Predviđanje kvalitete vode

|  |  |
| --- | --- |
| |  | | --- | | **Marko Bonovil**  *Prirodoslovno-matematički fakultet*  *Sveučilište u Splitu*  *Ruđera Boškovića 33*  [*mbonovil@pmfst.ht*](mailto:mbonovil@pmfst.ht?subject=Predviđanje%20kvalitete%20vode) | |

***Sažetak***

*Tvar koja prekriva više od polovice planeta Zemlje jedna je od esencijalnih tvari bez koje postojanje civilizacije ne bi bilo moguće. Sukladno tome se ovim radom nastoji pridonijeti očuvanju čistoće vode i njezine djelotvornosti na cjelokupni ekosustav. U tu svrhu primjenom algoritama strojnog učenja učinjen je značajan korak u procesu rane detekcije onečišćenja vode. Eksperiment opisan u radu obavljen je na dvama skupovima podataka primjenom modela slučajnih šuma, logističke regresije, stroja potpornih vektora i K-najbližih susjeda. Kao najdjelotvorniji algoritam iskazali su se slučajne šume koje su u oba slučaje pokazale izuzetne performanse utemeljene na evaluaciji korištenih modela.*

***Ključne riječi:*** *Kvaliteta vode; strojno učenje; evaluacija; predviđanje.*

# 1 Uvod

Voda predstavlja jednu od temeljnih tvari o kojoj uvelike ovisi život na Zemlji, a svojom kvalitetom odražava se na zdravlje ljudi te biljnog i životinjskog svijeta. Kvaliteta vode se brzo mijenja, a na nju utječe brojni čimbenici kao što su: prirodne nepogode (npr. poplava, padaline i sl.), onečišćenja (npr. onečišćenje vodom iz kanalizacijskog sustava, izljev otrovnih tvari u vodovodni sustav, poljoprivredna onečišćenja i sl.). Stoga je bitno razviti učinkovite metode za predviđanje kvalitete vode, kako bi prepoznali moguće prijetnje za zdravlje ljudi i okoliša.

Istraživanje je temeljeno na ispitivanju korištenih metodologija za određivanje kvalitete vode. Tipični postupak analize kvalitete vode obavlja se pomoću laboratorijske i statističke analize, dok postoje i postupci koji koriste algoritme strojnog učenja kao pomoć u pronalaženju kvalitetnog i prikladnog rješenja. Različiti modeli koriste različite varijable kao ulazne parametre kao što su kemijski, mikrobiološki, hidrološki i meteorološki parametri kako bi se izvela precizna predviđanja.

Kao i u svakom istraživanju tako i kod predviđanja kvalitete vode dolazi do pojave ograničenja u mogućnosti predviđanja, a najveće zapreke pojavljuju se pri prikupljanju podataka, jer mnogi skupovi podataka nisu javno dostupni ili ne sadrže ključne parametre nužne za predviđanje. Stvaranje vlastitog skupa podataka je skup i dugotrajan proces zbog složenosti promatranih vodotokova, potrebnog vremena za prikupljanje podataka i održavanje cjelokupnog tehničkog sustava za prikupljanje podataka, temeljnog na stanici za prikupljanje podataka. Ovaj tehnički sustav sadrži različite senzore, kamere i sl. te sustav za pohranu baze podataka koja čuva prikupljene podatke. Uzimajući u obzir veliku količinu podataka i nejasnu povezanost pojedinih mjerenja, strojno učenje je dobar kandidat koji može dati doprinos u brzini određivanja procjene kvalitete vode, te smanjivanja utjecaja ljudske pogreške na donošenje zaključaka.

# 2 Pregled literature

Pretraživanje literature obavljeno je sukladno smjernicama koje se koriste u sustavnim pregledima literature odnosno u PRISMA (engl. *Preferred Reporting Item for Systematic review and Meta-Analyses*) modelu (Moher i dr.,2009). Cilj navedenog postupka je identificiranje dostupnih istraživanja s navedenom temom, a sam postupak proveden je opsežno i temeljito. Kao baza za pretragu istraživanja koristila se Web of Science citatna baza.

