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 13](#_Toc79528749)

[3.3 Parametri i učenje modela 14](#_Toc79528750)

[4. REZULTATI 15](#_Toc79528751)

[5. ZAKLJUČAK 16](#_Toc79528752)

# 1. UVOD

Pristup vodi sigurnoj za piće je neophodno za normalan život ljudi. Na lokacijama gdije i postoji dovoljno vode ne znači uvijek da je sigurna za piće, tako se danas sve više pristupa testiranje 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 urzati detekcija vode koja niju 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 sestirane parametre. Bitno je dobro pregledati skup podataka kako se nebi model pogrešno istrenira jer podatci 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 gdije je cilj odrediti nepoznatu funkcionalnu ovisnost između ulaznih veličina i izlazne veličine na temelju podatkonih 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.

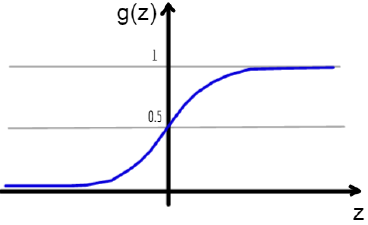
Algoritam nadziranog učenjea analizira trening podatke i definira funkciju koja se koristi za predviđanje novih izlaza s obzirom na nove ulaze. Optimalno algoritam će točno predvidjeti novi rezultat za podatke koje do sada nije vidio(nisu korišteni tijekom učenja). To predviđanje novih rezultata bi trebao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

Kod nadziranog učenja vrlo bitan kork je prikupljanje podataka i označavanje istih, kako ne bi došlo do velikih razlika imzeđu modela i svarnih primjena.

## 2.1 Logistička regresija

Logistička regresija se koristi kako bi se povećala robusnot linearne regresije kao klasifikatora, koristi se logistička funkcije umijesto linearne.

gdije je



Granica odluke :

ako je

ako je

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

Izlaz iz modela ogističke regresije se može skvatiti i kao vjerojatnost da podatak x pripada klasi.

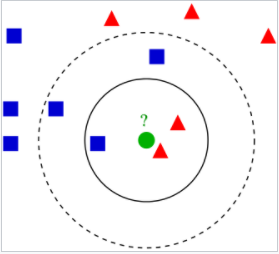
Kriteriska funkcija za određivanje parametara logističke regresije:

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

Klasifikacisk metoda K najbližih susjeda pripda grupi neprametarskih metoda. U takvim modelima ne postoje parametri koje je najprije potrebno procijeniti na temelju skupa učenja već se predikcija za neku novu vrijednost ulaznih beličin 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.

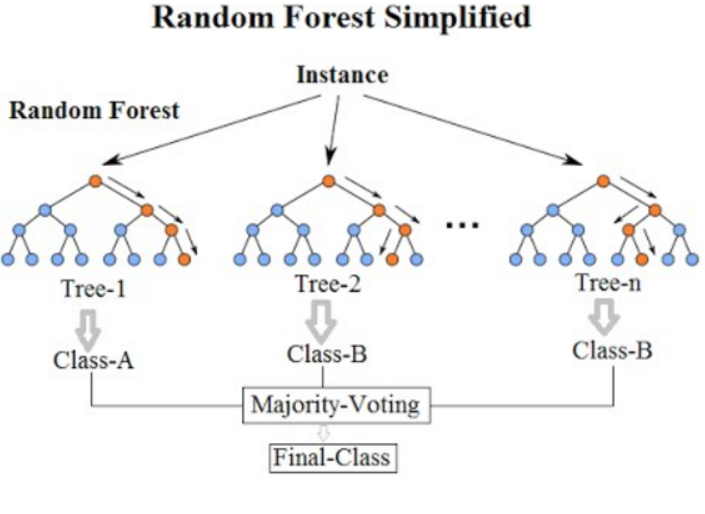


Primjer kako k-NN klasifikacije izgleda, Crvena točka je novi ulaz a plavi kvadrat i crveni trokut su već određene klase poznatih podataka.

