Analiza podataka-Predikcija zagađenja vazduha česticama PM2.5

Ilija Rakočević, IN59-2018, rakocevicilija7@gmail.com

I. Uvod

Šenjang, glavni grad provincije Liaoning sheng i najveći grad na severoistoku (bivša Mandžurija), a severno od reke Hun. Takođe jedan od najvećih industrijskih centara Kine. Prema proceni iz 2009. u gradu je živelo 3.543.444 stanovnika. Zagađenje vazduha skraćuje život za tri do četiri godine i izaziva 8,8 miliona prevremenih smrti, samim tim što Šenjang pripada industrijskoj zoni, zagađenost tog grada je iznad proseka. Pored ovog, postoje mnogi drugi razlozi, kao što su kisele kiše, kojima se priroda potpuno uništava. Veliki je značaj analize ovih pitanja jer mogu u potrebnim trenucima dati značajne odgovore. Kao na primer, ukoliko bi bio obučen model podacima iz ove oblasti, moglo bi se pomoći zaposlenima u industrijskim zonama i smanjila izloženost PM2.5 česticama.

II. Baza podataka

Baza sadrži podatke o atmosferskim parametrima za tačno određen sat, dan, mesec i godinu grada Šenjanga pri čemu je opseg godina od 2010. do 2015. godine. Dimenzije baze su (52584, 17), što znači da ima 52584 uzoraka i 17 obeležja. Manjinu čine kategorička obeležja koja su u kolonama: “year”, “month”, “season”, “cbwd”. Dok numeričkih obeležja ima 13, a to su: “No”, “day”, “hour”,“PM_US Post”,“PM_Taiyuanjie”,“PM_Xiaoheyan, “DEWP”,“TEMP”,“HUMI”,“PRES”,“Iws”,“precipitation” i “Iprec”.

III. Analiza podataka

A. *Nedostajuće i nepotrebne vrednosti u skupu*

Utvrđeno je da se nedostajuće vrednosti javljaju u velikoj meri kod obeležja „PM_US Post“ (30904), „PM_Taiyuanjie“ (28196), „PM_Xiaoheyan“ (27957), „precipitation“ (12763) i „Iprec“ (12763). Kod ostalih obeležja, nedostajuće vrednosti su oko 1%, što nije veliki procenat pa su iz tog razloga u sledečem koraku izbačene. Obeleže „No“ predstavlja redni broj uzorka, tako da nije od velikog značaja za analizu. Pored „No“, izbačena su i obeležja „PM_Taiyuanjie“ i „PM_Xioaheyan“, zbog zahteva zadatka. Primećeno je da u koloni DEWP, nelogična vrednost jeste temperatura od -97 stepeni celzijusa. Najverovatnije je u pitanju greška prilikom unosa, jer ova temperatura još uvek nije nigde zabeležena. Dodatno, pojavljuje se samo u jednom uzorku, tako da taj uzorak možemo otkloniti, jer nije od velike važnosti. Po Fujitsa skali, brzina vetra koja prelazi 322 km/h (~89.44 m/s) ima verovanoću događaja od 0.05%, tako da sve vrednosti obeležja „Iws“ koje su preko ove vrednosti (3.71%) imaju veliku verovatnoću da nisu ispravno unete, a pogotovo one koje su mnogo veće od 89.44 m/s jer je najrazorniji tornado koji je ikada zabeležen imao brzinu vetra od ~116m/s. i zove se „Daulatpur–Saturia“. Iz tog razloga, vrednosti koje su preko 89.44m/s možemo postaviti na nedostajuće vrednosti, kako bi ih mogli zameniti sa srednjom vrednošću; uzeti prethodne vrednosti i staviti na njihovo mesto... Ispostavilo se da je najbolje aproksimirana zamena za nedostajuće vrednosti u obeležju „Iws“, vrednost koja se dobila interpolacijom ostalih vrednosti obeležja „Iws“. To je odrađeno pomoću metode interpolate().

B. *Obeležje „cbwd“*

Pošto se treba odraditi obučavanje modela metodom linearne regresije, a ona radi sa numeričkim vrednostima; vrednosti obeležja „cbwd“ koje je po prirodi kategoričko, ćemo pretvoriti u obeležje sa numeričkim vrednostima. To je urađeno na takav način da su vrednosti „NE“, „NW“, „SE“, „SW“ i „cv“, konvertovane u vrednosti od 0 do 4, sledećom logikom. „cv“ = 0, „NE“ = 1, „NW“ = 2, „SE“ = 3 i „SW“ = 4.