Ključne riječi korištene za pretragu baza i pronalazak istraživanja su: "Water quality prediction", "Water quality", "Machine learning water quality prediction","Artificial intelligence water quality prediction". Riječi su odabrane sukladno pregledu dostupne literature kao i specifičnih riječi karakteristične za navedeno područje istraživanja. Literatura je pretražena do travnja 2023. godine bez ograničenja u pogledu godine publikacije ili jezika. Kriteriji za uključivanje i isključivanje istraživanja su jasno postavljeni, a proces odabira obuhvaćao je čitanje sažetaka radova ili cijelih radova, što se tiče duplikata oni su automatski isključivani iz istraživanja.

Ukoliko želimo da sustavni pregled literature provedemo kvalitetno i prema pravilima, potrebno je definirati i poštovati kriterije za uključivanje i isključivanje istraživanja. Navedeni korak se provodi nakon početnog pregleda literature te služi za definiranje PICO okvira, kao i određivanju karakterističnih, ključnih riječi po kojima će se pretraživati radove. Prije primjene kriterija za uključivanje i isključivanje u bazama podataka za pretragu radova utvrđeno je 18 052 rada koji su selektirani u različite kategorije, a detaljan pregled podjele radova prikazan je na slici 1.

Slika na kojoj se prikazuje tekst, snimka zaslona, Font, dizajn

Opis je automatski generiran

Slika 1 Podjela radova po kategorijama

Kriteriji za uključivanje koji su korišteni u ovom radu su sljedeći:

* Autori istraživanja su punoljetne osobe.
* Članak koristi postupke strojnog učenja i umjetne inteligencije.
* Članak je objavljen u časopisu ili na konferenciji.
* Sadrži detaljan opis postupka izvedbe istraživanja kao npr. opis skupa podataka, korištenih modela i sl.

Kriteriji za isključivanje korišteni u ovom radu su:

* Članci koji sadrže samo sažetak.
* Članci koji su duplicirani.
* Članak koji nije javno dostupan.
* Članci koji ne sadrže detaljan opis skupa podataka korištenih u istraživanju.
* Članci koji koriste više od jednog modela strojnog učenja.
* Članci koji ne koriste postupke strojnog učenja i umjetne inteligencije.

Pregled baze dostupnih radova obavljen je 6. travnja 2023. godine unosom ključnih riječi za odabrano područje istraživanja. Pretragom je utvrđeno da postoji 18 052 rada sa sličnom temom, primjenom kriterija za uključivanje i isključivanje taj se broj znatno smanjio. U prvom koraku isključeni su radovi koji nisu napisani na engleskom jeziku, a njih je u bazi radova Web of Science ukupno 217 čime je preostalo 17 835 radova za daljnje promatranje. Nadalje, kao kriterij se uzimao da područje istraživanja se odnosi na računalnu znanost (eng. *Research areas : Computer Science*) čime je uklonjeno dodatno 16 783 radova, te ih je preostalo 1 052 za daljnje selektiranje. Kao dodatan kriterij uključeno je da rad treba biti u obliku članka i da bude javno dostupan čime je odstranjeno dodatnih 799 radova iz ukupnog broja dobivenih radova, i na samom kraju kao dodatni selekcijski parametar korišteno je da kategorije koje pokrivaju radovi budu iz područja umjetne inteligencije, interdisciplinarne primjene računalne znanosti, programskog inženjerstva ili informacijskog sustava čime je uklonjeno dodatna 3 rada te je kao rezultat dobiven broj od 250 radova koja se proučavaju u daljnjem postupku.

Slika na kojoj se prikazuje tekst, snimka zaslona, paralelno, Font

Opis je automatski generiran

Slika 2 PRISMA model korišten u radu

Sukladno navedenom, na slici 2 detaljno je prikazan PRISMA model sa količinom radova u početnoj fazi istraživanja te razlozima uklanjanja određenih radova koji ne zadovoljavaju definirane kriterije istraživanja.