## 2.3 Algoritam slučajne šume

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

Slučajne šume rješava problem pretjeranog usklađivanja kod stabala odluke.

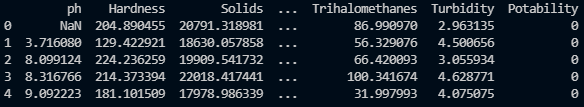


# 3. RJEŠENJE PROBLEMA

## 3.1 Skup Podataka

Skup podataka se stoji od različitih parametara s kojima se mjeri kvaliteta vode, je li voda za piće ili ne.

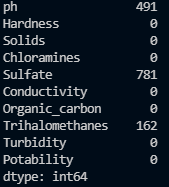
1. **ph**: pH vode (0 to 14).
2. **Tvrdoča (eng. Hardness)**: Kapacitet vode za taloženje sapuna u mg/L.
3. **Krute tvari (eng. Solid)s**: ukupne otopljene krute tvari u ppm.
4. **Kloramini (eng. Chloramines)**: Količina kloramina u ppm.
5. **Sulfat**: količina sulfata otopljenih u mg/L.
6. **Vodljivost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



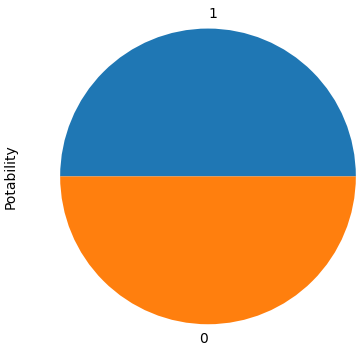
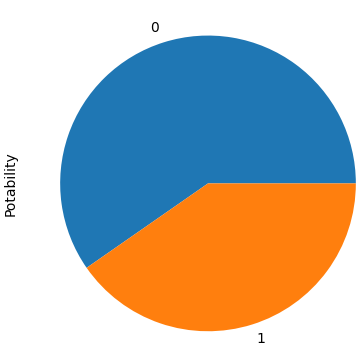
Ukupno ima 3276 vektora podataka po 10 Parametara u svakom vektoru.



Pošto je ovakav skup podataka osjetljiv na nedostajuče parametre, takvi podatci se brišu iz algoritma učenja jer se njihove nedostajuče rijednosti nemogu nadoknaditi s srednom vrijednošću ili sličnom metodom.



