PRIMENA ISTORIJSKIH PODATAKA SCADA SISTEMA ZA IZRADU MODELA MAŠINSKOG UČENJA U DETEKCIJI I SMANJENJU GUBITAKA VODE

APPLICATION OF SCADA SYSTEM HISTORICAL DATA FOR DEVELOPING MACHINE LEARNING MODELS IN DETECTION AND REDUCTION OF WATER LOSS

OGNJEN IVIĆ[1]

Stručni rad
DOI: 10.5937/GV25077I

Rezime: Upravljanje gubicima vode u vodosnabdevanju je ključno za poboljšanje efikasnosti distribucije i smanjenje troškova. Ovaj rad istražuje primenu istorijskih podataka SCADA sistema i primena mašinskog učenja u izradi modela, sa fokusom na detekciju, analizu i smanjenje gubitaka. Korišćenjem podataka o pritiscima, protoku i ključnim parametrima razvijen je model zasnovan na Random Forest algoritmu koji predviđa gubitke i identifikuje trenutke visokog rizika. Analiza grešaka između predviđenih i stvarnih vrednosti omogućava brzo prepoznavanje nepravilnosti, kao što su curenje ili krađa vode, čime se omogućava proaktivno reagovanje na potencijalne probleme.
Ključne reči: gubici vode, SCADA sistem, prediktivno modeliranje, optimizacija sistema, efikasnost distribucije

Abstract: Water loss management in water supply systems is crucial for improving distribution efficiency and reducing costs. This paper investigates the application of SCADA system historical data and the use of machine learning in model development, with a focus on detection, analysis, and reduction of losses. By using data on pressure, flow, and key parameters, a model based on the Random Forest algorithm has been developed to predict losses and identify high-risk moments. The analysis of residuals between predicted and actual values enables quick detection of irregularities, such as leaks or water theft, allowing proactive responses to potential issues.
Key Words: Water losses, SCADA system, Predictive modeling, System optimization, Distribution efficiency

[1] Ognjen Ivić, JKP „Vodovod i kanalizacija“ Subotica, Trg Lazara Nešića 9/a, Subotica, ognjen@vodovodsu.rs, ORCID: 0009-0005-3005-512X

1. Uvod

U savremenom upravljanju vodovodnim sistemima, poseban izazov predstavlja rano otkrivanje i smanjenje gubitaka vode. Pored tradicionalnih metoda mogu se primeniti i nove metode radi pravovremenog prepoznavanja anomalija koje ukazuju na curenja, kvarove ili neovlašćeno korišćenje vode. U tom kontekstu, mašinsko učenje i vizuelne analize podataka, sve više dobijaju na značaju. Mašinsko učenje predstavlja jedan od najbrže rastućih pravaca u obradi i analizi podataka, posebno u inženjerskim disciplinama. U ovom radu razmatra se primena metoda nadgledanog učenja, konkretno regresionih modela, za procenu i predviđanje izlaznog protoka vode na vodozahvatu u Bačkim Vinogradima u Subotici, a na osnovu dostupnih istorijskih podataka koje beleži SCADA sistem. Ulazni podaci obuhvataju vreme merenja (vreme - „time“) i izlazni pritisak (pritisak - „pressure“), dok se kao ciljna promenljiva koristi izlazni protok (protok - „flow“).

2. Opis skupa podataka i predprocesiranje

Skup istorijskih podataka obuhvata više podataka (oko 300 sirovih datoteka sa približno 86.400 zapisa) pri čemu svaki sadrži tri ključna parametra: vreme merenja („time“), izlazni pritisak („pressure“) i izlazni protok („flow“). Analiza se vrši pojedinačno za svaki istorijski podatak, pri čemu se iz svakog formira poseban model, kao i odabir najboljeg modela od svih. Pre izrade modela, podaci se procesuiraju – uklanjaju se redovi sa nedostajućim vrednostima (NaN), a zatim se vrši podela skupa podataka na trening (80%) i test (20%) uz pomoć funkcije „train test split“. Ova podela je ključna za razvoj pouzdanih modela mašinskog učenja, jer omogućava obuku modela na jednom delu podataka, dok se evaluacija ili ocena vrši na zasebnom, prethodno neviđenom skupu. Time se smanjuje rizik od prenaučenosti („overfitting“) i obezbeđuje realna procena sposobnosti modela da generalizuje na nove podatke.

U ovom radu, ulazni parametri – vreme i pritisak – predstavljaju nezavisne promenljive, dok je protok zavisna promenljiva koju model pokušava da predvidi. Primenjuje se metod nadgledanog učenja („supervised learning“), pri čemu model u fazi obuke uči funkcionalnu zavisnost između ulaznih i izlaznih parametara. Nakon toga, predviđanje se vrši nad test skupom, što omogućava ocenu tačnosti i robusnosti razvijenog modela.

