Izveštaj podataka - koncetracija PM2.5 čestica

Jovana Nedeljković, in32-2019, jovananedeljkovic30@gmail.com

I. Opis baze podataka

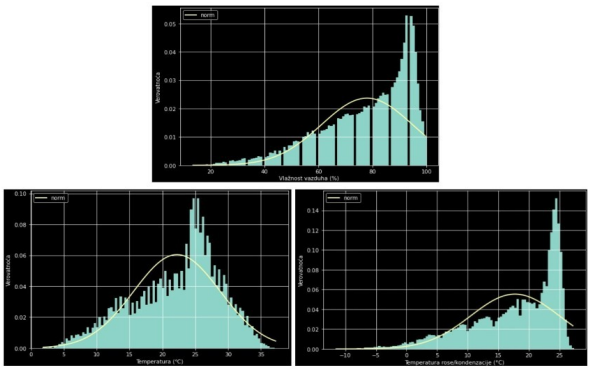
Ovaj izveštaj se bavi analizom podataka vezanih za koncetraciju PM2.5 čestica na nekoliko lokacija. Baza sadrži podatke o 52.584 merenja. Posmatrana su obeležja: vreme (sezona, godina, mesec, dan i sat), kondenzacija, temperatura, vlažnost vazduha, vazdušni pritisak, pravac vetra, kumulativna brzina vetra, padavine na sat i kumulativne padavine.
Cilj rada je da se uoče eventualne pravilnosti i zavisnosti između koncetracije čestica i prirodnih faktora na lokacijama koncentrisanih čestica.

II. Analiza podataka

Prilikom analize podataka, izostavljeni su podaci gde je vrednost koncetracije PM2.5 ('PM_US Post') čestica nepoznata, takođe i podatak gde je vrednost sezone 0, zato što ima nedostajuće vrednosti i za druga obeležja. Izostavljeni su i podaci gde vlažnost vazduha i kondezacija imaju nevalidne vrednosti. Baza je svedena na 32.351 podatak sa 14 obeležja, od kojih je jedno kategoričko.

A. Raspodele obeležja

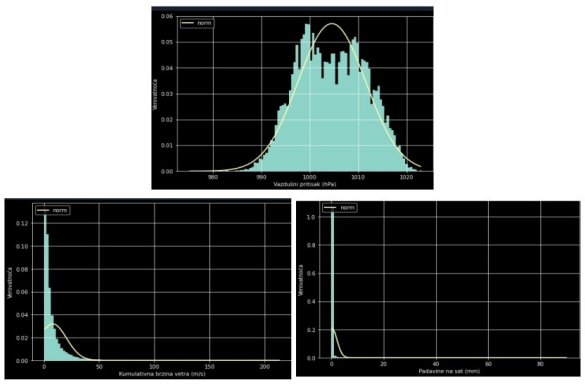
Na slici 1, prikazane su raspodele obeležja sa negativnom asimetrijom i upoređene su sa normalnom raspodelom. To su obeležja: temperatura, vlažnost vazduha i kondenzacija. Temperatura je sa malim koeficijentom spljoštena u odnosu na normalnu raspodelu, ostala dva obeležja su izdužena u odnosu na istu.



Slika 1. Obeležja sa negativnom raspodelom

Na slici 2, prikazana je raspodela ostalih numeričkih obeležja. Na gornjem delu slike nalazi se raspodela obeležja – vazdušni pritisak, koja se odlikuje simetrijom i spljoštanošću u odnosu na normalnu raspodelu. Na donjem delu slike prikaza su obeležja – kumulativna brzina vetra (levo) i padavine na sat (desno). Obe su izrazito pozitivno simetrične i izdužene u odnosu na normalnu raspodelu.

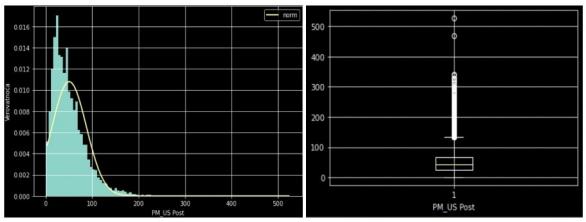
Uticaj obeležja – kumulativne padavine zanemarljivo se razlikuje od uticaja obeležja padavine na sat, pa je izostavljeno u analizi podataka zbog ponavljanja.



