|  |  |  |
| --- | --- | --- |
|  | Univerzitet u Beogradu - Elektrotehnički fakultet  Katedra za signale i sisteme |  |



**Izveštaj**

**Primena stabala odlučivanja na dijagnostiku dijabetesa**

**Kandidat**

Aleksandar Cvetković, 2018/0323

**Mentor**

Doc. dr Aleksandra Krstić

Beograd, *decembar* 2022. godine

# Sadržaj

[Sadržaj 2](#_Toc121782881)

[1 Uvod 4](#_Toc121782882)

[2 Analiza postojećih istraživanja 6](#_Toc121782883)

[3 Metodologija rada 7](#_Toc121782884)

[3.1 Podela skupa podataka 8](#_Toc121782885)

[3.2 Funkcije nečistoće 8](#_Toc121782886)

[3.2.1 Entropija – IH 9](#_Toc121782887)

[3.2.2. Gini nečistoća 10](#_Toc121782888)

[3.2.3. Greška klasifikacije – IE 11](#_Toc121782889)

[3.2.4. Hi-kvadrat test – Ichi 12](#_Toc121782890)

[3.3. Diskretizacija kontinualnih promenjivih 14](#_Toc121782891)

[3.3.1. Podela na skupove jednakih veličina 14](#_Toc121782892)

[3.3.2. Podela na skupove sa jednakim brojem primera u njima 15](#_Toc121782893)

[3.3.3. Dinamička podela atributa na skupove 15](#_Toc121782894)

[3.4. Metode za sprečavanje preobučavanja stabla 16](#_Toc121782895)

[3.4.1. Ograničavanje maksimalne dubine stabla: 17](#_Toc121782896)

[3.4.2. Ograničavanje minimalnog broja primera u čvoru za podelu: 17](#_Toc121782897)

[3.4.3. Orezivanje stabla za smanjenje greške (*Reduced Error Pruning)* 17](#_Toc121782898)

[3.4.4. Pesimistično orezivanje stabla: 18](#_Toc121782899)

[3.5. Nedostajući podaci 20](#_Toc121782900)

[4 Rezultati 21](#_Toc121782901)

[5 Diskusija i zaključak 22](#_Toc121782902)

[Literatura 23](#_Toc121782903)

# 1 Uvod

Tumor mozga predstavlja jedan od najčešćih vrsta tumora koji se javljaju u Srbiji i deseti je po broju prouzrokovanih smrti. Prema podacima instituta za javno zdravlje Srbije „Dr Milan Jovanović Batut“ tumori mozga su sa **52.9%** najčešći razlog smrti uzrokovane tumorom kod dece do 15 godina. Zbog dugoročnih zdravstvenih problema koje može izazvati ukoliko ostane neprimećen, tumor mozga je nužno otkriti i ispravno odrediti njegov položaj u što ranijem stadijumu. Ove činjenice jasno ukazuju da je tumor mozga ozbiljan problem današnjice i da je potrebno podići svest o važnosti pregleda kao i da same preglede treba učiniti kvalitetnijim i dostupnijim.

Tumor mozga specifičan je po tome što za njegovo otkriće ne može koristiti krvna slika i što još uvek nema poznatih tumor markera koji bi otkrili njegovo postojanje. Simptomi se neretko mogu ispoljiti samo jednom i usled velikog broja bolesti koje ispoljavaju slične simptome teško ih je koristiti za dijagnostiku. Zbog svog položaja unutar lobanje spoljašnji pregled je u potpunosti nemoguć dok je snimak rendgenom u osnovnom obliku od male koristi. Najpouzdanija metoda za dijagnostiku tumora mozga jeste skrining, koji se zbog posebne strukture mozga u najčešćem broju slučajeva sprovodi uz pomoć magnetne rezonance [1].

Magnetna rezonanca pogodna je za detekciju i klasifikaciju tumora pre svega zbog svoje preciznosti. Pored toga jako je bitno napomenuti i odsustvo jonizujućeg zračenja kao i isključenje rizika od svake vrste alergije. Još jedna jako bitna karakteristika magnetne rezonance je i što nudi rekonstrukciju snimaka u više ravni i omogućava 3D modelovanje položaja raka čime daje jasan uvid u njegov oblik i dimenzije. Ovakvo 3D modelovanje može biti jako korisno u slučajevima gde nije lako utvrditi mesto odakle će se pristupiti odstranjivanju raka: zbog osetljivosti mozga na ovakve vrste zahvata, pre operacije je neophodno napraviti detaljan plan pristupa mestu na kome se rak nalazi kako bi se smanjilo oštećenje okolnog tkiva. Pokazuje se da je ispravnim planiranjem operacije moguće smanjiti šansu za komplikacije prilikom zahvata na ispod **1%**.

Za preciznu dijagnostiku vrste i pozicije tumora potreban je iskusan radiolog. U Srbiji, prihvaćeni su programi skrininga tumora grlića materice, dojke i debelog creva, međutim program skrininga za rak mozga još uvek nije prihvaćen. Program skrininga mogao bi da bude sproveden bez dodatnog opterećivanja radiologa uz korišćenje veštačke inteligencije za automatsko čitanje snimaka magnetne rezonance.

