Analiza podataka – Predikcija zagađenja vazduha PM cesticama

Nikola Savić, IN35/2018,nikolaakv@gmail.com

I. Uvod

Ovaj izveštaj se bavi analizom podataka vezanih za zagađenost vazduha PM česticama. Takođe baza se bavi I predviđanjem koncentracije čestica PM2.5 koje imaju prečnik manji od 2.5 milimetara.

Čestice PM2.5 proizvode h toplane, motorna vozila, požari itd. Njihova velicina je uzrok mnogih kardiovaskularnih ili h toplane, motorna vozila, požari itd pulmonalnih bolesti.

Analizom ovih podataka i nakon utvrđivanja atributa koje utiču na koncentraciju čestica PM2.5 u vazduhu , moguće je kreiranje modela za njenu predikciju. Model bi se mogao koristiti za previđanje zagađenje vazduha česticama PM2.5 na osnovu meteorološke vrednosti.

II. Baza podataka

Baza sadrži podatke o 52584 uzoraka I 18 obeležja. od kojih su 11 numerička obeležja: PM: koncentracija PM2.5 čestica na nekoliko lokacija (ug/m3), DEWP: temperatura rose/kondenzacije (stepeni Celzijusa), TEMP: temperatura (stepeni Celzijusa), HUMI: vlažnost vazduha (%), PRES: vazdušni pritisak (hPa), Iws: kumulativna brzina vetra (m/s), precipitation: padavine na sat (mm), Iprec: kumulativne padavine (mm). Kategorička obeležja su: season: godišnje doba, redni broj merenja, godina, mesec, dan, sat, pravac vetra. Jedinica koncentracije čestice u vazduhu je mikrogram po metru kubnom, dok je temperatura rose izražena u celzijusima.

III. Analiza podataka

Analizom podataka utvrđeno je da nedostajući podaci javljaju za obeležja „PM_Dongsihuan“ (32076), „ PM_Nongzhanguan “ (27653), „PM_Dongsi “ (27532), „PM_US Post “ (2197), „ precipitation “ (484), „ Iprec “ (484), „ HUMI “ (339), „PRES “ (339), „ DEWP “ (5), „ TEMP “ (5), „ cbwd “ (5), „ Iws “ (5),

Najveći nedostatak podataka poseduju obeležja „PM_Dongsihuan “ 60.99%, „PM_Nongzhanguan “ 52.59% I „PM_Dongsi“ 52.59%, dok obeležje „PM_US Post“ , poseduje 4.2% nedostajućih vrednosti. Obelezja kod kojih fali nešto manje od 1% su : “precipitation” 0.92 %, “Iprec” 0.92 % , “HUMI” 0.64%, “PRES” 0.64%, “DEWP”,”TEMP”,”cbwd”,”Iws” 0.0096%.

Obeležja „PM_Dongsihuan“ (32076), „ PM_Nongzhanguan “ (27653), „PM_Dongsi “ (27532) u izostavljena iz razmatranja.

Takodje obeležje 'No' je izostavljeno iz razmatranja, jer predstavlja redni broj uzorka, samim tim ne utiču na analizu.

A. Nedostajuće vrednosti

Primetili smo da nasa baza u većini kolona ima bas mali procenat nedostajućih vrednosti. Samim tim obirisani su uzorci u kolonama “Iws”, “cbwd” , “TEMP” , “DEWP” , “PRES” , “HUMI” , “Iprec”, “precipitation”.

Nedostajuće vrednosti u koloni „PM_US Post“ su zamenjene metodom popunjavanja unazad.

Ovime smo rešili problem nedostajućih vrednosti u našem DataFrame.

Nakon ovih korekcija, baza se je svela na 51765 uzoraka i 14 obeležja.

