Domaći zadatak

Vladimir Vincan, DE4-2021, vladimirvincan@gmail.com

Uvod

Tema projekta je predviđanje koncentracije PM 2.5 čestica u vazduhu, na osnovu poznatih meteoroloških podataka. PM 2.5 (eng. Particulate Matter) [1] čestice su čestice manje od 2.5 mikrometara u prečniku (poređenja radi, virusi mogu biti veličine do 1 μm [2]). One mogu biti različitog porekla, kao što su emisije automobila, industrijska proizvodnja, sagorevanje drveta i slično. Te čestice predstavljaju poseban razlog za brigu zbog toga što su dovoljno male da se mogu duboko udahnuti u pluća, čime uzrokuju respiratorne i druge zdravstvene probleme. Predviđanje koncentracije PM 2.5 čestica u vazduhu može pomoći pojedincima da informisano odlučuju o svom zdravlju, kao i da pomogne vlastima i zajednicama da naprave ispravne odluke prilikom pokušaja smanjenja koncentracije PM 2.5 čestica i poboljšanja kvaliteta vazduha. U ovom projektu će se implementirati regresioni model za predviđanje koncentracije čestica, kao i klasifikacioni model za opisivanje trenutnih meteoroloških uslova sa „bezbedno“, „nebezbedno“ i „opasno“.

I. Baza podataka

Baza podataka sadrži podatke o koncentraciji PM 2.5 čestica u pet kineskih gradova – Peking, Čengdu, Guangdžou, Šangaj i Šenjang. U ovom radu su analizirani samo podaci za grad Šenjang.

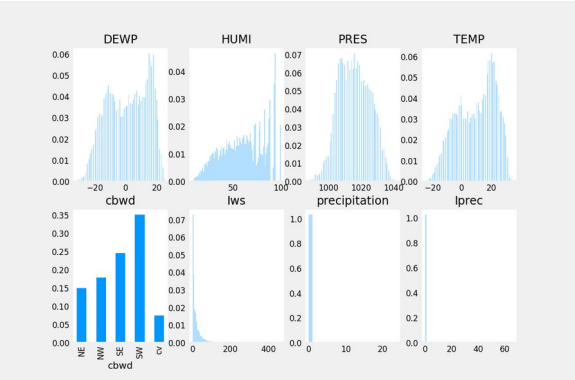
Baza podataka za Šenjang sadrži podatke skupljane od 2010. do 2015. godine i sadrži merenja obavljana svakog sata o koncentracijama PM 2.5 čestica u vazduhu, kao i različite meteorološke varijable kao što su temperatura, vlažnost vazduha i brzina vetra, kao i lokaciju na kojoj su podaci skupljani. U bazi ima ukupno 52584 uzorka, gde svaka vrsta predstavlja podatke skupljene u jednom satu i 11 obeležja. Sledeća obeležja postoje u bazi:

1. No – prvo obeležje, predstavlja redni broj uzorka;
2. year – godina u kojoj je uzorak izmeren;
3. month – mesec u kojem je uzorak izmeren;
4. day – dan u kom je uzorak izmeren;
5. hour – sat u kom je uzorak izmeren;
6. season – godišnje doba. Odabrano je da prvo doba počinje 1. marta, drugo 1. juna, treće 1. septembra i četvrto 1. decembra;
7. PM_Taiyuanjie - koncentracija PM2.5 čestica na lokaciji Taiyuanjie. Merenja su predstavljena u mikrogramima po metru kubnom;
8. PM_US Post - koncentracija PM2.5 čestica na lokaciji US Post. Merenja su predstavljena u mikrogramima po metru kubnom;
9. PM_Xiaoheyan - koncentracija PM2.5 čestica na lokaciji Xiaoheyan. Merenja su predstavljena u mikrogramima po metru kubnom;
10. DEWP - temperatura rose/kondenzacije. Merenja su prikazana u stepenu celzijusa;
11. HUMI - vlažnost vazduha, prikazano u procentima;
12. PRES - vazdušni pritisak. Merenja su prikazana u hekto-Paskalima;
13. TEMP – temperatura u hladu. Merenja su prikazana u stepenima Celzijusa;
14. cbwd - pravac vetra. Slovo predstavlja smer vetra i postoji 5 mogućnosti: N – sever, S – jug, E – istok, W – zapad, cv – mirno (eng. calm/variable);
15. Iws - kumulativna brzina vetra. Merenja su prikazana u metrima po sekundi;
16. precipitation – količina padavina na sat. Merenja su prikazana u milimetrima;
17. Iprec - kumulativne padavine. Merenja su prikazana u milimetrima;

Od datih obeležja, kategorička su godina, mesec, dan, sati, sezona i cbwd. Numerička obeležja su PM, DEWP, HUMI, PRES, TEMP, lws, precipitation i lprec.

