VYUŽITIE H2O ML A FEATURE ENGINEERINGU PRE PREDIKCIU TEPLOTY VODY V RIEKE LITAVA

ZBYNEK BAJTEK

Institute of Hydrology, SAS, Dúbravská cesta č. 9, 841 04 Bratislava, Slovakia

Water temperature is often perceived as a primary indicator of water quality, hence the need for accurate prediction tools for streams without direct measurements that can accurately predict its evolution based on other readily available meteorological characteristics, for example from climate stations. In recent years, together with the development of computational technology, there has been a rapid development of various mathematical tools and models that allow us to predict stream water temperature with a certain degree of accuracy. Nowadays, we can say that MachineLearning (ML) methods, among others, are experiencing a great growth. However, the methods themselves are only as good as the inputs to these models. Another essential factor in their application is so called Feature Engineering. This element does not only involve the actual analysis of the input data and the cleaning of missing features but also the creation of new variables that would provide the model with an additional feature that would better describe the data and the relationship between them. Therefore, in addition to the actual application of ML in this case H2O ML, the paper focuses on the impact of selected variables on the accuracy of the models.

Teplota vody je často vnímaná ako primárny indikátor kvality vody, z čoho vyplýva potreba presných predpovedných nástrojov pre toky bez priamych meraní, ktoré by dokázali presne predpovedať jej vývoj na základe iných ľahko dostupných meteorologických charakteristík - napríklad z klimatických staníc. V posledných rokoch spolu z rozvojom výpočtovej techniky nastal aj rýchly nástup rozvoj rôznych matematických nástrojov a modelov, ktoré nám umožňujú s určitou mierou presnosti predpovedať teplotu vody v tokoch. V súčasnosti môžeme konštatovať, že zaznamenávajú veľký nárast medzi inými aj metódy Machine Learning - strojového učenia (SU). Samotné metódy sú však len natoľko dobré, ako sú dobré vstupy do týchto modelov. Ďalší podstatný faktor v ich aplikácii predstavuje tzv. Feature Engeniering, čo by sme mohli preložiť ako Inžinierstvo Parametrov (IP). Tento prvok nezahŕňa len samotnú analýzu vstupných údajov a očistenie od chýbajúcich prvkov, ale tiež tvorbu nových premenných, ktoré by modelu poskytli ďalší prvok, ktorý by lepšie popísal dané údaje a vzťah medzi nimi. Preto sa príspevok okrem samotnej aplikácie SU, v tomto prípade H2O ML, zameriava na vplyv vybraných premenných, ale aj na presnosť modelov.

Key words: Machine Learning, Feature Engineering, river water temperature, prediction

ÚVOD

Rieky sú hybnou silou našej planéty ako zdroj pitnej vody, alebo ako zdroj vody pre závlahu a priemysel. V dôsledku klimatickej zmeny spôsobenou ľudskou činnosťou sa však teplota vody v riekach mení, čo má významný vplyv na ich ekosystémy a celú ľudskú spoločnosť. Predpovedanie týchto teplotných zmien je kľúčové pre správne riadenie vodných zdrojov a ochranu biodiverzity. V tomto článku sa zaoberáme metódami a technológiami používanými na predpovedanie teploty vody v povrchových tokoch. Vývoj nástrojov na predpovedanie teploty vody v riekach prešiel výraznými zmenami. Spočiatku sa na predpovedanie teploty vody v rieke ako regresor používala najmä teplota vzduchu (Mohseni and Stefan, 1998; Zhu et al., 2019). Neskôr sa začali používať širšie súbory vstupných premenných, ako aj niektoré modifikácie štandardného lineárneho regresného modelu (Morrill et al., 2005) a nelineárne regresné modely (Bajtek et al., 2022; Mohseni and Stefan, 1998). Rovnako sa uplatňujú aj SARIMA modely (Seasonal Autoregressive Integrated Moving Average) na predpovedanie teploty vody v riekach (Pekárová et al., 2023). Ďalším modelom je napríklad air2stream hybridný model na predpovedanie teploty riečnej vody, ktorý kombinuje fyzikálne založenú štruktúru so stochastickou kalibráciou parametrov (Toffolon and Piccolroaz, 2015).

V posledných rokoch sa s rozvojom výpočtového výkonu a umelej inteligencie postupne začali používať nové modely na predpovedanie teploty riečnej vody. Sem môžeme zaradiť rôzne metódy strojového učenia, ako napríklad kroková lineárna regresia, random forest(RF), eXtreme Gradient Boosting (XGBoost), dopredné neurónové siete (FNN) a dva typy rekurentných neurónových sietí (RNN). V štúdii prezentovanej v práci (Feigl et al., 2021) sa hodnotí výkonnosť šiestich rôznych modelov strojového učenia na predpovedanie teploty vody v riekach. Autori použili viacero súborov vstupných údajov na rôznych veľkostiach povodia. Práca je originálna v tom, že sa použili rôzne klimatické premenné a ich kombinácie ako vstupy do modelov. Bayesovská optimalizácia bola tiež použitá na objektívny odhad hyperparametrov modelov. Výsledná výkonnosť všetkých modelov bola porovnaná s dvoma referenčnými modelmi, aby sa dosiahli porovnateľné výsledky. Autori v práci dospeli k záveru, že testované modely by mohli výrazne zlepšiť predpovedanie teploty vody v porovnaní s lineárnou regresiou a modelom air2stream. Rajesh and Rehana, (2021) skúmali použitie modelov strojového učenia v spojení s analýzou globálnej citlivosti na predpovedanie teploty vody v rieke. V štúdii bola testovaná výkonnosť ridge regression (RR) , K-nearest neighbors (KNN), random forest (RF), and support vector regression (SVR), spolu Sobol' global sensitivity analysis (GSA). V štúdii sa dospelo

