Analiza podataka – $PM_{2.5}$ čestice

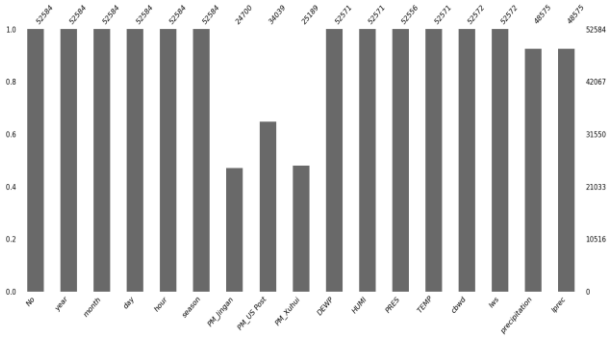
Aleksandra Mitro, IN8/2018, sandramitro99@gmail.com

1. OPIS BAZE PODATAKA

Ovaj izveštaj se bavi analizom podataka vezanih za količinu $PM_{2.5}$ čestica u vazduhu u Šangaju. Istraživanja su pokazala tesnu vezu između izlaganja $PM_{2.5}$ česticama i prevremenoj smrti od kardiovaskularnih i pulmonalnih bolesti. Koncentracija $PM_{2.5}$ čestica se smatra nezdravom ukoliko je vrednost preko 35.4 $\mu g/m^3$. Baza sadrži podatke o izmerenim vremenskim uslovima u toku svakog sata u periodu od 2010. do 2015. godine, ukupno 52584 uzoraka. Svakog sata su mereni sledeći parametri : koncentracija $PM_{2.5}$ čestica na 3 lokacije, temperature kondenzacije, temperatura, vlažnost vazduha, vazdušni pritisak, pravac vetra, kumulativna brzina vetra, količina padavina i kumulativna količina padavina. Pored prethodno navedenih obeležja, u bazi postoje i kategorička obeležja za redni broj merenja, godinu, mesec, dan, sat i godišnje doba, a takođe je i pravac vetra kategoričko obeležje.

2. ANALIZA PODATAKA

A. Osnovna analiza podataka



SLIKA 1 : Prikaz broja nedostajajućih podataka za svako od obeležja

Prilikom analize podataka uklonjena su obeležja za izmerenu količinu $PM_{2.5}$ čestica na lokacijama Jingan i Xuhui jer je više od 50% uzoraka za pomenuta obeležja imalo nedostajajuće vrednosti. Za obeležja koja su imala manje od 1% nedostajajućih vrednosti uzorci za koje su vrednosti tih obeležja nedostajale su uklonjeni. Dok su za obeležja koja su imala oko 7% nedostajajućih vrednosti uzorci za koje su vrednosti tih obeležja nedostajale zamenjeni prethodnom validnom vrednošću tog obeležja. Za obeležje $PM_{2.5}$ na lokaciji US Post nedostajalo je oko 35% vrednosti koje su zamenjene medijanom obeležja. Takođe sam uklonila obeležje koje je predstavljalo redni broj uzorka u bazi. Nakon primenjenih korekcija u bazi imamo 14 obeležja i 52555 uzoraka. Vrednost kategoričkog obeležja pravca vetra je zamenjena numeričkom vrednošću ugla.

B. Dinamički i interkvartalni opseg

TABELA 1 : Dinamički i interkvartalni opseg atributa

Atribut	Dinamički opseg	IQR opseg
Količina PM2.5 čestica	729	16
Temperatura kondenzacije	49	17
Temperatura	46	15
Vlažnost vazduha	88.68	25.44
Vazdušni pritisak	50	14.9
Kumulativna brzina vetra	1110	55
Količina padavina	61.6	0
Kumulativna količina padavina	226.4	0

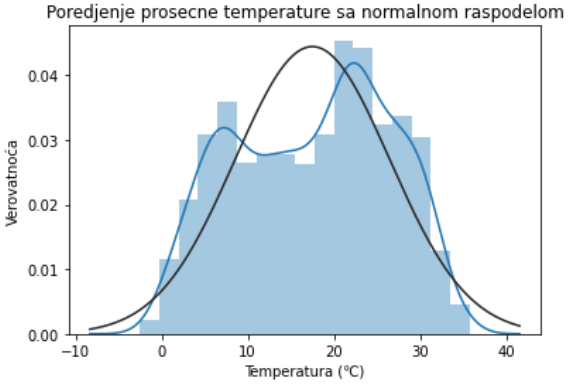
Iz tabele 1 se može zaključiti da dinamički opseg nije u potpunosti merodavan za procenu intervala koji vrednosti atributa zauzimaju. Bolje informacije nam daje interkvartalni opseg. Na primer, dinamički opseg temperature je 49 stepeni Celzijusa međutim 50% izmerenih temperatura se nalazi u opsegu od 17 stepeni Celzijusa. Takođe, ukoliko posmatramo izmerene vrednosti $PM_{2.5}$ čestica njihov dinamički opseg je čak 729 $\mu g/m^3$ ali se kod 50% uzoraka vrednost ovog atributa nalazi u opsegu od 16 $\mu g/m^3$.

