Analiza vremeskih uslova i predviđanje PM2.5 čestica u gradu Guangdžou

Isidora Tulumbić, IN52/2019, itulumbic@gmail.com

I. UVOD

Ovaj izveštaj se bavi analizom podataka vezanih za vremenske uslove i predviđanjem koncentracije čestica PM2.5 u gradu Guangžou (Kina). PM (Particulate matter) tj. suspendovane čestice predstavljaju mešavinu čvrstih čestica dima, čađi, prašine i kiseline, uz teške metale poput olova, kadmijuma, nikla i arsena, a nastaju kao posledica sagorevanja uglja, nafte i ostalih derivata zarad proizvodnje struje i grejanja, zatim sagorevanja goriva u avio, auto i brodskom saobraćaju, industrije, šumskih požara itd. Ovde se konkretno izučava analiza PM2.5 čestica tj. čestica koje su veličine 2.5 mikrometara. Sve PM čestice jako loše utiču na zdravlje ljudi, pre svega na pluća i krvotok do kojih čestice veličine 2.5 mikrometara vrlo lako dostižu. Njihov sastav je takav da je stoprocentno štetan za ljudski organizam, zbog toga analiza ovih čestica može biti veoma korisna. Kreiranjem modela regresije/klasifikacije može se predvideti koncentracija PM2.5 čestica i time skrenuti pažnja na nastali problem kako bi se što bolje redukovala koncentracija samih čestica.

II. BAZA PODATAKA

Baza podataka u kojoj se nalazi skup podataka koji se obrađuje sadrži 17 obeležja i 52 584 uzoraka. Jedan uzorak baze predstavlja koncentracija različitih vremenskih uslova u gradu Guangdžou u nekom danu kao što su npr. temperatura rose, vlažnost vazduha, vazdušni pritisak itd. kao i koncentracija PM2.5 čestica od 2010. do 2015. godine. Kategorička obeležja koja se pojavljuju su: **year**, **month**, **day**, **hour**, **season, cbdw** (pravac vetra), dok su numerička: **No**, **PM_City Station**, **PM_5th Middle School**, **PM_US Post** (koncentracija PM2.5 čestica na nekoliko lokacija (μg/m3)), **DEWP** (temperatura rose/kondenzacije (°C)), **HUMI** (vlažnost vazduha (%)), **PRES** (vazdušni pritisak (hPa)), **TEMP** (temperatura(°C)), **Iws** (kumulativna brzina vetra (m/s)), precipitation (padavine na sat (mm)) *i* **Iprec** (kumulativne padavine (mm)). Jedino obeležje koje je dato slovnim vrednostima je cbwd. Neki algoritmi zahtevaju da sva obeležja imaju numeričke vrednosti kako bi se model mogao kreirati (npr. linearna regresija koja će biti korićena u nastavku rada), pa je potrebno vrednosti ovog obeležja pretvoriti u numeričke vrednosti. Za pretvaranje kategoričkih u numeričke vrednosti korišćeno je formiranje tzv. *dummy* varijabli. Ideja je da se od obeležja p koje ima K mogućih vrednosti (kategorija) formira K novih obeležja, pri čemu svako prestavlja jednu kategoriju. U tom slučaju za svaki uzorak K – 1 novih obeležja ima vrednost 0, dok samo jedno novo obeležje koje ukazuje na kategoriju ima vrednost 1. S obzirom da obeležje cbwd ima 5 mogućih vrednosti ovaj pristup je prilično dobar jer ne povećava mnogo dimenzionalnost DataFrame-a.

III. ANALIZA PODATAKA

A. *Predobrada nedostajućih podataka*

Pre svega su obrisana obeležja PM_City Station i PM_5h Middle School kao što je traženo u zadatku, tako da je samim tim rešen i problem nedostajućih podataka za ta obeležja.

Dalje je uočeno da postoje null vrednosti za 10 obeležja, gde u 9 obeležja postoji samo po 1 null vrednost. Pre svega se izvršila provera koji uzorak ima null vrednost za obeležje Season. Tom proverom je zaključeno da je to uzorak No 52584. Takođe je u rezultatu uviđeno da u tom uzorku postoje i null vrednosti za 8 preostalih obeležja u kojima se nalazi po jedna null vrednost. Za rešenje je odabrano da se taj uzorak obriše, jer se njegovim brisanjem ne unosi velika greška u skup podataka.