k záveru, že SVR je najrobustnejší ML model na predpovedanie RWT v mesačnej časovej škále. V článku navrhujú nový prístup na predpovedanie indexu a triedy kvality vody pomocou algoritmov strojového učenia. Navrhovaná metóda je založená na štyroch parametroch vody: teplota, pH, zákal a koliformných baktériách. Kombinuje strojové učenie (ML) a analýzu citlivosti (DA) s cieľom zlepšiť predpovede na základe meraných údajov. Navrhnutý algoritmus aplikovali na merané údaje o teplote vody v rieke. Hlavným cieľom štúdie bolo identifikovať najviac ovplyvňujúcu premennú pomocou algoritmu GSA a následne aplikovať rôzne ML modely na predpovedanie riečneho toku. Podobne aj Souaissi et al., (2023) porovnáva dva modely strojového učenia - random forest (RF) a extreme gradient boosting (XGBoost) s neparametrickými viacrozmernými adaptívnymi regresnými spline (MARS) a semi parametrickými zovšeobecnenými aditívnymi modelmi (GAM), pričom sa zamerali na predpovedanie maximálnych teplôt riek. Predpovede sa robia na nemeraných lokalitách, čo znamená, že použili len malé množstvo údajov alebo žiadne. Prezentované výsledky ukazujú najväčšiu presnosť u modelov GAM a MARS v kombinácii NLCCA+GAM, nasledovanou CCA+MARS. V kontexte premenných aplikovaných v modeloch predstavuje práca (Wade et al., 2023) zaujímavý pohľad, nakoľko sa autori zaoberali teplotnými charakteristikami tokov na území USA, pričom skúmali vplyv rôznych faktorov na tieto teplotné režimy. V tomto prípade použili verejne dostupné záznamy o teplote vody v tokoch z 410 lokalít, pričom sa zamerali na vplyv klímy, charakteristík povodia, hydrologických parametrov a antropogénnych vplyvov. Cieľom bolo zistiť, ako sa vplyv týchto faktorov mení v rôznych ročných obdobiach a priestorových oblastiach. V rámci štúdie sa zameriavali na dve základné metriky - maximálnu teplotu vody a teplotnú citlivosť - a použili prístup SU na odvodenie týchto parametrov teploty vody. Čo sa týka samotného H2O AutoML, Madni et al., (2023) aplikovali túto metódu na predpovedanie kvality pitnej vody. Experimenty vykonali s použitím rôznych scenárov s cieľom vytvoriť efektívny a presný automatizovaný systém na klasifikáciu a predpoveď kvality vody. Ďalším príkladom použitia H2O ML je napríklad práca Považanová et al., (2023), kde bol tento nástroj použitý na predikciu evapotranspirácie. Klasifikácia a predpovedanie kvality vody je dôležitým problémom, ktorý sa pokúša riešiť pomocou modelovania a strojového učenia. Existujú však výzvy spojené s nízkou presnosťou a chýbajúcimi hodnotami v údajoch.

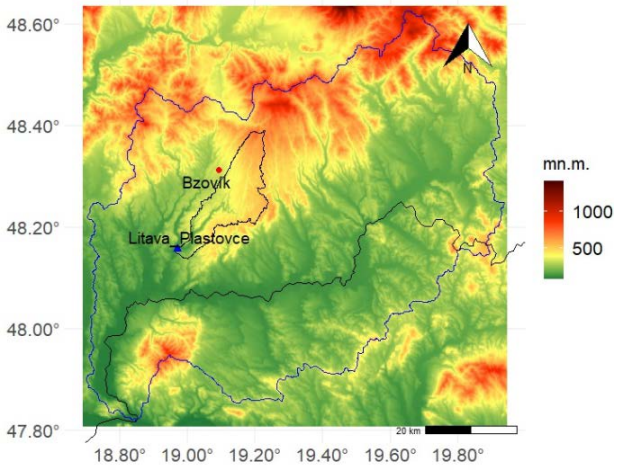
Tento článok sa zameriava na hľadanie parametrov v rámci IP, pričom vychádza z meraných veličín, ktoré boli doplnené o ďalšie syntetizované veličiny a následne bol testovaný ich vplyv na presnosť modelu a predpovedí. Pridanie týchto nových premenných spresnilo výsledky modelu a tiež možnosť aplikácie takýchto modelov na toky, kde nie sú priame merania teploty vody.

POPIS LOKALITY A POUŽITÝCH DÁT

V našom prípade sme pracovali s dátami (zdroj SHMÚ) z klimatologickej stanice Id. 11902 – Bzovik, z ktorej bola použitá teplota vzduchu (*Ta)* v dennom kroku (roky 2006 – 2020) a z vodomernej stanice na rieke Litava Id. 7600 – Litava-Plášťovce. Jednalo sa o denné hodnoty prietokov vody *Qd* a teploty vody *To* za roky 2006 – 2020. Rieka Litava je významný ľavostranný prítok Krupinice. Pramení v Krupinskej planine pod vrchom Javorok (695,0 m n. m.) v nadmorskej výške okolo 650 m n. m. Tečie najprv na juh cez obec Senohrad, zľava priberá Litavicu (451,5 m n. m.), ďalej preteká obcami Lackov a Litava, kde priberá rieku Malá Litava z ľavej strany. Má celkovú dĺžku 48 km.