C. *Obeležje „precipitation“*

Analizom je utvrđeno da je prisutan značajan procenat vrednosti 0 u ovom obeležju. Pored nula, zastupljeno je i ~24,2% nedostajućih vrednosti. Iz tog razloga kada se posmatraju osnovni parametri deskriptivne statistike, metoda df.describe(), zapažaju se nule na mestima min, 25%, medijana i 75%. Pošto 0 predstavlja: nema padavina po satu, NaN: nepoznata vrednost, a ostale vrednosti neku količinu padavina po satu. Urađen je isti postupak kao i sa brzinom vetra u prethodnom slučaju i uvedene su sledeće kategorije:

0 - nema padavina (0)
0.01 - 5 mm - mala količina padavina (1)
5.1 - 10 mm - srednja količina padavina (2)
10.1 - 15 mm - velika količina padavina (3)

15.1mm > - ogromna količina padavina (4)

Pošto imamo veliki broj nedostupnih vrednosti u ovoj koloni, vrednosti koje nedostaju, a koje su u periodu između oktobra i marta ćemo popuniti sa mean() vrednošću obeležja „precipitation“ jer je tada povećana verovatnoća bilo kakvih padavina, dok ćemo u ostale mesece staviti vrednost 0. Pošto su nam uzorci takvi da pokazuju podatke u određenom satu u toku dana, nećemo čuvati podatak o ukupnoj količini padavina jer je suvišan. Iz ovoga sledi da obeležje „Iprec“ otklanjamo.

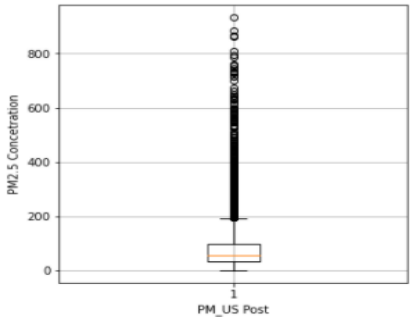
D. Nedostajuće vrednosti obeležja „PM_US Post“

Izračunat je broj nedostajućih vrednosti po godinama nebili došli do određenih zaključaka. Zapaženo je da godine 2010, 2011 i 2012 nemaju ni jednu poznatu vrednost za obeležje „PM_US Post“ (100% nepoznatih vrednosti). Godina 2013 ima ~38.5% nedostajućih vrednosti, dok godina 2014 ima 4.18%, a godina 2015 9.74%. Pošto 2014. i 2015. godina imaju korektan procenat nedostupnih vrednosti, svaka ta vrednost je popunjena prethodnom poznatom. Ovo je odrađeno metodom fill(). Sa druge strane, u 2013. godini je zapaženo da neki meseci imaju ogroman broj nedostajućih vrednosti (95% pa na više), a drugi meseci poprilično mali broj nedostajućih vrednosti. Napravljen je filter koji je otklonio sve mesece iz 2013. godine koji su imali preko 10% nedostajućih vrednosti. Tako da u daljoj analizi nisu korišćeni meseci „Januar“, „Februar“, „Mart“, „April“, „Jul“ i „Avgust“. One nedostajuće vrednosti u mesecima koji su prošli filter, ali su ih imali; su popunjene korišćenjem metode interpolate().

IV. ANALIZA OBELEŽJA „PM_US POST“

Tabela 1: Statistika obeležja „PM_US P“

Minimalna vrednost	1.0
Medijana	56.0
IQR opseg	33.0 – 97.0
Maksimalna vrednost	932.0



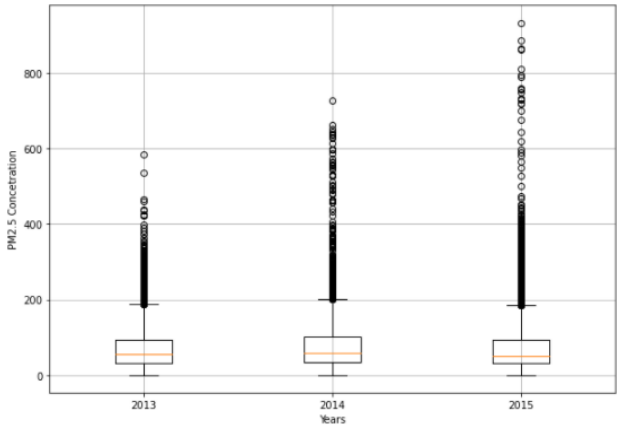
Sl. 1. Boxplot koji prikazuje opseg vrednosti

Sa slike (Sl. 1.) se može videti da obeležje poseduje određene outlayer-e kod velikih vrednosti. To potvrđuje činjenica da 75% vrednosti obeležja ima vrednost ispod 97 ug/m3.