# 3 Metodologija

Svrha ovog istraživanja je odrediti kvalitetu vode korištenjem suvremenih tehnika strojnog učenja. Za tu svrhu koriste se dva prikupljena skupa podataka: Pitkost vode za piće ("Aditya Kadiwal. Drinking water potability", 2021) i Skup podataka za klasifikaciju kvalitete vode ("Mssmartypants. Water Quality", 2021) (eng. *Drinking water potability, Dataset for water quality classification*), dok su autori promatranih istraživanja koristili različite skupove podataka za određivanje kvalitete vode kao što su rezultati uzoraka vode iz rijeke Songhua u Kini, rijeke Eufrat u Iraku, jezera Gulshan u Bangladešu, rezultate uzoraka morske vode i mnoge druge. U tablici 1 nalazi se cjelokupni pregled promatranih radova sa podacima o skupovima podataka i algoritmima korištenim u istraživanju.

|  |  |  |
| --- | --- | --- |
| AUTOR | SKUP PODATAKA | ALGORITAM |
| (Liu i dr., 2018) | Yixing | DT, SVM, MC |
| (Xu i dr., 2022) | Australian ports | LR, DT, KNN, SVM, MLPC, NB |
| (Sheng i dr., 2020) | Songhua River and Victoria Bay | RNN, LSTM, BPNN |
| (Rajesh i dr., 2021) | Tunga-Bhadra river | RR, KNN, SVR, RF |
| (Al-Sulttani i dr., 2021) | Euphrates Rive | RF, SVM, GBM |
| (Aslam i dr., 2022) | Pakistan water | M5P, RT, RF,ANN,BPNN |
| (Ahmed i dr., 2019) | CRWR | MLR, PR, RF, GB, SVM, RR, LR |
| (Rustam i dr., 2022) | ("Mssmartypants. Water quality", 2021) | DT, RF, LR, SVM, ETC, AB, CNN, LSTM, GRU |
| (Liu i dr., 2013) | YiXing city | BPNN, SVR, RGA |
| (Khan i dr., 2022) | Gulshan Lake | MLR, SVR, GBR, RFR |
| (Grbčić i dr., 2022) | Rijeka Bay | GB, RF, SVR, ANN |
| (Suwadi i dr., 2022) | The Langat Basin in Selangor | ANN, SVM, RF, NB |
| (Hussein i dr., 2020) | GRACE | MLR, MP, RF, XGB, SVM, SVR |
| (Chinnappan i dr., 2023) | Raw dataset | KNN, SVM, DT |

## Tablica Promatrana istraživanja sa korištenim skupovima podataka i algoritmima

Eksperimentalna analiza i procjena kvalitete uzoraka vode izvršava se korištenjem modela strojnog učenja, a to su: slučajne šume (eng. *random forests*), stroj potpornih vektora (eng. *support vector machine*), logistička regresija (eng. *logistic regression*) i K-najbliži susjeda (eng. *K-nearest neighbors*). Korišteni modeli i skupovi podataka detaljnije su opisani u nastavku rada.

3.1 Skupovi podataka

Skup podataka pod nazivom Pitkost vode za piće ("Aditya Kadiwal. Water Quality", 2021) preuzet je sa internetske stranice Kaggle i sadrži 3276 zapisa testiranih uzoraka vode stoga u tablici 2 nalaze se mjerene komponente u uzorcima vode i njihove granične vrijednosti prema Svjetskoj zdravstvenoj organizaciji (World Health Organization, 2022).

|  |  |
| --- | --- |
| KOMPONENTE | DOPUŠTENE RAZINE[PO LITRI VODE] |
| pH | 6.5-8.5 |
| Tvrdoća | 100-300 |
| TDS | 500-1000 |
| Kloramin | 0-4 |
| Sulfat | 250-1000 |
| Provodljivost | 0-400 |
| Organski ugljik | 0-4 |
| Trihalometani | 0-80 |
| Zamućenost | 0-5 |

## Tablica Komponente mjerene u uzorcima vode i njihove granične vrijednosti po litri vode (Pitkost vode za piće)

Kao konačna vrijednost navodi se pitkost vode koja je označena brojem 0 ili 1, a odnosi se na to je li voda sigurna za piće. Ukoliko je voda sigurna za piće označena je brojem 1, a ako nije označena je brojem 0. Na slici 3 vizualno je prikazana distribucija pojedine komponente u cjelokupnom skupu podataka.