Cilj ovog rada je da obezbedi model za segmentaciju snimaka magnetne rezonance mozga pri čemu će svakom pikselu sa slike biti pridružena vrednost 1 ukoliko on predstavlja ćeliju raka ili vrednost 0 ukoliko on ne predstavlja ćeliju raka. Pri tome je neophodno da uspešnost ove segmentacije bude u rangu uspešnosti koju pri otkrivanju raka imaju stručni radiolozi. Na ovaj način obrada snimaka magnetne rezonance postala bi automatizovana uz očuvanje preciznosti. U narednim sekcijama biće detaljnije opisan proces segmentacije i principi korišćeni za njeno unapređenje.

Pregled radova u kojima su opisane različite vrste metoda za semantičku segmentaciju tumora mozga dat je u drugom poglavlju.

U trećem poglavlju predstavljena je sva metodologija i analiza algoritama i arhitektura korišćenih u ovom radu. Na početku poglavlja objašnjeni su opšti principi funkcionisanja neuralnih mreža i konvolucionih neuralnih mreža. Zatim je izložen problem semantičke segmentacije i predstavljena je arhitektura UNet mreže. U narednom delu izložene su korišćene funkcije gubitaka i metrike, zatim sledi izlaganje tehnika augmentacije i ansambla dok je na kraju dat uvid u skup podataka korišćen za trening.

U četvrtom poglavlju izloženi su rezultati dobijeni različitim metodama opisanim u ovom radu i zatim je najbolji od dobijenih rezultata upoređen sa rezultatima iz radova navedenih u drugom poglavlju.

U petom poglavlju odrađena je diskusija korišćenja postignuća iz ovog rada u praksi. Dati su konkretni primeri segmentacije gde korišćeni postupak radi dobro i primeri gde postupak ima lošije performanse. Predstavljene su mogućnosti za dalje unapređenje postignutih rezultata i mogućnosti za dalji nastavak istraživanja.

# 2 Analiza postojećih istraživanja

Postoje različita istraživanja na polju segmentacije tumora mozga.

Klasičan način segmentacije slike u oblasti mašinskog učenja obuhvata izvlačenje karakteristika primenom različitih metoda za obradu slike, pri čemu se koriste različite karakteristike za različite vrste tkiva. Ove karakteristike mogu da predstavljaju intenzitete vrednosti piksela, pojava lokalnih tekstura, razlike u intenzitetu, karakteristike koje opisuju asimetriju slike i statistički parametri slike. Najbolji metod koji koristi ovaj tip karakteristika opisan je u radu koji su napisali Tustison i saradnici [2]. Ove karakteristike korišćene su za trening konkateniranih modela slučajnih šuma, pri čemu je rezultat jednog modela prosleđivan kao ulaz sledećeg.

U modernim sistemima iz oblasti mašinskog učenja za proces segmentacije koriste se duboke neuralne mreže. Njihova velika prednost je u tome što za njihovo korišćenje nije neophodno znanje eksperata za izvlačenje karakteristika tkiva, već se ove karakteristike otkrivaju automatski [3]. Shodno tome razlika između ovih modela ne ogleda se u korišćenim karakteristikama već u njihovoj arhitekturi. Urban i saradnici [4] su u svom radu opisali jednu ovakvu arhitekturu koja kao ulaz koristi 3D kocke voksela. Ovi vokseli daju se na ulaz četvoroslojnoj konvolucionoj neuralnoj mreži koja predviđa labelu središnjeg voksela. Ovakvo 3D predstavljanje ulaznih parametara dobro oslikava trodimenzionalne zavisnosti između ćelija raka koje se javljaju i u prirodi, ali je zbog svoje visokodimenzionalnosti ovaj postupak izuzetno računarski zahtevan.

Žikić i saradnici [5] postigli su znatno veću računarsku efikasnost korišćenjem 2D modela konvolucionih neuralnih mreža. Velika razlila između ovog i prošlog modela je i u tome što je ovde korišćena ReLU aktivaciona funkcija dok je do tada za proces segmentacije uglavnom korišćena aktivaciona funkcija tanh.

Moderan pristup problemu semantičke segmentacije ogleda se u korišćenju dubljih konvolucionih neuralnih mreža sa manjom veličinom filtera. Pereira i saradnici [6] u svom radu opisali su konvolucionu neuralnu mrežu sa 11 slojeva koja koristi filtere veličine 3x3 piksela. U ovom radu umesto ReLU aktivacione funkcije korišćena je njena modifikacija leaky ReLU što je smanjilo ispoljavanje problema odumiranja neurona. U ovom radu primenjene su i tehnike postprocesiranja gde su metodama klasterizacije finije podešene granice između čistog tkiva i tkiva tumora.

# 3 Metodologija rada

Stabla odlučivanja su model mašinskog učenja za analizu podataka i predikciju koja se zasniva na postavljanju niza pitanja. Podaci za analizu predstavljeni su u vidu skupa podataka u kom svaki primer ima svoje odlike (osobine, prediktore) i labelu (oznaku) koja predstavlja ciljnu vrednost koju bismo želeli da predvidimo. U svakom koraku skup podataka se grana na više delova koji se nazivaju čvorovi. u zavisnosti od vrednosti svojih odlika. Grananje prestaje kada nema više odlika na osnovu kojih bismo dalje delili skup ili kada metrike performansi stabla budu dostignute. Čvorovi koji nemaju daljih grana i vraćaju oznaku klase koju označavaju nazivaju se **listovima**.