3. Primenjeni modeli

3.1. Primenjeni modeli u procesu modelovanja

U ovom algoritmu testirani su sledeći modeli mašinskog učenja: Linearna regresija, Ridge regresija, Lasso regresija i Random Forest regresor. Svaki od ovih

modela ima specifične karakteristike koje direktno utiču na njihovu sposobnost generalizacije, interpretacije rezultata i efikasnost u različitim kontekstima:

Linear Regression (Linearna regresija) koristi sledeću formulu za modeliranje odnosa između ulaznih (nezavisnih) varijabli X i izlazne (zavisne) varijable Y:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon \quad (1)$$

Y predstavlja ciljnu varijablu (izlazni protok)

$X_1, X_2 \ldots\ldots, X_n$ su ulazne varijable (u ovom slučaju vreme i izlazni protisak)

β_o je slobodni član („intercept“)

$\beta_1, \ldots \beta_n$ su koeficijenti koji pokazuju uticaj svake ulazne varijable na izlaz

$\in$ je slučajna greška

Linearni regresioni model pretpostavlja pravolinijski odnos između ulaznih parametara i ciljne promenljive, pa ne prepoznaje nelinearne obrasce. Zbog složenih veza između pritiska, vremena i protoka, njegova tačnost može biti ograničena. Ipak, model je lako interpretirati jer koeficijenti jasno prikazuju uticaj svakog parametra na promenu protoka.

Ridge regresija: Proširena linearna regresija sa L2 regularizacijom radi sprečavanja prenaučenosti.

$$\hat{\beta} = \arg\min_{\beta} \left(\sum_{i=1}^{m} (y_i - X_i \beta)^2 + \lambda \sum_{j=1}^{n} \beta_j^2 \right) \quad (2)$$

λ parametar koji određuje jačinu regularizacije, tj. koliko model treba da bude jednostavan i izbegne prenaučenost, dok ostatak formule predstavlja zbir kvadratnih razlika između stvarnih i predviđenih vrednosti, što meri tačnost modela na trening podacima.

Ridge regresija koristi L2 regularizaciju kako bi smanjila prenaučenost, tako što kažnjava velike vrednosti koeficijenata. Iako ne može da modeluje nelinearne odnose, poboljšava stabilnost predikcija u prisustvu visoko korelisanih ulaznih varijabli. Pritom zadržava dobru interpretabilnost, što olakšava tumačenje modela.

Lasso regresija: Lasso regresija uvodi regularizaciju korišćenjem L1 norme, za razliku od Ridge regresije koja koristi L2. Ova vrsta kažnjavanja može dovesti do toga da pojedini koeficijenti budu tačno nula, čime model automatski vrši selekciju značajnih promenljivih. Formula Lasso regresije je:

$$\hat{\beta} = \arg\min_{\beta} \left(\sum_{i=1}^{m} (y_i - X_i \beta)^2 + \lambda \sum_{j=1}^{n} |\beta_j| \right) \quad (3)$$

Lasso regresija poboljšava stabilnost modela i automatski eliminiše manje značajne ulazne promenljive tako što njihove koeficijente postavlja na nulu. Ovakav pristup je posebno koristan kada postoji veliki broj ulaznih podataka, jer pomaže u izdvajanju najvažnijih parametara. Model pri tome zadržava jednostavnost i lakoću tumačenja.

Random Forest Regressor: Random Forest koristi „ensemble“ metod, gde se koristi niz odlučujućih stabala, a konačno predviđanje je srednja vrednost svih predviđanja tih stabala. Nema jedinstvene matematičke formule kao kod linearnih metoda, ali osnovna ideja je:

$$y_{\text{pred}} = \frac{1}{T}\sum_{t=1}^{T} f_t(X) \qquad (4)$$

y_{pred} je predviđeni izlazi protok

$f_t(X)$ je predviđanje svakog od stabla t

T je broj stabala u šumi.

Random Forest koristi više stabala koja glasaju za rezultat, što omogućava tačna predviđanja i dobro radi sa nelinearnim podacima. Pogodan je za složene zadatke poput promene pritiska i protoka, ali je teže objasniti kako donosi odluke, pa se često naziva „crna kutija“.

3.2. Ocena preciznosti modela za predviđanje protoka

Za ocenu performansi svakog modela, koriste se nekoliko metoda i načina, kao što je R^2 (koeficijent determinancije) i RMSE (koren srednje kvadratne greške). R^2 Score meri koliko varijanse u podacima može biti objašnjeno modelom. Vrednosti blizu 1 znači da model dobro odgovara podacima.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (5)$$

y_i su stvarne vrednosti, $\hat{y}_i$ su predviđene vrednosti, $\bar{y}$ je prosek stvarnih vrednosti.