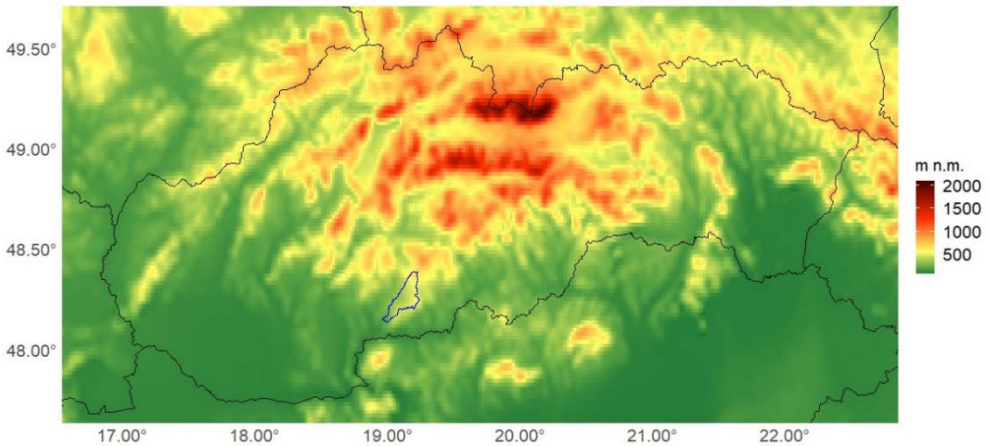
Obrázok 2. Povodie Litavy v rámci povodia Ipľa.

Figure 2. The Litava river basin within the Ipeľ river basin.



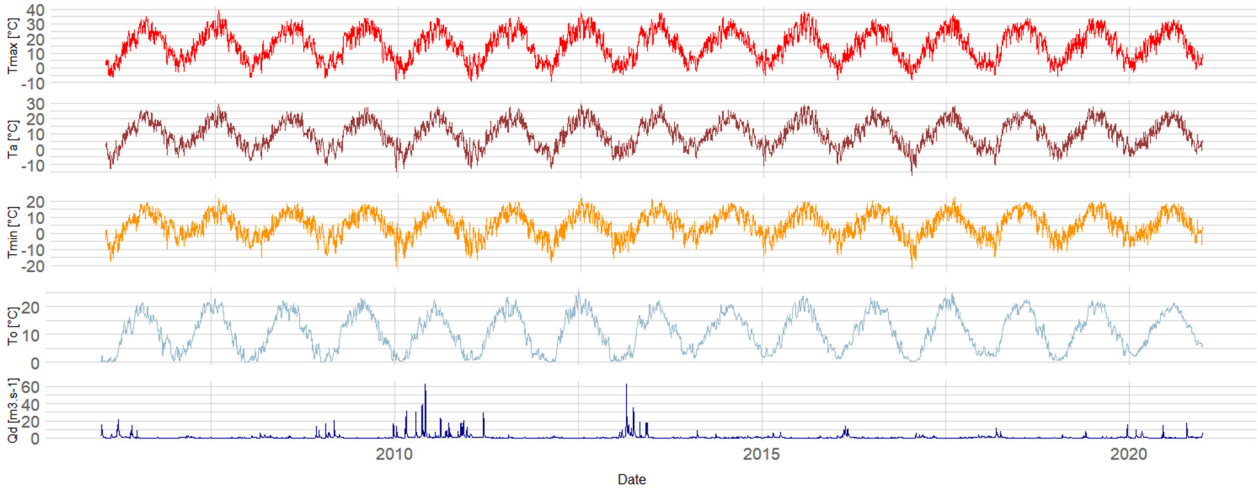
Obrázok 1. Povodie Litavy.

Figure 1. Location of the Litava river basin.



Obrázok 3. Priebeh teplôt Ta, Tmax, Tmin zo stanice Bzovík a teploty vody To a Qd denných prietokov zo stanice Litava-Plášťovce.

Figure 3. Temperatures Ta, Tmax, Tmin from station Bzovik and water temperature To and Qd of daily flows from station Litava-Plášťovce.



POPIS H2O ML MODELU

H2O AutoML (LeDell and Poirier, 2020) je open source platforma pre strojové učenie a prediktívnu analýzu, ktorá umožňuje vytvárať modely SU na veľkých objemoch dát a poskytuje jednoduchú aplikáciu týchto modelov v rôznych programovacích prostrediach. Jadro kódu H2O je napísané v jazyku Java. Vnútri H2O sa na prístup k údajom, modelom, objektom atď. a odkazovanie na ne vo všetkých uzloch a strojoch používa distribuované úložisko kľúčov/hodnôt. Algoritmy sú implementované nad distribuovaným rámcom H2O a využívajú rámec Java Fork/Join na viacvláknové spracovanie. Údaje sa čítajú paralelne, sú distribuované v rámci klastra a ukladajú sa do pamäte v stĺpcovom formáte komprimovaným spôsobom. Parser údajov H2O má zabudovanú inteligenciu na odhadnutie schémy prichádzajúceho súboru údajov a podporuje prijímanie údajov z viacerých zdrojov v rôznych formátoch (CSV, Excel, XML).