Stabla odlučivanja mogu se koristiti za rešavanje problema regresije – kada je ciljna promenjiva realan broj i klasifikacije – kada ciljna promenjiva iz diskretnog skupa uzima vrednost koja predstavlja njenu klasu. Tematika ovog rada baziraće se na upotrebi stabala odlučivanja za potrebe klasifikacije.

Stabla odlučivanja spadaju u nadgledane modele mašinskog učenja, što znači da nam za svaki primer u našem skupu podataka trebaju odlike datog primera kao i labela koja označava ispravnu vrednost ciljne promenjive.

Ono što čini stabla odlučivanja posebno pogodnim za primenu u medicini predstavlja njihova izuzetna interpretabilnost koja omogućava lekarima da na osnovu formiranog stabla donesu zaključke i uvide veze između odlika pacijenata kojih pre nisu bili svesni.

Na slici 1 možemo da vidimo primer stabla koje se koristi za dijagnostiku dijabetesa:

Diagram

Description automatically generated

Slika 1. Primer stabla odlučivanja za dijagnostiku dijabetesa

Semantička segmentacija predstavlja kompleksan problem u kom je neophodno da model nauči da iz vrednosti piksela prepozna ćelije tumora na slici. Funkcija za prepoznavanje ovakvih vrednosti izuzetno je složena i za učenje ovakve funkcije neophodan je model dovoljne složenosti. U narednom poglavlju najpre će u [delu 3.1](#_3.1_Neuralne_mreže) biti opisane neuralne mreže kao model koji je u stanju da opiše funkcije velike složenosti. Zatim će u [delu 3.2](#_3.2_Obrada_slike) biti opisane konvolucione neuralne mreže kao unapređenje modela neuralnih mreža specijalizovano za rad sa slikama. U [delu 3.3](#_3.3_Semantička_segmentacija) biće detaljnije opisan problem semantičke segmentacije i arhitektura UNet mreže koja se primenjuje u ovoj oblasti. U [delu 3.4](#_3.4_Funkcije_gubitka) biće predstavljene funkcije gubitka koje su specijalizovane za problem semantičke segmentacije i metrike koje se koriste za praćenje uspešnosti modela. U [delu 3.5](#_3.5_Tehnike_za) opisane su tehnike za rešavanje problema preobučavanja koji je neizostavni problem koji se javlja pri treniranju konvolucionih neuralnih mreža. Na kraju, u [delu 3.6](#_3.6_Skup_podataka) opisan je korišćeni skup podataka zajedno sa tehnikama pretprocesiranja korišćenim nad tim skupom podataka.

## 3.1 Podela skupa podataka

Cilj stabala odlučivanja je konstrukcija stabla koje na datom skupu podataka ostvaruje najbolju moguću tačnost. Teoretski, ukoliko bismo dozvolili stablu odlučivanja da se neograničeno grana ono bi moglo da postigne stopostotnu tačnost tako što bi dugom serijom pitanja savršeno razvojila primerke različitih klasa u posebne čvorove (ovo je moguće ukoliko u skupu podataka nema izuzetnih slučajeva gde su dva primera istih odlika označeni različitom klasom). Ipak, iako bi ovakvo stablo postiglo zavidan nivo tačnosti, ta tačnost bi se zasnivala na prevelikom prilagođavanju primerima u skupu podataka umesto na izvlačenju generalnih zaključaka koji bi nam koristili i za do tad neviđene primere. Ova pojava se u mašinskom učenju naziva **preobučenost**. Da bismo detektovali preobučenost pri konstrukciji stabla, skup podataka delimo na dva dela: jedan skup nad kojim će se vršiti konstrukcija stabla, koju nazivamo i treningom, i drugi skup koji neće učestvovati u treningu već će nam služiti za proveru performansi stabla na primerima koji do tada nisu viđeni. Ove skupove podataka nazivamo trening skup i validacioni skup i glavni indikator preobučenosti biće velika tačnost modela na trening skupu i mala tačnost modela na validacionom skupu podataka.

## 3.2 Funkcije nečistoće

U saglasnosti sa načinom podele predstavljenim u prethodnom poglavlju cilj rada modela jeste stablo koje ostvaruje maksimalnu tačnost na validacionom skupu podataka. Da bismo obezbedili visoke performanse na validacionom skupu i sprečili preobučavanje potrebno je veliku pažnju posvetiti redosledu odlika na osnovu kojih se vrši grananje unutar stabla. Jedan od načina da izaberemo redosled odlika za grananje je da probamo svaku moguću kombinaciju redosleda odlika. Iako bi ovakav postupak sa sigurnošću doveo do optimalnog rešenja on je izuzetno vremenski skup i postao bi nemoguć za skupove podataka sa velikim brojem odlika.

Umesto isprobavanja svake moguće kombinacije, stabla odluke u svakom koraku biraju odliku koja odvaja primere u grane koje su u najvećoj meri „**čiste**” tj. sastoje se iz većeg dela primera iste klase. Za procenu „čistoće” grana koriste se funkcije koje se nazivaju **funkcije čistoće**. U daljem delu biće predstavljene najčešće korišćene funkcije čistoće u stablima odlučivanja.

