Ekspertni sustav za klasificiranje podataka

Gaće, Marin

Master's thesis / Diplomski rad

2021

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **Josip Juraj Strossmayer University of Osijek, Faculty of Electrical Engineering, Computer Science and Information Technology Osijek / Sveučilište Josipa Jurja Strossmayera u Osijeku, Fakultet elektrotehnike, računarstva i informacijskih tehnologija Osijek**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:200:941361

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-11-22**

Repository / Repozitorij:

Faculty of Electrical Engineering, Computer Science and Information Technology Osijek

SVEUČILIŠTE JOSIPA JURJA STROSSMAYERA U OSIJEKU

FAKULTET ELEKTROTEHNIKE, RAČUNARSTVA I INFORMACIJSKIH TEHNOLOGIJA OSIJEK

Sveučilišni diplomski studij računarstva

EKSPERTNI SUSTAV ZA KLASIFICIRANJE PODATAKA

Diplomski rad

Marin Gaće

Osijek, 2021.

SADRŽAJ

1. UVOD

Klasifikacija se može opisati kao razvrstavanje nekoga skupa na manje skupove. Ako promatramo neku jedinku klasifikacijom tu jedinku svrstavamo u neki skup. Svrstavanjem jedinke u skup definiramo samu jedinku. Ljudi sve klasificiraju jer na taj način lakše razumiju stvari. Ako nešto klasificiramo odmah nam je lakše shvatiti kako stvari funkcioniraju, znamo s čime trebamo neku jedinku uspoređivati, što je razlikuje od drugih, a što dijeli s drugim jedinkama.

Stari filozofi su rekli da je klasifikacija svuda oko nas, što je i istina ako pogledamo primjere iz svakodnevnog života kao što su prijevozna sredstva, cjelokupan živi svijet, priroda, svemirska tijela, imaginarni pojmovi, brojevi, protokoli i slično. Drugim riječima, može se reći da gdje god postoji podjela postoji i klasifikacija. Kako bi se nešto pravilno klasificiralo moraju se znati neka pravila klasificiranja. Neka pravila su univerzalna za sve klasifikacije dok se druga razlikuju ovisno o tome što se želi klasificirati. Osnovno pravilo klasificiranja je da se jasno definiraju kriteriji koji će se koristiti za razlikovanje klasa.

U radu su opisani klasifikatori koji služe kao alat za klasifikaciju, a razlikuju se s obzirom na područje na koje se primjenjuju. Osim te podijele, postoji i niz drugih podjela klasifikatora prema načinu rada, prema odnosu rada, odnosno, vidi li se kako rade ili ne. Neki klasifikatori mogu učiti tako da što ih više primjenjujemo oni daju bolje rezultate.

Klasifikacija podataka u medicini vrlo je važna. Imati programsko rješenje koje klasificira medicinske podatke skraćuje vrijeme potrebno za obradu podataka, pomaže pri odlučivanju, a može služiti i u edukativne svrhe.

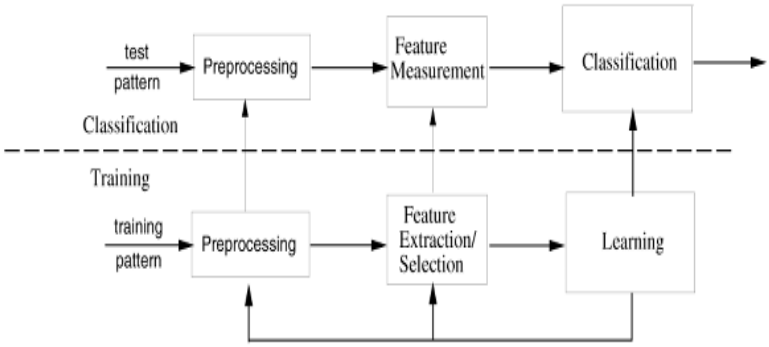
Klasifikatori koji su implementirani u programskom rješenju jesu klasifikatori temeljeni na udaljenosti. Takvi klasifikatori daju dobre rezultate te se mogu primijeniti na širokom području, što je razlog zašto su odabrani za implementaciju u ovom radu. Uz to, implementiran je i Bayesov klasifikator koji je jedan on najpoznatijih i najraširenijih klasifikatora na kojem je lako pokazati princip i osnove klasificiranja kako bi se dobiveni rezultati mogu usporediti.

U drugom poglavlju „Pregled područja klasifikacije podataka“ opisano je što je to klasifikacija, što su klasifikatori, kako i kada se koriste te kako se vrednuju. U trećem poglavlju opisano je kako se klasifikatori primjenjuju u medicini i ostalim znanstvenim granama te su opisani neki od klasifikatora koji se koriste. Za navedene klasifikatore opisane su karakteristike pojedinih klasifikatora kao i algoritmi preko kojih se vrši klasifikacija. Četvrto poglavlje daje osvrt na programsko rješenje za klasificiranje. Opisana je ideja, korišteni algoritmi i metode te su dane i upute za korištenje programskog alata. U istom poglavlju opisano je i provedeno testiranje kao i analiza rezultata. U zadnjem poglavlju dan je pregled cijelog rada s naglašenim opažanjima i zaključcima.

2. PREGLED PODRUČJA KLASIFIKACIJE PODATAKA

Klasifikacija je proces donošenja odluke o pripadnosti objekta predefiniranoj skupini ili klasi, koji se temelji na promatranim atributima toga objekta. Prema [1] klasifikacija je postupak donošenja odluke o pripadnosti objekta odnosno podatka ranije definiranoj skupini ili klasi, koji se temelji na promatranim atributima (osobinama) toga objekta. U području klasifikacije medicinskih podataka termin klasifikacija odnosi se i na tehniku kopanja (istraživanja) podataka (engl. *data mining*) koja se koristi za predviđanje pripadnosti grupi podataka.

Klasificiranje se sastoji od dvije faze to su faza treninga i faza testiranja. Što znači da svaka klasifikacija posjeduje dva skupa podataka, jedan je trening skup drugi je testni skup. Procesi i faze klasifikacije mogu se vidjeti na slici 2.1. iz koje se vidi da svaka podfaza trening faze direktno utječe na krajnji rezultat klasificiranja. Trening skup je skup jedinki za koje već ranije imamo definiran tip odnosno grupu kojoj pripada, kao i sva obilježja. Obilježja su osobine koje tu jedinku čine pripadnicom upravo toga skupa. Na temelju trening skupa najčešće se određuje kojoj klasi pripada neka nova jedinka te se na osnovi istoga podešavaju parametri klasifikatora kako bi se dobila bolja rješenja. Na trening skupu klasifikator, ako je u mogućnosti, uči. Testni skup je skup jedinki koje se trebaju klasificirati. Sam proces izrade klasifikatora može se prema [2] opisati u dva glavna koraka. Prvi korak je izrada modela prema kojem će se podaci klasificirati dok je drugi korak korištenje klasifikatora za određivanje pripadnosti podataka. Klasifikator se izrađuje pomoću odabranih algoritama te se trenira na trening skupu. Na osnovu rezultata testiranja koriste se algoritmi za procjenu točnosti te ako je točnost prihvatljiva klasifikator se može upotrijebiti na novim skupovima.



Slika 2.1. *Faze i procesi statističkoga prepoznavanja uzorka. [3]*

Pri klasifikaciji potrebno je dodatnu pažnju posvetiti pripremi, odnosno, pred-obradi podataka (engl. *preprocessing*). Stoga se treba provesti normalizacija atributa analizom relevantnosti atributa tako što da se odrediti koliki je utjecaj pojedinog atributa na rezultat klasificiranja. Ponekad se neki atributi mogu i izostaviti ukoliko imaju mali utjecaj na rezultat ili ako uopće ne

utječu na rezultat klasificiranja. U nekim slučajevima izostavljanje nekog od atributa može dati bolje rezultate. Za primjer se može uzeti pregled vida kod čovjeka. Boja šarenice oka je osobina čovjekovog oka ali nema utjecaj na kvalitetu ljudskog vida. Stoga će testiranje ljudi na vid koje ne uključuje boju očiju dati bolje rezultate. U većini slučajeva izostavljanjem takvih atributa skraćuje se vrijeme potrebno za klasificiranje. Tako se na neki način podaci očiste od "šuma" te se obrađuju oni atributi koji stvarno utječu na rezultat. Kako bi klasifikacija dala što bolje rezultate važno je prije same klasifikacije predvidjeti pogreške, odnosno, procijeniti točnosti klasifikatora kako bi se moglo utvrditi zadovoljava li trening skup zahtjeve testnog skupa. Glavno pitanje na koje treba odgovoriti je ima li u trening skupu dovoljno jedinki kako bi klasifikator mogao točno klasificirati jedinke testnog skupa. Veličina trening skupa koji zadovoljava testiranje ovisi o više parametara kao što su broj i raspon veličine atributa koji se koristi za klasificiranje, algoritmi klasificiranja, gustoći skupa, odnosno, kolika je razlika u klasama skupa. Potrebno je napomenuti da je najbitniji parametar broj jedinki testnog skupa, odnosno, odnos veličine testnog i trening skupa. Baza podataka može se definirati kao sveukupni skup jedinki tako da su u njoj sadržani i testni i trening skup. Uobičajen odnos bio bi da 80% sveukupnih podataka čini trening skup dok 20% podataka čini testni skup. Ovaj odnos nije uvijek odgovarajući pa tako ako je baza podataka velika omjer podataka sadržanih u trening i testnom skupu može biti i 70% trening skup, a 30% testni skup. Stoga se može zaključiti da se povećanjem broja jedinki baze podataka postotak trening skupa može smanjiti jer još uvijek u trening skupu ostaje dovoljan broj jedinki za klasifikaciju. Kod baza podataka s malim brojem sveukupnih jedinki postotak trening skupa raste, a postotak testnog skupa opada pa se tako za skup od 100 jedinki preporučuje omjer 90% trening a, 10% testni skup. Problem premalog trening skupa rješava bootstrap procedura predstavljena od Brad Effrona 1980 godine. Zasnovana je na engleskoj frazi: „*to pull oneself up by one's bootstraps*" što u slobodnom prijevodu znači „snađi se sa onim što imaš". Radi tako da od već postojećih uzoraka trening skupa napravi nove s tim da novom uzorku pridruži nasumične vrijednosti atributa već postojećih uzoraka te klase. Na taj način prividno povećava broj trening uzoraka. Nedostatak bootstrap procedure je u tom da ne moraju svi uzorci trening skupa biti zastupljeni, dok se neki mogu pojavljivati više puta. Bootstrap procedura koristi estimatore za predviđanje pogrešaka klasifikatora. Estimatori koje koristi bootstrap procedura opisani su u [4], a zasnivaju se na sklonosti učestalosti pogreške. Prema [5] najčešći estimatori su E0 estimator koji broji trening uzorke koji su pogrešno klasificirani i nisu se pojavili u bootstrap uzorku. Procjena se dobije kvocijentom sume pogrešno klasificiranih uzoraka svih bootstrap uzoraka te sume ukupnog broja trening uzoraka koji se ne pojavljuju u bootstrap uzorku. Drugi estimator koji se koristi je E632, a baziran je na asimptotičkoj vjerojatnosti da uzorak neće biti uključen u bootstrap test te uključuje i uzorke estimatora E0 koji su na odgovarajućoj udaljenosti od trening skupa. Postoji još metoda

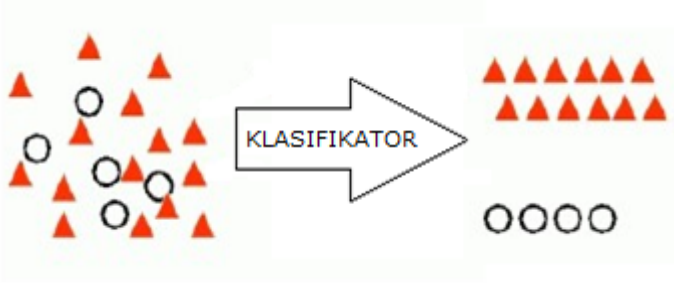
za estimaciju pogrešaka od kojih valja izdvojiti metodu unakrsne provjere (engl. *cross validation*) koja se temelji na podijeli trening skupa na podskupove nad kojima se vrši testiranje. Najčešći algoritam koji se primjenjuje je k-struka unakrsna provjera (engl. *k-folded cross validation*). Algoritam k-struke unakrsne provjere radi tako da trening skup podijeli na k podskupova gdje je k proizvoljan broj. Veličina broja k ovisi o veličini trening skupa, a prema [4] najčešće se uzima između 5 i 10. Nakon podjele na podskupove uzima se prvi podskup i na njemu se vrši testiranje kao da je on sada testni skup, a ostali podskupovi trening skup te se računa pogreška. Taj jedan korak naziva se preklop. Zatim se uzima drugi podskup te se na njemu vrši testiranje i tako k puta. Ukupna pogreška dobije se kao prosjek pogreške svih preklopa. Rad algoritma k-struke unakrsne provjere prikazan je na slici 2.2. Osim k-struke unakrsne provjere postoji i algoritam izostavi jednog unakrsne validacije (engl. *Leave one out cross validation*). Izostavi jednog algoritam uzima samo jedan uzorak iz cjelokupnog trening skupa i na njemu vrši testiranje pa je tako broj preklopa, odnosno, koraka jednak broju uzoraka u trening skupa. To ga čini pogodnim samo za male skupove.

Slika 2.2. *Algoritam k-struke unakrsne provjere. [4]*

2.1. Klasifikatori

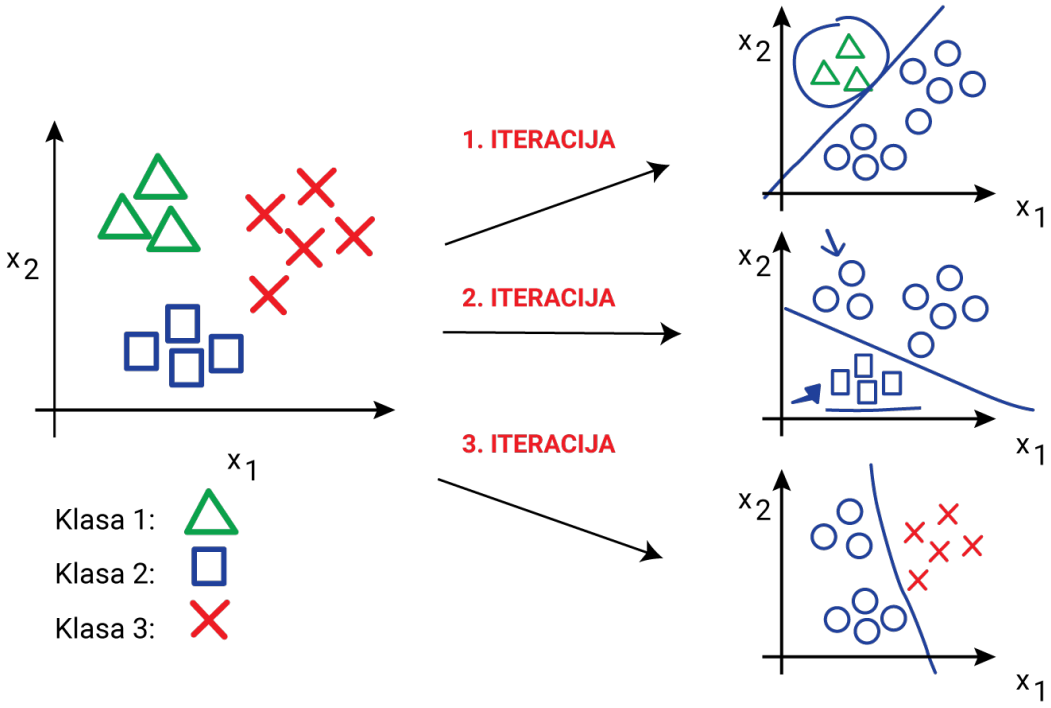
Klasifikator je algoritam koji se koristi za prepoznavanje uzorka i određivanje njegove pripadnosti jednom od skupova čiji je broj konačan i ranije određen. Prema [6] to je alat strojnog učenja koji se koristi nakon procesa učenja pri klasifikaciji novih podataka dodjeljujući im nabolje atribute po kojima se određuje pripadnost skupu. Klasifikatori se razlikuju po više obilježja po okruženju u kojem rade te načinu obrade podataka. Prema okruženu u kojem rade mogu se podijeliti na one koje rade u okruženju mutne logike (engl. *fuzzy logic*) i one koji se baziraju na Booleovoj logici (engl. *boolean logic*). Po načinu obrade podataka mogu se podijeliti na lijeno-učeće (engl. *lazy-*

learning) i željno-učeće (engl. *eager-learning*). Lijeno-učeći odgađaju indukciju generalizacije procesa sve dok klasifikacija svih uzoraka nije dovršena. Zahtijevaju manje vremena tokom trening faze za razliku od željno-učećih ali zahtijevaju više vremena u fazi klasifikacije. Željno-učeći odmah po početku rada stvaraju model klasifikatora za detekciju uzoraka. Klasifikatori se također razlikuju i po načinu učenja pa se tako razlikuju klasifikatori koji uče pod nadzorom (engl. *supervised learning*) i oni koji uče bez nadzora (engl. *unsupervised learning*). Neke vrste klasifikatora kao klasifikatori zasnovani na neuronskim mrežama mogu biti izvedeni za učenje s nadzorom ali i bez njega. Najvažnija podjela klasifikatora je ona po broju klasa koje određuju. Tako razlikujemo binarne i višeklasne klasifikatore. Binarni klasifikatori imaju samo dva moguća rezultata i daju odgovor pripada li podatak nekome skupu ili ne. Rade po principu ako uzorak nije tražene klase onda može biti samo druge klase. Binarni klasifikator prikazan je slikom 2.3.



Slika 2.3. *Binarni klasifikator.*

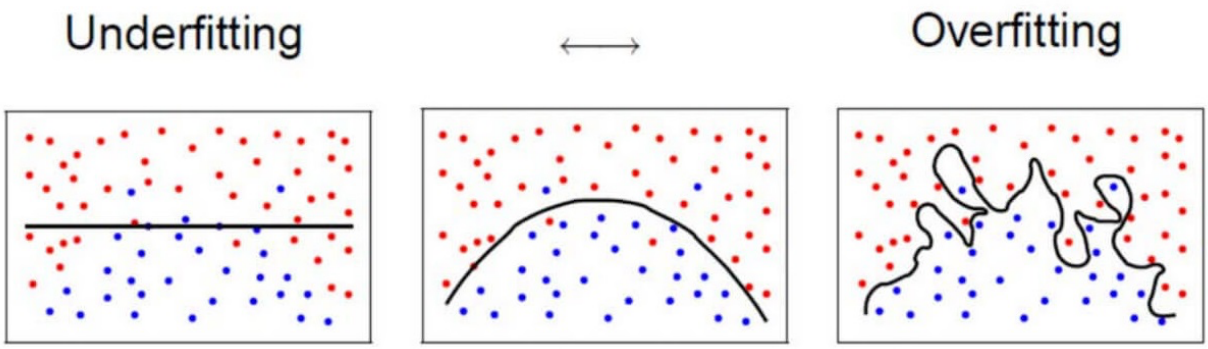
Višeklasni klasifikatori kao rješenje imaju tri ili više klasa. Najčešće rade na način da se prvo odrede jedinke koje pripadaju jednoj klasi, a zatim se u drugoj iteraciji određuju pripadnici druge klase i tako dalje. Princip rada takvih klasifikatora prikazan je na slici 2.4. Binarni klasifikatori mogli bi se primijeniti za klasificiranje višeklasnih podataka na sličan način tako da se prvo odrede pripadnici jedne klase, a zatim da se ta klasa makne iz baze podataka. Taj postupak se ponavlja sve dok ne ostanu samo dvije klase iako to nije praktično jer treba velika pred-obrada podataka te su greške klasificiranja veće nego kod višeklasnih klasifikatora. Važno je napomenuti da kod ovakvih klasifikatora klase moraju biti disjunktni skupovi jer se u protivnom radi o klasifikaciji jedan na više koja se vrlo rijetko koristi. Takav način klasifikacije naziva se klasifikacija s višestrukim oznakama (engl. *multilabel classification*).



