Analiza podataka – koncentracija $PM_{2.5}$ čestica u vazduhu

Anastasija Popović, IN46/2018, popovicanastasija17@gmail.com

I. Opis baze podataka

U ovom izveštaju prikazaćemo analizu podataka o zagađenosti vazduha u gradu Čengdu, Kina. $PM_{2.5}$ čestice nastaju sagorevanjem fosilnih goriva i njihovu koncentraciju smatramo nezdravom ako im je vrednost veća od 35.4 $\mu g/m^3$. Imaju tu sposobnost da mogu dovesti do prevremene smrti od kardiovaskularnih i pulmonalnih bolesti. Za analizu korišćena je baza podataka sa 52584 uzoraka, koja pruža podatke od 2010. do 2015. godine u različitim vremenskim uslovima tokom svakog sata. U bazi uočavamo sledeća kategorička obeležja: redni broj merenja, godina, mesec, dan, sat, godišnje doba i pravac vetra. Takođe imamo i numerička obeležja: koncentracija $PM_{2.5}$ čestica (PM_Caotangsi, PM_Shahepu, PM_US Post), temeprtura rose/kondenzacije, vlažnost vazduha u %, vazdušni pritisak u hPa, temperatura u ^oC, kumulativna brzina vetra u m/s, padavine na sat u mm, kumulativne padavine u mm. Analizu vršimo u cilju kreiranja modela za predikciju koncentracije $PM_{2.5}$ čestica.

II. Analiza podataka

Prilikom analize podataka uklonjena su obeležja PM_Caotangsi i PM_Shaheou, jer je vise od 50% uzoraka nedostajalo, kao i obeležje No koje nam nije neophodno tokom analize.

A. *Nedostajući podaci*

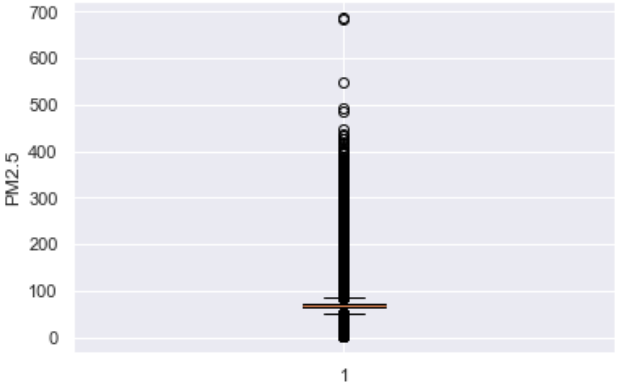
Obeležja koja su imala približno 1% nedostajućih uzoraka su uklonjena. Imamo i obeležja kojima je nedostajalo oko 6% uzoraka i njih smo popunili sa prethodnim validnim vrednostima. Obeležje US_Post $PM_{2.5}$ imalo je 45% nedostajućih podataka i te vrednosti smo zamenili srednjom vrednošću obeležja. Nakon prethodno pomenutih korekcija, modifikovana baza ima 52037 uzoraka i 14 obeležja.

B. *Dodela numeričkih vrednosti*

Kako je cbwd, tj. pravac vetra, jedino kategoričko obeležje koje nema numeričke vrednosti njegov pravac ćemo predstaviti preko uglova. Pored vrednosti 'NE' (45^o), 'SE' (135^o), 'NW' (315^o) i 'SW' (225^o) imamo i neodređeni pravac što smo postavili na vrednost 360^o.

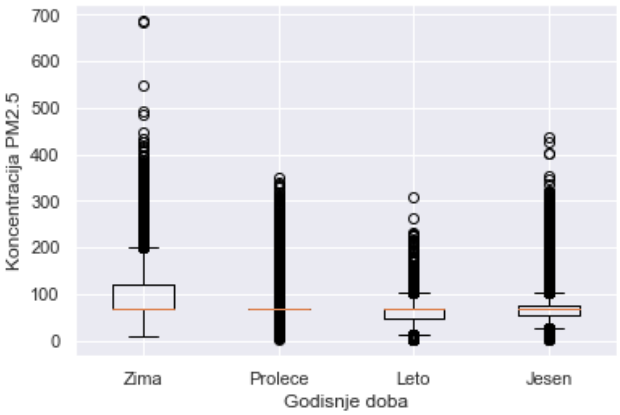
C. *Analiza koncentracije $PM_{2.5}$ čestica*

Sada prikazujemo putem boxplot funkcije da je interakvartilni opseg koncentracije čestica između 64 $\mu g/m^3$ i 73 $\mu g/m^3$. Primećujemo da postoji veliki broj autlajera koji imaju slične vrednosti. Maksimalna vrednost uočena u bazi je 688 $\mu g/m^3$.



