Analiza podataka – količina PM2.5 čestica u vazduhu

Vladimir Trpka, IN41-2018, vtrpka@gmail.com

I. Uvod

Domaći zadatak se bavi analizom PM2.5 čestica u vazduhu, kao i njihovim predviđanjem. Čestice manje od 2.5 mikrometra, jako su opasne. Nastaju sagorevanjem fosilnih goriva – auspusi automobila, grejanje na drva i ugalj, itd. Izazivaju teške plućne bolesti ulaze dikektno u krvotok. Njihov sastav je stoprocentno štetan za organizam. Čestice imaju u sebi supstance među kojima su i teški i toksični metali koji ulaze kao takvi u organizam. Koncentracija PM2.5 čestica se smatra nezdravom ukoliko je vrednost preko 35 $\mu g/m^3$. Analizom podataka se može doći do modela za predikciju čestica. Model bi pomogao u predviđanju zagađenosti vazduha PM2.5 česticama na osnovu klimatskih faktora.

II. Baza podataka

Baza podataka ima 52484 uzoraka i 17 obeležja. Jedan uzorak u bazi predstavlja jedan sat u toku jednog dana u godini sa zabeleženim podacima o klimatskim faktorima i količini PM2.5 čestica. Kategorička obeležja su : year-godina, month-mesec, day-dan u mesecu, hour-sat u danu, season-godišnje doba, cbwd-kombinovani pravac vetra. Numerička obeležja su : No-redni broj vrste ,PM_Caotangsi- koncentracija PM2.5 čestica, PM_Shahepu- koncentracija PM2.5 čestica, PM_US Post-koncentracija PM2.5 čestica, DEWP-temperatura rose, TEMP-temperatura, HUMI-vlažnost vazduha, PRES-vazdušni pritisak, Iws-kumulativna brzina vetra, precipitation-padavine na sat, Iprec-kumulativne padavine.

III. Analiza podataka

Primećeno je da u bazi podataka postoje nedostajući podaci za obeležja: PM_Caotangsi(28164), PM_Shahepu(27990), PM_US Post(23684), DEWP(529), HUMI(535), PRES(521), TEMP(527), cbwd(521), Iws(533), precipitation(2955), Iprec(2955). Obeležja PM_Caotangsi i PM_Shahepu su uklonjena iz baze jer nedostaje preko 50% podataka. Obeležje No-redni broj je uklonjeno jer nije relevantno za analizu. Nedostajući podaci za preostala gore navedena obeležja su takođe uklonjeni iz baze . Popunjavanje podacima bi narušilo preciznost predviđanja. Nakon korekcija ostaje 27368 uzoraka i 14 obeležja. Više nema uzoraka iz 2010 I 2011 godine. Prvi uzorci u skupu podataka nakon korekcije su iz maja 2012 godine I nad njima će se raditi dalja analiza. Vrednosti obeležja cbwd su zamenjena sa numeričkim vrednostima zbog lakšeg rukovanja podacima.

A. Dinamički i interkvartilni opseg

TABELA 1: Dinamički opseg atributa

Atribut	Dinamički opseg
Koncentracija PM2.5 čestica	687
Temperatura rose	44
Temperatura	40
Vlažnost vazduha	87.22
Vazdušni pritisak	50
Kumulativna brzina vetra	93
Padavine na sat	51.7
Kumulativne padavine	169.4

TABELA 2: Interkvartilni opseg atributa

Atribut	Interkvartilni opseg
Koncentracija PM2.5 čestica	63
Temperatura rose	13
Temperatura	12
Vlažnost vazduha	27.24
Vazdušni pritisak	13
Kumulativna brzina vetra	4
Padavine na sat	0
Kumulativne padavine	0

Iz ovih podatak iz ove dve tabele može da se zaključi da dinamički opseg nije merodavan za procenu intervala koji vrednosti nekog atributa zauzimaju. Bolje informacije daje interkvartilni opeg. Na primer dinamički opseg temperature rose je 44, ali se 50% tih vrednosti nalazi u opsegu 13.