Slika 2.4. *Višeklasni klasifikator.*

2.2. Vrednovanje klasifikatora

Kako bi se moglo odrediti zadovoljava li kasifikator zahtjeve klasifikacije, odnosno, hoće li rješenja koja se dobiju biti zadovoljavajuća potrebno je provesti razna mjerenja. Mjerenjem istih veličina moguće je usporediti dva klasifikatora koji se razlikuju po izvedbi. Prije toga potrebno je navesti dva glavna problema koja treba riješiti prilikom izrade klasifikatora. Prvi problem je problem nedovoljne podešenosti (engl. *underfitting*) koji nastaje kada klasifikator ne uspijeva aproksimirati podatke iz trening skupa te se kao rezultat javljaju pogreške pri klasifikaciji uzoraka testnog skupa. Uzrok tome može biti u greški prilikom pred-obrade baze podataka, malom broju uzoraka u trening skupu, malom broju atributa koji se koriste za klasifikaciju ili u odabiru krivog trening skupa. Drugi problem je suprotan tome, a to je problem prevelike podešenosti (engl. *overfitting*). On nastaje kada klasifikator savršeno nauči prepoznavanje uzoraka trening skupa, a ne može prepoznati uzorke koji se razlikuju od već naučenih što može nastati usred pretreniranosti. Optimalan klasifikator trebao bi biti negdje između nedovoljne podešenosti i prevelike podešenosti, kako je prikazano na slici 2.5.



Slika 2.5 *Glavni problemi klasifikacije. Slika nastala prema [2].*

Prva mjera vrednovanja klasifikatora je vrijeme potrebno kako bi se napravila klasifikacija. Vrijeme se počinje mjeriti od početka klasifikacije do prikaza rezultata što znači da vrijeme potrebno za pred-obradu podataka ne ulazi u vrijeme klasificiranja iako predstavlja važan čimbenik u odabiru algoritma klasifikatora. Vrijeme klasificiranja je važno prilikom klasificiranja u sustavima stvarnog vremena, odnosno, onim sustavima kod kojih rezultati trebaju stići prije vremena nužnog završetka (engl. *deadline*) jer u protivnom ti rezultati ne vrijede. Najvažnija veličina vrednovanja klasifikatora je njegova točnost (engl. *accuracy*) koja predstavlja omjer točno klasificiranih uzoraka i ukupnog broja uzoraka testnog skupa. Točnost je izražena u postocima, a izračunava se formulom (2-1).

$$Acc = \frac{BROJ\ TOČNO\ KLASIFICIRANIH\ UZORAKA\ SKUPA}{UKUPAN\ BROJ\ UZORAKA\ SKUPA} * 100\% \qquad (2\text{-}1)$$

Sljedeća veličina je standardna devijacija (engl. *standard deviation*) koja predstavlja mjeru raspršenosti u skupu. Ona govori koliko je prosječno odstupanje vrijednosti od aritmetičke sredine skupa, a označava se s grčkim slovom *sigma* (σ), te je dana izrazom (2-2) gdje je N - broj elemenata u skupu, μ - aritmetička sredina skupa, a x_i – i-ti član skupa (pa je i = 1,2,3,..., N).

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \qquad (2\text{-}2)$$

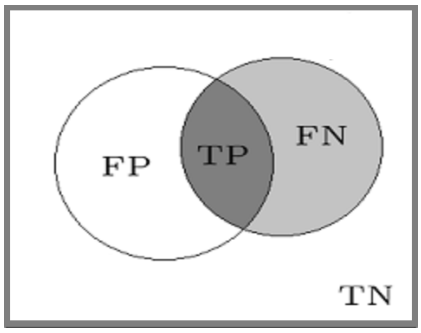
Ostale veličine koje se koriste prilikom vrednovanja klasifikatora, ako se vrši više testiranja, jesu maksimalna točnost, minimalna točnost, te razlika maksimalne i minimalne točnosti. Maksimalna točnost predstavlja najveću točnost koja se dobila prilikom testiranja, zatim minimalna točnost predstavlja najmanju točnost dobivenu prilikom testiranja, dok minimalna točnost predstavlja najmanju točnost dobivenu prilikom testiranja. Minimalna točnost je vrlo važna kod klasifikacije jer predstavlja najgori slučaj pa tako može direktno utjecati na odabir klasifikatora. Ako klasifikator ne ispunjava zahtjev za minimalnom točnosti on se ne može koristiti za klasifikaciju

odabranog skupa. Za vrednovanje klasifikatora uz navedene veličine koriste se još i metode za procjenu vrijednosti klasifikatora od kojih su najčešće *tablica zabune* i *ROC graf*.

2.2.1. Tablica zabune klasifikatora

Tablica zabune (engl. *confusion matrix*) omogućava detaljni pregled i vizualizaciju performansi klasifikatora, a zasniva se na informacijama o stvarnom stanju, odnosno, klasi uzoraka i stanju uzoraka predviđenim nastalim klasifikacijom algoritma koji se vrednuje. Vizualizacija rada tablice zabune binarnog klasifikatora prikazana je na slici 2.6. Kod tablice zabune ukupan skup testnih uzoraka podijeljen je na četiri cjeline pa se tako razlikuje broj točno pozitivnih uzoraka (engl. *true positive, TP*), broj točno klasificiranih negativnih uzoraka (engl. *true negative, TN*), broj netočno klasificiranih pozitivnih uzoraka (engl. *false negative, FP*) i broj netočno klasificiranih negativnih (engl. *false negative, FN*) uzoraka. Ukoliko se tablica zabune primjenjuje u višeklasnoj klasifikaciji treba se primijeniti za svaku klasu uzoraka tako da se pozitivno označe uzorci koji pripadaju toj klasi, a negativno oni koji pripadaju bilo kojoj drugoj klasi. Broj klasa odnosno tipova podataka određivati će broj iteracija.

		stvarno stanje	
		pozitivno	negativno
predviđeno stanje	pozitivno	TP	FP
	negativno	FN	TN



a) **b)**

Slika 2.6. *Tablica zabune binarnog klasifikatora prikazana **a)** tablično, **b)** dijagramima. [7]*

Uz točnost kao glavni pokazatelj tablica zabune prikazuje i druge parametre kvalitete klasifikatora. Tako je prema [7] preciznost (engl. *precision*) definirana kao udio točno klasificiranih uzoraka u skupu pozitivno klasificiranih primjera, a označena je s *P* te dana izrazom (2-3). Preciznost odgovara na pitanje: „*Od svih uzoraka koji su označeni kao pozitivni koji su stvarno pozitivni*“. Preciznost se u nekim literaturama [8, 9] naziva pozitivna predviđena vrijednost (engl. *positive predictive value, PPV*). Odziv (engl. *response*) oznake *R* prikazan formulom (2-4), a definiran je kao udio točno klasificiranih uzoraka u skupu svih pozitivnih uzorka. Odziv se u nekim literaturama [2] naziva osjetljivost (engl. *sensitivity*), ili prema [8, 9] stvarna pozitivna vrijednost (engl. *true positive rate, TPR*). Specifičnost (engl. *specificity*) ili prema [8, 9] stvarna negativna vrijednost (engl. *true negative rate, TNR)* definirana je kao udio točno klasificiranih uzoraka u

skupu svih negativnih uzoraka te prikazana formulom (2-5).

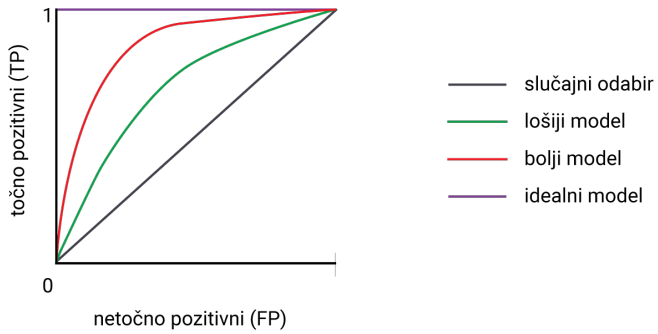
$$P = \frac{TP}{TP + FP} \tag{2-3}$$

$$R = \frac{TP}{TP + FN} \tag{2-4}$$

$$Spc = \frac{TN}{TN + FP} \tag{2-5}$$

2.2.2. Krivulja operativnih karakteristika

Krivulja operativnih karakteristika (engl. *Reciever Operating Characteristi, ROC*) je graf koji pokazuje odnos udjela netočno pozitivnih uzoraka u odnosu na udio točno pozitivnih uzoraka. Na apcisi grafa nalaze se netočno pozitivni dok na osi ordinata točno pozitivni udio uzoraka. S obzirom na to svaka je točka ROC krivulje definirana parom (FP, TP). Slikom 2.5. prikazana je ROC krivulja za četiri modela. Prema [9] površinom ispod ROC krivulje (engl. *Area Under the Curve, AUC*) može se mjeriti sposobnost klasifikatora da razlikuje uzorke koji pripadaju različitom klasama. Vrijednost AUC kreće se u intervalu od 0 do 1 gdje vrijednost 1 ima idealan model klasifikatora. Kako metoda slučajnog izbora ima AUC=0.5 tako sve modele klasifikatora koji imaju površinu ispod krivulje manju 0.5, odnosno, od metode slučajnog izbora treba odbaciti. Na osnovi površine ispod krivulje modeli klasifikatora mogu se ocijeniti prema [9] kao: izvrsni, ako im je AUC u intervalu od 0.90 do 1, dobri, ako im je AUC u intervalu od 0.80 do 0.90, srednji, ako im je AUC u intervalu od 0.70 do 0.80, slabi, ako im je AUC u intervalu od 0.60 do 0.70, te loši, ako im je AUC u intervalu od 0.50 do 0.60. Iz slike 2.7. može se zaključiti da što je krivulja okomitija to je veća i površina ispod krivulje pa je model klasifikatora bolji.



Slika 2.7. *ROC krivulja za razne modele. [9]*

2.3. Primjena klasifikatora

Osim pri klasifikaciji medicinskih podataka klasifikatori se naširoko primjenjuju u svim granama djelatnosti gdje je potrebno klasificiranje kao što su prepoznavanje uzoraka ili kopanje podataka. Koriste se prilikom analize financijskih podataka za dizajn i konstrukciju podatkovnih skladišta (engl. *datawarehouse*), višedimenzionalnu analizu podataka, predviđanje mogućnosti otplate kredita i ostalo. Klasifikatori se sve više koriste kod klasifikacije korisnika za ciljani marketing te otkrivanje tokova novca. Široku su primjenu našli i u telekomunikaciji prilikom višedimenzionalne analize podataka, analize ispravnosti podataka i detekcije prevare. Koriste se prilikom identifikacije neuobičajenih uzoraka i analiza velikih količina sekvencijalnih podataka. U prodajnoj industriji koriste se za ispitivanje tržišta (analiza podataka o prodaji, kupcima, proizvodima), preporuku proizvoda, savjetovanje korisnika, procjenu efektivnosti kampanja. Klasifikatori se također koriste i prilikom analize bioloških podataka u područjima analize genomskih baza podataka te analize sekvenci nukleotida. Veliku ulogu imaju u informacijskim sustavima prilikom analiza prometa podataka kao i kod korelacije vrsta podataka. U informacijskim sustavima koriste se još u svrhe kopanja podataka, izrade alata za vizualizaciju te pretraživanje Interneta. Široka je primjena i u prepoznavanju multimedijskih uzoraka kao što su prepoznavanje lica, prepoznavanje rukopisa, analiza glasa i prepoznavanje govora te prepoznavanje i pretraživanje video zapisa. Također, klasifikatori se koriste i za određivanje pripadnosti satelitskih snimaka određenoj geolokaciji. Poznata su i područja primjene u industrijskoj informatici i automatici, npr. pri kontroli proizvodnje elektroničkih elemenata, detekciji defektnih proizvoda na pokretnoj traci, interpretaciji rezultata senzora daljine, kao i predviđanje odziva neumreženih strojeva na temelju umreženih senzora i arhivskih podataka, kako je opisano u [3]. Klasifikatori su primjenu našli i u autoindustriji [10] gdje je opisano istraživanje povezano s povećanjem faktora ekonomičnosti benzinskih motora putem detekcije grešaka u intervalima okidanja svjećica.

3. POSTUPCI KLASIFICIRANJA MEDICINSKIH PODATAKA

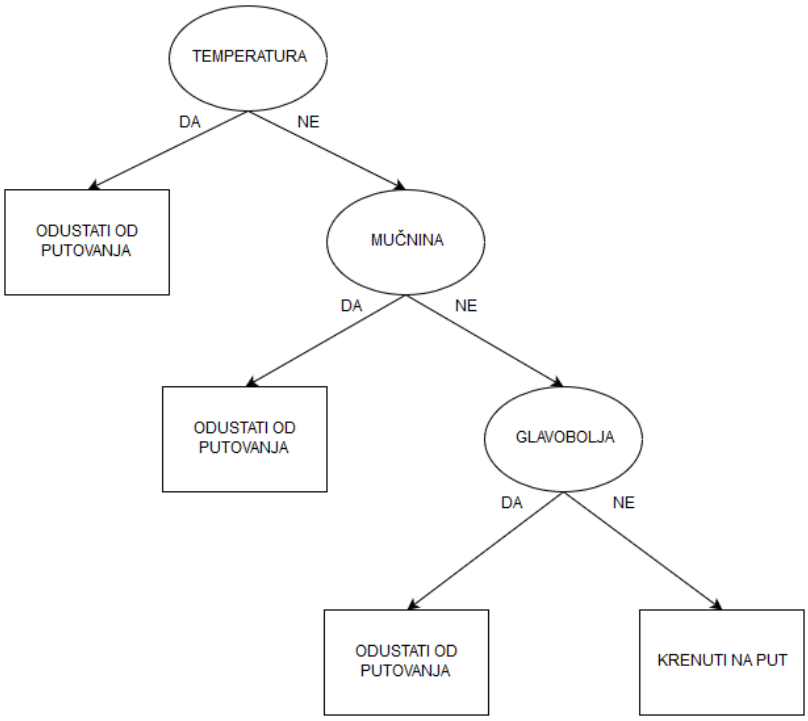
Klasifikatori u medicini mogu se koristiti za pretragu i obradu popisa pacijenata kao i promatranje i određivanje uzroka raznih bolesti. Klasifikatori se često koriste za brže i točnije donošenje odluka na temelju velikih količina naizgled međusobno neovisnih podataka. Tako je u [11] opisan projekt detekcija pacijenata starije životne dobi kojima bi rehabilitacijsko liječenje pomoglo. Skup podataka nad kojem je provedeno istraživanje je baza podataka zdravstvenih procjena pacijenata na kućnoj njezi s brojem uzoraka većim od 24000. Koristeći kNN (engl. *k-nearest neighbors, kNN*) klasifikator dokazano je poboljšanje brzine obrade podataka i povećanje točnosti predloženih terapija te se može reći da je došlo do smanjenja greške pri dijagnozi. Neki od drugih primjera primjene su i klasično prepoznavanje uzoraka, prepoznavanje slika i videa pri analizi rendgenskih i CT snimaka te analiza i klasifikacija podataka o elektromagnetskim valova kod elektro-kardiograma (engl. *ElectroCardioGram, ECG*). Dodatno, klasifikatori se koriste pri određivanju pokreta mišića koji mogu biti povezani s trzajnim ozljedama vrata i kralježnice. Osim navedenih primjera klasifikatori se primjenjuju u mnogo drugih istraživanja poput otkrivanja raka dojke, degenerativnih bolesti, istraživanju dijabetesa, Parkinsonove bolesti, raznih očnih bolesti te u mnogim drugim područjima.

U medicini se većinom koriste binarni klasifikatori (višeklasni klasifikatori koriste se rijetko npr. prilikom obrade popisa pacijenata) iz razloga jer postoje samo dva slučaja, a to su postojanje bolesti, odnosno, defekta ili ne postojanje bolesti. Modeli klasifikatora koji se koriste ovise isključivo o parametrima skupova. Najčešći klasifikatori koje se koriste u svrhu klasificiranja medicinskih podatak jesu klasifikatori zasnovani na udaljenostima, klasifikatori temeljeni na neuronskim mrežama, stablima odlučivanja, klasifikatori temeljeni na vektorima potpore te Bayesov klasifikator.

3.1. Stabla odlučivanja

Stablo odluke (engl. *decesion tree*) temelji se na strukturi stabla, a sastoji se od izvornog ili početnog čvora (engl. *root node*), grana i pod-čvorova. Postoje dvije vrste čvorova, krajnji čvor (engl. *leaf node*) i čvor odluke (engl. *decision node*). Krajnji čvor je čvor kojim završava grananje stabla te on određuje klasu kojoj uzorak pripada. Čvor odluke predstavlja određeni uvjet kojeg uzorak treba zadovoljiti. Uvjet je u obliku vrijednosti određenog atributa. U čvoru odluke na osnovi vrijednosti određenog atributa odlučuje se sljedeći korak, odnosno, na koju će se granu ići. Grane povezuju čvorove, a predstavljaju rezultate uvjeta čvorova. Prema [3] stablo odluke opisano je tako da svaki pod-čvor predstavlja test, a svaka grana rezultat toga testa. Nadalje, sljedeći pod-čvor predstavlja sljedeći test, a grana rezultat sve do krajnjega čvora koji predstavlja klasu odnosno

odluku. Stablo odluke ne zahtijeva dubinsko poznavanje problematike, a koraci učenja i klasifikacije su u pravilu brzi i jednostavni. Prednost stabla odlučivanja u odnosu na ostale klasifikatore je i u činjenici da nude model koji se lako može interpretirati u razumljivom obliku običnim jezikom putem pravila. Stoga se lako mogu povezivati s raznim bazama podataka (kao što su SQL baze). Kod stabla odlučivanja nema povratka na prethodno stanje pa nema ponavljanja testova već se algoritam rekurzivno izvršava odozgo prema dolje (engl. *top-down*). Binarna stabla odlučivanja karakterizira obilježje da svaki čvor odluke može imati samo dva pod-čvora. Slikom 3.1. prikazano je binarno stablo odlučivanja za primjer odluke treba li krenuti na put u odnosu na zdravstveno stanje osobe.