Slika na kojoj se prikazuje tekst, dijagram, radnja, crta

Opis je automatski generiran

## Slika Distribucija komponenti u skupu podataka Pitkost vode

Kao drugi skup podataka na kojem je temeljeno istraživanje korišten je Skup podataka za klasifikaciju kvalitete vode ("Mssmartypants. Water quality", 2021). Ovaj skup je preuzet sa internetske stranice Kagle, kao i prethodno definirani skup, a sadrži 7996 zapisa o testiranju komponenata na temelju kojih se donio zaključak je li voda prihvatljiva za piće ili ne. Jedina mjerna komponenta koja se pojavljuje u oba korištena skupa je Kloramin, a kako bi se bolje upoznali sa cjelokupnim spektrom testiranih tvari iz uzoraka vode u tablici 3 (Rustam i dr, 2022) nalazi se cijeli popis komponenata i dopuštenih razina koncentracije u litri vode.

|  |  |
| --- | --- |
| KOMPONENTE | DOPUŠTENE RAZINE [PO LITRI VODE] |
| Aluminij | 0-5.05 |
| Amonijak | -0.08-29.8 |
| Arsen | 0-1.05 |
| Barij | 0-4.94 |
| Kadmij | 0-0.13 |
| Kloramin | 0-8.68 |
| Krom | 0-0.9 |
| Bakar | 0-2 |
| Fluorid | 0-1.5 |
| Bakterije | 0-1 |
| Virusi | 0-1 |
| Voditi | 0-0.2 |
| Nitrati | 0-19.8 |
| Nitriti | 0-2.93 |
| Merkur | 0-0.1 |
| Perklorat | 0-60 |
| Radij | 0-7.99 |
| Selen | 0-0.1 |
| Srebro | 0-0.5 |
| Uran | 0-0.9 |

Tablica 3 Komponente mjerene u uzorcima vode i njihove granične vrijednosti po litri vode (Skupa podataka za klasifikaciju)

Kao i za prethodni skup, konačni rezultat je li voda za piće ili ne nalazi se u posljednjem stupcu i označena je brojem 0 ili 1 gdje su sa 1 označeni uzorci vode koja se može koristiti za piće, a sa 0 ona voda koja nije za piće.

3.2 Odabrani algoritmi strojnog učenja

U svrhu određivanja kvalitete vode korišteno je nekoliko algoritama strojnog učenja koji pomoću klasifikacijskog ili regresijskog pristupa rješavanju problema dolaze do konačnih rezultata. Kao što je navedeno u ovom eksperimentu korišteni su sljedeći algoritmi : logistička regresija, stroj potpornih vektora, slučajne šume i K-najbliži susjeda, a u daljnjem tekstu detaljnije je opisan svaki od njih prema osnovi načina izvršavanja i principa rada.

**3.2.1 Logistička regresija**

Logistička regresija predstavlja klasifikacijski tip modela strojnog učenja koji se korist za određivanje vrijednosti konačne varijable na temelju dostupnih podataka. Bazirana je na logističkoj ili sigmoidnoj funkciji stoga je poznata i kao sigmoida, dok je sigmoidna funkcija matematička funkcija koju karakterizira krivulja u obliku slova S poznatu pod imenom sigmoidna krivulja. Kao takav ima brojne primjene, a neke od njih su: funkcija aktivacije umjetnih neurona, funkcija kumulativne distribucije i sl. (Hosmer i dr.,2013).

**3.2.2 Slučajne šume**

Slučajne šume su vrsta modela strojnog učenja koji u osnovi koristi više osnovnih modela, tj. više stabala odlučivanja, kako bi došao do konačnog rezultata (Liaw, Wiener, 2002). Podaci na kojima se treniraju slučajne šume se stalno uzorkuju iz definiranog skupa kako bi se izbjegla pretreniranost i pristranost modela, a ta metoda poznata je pod nazivom bootstrapping. Leo Breiman smatra jednim od ključnih znanstvenika u teoriji slučajnih šuma jer je uvidio da se korištenjem bootstrapping metode povećava točnost i stabilnost algoritma (Breiman, 2001). Do pretreniranosti dolazi u slučajevima kada funkcija izvrsno procjenjuje rezultate, ali u slučaju promjene ulaznih podataka neće dati tako dobre rezultate stoga se podaci uvijek uzorkuju (Fitzgerald i dr., 2013).