Slika 1. Boxplot koncentracije $PM_{2.5}$ čestica

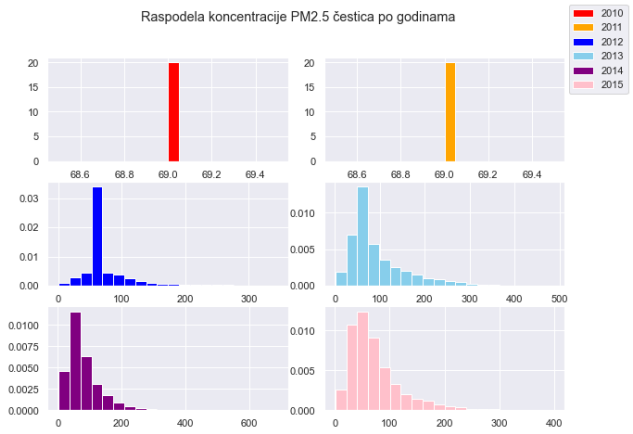
D. *Analiza $PM_{2.5}$ čestica kroz godišnja doba*



Slika 2. Boxplot $PM_{2.5}$ čestica kroz godišnja doba

Zanimljivo zapažanje imamo tokom proleća kada 50% uzoraka ima istu vrednost koncentracije čestica. Koncentracija čestica je najmanja tokom leta i jeseni a maksimalna zimi što nas dovodi do zaključka da je vazduh najzagađeniji tokom zime a razlozi za to su korišćenje raznih sirovina prilikom grejanja zatvorenih prostora.

E. *Analiza $PM_{2.5}$ čestica kroz godine analize*



Slika 3. Koncentracija $PM_{2.5}$ čestica kroz godine analize

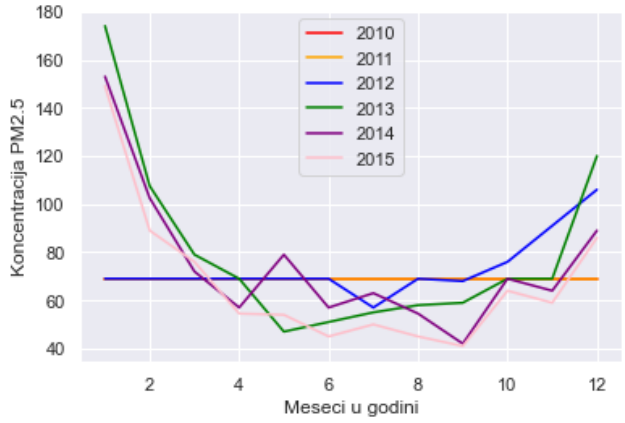
Koncentracija čestica tokom 2010. i 2011. godine je ista, dok za 2012., 2013., 2014. i 2015. godinu je veoma slična sa tim da godinama opada. Za ove četiri godine uočavamo sa slike levu asimetričnu raspodelu. Ova pojava istih vrednosti tokom prve dve godine rezultat je nedostajućih vrednosti u početnoj bazi podataka koje smo popunili srednjom vrednošću obeležja.

Tabela 1: IQR opseg tokom godina analize

Godina	IQR opseg
2010.	69.0 - 69.0 µg/m3
2011.	69.0 - 69.0 µg/m3
2012.	69.0 - 76.0 µg/m3
2013.	54.0 - 117.0 µg/m3
2014.	45.0 - 101.0 µg/m3
2015.	39.0 - 91.0 µg/m3

Iz tabele takođe primetimo da je opseg isti tokom 2010. i 2011. godine zato što smo nedostajuće vrednosti popunili srednjom vrednošću.