Slika 3.1. *Binarno stablo odlučivanja.*

Iz primjera sa slike 3.1. vidi se da postoje dvije klase, a to je prisutnost bolesti koja rezultira odustajanjem od putovanja i klasa bez prisutnosti bolesti koja rezultira odlaskom na put. Nadalje, vidi se da postoje tri atributa koja su u primjeru prikazana prisustvima bolesti (povišena temperatura, mučnina i glavobolja). Važan faktor pri izradi algoritma stabla odlučivanja je redoslijed ispitivanja atributa. Prema [3] cilj je postaviti najmanje stablo odnosno stablo koje ima najmanju težinu. Odabir atributa koji će se prvi ispitati određuje se preko entropije (engl. *entropy*) skupa. Entropija skupa definira se kao mjera homogenosti skupa, označava se sa (H), a dana je izrazom (3-1) gdje je S skup svih uzoraka skupa, N broj uzoraka skupa i p_i vjerojatnost događaja.

$$H(S) = -\sum_{i=1}^{N} p_i log_2 p_i \tag{3-1}$$

Vrijednost entropije je u intervalu od 0 do 1. Ako svi uzorci nekog skupa pripadaju istoj klasi vrijednost entropije sustava biti će jednaka nuli dok za skup kod kojeg je jednak broj uzoraka svih klasa vrijednost entropije biti će jednaka 1. Preko entropije se izračunava informacijska dobit (engl. *gain*) pojedinog atributa na skupu. Informacijska dobit (*Gain(A,* S*)*) definirana je kao količina informacija koja se dobije poznavanjem određenog atributa (*A*) na skupu uzoraka (*S*). Prema [12] informacijska dobit predstavlja razliku entropije prije grananja i entropije poslije grananja preko atributa *A*. Atribut s najvećim informacijskom dobiti trebao bi biti početni čvor. Postupak se ponavlja prilikom svakog slijedećeg grananja. Informacijska dobit dana je izrazom (3-2), gdje je: *H(S)* entropija skupa, $|S_i|$ broj uzoraka sa i-tom vrijednošću atributa, $|S|$ ukupan broj uzoraka u skupu S, *v* skup vrijednosti atributa A, $H(S_i)$ entropija podskupa uzoraka sa atributom A i *H(A, S)* entropija atributa A.

$$Gain(A,S) = H(S) - \sum_{i=1}^{v} \frac{|S_i|}{|S|} * H(S_j) = H(S) - H(A,S) \tag{3-2}$$

3.2. Bayesov klasifikator

Bayesov klasifikator je statistički klasifikator koji se temelji na Bayesovom teoremu. Prema [6] zasnovan je na temelju međuodnosa atributa. Ime je dobio po engleskom statističaru i filozofu Thomasu Bayesu (1701-1761) koji je proučavao međusobnu povezanost događaja. Ako odgađaj *A* uzrokuje događaj *B* onda događaj *A* implicira događaj *B*. To pravilo može se primijeniti i na klasifikaciju na način da se atributi postave kao događaji koji impliciraju klasu. Kako je prema klasičnoj definiciji vjerojatnost (engl. *probability, P*) broj pojave određenog događaja u odnosu na skup svih događaja vrijedi izraz (3-3). Ako se s n_A označi broj događaja *A*, sa n_B broj događaja *B* i s *n* ukupan broj događaja koji je konačan te se primjeni teorem uvjetne vjerojatnosti dobije se izraz (3-4) za vjerojatnost događaja *B* ako se dogodio događaj *A*.

$$P(A) = \frac{n_A}{n}, P(B) = \frac{n_B}{n}, P(AB) = P(A) + P(B) \rightarrow P(AB) = \frac{n_{AB}}{n} \tag{3-3}$$

$$P(B|A) = \frac{n_{BA}}{n_A} = \frac{\frac{n_{BA}}{n}}{\frac{n_A}{n}} = \frac{P(AB)}{P(A)}, \qquad P(A) > 0 \tag{3-4}$$

Ako se definiraju međusobno neovisni događaji kao hipoteze (engl. *hypothesis*) na konačnom skupu i označe sa $H_1, H_2,\ldots,H_n$ tada za svaki događaj E vrijede izrazi (3-5) i (3-6). Ako se događaj E definira kao dokaz (engl. *evidence*) u nekim literaturama [2] nazvan opažaj vezan za hipotezu H, izraz (3-7) konačni je izraz Bayesovog teorema.

$$P(E) = \sum_{i=1}^{n} P(H_i) * P(E|H_i) \tag{3-5}$$

$$P(H_i|E) = P(H_i) * P(E|H_i) = P(E) * P(H_i|E), \qquad i = 1,2,\ldots,n \tag{3-6}$$

$$P(H_i|E) = \frac{P(H_i) * P(E|H_i)}{P(E)} \tag{3-7}$$

Izraz *P(H)* naziva se *a priori* vjerojatnost hipoteze *H*, dok izraz *P(H|E)* je *a posteriori* vjerojatnost hipoteze *H* i predstavlja uvjetnu vjerojatnost hipoteze *H* ako se dogodio dokaz *E*. Izraz *P(E|H)* naziva se vjerodostojnost (engl. *likelihood*) ili izglednost prema [2], a predstavlja uvjetnu vjerojatnost dokaza *E* ako se dogodila hipoteza *H*.

Prema Bayesovom teoremu hipoteza (događaj) s najvećom uvjetnom vjerojatnosti će se i dogoditi. Po tom principu, klasa kojoj pripada uzorak je klasa s najvećom uvjetnom vjerojatnosti što znači da se za svaku klasu treba računati uvjetna vjerojatnost po svakom atributu. Putem teorema standardne normalne razdiobe mogu se izračunati sve uvjetne vjerojatnosti preko izraza (3-8). Prvo je potrebno izračunati pokazatelje samog trening skupa na osnovu kojeg će se vršiti klasifikacija. Tako je potrebno izračunati prosječnu vrijednost i varijancu svakog atributa za svaku klasu te vjerojatnost pojavljivanja svake klase u trening skupu. Varijanca (engl. *variance*) je mjera disperzije veličine, a jednaka je kvadratu standardne devijacije dane izrazom (2-2). Vjerojatnost pojavljivanja klase u skupu računa se preko izraza (3-3), a označena je s *P*. Ako je klasa označena sa *C*, a atribut s *A*, slijedi izraz (3-8).

$$p(A|C) = \frac{1}{\sqrt{2\pi\sigma^2}} * e^{-\frac{(A-\mu_A)^2}{2\sigma^2}} \tag{3-8}$$

Gdje je μ_A prosječna vrijednost atributa za klasu. Kako je dokaz ukupan umnožak vjerojatnosti i uvjetnih vjerojatnosti svake klase jer se nalaze u istom skupu on je konstanata za taj trening skup. Stoga, vjerojatnost klase da će uzorak biti određene klase ovisi o umnošku vjerojatnosti te klase i umnošku uvjetnih vjerojatnosti svih atributa uzorka. Ta vrijednost naziva se *posterior numerator*, a prikazana je izrazom (3-9) gdje *n* predstavlja broj atributa *A* klase *C*. Za uzorak kojem se želi odrediti klasa računa se posterior numerator za svaku klasu, a uzorak je klase s najvećim posterior numeratorom.

$$posterior\ numerator(C) = P(C) * \prod_{i=1}^{n} p(A_i|C) \tag{3-9}$$

Bayesov klasifikator karakterizira jednostavnost izvedbe i velika brzina klasificiranja. Nije memorijski zahtjevan, daje dobre rezultate čak i kada trening skup sadrži manju količinu uzoraka. Rezultati mu slabe kada se koristi na bazama podataka kod kojih je broj atributa mali, a bolje radi kada se uzima više atributa u obzir. Ako trening skup podataka sadrži dosta uzoraka koji odskaču od prosjeka (izuzeci) rezultati klasificiranja Bayesovim klasifikatorom znatno opadaju.

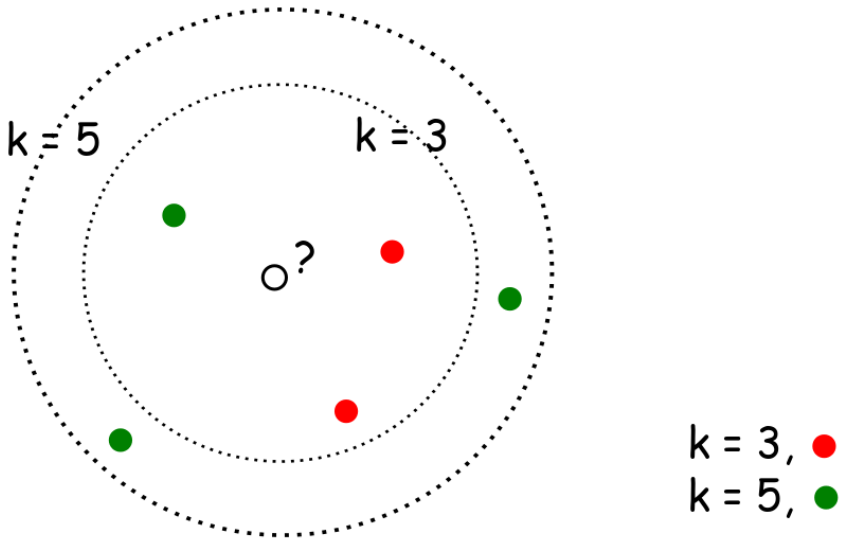
3.3. Klasifikator k - najbližih susjeda

Klasifikator k - najbližih susjeda najraširenij je klasifikator baziran na udaljenostima. Zasnovan je na ideji da se testni uzorak klasificira tako da se pogledaju njemu najbliži poznati primjeri te na osnovi zastupljenosti njihovih klasa odredi klasa tog testnog uzorka. Može se reći da radi po principu „s kim si takav si“. Stoga su uzorci predstavljeni kao točke u *n*-dimenzionalnom prostoru R^n gdje je broj dimenzija *n* određen brojem atributa korištenih za klasifikaciju. Broj *k* je hiperparametar koji određuje koliko je veliko susjedstvo, odnosno, koliko najbližih susjeda utječe na određivanje klase testnog uzorka. Općeniti način rada kNN klasifikatora opisan je slijedećim pseudokodom:

Za svaki uzorak testnog skupa čini:

- *Izračunaj udaljenost testnog uzorka od svih primjera trening skupa;*
- *Sortiraj listu udaljenosti od najmanje udaljenosti prema najvećoj;*
- *Odaberi k uzoraka trening skupa s najmanjim udaljenostima i stavi ih u grupu susjeda;*
- *Prebroj broj jedinki svake klase u grupi susjeda;*
- *Klasificiraj testni uzorak kao klasu koja ima najveći broj jedinki u susjedstvu;*

Takav princip odlučivanja naziva se odlučivanje većinskim proglašavanjem (engl. *majority vote*). Utjecaj parametra k na rezultat klasificiranja vidi se na slici 3.2. Iz načina rada može se zaključiti da parametar k u binarnoj klasifikaciji treba biti neparan broj kako ne bi došlo do toga da dvije klase različitih tipova imaju podjednak broj jedinki uzoraka u susjedstvu, pa odlučivanje o tipu klase ne bi bilo moguće. Povećanje parametra k u većini slučajeva rješava problem overfittinga, no treba imati na umu da se povećava i vrijeme trajanja klasifikacije. Prema [13, 14] najčešće je k=5 ili k=7. U slučaju ako je k=1 govori se o posebnom tipu klasifikatora temeljenom na udaljenosti, a to je algoritam najbližeg susjeda (engl. *nearest neighbor, 1NN*).



Slika 3.2. *Princip rada kNN klasifikatora.[13]*

Na primjeru prikazanom slikom 3.2. vidljive se dvije klase, crvena i zelena, te je potrebno odrediti klasu uzorka bez boje primjenom kNN klasifikatora. Vidi se utjecaj veličine parametra k na način da će susjedstvo testnog uzorka sadržavati veći broj crvenih jedinki trening skupa za k=3, omjer je 2 naprema 1 u korist crvenih. Povećanjem parametra k na 5 (k=5) broj crvenih susjeda ostaje isti dok broj zelenih raste te se omjer mijenja u 3 naprema 2 u korist zelenih što rezultira i promjenom u određivanju klase testnog uzorka. Iz iste slike može se vidjeti da povećanjem parametra k u susjedstvo testnog uzorka ulaze samo nove jedinke trening uzorka. Susjedi koji su bili i pri manjem broju k ostaju u susjedstvu jer su oni „bliži“ od novih susjeda. Odabir veličine parametra k ovisi o osobinama skupa koji testiramo.

Osim za klasificiranje podataka kNN klasifikator može se koristiti za nadopunjavanje vrijednosti atributa u uzorcima trening skupa kojima nedostaj neki atribut. To se radi na način da se izračuna prosječna vrijednost tog atributa u susjedstvu svih jedinki trening skupa koji dijele klasu s uzorkom kojem nedostaje atribut te se ta vrijednost dodijeli traženom atributu uzorka kojem nedostaje atribut. Prilikom računanja udaljenosti traženi atribut se izostavlja.

Kako je kNN klasifikator vrlo raširen i pogodan za klasifikaciju raznih baza podataka razvijeni su načini njegovog unapređenja. Razni načini unapređenja kNN klasifikatora opisani su u [13, 14]. Jedan od najčešćih načina unapređenja kNN klasifikatora je uvođenje težinskog faktora gdje se onda takav klasifikator naziva težinski klasifikator k najbližih susjeda (engl. *weigted k nearest neighbors, WkNN*). Težinski kNN klasifikator vodi se principom da ne utječe svaki susjed jednako na testni uzorak već onaj koji je bliže ima i veći utjecaj. Rezultat uvođenja težinskog faktora u binarnoj klasifikaciji je da parametar k više ne treba biti neparan broj. Odabir kako će se smanjivati utjecaj ovisi o prirodi baze podataka, a najčešće se odabire da utjecaj opada s udaljenosti ili s

kvadratom udaljenosti. Postoji još i način da utjecaj opada s brojem bližih susjeda. Drugi način unaprjeđenja kNN klasifikatora je implementacija stabla odluke u kNN klasifikator po principu da je svaki čvor *n*-dimenzionalna točka. Na taj način ubrzava se rad kNN klasifikatora jer se algoritmom stabla vrlo efektivno pronalaze susjedi. Ostali načini poboljšanja kNN klasifikatora jesu primjena modela radijalne bazne funkcije, model reduciranog kNN klasifikatora i lokalizirani kNN klasifikator. Svim tim unaprjeđenjima pokušava se riješiti glavni nedostatak kNN klasifikatora, a to je induktivna pristranost, odnosno, pretpostavka da je klasifikacija uzorka jednaka klasifikaciji primjera u blizini uzoraka te skraćivanje vremena potrebnog za klasifikaciju.

3.3.1 Mjere udaljenosti

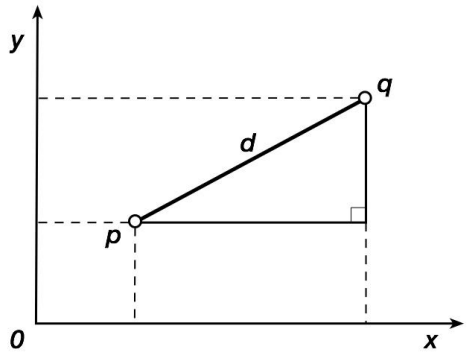
Odabir mjera udaljenosti, odnosno, načina na koji se određuje koliko su susjedi udaljeni od testnog uzorka može imati velik utjecaj i na to koji se uzorci ubrajaju u susjede. Odabir mjera udaljenosti ovisi o osobinama baze podataka nad kojom se vrši klasifikacija i o ciljevima koji se klasifikacijom žele postići. Najčešće mjere udaljenosti su euklidska, kvadratna euklidska, Manhattan, Chebyshevljeva, kosinusna, canberra, Bray-Curtisova udaljenost i ostale.

Euklidska mjera udaljenosti (engl. *euclidean distance*) predstavlja najkraću udaljenost između dvije točke u jednom prostoru. Kako su uzorci skupa u kod kNN klasifikatora predstavljeni kao točke u prostoru, a broj dimenzija prostora predstavljen je brojem atributa dolazi se do izraza (3-10) koji predstavlja euklidsku udaljenost dva uzorka.

$$d(p,q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \qquad (3\text{-}10)$$

Gdje je *d(p,q)* udaljenost između dvije točke p i q koje su zadane svojim atributima $p=(p_1, p_2,..., p_n)$ odnosno $q=(q_1, q_2,..., q_n)$. Euklidsku odaljenost u dvodimenzionalnom prostoru možemo predočiti Pitagorinim poučkom prikazanim na slici 3.3. **Kvadratna euklidska mjera udaljenosti** (engl. *squared euclidean distance*) temelji se na euklidskoj udaljenosti, a računa se kao kvadrat euklidske udaljenosti prema izrazu (3-11). Iako nije metrička udaljenost, kvadratna euklidska udaljenost koristi se kada je potrebno dati veću važnost udaljenosti, odnosno, njenom primjenom sam faktor udaljenosti ima veću težinu.

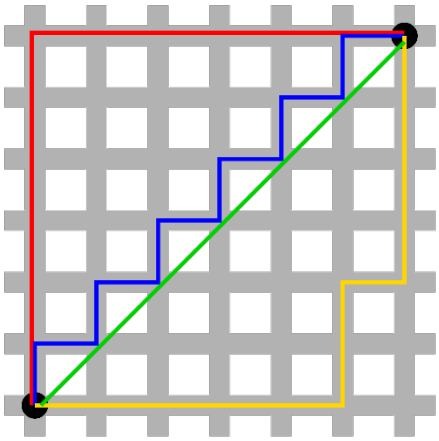
$$d(p,q) = (q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2 = \sum_{i=1}^{n} (q_i - p_i)^2 \qquad (3\text{-}11)$$



Slika 3.3. *Euklidska udaljenost u dvodimenzionalnom prostoru.*

Manhattan udaljenost (engl. *Manhattan distance*) ime je dobila po izgledu većine ulica sastavljenih od blokova kuća na Manhattanu. Prvobitno je zamišljena za mjerenje ruta taxija pa se naziva još i taxi udaljenost (engl. *taxicab distance*) ili udaljenost gradskih blokova (engl. *city block distance*). Cilj je udaljenost prezentirati u realnome svijetu pa taxi ne može proći kroz blokove već se oni trebaju zaobići. Rezultat toga je da postoji više putanja s jednakom udaljenosti, te udaljenost mjerena na ovakav način biti će veća od euklidske udaljenosti istih točaka. Slikom 3.4. prikazane su razne putanje jednake duljine Manhattan udaljenosti između dvije točke dvodimenzionalnog prostora dok je zelenom bojom na slici označena euklidska udaljenost istih točaka. Manhattan udaljenost izražena je formulom (3-12).

$$d(p,q) = ||p - q|| = \sum_{i=1}^{n} |p - q| \tag{3-12}$$

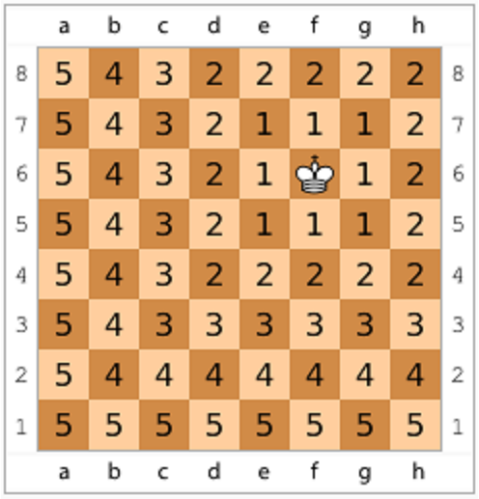


Slika 3.4. *Manhattan udaljenost. Preuzeto iz [15]*

Chebyshevljeva mjera udaljenosti (engl. *Chebyshev distance*) zasniva se na vektorskoj udaljenosti. Slično kao i kod Manhattan udaljenosti prostor je podijeljen u blokove ali se putanja kreće po blokovima. Najlakše se može opisati kao pomicanje kralja u šahu kako je prikazano na

slici 3.5. stoga se još naziva i udaljenost prema šahovskoj ploči (engl. *chessboard distance*). Udaljenost se prikazuje kao broj poteza (koraka) koje treba napraviti kralj da dođe do određenog polja (točke). Udaljenost se računa kao maksimalna apsolutna razlika koordinata točki pri čemu koordinate predstavljaju atribute. Chebyshevljeva udaljenost dana je izrazom (3-13).