RMSE meri prosečnu veličinu greške između predviđenih i stvarnih vrednosti. Niže vrednosti znače bolju preciznost modela.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (6)$$

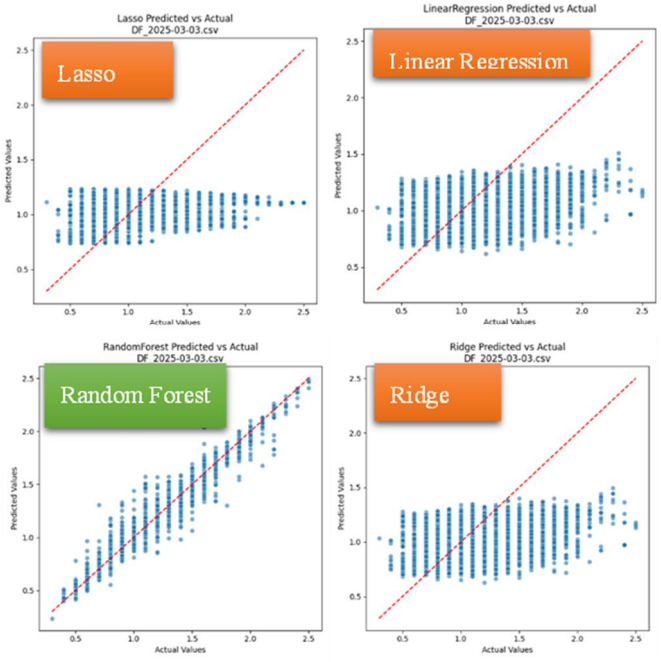
Za svaki model, rezultati su prikazani kroz dijagrame poređenjem stvarnih i predviđenih vrednosti, gde se na *X-osi* prikazuje stvarni protok, a na *Y-osi* predviđanje modela. Idealan model bi imao tačke koje leže duž dijagonale. Ovi

dijagrami pomažu da se brzo uoče sistematske greške u predviđanju, kao i potencijalna odstupanja.

4. Vizualizacija strukture podataka

Radi bolje procene tačnosti modela i lakšeg tumačenja rezultata, u ovom radu su u vidu dijagrama prikazani ulazni parametri. Rezultati svakog modela prikazani su putem dijagrama disperzije (scatter plot), gde osa X predstavlja stvarne vrednosti izlaznog protoka, a osa Y vrednosti koje je model predvideo.

Dijagonalna crvena linija predstavlja idealnu korelaciju između stvarnih i predviđenih vrednosti, što omogućava lako vizuelno poređenje tačnosti modela – tačke koje se nalaze bliže toj liniji ukazuju na veću preciznost, slika 1.



Slika 1. Dijagram disperzije tačnosti modela u predviđanju (podaci 03.03.2025)

5. Implementacija najboljeg modela i analiza rezultata

5.1. Poređenje Random Forest modela sa linearnim modelima (Linear Regression, Ridge, Lasso)

Random Forest je moćan model mašinskog učenja koji koristi više stabala odlučivanja za modelovanje i efikasno prepoznaje nelinearne odnose među

promenljivima. Za razliku od linearnih modela (kao što su Linear Regression, Ridge, i Lasso), koji pretpostavljaju linearnu zavisnost, Random Forest automatski otkriva složene interakcije između ulaznih parametara. Korišćenjem tehnike „bagging“, koja poboljšava preciznost smanjenjem varijanse i pomaže u sprečavanju prenaučenosti, model je otporan na overfitting i otporniji na promene (robustniji) prema ekstremnim vrednostima. Ne koristi ceo skup podataka za svako stablo, što smanjuje osetljivost na outliere. Za razliku od Ridge i Lasso modela koji zahtevaju regularizaciju, Random Forest prirodno selektuje značajne varijable tokom treniranja, čime omogućava stabilne performanse i sa velikim brojem ulaznih podataka. Takođe, ne zahteva unapred definisane funkcionalne oblike, jer svako stablo samostalno bira najinformativnije podele, što čini model fleksibilnim i prilagodljivim.