B. Kategorička obeležja

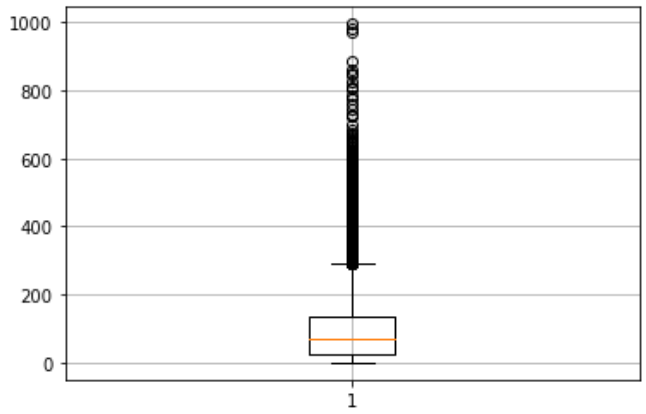
Obeležje “cbwd” je jedino kategoričko obeležje koje sadrži nenumeričke vrednosti. Zbog dalje analize kasnije je izvršena transformacija pravca vetra u stepene od 0 do 3.

C. Statističke veličine obeležja

Iz statističkih veličina obeležja možemo videti da za sve vrste zagađenja postoje outlineri, tj postoje izuzetno visoke vrednosti zagađenja koje se retko pojavljuju. Za svako od tih obeležja trebalo bi se konsultovati sa stručnjakom, da li su velike vrednosti posledica lošeg unosa ili stvarne vrednosti.

Sva obeležja koja imaju desnu asimetričnu raspodelu su: godina, mesec, dan, sezona, PM_US Post, PRES, cbwd, Iws, precipitation, Iprec, jer je relativno manji broj rezultata manji od aritmetičke sredine koji su znatno udaljeniji od nje. Sva ostala obeležja imaju levu asimetričnu raspodelu.

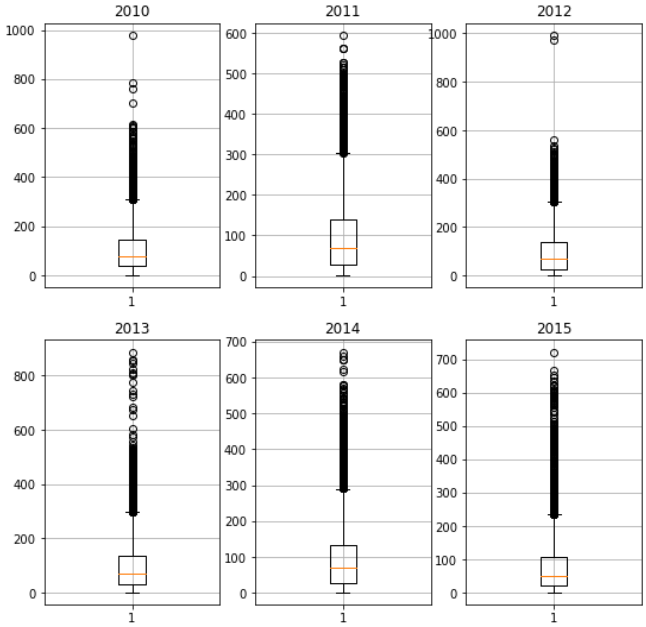
D. Analiza obeležja PM_US Post



Sl. Boxplot za obeležje PM_US Post

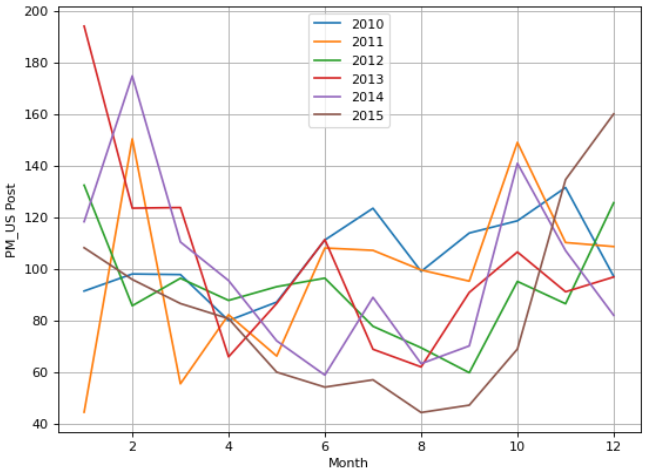
Neke od statistički veličina za obeležje PM_US Post su maksimalna vrednost 994, srednja vrednost 95, IQR opseg od 28 do 134. Takođe uočavamo da 75% uzoraka ima vrednost obeležja manju od 134 mikrograma po metru kubnom dok je maksimalna vrednost za ovo obeležje dostigla čak 994 mikrograma po metru kubnom.