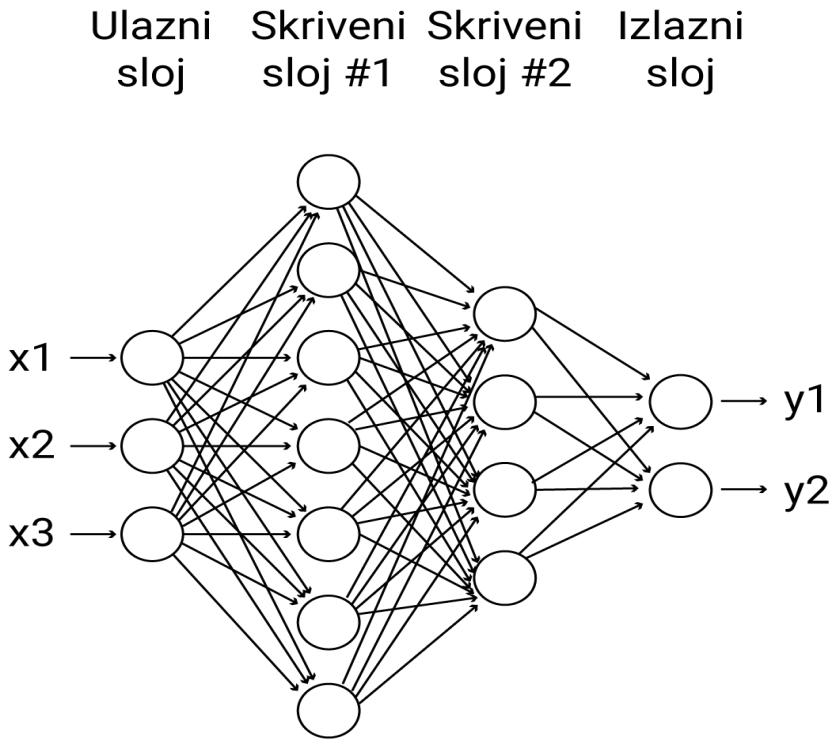
$$d(p,q) = \max(|p_i - q_i|), \qquad i = 1,2,\dots,n \tag{3-13}$$



Slika 3.5. *Chebyshevljeva udaljenost. [15]*

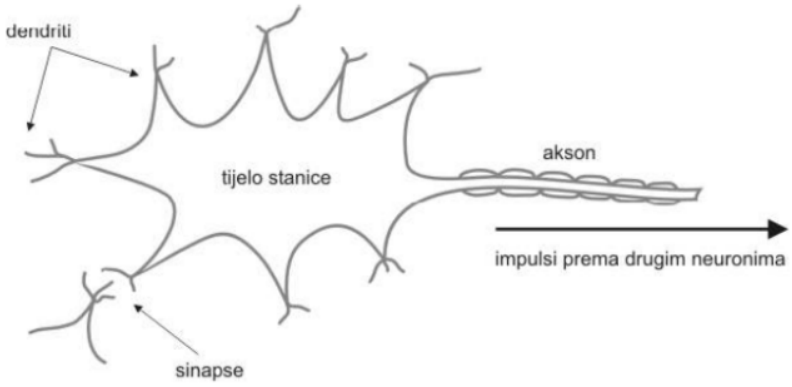
3.4. Klasifikatori temeljeni na neuronskim mrežama

Umjetna neuronska mreža (engl. *artificial neural network, ANN*) je skup međusobno povezanih umjetnih neurona (čvorova) koji oponašaju rad ljudskog mozga. Stoga i građa umjetne neuronske mreža oponaša građu ljudskog mozga. Neuronska mreža sastoji se od tri vrste slojeva, a to su ulazni sloj, izlazni sloj i skriveni slojevi. Ulazni sloj prima ulazne podatke u slučaju klasifikacije, a to bi bili atributi uzoraka te ih predaje prvom skrivenom sloju koji podatke obrađuje i predaju drugom sloju. Postupak se nastavlja sve do izlaznog sloja. Izlazni sloj odlučuje o tipu podataka pa je tako broj izlaznih čvorova određen brojem tipova klasa. Na slici 3.6. prikazana je umjetna neuronska mreža za binarnu klasifikaciju gdje se izlazni sloj sastoji od dva izlazna čvora y_1 i y_2. Broj slojeva skrivenih čvorova mreže je proizvoljan kao i broj čvorova u svakom skrivenom sloju, a ovisi o osobinama skupa nad kojim će se vršiti testiranje. Prema [16] uzorak koji se klasificira biti će one klase čiji izlaz ima veću vrijednost.

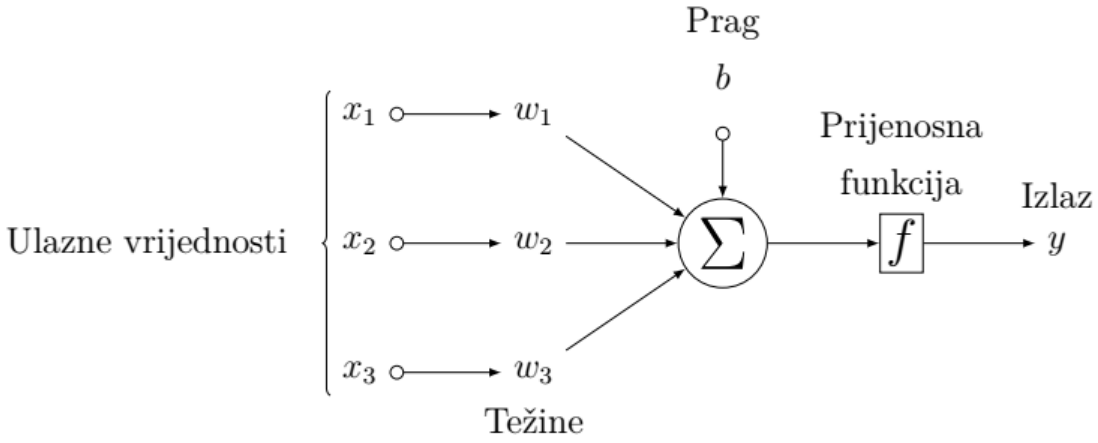


Slika 3.6. *Umjetna neuronska mreža.*

Kako građa umjetne neuronske mreže oponaša građu ljudskog mozga tako i građa samih umjetnih neurona nalikuje građi neurona ljudskog mozga. Sličnost se vidi usporedbom slika 3.7. i 3.8. Prirodni neuron (živčana stanica) prikazan je slikom 3.7., a sastoji se od dendrita koji prenose podražaj u stanicu, tijela stanice i aksona koji prenosi impuls na drugu stanicu. Umjetni neuron prikazan na slici 3.8. također ima dijelove zadužene za primanje ulaznih podataka, središnji dio te funkciju za prijenos podataka.



Slika 3.7. *Građa prirodnog neurona. Preuzeto iz [2].*



Slika 3.8. *Građa umjetnog neurona. Preuzeto iz[16].*

Neuroni su povezani putevima različitih težina, a izlaz umjetnog neurona dana je izrazom (3-14). Gdje je *f* prijenosna funkcija, *w* težina (engl. *weight*), a *b* prag (engl. *bias*). Prijenosna funkcija može biti linearna ili nelinearna. Nelinearne prijenosne funkcije potrebne su za rad s nelinearnim podacima koji se često susreću u klasifikaciji. Najčešći oblici prijenosne funkcije su funkcija skoka, linearna funkcija, logistička funkcija u nekim literaturama [16] nazvana sigmoidalna te funkcija hiperboličnog tangensa.

Samo učenje odnosno treniranje umjetnih neuronskih mreža može biti nenadzirano (engl. *unsupervised learning*) i nadzirano (engl. *supervised learning*). Nenadzirano učenje odvija se bez poznavanja vrijednosti izlaza, dok je nadzirano učenje, učenje kod kojeg mreža uči pomoću učitelja na način da joj se da niz ulaznih podataka kojima su poznate klase i atributi na osnovu kojih algoritam treba odrediti težine veza kako bi rezultati bili što točniji. Pri učenju treba paziti da ne dođe do pretreniranosti tako što se ograniči broj ponavljanja. Načelo rada umjetne neuronske mreže je vođenje prema naprijed (engl. *feedforward*), a očituje se u tome da je izlaz iz jednog neurona ulaz u drugi neuron slijedećeg sloja. Postoje i neuronske mreže koji imaju povratnu vezu. Najčešći algoritam za procjenu pogreške umjetnih neuronskih mreža je algoritam unazadne propagacije pogreške (engl. *backpropagation algorithm*). Algoritam unazadne propagacije predstavlja izlaz neuronske mreža kao funkciju svih ulaza i svih težina. Prema [16] varijable su težine puteva, a izlaz iz funkcije predstavlja pogrešku. Prvo se dobije odziv mreže za određeni ulaz, a potom se računaju greške za izlazni sloj te se na taj način određuje utjecaj prethodnog sloja na grešku. Zatim se ažuriraju težine da bi greška bila što manja te se prelazi na slijedeći sloj. Algoritam kreće od izlaznog sloja prema nazad, odakle i naziv unazadna propagacija.

Za rad s višedimenzionalnim podacima poput slika, koriste se konvolucijske umjetne neuronske mreže. Konvolucijske neuronske mreže rade po istom principu kao i neuronske mreže za skalarne vrijednosti ali imaju dva dodatna sloja koja se nalaze između ulaznog i prvog skrivenog sloja. To

su sloj konvolucije (engl. *convolutional layer*) i sloj sažimanja (engl. *pooling layer*). Ti slojevi mogu se sastojati od više podslojeva. Zadaća konvolucijskog sloja je izvlačenje parametara bitnih za klasifikaciju dok je sloj sažimanja zadužen za reduciranje parametara u jednu vrijednost.

3.5. Klasifikatori zasnovani na stroju s potpornim vektorima

Klasifikatori zasnovani na stroju s potpornim vektorima (engl. *support vector machine, SVM*) binarni su klasfikatori dizajnirani tako da za zadani skup podataka pronađu hiperravninu (engl. *hyperplane*) koja će podijeliti skup na dvije klase. Tako se pripadnost klasi određuje s obzirom na položaj uzorka u odnosu na hiperravninu te može iznositi 1 ili -1. Osnovni uvjet je da su klase linearno razdvojive. Svaki uzorak predstavljen je kao točka *n*-dimenzionalnog prostora, a broj dimenzija jednak je broju atributa. Za trodimenzionalan prostor hiperravnina predstavlja ravninu, a za dvodimenzionalan prostor hiperravnina je predstavljena pravcem zadanim s izrazom (3-14), gdje je *x* koordinata uzorka, *w* normala hiperravnine, a *b* udaljenost hiperravnine od ishodišta koordinatnog sustava. Na osnovu toga klasu *y* definiramo izrazima (3-14) i (3-15), a konačni oblik (3-16) nastao je spajanjem oblika (3-14) i (3-15).

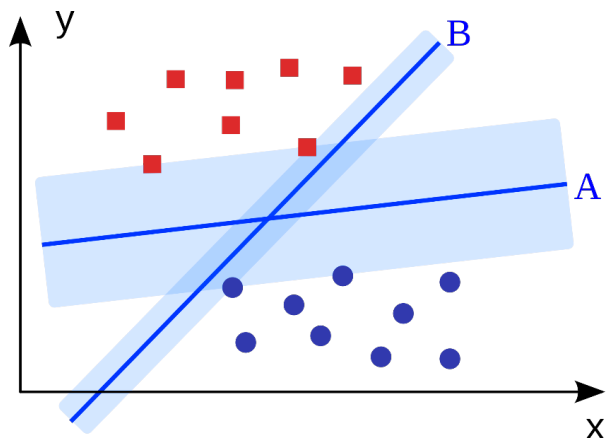
$$w * x + b = 0 \tag{3-13}$$

$$x_i * w + b \geq 1, \qquad za\ y_i = 1 \tag{3-14}$$

$$x_i * w + b \leq -1, \qquad za\ y_i = -1 \tag{3-15}$$

$$y_i(x_i * w + b) - 1 \geq 0 \tag{3-16}$$

Slikom 3.9. prikazana su pravci koji predstavljaju hiperravnine za neki zadani skup. Iz slike 3.9. se vidi da postoji više pravaca koji su u mogućnosti podijeliti skup na dvije klase. Na slici 3.9. to su pravci *A* i *B*. Oba pravca dijele trening skup no pitanje je koji će pravac biti bolje rješenje. Odgovor je pravac *A* jer on ima veću marginu, odnosno, veći razmak od najbližih uzoraka svake klase.



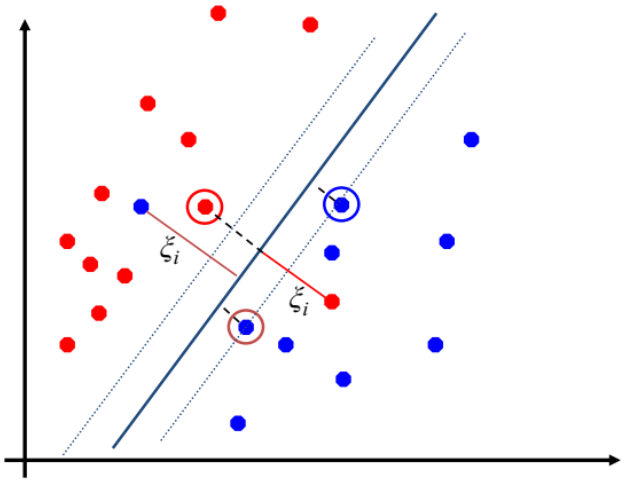
Slika 3.9. *Vektori potpore i njihove hiperravnine. [17].*

Podrazumijeva se da su margine udaljenost jedne i druge strane pravca (hiperravnine) jednake. Vektori koji sadrže sve uzorke pojedine klase koji su najmanje udaljeni od hiperravnine nazivaju se potporni vektori (engl. *support vectors*). Kako bi se maksimizirala margina, odnosno, udaljenost potpornih vektora od pravca (hiperravnine) treba se zadovoljiti izraz (3-17).

$$\min||w||, \quad takav\ da\ vrijedi\ y_i(\ x_i * w + b) - 1 \geq 0 \tag{3-17}$$

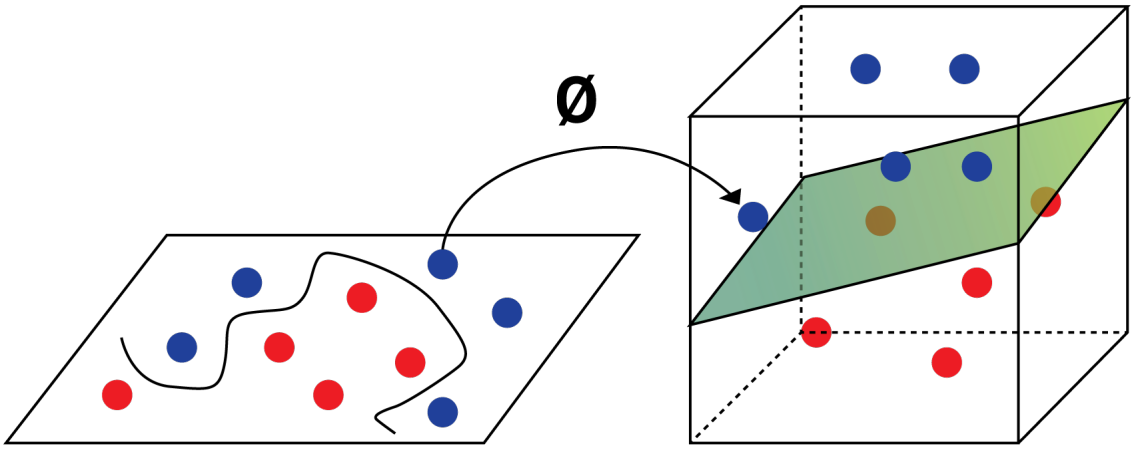
Ukoliko klase nisu linearno razdvojive može se primijeniti metoda mekih granica ili metoda trika jezgri. Metoda mekih granica (engl. *soft margin method*) prema [17] dozvoljava uvođenje nenegativne vrijednosti ξ. Tako se dozvoljava pogreška pri klasifikaciji, a izraz poprima oblik pod (3-18). Osim same veličine margine potrebno je pronaći i faktor pogreške.

$$(\ x_i * w + b) - 1 + \xi_i \geq 0 \tag{3-18}$$



Slika 3.10. *Metoda mekih granica. Preuzeto iz [17].*

Jezgreni trik (engl. *kernel trick*) je metoda vođena idejom da se originalni prostor zamijeni nekim drugim prostorom u kojem će klase biti linearno odvojive. Prema [17] potrebno je zamijeniti vektor značajki x_i s funkcijom $\Phi(x_i)$ koja ga preslikava iz *n*-dimenzionalnog prostora u *m*-dimenzionalni prostor, a pri tome je m>n po principu Hilbertovih prostora (engl. *Hilbert space*). Funkcija pretvorbe može biti raznih oblika, a najčešće su: polinomne funkcije, radijalne bazne funkcije, racionalne kvadratne funkcije i sigmoidalna funkcija. Slikom 3.11 prikazana je ideja pretvorbe iz dvodimenzionalnog u trodimenzionalni prostor.



Slika 3.11. *Prelazak između prostora pomoću jezgrenog trika. [17]*

4. PROGRAMSKO RJEŠENJE

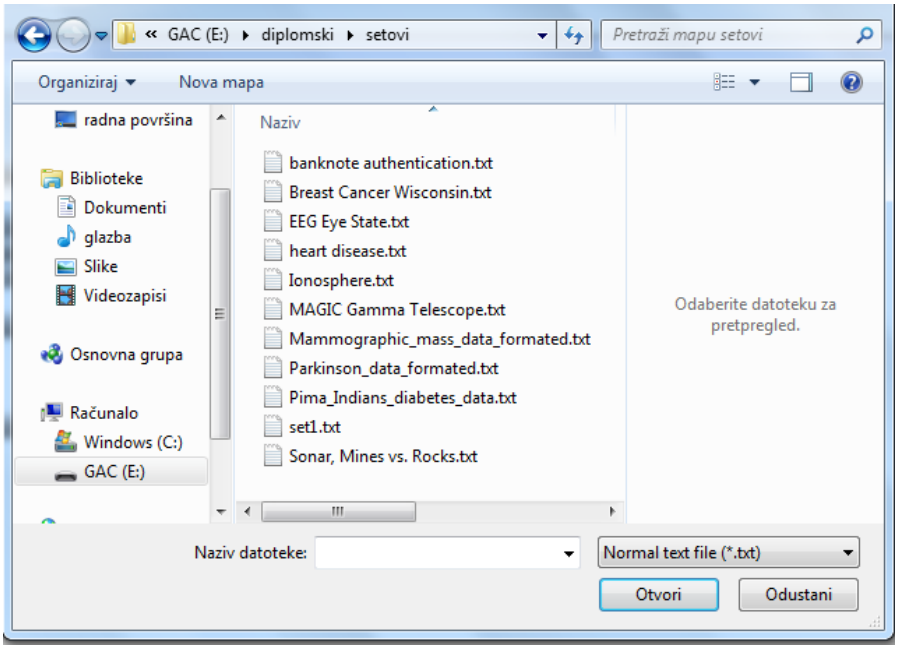
Aplikacija predstavlja klasifikator za binarno klasificiranje podataka napisana je u programskom jeziku C#. Uz klasificiranje medicinskih podataka može se koristiti i za klasificiranje drugih podataka koji zadovoljavaju parametre, odnosno, koji su napisani u obliku koji odgovara radu aplikacije. Tako se mogu klasificirati bilo koji podaci koji se nalaze u tekstualnoj datoteci u formatu vidljivom na slici 4.1. Format je definiran tako da se u prvom redu nalaze osobine i to na prvom mjestu ID uzorka, na drugom mjestu tip klase uzorka, a zatim nazivi parametara. Parametri trebaju biti prikazani realnim brojem, a odvojeni znakom točke s zarezom (;). Aplikacija je napravljena kao alat kako bi se ispitalo ponašanje i utjecaj broja k kNN klasifikatora na rezultat točnosti. Ispituje se i koliko različite mjere udaljenosti utječu na rezultate, odnosno, koja mjera udaljenosti je dobra za koje parametre te koje su im prednosti i mane. Aplikacijom se može vršiti i klasifikacija s Bayesovim klasifikatorom kako bi se mogli usporediti rezultati dobiveni kNN klasifikatorom.

```
id;tip;radius;texture;perimeter;area;smoothness;compactness;concavity;concave points;symmetry;fractal dimension
842302;M;17,99;10,38;122,8;1001;0,1184;0,2776;0,3001;0,1471;0,2419;0,07871
842517;M;20,57;17,77;132,9;1326;0,08474;0,07864;0,0869;0,07017;0,1812;0,05667
84300903;M;19,69;21,25;130;1203;0,1096;0,1599;0,1974;0,1279;0,2069;0,05999
84348301;M;11,42;20,38;77,58;386,1;0,1425;0,2839;0,2414;0,1052;0,2597;0,09744
8510426;B;13,54;14,36;87,46;566,3;0,09779;0,08129;0,06664;0,04781;0,1885;0,05766
8510653;B;13,08;15,71;85,63;520;0,1075;0,127;0,04568;0,0311;0,1967;0,06811
8510824;B;9,504;12,44;60,34;273,9;0,1024;0,06492;0,02956;0,02076;0,1815;0,06905
851509;M;21,16;23,04;137,2;1404;0,09428;0,1022;0,1097;0,08632;0,1769;0,05278
```