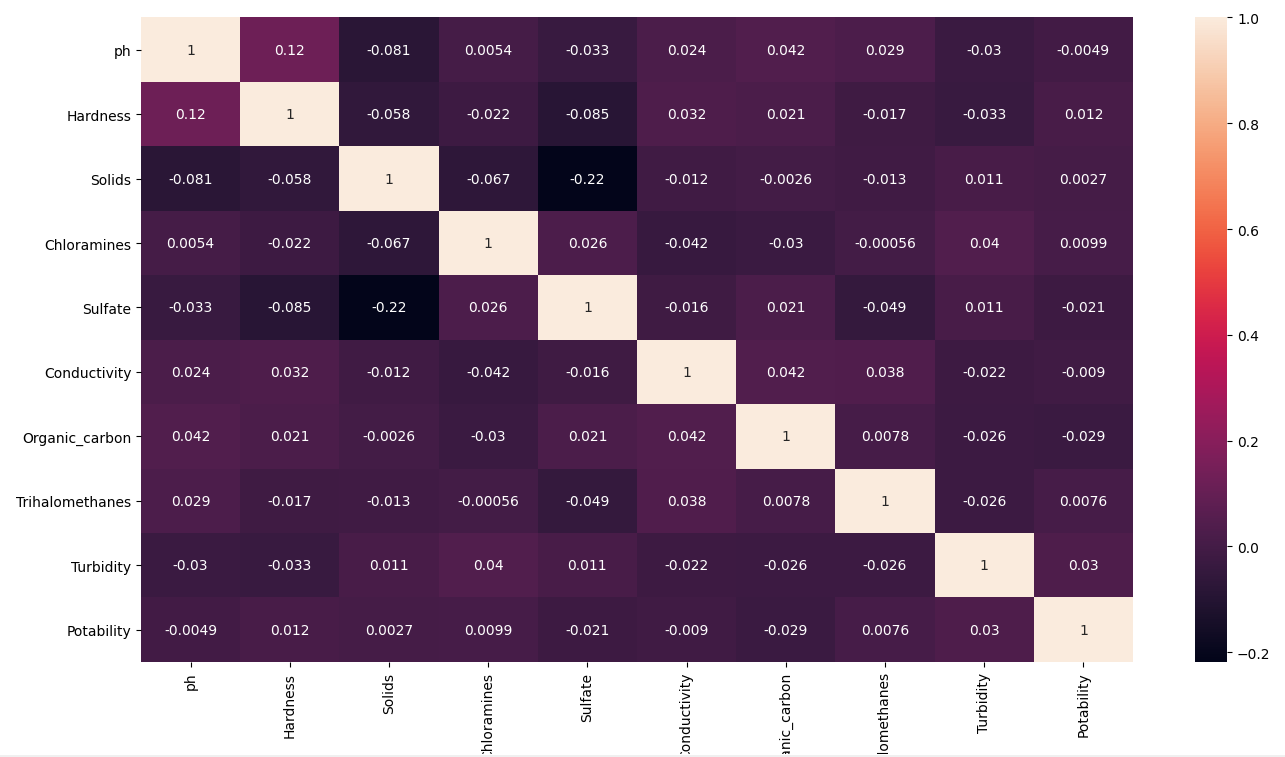
Podaci sadrže znatno više primjera kada voda nije pitak, što se vidi i na „pi chart“ slici ispod.



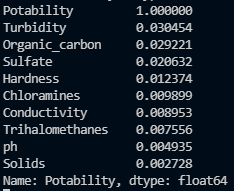
Podatke za ptiku vodu ćemo nadomjestiti „upsample“ metodom tj. neke od podataka kada je voda pitak (potability == 1) će mo ponovno iskoristiti.



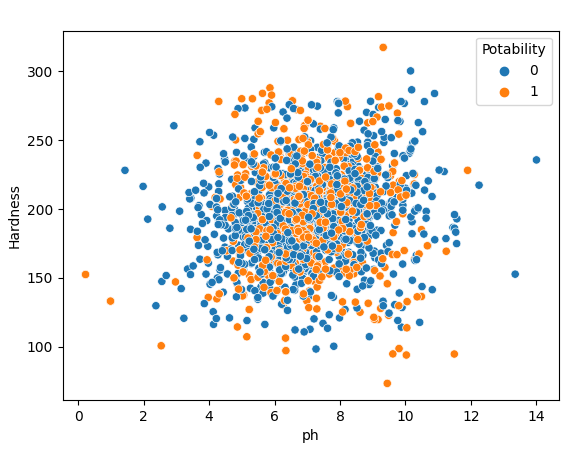
Zatim je provjereno je li posoji korelacija između parametrima. Na slici ispod se vidi da nema značajne korelacije između parametara, što znači da nemožemo smanjiti dimenzionalnost skupa.

