SVEUČILIŠTE JOSIPA JURJA STROSSMAYERA U OSIJEKU

FAKULTET ELEKTROTEHNIKE, RAČUNARSTVA I INFORMACIJSKIH TEHNOLOGIJA

Klasifikacija pitkosti vode

Seminarski rad

Raspoznavanje uzoraka i strojno učenje

Ivan Bačić

Osijek, 2021.

SADRŽAJ

[1. UVOD 3](#_Toc79528742)

[2. NADZIRANO UČENJE 4](#_Toc79528743)

[2.1 Logistička regresija 5](#_Toc79528744)

[2.2 K-NN klasifikacija 6](#_Toc79528745)

[2.3 Algoritam slučajne šume 7](#_Toc79528746)

[3. RJEŠENJE PROBLEMA 8](#_Toc79528747)

[3.1 Skup Podataka 8](#_Toc79528748)

[3.2 Biblioteke 12](#_Toc79528749)

[3.3 Parametri i učenje modela 13](#_Toc79528750)

[4. REZULTATI 14](#_Toc79528751)

[5. ZAKLJUČAK 15](#_Toc79528752)

# 1. UVOD

Pristup vodi sigurnoj za piće je neophodan za normalan život ljudi. Na lokacijama gdje i postoji dovoljno vode ne znači uvijek da je sigurna za piće, tako se danas sve više pristupa testiranju izvora vode za piće. U ovom seminaru će se uzeti takav jedan skup podataka o kvaliteti vode kako bi se izgradio model za klasifikaciju vode s obzirom na njezina zabilježena svojstva. Ovakavim pristupom se pokušava pojednostaviti detekciju vode koja nije sigurna za piće i kako bi se jednostavnije moglo pristupiti rješavanje problema zagađenja izvora vode.

Glavna ideja ovog rada je trenirati model koji će klasificirati je li voda pitka s obzirom na testirane parametre. Bitno je dobro pregledati skup podataka kako se model ne bi pogrešno istrenirao jer podaci nisu pripremljeni za ovakvu obradu. Klasifikcije će se vršiti pomoći Logističke regresije, K- najbližih susjeda i „Random Forest“ postupka.

# 2. NADZIRANO UČENJE

Nadzirano učenje je način strojnog učenja gdje je cilj odrediti nepoznatu funkcionalnu ovisnost između ulaznih veličina i izlazne veličine na temelju podatkovnih primjera. Podatkovni primjeri su parovi koji se sastoje od ulaznih vektora i vrijednosti izlazne veličine.

Kad je *y* (izlazna veličina) diskretna tada se ova metoda naziva klasifikacija, a kada je *y* kontinuirana veličina tada se metoda naziva regresija.

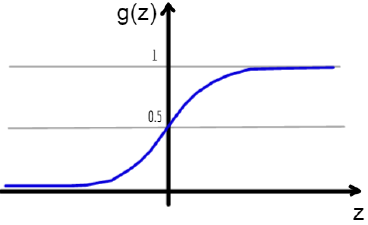
Algoritam nadziranog učenja analizira trening podatke i definira funkciju koja se koristi za predviđanje novih izlaza s obzirom na nove ulaze. Optimalno, algoritam će točno predvidjeti novi rezultat za podatke koje do sada nije vidio (nisu korišteni tijekom učenja). To predviđanje novih rezultata bi trebao raditi sa što manjom greškom, mjerenje točnosti algoritma se mjeri generaliziranom greškom.

Kod nadziranog učenja vrlo bitan korak je prikupljanje podataka i označavanje istih, kako ne bi došlo do velikih razlika između modela i stvarnih primjena.

## 2.1 Logistička regresija

Logistička regresija se koristi kako bi se povećala robusnot linearne regresije kao klasifikatora, koristi se logistička funkcija umjesto linearne.

gdje je



Slika 1: log funkcija

Granica odluke :

ako je

ako je

Granica odluke je hiper ravnina (za više od tri ulazna parametra).

Izlaz iz modela logističke regresije se može shvatiti i kao vjerojatnost da podatak *x* pripada klasi.

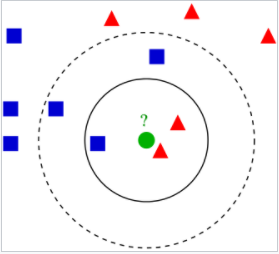
Kriterijska funkcija za određivanje parametara logističke regresije:

Rješenje ovog optimizacijskog problema ne postoji u zatvorenoj formi kao u slučaju regresije pa se moraju koristiti iterativne metode optimizacije.