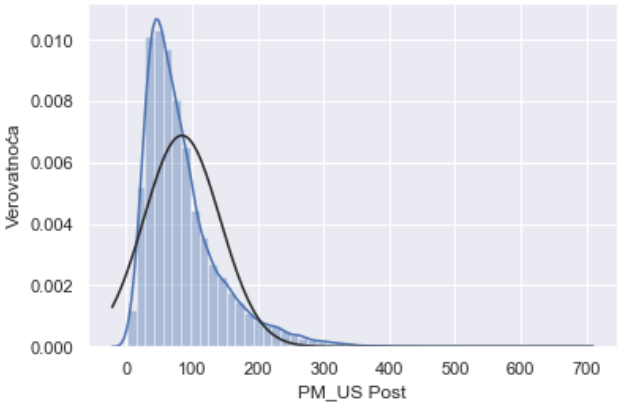
B. Da li raspodela PM2.5 čestica prati normalnu raspodelu?

TABELA 3: Koeficijenti

Koeficijent	Vrednost
Koeficijent asimetrije	1.7
Koeficijent spljoštenosti	4.24

Koeficijent spljoštenosti je veći od nule. Na osnovu toga zaključujemo da se navedena raspodela nalazi iznad modelovane normalne raspodele.

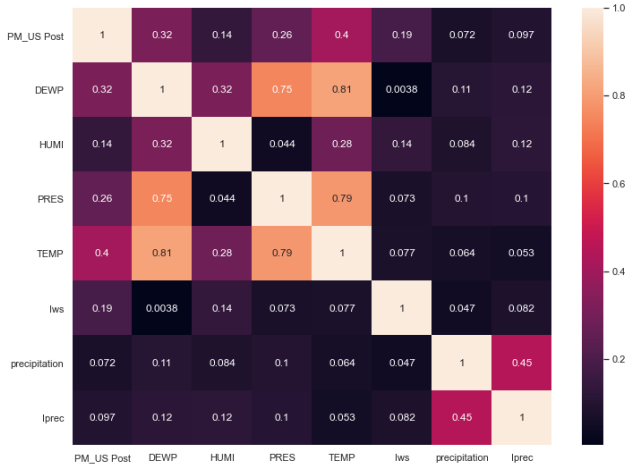
Koeficijent asimetrije iznosi 1.7 što nam govori da raspodela nije simetrična, već je pomerena u stranu.



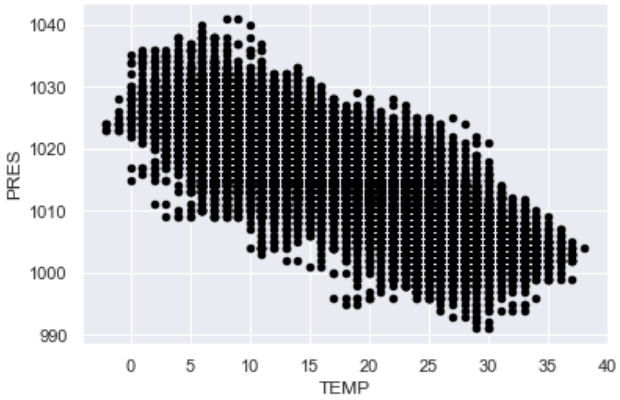
Slika 1. Raspodela PM2.5 čestica u poređenju sa normalnom raspodelom

C. Korelacija između obeležja

Primenom funkcije corr, dobijeni su parovi atributa sa najvećom korelacijom(međusobnom zavisnošću). To su temperatura i temperatura rose(jaka korelacija). Temperatura je takođe u jakoj korelaciji sa vazdušnim pritiskom. Vadušni pritisak je u jakoj korelaciji sa temperaturom rose. Obeležje PM_US Post je u najjačoj korelaciji sa obeležjem temperatura. Međusobna korelacija između svih ostalih parova obeležja je slaba i veoma slaba. Padavine na sat, kumulativne padavine i kumulativna brzina vetra ima malu korelaciju sa svim obeležjima, jer je količina padavina u ovoj regiji jako mala.



