Predikcija kvaliteta vazduha

Matej Mihailović, IN 19/2020, mihailovicmatej@gmail.com

Dušan Stojanović, IN 16/2020, dusan.a.stojanovic@gmail.com

I. Uvod

Predviđanje kvaliteta vazduha igra ključnu ulogu u zaštiti javnog zdravlja, zaštiti životne sredine i suočavanju sa globalnim izazovima kao što su klimatske promene. Preciznim predviđanjem kvaliteta vazduha, nadležni organi mogu donositi bitne odluke za prevenciju ovog problema i preduzeti proaktivne mere za poboljšanje kvaliteta vazduha i poboljšanje života zajednica koje su nastanjene u oblastima gde je vazduh zagađen. Sa ovim problemom se i sami suočavamo u našim gradovima, na ovim prostorima.

II. Baza podataka

Baza podataka sadrži 9358 uzoraka tj. prosečnih merenja po satu, od pet hemijskih senzora metalnih oksida, koji se nalaze u okviru višesenzorskog uređaja koji meri kvalitet vazduha. Ovi uređaji su pozicionirani na nivou puta, unutar značajno zagađene oblasti italijanskog grada. Podaci su zabeleženi u periodu od marta 2004. do februara 2005. godine, što predstavlja najduže slobodno dostupne snimke takvih senzora. Ova baza sadrži prosečne koncentracije po satu za CO, nemetanske ugljovodonike, benzen (C6H6), azotne okside (NOk) i azot-dioksid (NO2). Vrednosti koje nedostaju su označene sa -200. Na početku imamo 15 obeležja, među kojima su datum i vreme merenja, pri čemu su sva numerička. Obeležje koje predviđamo je C6H6 (benzen), veoma otrovna supstanca, za koju je utvrđeno da dovodi do razvoja kancera.

III. Analiza obeležja

U bazi imamo sveukupno 18 411 nedostajućih vrednosti. Kolona „NMHC(GT)“, koja opisuje merenja nemetanskih ugljovodonika, sadrži oko 90% nedost. vrednosti. S obzirom da imamo kolonu sa informacijama o koncentracijama benzena (C6H6) (benzen je ugljovodonik) kroz kolone C6H6(GT) i PT08.S2(NMHC), tako da izbacivši „NMHC“ kolonu, nismo u potpunosti izgubili merenja ugljovodonika. Nedost. vrednosti ostalih kolona smo najpre probali da dopunimo na naivan način, uzevši mean vrednost svake kolone, međutim kao mnogo bolji pristup pokazala se dopuna vrednosti kNN metodom. KNNImputer koristi metod k-najbližih suseda za popunjavanje nedostajućih vrednosti u skupu podataka, prvo izračunavajući rastojanje između tačaka (euklidsko rastojanje u ovom slučaju), zatim koristeći prosečnu vrednost tih tačaka kako bi imputirao nedostajuće vrednosti u datasetu.

Obeležja kao što su T (temperatura), RH i AH (vlažnost vazduha) nemaju jaku korelaciju sa drugim obeležjima, pa se izbacuju iz razmatranja. Obeležje NMHC ima skoro egzaktnu korelaciju sa obeležjem koje predviđamo, C6H6, što je i očekivano jer je benzen zapravo nemetanski ugljovodonik. Zbog problema multikolinearnosti i velikog broja nedostajućih vrednosti, rešili smo da uklonimo obeležje NMHC.

Urađena je klasična podela na trening i test skup, gde je uzeto 20% podataka test skup, a ostatak kao trening skup.

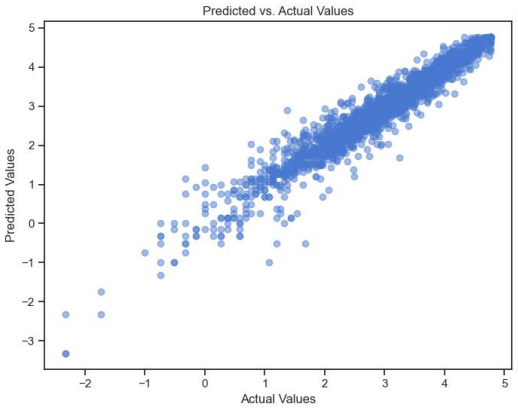
Zatim je urađeno skaliranje obeležja korišćenjem MinMax scaler-a. MinMax scaler transformiše obeležja tako što skalira svako od njih na dati opseg, obično [0, 1] ili [-1, 1] (u slučaju negativnih vrednosti).