**3.2.3 Stroj potpornih vektora**

Stroj potpornih vektora predstavlja jako učinkovit klasifikacijski i regresijski algoritam strojnog učenja čiji su kreatori Vladimir Vapnik i Alexey Chervonenkis (Vapnik, 1999a). Svrha stroja potpornih vektora je pronalazak linije razgraničenja, poznate i kao hiperravnina, klasa u podacima tako da udaljenost između točaka koje su najbliže liniji razgraničenja bude jednaka. Podaci nekada mogu biti nelinearno razdjeljivi stoga se u tu svrhu koriste polinomi višeg stupnja kako bi razdvojili podatke u klase te se taj proces naziva podizanje odnosno mapiranje u više dimenzije (Vapnik, 1999b).

**3.2.4 K-najbliži susjeda**

K-najbliži susjeda je algoritam strojnog učenja koji uglavnom rješava zadatke vezane uz klasifikaciju, zbog svog načina rada i izvedbe. Često se koristi kao prvi izbor u situacijama kada nemamo pretpostavki o distribuciji ulaznih podataka, a radi na principu da razdvaja podatke u različite klase odabirom parametra k koji predstavlja broj najbližih susjeda koji ulaze u proces klasifikacije. Algoritam karakterizira kasno učenje što znači da ne koristi trening podatke za bilokakvu generalizaciju stoga faza treniranja ili ne postoji ili je jako mala, zbog toga se cjelokupni skup podataka koristi za testiranje (Beyer, 1999).

3.3 Metrike

U svrhu ispitivanja točnosti korištenih modela strojnog učenja primjenjuju se određene metrike koje omogućavaju evaluaciju rezultata i mjerenje performansi korištenih algoritama. Prema tome u daljnjem tekstu opisane su metrike korištene u istraživanju.

**3.3.1 Točnost**

Točnost predstavlja jednu od najjednostavnijih metrika zbog toga što se računa kao omjer točnih rezultata, odnosno predviđanja podijeljen sa brojem ulaznih vrijednosti. Kao što je prikazano u jednadžbi 1 gdje su u brojniku istinski pozitivni i istinski negativni rezultati podijeljeni sa ukupnim brojem ulaza (Sokolova i dr., 2006).

(1)

**3.3.2 Preciznost**

Obzirom da točnost ne daje uvijek najsigurnije rezultate zbog nejednake raspodjele uzoraka. U tom slučaju koristi se preciznost koja predstavlja postotak ispravno klasificiranih instanci u odnosu na ukupan broj instanci, a jednadžba 2 prikazuje postupak kako se dolazi do iznosa preciznosti (Sokolova i dr., 2006).

(2)

**3.3.3 Opoziv**

Cilj računanja opoziva je dobivanje predodžbu o količini pozitivno klasificirane klase koji su ispravno klasificirani. Opoziv stoga može ukazati na to ima li algoritam sposobnosti pravile identifikacije relevantnih instanci. Sljedeća jednadžba (3) prikazuje postupak dobivanja vrijednosti opoziva gdje TP predstavlja istinski pozitivne instance dok FN predstavlja istinski negativne instance (Sokolova i dr., 2006).

(3)

**3.3.4 F1 rezultat**

F1 rezultat predstavlja harmonijsku sredinu preciznosti i opoziva budući da pojedinačno te dvije metrike ne pokrivaju sve aspekte točnosti. Stoga je svrha F1 rezultata da nam da cjelovitu sliku o izvedbi algoritma, a jednadžba (3) prikazuje postupak dobivanja F1 rezultata (Goutte i dr., 2005).

