Prognoziranje iznadprosečnih vrednosti kvaliteta vazduha u Novom Sadu korišćenjem Random Forest modela

Filip Arnaut, Vesna Cvetkov, Dragana Đurić

PROGNOZIRANJE IZNADPROSEČNIH VREDNOSTI KVALITETA VAZDUHA U NOVOM SADU KORIŠĆENJEM RANDOM FOREST MODELA

Filip Arnaut, Vesna Cvetkov, Dragana Đurić
Univerzitet u Beogradu, Rudarsko-geološki fakultet, Đušina, 7 11000 Beograd
filip.arnaut@rgf.rs

Izvod

Kvalitet vazduha je značajan parametar kvaliteta života i životne sredine celokupno, pa je praćenje emisija suspendovanih čestica u vazduhu od izuzetnog značaja za očuvanje životne sredine i unapređenje kvaliteta života. U ovom radu se istražuje primena mašinskog učenja za prognoziranje iznad ili ispodprosečnih vrednosti zagađujućih materija u vazduhu, sa fokusom na koncentraciju $PM_{2.5}$ čestica. Korišćeni Random Forest model se pokazao kao adekvatno rešenje za problem klasifikacije, sa tačnošću i preciznošću od 77%, čak i bez optimizacije modela. Analiza pokazuje da parametar SO_2 ima zanemarljivu korelaciju sa ostalim parametrima kvaliteta vazduha ($PM_{2.5}$, PM_{10} i NO_2), što omogućava izostavljanje ovog parametra u narednim modelima i smanjenje utrošenog računarskog vremena.

Ključne reči: Klasifikacija, Mašinsko učenje, Statistička analiza kvaliteta vazduha, Novi Sad.

Uvod

Povišene vrednosti zagađujućih materija u vazduhu predstavlja sve veći problem u urbanim sredinama širom sveta, a posebno u gradovima Srbije poput Beograda i Valjeva koji se često nalaze na listi najzagađenijih gradova na svetu (Air Visual). Potrebno je napomenuti da platforma Air Visual za rangiranje gradova po kvalitetu vazduha koristi podatke sa zvaničnih državnih mernih stanica. Ovo je posledica zavisnosti Republike Srbije od fosilnih goriva za proizvodnju energije, neregulisane industrije i povećanja broja motornih vozila. Kvalitet vazduha se najčešće meri kroz nivo suspendovanih čestica koje mogu negativno uticati na zdravlje ljudi, posebno na respiratorni sistem i kardiovaskularne funkcije. To su zagađujuće materije suspendovane u vazduhu u formi čvrstih čestica ili tečnih kapljica koje su značajan polutant u vazduhu sa snažnim uticajem na ljudsko zdravlje. Stoga, bilo kakvi pokušaji razumevanja uzročnika zagađujućih materija u vazduhu, kao i mogućnost njihovog prognoziranja mogu se smatrati značajnim po šire društvo.

Za prognoziranje koncentracije zagađujućih materija u vazduhu sve više se koriste metode prognoziranja vremenskih serija i mašinskog učenja. Primer toga je Prophet model [1-7], kao i određene hibridne metode [8].

U ovom radu je predstavljen pristup koji kombinuje metode mašinskog učenja i podatke o kvalitetu vazduha u Republici Srbiji kako bi se predvideo nivo zagađujućih materija u vazduhu u budućnosti. Za ovaj pristup korišćena je klasifikaciona metoda, a ne regresiona. Rezultati dobijeni ovim pristupom mogu predstavljati kvalitetnu uvodnu studiju za buduća opsežnija istraživanja prognoziranja kvaliteta vazduha u Srbiji. Ova studija se razlikuje od drugih pristupa koji koriste metode prognoziranja vremenskih serija ili hibridne metode. Primena mašinskog učenja na podatke o kvalitetu vazduha može biti korisna za razumevanje uzroka povišenih vrednosti zagađujućih materija u vazduhu i za dalje istraživanje prognoziranja koncentracija zagađujućih materija u vazduhu u Republici Srbiji.