### 3.2.1 Entropija – IH

U toeriji informacija, entropija predstavlja meru neodređenosti slučajne promenjive. Primenjena na stabla odlučivanja, onda predstavlja meru nečistoće nakon grananja i definisana je na sledeći način:

gde **t** predstavlja oznaku čvora, a **p(k|t)** predstavlja proporciju primera koji pripadaju klasi **k** u čvoru **t.**

Shodno ovoj definiciji, entropija je minimalna i iznosi 0 kada svi primeri u čvoru pripadaju istoj klasi (tj. čvor je „čist“). Entropija ima maksimalnu vrednost kada u čvoru ima podjednako mnogo primera različitih klasa. Koristeći entropiju kao funkciju čistoće, minimizacija entropije direktno bi obezbedila čistoću listova nakon grananja.

Za binarnu promenjivu koja uzima vrednost 1 sa verovatnoćom p dat je sledeći grafik vrednosti entropije:

A picture containing diagram

Description automatically generated

Slika 2. Vrednost entropije u zavisnosti od verovatnoće p binarne promenjive. Izvor: [wikipedia](https://en.wikipedia.org/wiki/Binary_entropy_function)

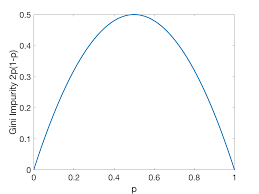
### 3.2.2. Gini nečistoća

Gini nečistoće je funkcija nečistoće koja je jako slična entropiji i takođe se zasniva na proporciji primera različitih klasa unutar čvora. Definisana je na sledeći način:

gde su oznake iste kao u jednačini entropije.

Gini nečistoća takođe ima minimalnu vrednost kada se u listu nalaze primeri samo jedne klase, dok je maksimalna kada se u listu nalazi jednak broj jedinki različitih klasa. Po vrednostima funkcija, Gini nečistoća je veoma slična skaliranoj verziji entropije i često prelazak sa jedne funkcije nečistoće na drugu ne dovodi ni do kakvih promena. Međutim, zbog lakšeg izračunavanja Gini nečistoće od entropije (za koju je neophodno nalaziti vrednosti logaritma sa osnovom 2) Gini nečistoće se češće primenjuje.

Za binarnu promenjivu koja uzima vrednost 1 sa verovatnoćom p dat je sledeći grafik vrednosti Gini nečistoće (skaliran sa 2):



Slika 3. Vrednost Gini nečistoće u zavisnosti od verovatnoće p binarne promenjive. Izvor: [cornell](https://www.cs.cornell.edu/courses/cs4780/2022sp/notes/Notes20.pdf)

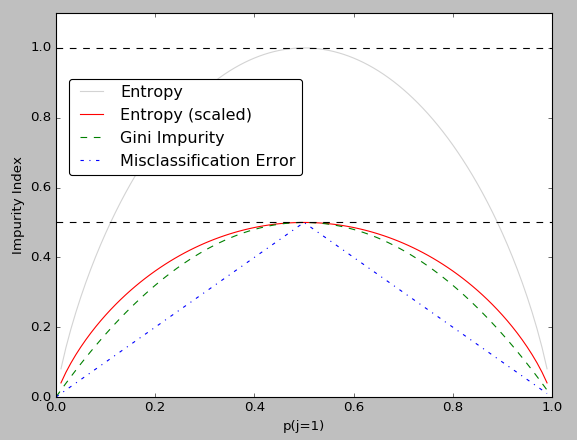
### 3.2.3. Greška klasifikacije – IE

Greška klasifikacije je funkcija nečistoće koja uzima u obzir samo proporciju primera najčešće klase u čvoru. Pošto će list koji nije čist imati oznaku upravo najčešće klase u njemu, onda će proporcija primera ostalih klasa predstavljati **grešku klasifikacije**.

Ona je definisana sledećom formulom:

Iako je jednostavna za računanje, greška klasifikacije nije dobar indikator čistoće čvora zbog manje osetljivost na promene u klasnim proporcijama unutar čvora. Ipak, ona svoju primenu nalazi prilikom postupaka potkresivanja.

Poređenje funkcija nečistoće: entropije, Gini i greške klasifikacije dato je na sledećoj slici:



Slika 4. Vrednosti funkcija nečistoće u zavisnosti od verovatnoće p binarne promenjive. Izvor: [scikit-learn](https://www.bogotobogo.com/python/scikit-learn/scikt_machine_learning_Decision_Tree_Learning_Informatioin_Gain_IG_Impurity_Entropy_Gini_Classification_Error.php)

### 3.2.4. Hi-kvadrat test – Ichi

Hi-kvadrat test predstavlja funkciju za poređenje grupa u odnosu na kategorijalnu zavisnu promenjivu. Ova funkcija opisuje razliku između posmatrane frekvencije - **fo** i očekivane frekvencije - **ft.** Očekivanu frekvenciju definišemo kao frekvenciju koja bi mogla da se očekuje potpuno slučajno, u skladu sa uslovnim verovatnoćama svakog ishoda.

Očekivanu frekvenciju za svaku od ćelija tabele određujemo proizvodom marginalnih frekvencija (sume odgovarajućeg reda i kolone) podeljenim ukupnom veličinom uzorka tj.:

Vrednost Hi-kvadrat testa određuje se kao suma odstupanja očekivanih od osmotrenih vrednosti frekvencija u svakoj ćeliji tj.:

gde r predstavlja odgovarajući red, k odgovarajuću kolonu, a N ukupan broj uzoraka.