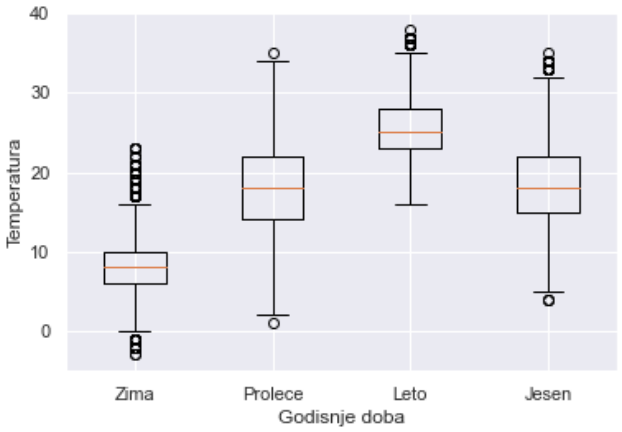
1) *Analiza čestica kroz godine za svaki mesec*



Slika 4. $PM_{2.5}$ čestice kroz godine analize za svaki mesec

Kao što smo već i objasnili, koncentracija čestica tokom 2010. i 2011. godine se ne menja te je nećemo razmatrati. Međutim, iz linijskog dijagrama sa slike 4 primećujemo da su čestice prisutne u najvećoj meri početkom godine i krajem godine (zimi) , dok je njihova koncentracija najmanja leti.

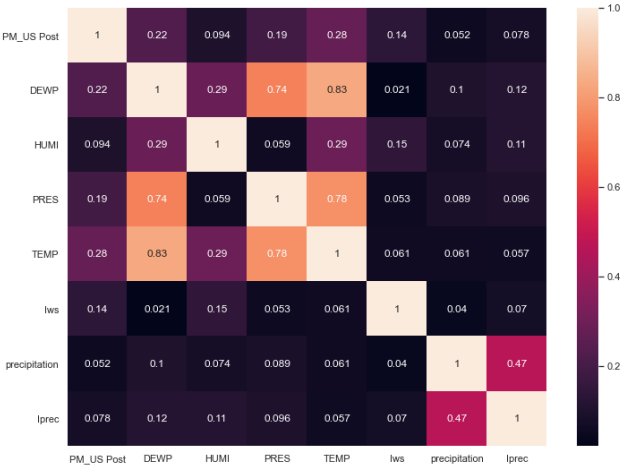
F. *Oscilacija temperature tokom godišnjih doba*



Slika 5. Boxplot temperatura tokom godišnjih doba

Ovde možemo da primetimo velike oscilacije temperatura između svakog godišnjeg doba. Zaključićemo da je klima blaga sa veoma toplim zimama i umereno toplim letima, dok prelazna godišnja doba imaju sličan opseg temperatura.

G. *Međukorelacija obeležja*



Slika 6. Korelacija između svih obeležja

Primetimo da najveću korelaciju sa koncentracijom $PM_{2.5}$ čestica ima obeležje temperatura i to 0,28% pa zatim temperatura rose sa 0,22%.

H. *Zavisnost $PM_{2.5}$ čestica sa ostalim obeležjima*

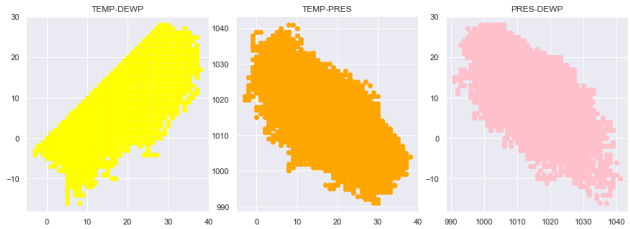
Slika 7. Zavisnost čestica sa ostalim obeležjima

Sa slike vidimo da se koncentracija čestica ponaša

slično u odnosu sa obeležjem koje predstavlja temperaturu kondenzacije i vazdušnog pritiska. Čestice naglo opadaju za temperature ispod nule za temperaturu rose. Takođe primetimo pad koncentracije čestica nakon temperature veće od 20 ºC. Koncentracija čestica sa obeležjem koje predstavlja vlažnost vazduha nam govori da koncentracija raste sa porastom vrednosti vlažnosti vazduha. U odnosu sa kumulativnom brzina vetra, koncentracija čestica je veća dok je brzina vetra mala.