Rozhranie REST API H2O umožňuje prístup ku všetkým možnostiam H2O z externého programu alebo skriptu prostredníctvom JSON cez HTTP. Toto rozhranie využíva webové rozhranie H2O (Flow UI) s prepojením na R (R Core Team, 2022) (H2O-R) alebo Python (H2O-Python).

FEATURE ENGINEERING

Feature Engineering alebo tiež inžinierstvo parametrov IP je proces vytvárania alebo transformácie premenných v ML s cieľom zlepšiť jeho výkon. Tento proces zahŕňa výber relevantných informácií zo surových údajov a ich transformáciu do formátu, ktorý je pre model ľahko zrozumiteľný. Cieľom je poskytnúť modelu zmysluplnejšie a relevantnejšie informácie, čo vedie k zvýšeniu jeho presnosti. Kvalita funkcií použitých na trénovanie modelov ML je kľúčová pre ich úspech. Feature Engineering poskytuje techniky na vytváranie nových premenných z existujúcich údajov. Tieto techniky pomáhajú identifikovať dôležité vzory a vzťahy v údajoch, čo zlepšuje efektivitu učenia modelu zo vstupných údajov. Feature sú individuálne merateľné vlastnosti údajov, ktoré sa používajú ako vstupy do algoritmov ML. Môžu byť číselné, kategorické alebo textové a predstavujú rôzne aspekty údajov relevantné pre daný problém.

V príspevku boli použité z meraných premenných: teplota vody v rieke *To* ako závislá premenná a teplota vzduchu *Ta* ako prediktor, okrem tej boli medzi prediktory zahrnuté parametre mesiac a týždeň roka, v ktorom bolo konkrétne meranie uskutočnené, pričom tieto boli transformované na „Faktor“. Faktory v R sú dátové štruktúry, ktoré sa používajú na reprezentáciu a kategorizáciu kategorických dát. Kategorické dáta sú premenné, ktoré nadobúdajú obmedzený počet rôznych hodnôt. Faktory môžu ukladať celé čísla a reťazce a majú atribút „levels“, ktorý obsahuje všetky možné hodnoty faktora. Ďalej boli použité dve premenné generované z *Ta*. Konkrétne *Tan,* kde parameter n predstavuje n-denný pohyblivý priemer a druhou premenou je predikovaná *Taexp* pomocou metódy exponenciálneho vyhladzovania časových radov, ktorá sa používa na predpovedanie budúcich hodnôt na základe histórie meraných údajov. Táto metóda priraďuje väčšiu váhu novším pozorovaniam a menšiu váhu starším pozorovaniam, pričom váha exponenciálne klesá s vekom pozorovania. Konkrétne, príkazom ets() v prostredí R vytvorí nový stĺpec *Taexp,* ktorý obsahuje prispôsobené hodnoty z modelu exponenciálneho vyhladzovania (ETS) aplikovaného na stĺpec *Ta*. Tieto prispôsobené hodnoty sú vlastne jednokrokové predpovede, ktoré sú založené na histórii meraných údajov v stĺpci *Ta*. V druhom kroku sme pridali medzi premenné ešte parameter *Qd* denné prietoky, ale tak isto sme ju transformovali konkrétne na *Qdn,* kde parameter n predstavuje n-denný pohyblivý priemer.

V tomto článku sme použili H2O ML aplikované v prostredí programovacieho jazyka R v prostredí Rstudia (RStudio Team, 2020). Rstudio je platforma, ktorá poskytuje rôzne open source balíky a nástroje, ktoré neslúžia len na samotné modelovanie, ale a aj na následnú štatistickú analýzu a grafickú prezentáciu výsledkov.

V prvom kroku je potrebná kontrola kvality údajov a neprerušenosti časových radov (chýbajúce údaje), následne boli vygenerované nové premenné. Ďalej bola potrebná transformácia dát z tabuľkovej formy na formu z H2O prostredia. Následne je potrebné si rozdeliť dáta na trénovacie a testovacie sady, v tomto prípade sú delené v pomere 80/20. Modelu sme zadefinovali závislú premennú a prediktory, teda premenné, na základe ktorých ju budeme modelovať. Následne bol spustený model, kde prediktormi boli: mesiac, týždeň (v roku), *Ta, Tan* a *Taexp* a v modeli bola aplikovaná 5 a 10 stupňová krížová validácia.

Samotná krížová validácia spočíva v rozdelení trénovacieho súboru údajov na podmnožiny, v našom prípade na 5, resp. 10, kde jedna podmnožina slúži ako testovacia množina. Ostatné podmnožiny slúžia ako tréningové množiny. Klasifikátor trénuje model na trénovacej množine a testovaciu množinu používa na testovanie presnosti a výkonnosti modelu. Tento proces sa opakuje niekoľkokrát, pričom tréningovú a testovaciu množinu tvorí vždy iná podmnožina. V prípade premennej *Tan* boli testované viaceré nastavenia kĺzavého priemeru a to 2, 3, 5, 7, 10, 12 dní. Po vyhodnotení bol v ďalšom kroku z prediktorov vyradený mesiac a nahradil ho *Qdn n*-denný kĺzavý priemerný prietok a použitá bola už len 10 stupňová krížová validácia.