## 2.2 K-NN klasifikacija

Klasifikacijska metoda K najbližih susjeda pripada grupi neparametarskih metoda. U takvim modelima ne postoje parametri koje je najprije potrebno procijeniti na temelju skupa učenja, već se predikcija za neku novu vrijednost ulaznih veličina *x* (testni uzorak) određuje izravno na temelju raspoloživih podataka (skupa učenja). Novi mjerni uzorak se kasificira na temelju klase njemu najbližih K susjeda. Npr. ako je od 5 najbližih susjeda (uzorka), 4 uzorka klase „0“, a 1 uzorak je klase „1“, tada se novi uzorak klasificira kao „0“. Pri tome se pronalaženje K najbližih susjeda svodi na izračunavanje udaljenosti novog (testnog) uzorka do svih uzoraka skupa za učenje. Najčešće se koristi euklidska udaljenost. Matematički se klasifikacija metodom K najbližih susjeda definira na način:

gdje skup S sadrži K najbližih susjeda uzorku *𝐱*, a 𝐼(𝑎=𝑏) je funkcija koja je jednaka 1 ako je 𝑎=𝑏, a u suprotnom je jednaka 0. K najbližih susjeda je vrlo jednostavan i popularni algoritam koji stvara nelinearnu granicu odluke u ulaznom prostoru i može se koristiti za složene probleme klasifikacije. Međutim, mane ovog algoritma su memorijska i računalna zahtjevnost.



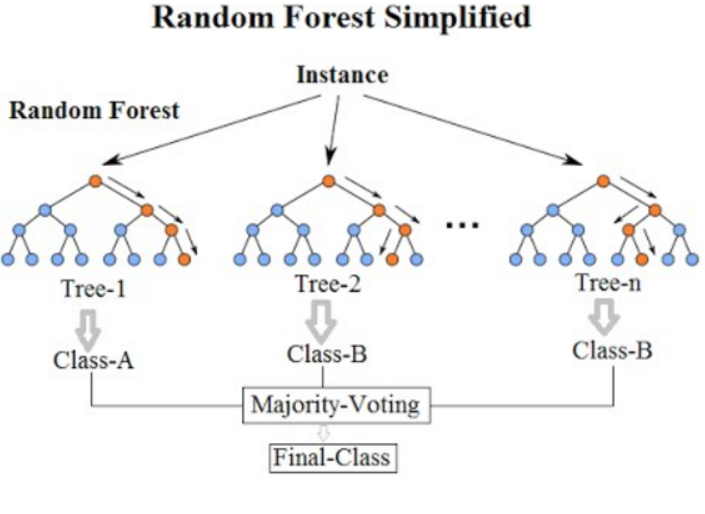
Slika 2: k-NN primjer

Primjer kako k-NN klasifikacije izgleda, crvena točka je novi ulaz, a plavi kvadrat i crveni trokut su već određene klase poznatih podataka.

## 2.3 Algoritam slučajne šume

Klasifikator slučajne šume je estimator koji uzima nekoliko klasifikatora stabla odlučivanja na različitim podskupovima podataka i koristi prosječna rješenja za regresiju ili kod klasifikacije odabire onu klasu koja je odabrana od najviše stabala za taj ulaz kako bi povećao preciznost i smanjio pretjerano usklađivanje (eng. „over-fitting“) na kontrolnom skupu podataka. Veličina podskupa se kontrolira s „max\_samples“ parametrom ili se koristi cijeli set trening podataka za svako podstablo.

Algoritam slučajne šume rješava problem pretjeranog usklađivanja kod stabala odluke.



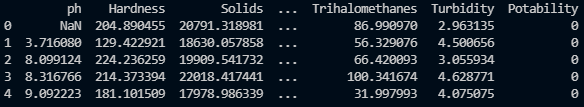
Slika 3: nausmična šuma primjer

# 3. RJEŠENJE PROBLEMA

## 3.1 Skup podataka

Skup podataka se sastoji od različitih parametara s kojima se mjeri kvaliteta vode; je li voda za piće ili nije.