1) Zavisnosti među drugim obeležjima

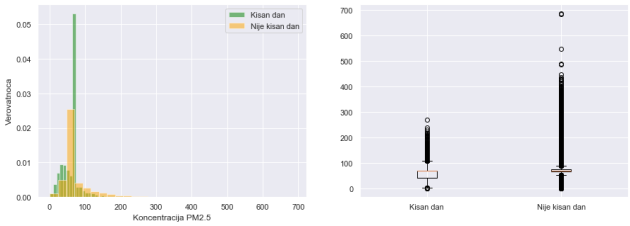
Primetili smo najveće zavisnosti među sledećim obeležjima:



Slika 8. Druge značajne zavisnosti

Sa ove slike možemo da zaključimo da su temperatura i temperatura rose pozitivno korelisana obeležja, temperatura i vazdušni pritisak su negativno korelisani kao i temperatura rose i vazdušni pritisak.

I. Anliza padavina

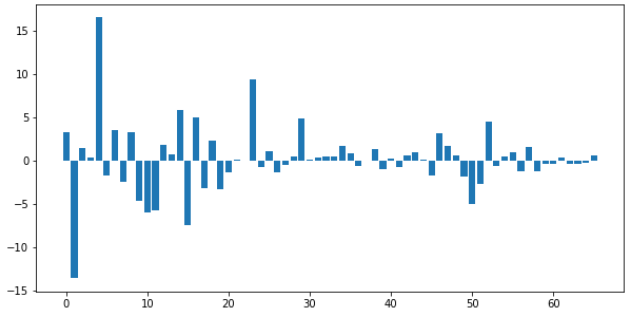


Slika 9. Analiza padavina

Koncentracija čestica za vreme kiše je između 43 $\mu g/m^3$ i 69 $\mu g/m^3$, a kada nema kiše koncentracija je između 67 $\mu g/m^3$ i 76 $\mu g/m^3$. Zaključujemo da je veća verovatnoća da će biti izmerena veća koncentracija štetnih čestica kada nema kiše a na to nam ukazuje veliki broj autlajera.

J. Predikcija količine $PM_{2.5}$ čestica

Izbacivanjem obeležja PM_US Post iz skupa za obuku, ubacujem u skup za testiranje. Za obuku modela korišćeno je 90% uzoraka a preostalih 10% je stavljeno u test skup. Prilikom obuke modela najbolji rezultati test skupa su dobijeni linearnom regresijom sa hipotezom interakcije i kvadrata. Takođe izbačena su i obeležja koja predstavljaju temperaturu i količinu padavina jer su imali p vrednosti redom 0,301 i 0,187. Tokom obuke vrednost R^2 skor se približavala vrednosti 1, sve do cetvrtog stepena. Polinomijalna funkcija petog stepena izbacuje R^2 skor približno -6, zato smo tu stali sa obukom.



Slika 10. Prikaz koeficijenata linearne regresije sa hipotezom interakcije i kvadrata

Sa slike uočavamo da je većina koeficijenata u opsegu od -5 do 5, dok nam se javljaju pojedine ekstremne vrednosti koeficijenata koje u većoj meri utiču na pogrešnu procenu nad podacima u test skupu.

III. Zaključak

Nakon celokupne analize koncentracije $PM_{2.5}$ čestica u gradu Čengdu došli smo do zaključka da posmatranjem vremenskih uslova (temperature, padavina, temperature kondenzacije...) ne možemo doći do konkretnih zaključaka za štetne čestice koje se nalaze u vazduhu. Zato su potrebne dublje analize kako bi dobili tačnije rezultate linearne regresije. Kako su čestice male i neprimetne, duže ostaju i lete u vazduhu i teško je kontrolisati njihovu koncentraciju, pogotovo u milionskim gradovima gde je velika koncentracija izduvnih gasova i drugih štetnih materija.