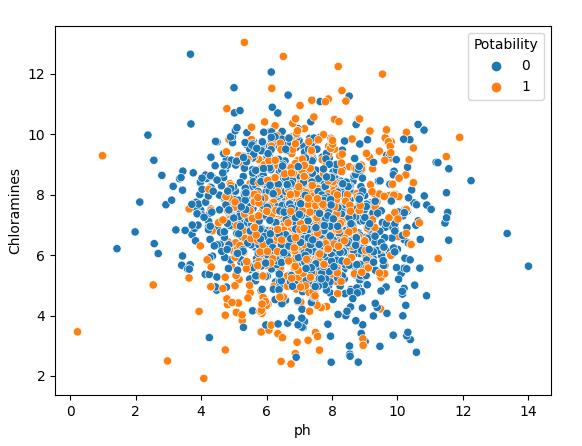


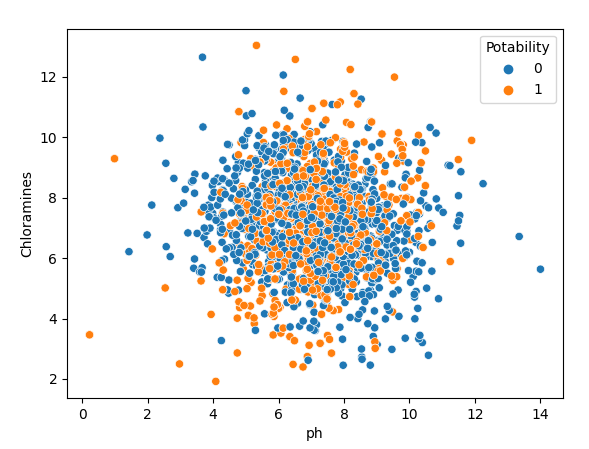
Također važno je provjeriti korelaciju izmežu veličine za koju trniramo model, Korelacije ponovno nema.



Točkastim dijagramaom je provjereno postoji li uzorak međuu parametrima, nakon porvjere nekoliko od njih, viđeno je da uzorka nema.







# 3.2 Biblioteke

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

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

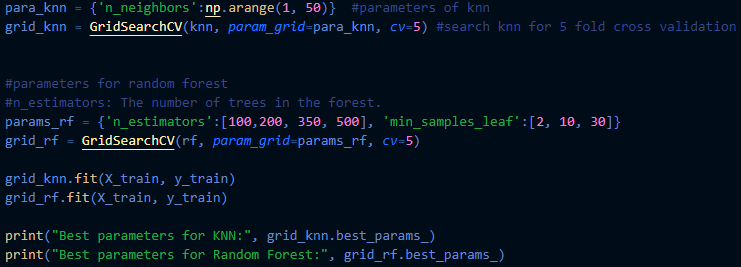
**Matplotlib:** biblioteka koja dodaje funkcionalnsti za prikaz i analizu podataka sličnim kao u Matlab-u, u ovom radu se koristi za prikaz različitih grafova.

**Seaborn:** proširenje Mtplotlib bilbioteke.

**Sklearn:** Scikit-learn je biblioteka koja se koristi u Python programskom jeziku. Sadrži brojne algoritme strojnog učenja, kao što su algoritmi koji se koriste u ovom sminaru (Logistic regression, random forest i kNN).

## 3.3 Parametri i učenje modela

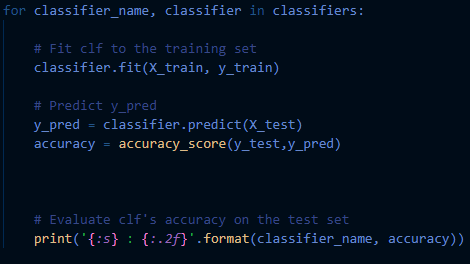
Traženje optimalnih parametara je rađeno pomoću funkcije GridSearchCV, pronađeni su parametri za kNN i Random Forest, dok su za logističku regresiju korišteni početni parametri.





# 4. REZULTATI

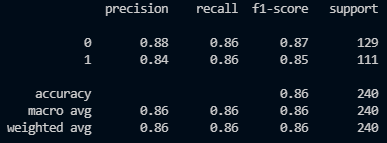
Preciznost modela se procijenjuje pooću funkcije „accuracy\_score“ koja vraća preciznost svakog modela.



Najbolji rezultat daje Algoritam slučajne šume s preciznošću od 86%, dok je logističa regresije najlošija i ne uspijeva točno klacificirati velik broj testinh podataka.



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



# 5. ZAKLJUČAK