(4)

# 4 Rezultati

Primjenom modela strojnog učenja na skupove podataka koji su opisani u tekstu postigli su se rezultati koji su opisani u nastavku. Eksperiment je obavljen pomoću Google Colab bilježnice i programskog jezika Python.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Alg. | Točnost | Preciznost | Opoziv | F1 |
| LR | 0.62 | 0.76 | 0.76 | 0.75 |
| RF | 0.76 | 0.76 | 0.76 | 0.75 |
| SVM | 0.63 | 0.76 | 0.76 | 0.75 |
| KNN | 0.60 | 0.54 | 0.60 | 0.54 |

Tablica 4 Rezultati modela strojnog učenja na skupu Pitkost vode za piće

Cilj istraživanja je bio utvrditi je li voda čiji je sastav opisan u skupu podataka sigurna za piće ili ne. Rezultati eksperimentalne analize za prvi skup podataka prikazani su u tablici 4. Kao što prikazuju podaci, što se točnosti tiče slučajne šume imaju znatno bolje rezultate u odnosu na sve ostale dok se u cjelokupnom pogledu kao najlošiji algoritam pokazao KNN tj. K-najbliži susjeda. Nadalje što se F1 rezultata tiče slučajne šume RF, logistička regresija LR i stroj potpornih vektora SVM imaju iste rezultate dok K-najbliži susjed ima znatno manju vrijednost.

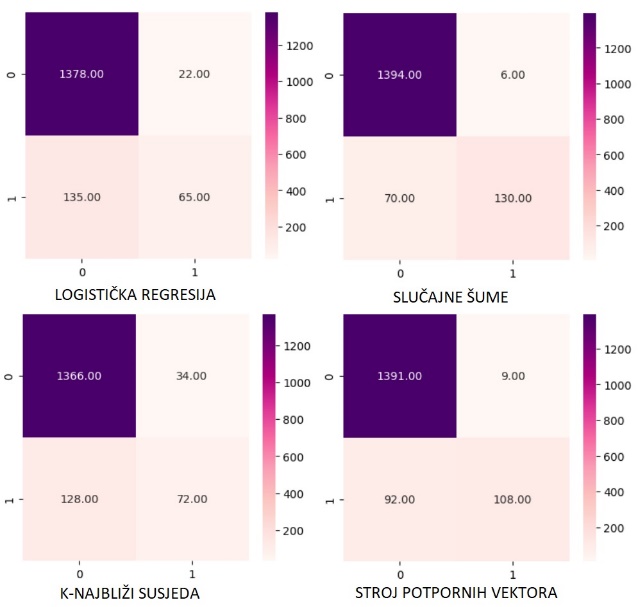
Rezultati eksperimenata vezanih uz Skup podataka za klasifikaciju kvalitete vode, prikazani u tablici 5, dobiveni su korištenjem istih algoritama kao i za prethodni skup podataka.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Alg. | Točnost | | K | Preciznost | Opoziv | F1 |
| LR | 0.90 | 0 | | 0.91 | 0.98 | 0.95 |
| 1 | | 0.75 | 0.33 | 0.45 |
| RF | 0.95 | 0 | | 0.95 | 1.00 | 0.97 |
| 1 | | 0.96 | 0.65 | 0.77 |
| SVM | 0.89 | 0 | | 0.94 | 0.99 | 0.96 |
| 1 | | 0.92 | 0.54 | 0.68 |
| KNN | 0.93 | 0 | | 0.91 | 0.98 | 0.94 |
| 1 | | 0.68 | 0.36 | 0.47 |

Tablica 5 Rezultati modela strojnog učenja na Skup podataka za klasifikaciju kvalitete vode)

Na osnovu prikazanih rezultata dolazi se do spoznaje da model slučajne šume tj. RF ima najbolje rezultate, koji se odnose na točnost i F1 rezultat, naspram ostalih korištenih modela. Model slučajnih šuma pokazao je znatnu razliku i u pogledu rezultata dobivenih na manjinskom skupu podataka, odnosno klasificiranju da je voda sigurna za piće zbog toga što takvi uzorci sačinjavaju manje od 12% cjelokupnog skupa podataka. Ostali modeli nisu pokazali zadivljujuće rezultate u tom pogledu jer samo stroj potpornih vektora u istom aspektu ima F1 rezultat veći od 0.50. Što se tiče većinskog dijela skupa podataka svi algoritmi imaju jako dobre rezultate npr. u pogledu preciznosti i F1 rezultata modeli su pokazali izuzetne sposobnosti.