Slika 2. Raspodela atributa

B. Detaljna analiza obeležja PM_US POST

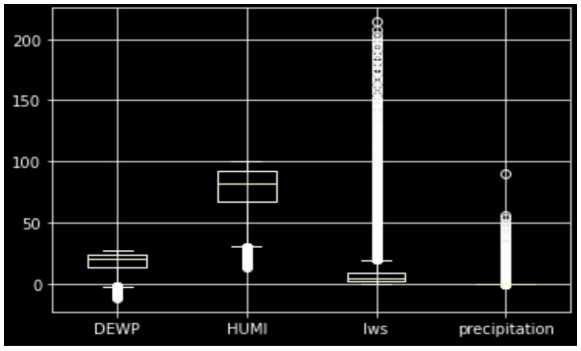
Na slici 3 je prikazana detaljna analiza obeležja PM_US POST. Obeležje koncetracije PM2.5 čestica ima raspodelu koja je pozitivno simetrična i izdužena u odnosu na normalnu raspodelu. Odlikuje se outlier-ima čija vrednost doseže i do 8 puta veće od one koju sadrži oko 50% podataka. Ta vrednosti (43), manja je od dinamičkog opsega (525) i znači da je interkvartilni opseg značajniji od ovog opsega. Srednja vrednost je 50,8 i standardna devijacija 36,93.



Slika 3. Analiza obeležja PM2.5

C. Ouliers i interkvartilni opseg

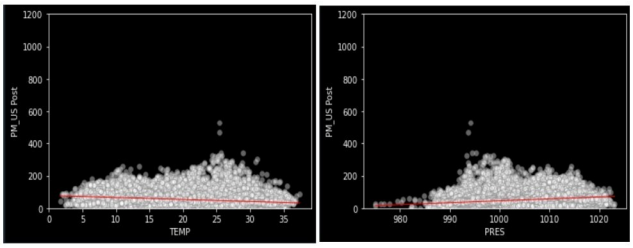
Na slici 4 su prikazani boxplot-ovi obeležja – kondenzancija, vlažnost vazduha, kumulativna brzina vetra i padavine na sat, zbog izraženog postojanja outlier-a koji se sa njih primećuju. Osim kod prvog obeležja, sa boxplot-ova se takođe vidi da je bolje analizu vršiti po interkvartilnom opsegu ostala 3 obeležja, jer se primećuje da se preko 50% vrednosti nalaze blizu gornje granice (kod drugog obeležja) i blizu donje kod ostala dva.



Slika 4. Obeležja sa izrazitim outlier-ima

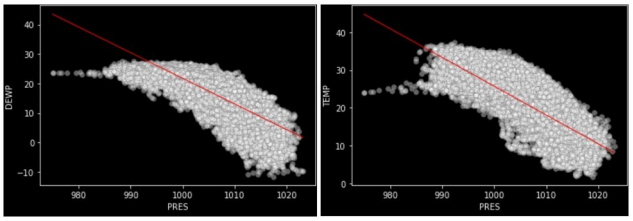
D. Korelacije između obeležja

Na slici 5, prikazane su korelacije između PM2.5 čestica i, redom, temperature i vazdušnog pritiska. Obe korelacije su slabe i čestice nemaju boljih korelacija od njih ni sa jednim drugim atributom.



Slika 5. Korelacija PM2.5 sa dva atributa

Na slici 6, prikazani su parovi atributa sa najvećim korelacijama – kondenzacija i temperetura i vazdušni pritisak i temperatura.



Slika 6. Atributi sa najvećim korelacijama

III. LINEARNA REGRESIJA

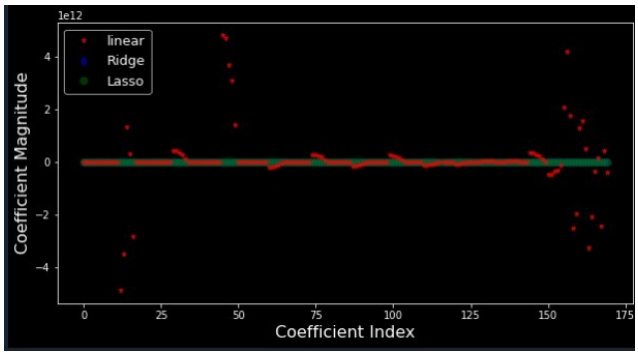
Model predviđa koncentraciju PM2.5 čestica.

Model je prvo treniran osnovnim oblikom linerne regresije, zatim metodom selekcije obeležja unazad. Izvršena je standardizacija obeležja i ponovljeno je treniranje.

Utvrđeno je da postoji korelacija između obeležja i da ima smisla gledati njihov zajednički uticaj. Zbog toga su isprobane hipoteze, korišćenjem *PolynomialFeatures*, sa i bez definisanja parametra *degree* u okviru nje.

Na kraju je izvršena regularizacija korišćenjem metoda *Lasso* i *Ridge*.

Rezultati su prikazani na slici 7.



Slika 7. Linerna regresija

Izabrani model je Lasso. Ovom metodom obezbeđeno je da su koeficijenti ograničeni, čime je sprečeno nadprilagođavanje modela podacima. Obeležjima s malim koeficijentima su dodeljene nule za nove vrednosti istih i time je izvršena selekcija. Srednja kvadratna greška i srednja apsolutna greška su najmanje, a koeficijent determinacije je najveći.