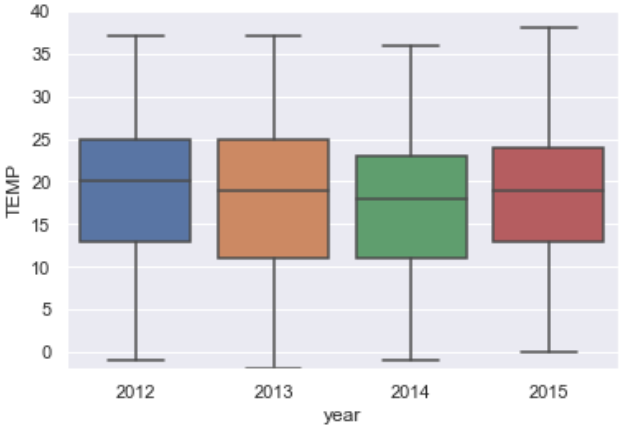
Slika 2. Korelacija između obeležja



Slika 3. Scatterplot između temperature i vazdušnog pritiska

D. Analiza obeležja TEMP

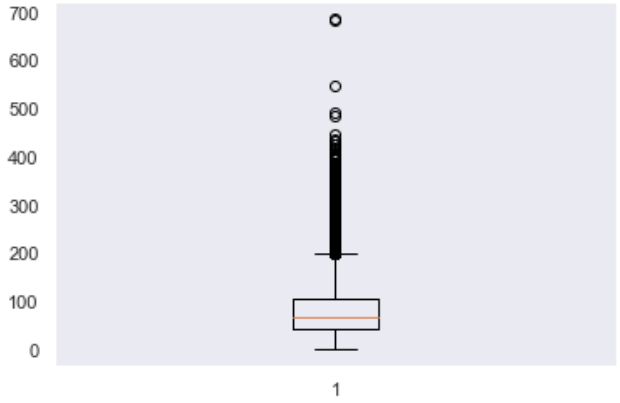
Minimalna vrednost temperature iznosi -2 stepena celzijusa dok je maksimalna zabeležena temperature 38 stepena celzijusa. Zaključujemo da ovo područje ima blage zime I umereno topla leta.



Slika 4. Prikaz temperatura po godinama

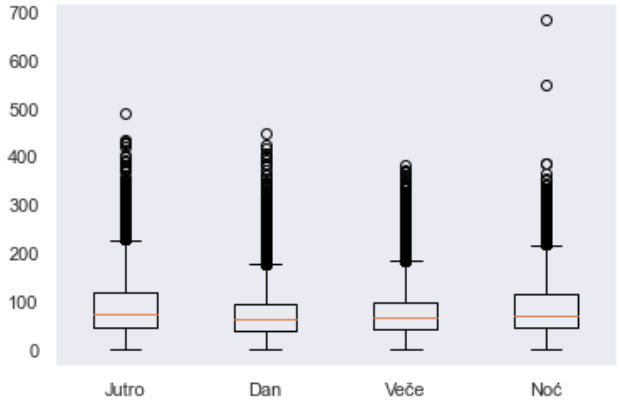
E. Analiza obeležja PM_US Post

Minimalna vrednost obeležja je 1 $\mu g/m^3$, dok maskimalna vrednost dostiže 688 $\mu g/m^3$. 75% uzoraka ne prelazi vrednost veću od 107 $\mu g/m^3$. Obeležje poseduje autlajere na visokim vrednostima.



Slika 5. Boxplot za obeležje PM_US Post

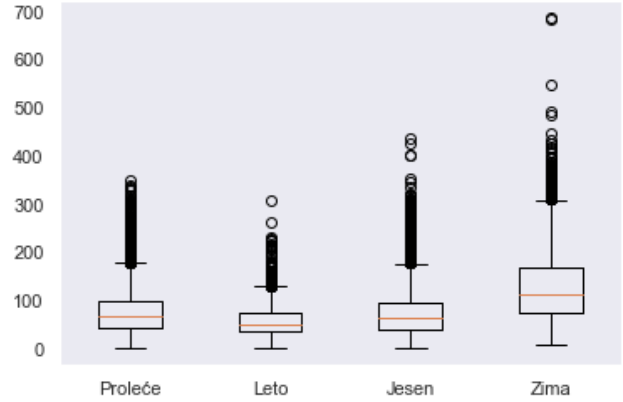
Na slici 6 vidimo kako se kreće koncentracija PM2.5 čestica tokom dana. Uočava se da su najviše prisutne u jutarnjem periodu I tokom noći. Preko dana i u večernjim satima koncentracija PM2.5 čestica je znatno manja.