VÝSLEDKY

Sofvérový balík H2O ML poskytuje možnosť nastavenie výstupu pre viaceré parametre výkonu daných modelov. V tejto práci boli porovnávané nasledovné parametre:

R2 (Koeficient determinácie) - Hodnota R2 predstavuje mieru, v akej sa predpovedaná hodnota a skutočná hodnota pohybujú súčasne. Hodnota R2 sa pohybuje medzi 0 a 1, kde 0 predstavuje žiadnu koreláciu medzi predpovedanou a skutočnou hodnotou a 1 predstavuje úplnú koreláciu.

MSE (Priemerná kvadratická chyba) - Metrika MSE meria priemer štvorcov, chýb alebo odchýlok. MSE berie vzdialenosti od bodov k regresnej línii (tieto vzdialenosti sú „chyby“) a odstraňuje akékoľvek záporné znamienka pomocou druhej mocniny.. MSE zahŕňa aj variáciu a skreslenie prediktoru. MSE tiež dáva väčšiu váhu väčším rozdielom. Čím väčšia je chyba, tým viac je penalizovaná. Čím menšia je MSE, tým lepší je výkon modelu.

$$\mathrm{MSE} = \frac{1}{\mathrm{N}}\sum(\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i})^2, \quad (1)$$

kde *N* je celkový počet pozorovaní vo vašom príslušnom dátovom súbore, *yi* je meraná hodnota, *ŷi* je predpovedaná hodnota.

RMSE (Koreň z priemernej kvadratickej chyby) - Metrika RMSE hodnotí, ako dobre môže model predpovedať kontinuálnu hodnotu. Jednotky RMSE sú rovnaké ako predpovedaný cieľ. Čím menšia je RMSE, tým lepší je výkon modelu.

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{N}\sum(\mathrm{y_i}-\hat{\mathrm{y}}_\mathrm{i})^2}}, \quad (2)$$

kde *N* je celkový počet pozorovaní vo vašom príslušnom dátovom súbore *yi* je meraná hodnota, ŷi je predpovedaná hodnota.

RMSLE (Koreň z priemernej kvadratickej logaritmické chyby) je metrika, ktorá meria pomer medzi skutočnými hodnotami a predpovedanými hodnotami. Táto metrika vypočíta logaritmus predpovedaných a skutočných hodnôt a následne vypočíta kvadratickú chybu týchto logaritmických hodnôt.

$$\mathrm{RMSLE} = \sqrt{\frac{1}{\mathrm{N}\sum(\ln(\mathrm{y_i}+1)-\ln(\hat{\mathrm{y}}_\mathrm{i}+1))^2}}, \quad (3)$$

kde: *N* je celkový počet pozorovaní vo vašom príslušnom dátovom súbore, *yi* je meraná hodnota, *ŷi* je predpovedaná hodnota.

MAE (Priemerná absolútna chyba) - Priemerná absolútna chyba je priemer absolútnych chýb. Jednotky MAE sú rovnaké ako predpovedaný cieľ, čo je užitočné pre pochopenie. Čím menšia je MAE, tým lepší je výkon modelu.

$$\mathrm{MAE} = \frac{1}{\mathrm{N}\sum|\mathrm{x_i}-\mathrm{x}|}, \quad (4)$$

kde: *N* je celkový počet chýb, *xi* sú merané hodnoty, *x* sú predpovedané hodnoty a tento vzorec vypočíta priemernú absolútnu hodnotu rozdielov medzi skutočnými a predpovedanými hodnotami.

VÝSTUPY Z MODELOV

Ako už bolo spomenuté, modelovanie prebehlo v dvoch krokoch. V prvom kroku bolo použitých 5 premenných, pričom pri premennej *Tan* bol menený počet dní, z ktorých sa počítal kĺzavý priemer a to 2, 3, 5, 7, 10 a 12 dní. Ostatné premenné boli pre všetky spustenia modelu rovnaké. Ďalej bol testovaný vplyv cross validácie na výstupy z modelov. Čo sa týka parametru krížovej validácie pre všetky alternatívy *Tan,* výstup s použitím 10 stupňovej CV bol vždy lepší ako pri použití 5 stupňovej CV. Čo sa týka parametra *Tan,* teda kĺzavého priemeru denných teplôt vzduchu, najlepšie výsledky boli dosiahnuté pri 10 dňovom kĺzavom priemere. Tieto výsledky sú zosumarizované v Tab 1. V druhom kroku bola vypustená premenná „Mesiac“ a nahradila ju premenná *Qdn,* teda kĺzavý priemer denných prietokov za n dní. V tomto prípade parameter *Tan* bol ponechaný na 10 dňoch a testovaná bola zmena u premennej *Qdn,* konkrétne 5, 7, 10 a 12 dní a následne ešte 15 dní.