1) Analiza obeležja PM2.5 u odnosu na godinu



S2. 6 boxplotova za obeležje PM_US Post po godinama

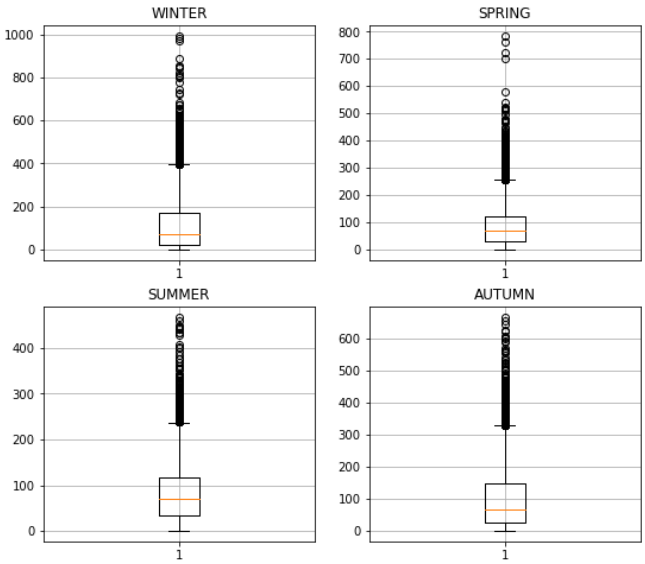
Na svakom od ovih boxplotova vidimo različita kretanja medijane po godinama. Međutim IQR nam pokazuje da se naši uzorci najčešće nalaze u sličnim opsezima tokom svih prikazanih godina. Ovim zaključujemo da je koncentracija čestica tokom godina približno ista.



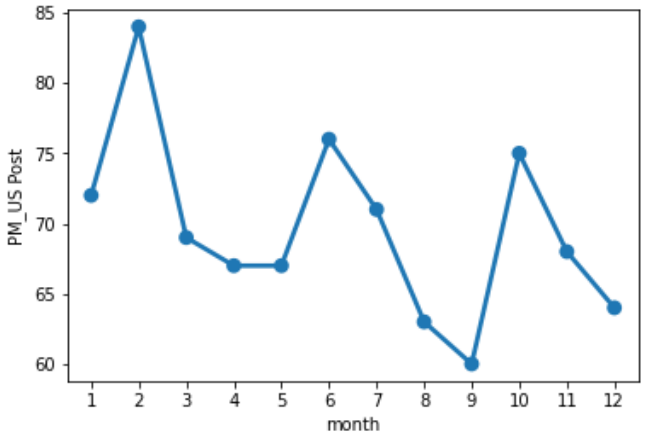
S3. Iscrtane srednje vrednosti za svaku od godina po mesecima

Ovim smo dokazali da nam je i srednja vrednost zagađenosti za svaku od godina približno jednaka u odnosu na mesece.