Posmatrajmo sledeću sliku:

Chart, bubble chart

Description automatically generated

Slika 5. Tabela zadovoljstva studenata uslugama studentskog restorana gde nema razlike u raspodeli. Izvor: [filozofski fakultet](http://psihologija.ff.uns.ac.rs/viz/hi-kvadrat-test)

Na slici je data tabela čiji redovi predstavljaju fakultete: Filozofski fakultet (FF) i Prirodno-matematički fakultet (PMF). U kolonama je izraženo zadovoljstvo studenata uslugama studentskog restorana: 1-nezadovoljni, 2-zadovoljni, 3-vrlo zadovoljni. Iako je više studenata sa FF-a učestvovalo u anketi možemo da primetimo da su međusobni odnosi broja studenata u sve tri kategorije zadovoljstva jednaki.

Iz datog primera možemo da vidimo da ni za jednu ćeliju tabele nema razlike između posmatrane i očekivane frekvencije. Na primer, za ćeliju FF1 očekivana frekvencija bila bi 15x70/105=10 što se upravo slaže sa sadržajem ćelije. Samim tim suma odstupanja vrednosti očekivanih i osmotrenih frekvencija u datoj tabeli je 0 što bi značilo da ne postoje statistički značajne razlike između distribucija studenata sa FF-a i PMF-a.

U kontekstu stabala odluke ovakva situacija odgovarala bi slučaju gde smo nakon izvršene podele po zadovoljstvu studenata uslugama restorana uvideli da novonastale grupe imaju podjednak odnos frekvencija studenata sa FF-a i PMF-a kao što je ona bila i pre podele tj. grupe nastale podelom po zadovoljstvu nisu postale čistije. Iz ovog primera možemo da izvučemo zaključak da bi manje vrednosti Hi-kvadrat testa odgovarale lošijoj podeli po datom atributu u stablu odlučivanja.

Na sledećoj slici dat je primer kada između grupa postoje znatne statističke razlike u distribuciji studenata:

Chart, bubble chart

Description automatically generated

Slika 6. Slika 5. Tabela zadovoljstva studenata uslugama studentskog restorana gde ima razlike u raspodeli. Izvor: [filozofski fakultet](http://psihologija.ff.uns.ac.rs/viz/hi-kvadrat-test)

Na slici 6. možemo da vidimo da u ćelijama u prvoj i trećoj koloni postoje znatne statističke razlike u distribuciji studenata. Na primer očekivana frekvencija studenata u ćeliji FF1 iznosi: 50x60/120=25 ali je njena stvarna vrednost 10. Ovakve razlike akumuliraju se u Hi-kvadrat testu čija vrednost sada iznosi 36. U kontekstu stabala odluke ovakva situacija odgovarala bi slučaju gde smo nakon izvršene podele po zadovoljstvu studenata uslugama restorana uvideli da novonastale grupe veoma različit odnos frekvencija studenata sa FF-a i PMF-a nego što je ona bila i pre podele tj. grupe nastale podelom po zadovoljstvu su **čistije**. Iz ovog primera možemo da izvučemo zaključak da bi veće vrednosti Hi-kvadrat testa odgovarale boljoj podeli po datom atributu u stablu odlučivanja.

Kao što smo videli iz datih primera Hi-kvadrat test može se efikasno koristiti da opiše kvalitet podele po određenom atributu u stablu, pri čemu veće vrednosti označavaju bolju podelu po datom atributu. Hi-kvadrat test se, osim za izbor atributa za podelu tokom konstrukcije stabla može koristiti i za potkresivanje stabala o čemu će biti reči u daljim poglavljima.

## 3.3. Diskretizacija kontinualnih promenjivih

Pri grananju skupa na delove na osnovu vrednosti atributa podrazumevali smo da su te vrednosti kategoričke tj. da uzimaju vrednosti iz diskretnog skupa: kao što studenti mogu da se izjasne za jednu od tačno tri vrednosti zadovoljstva: nezadovoljan, zadovoljan, vrlo zadovoljan. Međutim, prilikom primena stabla odlučivanja na medicinske podatke atributi često mogu imati vrednosti koje predstavljaju cele ili realne brojeve: broj godina, krvni pritisak, visina, težina, koncentracija glukoze u krvi itd.. Posmatranje ovih vrednosti na isti način kao i kategoričke vrednosti dovelo bi do toga da skoro svaki pacijent ima svoju vrednost datog atributa što bi dovelo do toga da se model prilagođava svakom posebnom pacijentu, umesto da sagledava generalnu sliku. Ovakav postupak doveo bi do preobučavanja. Da bismo sprečili preobučavanje, potrebno je da realnu vrednost pretvorimo u kategoričku primenom nekog od algoritama diskretizacije: na primer umesto da visinu izražavamo u centimetrima mogli bismo da je podelimo na dve kategorije ili podeoke: niži od 180cm i viši od 180cm. U daljem delu ovog poglavlja biće objašnjeni načini za diskretizaciju kontinualnih promenjivih.

### 3.3.1. Podela na skupove jednakih veličina

Prvi i najjednostavniji vid podele zasniva se na tome da podelimo skup mogućih vrednosti na delove jednakih veličina. Na primeru visine izražene u centimetara podela na skupove jednakih veličina bila bi određena na sledeći način:

cm

190

210

170

150

Visina ispitanika kreće se u granicama od 150cm do 210cm. Odabraćemo da broj skupova za podelu bude 3. Na ovaj način podela će biti izvršena na skupove: niži od 170cm, između 170 i 190cm i viši od 190cm.