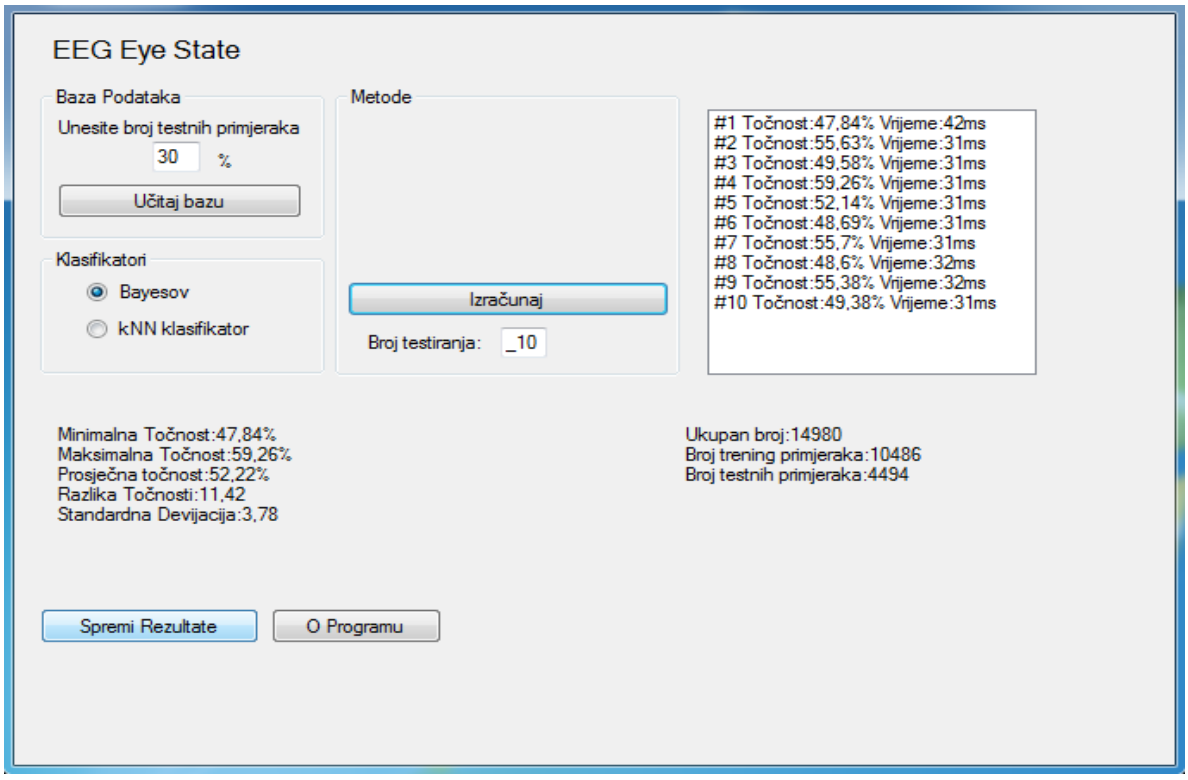
Slika 4.1. *Format baze podataka: Breast Cancer Wisconsin.*

4.1. Opis programskog rješenja

Proces počinje odabirom baze podataka nad kojom će se vršiti testiranje. Pritiskom na gumb „Učitaj bazu“ otvara se prozor za učitavanje baze prikazano na slici 4.2. Odabirom postotka određuje se koji postotak elemenata će se uzeti za trening skup, odnosno, veličina trening skupa. Ako pri podijeli broj elemenata testnog skupa ne bude cijeli broj veličina testnog skupa zaokružuje se na prvi manji cijeli broj.



Slika 4.2. *Odabir baze podataka.*



Slika 4.3. *Sučelje aplikacije Bayesov klasifikator.*

Na slici 4.3 prikazano je glavno sučelje aplikacije na kojemu se unose parametri testiranja, prikazuju rezultati testiranja i vrši spremanje rezultata. Odmah ispod okvira za podešavanje parametara baze podataka nalazi se radio gumbi za odabir klasifikatora. Može se odabrati Bayesov klasifikator ili kNN klasifikator. U središnjem dijelu je okvir za odabir parametara klasifikatora i broja testiranja. Kod Bayesovog klasifikatora nema odabira metoda već samo odabir broja

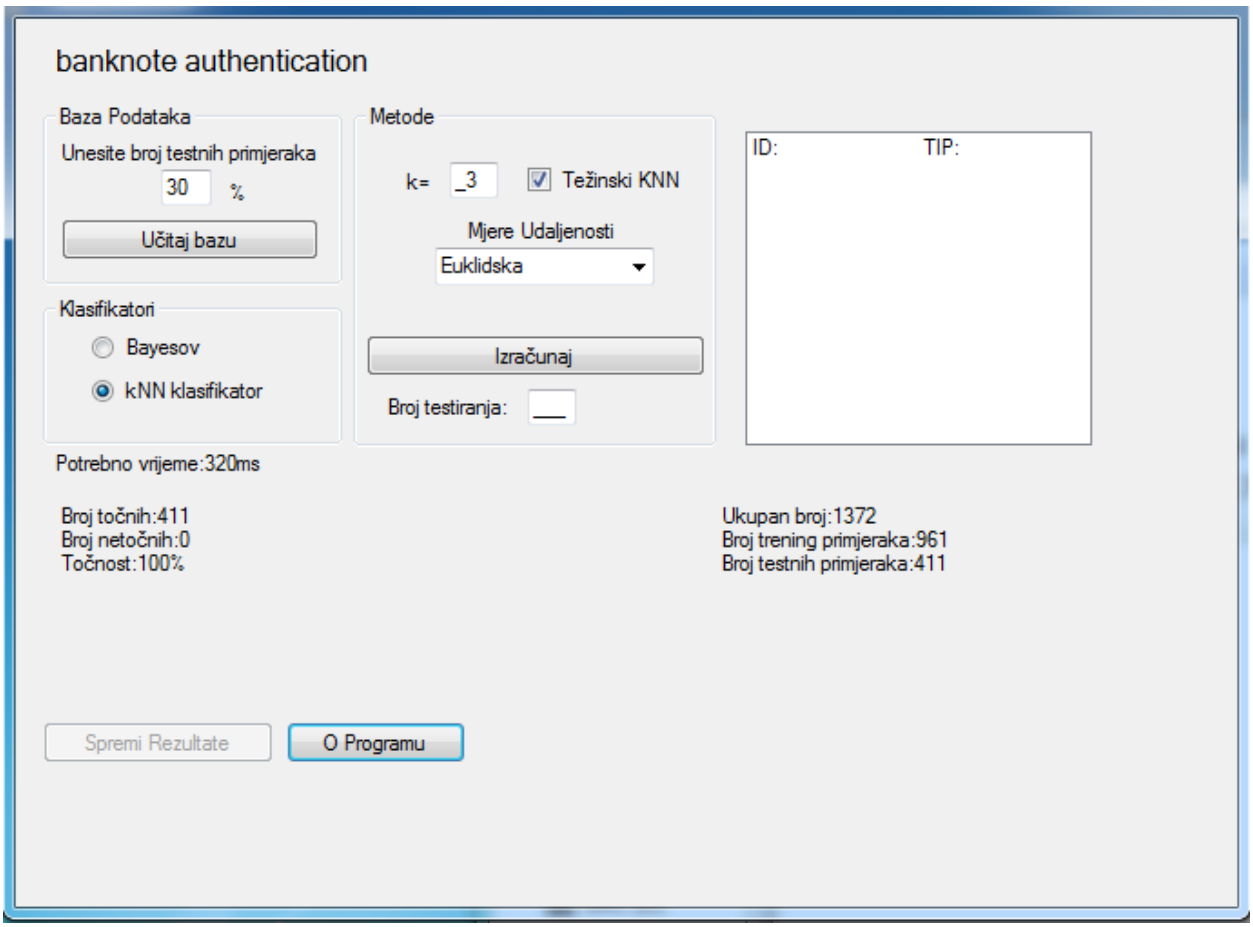
testiranja. Desno se nalazi okvir za prikaz rezultata. Ukoliko je odabrano više testiranja u okviru rezultata prikazuje se točnost i vrijeme potrebno za svako pojedino testiranje. U donjem dijelu sučelja prikazuju se statistički podaci za provedena testiranja kako bi se kasnije mogla vršiti analiza. Statistički podaci koji se računaju jesu:

- Minimalna točnost
- Maksimalna točnost
- Prosječna točnost
- Razlika maksimalne i minimalne točnosti
- Standardna devijacija
- Vrijeme trajanja testiranja

Prikazuju se još i podaci o ukupnom broju elemenata u bazi podataka kao i podaci o veličini testnog i trening skupa. Ako se umjesto više testiranja odabere provođenje samo jednog testa tada se u okviru rezultata umjesto prikaza točnosti i vremena pojedinih testova prikazuju pojedini uzorci za koje smo dobili netočne rezultate što je prikazano na slici 4.4. Izdvajaju se pojedini netočni rezultati kako bi se lakše utvrdili razlozi zbog kojih se dobije netočan rezultat. Tako se može provjeriti koje su osobine parametara netočno određenih uzoraka testnog skupa, a isto tako ako se ponovi test može se vidjeti griješi li algoritam uvijek na istim uzorcima i što je kod ti uzoraka specifično. Kao statistički podaci prikazuju se:

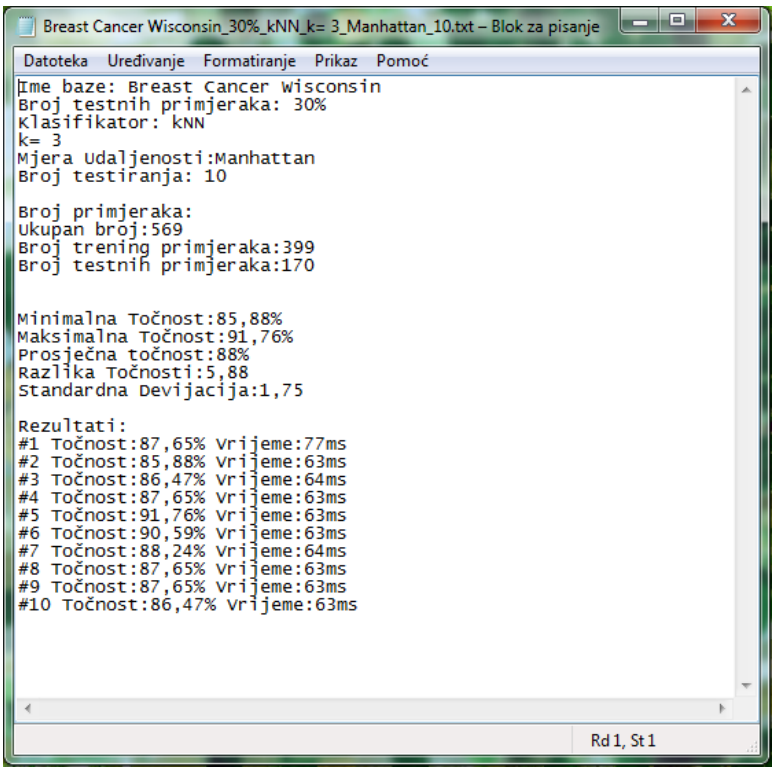
- vrijeme
- točnost
- broj točnih
- broj netočnih

Na slici 4.4 također je prikazan i odabir parametara za kNN klasifikator. Odabire se broj k, mjera udaljenosti te računa li se težinski faktor ili ne. Ako se odabere kNN bez težinskog faktora i za parametar k postavi parni broji te u susjedstvu bude podjednak broj uzoraka obje klase testnom uzorku dodjeljuje se tip klase najbližeg susjeda.



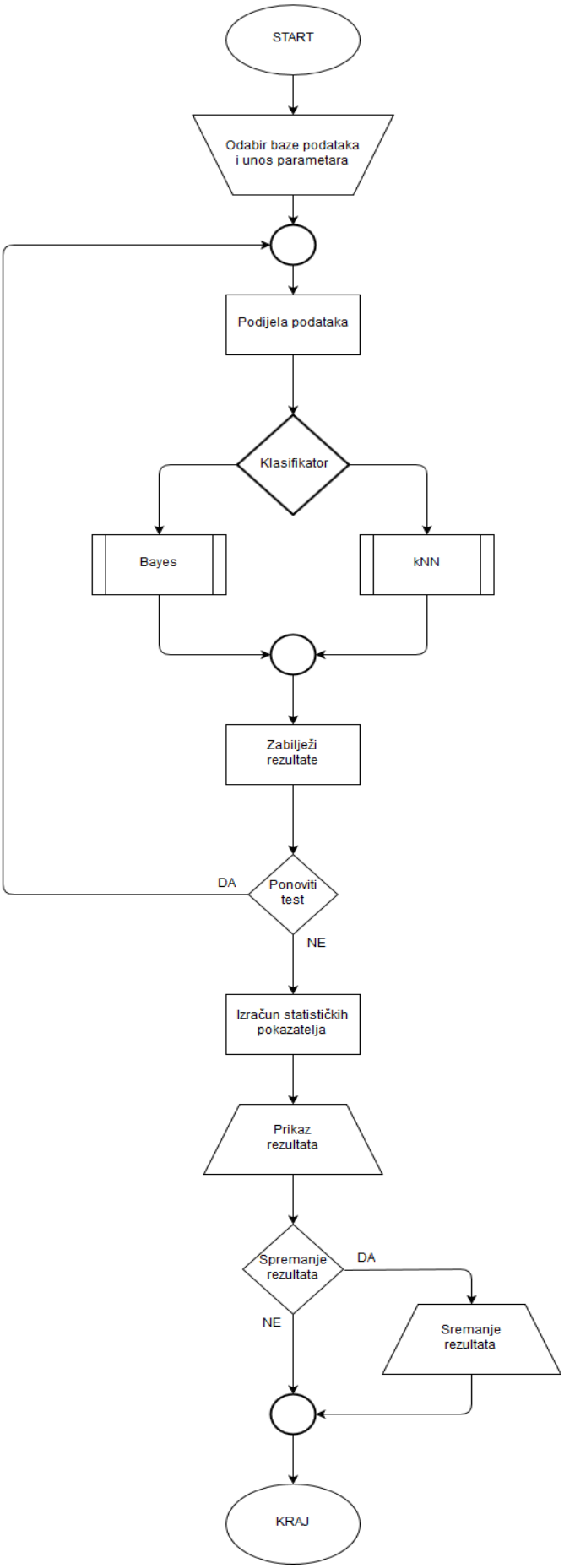
Slika 4.4. *Sučelje aplikacije kNN klasifikator.*

Nakon završetka testiranja nudi se opcija spremanja rezultata koji se odabire gumbom „Spremi Rezultate“. Rezultati se spremaju u tekstualnu datoteku sa svim parametrima i rezultatima u obliku prikazanom na slici 4.5. Ime datoteke generira se automatski tako da se zabilježe svi parametri testiranja (ime baze, postotak testnog skupa, klasifikator, metoda, broj testiranja, te koristili se težinski faktor) ili se ime može ručno upisati po želji.



Slika 4.5. *Format spremanja rezultata.*

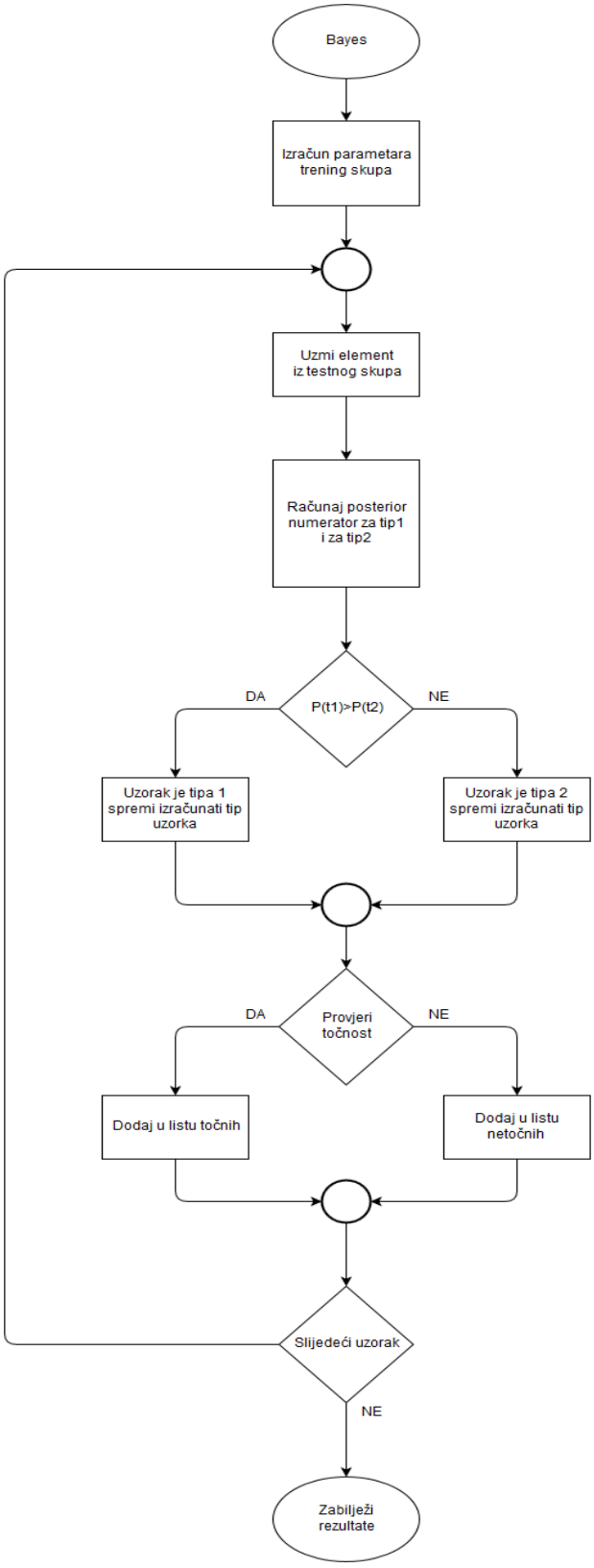
Rad klasifikatora prikazuje se slijedećim dijagramima stanja (slike 4.6.-4.9.). Dijagrami stanja napravljeni su prema [12] te daju uvid u rad algoritma klasifikacije kao i rad cijele aplikacije. Dijagram stanja cjelokupne aplikacije bio bi prevelik i kompliciran za prikaz na samo jednoj slici stoga je prema pravilima razdijeljen u više slika.