1. **ph**: pH vode (0 to 14).
2. **Tvrdoća (eng. Hardness)**: kapacitet vode za taloženje sapuna u mg/L.
3. **Krute tvari (eng. Solid)s**: ukupne otopljene krute tvari u ppm.
4. **Kloramini (eng. Chloramines)**: količina kloramina u ppm.
5. **Sulfat**: količina sulfata otopljenih u mg/L.
6. **Vodljivost (eng. Conductivity)**: električna vodljivost vode u μS/cm.
7. **Organski ugljik (eng. Organic\_carbon)**: količina organskog ugljika u ppm.
8. **Trihalometani (eng. Trihalomethanes)**: količina trihalometana u μg/L.
9. **Zamućenost (eng. Turbidity)**: mjera svojstva emitiranja svjetlosti vode u NTU.
10. **Pitkost (eng. Potability)**: pokazuje je li voda sigurna za piće. Pitka - 1 i nije pitka – 0.



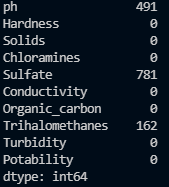
Slika 4: prvih pra podataka iz skupa

Ukupno ima 3276 vektora podataka po 10 parametara u svakom vektoru.



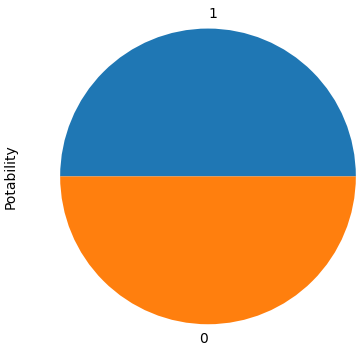
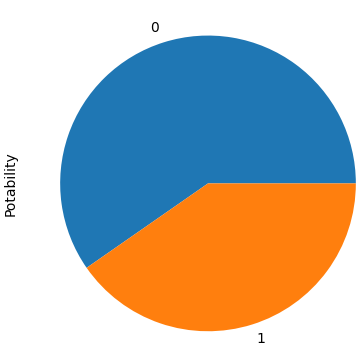
Slika 5: dimenzije podataka

S obzirom da je ovakav skup podataka osjetljiv na nedostajuće parametre, takvi podaci se brišu iz algoritma učenja jer se njihove nedostajuće vrijednosti ne mogu nadoknaditi sa srednjom vrijednošću ili sličnom metodom.



Slika 6: nepostojući podaci

Podaci sadrže znatno više primjera kada voda nije pitka, što se vidi i na „pi chart“ slici..