A. PM2.5 koncentracija po godinama

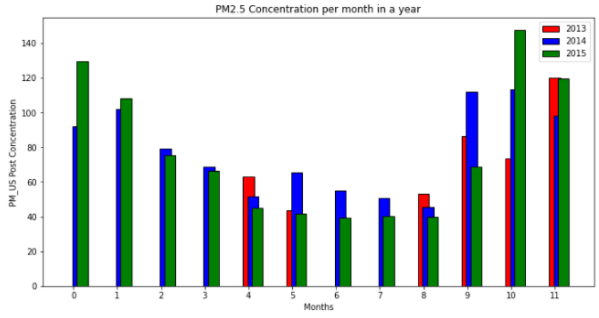
Tabela 2: Statistika obeležja „PM_US Post“ po godinama

Godine	2013	2014	2015
Minimalna	1.0	1.0	1.0
Medijana	55.0	60.0	52.0
IQR opseg	31.0 – 94.0	34.0- 101.0	33.0- 94.0
Maksimalna	583.0	725.0	932.0



Sl. 2. Prikaz koncentracije PM2.5 tokom godina

Opseg koji koncentracija PM2.5 postiže raste sa godinama i to primetno, i moglo bi se reći da dolazi do rasta koncentracije. Međutim, medijana i IQR opseg ostaju identični, uz neke manje razlike tokom godina. U suštini koncentracija je konstantna.

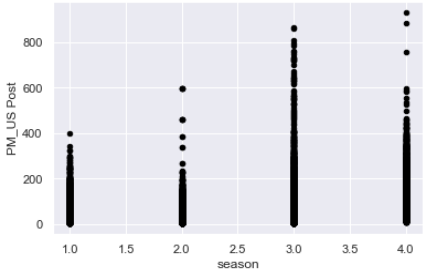


Sl. 3. Koncentracija PM2.5 kroz mesece u sve 3 godine

Na sledećoj slici (Sl. 3.) možemo primetiti da tokom godine koncentracija PM2.5 čestica varira. U letnjim mesecima koncentracija je drastično manja, dok u zimskim i jesenjim mesecima je povišena. To moze biti posledica ogreva u domaćinstvima i rada termoelektrana zbog grejanja gradova...

Tabela 3: PM2.5 po sezonama

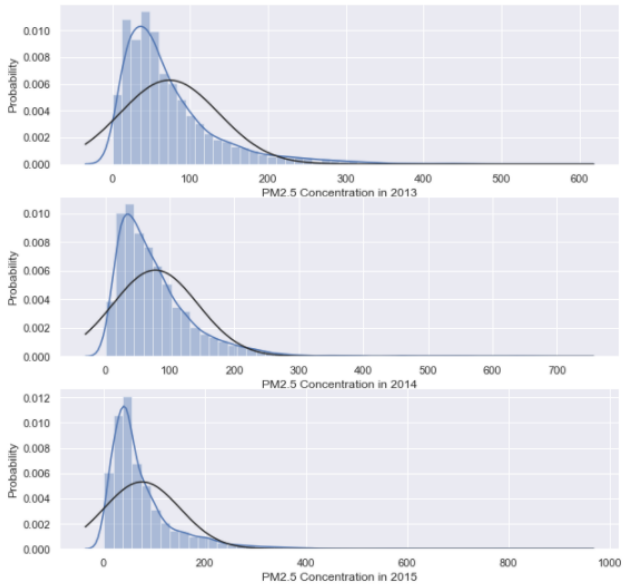
Sezone	1(Proleće)	2(Leto)	3(Jesen)	4(Zima)
Minimalna	2.0	1.0	1.0	7.0
Medijana	56.0	46.0	61.0	86.0



IQR opseg	33.0-85.0	27.0-58.0	30.0-106.0	47.0-154.0
Maksimalna	398.0	597.0	864	932

Sl. 4. Prikaz PM2.5 po sezonama

Slika (Sl. 4.) govori u prilog prethodnoj slici (Sl. 3.), a to je da koncentracija PM2.5 je u porastu kada su zimske i jesenje sezone, dok opada kada su letnje i prolećne.