5.2. Princip rada modela

Svi ovi modeli pokušavaju da predviđaju neku vrednost protoka, ali sa različitim metodama i pretpostavkama. U tom konkretnom slučaju, da bismo tvrdili da je Random Forest bolji od drugih modela, moramo uzeti u obzir nekoliko faktora koji se odnose na podatke i problem koji se rešava, na sledeći način:

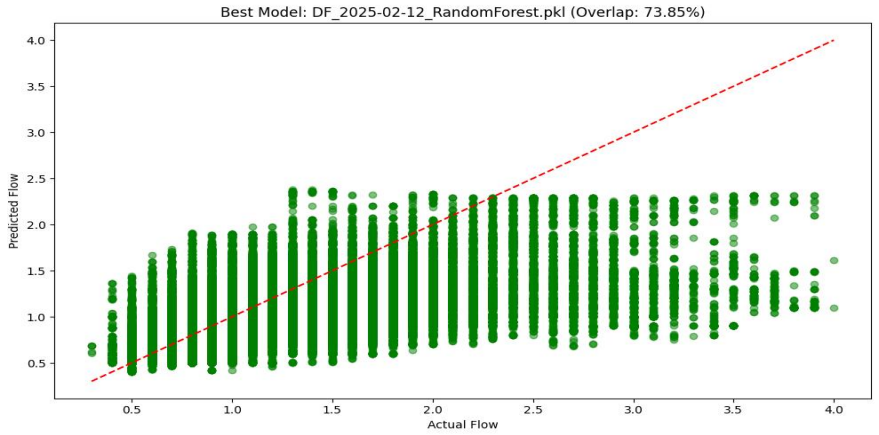
- Linear Regression pretpostavlja da postoji linearna veza između ulaznih i izlaznih varijabli (promenljivi).
- Ridge i Lasso su modifikacije linearnog modela koje uvode regularizaciju (smanjenje složenosti modela) kako bi se smanjio uticaj prekomerne složenosti modela (overfitting).
- Random Forest koristi skup stabala odluke da bi napravio predikcije, s tim da svako stablo donosi svoju predikciju i na kraju se koristi prosečna vrednost svih predikcija kao konačna procena.

Algoritam automatski preuzima podatke iz SCADA sistema, konvertuje vreme u sekunde radi lakše obrade i izdvaja relevantne ulazne i ciljne varijable. Podaci se zatim dele na skupove za obuku i testiranje, nakon čega se treniraju sva četiri modela. Svaki model se ocenjuje pomoću R^2 i RMSE. Nakon ocene, svi modeli se čuvaju, a najbolji – sa najvećim R^2 i najmanjim RMSE – se posebno izdvaja i čuva za buduću upotrebu. Algoritam omogućava brzu i efikasnu procenu više modela, uz automatski izbor i čuvanje najpreciznijeg i najboljeg.

5.3. Korišćenje Random Forest modela

U okviru rada razvijen je algoritam od strane autora rada, koji omogućava ocenu više treniranih modela mašinskog učenja (u ovom slučaju oko 196 Random Forest modela sa Lasso regularizacijom) na podacima za željeni datum (28.04.2025). Cilj algoritma je da proceni tačnost svakog modela na osnovu podataka prikupljenih za određeni dan, kako bi se identifikovao model koji najbolje predviđa stvarne vrednosti protoka vode u sistemu. Učitavanje realnih podataka iz fajla koji sadrži

merenja za izabrani datum. Posebno se izdvajaju kolone koje predstavljaju vreme, ulazni pritisak i stvarni protok. Formiranje ulaznih podataka (X) za modele koristi vreme i ulazni pritisak, dok je ciljna promenljiva (Y) stvarni protok. Učitavanjem svih modela iz zadatog direktorijuma – algoritam pretražuje i automatski učitava sve modele koji su sačuvani kao „RandomForest.pkl“. Izračunava se prosečno poklapanja („overlap“) između predviđanih i stvarnih vrednosti, pri čemu se koristi procentualna greška. Poklapanja se potom svrstavaju u kategorije tačnosti od 0–25%, 25.1–50%, ..., do 90.1–100%. Ovaj pristup omogućava brzu i efikasnu ocenu performansi ovog modela na realnim podacima i može poslužiti kao osnova za automatski izbor najpouzdanijeg modela za dalju operativnu primenu. Korišćenjem modela Random Forest, najveći broj modela (161 od ukupno 196, tj. 82.14%) ostvario je tačnost poklapanja u opsegu od 70.1% do 80%, što ukazuje na stabilne i konzistentne performanse većine Random Forest modela. Vrlo mali broj modela postigao je izuzetno visoku tačnost (90.1–100%), dok su samo 11 modela imali slabije performanse sa poklapanjem manjim od 70%. Značajno je i to što nijedan model nije imao tačnost ispod 25%, što pokazuje da ne postoje modeli sa izrazito lošim performansama u ovom skupu, slika 2.