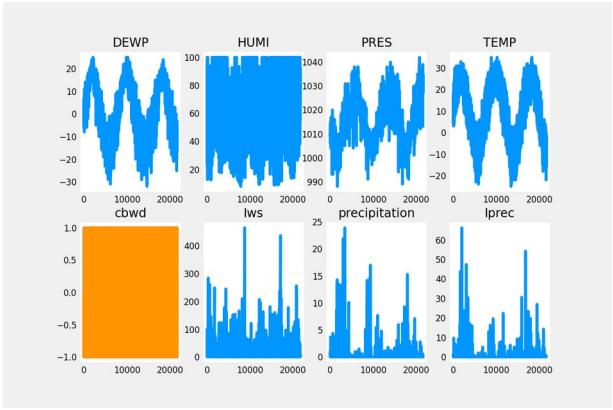
U radu će biti analizarano obeležje PM_US Post, dok će obeležja PM_Taiyuanjie i PM_Xiaoheyan biti uklonjena.

U datoj tabeli, prikazani su postojeći podaci za svaki sat i svaki dan od 1. januara 2010. do 31. decembra 2015. Međutim, nisu u svakoj vrsti sva obeležja popunjena. Tako na primer, postoje nedostajući podaci za PM_US Post - 30904 uzoraka je prazno, odnosno oko 58.77%. Pored toga, postoji jedna vrsta kod koje nedostaje većina podataka za meteorološke uslove, a da ima popunjeno PM_US 2.5 obeležje. Sve te vrste će biti uklonjene iz daljeg razmatranja. Takođe, postoji veći broj uzoraka kod kojih nisu popunjeni podaci za precipitation (5.66%) i lprec (5.66%). Međutim, pošto većina postojećih podataka za ta obeležja ima nultu vrednost (ne pada kiša, slika 1), nepostojeći podaci će biti popunjeni sa nulama.



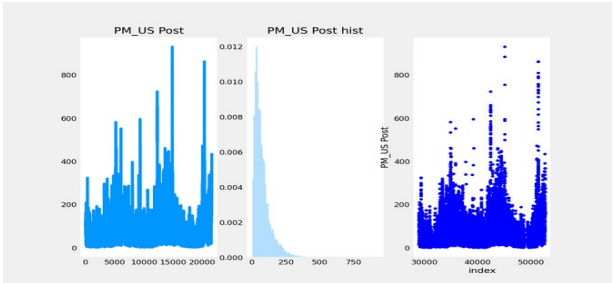
Slika 1: Histogram koncentracije obeležja

Vrednosti su smislene za temperaturu (nikad ne prelazi najveće vrednosti ikad izmerene u tom gradu [3, 4]), smer vetra ima smisla da pretežno ide ka zapadu (jer je na istoku planina, vetar duva sa planine), vlažnost vazduha ima smisla iako vrednost može dostići 100% [5], kao i vazdušni pritisak. Međutim, brzina vetra dostiže 465 m/s, što iznosi 1674 km/h. Koncentracije PM_US 2.5 čestica dostižu vrednosti od 932 $\mu g/m^3$, iako se najčešće te vrednosti mere do 500 $\mu g/m^3$ [6]. Svi ti unosi su ostavljeni zbog nepoznavanja načina dobijanja tih podataka i moguće razlike u razumevanju podataka i njihovog stvarnog značenja. Nakon svih izmena podataka, ostalo je 21679 uzoraka.



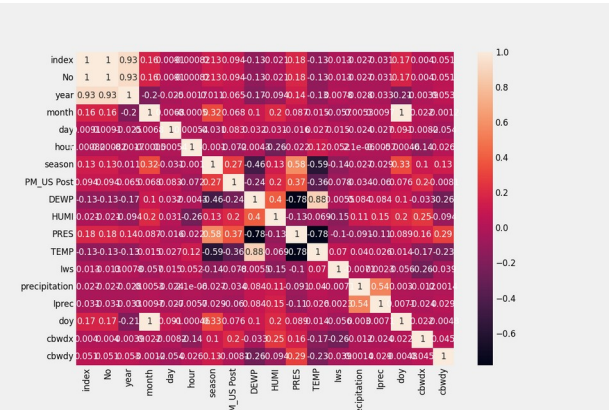
Slika 2: Promena meteoroloških parametara tokom vremena

