima (narandžasto) za linearni regresor prvog reda.

Iako su koeficijenti ovog modela u razumnom opsegu, rezultati se nisu najbolje pokazali.

Nakon selekcije obeležja unazad, odnosno, nakon što iz baze podataka izbacimo obeležja Pressure i Temperature, dobijamo dosta lošije rezultate.



Slika 10 - Rezultat predikcije modela bez obeležja Pressure i Temperature.

Odavde je očigledno da su ova obeležja, za ovaj model regresije, dosta važna, i da znatno utiču na sposobnosti uspešnog predviđanja budućih vrednosti.

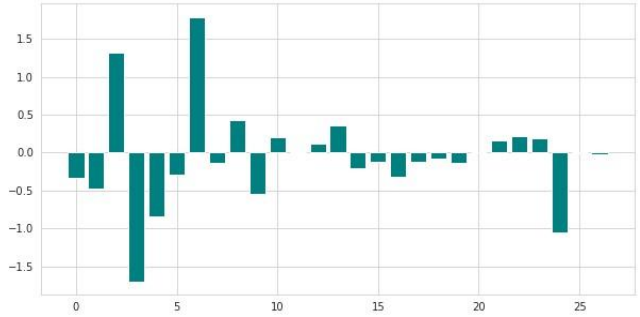
Za drugi model regresije biramo Ridge regresiju, za koju standardizujemo trening, validacioni i test skup podataka nultom srednjom vrednošću i jediničnom varijansom, što neće promeniti rezultat regresije. Takođe, možemo povećati broj parametara modela do kvadrata, što će bolje prilagoditi model na trening skup podataka.

Pravimo više modela Ridge regresije tako što menjamo parametre regresije, naime, uzimamo različite vrednosti za hiperparametar Alpha, koristimo različite ugrađene algoritme za računske rutine unutar biblioteke, i različiti maksimalan broj iteracija za algoritam gradijentnog silaska. Nakon procesiranja, dobijamo 45 modela, od kojih biramo model sa najmanjom izabranom greškom, u ovom slučaju R2 skor. Izabrani model je imao sledeće parametre :

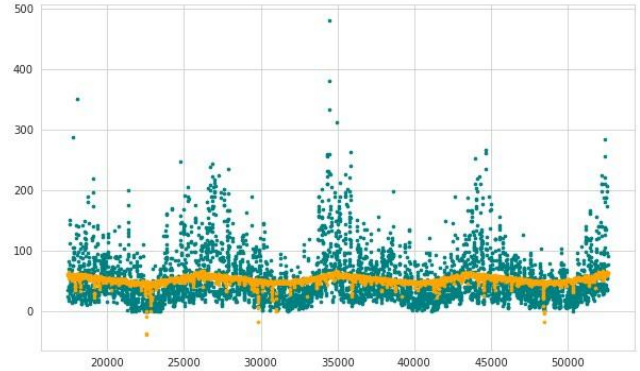
- Alpha = 5
- Solver = SAGA
- Max Iterations = 100

Nakon što ponovo kombinujemo trening i validacioni skup, i ponovo uradimo predikciju nad test skupom, taj izabrani model dobija sledeće rezultate :

- MSE = 1539.95
- MAE = 27.83
- RMSE = 39.24
- R2 / R2 Adjusted = 0.096



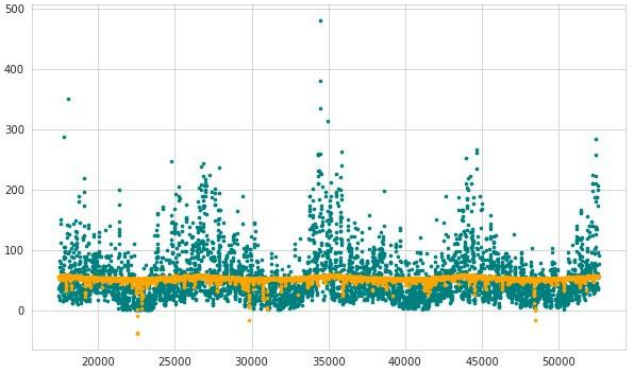
Slika 11 - Koeficijenti modela Ridge regresije.



Slika 12 - Poređenje pravih vrednosti (plavo) sa predviđenim vrednostima (narandžasto) za Ridge regresor.

Očigledno je da su rezultati linearnog regresora bolji za predviđanje novih vrednosti koncentracije PM čestica od Ridge regresora. U slučaju da smo uključili Wind Direction u regresiju, i da smo pravili modele sa većim stepenima, rezultat regresije bi sigurno bio bolji, međutim, mogućnost generalizacije modela za nove uzorke ne bi bila dobra kao što bi bila kod jednostavnijeg modela.

Takođe, kada bi za izabrani model Ridge regresije primenili istu selekciju obeležja, imali bi lošije rezultate.



Slika 13 - Rezultat predikcije modela bez obeležja Pressure i Temperature.