Razlike između slučajnih šuma i ostalih modela može se okarakterizirati tim da za uvježbavanje slučajnih šuma nisu potrebni veliki skupovi podataka kao za ostale korištene modele. Bolju predodžbu o tome koliko je svaki od korištenih modela kvalitetno odradio klasifikaciju daju nam matrice konfuzije prikazane na slici 4.



Slika 4 Matrice konfuzije za korištene modele strojnog učenja na Skupu podataka za klasifikaciju kvalitete vode

Ponovno se model slučajnih šuma ističe točnim predviđanima kojih je ukupno imao 1524 dok je 76 neispravno klasificiranih uzoraka. Sljedeći model prema broju ispravno klasificiranih uzoraka je stroj potpornih vektora koji ima ukupno 1499 ispravno klasificiranih uzoraka naspram 101 neispravno klasificiranih. Model K-najbližih susjeda se opet pokazao kao najlošiji s obzirom da ima 1438 ispravno klasificiranih uzoraka naspram 162 neispravno klasificirana.

**5 Zaključak**

Voda predstavlja jednu od ključnih komponenti o kojoj ovisi postojanje života na zemlji, zbog toga se ovom studijom pokušalo doprinijeti njenom očuvanju i zaštiti. Primjenom algoritama strojnog učenja postiglo se na brz i efikasan način, konačne rezultate, u odnosu na tradicionalne postupke koji zahtijevaju duži vremenski period i kod kojih postoji utjecaj ljudske pogreške na krajnji rezultat. Algoritmi strojnog učenja testirani su na dvama skupovima podataka i svojim učinkom ostvaruju mogućnost sve veće implementacije algoritama u određivanje kvalitete vode. Kao najznačajniji algoritam korišten u eksperimentalnoj analizi pokazao se model slučajne šume koji svojom strukturom i rezultatima pokazao da veličina klase podataka ne utječe na model toliko izražajno kao na ostale korištene modele. Ovaj eksperiment otvara mogućnosti za daljnji napredak i usavršavanje kako korištenih algoritama tako i brojnih drugih kojima obiluje strojno učenje.

**Literatura**

Aditya Kadiwal. Drinking water potability (2021). Preuzeto 12. 04. 2023. sa <https://www.kaggle.com/datasets/adityakadiwal/water-potability>

Ahmed, U., Mumtaz, R., Anwar, H., Shah, A. A., Irfan, R., & García-Nieto, J. (2019). Efficient water quality prediction using supervised machine learning. Water, 11(11), 2210.

Al-Sulttani, A. O., Al-Mukhtar, M., Roomi, A. B., Farooque, A. A., Khedher, K. M., & Yaseen, Z. M. (2021). Proposition of new ensemble data-intelligence models for surface water quality prediction. IEEE Access, 9, 108527-108541.

Aslam, B., Maqsoom, A., Cheema, A. H., Ullah, F., Alharbi, A., & Imran, M. (2022). Water quality management using hybrid machine learning and data mining algorithms: An indexing approach. IEEE Access, 10, 119692-119705.

Beyer, K., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1999). When is “nearest neighbor” meaningful?. In Database Theory—ICDT’99: 7th International Conference Jerusalem, Israel, January 10–12, 1999 Proceedings 7 (pp. 217-235). Springer Berlin Heidelberg.

Breiman, L. (2001). Slučajne šume. Strojno učenje , 45 , 5-32.

Chinnappan, C. V., John William, A. D., Nidamanuri, S. K. C., Jayalakshmi, S., Bogani, R., Thanapal, P., ... & Syed Masood, J. A. I. (2023). IoT-Enabled Chlorine Level Assessment and Prediction in Water Monitoring System Using Machine Learning. Electronics, 12(6), 1458.

Fitzgerald, J., Azad, R. M. A., & Ryan, C. (2013, July). A bootstrapping approach to reduce over-fitting in genetic programming. In Proceedings of the 15th annual conference companion on Genetic and evolutionary computation (pp. 1113-1120).