2) Analiza obeležja PM2.5 u odnosu na godišnje doba

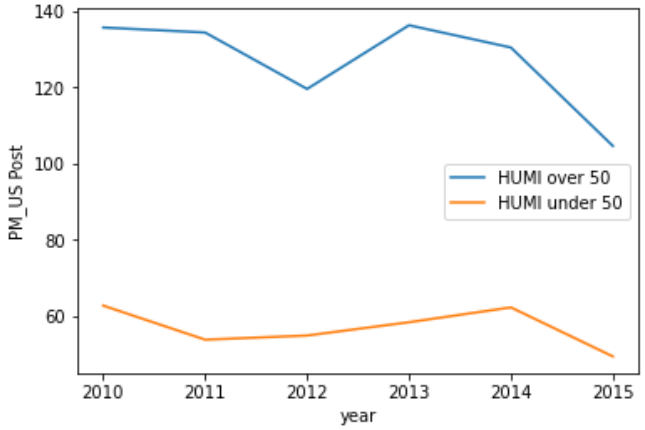


S4. 4 boxplotova za obeležje PM_US Post po godišnjim dobima



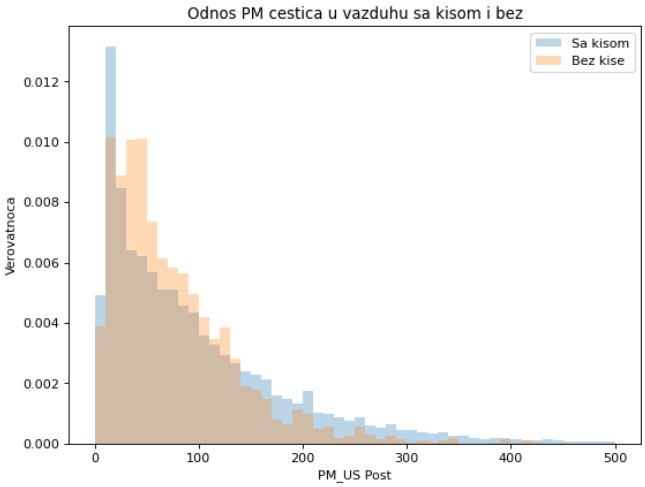
S5. Medijani obeležja PM_US Post po mesecima

Iz uporedne analize boxplot-ova može se primetiti da godišnja doba imaju uticaja na koncentraciju čestica PM_US Post. Vidimo nakon ovih istraživanja da je koncentracija čestica povećana u zimskom delu godine, ali vidimo i da se krajem svake zime koncentracija čestica smanjuje. Istu situaciju imamo i za leto, gde se na početku povećava a na kraju leta smanjuje na minimum što nam ukazuje na najmanju vrednost medijane u mesecu septembru. Početak jeseni nam nagoveštava ponovni rast nezdravih čestica.



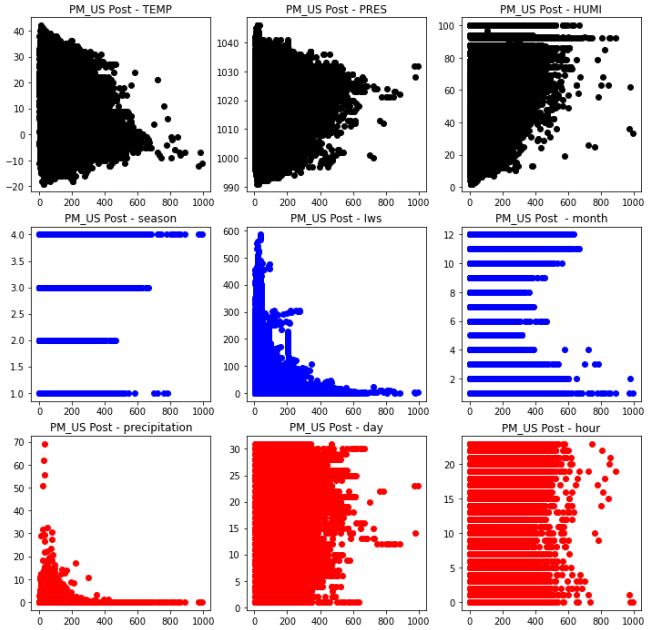
S6. Iscrtana koncentracija PM čestica kada je HUMI preko i ispod 50

Ovim smo pokazali da postoji zavisnost između HUMI i PM_US Post tako da što je vrednost za HUMI veća kroz godine to se koncentracija čestica srazmerno povećava.