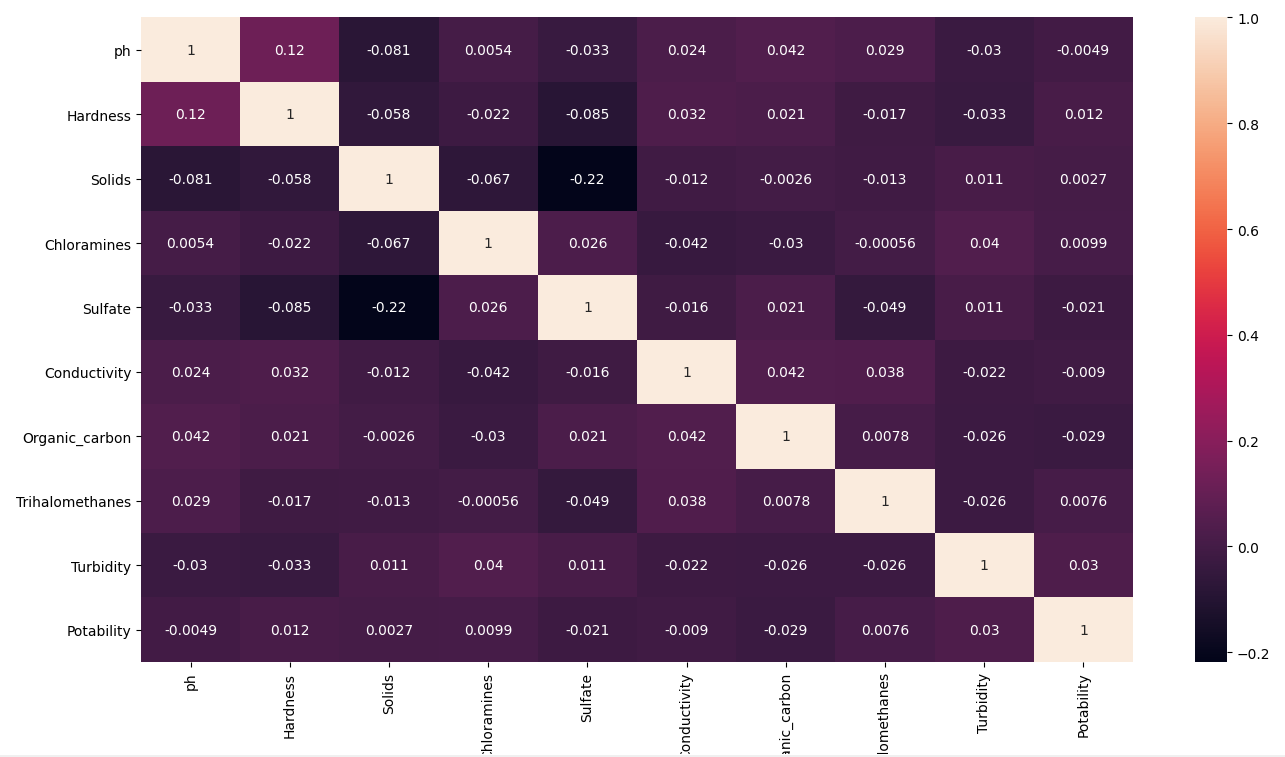
Slika 7: omjer klasa

Podatke za pitku vodu ćemo nadomjestiti „upsample“ metodom tj. neke od podataka kada je voda pitka (potability == 1) ponovno ćemo iskoristiti.



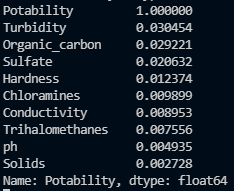
Slika 8: nadomještanje podataka

Zatim je provjereno postoji li korelacija među parametrima. Na slici 9. vidi se da nema značajne korelacije između parametara, što znači da ne možemo smanjiti dimenzionalnost skupa.



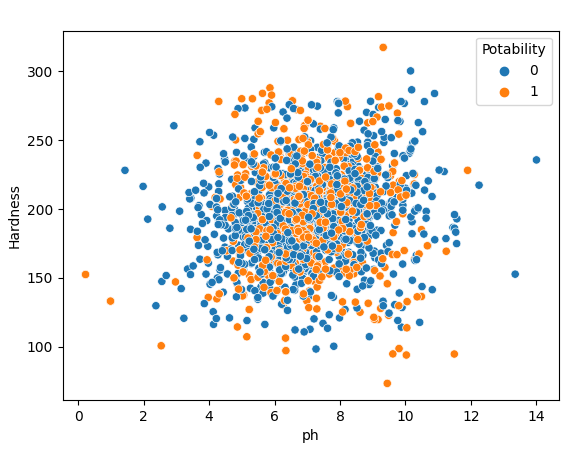
Slika 9: matrica korelacija parametara

Također, važno je provjeriti korelaciju između veličine za koju treniramo model -> korelacije ponovno nema.

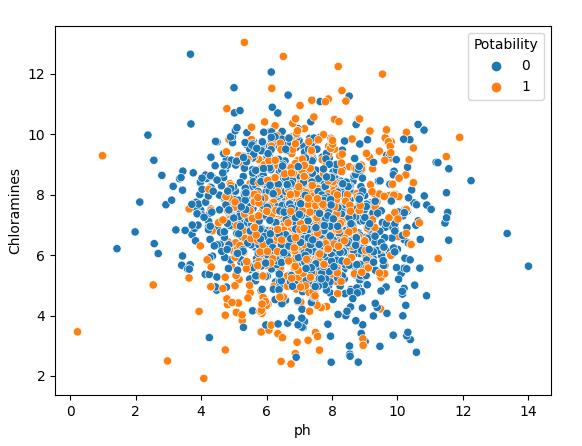


Slika 10: korelacija pitkosti s ostalim parametrima

Točkastim dijagramom provjereno je postoji li uzorak među parametrima, nakon provjere nekoliko od njih, zaključeno je da uzorka nema.



Slika 11: točkasti diagram, tvrdoća i ph



Slika 12: točkasti diagram, kloramini i ph

# 3.2 Biblioteke

**Numpy:** biblioteka za efikasniji rad s velikim vektorima i matricama.

**Pandas:** koristi se za učitavanje podataka iz datoteke.

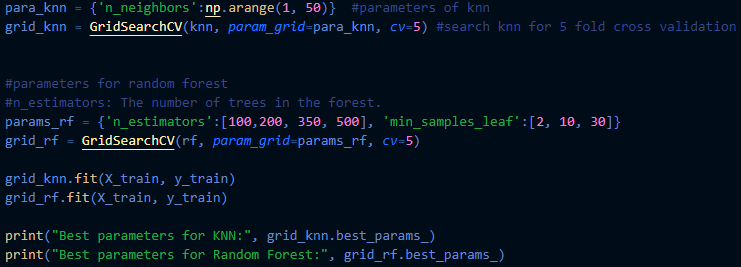
**Matplotlib:** biblioteka koja dodaje funkcionalnosti za prikaz i analizu podataka sličnim kao u Matlab-u; u ovom radu se koristi za prikaz različitih grafova.