C. Analiza temperature

TABELA 2 : Prikaz minimalnih, prosečnih i maksimalnih temperatura u zavisnosti od godišnjeg doba

Godišnje doba	Minimalna temperatura	Prosečna temperatura	Maksimalna temperatura
Proleće	-2.0	15.98	36
Leto	17	27.71	41
Jesen	-2	19.77	36
Zima	-5	6.23	24

Iz tabele 2 uočavamo da postoje velike amplitude u temperaturama u toku svakog godišnjeg doba, i takođe da je klima u Šangaju blaga sa toplim zimama i umereno toplim letima, dok su prelazna godišnja doba topla i u toku njih se dešavaju najveće temperaturne razlike.



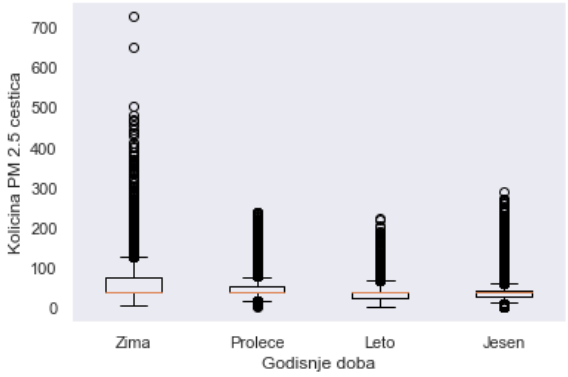
SLIKA 2 : Poređenje raspodele prosečnih dnevnih temperatura sa normalnom raspodelom

Sa slike 2 uočavamo da je raspodela prosečnih temperatura spljoštenija od normalne raspodele i da odstupa od normalne raspodele u tački maksimuma funkcije normalne raspodele i takođe sa slike 2 uočavamo izražen negativan koeficijent asimetrije koji iznosi -1,09.

D. Analiza količine $PM_{2.5}$ čestica

SLIKA 3 : Prosečna količina $PM_{2.5}$ čestica u toku godina

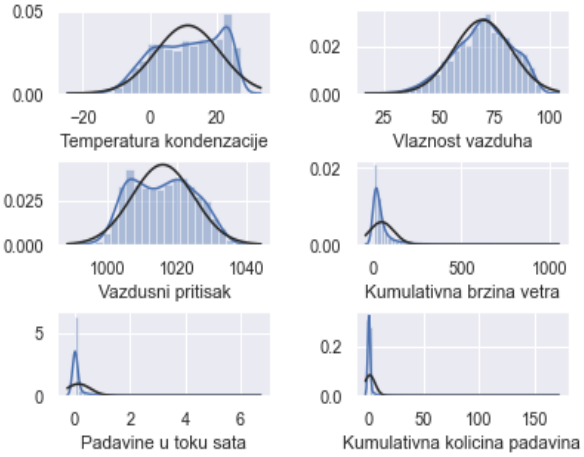
Sa slike 3 primećujemo porast u prosečnoj količini izmerenih $PM_{2.5}$ čestica od 2011. godine do 2013. u kojoj je bila najveća prosečna količina čestica koja zatim postepeno opada. Jedan od razloga za smanjenje koncentracije štetnih $PM_{2.5}$ čestica se ogleda i u tome što se od 2013. godine se u Kini primenjuju nova ekološka pravila usmerena na redukciju količine loših čestica koje se ispuštaju u atmosferu. Neke od mera koje su se pokazale kao efikasne u cilju smanjenja $PM_{2.5}$ čestica su jačanje standarda industrijske emisije, nadogradnja industrijskih kotlova, postepeno gašenje zastarelih industrijskih kapaciteta i promovisanje čistih goriva u stambenom sektoru.