IV. Linearna regresija

Linearna regresija je metoda nadgledanog učenja koja se koristi za modeliranje linearnih veza između nezavisnih i zavisnih obeležja. Osnovna pretpostavka linearne regresije je da postoji linearna veza između nezavisnih obeležja (ulaza) i zavisnog obeležja (izlaza). Cilj linearne regresije je pronaći optimalne vrednosti koeficijenata (težine) kako bi se minimizovala razlika između stvarnih i predviđenih vrednosti zavisnog obeležja. Ovo se obično postiže primenom metode najmanjih kvadrata, koja minimizuje kvadratne greške između stvarnih i predviđenih vrednosti. Nakon što su procenjeni koeficijenti, model može biti korišćen za predviđanje vrednosti zavisnog obeležja za nove ulazne podatke.

Parametar linearne regresije:

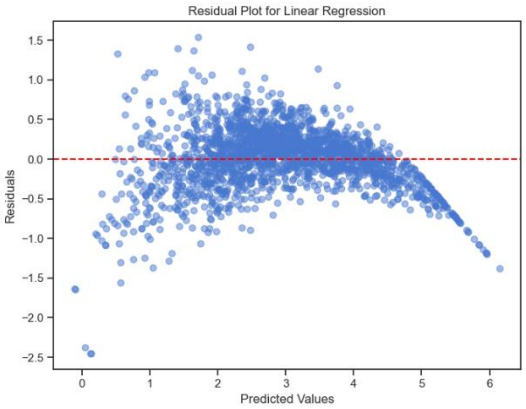
- fit_intercept - određuje da li model ima slobodan član ili ne



Predikcije vs. Prave vrednosti za linearnu regresiju

Dijagram rasipanja pokazuje pozitivnu korelaciju između predviđenih i stvarnih vrednosti. Gusti skup

tačaka duž dijagonalne linije ukazuje na to da su predviđanja modela linearne regresije generalno tačna. Međutim, postoje neki outlier-i, pojedinačne tačke koje značajno odstupaju od dijagonalne linije. One predstavljaju slučajeve u kojima su predviđanja našeg modela bila manje tačna.



Reziduali (Linearna regresija)

Čini se da reziduali imaju obrazac, a ne da su nasumično rasuti oko nule. Postoji zakrivljeni trend u rezidualima, što sugeriše da model linearne regresije možda nije najbolji za moje podatke. Crvena isprekidana linija na nuli predstavlja gde bi bili reziduali da su predviđanja savršena. Još neke od metrika koje su korišćenje za sve modele su: RMSE, MAE, R2 skor. RMSE (Root Mean Square Error) meri prosečnu grešku predviđanja modela. Izračunava kvadratni koren proseka kvadrata razlika između stvarnih i predviđenih vrednosti.

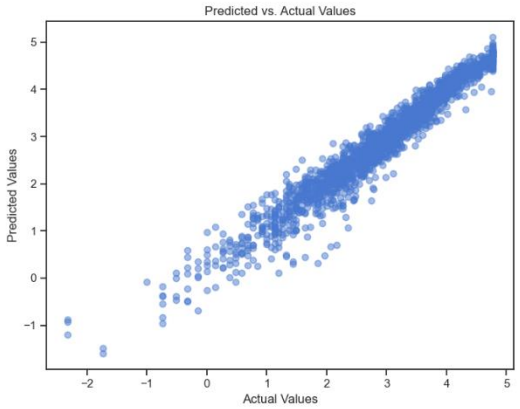
Niže RMSE vrednosti ukazuju na bolje performanse modela. MAE (mean absolute error) predstavlja prosečnu apsolutnu razliku između predviđenih i stvarnih vrednosti. Manje je osetljiv na outlier vrednosti od RMSE. Manje vrednosti MAE ukazuju na bolju moć predikcije modela. R2 ocena meri koliko dobro model objašnjava varijansu u ciljnoj promenljivoj. Ona se kreće od 0 do 1, gde 1 označava savršeno uklapanje. Više vrednosti R2 ocene impliciraju bolje performanse modela.