Metodologija i podaci

Deskriptivna statistika predstavlja standardnu proceduru prilikom statističkih istraživanja. Parametri koji spadaju pod deskritpivnu statistiku pomažu prilikom „upoznavanja sa podacima“. Deskriptivnu statistiku mogu označavati vrednosti minimuma i maksimuma podataka, kao i modaliteta, medijane i prosečne vrednosti. Mere koje kvantifikuju disperziju podataka, takođe, pripadaju koracima deskriptivne statistike, kao što su koeficijent varijacije, standardna devijacija i varijansa. Konstrukcija raspodela (histograma) podataka i prikazivanje koeficijenta asimetrije i koeficijenta spljoštenosti spadaju takođe pod deskriptivnu statistiku. U radu sa vremenskim serijama potrebno je poznavati ukupan broj podataka i broj nedostajućih (preskočenih) opservacija da bi se odredio kompletan set podataka i odabrala adekvatna metoda imputacije ili amputacije podataka.

Koeficijenti korelacije predstavljaju kvantitativnu meru korelacije između dve grupe podataka. Pirsonov koeficijent korelacije (r) meri intenzitet i smer linearnog odnosa između dve grupe podataka, dok Spirmanov koeficijent korelacije (r_s) kvantitativno ocenjuje intenzitet i smer korelacije između dve grupe podataka bez pretpostavki o raspodeli ili vrsti povezanosti. Opseg u kom se Pirsonov koeficijent korelacije nalazi je od -1 (negativna korelacija) do 1 (pozitivna korelacija). Vrednost od 0 kod Pirsonovog koeficijenta korelacije predstavlja ne postojanje korelacije izeđu dve grupe podataka. Pretpostavka kod Pirsonovog koeficijenta korelacije je da su podaci u linearnoj zavisnosti i da su obe grupe podataka u normalnoj raspodeli.

Spirmanov koeficijent korelacije (r_s) kvantitativno ocenjuje intenzitet i smer korelacije između dve grupe podataka. Slično Pirsonovom koeficijentu korelacije, opseg u kome se prikazuje korelacija se nalazi od -1 do 1, gde vrednost od 0 se smatra nepostojanjem korelacije. Prilikom istraživanja, korišćena su oba koeficijenta korelacije sa tim da je veći akcenat stavljen na Spirmanov koeficijent korelacije u slučaju ne-normalne raspodele podataka.

Random Forest model prvi put je prikazan 2001. godine [9], i od tada predstavlja jedan od najprimenljivijih metoda mašinskog učenja. Random Forest model svoju primenu je našao u klasifikacionim kao i u regresorskim problemima. Random Forest model zasniva se na kombinaciji većeg broja stabala time da se međusobnim glasanjem stabala dolazi do jedinstvenog rešenja koje je predstavljeno određenom klasom [9-10]. Kao hiperparametar ili kao regularizacioni parametar se može posmatrati broj stabala, sa tim da su veće vrednosti uglavnom bolje ukoliko postoji adekvatno računarsko vreme na raspolaganju [11].

Podaci za ovo istraživanje preuzeti su od strane Agencije za zaštitu životne sredine Republike Srbije. Baza podataka za 2021. godinu sadrži merene parametre kvaliteta vazduha kao što su SO_2, NO_2, $PM_{2.5}$, PM_{10}, O_3, CO i druge za sve merne stanice u Republici Srbiji. Podaci za mernu stanicu Novi Sad- Rumenačka su preuzeti iz baze podataka i korišćeni za ovo istraživanje.