IV. KNN KLASIFIKATOR

Napravljen je klasifikator koji koristi KNN metodu za klasifikaciju uzorka u jednu od tri klase.

Metodom unakrsne validacije određeni su najpovoljniji parametri za obučavanje modela (za *metric – manhattan,* za n_*neighbors* - 14)

A. Akumulirana matrica konfuzije

Akumulirana matrica konfuzije dobijena je akumulacijom matrica konfuzije iz svih iteracija unakrsne validacije i prikazana je na slici 8.

	0	1	2
0	17647	84	0
1	262	8878	11
2	0	23	589

Slika 8. Akumulirana matrica konfuzije

Analiza matrice:

- 17.647 bezbednih čestica predviđenih kao bezbednih (u matrici: indexu 0, odgovara labela dodeljena bezbednim česticama)
- 8.878 nebezbednih čestica predviđenih kao nebezbednih (u matrici: indexu 0, odgovara labela dodeljena nebezbednih česticama)
- 589 opasnih čestica predviđenih kao opasnih (u matrici: indexu 0, odgovara labela dodeljena opasnim česticama)
- 80 bezbednih čestica čestica predviđenih kao nebezbednih
- 0 bezbednih čestica predviđenih kao opasnih

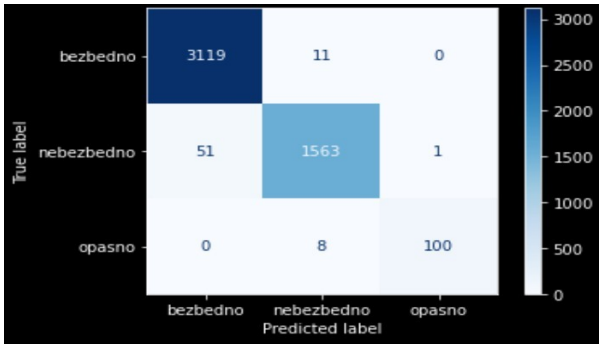
- 279 nebezbednih čestica predviđenih kao bezbednih
- 8 nebezbednih čestica predviđenih kao opasnih
- 0 opasnih čestica predviđenih kao bezbednih
- 25 opasnih čestica predviđenih kao nebezbednih

Iz ove matrice računa se prosečna tačnost klasifikatora koja iznosi: 0.98, dok tačnosti svake od klasa iznose:

1) za bezbedne čestice – 0.99;
2) za nebezbedne čestice – 0.97;
3) za opasne čestice – 0.96.

B. Konačna matrica konfuzije

Primenom metode na 85% dataset-a i njenim testiranjem na ostalih 15%, dobija se konačna matrica konfuzije, prikazana na slici 9.



Slika 9. Finalna matrica konfuzije

Analiza matrice:

- 3.119 bezbednih čestica predviđenih kao bezbednih
- 1.563 nebezbednih čestica predviđenih kao nebezbednih
- 100 opasnih čestica predviđenih kao opasnih
- 11 bezbednih čestica čestica predviđenih kao nebezbednih
- 0 bezbednih čestica predviđenih kao opasnih
- 51 nebezbednih čestica predviđenih kao bezbednih
- 1 nebezbednih čestica predviđenih kao opasnih
- 0 opasnih čestica predviđenih kao bezbednih
- 8 opasnih čestica predviđenih kao nebezbednih

Mere uspešnosti:

1) preciznost – 0.99, što znači da je udeo pravih pozitiva među svim pozitivima 99%,
2) tačnost – 0.98, što znači da je udeo tačno klasifikovanih uzoraka 98%,
3) osetljivost – 0.97, što znači da je udeo pravih pozitiva među onome što je originalno bilo pozitiv 97%,
4) specifičnost – 0.99, što znači da je udeo pravih negativa među onome što je originalno bilo negativ 99%,
5) F score – 2.0, predstavlja meru tačnosti testa.

Mere uspešnosti po klasama prikazane su na slici 10.

	precision	recall	f1-score	support
bezbedno	0.98	1.00	0.99	3130
nebezbedno	0.99	0.97	0.98	1615
opasno	0.99	0.93	0.96	108

Slika 10. Mere uspešnosti modela po klasama

Analiza:

1) 98% bezbednih čestica koje je model predvideo da su bezbedne je dobro klasifikovano
1- 1) 99% nebezbednih čestica koje je model predvideo da su nebezbedne je dobro klasifikovano
1- 2) 99% opasnih čestica koje je model predvideo da su opasne je dobro klasifikovano
2) Od svih čestica koje su zapravo bezbedne, model je tačno predvideo svih 100%
2-1) Od svih čestica koje su zapravo nebezbedne, model je tačno predvideo 97% njih
2-2) Od svih čestica koje su zapravo bezbedne, model je tačno predvideo 93% njih

Klasifikator tačno prevideo skoro sve podatke, što se zaključuje iz vrednosti f1-score-a, koji je blizu 1.