Ovaj način podele je lak za implementaciju i izuzetno brz za realizaciju ali se može dogoditi da za pojedine podeoke ne postoji ni jedan primer koji im pripada (u našem primeru nije se pojavio ni jedan ispitanik visine između 170 i 190). Verovatnoća javljanja ovakvog podeoka raste sa povećanjem broja podeoka. Takođe, neophodno je precizno podesiti broj podeoka za svaki atribut.

### 3.3.2. Podela na skupove sa jednakim brojem primera u njima

Ovaj način podele podrazumeva da skup vrednosti bude podeljen na podeoke od kojih će svaki sadržati približno jednak broj vrednosti. Na primeru visine izražene u centimetara podela na skupove sa jednakim brojem primera bila bi određena na sledeći način:

cm

170

160

210

150

Visina ispitanika kreće se u granicama od 150cm do 210cm. Odabraćemo da broj skupova za podelu bude 3. Na ovaj način podela će biti izvršena na skupove: niži od 160cm, između 160 i 170 i viši od 170cm.

Na ovaj način nema skupova koji u sebi ne sadrže ni jedan primer ali je i dalje potrebno podesiti broj podeoka za svaki atribut.

### 3.3.3. Dinamička podela atributa na skupove

Dinamička podela atributa na skupove rešava problem izbora broja podeoka time što sama određuje podeoke na koje skup podataka treba da bude podeljen. Ovaj vid podele skupa pored vrednosti atributa primera uzima u obzir i njihovu klasu.

Algoritam deljenja funkcioniše na sledeći način:

* Inicijalno svaki od primera u datom skupu predstavlja po jedan podeok – ovi podeoci nazicaju se „**slabim**”
* Zatim se slabi podeoci iste klase koji se nalaze jedan pored drugog grupišu u veće podeoke koji se nazivaju „**jakim**” podeocima
* Na kraju se iterativno svi slabi podeoci koji se nalaze između dva jaka podeoka iste klase grupišu zajedno sa ta 2 jaka podeoka u još veći jaki podeok

Rad algoritma za dinamičku podelu predstavljen je po koracima na sledećoj slici:

Diagram

Description automatically generated

Slika 7. Primer rada algoritma za dinamičku podelu kontinualnog skupa vrednosti. Izvor: rad – dodati

Velika prednost ovog načina podele je što nema potrebe za eksplicitnim podešavanjem parametra broja podeoka. Ipak, ovaj algoritam pokazuje slabije rezultate u višeklasnoj klasifikaciji gde je broj klasa veliki jer uzrokuje manje podeoke.

## 3.4. Metode za sprečavanje preobučavanja stabla

Kao što je bilo reči u poglavlju 3.1. stabla odlučivanja čiju konstrukciju ne ograničimo razgranaće se tako da će se povinovati svakom dostupnom primeru u skupu pdataka za trening. Ovaj postupak će učiniti da greška na trening skupu bude mala, jer mu je stablo prilagođeno, ali će greška na validacionom skupu biti velika. S’ obzirom da validacioni skup podataka predstavlja primere koje stablo nije videlo prilikom procesa treninga ovaj skup mnogo bolje oslikava dešavanja u praktičnoj primeni stabla, kada dobijemo nove podatke od novog pacijenta kakve stablo do tada nije videlo. Zbog osobine stabla da se prilagođava svakom trening skupu koji dobije, stabla će se veoma razlikovati po svojoj konstrukciji u zavisnosti od toga kakvi podaci se nalaze u trening skupu. Zbog velike razlike u izgledu stabla u odnosu na podatke u trening skupu kažemo da ovaj algoritam ima **veliku varijansu.**

U daljem delu ovog poglavlja biće reči o tehnikama za rešavanje problema preobučavanja stabla. Prvo će biti izložene tehnike za ograničavanja grananja stabla tokom njegove konstrukcije a zatim će biti obrađeni postupci za smanjenje dubine stabla nakon njegove konstrukcije metodama potkresivanja (orezivanja).

Sprečavanjem stabla da raste neograničeno, listovi stabla se više neće sastojati od primera samo jedne klase čime će prestati da budu čisti. Za ovakve listove klasa će biti određena klasom primera kojih ima najviše u datom listu.

### 3.4.1. Ograničavanje maksimalne dubine stabla:

U poglavlju 3.4. objašnjena je tendencija stabla odlučivanja da se grana sve dok se ne povinuje svakom pojedinačnom primeru iz trening skupa, što dovodi do preobučavanja. Ideja ograničavanja maksimalne dubine stabla je da spreči neograničeno grananje stabla čime će smanjiti preobučavanje. Ovo ograničavanje sprovodi se preko parametra **max\_depth,** koji označava maksimalnu dubinu do koje puštamo stablo da raste. Na ovaj način podela će biti izvršena samo po najinformativnijih **max\_depth** atributa, što sprečava da stablo nakon toga krene da donosi odluke po neinformativnim atributima čije vrednosti nemaju direktne veze sa izborom klase.

### 3.4.2. Ograničavanje minimalnog broja primera u čvoru za podelu:

Slično kao ograničavanje maksimalne dubine stabla, i ova metoda pokušava da spreči grananje na osnovu neinformativnih atributa. Primenom ove metode, prestaće se sa daljim grananjem čvora koji sadrži manje od parametra koji označava minimalni broj primera. Ovo će dovesti do toga da se primeri koje nije moguće „razumno“ podeliti na posebne čvorove ne dele dalje koristeći neinformativne atribute čije vrednosti nemaju direktne veze sa izborom klase.