Da bi podaci bili adekvatni za klasifikacioni problem mašinskog učenja, merene vrednosti PM_{10}, $PM_{2.5}$, SO_2 i NO_2 su prebačene u vrednosti klasa (*TRUE* ili *FALSE*) ukoliko je merena vrednost za dati sat veća od prosečne vrednosti tog parametra. Parametar koji je bio prognoziran je koncentracija $PM_{2.5}$, dok su PM_{10}, SO_2 i NO_2 korišćeni kao parametri za klasifikaciju $PM_{2.5}$ vrednosti. Time se omogućila primena Random Forest klasifikacionog modela, odnosno, omogućuje se klasifikacija budućih $PM_{2.5}$ vrednosti na osnovu iznad ili ispodprosečnih vrednosti PM_{10}, NO_2 i SO_2. Takođe, da bi podaci bili prikladni za Random Forest model, podeljeni su na tri grupe: podaci za treniranje modela „*Train*“, podaci za validaciju modela „*Validation*“ i podaci za testiranje modela, odnosno „*Test*“. Podele su izvršene prema 64%- 16%- 20% odnosu. Test grupa podataka korišćena je za verifikaciju i statističku evaluaciju Random Forest modela.

Rezultati i diskusija

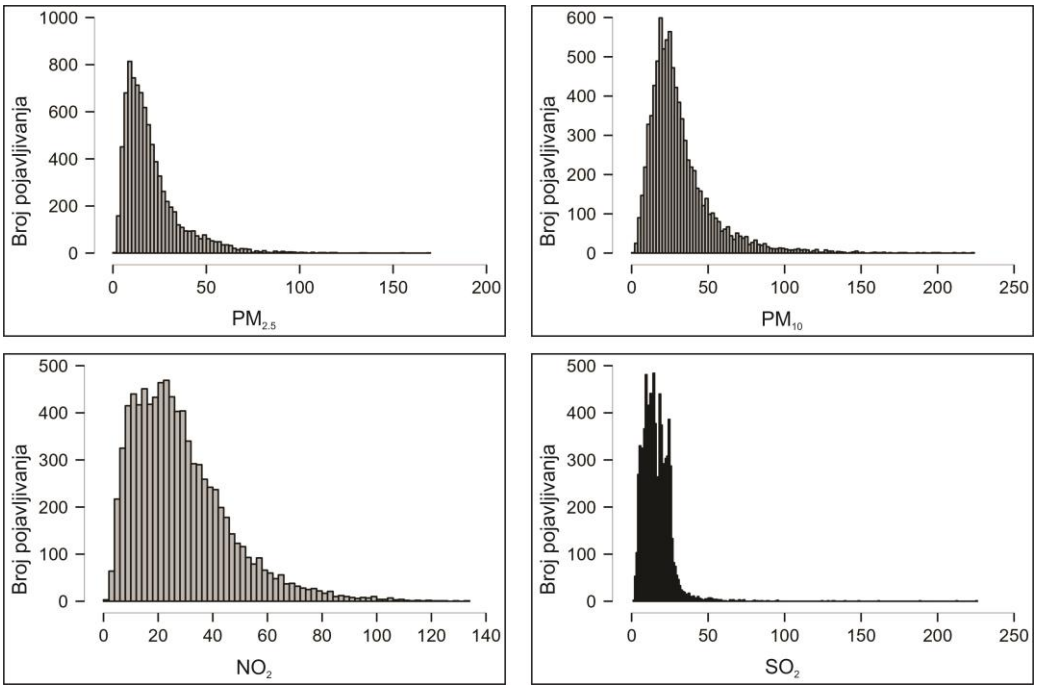
Statistička analiza podataka

U tabeli 1 prikazani su parametri deskriptivne statistike sračunate za koncentracije PM_{10}, $PM_{2.5}$, SO_2 i NO_2 u vazduhu merenoj na mernoj stanici Rumenička u Novom Sadu u 2021. godini. Ukupan broj podataka koji može biti meren u godini (u slučaju da godina nije prestupna) iznosi 8760 podataka, pod pretpostavkom da su podaci mereni sa rezolucijom od 1h, kontinualno. Procenat nedostajućih podataka sa merne stanice Novi Sad- Rumenička nalazi se u opsegu od 1.4% ($PM_{2.5}$ i PM_{10}) do 1.59% (NO_2), što se za analizu primenom metoda mašinskog učenja u ovom istraživanju može smatrati kontinualnim zapisom podataka. Takođe, maksimalan broj uzastopno prekinutih opservacija se nalazi u redu veličine od 100 podataka (sati), zbog toga se, set podataka može smatrati uslovno kontinualnim za svrhe ovog istraživanja, a metode imputacije i amputacije podataka neće biti primenjene.