SLIKA 4 : Kombinovani boxplotovi kolicine $PM_{2.5}$ čestica u toku godišnjih doba

Boxplot-ovi daju pregled medijana, interkvartilnih opsega i prisustva outlier-a za svaki atribut. Sa slike 4 uočavamo da je najveća količina $PM_{2.5}$ čestica izmerena u toku zime što je posledica atmosferskih uslova i toga što se koriste razne vrste goriva za grejanje zatvorenih prostora. Takođe uočavamo da je u toku proleća i leta prilično jednaka količina $PM_{2.5}$ čestica što je posledica toga što je većina dana vedra kada je koncentracija pomenutih čestica najniža. U toku svakog godišnjeg doba imamo prisutan velik broj outlier-a a najviše u toku zime. Primetno je da u toku svakog godišnjeg doba outlier-i zauzimaju višestruko veći opseg od interkvartalnog opsega. U uvodu je pomenuto da je nezdravo ukoliko koncentracija $PM_{2.5}$ čestica bude veća od 35.4 $\mu g/m^3$, daljom analizom je utvrđeno da je u toku 2010. i 2011. godine svaki dan koncentracija štetnih čestica bila veća od dozvoljene dok je taj broj opao na 230 dana u toku 2015. godini što je posledica novih ekoloških zakona koji su doneseni u 2013. godini.

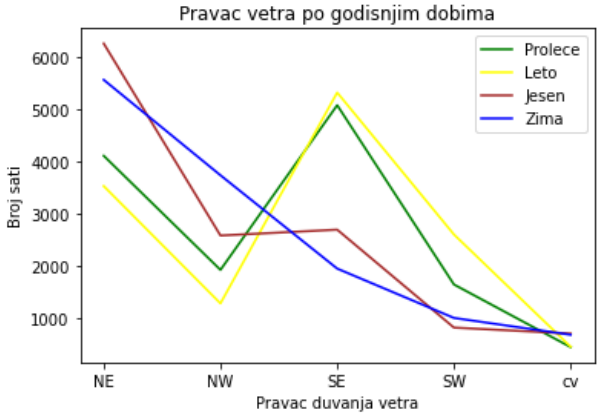
E. Analiza ostalih obeležja



SLIKA 5 : Poređenje funkcije raspodele srednjih dnevnih vrednosti obeležja sa normalnom raspodelom

Sa slike 5 uočavamo da funkcije raspodele obeležja temperature kondenzacije, vlažnosti vazduha i vazdušnog pritiska ne odstupaju mnogo od funkcije normalne raspodele. Sa druge strane, funkcija raspodele kumulativne brzine vetra, padavina i kumulativne količine padavine su spljoštenije u odnosu na funkciju normalne raspodele.

F. Analiza pravca vetra

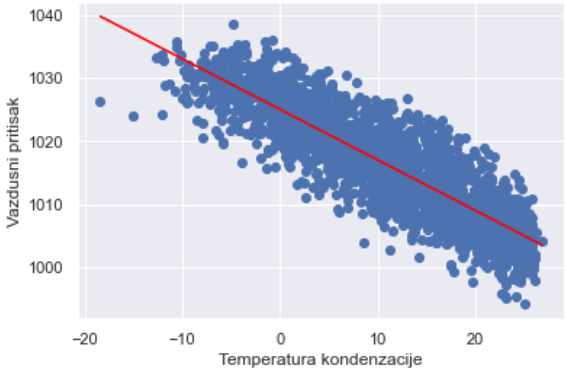


SLIKA 6 : Pravac duvanja vetra u toku svakog godišnjeg doba

Sa slike 6 uočavamo da u toku proleća i leta najčešće duva jugo-istočni vetar, dok u toku jeseni i zime najčešće duva severo-istočni vetar. Takođe je primetno da zapadni vetrovi retko duvaju, sem severo-zapadnog vetra u toku zime. Pošto vetar nosi sa sobom štetne čestice potrebno je voditi računa o tome sa koje strane grada se nalaze fabrike kako se ne bi uz pomoć vetra raznosile štetne čestice po gradu, a sa prethodne slike uočavamo da se fabrike u Šangaju ne bi trebale nalaziti sa istočne strane.

F. Obeležja sa najvećom korelacijom

Primenom funkcije corr utvrđeni su parovi obeležja sa najvećom korelacijom tj. međusobnom zavisnošću. Utvđeno je da je najveća korelacija između vazdušnog pritiska i temperature kondenzacije koja iznosi -0.85.

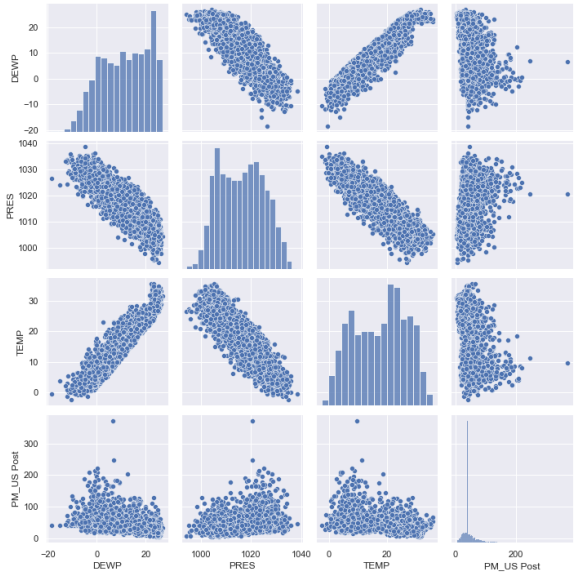


SLIKA 6 : Korelacija dnevnih vrednosti vazdušnog pritiska i temperature kondenzacije

Sa slike 6 uočavamo da sa opadanjem vazdušnog pritiska raste temperatura kondenzacije.