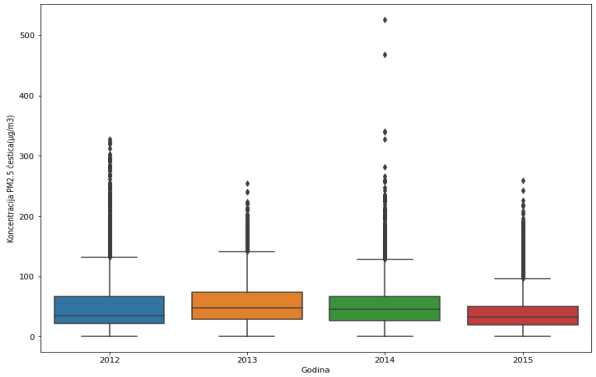
Za obeležje PM_US Post postoji nešto više od 38% null vrednosti. Ovo obeležje je dosta osetljivo jer se kasnije koristilo za predviđanje, pa je kao najbolje rešenje odlučeno da se obrišu godine 2010. i 2011. jer imaju više od 80% nedostajućih podataka. Ostali nedostajući podaci za ovo obeležje dopunjeni su susednim vrednostima metodom ‘ffill’.

Pored nedostajućih podataka mogu postojati i podaci koji su netipični za dato obeležje, a koje bi trebalo izuzeti iz daljeg razmatranja jer previše odstupaju od očekivanog uzorka. Analizom statističkih podataka uviđeno je da obeležja dewp i humi imaju vrednost -9999 što nije uobičajena vrednost za ta obeležja. Ovaj problem je rešen tako što su uzorci sa nevalidnim vrednostima zamenjeni null vrednostima i nakon toga su obrisani. Njihovim brisanjem se ne unosi velika greška u skup podataka, jer poseduju manje od 1% null vrednosti. Za ostala obeležja je utvrđeno da se nalaze unutar dozvoljenih vrednosti. Nakon sređivanja nedostajućih i nevalidnih podataka novi skup podataka sadrži 15 obeležja i 35 059 uzoraka .

B. Analiza i vizuelizacija zavisnosti promene PM2.5 od ostalih obeležja

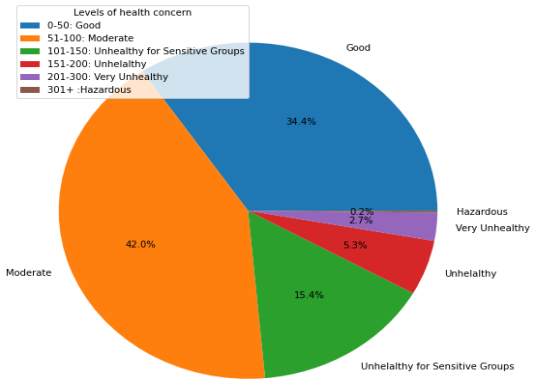
Klimatske varijacije značajno mogu uticati na koncentracije i hemijski sastav PM čestica, nezavisno od izvora zagađenja.

Pre svega analizirana je koncentracija PM2.5 čestica tokom godina koje su date u bazi. Na grafikonu (Slika 1) može se uočiti da je nešto manja koncentracija ovih čestica bila zastupljena tokom 2012. i 2015. godine u odnosu na 2013. i 2014. godinu. Takođe, uočava se da se za sve četiri godine u 50% slučajeva koncentracija čestica kreće do nešto manje od 100 µg/m3. Gornji opseg (maksimalne vrednosti) za 2012, 2013. i 2014. godinu je nešto manji od 150 µg/m3, dok za 2015. godinu iznosi oko 110 µg/m3. Izuzeci se uglavnom javljaju iznad gornjeg opsega. Najveći izuzeci zabeleženi su tokom 2014. godine. Može se uočiti da je najveća koncentracija PM2.5 čestice prelazi vrednost od 500µg/m3 što je veoma opasno po životnu sredinu. Ova činjenica je takođe prikazana na dijagramu koji prikazuje procentualnu koncentraciju PM2.5 čestica u i kvalitet vazduha po intervalima (Slika 2). Mere opreza koje se preporučuju u ovakvim situacijama jeste privremeno napuštanje zagađenog područja. Tokom 2012. godine takođe postoji dosta opasnosti, dok je tokom 2013. i 2014. godine situacija malo stagnirala.