Na slici 2 su prikazani meteorološki podaci tokom vremena. Može se primetiti da temperatura rose, pritisak i temperatura imaju sinusni oblik što odgovara periodi od 365 dana (promena godišnjih doba). Vlažnost vazduha ima takođe blagi sinusni oblik sa većom frekvencijom i mnogo većim šumom. Brzina vetra je pretvorena iz tekstualnog oblika u numerički: cwbdx ima vrednost između -1 i 1, u zavisnosti od toga da li vetar duva ka zapadu i istoku, i slično cwbdy ima vrednosti između -1 i 1 u zavisnosti od toga da li vetar duva ka severu i jugu (ili miruje, kad je vrednost 0). U donjem levom uglu na slici dva možemo videti da se smer vetra stalno smenjuje. Takođe, brzina vetra i količina padavina imaju naizgled uniformni izgled i neka korelacija se ne može odmah zaključiti.



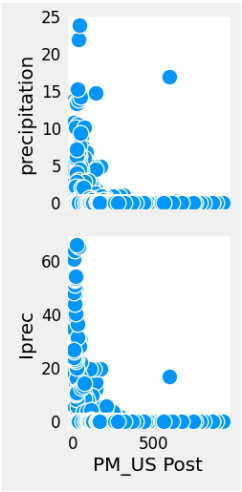
Slika 3: Linearna interpolacija, histogram i promena koncentracije PM 2.5 čestica tokom vremena

Na slici 3 je prikazan histogram dobijenih vrednosti za PM 2.5 čestice u zavisnosti od koncentracije (srednja slika). Funkcija ima očekivani oblik, odnosno ima više uzoraka sa malom koncentracijom čestica nego sa velikom.



Slika 4: Korelisanost između obeležja

Međusobne zavisnosti između svih obeležja su analizirana korelacijom, odnosno toplotnom mapom (eng. Heatmap), slika 4. Sa date slike možemo primetiti da postoji velika srazmerna zavisnost između pritiska i temperature, kao i velika obrnuta srazmernost datih obeležja sa sezonom i temperaturom rose. Takođe, postoji primetna zavisnost između količine padavina na sat i kumulativnih padavina. Pošto sa date toplotne mape, najveća korelisanost postoji između temperature, pritiska, temperature rose i sezone sa koncentracijom PM 2.5 čestica, ta obeležja će se primarno koristiti za regresioni model. Postoji veća korelisanost između duvanja vetra po osi istok-zapad sa koncentracijom čestica nego po osi sever-jug, što je povezano sa činjenicom da se na istoku nalazi planina i da sa istoka najčešće duva vetar. Takođe, primećeno je da veće vrednosti PM 2.5 čestica nastaju gotovo isključivo pri malim količinama padavina (slika 5), što će biti korisnija informacija u klasifikaciji.



Slika 5: Zavisnost između količine padavina i koncentracije PM 2.5 čestica

II. Linearna regresija

Podaci su podeljeni na nasumičan način u sledećem odnosu: 70% trening skup, 15% validacioni i 15% test skup.

Zbog toga što postoji međusobna korelisanost između temperature, pritiska i temperature rose, uvedena je interakcija između tih obeležja u vidu množenja njihovih vrednosti (proizvod svakog sa svakim, kao i proizvod sve tri vrednosti zajedno za svaki uzorak). Temperature su prvo skalirane na pozitivne vrednosti kako ne bi bilo problema prilikom množenja negativnih brojeva.

Testirani su različiti modeli za linearnu regresiju, varirano je sledeće:

- metod učenja, tj. da li se koristi metod najmanjih kvadrata (eng. Mean Square Error) ili gradijentni silazak (eng. Gradient Descent);
- regularizacija, tj. da li se koristi Ridge ili Lasso metode;
- da li su podaci normalizovani ili ne;
- koeficijent učenja alfa – testiralo se za vrednosti 0.1, 0.01, 0.001, 0.0001;
- obeležja koja su se koristila;
- da li postoji interakcija između obeležja;
- da li postoji polinomijalna zavisnost i kog je stepena.

Prilikom evaluacije modela, mereni su sledeći podaci:

- srednja kvadratna greška;
- srednja apsolutna greška;
- RMSE (koren srednje vrednosti sume kvadrata);
- R2 norma;
- korigovana R2 norma.