Tabela 1. Metrike linearne regresije

	RMSE	MAE	R2
Linear	0.185	0.332	0.864

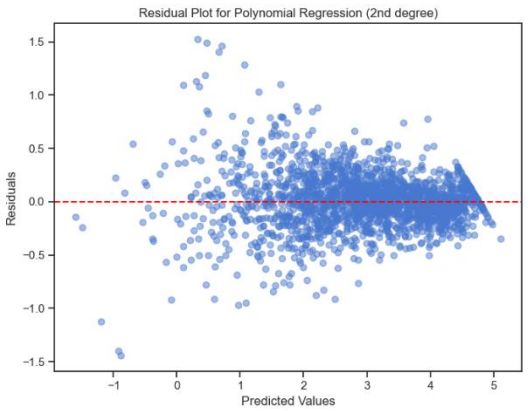
V. Polinomijalna Regresija

Nakon treniranja i testiranja linearne regresije, uradili smo treniranje i testiranje modela koji ne podrazumevaju linearnu vezu između zavisne i nezavisnih promenjivih. Umesto korišćenja linearne funkcije koja pretpostavlja linearnu vezu između promenljivih, polinomijalna regresija koristi polinomsku funkciju za modeliranje odnosa. Prvo smo isprobali model sa stepenom 2 i dobili bolje rezultate nego sa linearnom regresijom.



Predikcije vs. Prave vrednosti za polinomijalnu regresiju

Zakrivljeni trend sugeriše da kvadratni polinom (2. stepena) bolje modelira vezu od jednostavnog linearnog modela. Većina tačaka grupiše se oko regresione linije koja ide nagore. Ovo poravnanje sugeriše da su predviđanja modela generalno tačna. Međutim, postoji određena disperzija oko ove linije, što ukazuje na varijabilnost u predviđanjima.



Reziduali (Polinomijalna regresija)

Disperzija oko crvene isprekidane linije ukazuje da model nije savršen, ali i dalje radi prilično dobro.

Tabela 2. Metrike polinomijalne regresije stepena 2

	RMSE	MAE	R2
2nd Degree	0.072	0.189	0.948

Parametri polinomijalne regresije su:

- Degree – formira polinomijalna obeležja do zadatog stepena
- Include_bias – uvodi i konstantu kao obeležje
- Interaction_only – uvodi interakciju između obeležja, ne i viši stepen obeležja

Za podešavanje hiperparametara polinomijalne regresije, i kasnije stabla odluke, korišćena je tehnika Grid Search, koja uzima kombinaciju svih parametara koje mu navedemo, uradi treniranje korišćenjem kros validacije i to za svaku od ovih

kombinacija i tako odredi koja kombinacija datih parametara daje najbolji rezultat. U našem slučaju urađena je kros validacija sa 5 particija. Dobijeno je da najbolje rezultate daje polinomijalna regresija stepena 4, bez dodatne konstante i interakcije između obeležja. Ridge i Lasso regularizacije su tehnike koje se koriste u regresiji radi smanjenja natprilagođenosti modela. Ridge regularizacija dodaje kvadratne zbir koeficijenata u funkciju gubitka, dok Lasso regularizacija koristi apsolutne vrednosti koeficijenata. U našem slučaju, polinomijalna regresija stepena 4, postigla je najbolje rezultate uz korišćenje Ridge regularizacije.

Tabela 3. Metrike polinomijalne regresije stepena 4

	RMSE	MAE	R2 Score
4th Degree (No reg.)	0.052	0.163	0.962
4th Degree (Ridge)	0.051	0.162	0.962
4th Degree (Lasso)	0.060	0.174	0.956

VI. Stablo Odluke

Stablo odlučivanja je nadgledani algoritam mašinskog učenja koji se koristi i za zadatke klasifikacije i za zadatke regresije.