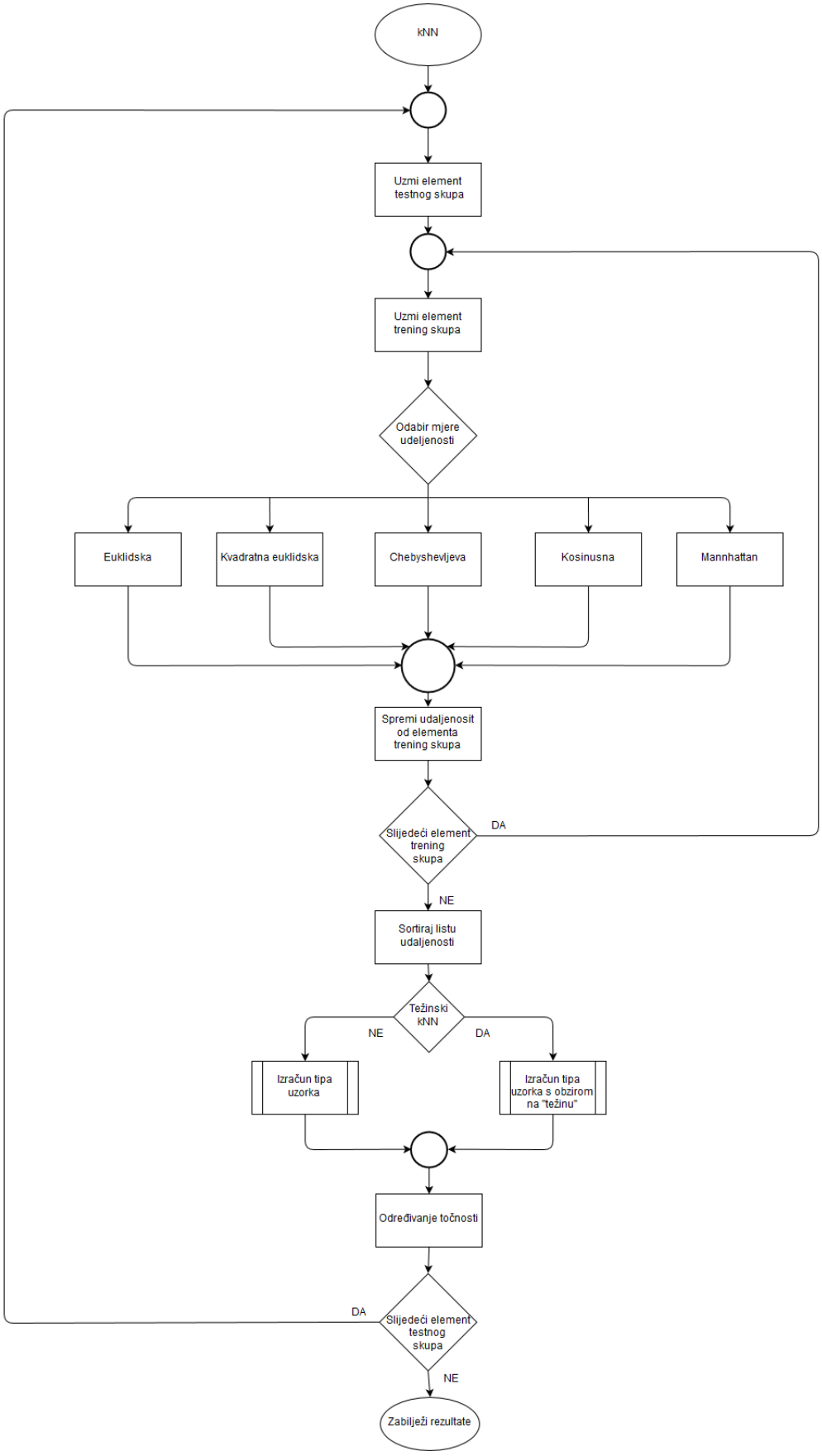
Slika 4.6. *Dijagram stanja aplikacije.*

Proces započinje odabirom baze podataka nad kojom se želi vršiti testiranje nakon čega se odabiru parametri testiranja potrebni za rad (koji postotak baze podataka će se uzeti za testni skup, broj testiranja, vrsta i parametri klasifikatora koji će se koristiti). Nakon učitavanja baze podataka od iste se na osnovi postotka zadanog u parametrima kreiraju dva skupa, testni skup i trening skup. Testni skup nad kojim će se vršiti testiranje stvara se tako da se uzmu nasumični uzorci iz ukupnog skupa broj elemenata (uzoraka) određuje se parametrima u postavkama aplikacije. Ostatak uzoraka čini testni skup. Tada se izvršava odabrani algoritam klasificiranja te se bilježe rezultati. Postupak se ponavlja sve dok se ne zadovolji broj ponavljanja ranije definiran u parametrima testiranja. Zatim se vrši izračun statističkih pokazatelja testiranja te se prikazuju rezultati. Ukoliko korisnik odabere opciju za spremanje rezultata, rezultati se spremaju u tekstualni dokument te se završava provedba testa. Rad algoritama klasifikatora opisan je slijedećim dijagramima prikazanim slikama 4.7. i 4.8.

Dijagram stanja rada Bayesovog klasifikatora karakterizira jednostavniji odabir određivanja tipa u odnosu na kNN klasifikator. Bazira se na usporedbi prosječnih parametara trening skupa i uzorka test skupa kojem se određuje tip. Bayesov klasifikator ima puno manje uspoređivanja od kNN klasifikatora što rezultira time da je njegov dijagram stanja puno pravocrtniji te ima samo jednu petlju koja prolazi kroz sve elemente testnog skupa. Dijagram stanja algoritma Bayesovog klasifikatora počinje izračunom parametara trening skupa. Zatim se uzima uzorak iz testnog skupa za koji se računa posterior numerator za tip 1 i za tip 2. Poslije se usporedbom posterior numeratora tipa 1 i tipa 2 dolazi do određivanja tipa uzorka. Zatim se provjerava točnost uzorka te se uzorak sprema kao element u tablicu točnih ako je određeni tip točan ili u tablicu netočnih ako je određeni tip pogrešan. Zatim se prelazi na sljedeći uzorak, odnosno, element testnog skupa ako isti postoji. Rezultati se bilježe kako bi se kasnije mogla vršiti analiza.



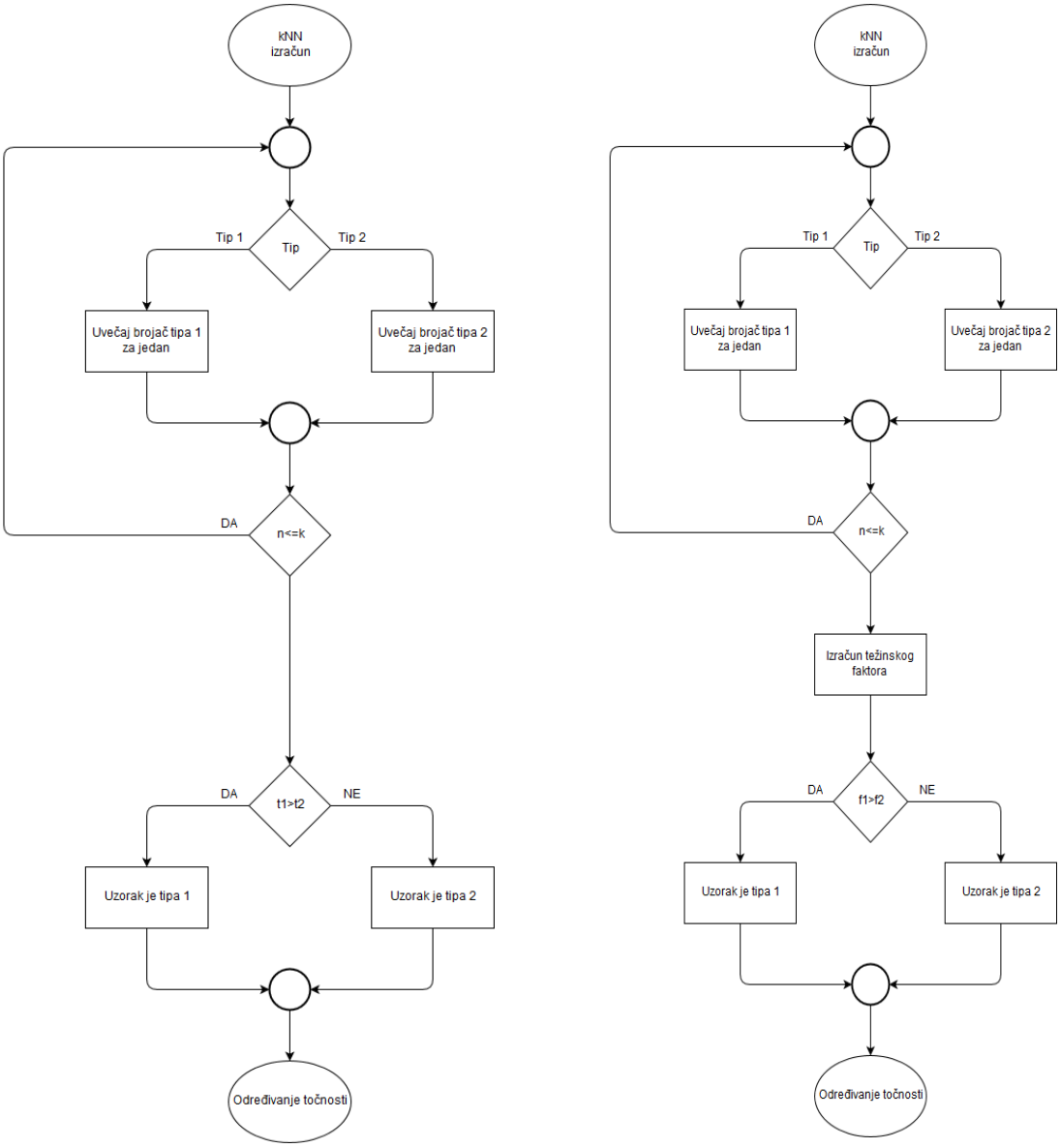
Slika 4.7. *Dijagram toka Bayesovog klasifikatora.*



Slika 4.8. *Dijagram toka kNN klasifikatora.*

Za razliku od dijagrama stanja Bayesovog klasifikatora, dijagram stanja kNN klasifikatora puno

je složeniji jer se za svaki testni primjerak računa udaljenost od svih elemenata testnog skupa. Što rezultira s više petlji, a uz to ima i odabir mjere udaljenosti kojim će se udaljenost računati te još i odabir radi li se o kNN klasifikatoru s težinskim faktorom ili bez njega. Algoritam za kNN klasifikator počinje tako što uzima uzorak iz testnog skupa te se na taj način ulazi u prvu petlju koja će prolaziti kroz sve elemente testnog skupa. Zatim se uzima element trening skupa što označava ulazak u drugu petlju koja prolazi kroz elemente trening skupa. Na temelju odabrane mjere udaljenosti računa se udaljenost ta dva elementa, a udaljenosti se spremaju na listu udaljenosti za svaki pojedini uzorak. Ukoliko nema više elemenata test skupa lista udaljenosti se sortira. Zatim se na osnovu odabira algoritma s težinskim faktorom ili bez njega određuje tip uzorka. Slijedi provjera točnosti određivanja tipa uzorka. Postupak se ponavlja za svaki element testnog skupa. Kada više nema elemenata testnog skupa rezultati se bilježe. Algoritmi određivanja tipa podataka prikazani su dijagramima na slici 4.9. Algoritmi za određivanje tipa uzoraka kNN klasifikatora razlikuju se vrlo malo, stoga su im i dijagrami slični. Algoritam s težinskim faktorom ima dodan blok za računanje težinskog faktora za razliku od algoritma bez težinskog faktora koji taj blok naredbi nema već on samo uspoređuje broj elemenata tipa 1 i tipa 2 u k elemenata sortirane liste udaljenosti od pojedinog uzorka. Algoritmi počinju prolaskom kroz petlju k puta tako da se iz sortirane liste udaljenosti uzme k elemenata te se prebroji broj elemenata tipa 1, odnosno, tipa 2. Pri određivanju tipa uzorka očituje se razlika jer se kod algoritma s težinskim faktorom računa težinski faktor, a tip uzorka jednak je tipu s većim težinskim faktorom. Dok se kod algoritma bez težinskog faktora tip uzorka određuje prebrojavanjem broja elemenata po tipovima po principu kojih tipova ima više u k elemenata sortirane liste udaljenosti, tog je tipa uzorak.



Slika 4.9. *Dijagrami stanja algoritama određivanja tipa uzoraka kNN klasifikatora* **a)** *bez težinskog faktora*, **b)** *s težinskim faktorom.*

4.2. Testiranje i analiza rezultata

Testiranje je provedeno na pet baza podataka, odnosno, skupova koji sadrže podatke o istraživanju raznih bolesti i stanja u medicini. Provedena su i testiranja na dodatne četiri baze podataka koje ne sadrže medicinske podatke kako bi se moglo pokazati da se aplikacija može primijeniti na raznim područjima. Baze podataka preuzete su sa repozitorija za strojno učenje [18]. Baze podataka potrebno je prije testiranja obraditi tako da zauzmu formu čitljivu aplikaciji za klasificiranje (slika 4.1.). Ukoliko u bazi podataka postoje uzorci kojima nedostaju vrijednosti tada se ti uzorci izostavljaju iz baze, a broj uzoraka baze se smanjuje za broj izostavljenih uzoraka. Skupovi s medicinskim podacima korišteni za testiranje su: Breast Cancer Wisconsin, Heart

Disease, Mammographic Mass, Parkinsons i EEG Eye State. Kao dodatni setovi podataka korišteni su Ionosphere, MAGIC Gamma Telescope, Sonar Mines vs. Rocks i Banknote Authentication data set. **Breast Cancer Wisconsin** je skup podataka koji sadrži rezultate testiranja raka dojke u Američkoj Saveznoj Državi Wisconsin. Parametri opisuju sliku jezgre stanice za koju se smatra da je zaražena. Parametri su promjer, tekstura, opseg, kompaktnost, konkavnost, konkavne točke, simetričnost i faktor fraktalne dimenzije. Raspon veličina parametara je različit. Tipovi klase su „M“ za maligni i „B“ za benigni. Broj uzoraka je 569 od čega je 357 benigni, a 212 maligni. **Heart Disease** je skup podataka dobiven mjerenjem različitih veličina prilikom vježbe kako bi se otkrile moguće bolesti srca ispitanika. Mjerene veličine predstavljaju atribute, a tipovi klasa su tip „1“ za zdravo srce i tip „2“ za prisutnost bolesti srca. Neki atributi su normirani pa tako atribut defekta talamusa poprima vrijednost 3 ako defekt nije prisutan, 6 ako je učvršćeni defekt (engl. *fixed defect*), a vrijednost 7 ako je reverzibilni defekt. Normirana je i prisutnost boli u prsima s rasponom vrijednosti od 1 do 4. Mjere se još godine ispitanika, spol, krvni tlak, razina kolesterola, maksimalan broj otkucaja srca, prisutnost upale te vrh EKG-a. U skupu nema vrijednosti koje nedostaju, a broj uzoraka je 270. **Mammographic Mass** je skup podataka dobiven mamografijom. Atributima se opisuju slike kvržica dobivenih pregledom. U skupu svi atributi imaju cjelobrojne vrijednosti stoga su svi atributi osim godina pacijenta (godine su po prirodi cjelobrojne) normirani. Atributi uz godine su faktor BI-RADS skale, oblik, margina, gustoća. Tip klase „0“ predstavlja benigni uzorak, a tip klase „1“ maligni uzorak. Klasificiranje ove baze podataka koristi se ponajviše za izbjegavanje nepotrebnih biopsija kvržica jer otprilike 70% nalaza biopsije kvržica pokazuje da su one benignog tipa. Na ovaj način želi se izbjeći nepotrebna biopsija koja je jako invazivan zahvat. U skupu postoje uzorci kojima nedostaju vrijednosti pa je originalni broj uzoraka s 961 smanjen na 830 izbacivanjem uzoraka kojima nedostaju vrijednosti. **Parkinsons** je skup podataka dobiven biomedicinskom analizom glasovnih uzoraka osoba kako bi se odredila prisutnost Parkinsonove bolesti. Uz godine i spol pacijenta te vrijeme sakupljanja uzorka kao atributi se uzimaju biomedicinske osobine govora (glasnoća, frekvencija, razna kašnjenja, omjeri tonska zvuka i šuma te fraktalni eksponent) i faktori unificirane skale Parkinsonove bolesti (engl. *Unified Parkinson's disease rating scale*, UPDRS). Broj jedinki uzoraka je smanjen s 197 na 195 zbog dva uzorka s nepotpunim vrijednostima, tip klase „0“ označava zdravog ispitanika, a tip klase „1“ prisutnost Parkinsonove bolesti. **EEG Eye State** je skup podataka o stanju ljudskog oka pacijenta pri snimanju glave EEG-om. Stanje oka je zabilježeno kamerom te određuje klasu. Tip klase poprima vrijednost „0“ za zatvoreno i „1“ za otvoreno oko. Atributi su mjerenja dobivena kanalima EEG-a. Skup je nastao kako bi potpomogao daljnja istraživanja. Skup je velik i sastoji se od 14980 mjerenja koja predstavljaju uzorke. Mjerenja su poredana kronološkim redom. **Ionosphere** je skup podataka dobivenih odzivom 16 visoko frekventnih antena na elektrone u

ionosferi kako bi se detektirala struktura. Svaka antena ima dva odziva koji predstavljaju atribute, a tipovi klase su „g“ ukoliko je pronađen objekt i „b“ ako objekt nije pronađen. **MAGIC Gamma Telescope** je skup podataka za otkrivanje gama zračenja putem teleskopa. Odziv teleskopa predstavlja atribute, a tip klase „g“ označava postojanje gama zračenja dok tip klase „h“ pokazuje da zračenja nema. Skup je velik i sadrži 19020 uzoraka od kojih 12332 sa zabilježenim gama zračenjem. **Sonar Mines vs. Rocks** je skup podataka koji sadrži odzive sonara za razne objekte. Odzivi predstavljaju atribute na osnovu kojih se određuje prisutnost podvodne mine koja predstavlja prvu klasu označenu s „M“. Ako je objekt stijena, ona predstavlja drugu klasu, a označava se s „R“. Skup sadrži 208 uzoraka, a karakterizira ga velik broj atributa njih 60. **Banknote Authentication** je skup podataka prikupljenim transformacijom iz slika za bankovnu autorizaciju. Skup sadrži 1372 uzorka klasificiranih na osnovi 5 atributa. Skup karakterizira široka granica između klasa te se pri njegovoj klasifikaciji očekuju najbolji rezultati.

Rezultati testiranja prikazani su tablicama i odgovarajućim grafovima. U ovom radu prikazani su samo odabrani testovi sa bitnim značajkama dok se sva testiranja (njih više od 500) zajedno sa svim značajkama nalaze u dodatku.

Prvi testovi koji su provedeni bili su testiranje kNN klasifikatora mijenjajući parametar k i mjeru udaljenosti. Testovi su provedeni na različitim skupovima kako bi se ustvrdilo ponašanje kNN klasifikatora tako da se vidi koja je mjera udaljenosti bolja za koji skup podataka i kako vrijednost parametra k utječe na rezultate. Omjer testnog i trening skupa bio je takav da se 20% ukupnog skupa odredi kao testni skup. U sljedećim tablicama (tablice 4.1.-4.5.) prikazani su rezultati prosječne točnosti od 20 testiranja za zadane parametre. Za parametar k uzete su vrijednosti koje se najčešće koriste, a uzeti su i neki parni brojevi kako bi se prikazao njihov utjecaj te vrijednost k=21 kako bi se prikazao utjecaj velikog broja k. Promatrala se točnost jer u medicini nije toliko važno jesu li netočno pozitivni ili netočno negativni. Razlog tomu je da se ne bi zdravom čovjeku pripisala neka invazivna terapija ili zahvat koji može dovesti do pogoršanja njegovog zdravstvenog stanja.