V tomto kroku bolo dosiahnuté ešte nepatrné zlepšenie, a to pri doplnení o 15 dňový kĺzavý priemer, pri ktorom boli dosiahnuté najlepšie hodnoty štatistík (Tab. 2). Treba poznamenať, že s výnimkou *Qdn* 15 dní bol pri ostatných spusteniach parameter *Tan* ponechaný na 10 dňoch a až v prípade *Qdn* 15 bol parameter *Tan* upravený na rovnaký počet dní. Po tomto kroku boli ešte otestované dané modely. Test spočíval vo vylúčení dát z roku 2019, kedy tieto dáta neboli ani medzi testovacími ani trénovacími dátami. V Tab. 3 je porovnanie štatistík z trénovania modelu výstupu pre roky, s ktorými model pracoval a predikcia *To* pre rok 2019 a následne predikcia pre kompletné dáta.

Tabuľka 1. Štatistické parametre pre dané konfigurácie parametrov.

Table 1. Statistical parameters for given parameter configurations.

	5 stupňová crossvalidácia Tan					10 stupňová crossvalidácia Tan					
	2 dni	3 dni	5 dní	7 dní	10 dní	2 dni	3 dni	5 dní	7 dní	10 dní	12 dní
MSE	1,244	0,939	0,675	0,474	0,427	1,240	0,951	0,650	0,512	0,411	0,462
RMSE	1,115	0,969	0,822	0,689	0,653	1,110	0,975	0,811	0,716	0,641	0,681
R2	0,973	0,979	0,985	0,989	0,991	0,973	0,979	0,985	0,989	0,991	0,990
MAE	0,848	0,741	0,626	0,527	0,491	0,851	0,741	0,616	0,543	0,481	0,514
RMSLE	0,200	0,190	0,154	0,127	0,126	0,203	0,185	0,151	0,134	0,119	0,119

Tabuľka 2. Štatistické parametre pre dané konfigurácie parametrov.

Table 2. Statistical parameters for given parameter configurations.

	10 stupňová crossvalidácia Qdn				
	5 dní	7 dní	10 dní	12 dní	15 dní
MSE	0,172	0,100	0,105	0,098	0,079
RMSE	0,415	0,317	0,325	0,313	0,281
R2	0,996	0,998	0,998	0,998	0,998
MAE	0,318	0,238	0,246	0,237	0,211
RMSLE	0,009	0,066	0,071	0,068	0,060

Tabuľka 3. Štatistické parametre pre model, testovanie vynechania údajov z roku 2019.

Table 3. Statistical parameters for the model, testing the omission of 2019 data.

	2006 – 2020	2006 – 2020 bez 2019	2019
MSE	0,069	0,096	0,434
RMSE	0,263	0,310	0,658
R2	0,998	0,998	0,988
MAE	0,201	0,233	0,511
RMSLE	0,059	0,060	0,098

Po tomto kroku bol z údajov vylúčený rok 2019, zvyšné roky boli rozdelené v pomere 80/20 % na trénovacie a testovacie dáta. Následne bol spustený a natrénovaný model, ktorý bol použitý na predikciu hodnôt *To* pre rok 2019. Vzájomné porovnanie štatistických parametrov nájdeme v Tab. 3, kde v prvom stĺpci sú štatistické parametre, kedy model pracoval s celým rozsahom dát. V druhom stĺpci sú parametre v prípade vylúčenia roku 2019 a v poslednom stĺpci sú štatistické parametre modelu. Použitý model na rok 2019 bol z dát vynechaný.

Na Obr. 4 môžeme vidieť závislosť teploty vzduchu a teploty vody v rieke. Zelenou farbou sú merané hodnoty a modrou farbou predpovedané hodnoty pre rok 2019. Ako vidíme, teploty vody predpovedané modelom sú mierne vyššie v prípade nízkych, resp. vysokých teplôt. Na Obr. 5 vidíme porovnanie meraných a predpovedaných hodnôt *To* pre rok 2019, kde *To* je meraná hodnota a Topredict1 je výstup z modelu, ktorý pracoval s celým rozsahom meraní. Topredict2 je predpoveď pre rok 2019 z modelu, v ktorom sme vylúčili dáta pri jeho trénovaní. Ako môžeme vidieť aj v Tab. 3, v prípade predikcie sa pre rok 2019 zhoršila predpoveď v štatistike MAE z hodnoty 0,201 °C, resp. 0,232 °C na hodnotu 0,511 °C.

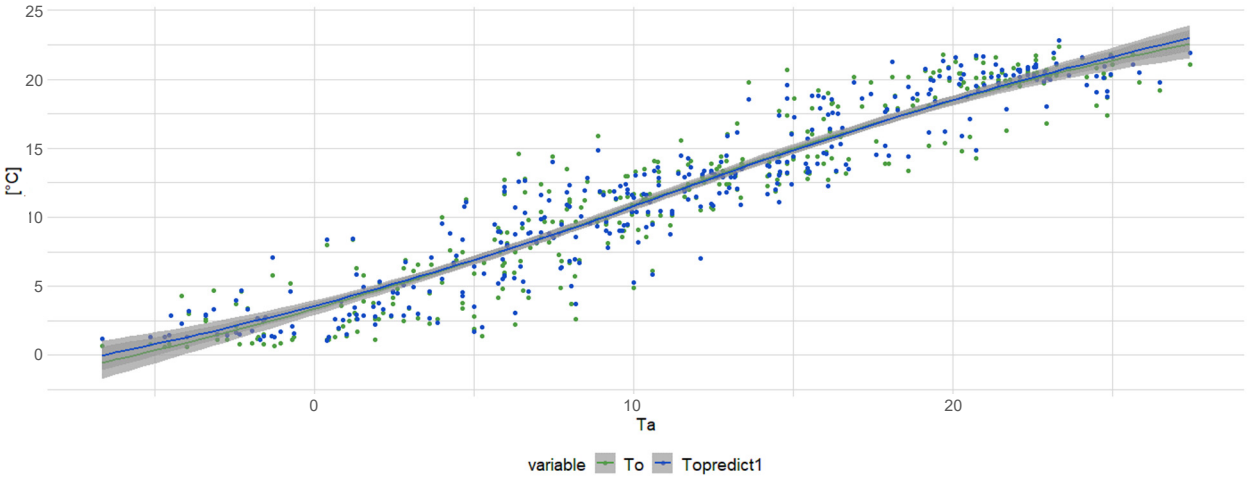
Na Obr. 6 sú zobrazené priemerné mesačné hodnoty *Ta* a *To* pre celé obdobie 2006 – 2020 spolu s predpovedanou hodnotou *Topredicted,* kde môžeme vidieť predpovedanú hodnotu *To,* ktorá kopíruje trend meranej hodnoty.