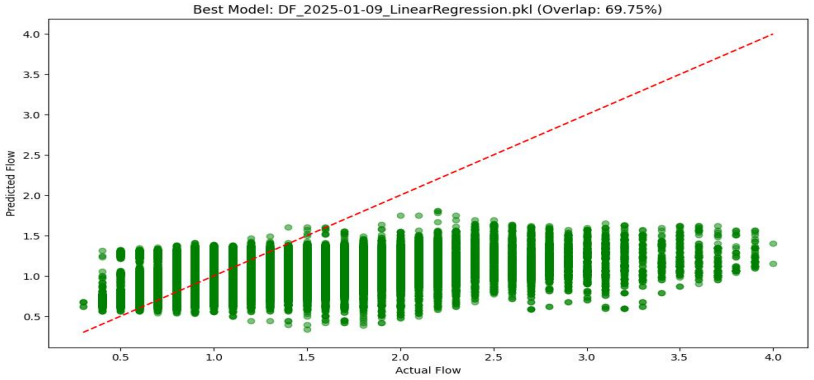


Slika 2. Prikaz disperzije podataka i preklapanje podataka na dan 12.02.2025. sa 73.85 % tačnosti

5.4. Korišćenje Linear Regression modela

Ocenom ukupno oko 197 modela treniranih pomoću Linearne regresije (LinearRegression.pkl) dobijeni su rezultati koji pokazuju da je apsolutna većina modela (95.43%) ostvarila prosečan stepen poklapanja sa stvarnim vrednostima u opsegu 60.1–70%, slika 3. Ovakva raspodela ukazuje na doslednost Linearne regresije u predviđanju, ali sa značajno nižim prosečnim učinkom u poređenju sa Random Forest modelima, koji su u velikom broju slučajeva dostigli viši stepen

poklapanja (70–80%). Nema modela sa tačnošću iznad 70%, što ukazuje na ograničenu sposobnost Linearne regresije da u potpunosti uhvati nelinearnosti u podacima. Takođe, pozitivno je što nema ni izrazito loših modela sa tačnošću ispod 25%.



Slika 3. Prikaz disperzije podataka i preklapanje podataka na dan 09.01.2025. sa 69.75 % tačnosti

5.5. Korišćenje Ridge Regression modela

Raspodela tačnosti za 197 Ridge Regression modela, izražena kroz procente poklapanja između stvarnih i predviđenih vrednosti, dovodi do toga da Ridge regresija pokazuje identične rezultate kao Linearna regresija na ovom skupu podataka. Gotovo svi modeli (95.43%) postižu tačnost u rasponu od 60.1–70%, što ukazuje da dodavanje L2 regularizacije nije značajno unapredilo performanse u odnosu na običnu linearnu regresiju. Takođe, kao i kod Linearne regresije, nema modela sa visokom tačnošću (preko 70%), što potvrđuje da modeli ovog tipa imaju ograničenja pri modeliranju kompleksnijih nelinearnih obrazaca u podacima. Ridge regresija, uprkos regularizaciji, ne donosi prednost u tačnosti u poređenju sa običnom Linear Regression. Ridge modeli dele ista ograničenja kao Linearni model, posebno kada su podaci nelinearni odnosi. Za bolje rezultate, poželjno je koristiti složenije algoritme poput Random Forest, koji su u prethodnoj analizi pokazali znatno bolje performanse.

5.6. Korišćenje Lasso modela

Uspešnost 197 Lasso regresionih modela, prema procentu poklapanja sa stvarnim vrednostima protoka, koji primenjuje L1 regularizaciju, pokazuje gotovo identične performanse kao Ridge i Linear Regression modeli, ali sa najvećim brojem modela (98.48%) koji se nalaze u opsegu poklapanja 60.1–70%. Ovo ukazuje da Lasso model dodatno „očisti“ koeficijente, što može smanjiti složenost modela, ali u ovom slučaju ne dovodi do većeg poboljšanja u predviđanju. Nema modela sa

visokom preciznošću (preko 70%), modeli nisu u stanju da uhvate kompleksnije nelinearne relacije koje su očigledno prisutne u stvarnim podacima.

Obzirom na veoma slične rezultate kod svih linearnih pristupa, može se zaključiti da je za analizu ovakvog tipa vremenskih i podataka vezanih za protok, neophodna primena nelinearnih modela poput Random Forest-a koji je u prethodnoj analizi ostvario mnogo veću tačnost i veću zastupljenost u višim kategorijama poklapanja.