**Seaborn:** proširenje Matplotlib biblioteke.

**Sklearn: s**adrži brojne algoritme strojnog učenja, kao što su algoritmi koji se koriste u ovom seminaru (Logistic regression, random forest i kNN).

## 3.3 Parametri i učenje modela

Traženje optimalnih parametara je obavljeno pomoću funkcije *GridSearchCV*; pronađeni su parametri za *kNN* i *Random Forest*, dok su za logističku regresiju korišteni početni parametri.



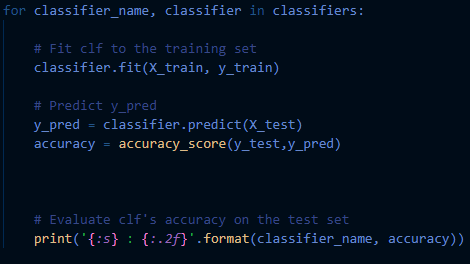
Slika 13: traženje optimalnih parametara modela



Slika 14: najbolji parametri

# 4. REZULTATI

Preciznost modela se procjenjuje pomoću funkcije „accuracy\_score“ koja vraća preciznost svakog modela.



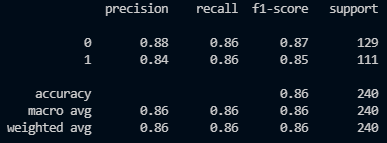
Slika 15: treniranje modela

Najbolji rezultat daje Algoritam slučajne šume s preciznošću od 86%, dok je logistička regresija najlošija i ne uspijeva točno klasificirati pola testnih podataka.



Slika 16: rezultati modela

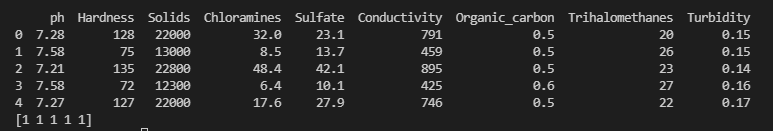
Kod daljne analize rezultata nasumične šume možemo vidjeti da kod klasifikacije preciznost kada je stvarna klasa 0 je 88% a kada je klasa 1 onda 84%, što znači da model točno predvidi klasu 0 s 88% točnosti a klasu 1 s 84% točnosti.

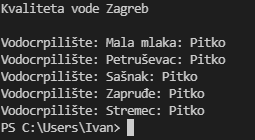


Slika 17: rezultati modela nasumične šume

## 4.1. Testiranje kvalitete vode Zagreb

Manji mjerni skup je preuzet sa službene stranice vodoopskrba zagreb i testirani su podaci kako bi se vidjela kvaliteta vode. Test je poroveden na naj boljem modelu Random Forest.





# 5. ZAKLJUČAK

Klasifikacija se danas sve više primjenjuje u raznim područjima života kako bi ubrzala i olakšala mnoge zadatke za koje su inače potrebni eksperti. To se postiže prikupljanjem dovoljnog broja podataka o problemu. Glavna prepreka primjeni ovakvih modela su podaci odnosno njihova točnost, mogućnost mjerenja širokog spektra radnog područja specifičnog problema. U korištenom skupu podataka glavni problem je što su neki od parametara uglavnom slični. 92% podataka vode se smatra tvrdom vodom, 98% vode ima veliku količinu sumpora. Ovakvi problemi se rješavaju boljim prikupljanjem podataka.

Najbolji klasifikator je nasumična šuma s preciznošću od 86%. Traženje parametara je vršeno pomoću funkcije „GridSearchCV“ koja pronalazi optimalne parametre za klasifikatore u zadanom intervalu kod k-NN ili odabire parameter iz zadanih vektora kod nasumične šume.

Logistička regresija ima preciznost od 51%, što je praktički pogađanje klase, na ovim podacima logistička regresija nije uspjela dobro predvidjeti klase podataka.