Tabela 1. Deskriptivna statistika parametara kvaliteta vazduha.

	$PM_{2.5}$ [µg/m^3]	**SO_2 [µg/m^3]**	**NO_2 [µg/m^3]**	**PM_{10} [µg/m^3]**
Broj validnih podataka	8639	8625	8623	8639
Broj nedostajućih podataka	121	135	137	121
Procenat nedostajućih podataka	1.40	1.57	1.59	1.40
Modalitet	10.4	14.5	10.5	24.7
Medijana	16.3	14.8	24.9	26.1
Prosečna vrednost	20.487	16.287	28.489	31.688
Standardna devijacija	15.304	10.064	17.883	21.566
Koeficijent asimetrije	2.153	5.211	1.321	2.397
Koeficijent spljoštenosti	7.332	71.721	2.486	9.356
Minimum	1.81	1.52	1.68	1.83
Maksimum	170	226	134	223

Podaci iz tabele 1 ukazuju na desno-asimetričnu raspodelu svih parametara kvaliteta vazduha, gde je vrednost modaliteta (najčešće vrednosti) manja od vrednosti medijane podataka, koja je pak manja od prosečne vrednosti. Koeficijenti asimetrije za sve četiri koncentracije ukazuju na desno-asimetričnu raspodelu, kao i izrazito velike vrednosti maksimuma u odnosu na prosečnu vrednost, što ukazuje na prisustvo ekstremnih vrednosti.



Slika 1. Raspodela parametara kvaliteta vazduha na mernoj stanici Novi Sad- Rumenička za 2021. godinu

Bitan parametar pri većini statističkih istraživanja predstavlja koeficijent korelacije, odnosno mera korelacije između podataka. U tabeli 2 su dati izračunati koeficijenti korelacije primenom Pirsonovog i Spirmanovog koeficijenta korelacije. Sa obzirom da ni jedan prethodno prikazan parametar kvaliteta vazduha nije predstavljen normalnom raspodelom, težinski će se više uzimati značajnost Spirmanovog koeficijenta korelacije. Parametar SO_2 prikazuje zanemarljive koeficijente korelacije sa NO_2 i PM_{10}, dok je sa $PM_{2.5}$ prikazana slaba korelacija. Sa druge strane, najveću (jaku) korelaciju prikazuju parametri PM_{10} i $PM_{2.5}$, dok su NO_2 i PM_{10} i $PM_{2.5}$ predstavljeni srednjom korelacijom.

Tabela 2. Koeficijenti korelacije parametara kvaliteta vazduha.

		$PM_{2.5}$ [µg/m³]	**SO_2 [µg/m³]**	**NO_2 [µg/m³]**
SO_2 [µg/m³]	Pirson (r)	-0.084*		
	Spirman (r_s)	-0.283**		
NO_2 [µg/m³]	Pirson (r)	0.518***	0.018*	
	Spirman (r_s)	0.46***	-0.012*	
PM_{10} [µg/m³]	Pirson (r)	0.834****	0.078*	0.64***
	Spirman (r_s)	0.793****	-0.019*	0.59***

*Zanemarljivo;**Slabo;***Srednje;****Jako;*****Vrlo jako
opsezi koeficijenata korelacije dati prema [12]; p-vrednosti su manje od 0.001

Random Forest klasifikacija

Random Forest klasifikacija je urađena sa podelom podataka tako da podaci za treniranje modela uzimaju ukupno 5606 sati iz godine (oko 64% od ukupnog seta podataka), podaci za validaciju modela uzimaju 1402 sata (16% celog seta podataka), dok podaci za testiranje uzimaju preostalih 1752 sati, odnosno 20% celog seta podataka. Random Forest model je imao ukupno 73 stabla.