5.7. Najbolji model u predviđanju protoka na realnim podacim

Zajednička tabela 1, poređenja modela po kategorijama tačnosti:

Tabela 1. Poređenje rezultata predviđanja protoka

Kategorija poklapanja	Random Forest	Linear Regression	Ridge Regression	Lasso Regression
0–25%	0	0	0	0
25.1–50%	7 (3.6%)	7 (3.5%)	7 (3.5%)	1 (0.5%)
50.1–60%	4 (2.0%)	2 (1.0%)	2 (1.0%)	2 (1.0%)
60.1–70%	22 (11.2%)	188 (95.4%)	188 (95.4%)	194 (98.4%)
70.1–80%	161 (82.1%)	0	0	0
80.1–90%	0	0	0	0
90.1–100%	2 (1.0%)	0	0	0
Ukupno modela	196	197	197	197

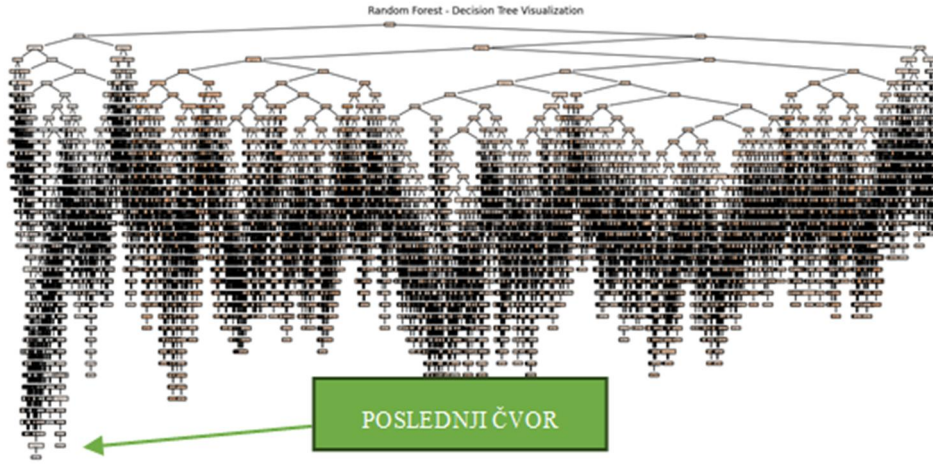
Analiza rezultata jasno pokazuje razliku između Random Forest modela i tri linearna modela (Linear, Ridge i Lasso):

- Random Forest ostvario je najviši kvalitet predviđanja, sa čak 82.14% modela u kategoriji poklapanja 70.1–80%, dok linearni modeli nisu imali nijedan model iznad 70% poklapanja.
- Linearni modeli su gotovo u potpunosti koncentrisani u kategoriji 60.1–70% (više od 95% slučajeva), što ukazuje na stabilnost, ali i ograničenu tačnost.
- Lasso Regression ima blago viši udeo u najdominantnijoj kategoriji (98.4%), ali ne prelazi prag od 70% poklapanja.
- Random Forest je jedini model koji je pokazao sposobnost da uhvati složenije nelinearne obrasce, što se ogleda u znatno višem prosečnom poklapanju i u prisustvu rezultata i u višim kategorijama (uključujući i 90.1–100%).

Random Forest se pokazao kao najprecizniji model za predviđanje protoka zahvaljujući sposobnosti da prepozna nelinearne odnose, pa će se tako koristiti za buduće analize i predviđanja tehničkih i vremenskih podataka.

5.8. Korišćenje najboljeg modela u predviđanju protoka

Nakon obrade i pripreme podataka (oko 397 modela), koristi se najbolji trenirani Random Forest Regressor za predviđanje protoka na osnovu ulaznih parametara (vreme i pritisak), primenom funkcije model.predict(X). Model koristi 100 stabala bez ograničenja dubine, sa minimalno 2 uzorka za podelu i 1 na listu, a „random state“ je 42 radi reproduktivnosti. Analiza važnosti pokazuje da prva osobina dominira (0.915), dok druga ima manji uticaj (0.084), što potvrđuje njen značaj za tačnost modela. Slika opadajućeg stabla, slika 4, prikazuje strukturu stabla sa roditeljskim čvorovima na višim nivoima, dok su podčvorovi niže, stvarajući efekat opadanja.



Slika 4. Prikaz skupa odlučujućeg stabla modela

Čvorovi se pozicioniraju tako da su roditelji iznad svojih podčvorova, a veze između njih prikazane linijama. Poslednji čvor daje konačnu vrednost koju treba doneti na osnovu prethodnih odluka.

5.8.1. Mesečna predviđanja protoka

Nakon što je izvršeno predviđanje pomoću najboljeg modela na kompletnim podacima za mesec april, izvršena je ocena tačnosti rezultata na osnovu procenta preklapanja između stvarnih i predviđenih vrednosti.

Najveći broj podataka (44.83%) se nalazi u opsegu tačnosti od 70.1% do 80%, što ukazuje da model daje relativno dobra predviđanja za većinu podataka. Trećina podataka (31.03%) se nalazi u srednjoj kategoriji tačnosti (60.1%–70%), što je još uvek prihvatljiv nivo za tehničku analizu. Zabrinjavajuće je što četvrtina (24.14%) ima vrlo nisku tačnost ispod 25%, što ukazuje na to da u tim slučajevima model nije

uspeo da uhvati obrazac u podacima – moguće zbog atipičnog ponašanja u sistemu, lošeg kvaliteta ulaznih podataka ili vanrednih uslova. Model nije postigao visoku tačnost (iznad 80%) ni za jedan datum, što pokazuje da još ima prostora za unapređenje modela ili dodatnu analizu najproblematičnijih fajlova.