ZÁVER

V príspevku bol použitý H2O AutoML, algoritmus na automatické strojové učenie, ktorý je súčasťou platformy strojového učenia H2O. Okrem toho sa príspevok zameral na vplyv IP v spresnení modelov H2O ML, pričom boli testované ako prediktory rôzne merané aj generované premenné a tiež kategorické premenné. Ich kombinácia viedla k zlepšeniu predpovedí modelu. Významným faktom je, že v porovnaní s inými metódami venujúcimi sa predikcii

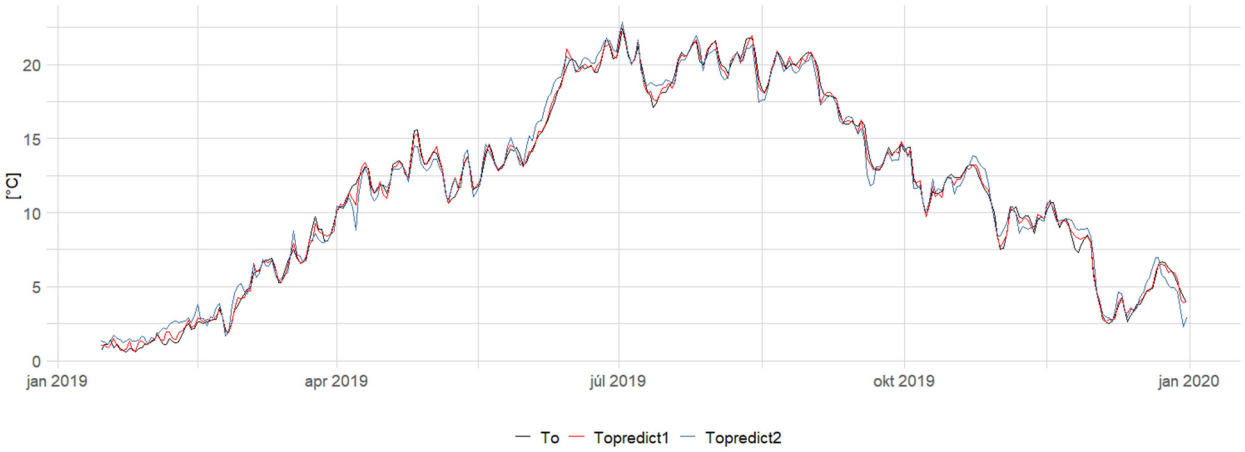
Obrázok 4. Namerané a predikované hodnoty To pre rok 2019 a teploty vzduchu Ta.

Figure 4. Observed and predicted values of To for 2019 and air temperature Ta.



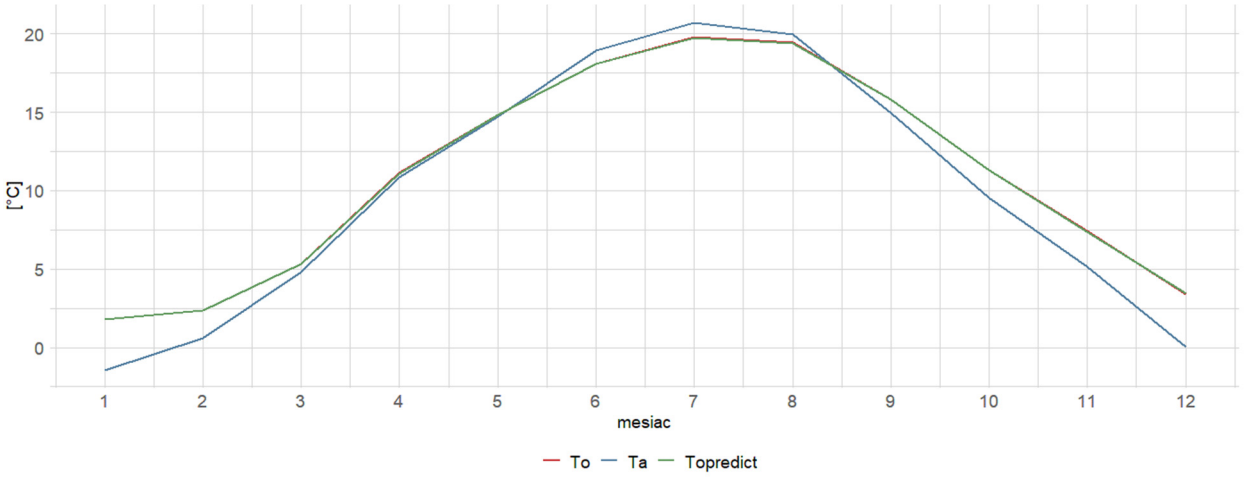
Obrázok 5. Porovnanie predpovedaných hodnôt Topredict1 a Topredict2 s meranou hodnotou To.