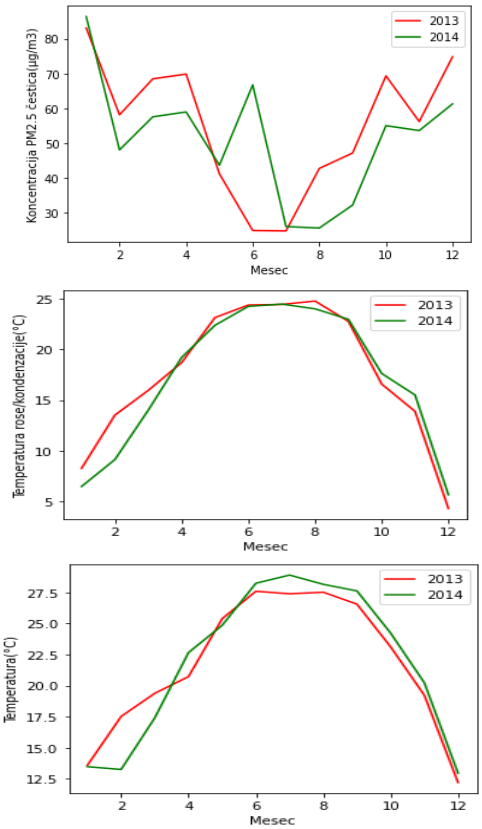
Slika 1: Koncentracija PM2.5 čestica od 2012. do 2015. godine

Sledeći dijagram prikazuje procentualnu koncentraciju PM2.5 čestica i kvalitet vazduha po intervalima (Slika 2). Zaključuje se da je grad Guangžou u periodu od 2012-2015. godine imao najveću koncentraciju PM2.5 čestica u intervalu od 51-100µg/m3. Pripadnost ovom skupu iznosi 42% od ukupnog broja uzoraka. Takođe, nešto manju koncentraciju ovih čestica grad je imao i u intervalu do 50 µg/m3, kao i intervalu od 101-150 µg/m3. Pripadnost ovim skupovima iznosi redom 34.4% i 15.4% od ukupnog broja uzoraka. Vazduh sa ovolikim koncetracijama PM2.5 čestica se smatra srednje zagađenim , gde se aktivnost na otvorenom ne moraju ograničiti, ali bi bilo poželjno da se čuvaju ljudi koji spadaju u osetljive grupe.



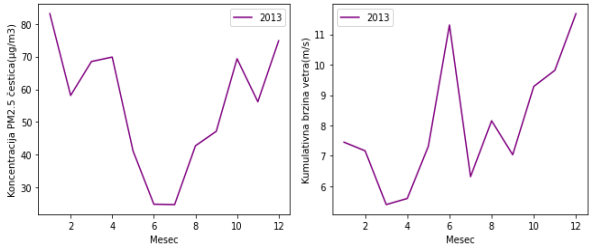
Slika 2: Prikaz procentualne koncentracije PM2.5 čestica i kvalitet vazduha po intervalima

Na sledeća tri dijagrama (Slika 4) prikazana je multivarijantna analiza koncentracije PM2.5 čestica, zatim temperature rose/kondenzacije i temperatura samog vazduha tokom meseci 2012. i 2013. godine. Može se uočiti da se temperatura vazduha i temperatura rose prilično podjednako kreću. Takođe, može se uočiti da su promene koncentracije PM2.5 suprotne od promene temperature vazduha i temperature rose, kao i da je tokom zimskih meseci kada je temperatura vazduha i rose mala, koncentracija PM2.5 čestica veća, a tokom leta kada je temperatura vazduha i rose velika, koncentracija PM2.5 čestica mala.



Slika 4: Prikaz promene PM2.5 čestica, temperature i temperature rose/kondenzacije tokom 2013 i 2014. godine

Na sledećim grafikonima (Slika 5) se može videti uporedno kretanje PM2.5 čestica i kumulativne brzine vetra tokom 2013. godine. Na prvom grafikonu se uočava da je koncentracija PM2.5 čestica znatno veća tokom prvih pet meseci i poslednja tri meseca u odnosu na preostale mesece. Na drugom grafikonu se uočava manja brzina vetra tokom prva četiri, kao i od sedmog do desetog meseca, dok je u preostalim mesecima bila znatno veća. Što se tiče poređenja ova dva grafikona vidimo da je koncentracija PM2.5 čestica veća prilikom manje brzine vetra za ovu godinu. Određene studije tvrde da je ovaj slučaj negativne korelisanosti čest kada su u pitanju PM2.5. Tek u slučaju čestica većih od 2.5 mikrometara nastupa pozitivna korelacija jer se usled resuspenzije prašine pod jakim vetrom povećava koncentracija istih.