Nakon primene najboljeg modela na podacima za mart, izvršena je analiza tačnosti po kategorijama procentualne tačnosti. Najveći broj (45.16%) se nalazi u tačnosti između 60.1% i 70%, što ukazuje da model u većini slučajeva daje umereno tačna predviđanja. Sledeća najzastupljenija kategorija je 70.1–80% sa 25.81% fajlova, što su dobre, ali ne odlične. Skoro trećina podataka (29.03%) ima veoma nisku tačnost (do 25%), što je zabrinjavajuće i zahteva dodatnu pažnju. Nema nijedne vrednosti u visokoj tačnosti (iznad 80%), što znači da model ne dostiže visoke performanse ni za jedan slučaj.

Na podacima za februar, izvršena je analiza tačnosti gde 53.57% podataka spada u kategoriju 60.1–70% tačnosti, što pokazuje da model u ima dobru preciznost, ali nije postigao visoke rezultate. Sledeća najzastupljenija kategorija je 70.1–80% sa 46.43%, što znači da je model ostvario solidne rezultate, ali i dalje postoji prostor za poboljšanje. Nema podataka sa izuzetno tačnim predviđanjima (iznad 80%), što je i dalje ograničenje u performansama modela za februar.

Na podacima za januar, izvršena je analiza tačnosti predviđanja po kategorijama procentualne tačnosti. U januaru je model ostvario najveći broj predviđanja u kategoriji 60.1–70% (64.52%) i solidan broj u kategoriji 70.1–80% (25.81%). Solidne rezultate sa 53.57% podataka u kategoriji 60.1–70% i nešto bolji učinak u kategoriji 70.1–80% (46.43%).

U martu i aprilu je zabeležen pad u tačnosti, sa više podataka u kategoriji 0,0 – 25,0% u martu (29.03%) i smanjenjem ukupne tačnosti u aprilu.

5.8.2. Zaključak i mišljenje o mesečnim predviđanjima

Na osnovu analize predviđanja za svaki mesec, najbolji mesec korišćenjem najboljeg modela je februar. Ovaj mesec je imao najpovoljniji rezultat, kada su u pitanju predviđanja u kategorijama 60.1–70% i 70.1–80%, sa velikim brojem podataka koji spadaju u ove kategorije tačnosti. U februaru, model je ostvario 53.57% u kategoriji 60.1–70% i 46.43% u kategoriji 70.1–80%. Ovi rezultati ukazuju na dobru preciznost modela, sa velikim brojem predviđanja, koja su bila tačna u rasponu od 60% do 80%.

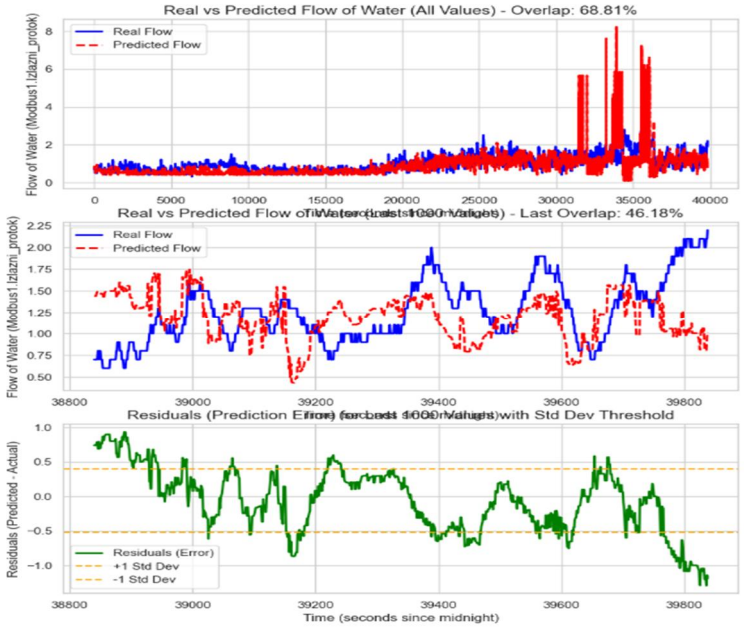
Februar se istakao najvećom tačnošću predviđanja zbog mogućih stabilnijih vremenskih uslova, manjeg broja eksternih uticaja i ujednačenijih radnih uslova vodozahvata. U poređenju sa drugim mesecima, manja varijabilnost podataka u februaru smanjila je šum i omogućila precizniju analizu. Sa 53.57% predviđanja u

kategoriji 60.1–70% i 46.43% u 70.1–80%, februar je postigao najbolji balans između preciznosti i tačnosti, čineći ga najpouzdanijim za primenu modela.