Tabela 3 prikazuje matricu konfuzije koja je sačinjena nakon treniranja i validacije modela sa podacima za testiranje. Iz tabele 3 se može videti da model u 47% slučajeva prognozira da će nova merena vrednost $PM_{2.5}$ koncentracije biti manja od srednje vrednosti $PM_{2.5}$ koncentracije za godinu. Tačnu prognozu model daje u 35% slučajeva dok u ostalih 12% istinita vrednost $PM_{2.5}$ koncentracije je veća od prosečne vrednosti. Sa druge strane, model u 53% slučajeva prognozira da će nova merena vrednost $PM_{2.5}$ koncentracije biti veća od prosečne vrednosti. Tačnu prognozu model pravi u 42% slučajeva, dok u preostalih 11% pravi pogrešnu prognozu. Ukupno, u 77% slučajeva model pravi tačnu prognozu, a u 23% slučajeva pogrešnu prema podacima iz Tabele 3.

Tabela 3. Matrica konfuzije za Random Forest klasifikacioni model.

		Prognozirano			
		FALSE [%]	**FALSE**	**TRUE [%]**	**TRUE**
Istinito	**FALSE**	35	615	11	187
	TRUE	12	209	42	741

Da bi se dobila kompletnija slika o sposobnostima modela prikazani su dodatni parametri evaluacije Random Forest modela (Tabela 4). Ukupna tačnost modela (0.774) prikazuje da je model relativno adekvatan, ali sa prostorom za njegovo unapređenje i poboljšanje tačnosti. Parametar preciznosti za „*TRUE*“ klasu uzima veće vrednosti nego za „*FALSE*“ klasu. Drugim rečima, kada model napravi prognozu da novi mereni podatak $PM_{2.5}$ koncentracije neće biti veći od prosečne vrednosti, ima preciznost od 74.6%, dok kada model napravi prognozu da će novi mereni podatak biti veći od prosečne vrednosti njegova preciznost će biti veća sa 79.8%.

Vrednosti F1 mere za obe klase su relativno bliske, što govori o tome da model ima balansiran kompromis između preciznosti i odziva za obe klase. Stopa lažno pozitivnih rezultata prikazuje da model netačno prognozira iznadprosečne vrednosti kvaliteta vazduha u 23.3% vremena kada je stvarna vrednost ispod prosečne, i obrnuto (stopa lažno negativnih rezultata). Statistički paritet za model prikazuje malu pristrastnost ka „*TRUE*“ klasi, odnosno da model ima malo veću verovatnoću prognoziranja novih merenih vrednosti kao iznadprosečnih.

Tabela 4. Evaluacija Random Forest klasifikacionog modela.

	FALSE	**TRUE**	**Prosek/Ukupno**
Broj podataka	802	950	1752
F1 mera	0.756	0.789	0.774
Tačnost	0.774	0.774	0.774
Preciznost	0.746	0.798	0.775
Odziv	0.767	0.78	0.774
Stopa lažno pozitivnih rezultata	0.22	0.233	0.227
Stopa lažno negativnih rezultata	0.233	0.22	0.227
Statistički paritet	0.47	0.53	1

Bitan parametar nakon treniranja i evaluacije modela predstavlja parametar prosečnog smanjenja tačnosti koji u ovom slučaju prikazuje da na vrednosti prognoze iznadprosečnih ili ispodprosečnih vrednosti $PM_{2.5}$ koncentracije najviše utiče parametar koncentracije PM_{10}, nakon čega parametar koncentracije NO_2 i na kraju SO_2.

Drugim rečima, na kvalitet dobijene prognoze najviše doprinose vrednosti koncentracije PM_{10}, a najmanje vrednosti koncentracije SO_2. U velikoj meri se parametar prosečnog smanjenja tačnosti poklapa sa koeficijentom korelacije dobijenim u tabeli 2, gde je prikazano da parametar koncentracije SO_2 ima zanemarljive do slabe korelacije sa ostalim parametrima kvaliteta vazduha prema Spirmanovom koeficijentu korelacije. Sa druge strane, najveću korelaciju parametar $PM_{2.5}$ ima upravo sa parametrom PM_{10}, što u najvećoj meri objašnjava i veliki uticaj parametra PM_{10} na prosečno smanjenje tačnosti. U budućim istraživanjima i primeni Random Forest klasifikacionog modela, moguće je izbaciti parameter SO_2 iz analize zbog njegovog zanemarljivog doprinosa tačnosti. Time bi se, pri analizi velikog broja podataka (na primer istraživanje sa podacima od više godina i parametara), smanjilo računarsko vreme koje je potrebno za analizu.