On konstruiše strukturu stabla nalik dijagramu toka gde svaki unutrašnji čvor predstavlja osobinu, grane označavaju pravila, a u listovima se nalaze rezultati algoritma (oznaku klase ili numeričku vrednost).

Stabla odlučivanja se konstruišu rekurzivnom podelom podataka obuke u podskupove na osnovu vrednosti atributa. Algoritam bira najbolji atribut za podelu, sa ciljem da se maksimizuje dobitak informacija ili smanji nečistoća.

Stabla odlučivanja su laka za razumevanje i vizuelizaciju, ne pretpostavljaju specifične distribucije podataka i rukuju i sa kategoričkim i sa numeričkim podacima.

Parametri stabla odluke:

- Criterion - kriterijum podele; podrazumevana vrednost kod klasifikacionog stabla je gini, a može se odabrati i entropy
- Splitter - određuje strategiju koja se koristi za odabir podele u svakom čvoru; best (podrazumevano) bira najbolju podelu na osnovu kriterijuma kvaliteta (npr. Gini impurity ili entropy); random nasumično bira najbolju podelu, korisno za smanjenje preopterećenja.
- Max_depth - Parametar max_depth kontroliše maksimalnu dubinu stabla odlučivanja; Ako je podešeno na None, čvorovi se proširuju dok svi listovi ne budu čisti ili ne sadrže manje uzoraka od min_samples_split
- Min_samples_leaf – postavlja minimalni broj uzoraka koji je potreban da se nalaze na lisnom čvoru. Tačka podele čvora se uzima u obzir samo ako ostavlja najmanje ovoliko uzoraka za obuku u svakoj grani; pomaže da se izgladi model, posebno u zadacima regresije
- Min_samples_split – određuje minimalni broj uzoraka potrebnih za razdvajanje unutrašnjeg čvora
- Max_features - određuje broj karakteristika koje se uzimaju u obzir kada se traži najbolja podela čvora

Za pronalaženje hiperparametara stabla odluke je korišćen takođe Grid Search. U našem slučaju, kao najbolji parametri su se pokazali:

- criterion: friedman_mse,
- max_depth: None,
- max_features: sqrt,
- min_samples_leaf: 5,
- min_samples_split: 2,
- splitter: best

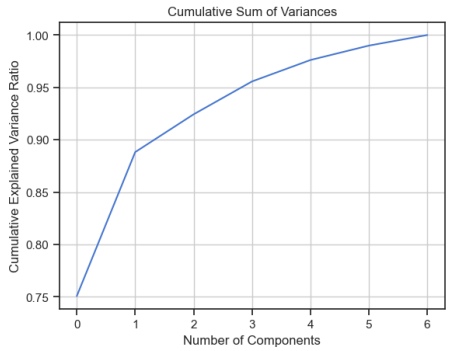
Tabela 4. Metrike stabla odluke

	RMSE	MAE	R2 Score
Decision Tree	0.096	0.227	0.929

VII. PCA

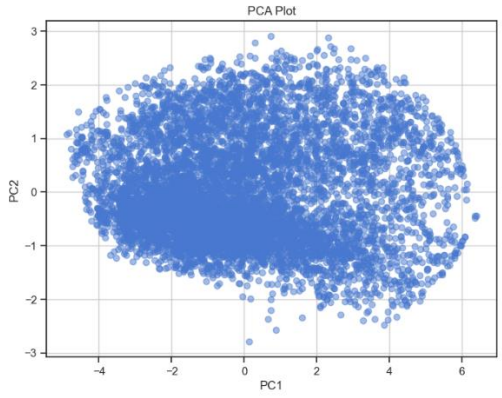
PCA (Principal Component Analysis) je metoda nenadgledanog učenja koja ima za cilj smanjenje dimenzionalnosti uzoraka. Osnovna svrha PCA je da transformiše originalni prostor podataka u novi prostor manjih dimenzija, omogućavajući prikazivanje podataka u prostoru sa manje komponenti, dok se istovremeno pokušava zadržati što veći deo informacija. Ova tehnika omogućava lakšu analizu, interpretaciju i vizualizaciju podataka.