6. Primena u realnom vremenu

Predstavljeni sistem je razvijen kao jednostavna „web“ aplikacija u Python okruženju (Flask), koja koristi podatke generisanih od strane SCADA sistema. Na osnovu prethodno treniranog modela mašinskog učenja, najbolji model, aplikacija vrši predviđanje očekivanih vrednosti izlaznog protoka u odnosu na vreme i trenutni pritisak, zatim ih upoređuje sa stvarnim izmerenim vrednostima, i vizuelno prikazuje razliku na grafikonima. Rezultati ove analize pomažu korisniku da lako prepozna odstupanja od uobičajenog ponašanja sistema i na osnovu toga identifikuju moguće gubitke. Na osnovu visine razlike (reziduala), mogu se identifikovati sumnjive tačke koje ukazuju na curenje ili neregularnu potrošnju.

Reziduali za poslednjih 1000 uzoraka prikazuju se grafički uz srednju vrednost i granice jedne standardne devijacije, slika 5, čime se vizuelno izdvajaju odstupanja od očekivanog, od -0,5 do +0,5. Ova metoda omogućava procenu tačnosti modela i ranu detekciju potencijalnih kvarova, curenja ili grešaka u sistemu, što može doprineti smanjenju gubitaka vode.



Slika 5. Prikaz kretanja realnog protoka i predviđenog sa prikazom reziduala standardne devijacije na dan 30.04.2025.

7. Moguća rešenja i primena modela

Ovaj sistem je praktično rešenje za male i srednje vodovode bez razvijenih analitičkih platformi, jer omogućava jednostavno praćenje mreže uz pomoć SCADA podataka i veštačke inteligencije, odnosno mašinskog učenja.

Daljim razvojem, poput uvođenja vremenskih faktora ili povezivanja sa GIS-om, može se dodatno unaprediti preciznost i efikasnost u smanjenju gubitaka vode. Rezultati rada pokazuju da modeli mašinskog učenja mogu uspešno predviđati protok u sistemima za nadzor čak i do preko 80%, uz preduslov adekvatnog predprocesiranja podataka.

Nelinearni model kao što je Random Forest pruža bolje rezultate na kompleksnim skupovima, dok analiza stvarnih i predviđenih vrednosti protoka može ukazivati na nekoliko ključnih problema u vodovodnim sistemima, kao što su:

- Gubici vode: Ako stvarni protok značajno odstupa od predviđenih vrednosti, to može sugerisati da je došlo do defekta u mreži, ili da postoji neki drugi oblik gubitka na samoj mreži.
- Kvarovi i blokade: Ukoliko su razlike između stvarnih i predviđenih vrednosti velike, to može ukazivati na tehničke probleme u sistemu, poput začepljenja, neispravnih pumpi ili drugih nepravilnosti u radu opreme.

Primena mašinskog učenja u vodovodnim sistemima može doneti brojne prednosti, među kojima se izdvajaju:

- Prepoznavanje anomalija: Korišćenjem analize predviđenih i stvarnih vrednosti moguće je brzo identifikovati neobične promene u protoku, što može ukazivati na prisutnost kvarova ili gubitaka.
- Optimizacija rada sistema: Pravovremeno predviđanje protoka omogućava bolju optimizaciju rada pumpi i ostale infrastrukture u mreži, čime se smanjuje potrošnja energije i povećava efikasnost celokupnog sistema.
- Poboljšanje održavanja: Kontinuirano praćenje i identifikacija problema u realnom vremenu omogućava preventivno održavanje i smanjuje potrebu za hitnim popravkama, čime se smanjuju troškovi i vreme zastoja u mreži.

Predloženi sistem za predviđanje protoka vode pomoću modela mašinskog učenja može značajno doprineti unapređenju efikasnosti u upravljanju vodovodnim sistemima. Korišćenjem ovih modela za analizu podataka u realnom vremenu, vodovodna preduzeća mogu brzo detektovati odstupanja od očekivanih vrednosti i pravovremeno reagovati na potencijalne gubitke ili kvarove.

8. Literatura

[1] Breiman, L. *Random Forests*, Machine Learning, 45(1), 5-32, 2001.

[2] James G, Witten D, Hastie T. & Tibshirani R, *An Introduction to Statistical Learning*, Springer, 2013.

[3] Zhang Y. & Wang H, Water Distribution Network Monitoring and Fault Detection Using Machine Learning Techniques, *Journal of Water Resources Planning and Management*, 146(4), 04020021, 2020.