Slika 5: Prikaz promene PM2.5 čestica i brzine vetra tokom 2013. godine

C. *Međusobne korelacije obeležja*

Zavisnost između dva obeležja može se izraziti kroz uzoračku korelaciju. U ovoj analizi, u cilju da se odredi međusobna korelacija obeležja, korišćena je tzv. toplotna mapa (engl. *heatmap*), koja je pogodna u slučaju ispitivanja većeg broja obeležja. Velika pozitivna korelacija uočava se između temperature rose/kondenzacije i temperature tj. kada raste temperatura rose, raste i temeratura i suprotno, što je i pokazano dijagramima na slici broj 4 (Slika 4). Velika negativna korelacija uočava se između temperature rose/kondenzacije i vazdušnog pritiska, kao i između temperature i vazdušnog pritiska, tj. kada vazdušni pritisak raste, temperature rose i temperatura opadaju i suprotno.

IV. LINEARNA REGRESIJA

Linearna regresija predstavlja metodu nadgledanog učenja koja se koristi za predviđanje vrednosti kontinualne izlazne promenljive. U ovom radu se pravi model linearne regresije koji predviđa koncengtraciju PM2.5 čestica.

Pre svega je iz skupa obeležja izbačeno obeležje PM_US Post jer se ono predviđa. Zatim je izvršena podela skupa podataka na test skup - 15% nasumično izabranih podataka, validacioni skup - 15% nasumično izabranih podataka i preostalih 70% podataka izdvojeno je za obuku modela.

Zatim je postavljena funkcija koja računa različite mere uspešnosti regresora, a koja će biti korišćena nakon svake obuke i testiranja. Nakon inicijalizacije i obuke modela gde je korišćena osnovna hipoteza utvrđeno je da ovaj model ne predviđa ni približno dobro, usled toga da su mere uspešnosti regresora daleko od idealnih vrednosti. (Slika 6). Vrednost R2 Score iznosi 0.091 što je daleko od 1 koja predstavlja idealan slučaj.

U cilju poboljšanja mere uspešnosti regresora dalje je rađena selekcija unazad, standardizacija obeležja, kao i Lasso i Ridge regularizacija.

```
Mean absolute error:  24.98275819435761
R2 score:             0.09155128792736
R2 adjusted score:    0.09081026647420343
```

Slika 6: Neke od mera uspešnosti regresora nakon obuke modela sa osnovnom hipotezom

A. *Selekcija unazad*

Selekcija obeležja podrazumeva da se iz skupa obeležja sukcesivno izbacuju obeležja čija je verovatnoća P veća od postavljene. Postavljena vrednost je 1%. Obeležja koja su imala veću verovatnoću od ove su obeležja cbdw_NE i cbdw_SE. Ova obeležja su uklonjena i ponovljena je obuka modela sa osnovnom hipotezom. Rezultati i dalje nisu poboljšani. Mere uspešnosti regresora su gotovo identične kao i mere uspešnosti regresora za prethodno ispitivanje (Slika 6).

B. *Standardizacija obeležja*

Standardizacija obeležja podrazumeva svođenje srednje vrednosti obeležja na nultu vrednost i jediničnu varijansu. Takođe, svodi se vrednost obeležja pod isti opseg. Nakon odrađene standardizacije obeležja ponovljena je obuka modela sa osnovnom hipotezom koja i dalje ne daje bolje rezultate, što je i očekivano s obzirom da standardizacija generalno ne poboljšava linearnu regresiju, već samo utiče na brzinu konvergencije algoritma opadanja gradijenta.