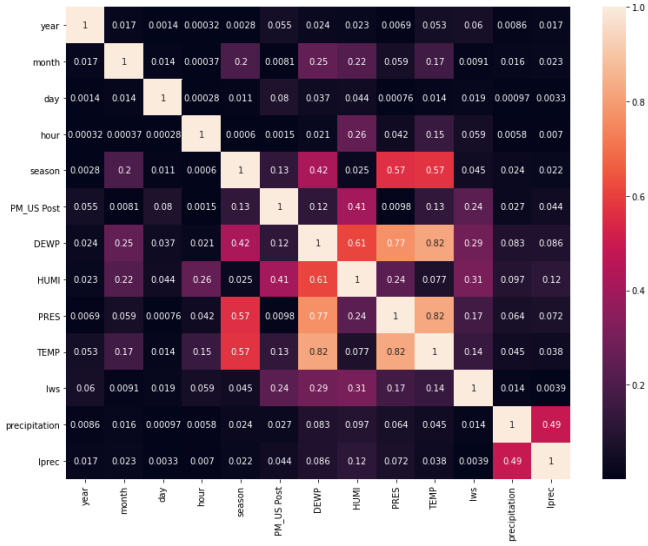


S7. Verovatnoća koncentracije čestica PM_US Post u toku kišnog/sušnog perioda

Raspodela u toku sušnog perioda je normalna raspodela, PM_US Post je najčešće između 35 mikrograma po metru kubnom, sa nešto manjom verovatnoćom između 0 i 150 mikrograma po metru kubnom, dok je puno manja verovatnoća da će koncentracija čestica PM_US Post biti veća od 150 mikrograma po metru kubnom. Raspodela u toku kišnog perioda nakrivljena u levo sa najčešćom koncentracijom čestica PM_US Post između 10 i 20 mikrograma po metru kubnom, dok su vrednosti iznad 280 mikrograma po metru kubnom sa vrlo malom verovatnoćom.

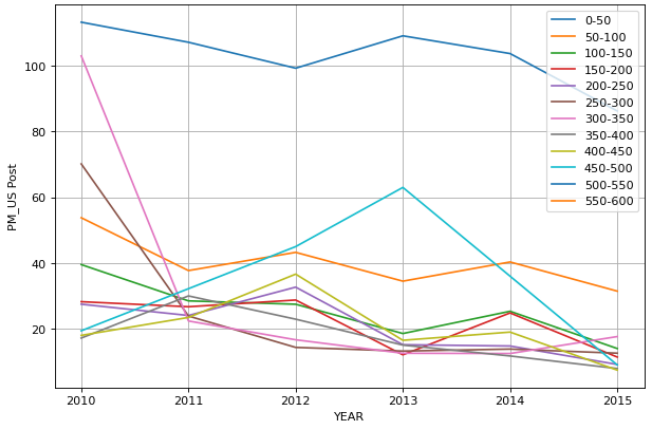


S8. 9. Scatter plotovi zavisnosti između PM_US Post i drugih obeležja



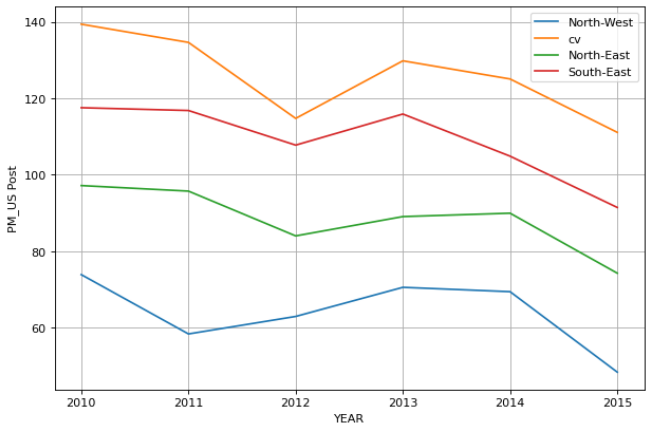
S9. Heatmap-a za prestavljanje korelacije između numeričkih obeležja

Vidimo da velika većina obeležja ima malu korelaciju sa obilježjem PM_US Post. Najveću korelaciju sa ovim obeležjem ima obeležje HUMI (0.41) a to se vidi na prvom scatter plotu. Nešto manju korelaciju sa ovim poljem imaju i polja Iws, TEMP i DEWP.