Zaključak

Kvalitet vazduha predstavlja bitan pokazatelj stanja životne sredine, naročito u urbanim sredinama. Povišene vrednosti zagađujućih materija u vazduhu mogu imati nepovoljne uticaje na respiratorne i kardiovaskularne bolesti kod stanovništva. Poznavanje budućih vrednosti koncentracija zagađujućih materija u vazduhu može biti značajna informacija za ugrožene grupe (decu, stariju populaciju i osobe sa hroničnim bolestima) kao i za celokupno stanovništvo.

U radu je prikazan alternativni postupak primene mašinskog učenja za prognoziranje budućih vrednosti kvaliteta vazduha. Problemu klasifikacije iznadprosečnih vrednosti parametara $PM_{2.5}$ pristupljeno je korišćenjem parametara PM_{10}, SO_2 i NO_2. Rezultati ukazuju da i modeli koji nisu u potpunosti optimizovani mogu dati relativno kvalitetne rezultate. Prikazan Random Forest model ostavlja prostor za dalju optimizaciju. Takođe je utvrđeno da, je moguće i korisno primeniti dodatna istraživanja sa optimizovanim modelima, većim brojem podataka i parametara radi evaluacije prognoze budućih vrednosti kvaliteta vazduha.

Literatura

[1] Taylor, S. J., & Letham, B. (2018). Forecasting at scale. *The American Statistician*, *72*, 37-45.

[2] Zhao, N., Liu, Y., Vanos, J. K., & Cao, G. (2018). Day-of-week and seasonal patterns of $PM_{2.5}$ concentrations over the United States: Time-series analyses using the Prophet procedure. *Atmospheric environment*, *192*, 116-127.

[3] Samal, K. R., Babu, K. S., Das, S. K., & Acharaya, A. (2019). Time series-based air pollution forecasting using SARIMA and Prophet model. *Proceedings of the 2019 international conference on information technology and computer communications*, 80-85.

[4] Ye, Z. (2019). Air pollutants prediction in Shenzhen based on ARIMA and Prophet method. *E3S Web of Conferences*. EDP Sciences.

[5] Shen, J., Valagolam, D., & McCalla, S. (2020). Prophet forecasting model: A machine learning approach to predict the concentration of air pollutants ($PM_{2.5}$, PM_{10}, O_3, NO_2, SO_2, CO) in Seoul, South Korea. *PeerJ*, 8.

[6] Zhoul, L., Chenl, M., & Ni, Q. (2020). A hybrid Prophet-LSTM model for prediction of air quality index. *2020 IEEE Symposium Series on Computational Intelligence,* 595-601, IEEE

[7] Tejasvini, K. N., Amith, G. R., & Shilpa, H. (2020). Air pollution forecasting using multiple time series approach. *Proceedings of the Global AI Congress 2019*, 91-100. Springer Singapore.

[8] Ejohwomu, O. A., Shamsideen Oshodi, O., Oladokun, M., Bukoye, O. T., Emekwuru, N., Sotunbo, A., & Adenuga, O. (2022). Modelling and forecasting temporal $PM_{2.5}$ concentration using ensemble machine learning methods. *Buildings*, *12*, 46.
[9] Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.
[10] Nosek, T., Brkljač, B., Despotović, D., Sečujski, M. & Lončar-Turukalo, T. (2020). *Praktikum iz mašinskog učenja*. Univerzitet u Novom Sadu, Fakultet Tehničkih Nauka, Katedra za telekomunikacije i obradu signala.
[11] Nikolić, M., & Zečević, A. (2019). *Mašinsko učenje*. Beograd: Matematički fakultet.
[12] Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, *126*, 1763-1768.