Sledeće što je razmatrano jeste da li možda treba da se upotrebi neka obuka sa interakcijama između obeležja što se radi ukoliko su neka obeležja korelisana. U delu analize je utvrđeno da postoji korelacija između nekih obeležja, tako da je ova obuka primenjena i dala je nešto malo bolje rezultate (Slika 7). Vidimo da je srednja apsolutna greška nešto veća u odnosu na vrednost sa slike 6, dok je R2 score bliži jedinici ovog puta.

```
Mean absolute error:  22.468308513530378
R2 score:  0.23961995927063895
R2 adjusted score:  0.23484946079884694
```

Slika 7: Neke od mera uspešnosti regresora nakon obuke modela sa interakcijama između obeležja

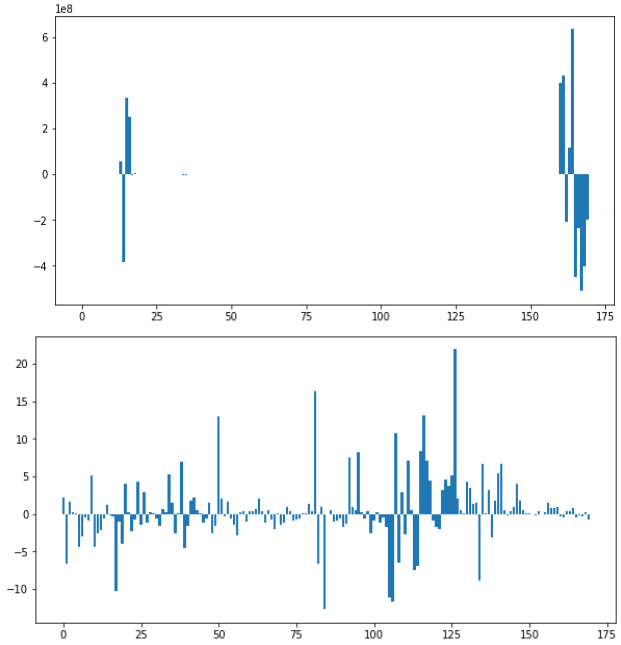
Dalje je primenjena obuka sa hipotezom koja pored interakcija između obeležja uključuje kvadrate pojedinačnih obeležja. Ona je dala neko minimalno poboljšanje vrednosti R2 score, što svakako ne dovodi do nekog većeg poboljšanja u obuci modela. Šta više, uviđa se da postoje veliki koeficijenti, a velike vrednosti uvek

treba sprečiti da bilo sprečeno da se model nadprilagodi. Zato je sledeći razmatran postupak zapravo postupak regularizacije.

C. Regularizacija

Regularizacija sprečava pojavu velikih vrednosti koeficijenata. Postoje dva pristupa regularizacije: Ridge i Lasso.

Prvo je isprobana Ridge regularizacija čija početna hipoteza je hipoteza sa kombinovanim obeležjima i sa kvadratima pojedinačnih obeležja. Mere uspešnosti regresora nakon Ridge regularizacije su uglavnom podjednake kao mere na Slici 7, tako da model nije puno poboljšan po tom pitanju. Međutim raspodele koeficijenata su sada dosta ravnomernije tako da je ovo za sada najbolji model linearne regresije (Slika 8).



Slika 8: Raspodela koeficijenata pre i nakon Ridge regularizacije

Nakon Ridge regularizacija isprobana je i Lasso regularizacija. Lasso regularizacija vrši i selekciju obeležja tj. koliko zaključi da je procenjen koeficijent za neko od obeležja jako blizu nuli, to obeležje izbacuje iz skupa. U ovom slučaju Lasso regularizacija nije dovela do poboljšanja mera uspešnosti modela u odnosu na Ridge regularizaciju.

Najbolje obučen model je model nad kojim je izvršena Ridge regularizacija sa hipotezom sa kombinovanim obeležjima i sa kvadratima pojedinačnih obeležja gde je stepen kažnjavanja ridge regularizacije 5.

V. KNN KLASIFIKACIJA

KNN klasifikacija (engl. nearest neighbours density estimation) je metoda za neparametarsku procenu gustine raspodele verovatnoće na osnovu *k* najbližih suseda.

Najpre je dodato novo obeležje Nivo bezbednosti koje poseduju podatke o tome da li je količina PM2.5 čestica bezbedna po životnu sredinu. Zatim je skup podataka podeljen na 15% uzoraka za testiranje finalnog klasifikatora, a preostalih 85% uzoraka za metodu unakrsne validacije sa 10 podskupova i definisana je funkcija koja koja računa mere uspešnosti klasifikatora.