S10. Iscrtane srednje vrednosti PM čestica za različite vrednosti kumulativne brzine vetra u odnosu na godine

Na osnovu ovoga dolazimo do zaključka da kada nema ili ima baš malo padavina imamo ogromnu prosečnu količinu PM čestica u vazduhu. Sve vrednosti brzine vetra preko 50 nam govore da u vazduhu postoji znatno manja količina PM čestica.



S11. Iscrtane srenje vrednosti za PM čestice u odnosu na različite smerove vetrova u odnosu na godine

U slučaju kada nemamo vetrova kao što smo i gore pokazali koncentracija PM čestica je povišena. Takođe vidimo da je visoka prosečna koncentracija PM čestica kada duva jugoistočni vetar. Najmanje zagađenje vazduha imamo kada duva severozapadni vetar.

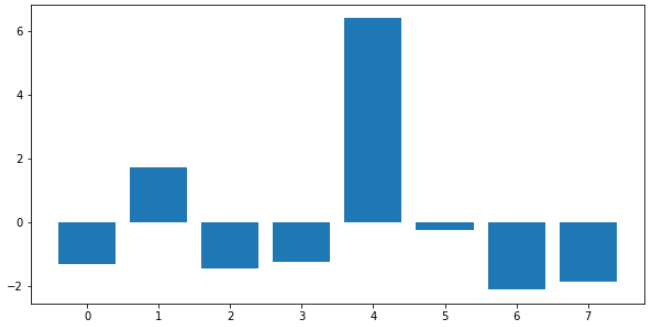
IV. LINEARNA REGRESIJA

A. Priprema podataka

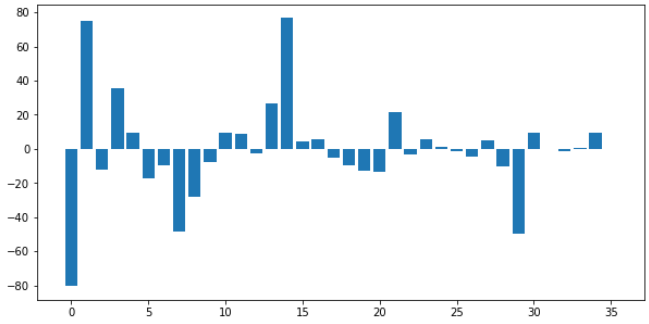
Početan skup uzoraka, podeljen je na dva podskupa, skup za obuku i skup za testiranje. Skup za obuku sadrži 90% uzoraka, dok test skup sadrži preostalih 10% nasumično izabranih uzoraka. Iz oba skupa izbačeno je obeležje godina, mesec, dan, sat, sezona, iz razloga što nam ova obeležja predstavljaju kategorička obeležja, osim ovog razloga obeležje godine je izbačeno kako bi model radio i za godine u budućnosti. Nakon toga izvršena je standardizacija obeležja, neophodna da bi se obeležja skalirala na isti opseg vrednosti.

B. Linearna regresija sa hipotezom y=b0+b1x1+b2x2+…

Model linearne regresije sa hipotezom y=b0+b1x1+b2x2+.. nakon regularizacije pokriva 77% udela korena srednje kvadratne greške, dok srednja apsolutna greška iznosi 56.78.



Promenu hipoteze korišćenjem “PolynomialFeatures” dolazimo do zaključka da je pokriveno manje procenata korena srednje kvadratne greške, a to je 74.91%, dok je srednja apsolutna greška 54.



Takođe promenom hipoteze na kvadratnu “PolynomialFeatures” dolazimo do srednjog korena kvadratne greške 73.44% dok je srednja apsolutna greška 52.19. Dolazimo do zaključka da je ovaj model najbolji za linearnu regresiju ovog data set-a.