Tablica 4.1. *Rezultati točnosti kNN klasifikatora za Breast Cancer Wisconsin.*

	k=1	k=2	k=3	k=5	k=6	k=7	k=21
EUKLIDSKA	85.75%	83.52%	88.67%	88.14%	83.43%	86.73%	88.80%
KVAD. EUK.	86.63%	86.73%	89.03%	88.14%	83.74%	87.43%	89.12%
MANHATTAN	87.43%	86.02%	89.12%	87.79%	83.43%	85.84%	89.32%
CHEBYSHEV	85.84%	84.94%	88.67%	88.59%	82.28%	88.59%	88.63%
KOSINUS	83.28%	82.30%	86.02%	87.44%	76.86%	86.73%	87.43%

Tablica 4.2. *Rezultati točnosti kNN klasifikatora za Heart Disease.*

	k=1	k=2	k=3	k=5	k=6	k=7	k=21
EUKLIDSKA	62.96%	57.41%	77.04%	74.54%	65.93%	74.07%	75.93%
KVAD. EUK.	64.81%	57.41%	74.04%	70.37%	68.52%	74.07%	75.12%
MANHATTAN	68.52%	74.07%	79.63%	70.18%	67.96%	77.78%	75.93%
CHEBYSHEV	60.37%	55.56%	64.84%	67.68%	65.275	75.93%	76.63%
KOSINUS	66.67%	64.81%	72.96%	69.07%	62.04%	66.67%	68.52%

Tablica 4.3. *Rezultati točnosti kNN klasifikatora za Mammographic Mass.*

	k=1	k=2	k=3	k=5	k=6	k=7	k=21
EUKLIDSKA	71.89%	70.23%	79.32%	82.33%	77.51%	80.40%	79.22%
KVAD. EUK.	74.70%	71.12%	80.38%	80.88%	74.70%	79.22%	78.84%
MANHATTAN	73.49%	69.76%	79.92%	79.96%	72.47%	80.44%	80.74%
CHEBYSHEV	71.49%	68.04%	76.27%	78.31%	69.83%	76.35%	76.89%
KOSINUS	71.08%	65.32%	76.02%	76.83%	65.63%	78.07%	78.80%

Tablica 4.4. *Rezultati točnosti kNN klasifikatora za Parkinsons.*

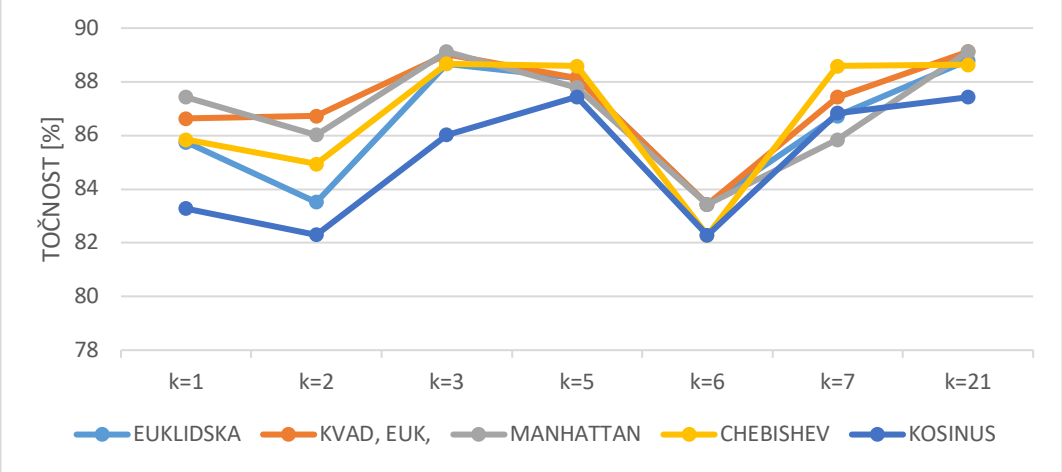
	k=1	k=2	k=3	k=5	k=6	k=7	k=21
EUKLIDSKA	75.86%	72.41%	80.17%	84.48%	77.59%	78.96%	83.02%
KVAD. EUK.	75.43%	70.69%	79.22%	82.34%	77.07%	80.34%	80.86%
MANHATTAN	80.17%	75.86%	84.05%	81.9%	74.14%	82.07%	81.38%
CHEBYSHEV	74.31%	67.24%	75.86%	78.36%	70.69%	80.00%	80.26%
KOSINUS	78.28%	66.52%	78.88%	79.83%	68.14%	79.05%	79.31%

Tablica 4.5. *Rezultati točnosti kNN klasifikatora za EEG Eye State.*

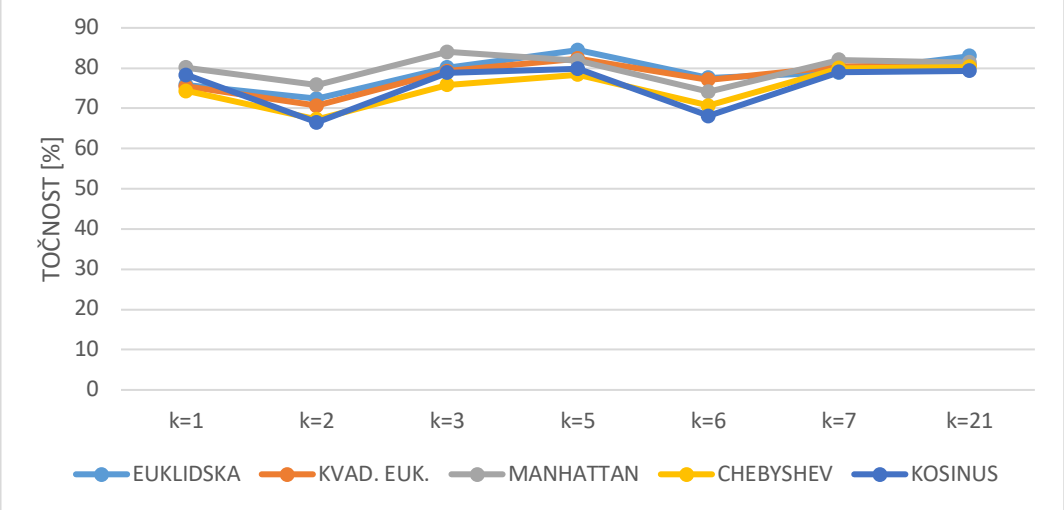
	k=1	k=2	k=3	k=5	k=6	k=7	k=21
EUKLIDSKA	94.56%	92.27%	96.80%	97.18%	91.31%	96.12%	96.07%
KVAD. EUK.	94.09%	91.89%	96.46%	97.04%	90.67%	96.32%	95.34%
MANHATTAN	93.23%	90.72%	96.55%	97.10%	91.28%	96.67%	95.73%
CHEBYSHEV	93.37%	90.31%	95.84%	95.89%	90.12%	94.87%	93.92%
KOSINUS	91.43%	89.31%	93.08%	94.13%	89.48%	93.34%	93.12%

Usporede li se rezultati u tablicama 4.1.-4.5. vidi se da vrijednosti točnosti variraju ovisno o skupu nad kojim je vršena klasifikacija. Najbolji rezultati bili su za skup EEG Eye State (EEG stanje

oka), a najlošiji za Heart Disease (bolesti srca). Odabir mjera udaljenosti također utječe na točnost rezultata. Različite mjere udaljenosti pokazale su se bolje za određene skupove podataka. Manhattan mjera udaljenosti najbolje rezultate daje za skup uzoraka Parkinsonove bolesti (tablica 4.4.), kvadratna euklidska za Breast Cancer Wisconsin (tablica 4.1.) dok euklidska mjera udaljenosti daje poprilično dobre rezultate za sve skupove podataka. Promotre li se tablice 4.1-4.5. zajedno može se primijetiti utjecaj parametra k na točnost rezultata i uočiti pravilnost, da najlošiji rezultati su za parne brojeve broja k (k=2 i k=6) što je zbog načina rada kNN klasifikatora. Nešto bolji rezultati su sa specijalni slučaj kada je k =1, dok su najbolji rezultati za male neparne brojeve kao što su k=3 i k=5. Za veliku vrijednost parametra k (k=21) rezultati su se pokazali dobri no potrebno je puno više vremena za klasifikaciju a i puno više memorije za izvođenje samog programa, a rezultati nisu toliko dobri da bi se to isplatilo. Utjecaj parametra k na točnost rezultata može se prikazati i grafom (slike 4.10 i 4.11.).



Slika 4.10. *Graf točnosti rezultata za Breast Cancer Wisconsin.*



Slika 4.11. *Graf točnosti rezultata za Parkinsonovu bolest.*

Iako je vidljivo da za parnu vrijednost parametra k kNN klasifikator daje najlošije rezultate oni ipak nisu toliko niski u odnosu prema rezultatima ostalih vrijednosti parametra k. Za očekivati bi

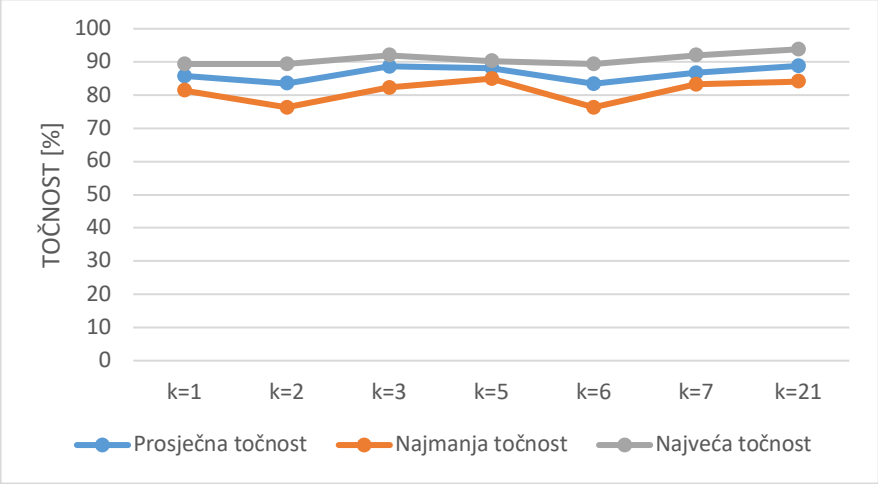
bilo da su rezultati još lošiji. Razlog tomu je taj što iako je broj susjeda paran nije uvijek slučaj da je u susjedstvu jednak broj klase tipa 1 i klase tipa 2. Jedino taj slučaj razlikuje parnu vrijednost parametra k od nepare prema rezultatu jer ukoliko je u susjedstvu više uzoraka koji su klase različite od stvarne klase testnog uzorka rezultat će u oba slučaja (i za paran i za neparan broj k) biti netočan. Bolji uvid u rezultate ovog slučaja dobije se usporedbom najlošijih rezultata točnosti prikazanih tablicom 4.6. i 4.7. te odgovarajućim grafom (slike 4.12. i 4.13.). Kako ponašanje rezultata vrijedi za sve skupove nad kojima su provedeni testovi i za sve mjere udaljenosti na grafu prikazuju se rezultati za euklidsku mjeru udaljenosti za skup podataka Breast Cancer Wisconsin i Parkinsons.

Tablica 4.6. *Najmanje točnosti kNN klasifikatora za Breast Cancer Wisconsin.*

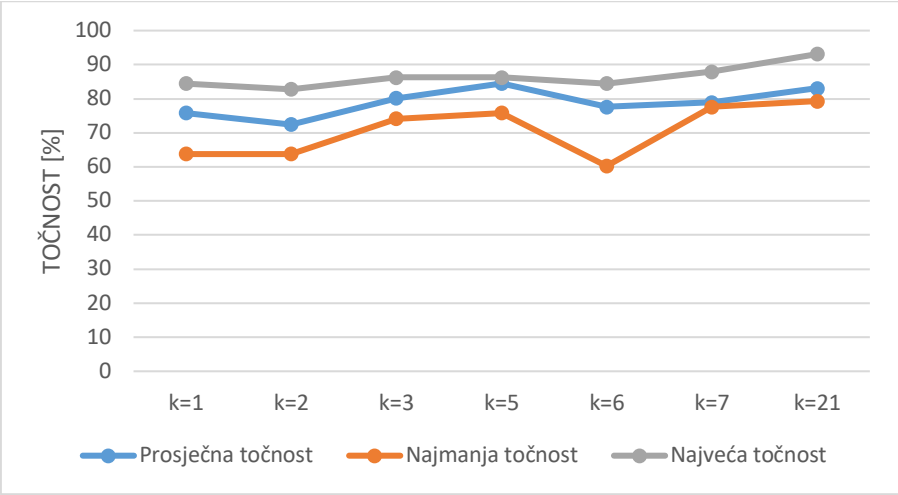
	k=1	k=2	k=3	k=5	k=6	k=7	k=21
EUKLIDSKA	81.42%	76.34%	82.30%	84.96%	76.34%	83.19%	84.07%
KVAD. EUK.	83.19%	78.42%	86.73%	83.19%	74.28%	81.42%	84.96%
MANHATTAN	76.99%	73.53%	86.73%	84.96%	72.73%	82.30%	88.50%
CHEBYSHEV	80,83%	76.53%	85.84%	84.07%	74.28%	84.96%	84.07%
KOSINUS	76.11%	72.88%	77.88%	82.30%	73.53%	80.83%	80.53%

Tablica 4.7. *Najmanje točnosti kNN klasifikatora za Parkinsons.*

	k=1	k=2	k=3	k=5	k=6	k=7	k=21
EUKLIDSKA	63.79%	63.79%	74.14%	75.86%	60.24%	77.59%	79.31%
KVAD. EUK.	63.79%	60.24%	72.41%	74.14%	59.42%	77.07%	75.86%
MANHATTAN	65.52%	63.79%	79.31%	75.86%	63.79%	74.14%	75.85%
CHEBYSHEV	61.71%	60.24%	68.97%	70.69%	61.71%	70.69%	74.14%
KOSINUS	62.38%	59.42%	70.69%	74.14%	59.42%	68.14%	72.41%



Slika. 4.12. *Graf točnosti euklidske mjere udaljenosti za Breast Cancer Wisconsin.*



Slika 4.13. *Graf točnosti euklidske mjere udaljenosti za Parkinsons.*

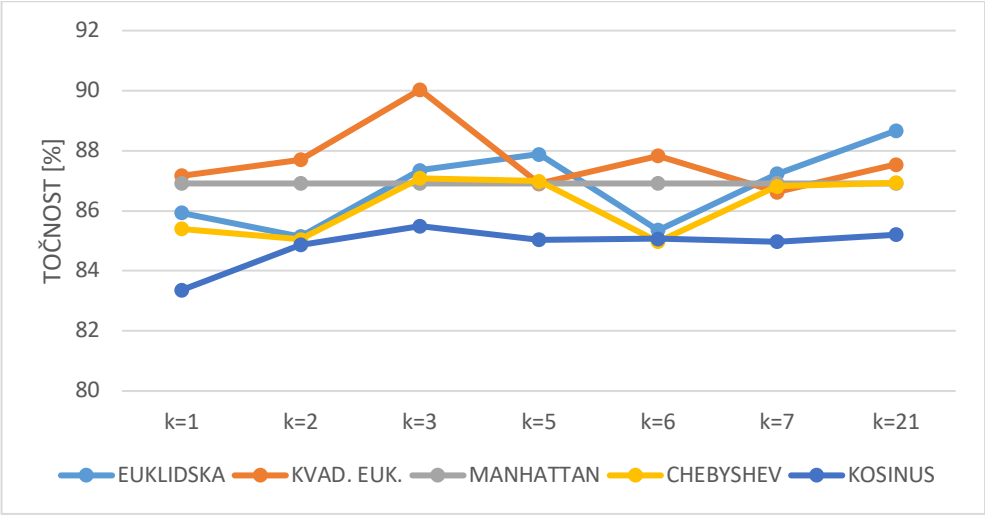
Iz grafova na slikama 4.12. i 4.13. možemo vidjeti da najmanja točnost najviše odstupa od prosječne točnosti za parne vrijednosti parametra k što je ujedno i glavni razlog zašto vrijednost parametra k treba biti neparan broj. Može se primijetiti da je i razlika točnosti za parne parametre k veća od one za neparne. Ukoliko se primjeni algoritam kNN klasifikatora s težinskim faktorom tada parametar k može biti paran broj i to neće imati toliko velik utjecaj na rezultat. Rezultati dobiveni klasifikacijom s težinskim faktorom prikazani su tablicama 4.8. i 4.9. a utjecaj na rezultat odgovarajućim grafom (slike 4.14. i 4.15.).

Tablica 4.8. *Prosječne točnosti WkNN klasifikatora za Breast Cancer Wisconsin.*

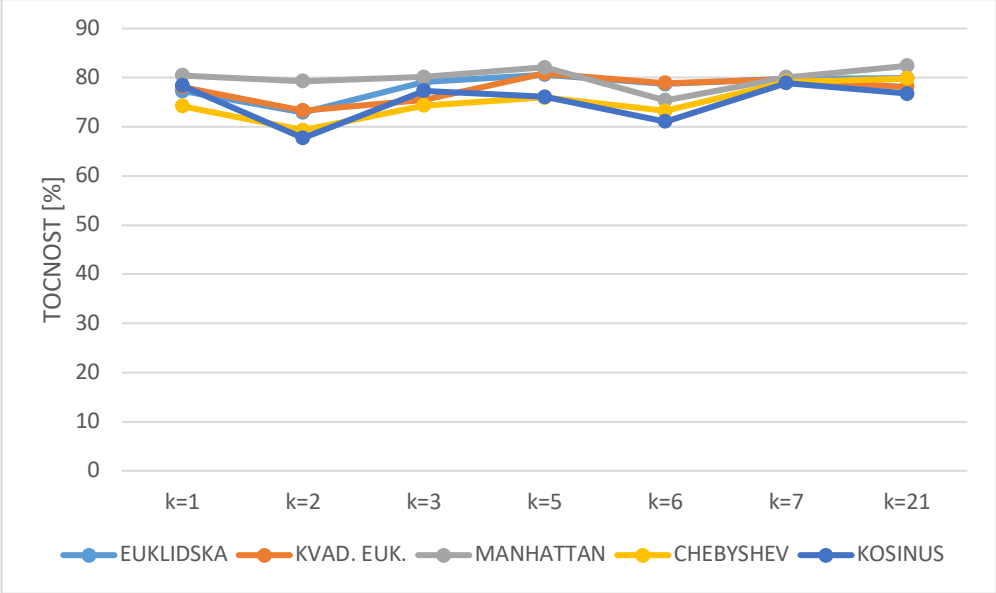
	k=1	k=2	k=3	k=5	k=6	k=7	k=21
EUKLIDSKA	85.93%	85.14%	87.35%	87.88%	85.36%	87.23%	88.67%
KVAD. EUK.	87.17%	87.70%	90.03%	86.90%	87.83%	86.63%	87.54%
MANHATTAN	86.91%	86.20%	89.47%	90.18%	86.49%	89.17%	89.76%
CHEBYSHEV	85.40%	85.05%	87.08%	86.99%	84.97%	86.82%	86.93%
KOSINUS	83.36%	84.87%	85.49%	85.04%	85.07%	84.97%	85.21%

Tablica 4.9. *Prosječne točnosti WkNN klasifikatora za Parkinsons.*

	k=1	k=2	k=3	k=5	k=6	k=7	k=21
EUKLIDSKA	77.26%	72.91%	79.14%	80.60%	78.72%	79.83%	79.91%
KVAD. EUK.	78.12%	73.29%	75.43%	80.86%	78.93%	79.65%	78.10%
MANHATTAN	80.47%	79.31%	80.17%	82.07%	75.37%	80.08%	82.41%
CHEBYSHEV	74.18%	69.34%	74.31%	75.95%	73.23%	79.14%	79.80%
KOSINUS	78.41%	67.72%	77.33%	76.12%	71.08%	78.89%	76.72%

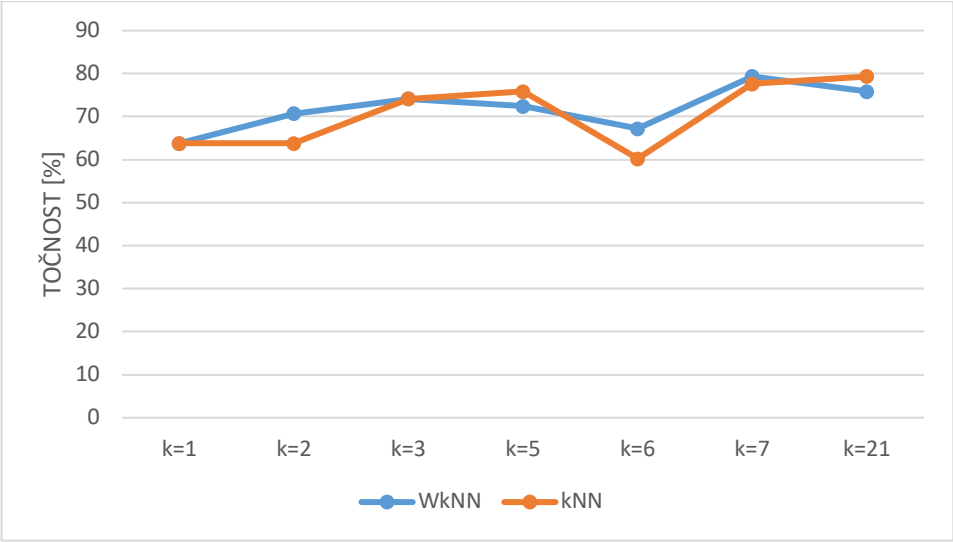


Slika 4.14. *Prosječna točnost WkNN klasifikatora za Breast Cancer Wisconsin.*



Slika 4.15. *Prosječna točnost WkNN klasifikatora za Parkinsons.*

Promotre li se tablice 4.8. i 4.9. te grafovi (slike 4.14. i 4.15.) može se vidjeti da se i dalje bolji rezultati postižu za neparne parametre k samo što je razlika točnosti za parametre k puno manja u odnosu na odgovarajuće rezultate dobivene kNN klasifikatorom bez težinskog faktora. Najveća razlika u primjeni težinskog faktora vidi se usporedbom najlošijih točnosti testiranja. Na slici 4.16. prikazana je usporedba najmanje točnosti euklidske mjere udaljenosti za kNN klasifikator sa i bez težinskog faktora na skupu Parkinsons. Na grafu se može prijetiti da su najveće razlike za parni parametar k.