### 3.4.3. Orezivanje stabla za smanjenje greške (*Reduced Error Pruning)*

Orezivanje stabla za smanjenje greške je tehnika orezivanja za koju je neophodno obezbediti skup podataka uz pomoć kog će se vršiti orezivanje. Podaci iz ovog skupa se ne smeju koristiti za trening stabla. Shodno tome trening skup podataka koji stablo dobije podeliće se na podatke koji će stvarno biti korišćeni za trening i na podatke koji će biti korišćeni za orezivanje.

Postoji više verzija ovog algoritama koje su nastale kao implementacija opisa algoritma iz rada (dodati rad). U ovom radu biće korišćena verzija algoritma koja je u radu (dodati rad) bila ocenjena kao jedina konzistentna sa opisima algoritma u radu (dodati rad) ali i u ostalim radovima (dodati autora).

Algoritam funkcioniše na sledeći način:

* Čvorovi se obilaze u post-order redosledu obilaska
* Listovi se ne obrađuju
* Za svaki čvor koji nije list (čvor odluke) radi se sledeće:
  + Određuje se greška koju trenutno stablo pravi na skupu podataka za orezivanje
  + Čvor se proglašava za list, njegova deca se orezuju i za klasu čvora uzima se klasa primera kojih u datom čvoru ima najviše
  + Ispituje se kolika je greška koju stablo pravi na skupu podataka za orezivanje nakon orezivanja
  + Ako je greška nakon orezivanja manja nego pre orezivanja onda se orezivanje zadržava, u suprotnom stablo se vraća na prethodno stanje

Ovakav vid orezivanja stabla obezbeđuje da rezultujuće stablo postiže minimalnu grešku na skupu podataka za orezivanje i da od svih stabala koje postižu ovu minimalnu grešku ovo stablo bude sa minimalnim brojem čvorova. Problem ovog načina orezivanja je to što je za njegovu implementaciju potreban poseban skup podataka za orezivanje.

### 3.4.4. Pesimistično orezivanje stabla:

Jedna od navedenih mana orezivanja stabla za smanjenje greške je to što nam je za njegovu implementaciju potreban poseban skup podataka na kome će biti testirane performanse pre i posle orezivanja.

Pesimistično orezivanje stabla je postupak u kom možemo da procenimo performanse stabla pre i posle orezivanja na samom skupu za trening, bez potrebe da uvodimo poseban skup podataka. Međutim, korišćenjem trening seta za ispitivanje performansi na uobičajen način do orezivanja nikada ne bi došlo jer se treningom maksimizuju performanse uravo na datom trening skupu. U svom radu (dodati rad) (dodati autora) tvrdi kako je performansa stabla na trening setu previše optimistična zbog toga što je stablo videlo date primere u toku treninga. Samim tim ovu procenu je potrebno učiniti *pesimističnijom.* U pesimističnom orezivanju ovo je postignuto dodavanjem korekcije kontinuiteta za binomnu raspodelu što dovodi do bolje procene greške na trening skupu.

Neka je za čvor **t** broj primera u datom čvoru predstavljen sa **N(t)** i neka je **e(t)** broj pogrešno klasifikovanih primera u datom čvoru. Tada je greška u datom čvoru data sa:

Kada se u ovu procenu uključi i korekcija kontinuiteta dobija se:

Shodno tome, greška u klasifikaciji za podstablo T definisano je sa:

gde suma po **i** ide po svim listovima u datom podstablu**,** pri čemu je ∑N(i) = N(t)

Pa će greška u klasifikaciji za podstablo T uz korekciju biti data sa:

gde je sa **Nt** označen broj listova u podstablu **T**

Kako je ∑N(i) = N(t) imenioci izraza za r’(t) i r’(T) biće isti pa je dovoljno upoređivati brojioce koji predstavljaju broj pogrešno klasifikovanih primera (sa korekcijom) označenih sa **n(t)**:

Iz ovih jednačina se vidi da kada orežemo stablo mi ćemo broju pogrešno klasifikovanih primera dodati ½ samo jednom, dok ćemo u slučaju kada imamo podstablo mi dodati ½ za svaki list. Dakle postojanje svakog lista se dodatno naplaćuje, ovim je obezbeđeno da se stablo obreže ukoliko postojanje datih listova nije obrazloženo preciznijom klasifikacijom.

Shodno sa prethodnim definicijama podstablo ne bi trebalo orezati ako je n’(T) < n(t) tj. ako podstablo ima bolje performanse od orezanog čvora. Međutim, i pored korišćenja korekcije, greška podstabla je i dalje optimistična u odnosu na grešku koju bismo dobili kada bismo stablo primenili na podatke koji nisu viđeni u treningu. Zato ćemo sačuvati podstablo samo ako bismo njegovim očuvanjem postigli performanse bolje za jednu standardnu grešku bolje u odnosu na orezani čvor. Tj. stablo nećemo orezati ako je **n’(T) + SE(n’(t)) < n(t)** pri čemu je standardna greška definisana sa:

I pored ove korekcije procena greške podstabla je i dalje blago optimistična u odnosu na stvarne rezultate koje bi stablo postiglo na do tad neviđenim primerima.