Sl. 5. Raspodela PM2.5 posebno za svaku godinu

Na slici (Sl. 5.) možemo videti raspodelu PM2.5 koncentracije čestica. Raspodela je iskrivljena u desnu stranu (pozitivno asimetrična) i izdužena je u odnosu na normalnu raspodelu. Možemo primetiti da se verovatnoća za određene vrednosti smanjuje tokom godina. Pogledajmo raspodelu iz 2013. i 2015. godine za opseg 0 – 100 mikrograma po metru kubnom, raspodela iz 2015. je izduženija i uža, što znači da je verovatnoća da su njene vrednosti iz tog opsega mnogo veće (manja je varijansa).

Sl. 6. Korelacija između obeležja

B. *Korelacija između obeležja*

Što se tiče korelacije između obeležja, prvo možemo primetiti, da su „TEMP“ i „DEWP“ u jakoj pozitivnoj korelaciji što se i moglo očekivati jer što je veća temperatura, to je veća i kondenzacija, i suprotno. Takođe možemo primetiti da je „season“ u pozitivnoj korelaciji sa „precipitation“, što i ima smisla jer kada su zimski i jesenji periodi, padavine su češće. Vidimo takodje da su „PRES“ i „season“ u korelaciji. Dalje, vazdušni pritisak u korelaciji sa padavinama po satu... Sa druge strane, „precipitation“ i „TEMP“ su u negativnoj korelaciji jer što su padavine veće, temperatura će biti manja; Odatle sledi da će „precipitation“ biti u negativnoj korelaciji i sa „DEWP“. Takodje „TEMP“ i „PRES“ u negativnoj korelaciji.

V. LINEARNA REGRESIJA

Iz modela je izbačeno obeležje „PM_US Post“ jer ćemo u linearnoj regresiji predviđati vrednosti za ovo obeležje. Oformljen je ciljan skup vrednosti od obeležja „PM_US Post“. Funkcija def model_evaluation(y, y_predicted, N, d) služi za ispisivanje vrednosti mere uspešnosti regresora i to : MSE, MAE, RMSE, R2 i R2adjusted. Izabrana je samo mera R2 po kojoj će se pratiti. Početni skup uzoraka, podeljen je na dva podskupa, skup za obuku i skup za testiranje. Test skup sadrži preostalih 10% nasumično izabranih uzoraka, koji su sakriveni od skupa za obuku kako bismo proverili model, dok skup za obuku sadrži 90% uzoraka. Prvo je pokrenut osnovni oblik linearne regresije sa hipotezom y=b0+b1x1+b2x2+...+bnxn, dobijene vrednosti su prikazani na slici ispod.

```
Mean squared error:  4034.5355671904626
Mean absolute error:  42.528727624558464
Root mean squared error:  63.51799404255823
R2 score:  0.20336655506050638
R2 adjusted score:  0.20288144385887874
```

Sl. 7. Rezultat mera uspešnosti za običnu LR

U sledećem koraku je izvršena selekcija obeležja pomoću metode OLS iz statsmodel paketa, selekcijom unazad, gde su otkačena obeležja koja nemaju uticaj na obučavanje.

Naredni korak je standardizacija obeležja, kojim smo sve vrednosti skalirali na srednju vrednost koja je jednaka 0 i varijansu koja je jednaka jedinici. Ovim smo kompresovali sva obeležja koja su imali prevelike težine. Nismo promenili rezultat, tako da je rezultat isti kao sa sl. 7.

#Linearna regresija sa hipotezom y=b0+b1x1+b2x2+...+bnxn+c1x1x2+c2x1x3+...Pošto je

regresija osetljiva na korelaciju mešu obeležjima, ovom hipotezom se smanjuje efekat korelacije. Dobili smo sledeće rezultate.

```
Mean squared error:  3523.0337615835124
Mean absolute error:  38.3003399446205 96
Root mean squared error:  59.35514941084314
R2 score:  0.30436441186641205
R2 adjusted score:  0.3016017043371646
```

Sl. 8. Hipoteza sa interakcijom obeležja

Linearna regresija sa hipotezom y=b0+b1x1+b2x2+...+bnxn+c1x1x2+c2x1x3+...+d1x1^2 +d2x2^2+...+dnxn^2

U ovom koraku smo dozvolili da učestvuju i kvadrati obeležja što nam je pozitivno uticalo na rezultat.

```
Mean squared error:  3437.3704190388744
Mean absolute error:  37.46630316376323
Root mean squared error:  58.62909191722889
R2 score:  0.3212789445405877
R2 adjusted score:  0.31816681416605397
```

Sl. 9. Rezultat sa hipotezom koja uključuje kvadrate