G. Analiza korelacije obeležja sa količinom $PM_{2.5}$ čestica

Korišćenjem korelacione matrice nad numeričkim obeležjima utvrđeno je da je količina $PM_{2.5}$ čestica u najizraženijoj korelaciji sa temperaturom kondenzacije, vazdušnim pritiskom i temperaturom, što je očekivano jer je u prethodnom tekstu uočena izvesna zavisnost između godišnjih doba i količine $PM_{2.5}$ čestica u vazduhu.



SLIKA 7 : Prikaz korelacije između obeležja koja imaju najveći uticaj na količinu $PM_{2.5}$ čestica

Sa slike 7 uočavamo da je prilično mala korelacija između obeležja koja imaju najistaknutiju numeričku vrednost korelacije sa količinom $PM_{2.5}$ čestica. Zaključak je da atmosferski uslovi nemaju mnogo uticaja na količinu štetnih čestica u vazduhu već razne vrste zagađivača, a da je period zadržavanja štetnih čestica u vazduhu u korelaciji sa atmosferskim uslovima.

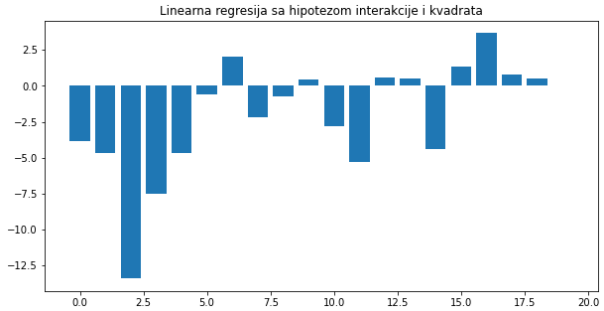
H. Predikcija količine $PM_{2.5}$ čestica

Za predikciju količine $PM_{2.5}$ čestica u vazduhu iz baze podataka su odbačena kategorička obeležja. Za obuku modela je korišćeno 90% uzoraka dok je preostalih 10% uzoraka stavljeno u test skup. Takođe je primenjena i selekcija obeležja korišćenjem OLS regresionog modela, prilikom koje su odbačena obeležja koja predstavljaju temperaturu kondenzacije i količinu padavina. Prilikom obuke modela korišćeno je više pristupa, od kojih su najbolje rezultate na test skupu imala linearna regresija sa hipotezom interakcije i kvadrata i Lasso regresija.

Mera uspešnosti regresora	Izračunata vrednost
Srednja kvadratna greška	1068.14
Srednja apsolutna greška	20.86
Koren srednje kvadratne greške	32.68
R^2	0.14
R^2 prilagođen	0.14

TABELA 3 : Mera uspešnosti linearne regresije sa hipotezom interakcije i kvadrata

Iz tabele 3 uočavamo na osnovu vrednosti R^2 mere, koja predstavlja udeo ukupne varijanse koju obučeni model pokriva, da naš model predviđa vrednosti koje su bliske srednjoj vrednosti zavisne promenjive u skupu za obuku.



SLIKA 8 : Ilustracija koeficijenata linearne regresije sa hipotezom interakcije i kvadrata

Sa slike 8 uočavamo da postoje ekstremne negativne vrednosti koeficijenata koji u velikoj meri utiču na pogrešnu procenu na test skupu.

3. ZAKLJUČAK

$PM_{2.5}$ čestice zbog svoje male mase ostaju dugo u vazduhu, često nošene vazdušnim strujama, što povećava verovatnoću da ih čovek ili životinja udahnu. $PM_{2.5}$ čestice prolaze barijeru nosa, grla i bronhija i dolaze u alveole pluća gde se talože i prelaze u krvotok, dolazeći do svih ćelija i na taj način mogu izazvati razne vrste oboljenja. Kroz analizu baze podataka dolazi se do zaključka da samo posmatranjem vremenskih uslova ne možemo doći do potpunih zaključaka u vezi sa štetnim materijama koje se nalaze u vazduhu već je potrebno posmatrati i nivo sagorevanja čvrstih i tečnih materija kako bi dobili tačnije rezultate linearne regresije. Preporuka na osnovu analize date baze podataka je da stanovnici Šangaja treba da nose maske otporne na $PM_{2.5}$ čestice u toku zimskih meseci kada je koncentracija štetnih materija najizraženija.