Goutte, C., & Gaussier, E. (2005). A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In Advances in Information Retrieval: 27th European Conference on IR Research, ECIR 2005, Santiago de Compostela, Spain, March 21-23, 2005. Proceedings 27 (pp. 345-359). Springer Berlin Heidelberg.

Grbčić, L., Družeta, S., Mauša, G., Lipić, T., Lušić, D. V., Alvir, M., ... & Kranjčević, L. (2022). Coastal water quality prediction based on machine learning with feature interpretation and spatio-temporal analysis. Environmental Modelling & Software, 155, 105458.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley & Sons.

Hussein, E. A., Thron, C., Ghaziasgar, M., Bagula, A., & Vaccari, M. (2020). Groundwater prediction using machine-learning tools. Algorithms, 13(11), 300.

Khan, M. S. I., Islam, N., Uddin, J., Islam, S., & Nasir, M. K. (2022). Water quality prediction and classification based on principal component regression and gradient boosting classifier approach. Journal of King Saud University-Computer and Information Sciences, 34(8), 4773-4781.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.

Liu, S., Tai, H., Ding, Q., Li, D., Xu, L., & Wei, Y. (2013). A hybrid approach of support vector regression with genetic algorithm optimization for aquaculture water quality prediction. Mathematical and Computer Modelling, 58(3-4), 458-465.

Liu, S., Xu, L., Li, Q., Zhao, X., & Li, D. (2018). Fault diagnosis of water quality monitoring devices based on multiclass support vector machines and rule-based decision trees. IEEE Access, 6, 22184-22195.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group\*, T. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Annals of internal medicine, 151(4), 264-269.

Mssmartypants. Dataset for water quality classification (2021). Preuzeto 12. 04. 2023. sa <https://www.kaggle.com/datasets/mssmartypants/water-quality?select=waterQuality1.csv>

Rajesh, M., & Rehana, S. (2021). Prediction of river water temperature using machine learning algorithms: A tropical river system of India. Journal of Hydroinformatics, 23(3), 605-626.

Rustam, F., Ishaq, A., Kokab, S. T., de la Torre Diez, I., Mazón, J. L. V., Rodríguez, C. L., & Ashraf, I. (2022). An Artificial Neural Network Model for Water Quality and Water Consumption Prediction. Water, 14(21), 3359.

Sheng, L., Zhou, J., Li, X., Pan, Y., & Liu, L. (2020). Water quality prediction method based on preferred classification. IET Cyber‐Physical Systems: Theory & Applications, 5(2), 176-180.

Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In AI 2006: Advances in Artificial Intelligence: 19th Australian Joint Conference on Artificial Intelligence, Hobart, Australia, December 4-8, 2006. Proceedings 19 (pp. 1015-1021). Springer Berlin Heidelberg.

Suwadi, N. A., Derbali, M., Sani, N. S., Lam, M. C., Arshad, H., Khan, I., & Ki-Il, K. (2022). An Optimized Approach for Predicting Water Quality Features Based on Machine Learning. Wireless Communications & Mobile Computing (Online), 2022.

Vapnik, V. (1999). The nature of statistical learning theory. Springer science & business media.

Vapnik, V. N. (1999). An overview of statistical learning theory. IEEE transactions on neural networks, 10(5), 988-999.

World Health Organization. (2022). Guidelines for drinking-water quality: incorporating the first and second addenda. World Health Organization.

Xu, X., Lai, T., Jahan, S., Farid, F., & Bello, A. (2022). A Machine Learning Predictive Model to Detect Water Quality and Pollution. Future Internet, 14(11), 324.

**Abstract**

The substance that covers more than half of the planet Earth is one of the essential substances without which the existence of civilization would not be possible. Accordingly, this work aims to contribute to the preservation of water purity and its effectiveness on the entire ecosystem. For this purpose, a significant step was taken in the process of early detection of water pollution by applying machine learning algorithms. The experiment described in the paper was performed on two data sets using the random forest model, logistic regression, support vector machine and K-nearest neighbors. The most effective algorithm was random forests, which in both cases showed exceptional performance based on the evaluation of the used models.

**Keywords:** Water quality; machine learning; evaluation; prediction.