Figure 5. Comparison of predicted values of Topredict1 and Topredict2 with measured value of To.



Obrázok 6. Porovnanie priemerných mesačných hodnôt To a Ta s predpovedanou priemernou mesačnou hodnotou Topredicted.

Figure 6. Comparison of average monthly values of To and Ta with predicted average monthly value of To predicted.



teploty vody v tokoch, ktoré dosahujú predpovede porovnateľnej presnosti, sú tieto metódy závislé na ďalších meraných veličinách, ako napríklad zrážky alebo intenzita globálneho žiarenia, ktoré nemusia byt vždy pre danú oblasť dostupné. Je predpoklad, že by postup popísaný v tomto príspevku podával veľmi podobné výsledky, aj keby sme na predikciu teploty vody v toku použili ako jedinú meranú veličinu teplotu vzduchu. Cieľom ďalšieho výskumu bude verifikácia tohto predpokladu, ako aj aplikácia na rôznych veľkostiach a typoch tokov a tiež možnosti aplikovať tieto modely spätne, napríklad na dopĺňanie chýbajúcich údajov.

Poďakovanie

Táto práca bola podporená projektami VEGA No. 2/0015/23; APVV-20-0374 and WATSIM „Water temperature simulation during summer low flow conditions in the Danube basin“.

LITERATÚRA

Bajtek, Z.– Pekárová, P.–Jeneiová, K.–Ridzoň, J., 2022, Analysis of the water temperature in the Litava River. Acta Hydrol. Slovaca 23, 296 – 304. https://doi.org/10.31577/ahs-2022-0023.02.0034

Feigl, M.–Lebiedzinski, K.–Herrnegger, M.–Schulz, K., 2021, Machine-learning methods for stream water temperature prediction. Hydrol. Earth Syst. Sci. 25, 2951 – 2977. https://doi.org/10.5194/hess-25-2951-2021

LeDell, E.–Poirier, S., 2020, H2O AutoML: Scalable Automatic Machine Learning. 7th ICML Workshop Autom. Mach. Learn. AutoML.

Madni, H.–Umer, M.–Ishaq, A.–Abuzinadah, N.–Saidani, O., Alsubai, S.–Hamdi, M.–Ashraf, I., 2023, Water-Quality Prediction Based on H2O AutoML and Explainable AI Techniques. Water 15, 475. https://doi.org/10.3390/w15030475

Mohseni, Stefan, H.G., 1998, Stream temperature/air temperature relationship: A physical interpretation.

Morrill, J.C.–Bales, R.C.–Conklin, M.H., 2005, Estimating Stream Temperature from Air Temperature: Implications for Future Water Quality. J. Environ. Eng. 131, 139 – 146. https://doi.org/10.1061/(ASCE)0733-9372(2005)131:1(139)

Pekárová, P.–Bajtek, Z.–Pekár, J.–Výleta, R.–Bonacci, O.–Miklánek, P.–Belz, J.U.–Gorbatchova, L., 2023, Monthly stream temperatures along the Danube River: Statistical analysis and predictive modelling with incremental climate change scenarios. J. Hydrol. Hydromech. 71, 382 – 398. https://doi.org/10.2478/johh-2023-0028

Považanová, B.–Čistý, M.–Bajtek, Z., 2023, Using feature engineering and machine learning in FAO reference evapotranspiration estimation. J. Hydrol. Hydromech. 71, 425 – 438. https://doi.org/10.2478/johh-2023-0032

R Core Team, 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rajesh, M.–Rehana, S., 2021, Prediction of river water temperature using machine learning algorithms: a tropical river system of India. J. Hydroinformatics 23, 605 – 626. https://doi.org/10.2166/hydro.2021.121

RStudio Team, 2020, RStudio: Integrated Development Environment for R. RStudio, PBC., Boston, MA.

Souaissi, Z.–Ouarda, T.–St-Hilaire, A., 2023, Non-parametric, semi-parametric, and machine learning models for river temperature frequency analysis at ungauged basins. Ecol. Inform. 75, 102107. https://doi.org/10.1016/j.ecoinf.2023.102107

Toffolon, M.–Piccolroaz, S., 2015, A hybrid model for river water temperature as a function of air temperature and discharge. Environ. Res. Lett. 10, 114011. https://doi.org/10.1088/1748-9326/10/11/114011

Wade, J.–Kelleher, C.–Hannah, D., 2023, Machine learning unravels controls on river water temperature regime dynamics. J. Hydrol. 623, 129821. https://doi.org/10.1016/j.jhydrol.2023.129821

Zhu, S.–Bonacci, O.–Oskoruš, D.–Hadzima-Nyarko, M.–Wu, S., 2019, Long term variations of river temperature and the influence of air temperature and river discharge: case study of Kupa River watershed in Croatia. J. Hydrol. Hydromech. 67, 305 – 313. https://doi.org/10.2478/johh-2019-0019