Formiranje PCA komponenti (glavne komponente) uključuje standardizaciju podataka radi svođenja srednje vrednosti na 0, nakon toga sledi izračunavanje kovarijansne matrice i određivanje njenih karakterističnih vektora i vrednosti. Vektor kojem odgovara najveća karakteristična vrednost biće proglašen za prvu PCA komponentu (PC1) , dok će svaka sledeća PCA komponenta odgovarati sledećoj najvećoj karakterističnoj vrednosti.



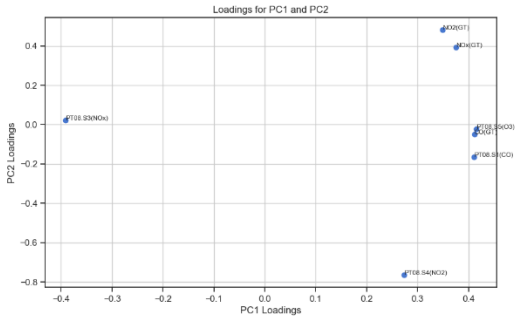
Kumulativna suma varijansi

PC1 i PC2 su prve dve PCA komponente i njihova kumulativna suma od oko 89% ukazuje da ove dve komponente zajedno objašnjavaju 89% ukupne varijanse u podacima. Ovo nam govori da su PC1 i PC2 značajne u predstavljanju i opisu podataka, čime se omogućava efikasno smanjenje dimenzionalnosti podataka bez gubitka značajnog dela informacija.



Raspored uzoraka u prostoru PC1 i PC2

Ovaj grafik predstavlja raspored svih uzoraka u novom prostoru koji je definisan komponentama PC1 i PC2 dobijenim primenom PCA algoritma. Vizualizacija ovog prostora omogućava uvid u raspored podataka i eventualne obrasce ili grupe koje se mogu identifikovati, što nam u ovom slučaju nije potrebno.



Raspodela težina obeležja za PC1 i PC2

Ovaj grafik prikazuje raspodelu težina (loadings), odnosno doprinosa originalnih obeležja u formiranju PC1 i PC2 u PCA analizi. Svaka tačka na grafiku predstavlja jedno od originalnih obeležja, a njena pozicija određena je sa koliko i u kom pravcu to obeležje doprinosi varijabilnosti u PC1 i PC2 komponentama. Obeležja koja su u neposrednoj blizini su pozitivno korelisana. Takođe možemo videti da sva obeležja doprinose formiranju PC1, dok nekoliko obeležja nema toliko značajan uticaj na formiranje PC2.

Tabela 4. Metrike polinomijalne regresije stepena 4 za PC1 i PC2

	RMSE	MAE	R2 Score
4th Degree (PCA)	0.097	0.231	0.929

Iako smo primetili lošije metrike modela, važno je napomenuti da je proces obuke značajno ubrzan zahvaljujući manjoj dimenzionalnosti. Ovo ubrzanje bi bilo još značajnije u slučaju većeg obima podataka ili obeležja.

VIII. Zaključak i rezultati

Tabela 6. Metrike svih modela

	RMSE	MAE	R2 Score
Linear	0.185	0.332	0.864
2nd Degree	0.072	0.189	0.948
4th Degree (No reg.)	0.052	0.163	0.962
4th Degree (Ridge)	0.051	0.162	0.962
4th Degree (Lasso)	0.060	0.174	0.956
Decision Tree	0.096	0.227	0.929
4th Degree (PCA)	0.097	0.231	0.929

Izbor optimalnog modela zavisiće od potrebe za balansom između preciznosti predikcije i složenosti modela. Ipak, u ovom slučaju, model polinomijalne regresije četvrtog stepena sa Ridge regularizacijom se pokazao kao najefikasniji i najpouzdaniji.