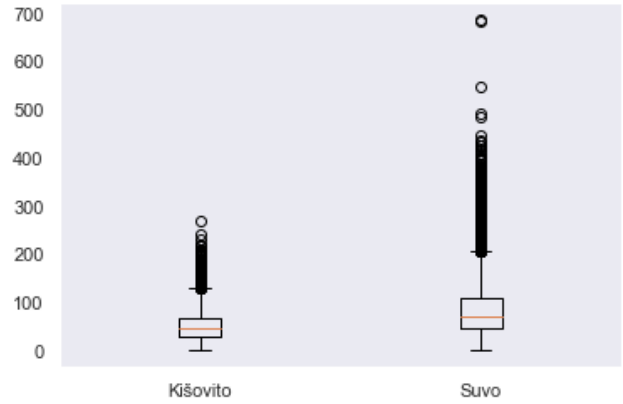
Slika 6. Boxplot pojavljivanja PM2.5 čestica tokom

dana

Prisutnost PM2.5 čestica tokom godine je najveća u zimskom periodu. Pretpostvlja se da se tada vazduh najviše zagađuje PM2.5 česticama zbog grejne sezone, paljenja uglja, drva itd. Tokom proleća i jeseni je koncentracija čestica približno ista, dok ih preko leta ima najmanje.



Slika 7. Boxplot pojavljivanja PM2.5 čestica tokom godine



Slika 8. Boxplot pojavljivanja PM2.5 čestica tokom kišovitog I suvog perioda

IV. LINEARNA REGRESIJA – PREDVIĐANJE PM2.5 ČESTICA U VAZDUHU

Za predviđanje količine PM2.5 čestica je iz celokupnog skupa uzoraka, 10% nasumično odabranih uzoraka korišćeno za testiranje a preostali uzorci su korišćeni za obuku modela. Korišćenje su različite hipoteze. Prikazane su dve od nekoliko korišćenih hipteza.

TABELA 4: Mere uspešnosti linearne regresije sa hipotezom y= b0+b1x1+b2x2+…+bnxn

Mera uspešnosti regresora	Vrednost
Srednja kvadratna greška	2495.209
Srednja apsolutna greška	36.501
Koren srednje kvadratene greške	49.952
R2	0,241
R2 prilagođen	0,240

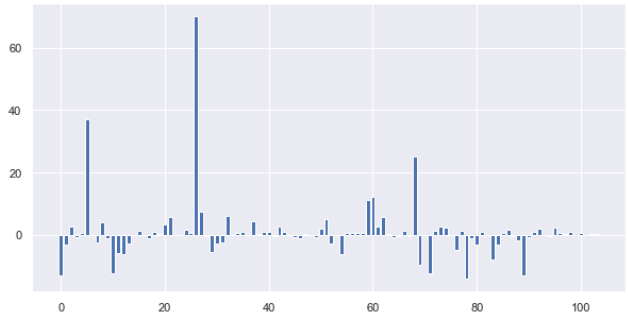
Linearnom regresijom sa hipotezom y= b0+b1x1+b2x2+…+bnxn dobijeno je pokrivanje varijanse od 42,1% sa srednjom apsolutnom greškom od 36,501.

Slika 9. Koeficijenti linearne regresije sa hipotezom y= b0+b1x1+b2x2+…+bnxn

TABELA 5: Mere uspešnosti Lasso regresije

Mera uspešnosti regresora	Vrednost
Srednja kvadratna greška	1868,352
Srednja apsolutna greška	31,666
Koren srednje kvadratene greške	43,224
R2	0,431
R2 prilagođen	0,429

Lasso regresijom dobijeno je pokrivanje varijanse od 43,1% sa srednjom apsolutnom greškom od 31,66. Na osnovu poređenja mera uspešnosti sa hipotezama koje su korišćene izabran je ovaj model kao najbolji.



Slika 10. Koeficijenti lasso regresije