Slika 4.16. *Graf najmanjih točnosti euklidske mjere udaljenosti za skup Parkinsons s obzirom na težinski faktor.*

Tijekom testiranja mjereno je vrijeme potrebno za klasificiranje kako bi se prikazao utjecaj parametara kNN klasifikatora na samo vrijeme klasificiranja. Vrijeme potrebno za klasificiranje prikazano je tablično, a vremena su izražena u milisekundama (ms). Prikazani su rezultati za skupove Mammographic Mass (tablica 4.10.) i EEG Eye State (tablica 4.11), a vremena ostalih skupova prate takav obrazac. Odabir težinskog faktora gotovo da i nema utjecaj na vrijeme trajanja klasifikacije. Dok mjere udaljenosti jako malo utječu na vrijeme klasifikacije. Na vrijeme potrebno za klasifikaciju utječu naravno i performanse samog stroja (računala) pomoću kojega se vrši klasifikacija. Prilikom testiranja korišteno je računalo s modelom procesora *Intel i5(2.5 GHz)* i 4 GB radne memorije.

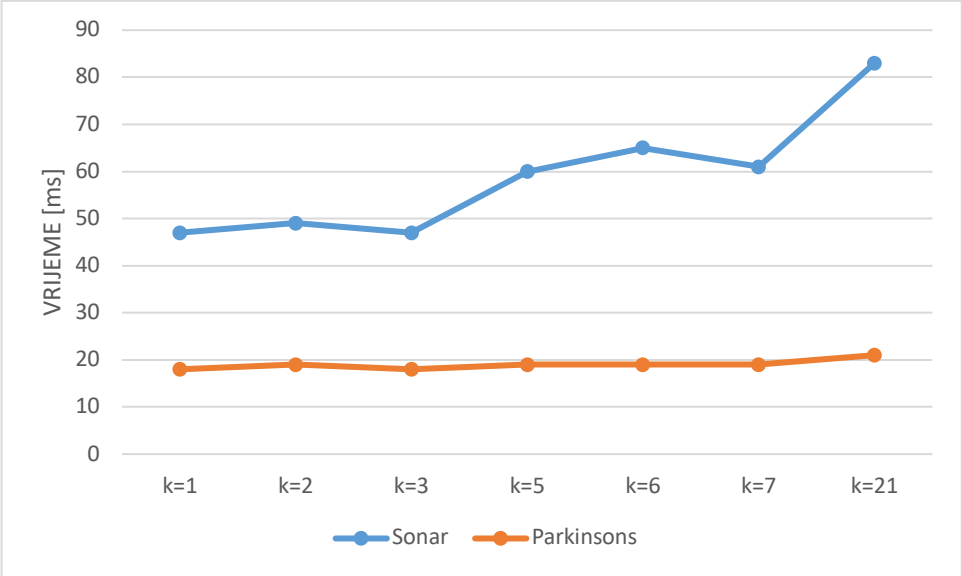
Tablica 4.10. *Vremena trajanja klasifikacije kNN klasifikatora za Mammographic Mass.*

	k=1	k=2	k=3	k=5	k=6	k=7	k=21
EUKLIDSKA	120	120	120	124	125	127	129
KVAD. EUK.	121	121	122	126	126	126	128
MANHATTAN	92	94	94	94	94	96	98
CHEBYSHEV	92	94	95	98	98	99	102
KOSINUS	161	162	164	165	166	168	169

Tablica 4.11. *Vremena trajanja klasifikacije kNN klasifikatora za EEG Eye State.*

	k=1	k=2	k=3	k=5	k=6	k=7	k=21
EUKLIDSKA	69805	69987	70155	70452	70494	70863	74398
KVAD. EUK.	70117	70174	70268	70623	70812	70996	73987
MANHATTAN	60396	60488	60527	60673	60698	61024	64653
CHEBYSHEV	60421	60423	60470	60538	60619	60898	63782
KOSINUS	73012	72996	73056	73890	73967	74023	79873

Promotre li se rezultati iz tablica 4.10. i 4.11. može se primijetiti da neke mjere udaljenosti zbog kompleksnijeg izraza za određivanje udaljenosti imaju i veća vremena trajanja klasifikacije. Također je vidljivo da se povećanjem parametra k povećava i vrijeme trajanja jer je potrebno prebrojati više uzoraka u susjedstvu. Vidi se da i veličina ukupnog skupa uvelike utječe na vrijeme trajanja pa tako skup EEG Eye State ima puna veća vremena od skupa Mammographic Mass jer je i njegov broj uzoraka puno veći. Parametar k više utječe na vrijeme ako je skup veći. Osim broja uzoraka, na trajanje utječe i broj atributa skupa. Utjecaj broja atributa na vrijeme trajanja prikazan je grafom na slici 4.16. Na grafu se vidi usporedba vremena trajanja za dva skupa približno iste veličine ali različitog broja parametara. Grafom su prikazana vremena trajanja klasifikacije kvadratne euklidske mjere udaljenosti za skup Parkinsons (195 primjeraka, 15 atributa) i skup podataka Sonar Mines vs. Rocks (208 primjeraka, 60 atributa).



Slika 4.16. *Utjecaj broja atributa i parametra k na vrijeme klasificiranja kNN klasifikatora.*

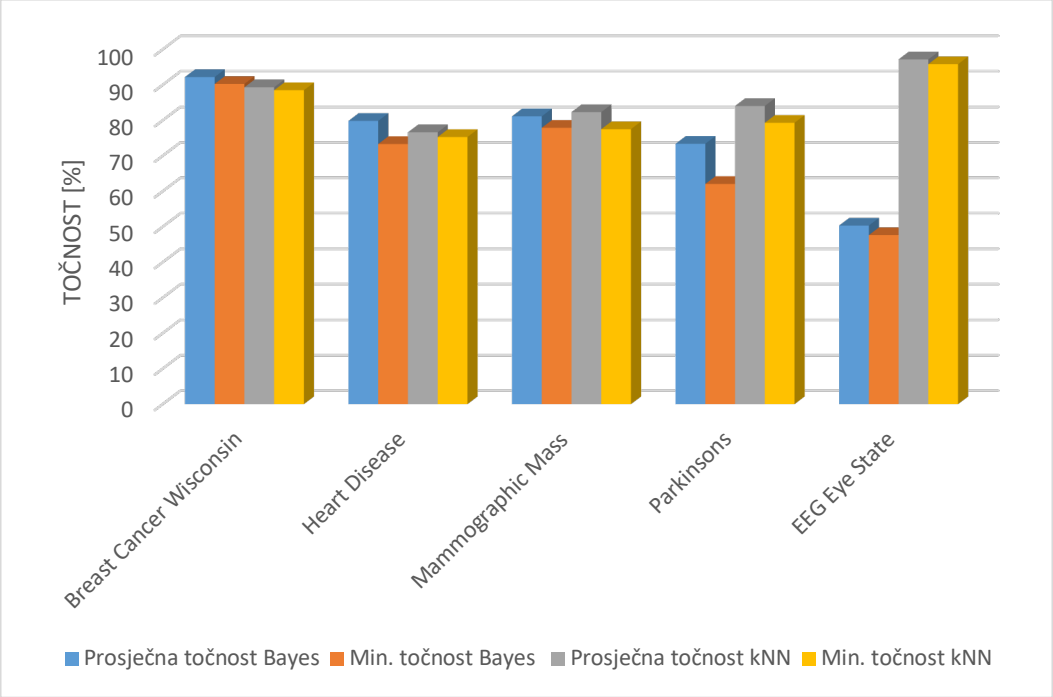
Provedena su i testiranja Bayesovog klasifikatora na istim skupovima podataka. Kako Bayesov klasifikator nema odabira parametara rezultati su prikazani po skupovima. Tablicom 4.12. prikazana su prosječna vremena i prosječne točnosti 20 testiranja putem Bayesovog klasifikatora

za odabrane skupove.

Tablica 4.12. *Vremena trajanja i rezultati točnosti klasifikacije Bayesovim klasifikatorom.*

SKUP PODATAKA	VRIJEME [ms]	MIN. TOČNOST	PROSJEČNA TOČNOST
Breast Cancer Wisconsin	2 ms	90.27 %	92.21 %
Heart Disease	2 ms	73.33 %	79. 85 %
Mammographic Mass	1 ms	77.91 %	81.14 %
Parkinsons	1 ms	62.07 %	73.39 %
EEG Eye State	30 ms	47.66 %	50. 34 %

Usporede li se rezultati Bayesovog klasifikatora s rezultatima kNN klasifikatora može se primijetiti da Bayesov klasifikator radi puno brže. Vremena klasifikacije također kao i kod kNN ovise o broju uzoraka skupa kao i o broju parametara. Točnost Bayesovog klasifikatora varira i za neke skupove je vrlo dobra (npr. Breast Cancer Wisconsin) dok je za neke vrlo loša (npr. EEG Eye State), a to ovisi o prirodi skupa. Glavna prednost kNN klasifikatora je ta što se on može prilagoditi skupu podešavanjem parametara dok Bayesov klasifikator to ne može. Tako je moguće za neke skupove dobiti i točnost od 100% za pravilno podešene parametre. Primjer je skup Banknote Authentication što je prikazano na slici 4.4. Usporedba rezultata testiranja prikazana je grafom na slici 4.17. Uspoređuju se prosječne i minimalne točnosti za pojedine skupove tako da se za kNN uzme najbolja vrijednost bez obzira na parametre.



Slika 4.17. *Usporedba točnosti rezultata kNN i Bayesovog klasifikatora.*

5. ZAKLJUČAK

Klasifikacija se nalazi u svim sferama ljudskog života. Ljudi sve klasificiraju i na taj način uče. Općenito je stvari lakše shvatiti ako se zna gdje pripadaju jer se odmah znaju i njihova obilježja. Klasifikacija se vrši prema određenim kriterijima. Nekad je te kriterije lako utvrditi, a nekad je teže. Što bolje poznajemo kriterije i što više znamo njihovu prirodu lakše nam je klasificirati neku stvar. Tako možemo i izraditi bolji klasifikator, odnosno, klasifikator koji daje bolje rezultate.

Što je klasifikator jednostavniji on radi brže, a što je klasifikator kompliciraniji, odnosno, uzima u obzir više parametara, treba mu više vremena za određivanje što ne mora nužno značiti da će dati točnije rezultate. Točnost rezultata ovisi o prirodi skupa koji testiramo. Zato je potrebno dobro proučiti i definirati koji su parametri bitni za klasifikaciju pojedinog skupa podataka, a koji nisu. Usporede li se Bayesov klasifikator i kNN klasifikator, može se zaključiti da Bayesov klasifikator radi brže. Prednost kNN klasifikatora je ta što se može prilagoditi skupu odabirom mjere udaljenosti i parametra k kako bi dao točnije rezultate, a i postavljanjem broja k na manji broj proces klasifikacije može se ubrzati.

Veliki problem pri klasificiranju predstavljaju ekstremi ali se pravim definiranjem parametara klasifikatora utjecaj tih ekstrema na određivanje klasa ostalih uzoraka da anulirati. Veći je problem određivanje klase samog ekstrema. Puno bolji rezultati postižu se na skupovima koji imaju veću razliku između klasa nego kod skupova malih razlika među klasama.

Ukoliko se klasifikator specijalizira za klasificiranje samo jedne baze podataka mogu se postići bolji rezultati. Klasifikator se može napraviti i tako da nemaju svi atributi jednak utjecaj na određivanje klase. U aplikaciji se mogu dodati i moduli za računanje dodatnih parametara kvalitete klasifikatora poput ROC krivulje i tablice zabune. Kako klasifikacija ima sve veću primjenu razvijen je i alat za klasificiranje WEKA.

LITERATURA

[1] G. P. Zhang: Neural Networks for Classification, IEEE Transactions on Systems, man, Cybernetics, 4, November 2000.

[2] D. Ivanković: Osnove statističke analize za medicinare, Zagreb, Medicinski Fakultet, 1988.

[3] A. K. Jain, R. P.W. Duin, J. Mao: Statistical Pattern Recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, NO. 1, January 2000.

[4] S. Garcia, J. Luengo, F. Herrera: Data Preprocessing in Data Mining, Springer Internacional Publishing Switzeland 2015.

[5] A. K. Jain, R. C. Dubes, C. C. Chen: Bootstrap Techniques for Error Estimation, IEEE Transactions on Pattern Analysis and machine Intelligence, vol. PAMI-9, NO. 5, 1987.

[6] T. N. Phyu: Survey of Classification Techniques in Data Mining, Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I IMECS 2009, March 18 - 20, 2009, Hong Kong.

[7] D. L. Sackett, R. B. Haynes: Evidence Base of Clinical Diagnosis: The architecture of diagnostic research, BMJ. 2002;324:539-41.

[8] J. Eggermont, A. E. Eiben, J. I. van Hemert: A comparison of genetic programming variants for data classification, Advances in intelligent data analysis, Lecture Notes in Computer Science, Volume 1642, Leiden University, Leiden, 1999.

[9] D. G. Altman: Practical Statistics for Medical Research. London. Chapman & Hall, 1991.

[10] A. Ali, O. Magnor, M. Schultalbers: Misfire Detection Using a Neural Network Based Pattern Recognition, International Conference on Artificial Intelligence and Computational Intelligence , Vol.2

[11] I. Saini, D. Singh, A. Khosla: QRS detection using K-Nearest Neighbor algorithm (KNN) and evaluation on standard ECG databases, Cairo University Journal of Advanced Research, Cairo 2012.

[12] J. Hopcroft i J. Ullman: Introduction to Automata Theory, Languarges, and Computation, Addison-Wesley Publishing Company, 1979.

[13] N. Bhatia: Survey of Nearest Neighbor Techniques, International Journal of Computer Science and Information Security, Vol. 8, No. 2, 201.

[14] L. Jiang, Z. Cai, D. Wang, S. Jiang: Survey of Improving K-Nearest-Neighbor for Classification, Faculty of Computer Science, China University of Geosciences, Wuhan

430074, China.

[15] C. Sammut, G. I. Webb. Encyclopedia of Machine Learning adresa: https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-30164-8_506 (pogledano 23.08.2021)

[16] K. Saravanan, S. Sasithra: Review on Classification Based on Artifical Neural Networks, International Journal of Ambient Systems and Applications (IJASA) Vol.2, No.4, December 2014.

[17] A. Bernal: Machine Learning and SVM Classification, ASc Program - ID:500134620.

[18] UCI Machine learning repository, adresa: http://archive.ics.uci.edu/ml/index.html (pogledano 15.09.2021)

[19] R. Kohavi: A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection Artificial Intelligence Proceedings 14th International Joint Conference, 20 -- 25. August 1995, Montréal, Québec, Canada, 1995.

[20] N. Bhatia: Survey of Nearest Neighbor Techniques, International Journal of Computer Science and Information Security, Vol. 8, No. 2, 201.

[21] R. Kohavi: A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the 14th International Joint Conference on Artificial Intelligence, Volume 2 , San Francisco, CA, USA, pp. 1137–1143. Morgan Kaufmann Publishers Inc, 1995.

[22] M. Zhua, W. Chena, J. P. Hirdesb, P. Stoleeb: The K-nearest neighbor algorithm redicted rehabilitation potential better than current Clinical Assessment Protocol, Journal of Clinical Epidemiology 60 (2007) 1015-1021.

[23] S. S. Dadhania, J. S. Dhobi: Improved kNN Algorithm by Optimizing Cross-validation, International Journal of Engineering Research & Technology (IJERT), Vol. 1 Issue 3, 2012.

[24] C. C. Bojarczuk: A constrained-syntax genetic programming system for discovering classification rules: application to medical data sets, Artificial Intelligence in Medicine, 30(1), 2004.

[25] G. P. Zhang: Proposed the use of harmony search and back propagation based ANN to classify breast cancer data, International Journal of Ambient Systems and Applications, 2000.

SAŽETAK

Klasifikacija je metoda učenja, shvaćanja i razumijevanja. Nalazi se u svim sferama života. Sve što postoji se klasificira. Klasifikacija je kategoriziranje uzoraka u skupine koji imaju neka zajednička obilježja koji ih razlikuju od drugih. Klasifikacija u medicini ubrzava proces obrade podataka te pomaže pri donošenju odluka. Klasifikatori su alati za klasificiranje. Razlikuju se po načinu djelovanja i izvedbi, iako im je cilj isti.

U radu je napravljen klasifikator baziran na mjerama udaljenosti k najbližih susjeda i Bayesov klasifikator koji nalaze primjenu u medicini kao i u drugim granama znanosti. Provedeni su testovi za ispitivanje rada klasifikatora. Objašnjeni su problemi koji nastaju prilikom klasifikacije te prilikom izrada samih klasifikatora.

Ključne riječi: algoritam k najbližih susjeda, atribut, klasa, klasifikacija, klasifikator, mjere, točnost udaljenosti

ABSTRACT

Expert system for data classification

Classification is a method of learning, understanding, and comprehension. It is located in all spheres of life. Everything that exists is classified. The classification is to categorize the group samples with some common characteristics that distinguish them from others. Classification in medicine speeds up the process of data processing helps in decision making. Classifiers are tools for classification. They differ in the way of activity and method, but their goal is the same. The classifier in this paper is made based on measures of distance k nearest neighbours and the Bayesian classifier used in medicine and other branches of science. Tests have been conducted to examine the work of the classifier. Explains the problems that arise when classifying and making the classifier.

Key words: accuracy, attribute, class, classification, classifier, distance measure, k-nearest neighbours algorithm.

ŽIVOTOPIS

Marin Gaće, rođen 20. travnja 1984. u Osijeku. od oca Branka i majke Rosande. Prvi razred OŠ upisuje 9. rujna 1991. u Umagu u osnovnoj školi „Maria i Lina“, a završava ga u OŠ „Stjepan Filipović“ u Opuzenu 1992. godine. Drugi razred osnovne škole upisuje u Osijeku u OŠ „Tin Ujević“ 1992. godine. Osnovnu školu završava s odličnim uspjehom. Potom završava III. prirodoslovno-matematičku gimnaziju u Osijeku. Školovanje nastavlja na Elektrotehničkom Fakultetu u Osijeku, na kojem završava preddiplomski studij računarstva. Na istom fakultetu upisuje diplomski studiji računarstva smjer procesno računarstvo. Tijekom osnovnoškolskog i srednjoškolskog obrazovanja sudjeluje na županijskim natjecanjima iz matematike, fizike i biologije.

Marin Gaće _____

potpis