U ovom radu se tražio model koji je minimizovao srednju kvadratnu grešku, međutim lako se može izmeniti kod kako bi se minimizovalo po nekoj drugoj normi.

Kako bi se vršila selekcija obeležja unapred ili unazad, korišćena je p-vrednost.

Model sa najboljim rezultatom srednje kvadratne greške iz validacionog skupa ima sledeće parametre:

Tabela 1: Parametri dobijenog modela

Parametri	Vrednost
Metoda	MSE
Regularizacija	Ne
Normalizacija	Da
Stepen polinoma	5
Koef. učenja	-
Osobine	TEMP, PRES, DEWP, season, HUMI, cbwdx

Parametri	Vrednost
Mean Square Error	2954.75
Mean Absolute Error	37.34
Root MSE	54.36
R2 score	0.33
R2 adjusted score	0.33

III. KNN klasifikator

IV. Formatiranje

A. Glavne dimenzije

Izveštaj treba da je pripremljen na A4 formatu (210 mm x 297 mm).

Rad treba pripremiti u dve kolone sa 5 mm razmaka između kolona. Strane ne treba numerisati.

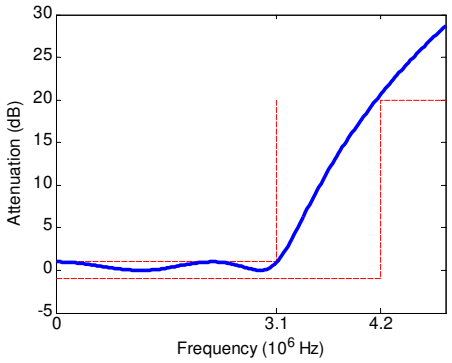
Prvi red svih paragrafa treba uvući za 3,6 mm. Za tekst u radu koristite Times New Roman 10 pt font.

B. Ostale dimenzije

Naslov izveštaja i ime autora treba da budu na prvoj strani centrirani po celoj širini strane. Veličina slova za naslov je 24 pt. Ime autora treba da bude ispod naslova sa veličinom slova 11 pt.

Tabela 1: Opis izgleda strane.

Format papira	A4
Gornja margina	20,0 mm
Donja margina	20,0 mm
Leva margina	19,0 mm
Desna margina	19,0 mm



Sl. 1. Ispod slike treba da stoji objašnjenje.

V. Ocenjivanje

Domaći zadatak nosi 25 bodova. Tih 25 bodova raspoređeno je na sledeći način: 10 bodova nosi sam izveštaj – preciznost, jasnoća, formatiranje i potpunost urađenog zadatka – da li je u kodu urađeno sve što je u zadatku traženo i da li je tačno urađeno (kod može biti pitan i na odbrani); 5 bodova nosi sam kod, a koji može biti proveravan na samoj odbrani (ispitivanje značenja funkcija, parametara ili da se traži da se nešto izmeni u kodu na licu mesta); preostalih 10 bodova nosi usmena odbrana – asistentova i profesorova procena studentovog razumevanja zadatka koje će biti provereno kroz usmena pitanja na samoj odbrani.

Iako su bodovi raspoređeni kako je u prethodnom pasusu napisano, bez predaje izveštaja kao ni bez usmene odbrane, student ne može dobiti bodove za domaći zadatak (dobija 0 bodova bez mogućnosti naknadne predaje).

VI. Zaključak

Zaključak nije neophodan. Iako zaključak može da sadrži pregled ključnih rezultata i objasni njihov značaj, nemojte da ponavljate deo opisan u uvodu. Nije loše navesti i korišćenu literaturu.

VII. Literatura

[1] https://www.epa.gov/pm-pollution/particulate-matter-pm-basics, pogledano dana 9. januar 2023.

[2] https://www.news-medical.net/health/The-Size-of-SARS-CoV-2-Compared-to-Other-Things.aspx, pogledano dana 9. januar 2023.

[3] https://www.extremeweatherwatch.com/cities/shenyang/lowest-temperatures, pogledano dana 9. januar 2023.

[4] https://www.extremeweatherwatch.com/cities/shenyang/highest-temperatures, pogledano dana 9. januar 2023.

[5] https://www.quora.com/What-is-the-highest-humidity-ever-recorded, pogledano dana 9. januar 2023.

[6] https://bewell.co.th/pm-2-5/#:~:text=Acceptable%20levels,truly%20safe%20level%20of%20PM2., pogledano dana 9. januar 2023.