Za razliku od orezivanja stabla za smanjenje greške, ovaj algoritam kreće rezanje od vrha stabla ka dnu. Ovo ga čini mnogo bržim jer kada se oreže podstablo koje je bliže vrhu više neće biti ispitivanja njegovih podstabala. Ipak, iteriranje kroz sva podstabla je nekad korisno jer se ekspertima može ponuditi serija stabala od kojih bi oni izabrali ono koje ima najviše fizičkog smisla.

## 3.5. Nedostajući podaci

Problem sa nedostajućim podacima javlja se u svim sferama analize podataka i mašinskog učenja. U oblasti primene modela u medicini nedostatak podataka je znatno češća pojava zbog toga što u specifičnim situacijama nije moguće obaviti sva neophodna merenja zbog fizičkog stanja pacijenta ili je vremenski nemoguće obaviti merenje. Postoje razni načini da se nedostajuće vrednosti procene iz ostatka podataka, ali je u ovom radu primenjen drugačiji pristup. Kada se u odluke desi situacija da nedostaje vrednost atributa u odnosu na koji se vrši podela u grane, put kroz stablo se nastavlja na sledeći način: grana za nastavak puta bira se nasumično, tako da grane koje sadrže veći broj primera imaju veću verovatnoću da budu odabrane.

Posmatrajmo sledeću sliku:

Diagram

Description automatically generated

Slika 8. Primer podstabla

Pretpostavimo da pokušavamo da odredimo klasu primera koji je došao do podstabla na slici 8. Neka za taj primer nedostaje informacija o indeksu telesne mase (*body weight mass – mbi).* Čvor **bmi** ima 3 sina sa ukupno 26+44+55=125 potomka.Verovatnoća nastavka puta do levog sina je 26/125, za srednjeg sina 44/125i za desnog sina 55/125.

# 4 Implementacija

# 5 Rezultati

U ovom poglavlju biće dati rezultati koji predstavljaju vrednosti funkcije gubitka i osnovne metrike koje će biti predstavljene grafički i tabelarno. Metrike IoU, dice i senzitivnost određivane su isključivo za slike na kojima ima tumora. Zbog konzistentnosti i poređenjima sa rezultatima dobijenim u drugim radovima ove metrike će na kraju biti odrađene za sve slike na skupu podataka za testiranje.

*Tabela 2. Prikaz metrika na skupu podataka za trening i validaciju pri korišćenju dice funkcije gubitka sa augmentacijom*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Skup | gubitak | IoU | dice | sensitivity | specificity | AUC | Hausdorff |
| Trening | 0.025 | 0.924 | 0.958 | 0.959 | 0.999 | 0.989 | 0.215 |
| Validacija | 0.185 | 0.564 | 0.652 | 0.626 | 0.994 | 0.9 | 0.584 |

# 6 Diskusija i zaključak

Zbog specifičnosti lokalnog tkiva i muskulature očitavanje magnetne rezonance mozga je težak problem koji zahteva puno pažnje i iskusnog radiologa. Metrike koje je postigao model opisan u ovom radu u rangu su sa metrikama koje radiolozi postižu ručnim obeležavanjem.

Ovakva preciznost modela koja se graniči sa preciznošću ručnog obeležavanja otvara vrata primeni automatske segmentacije tumora mozga u praksi. Primenom ovakvog modela u svrhu automatske segmentacije, moguće je povećati učestalost kontrolnih skrininga bez dodatnog opterećenja radiologa i bez uštrba na kvalitetu segmentacije. Ovim bi se omogućilo prihvatanje programa skrininga koji bi pomogao u popravljanju trenutne izuzetno negativne statistike koju tumor mozga ima u našoj zemlji.

Chart, histogram, scatter chart

Description automatically generated

Slika 9. Primer segmentacije gde model radi dobro: tumor koji se sastoji iz jednog dela

Najveći problem ovog modela predstavlja izražena pojava preobučavanja u treningu. U ovom radu predstavljene su neke od tehnika za rešavanje preobučavanja poput augmentacije i tehnike ansambla. Dodatni napredak može biti postignut primenama tehnika regularizacije i gašenjem (eng. *dropout*) delova mreže pri treningu. Najveći doprinos boljem radu modela imalo bi povećanje skupa podataka za trening, čime bi se model u treningu upoznao i prilagodio na što više različitih tipova i dimenzija tumora.

Chart, histogram

Description automatically generated

Slika 10. Primer segmentacije gde model radi lošije: tumor koji se sastoji iz više delova

Dodatno unapređenje priložene metode može biti postignuto uz korišćenje tehnika za postprocesiranje segmentacije koje uzimaju u obzir prostorne zavisnosti koje postoje između ćelija tumora. Na ovaj način moguće je poboljšati segmentaciju u slučajevima gde je potrebna finija granularnost dobijenih rezultata.

# Literatura

1. S.U. Aswathy, G. Glan Deva Dhas and S.S. Kumar, "**A Survey on detection of brain tumor from MRI Brain images**", 7th International Conference on Cloud Computing Data Science & Engineering –Confluence, 2017.
2. Tustison N. et al. "**Optimal symmetric multimodal templates and concatenated random forests for supervised brain tumor segmentation (simplified) with antsr**." Neuroinformatics (2015)
3. Ciresan D. et al. “**Deep neural networks segment neuronal membranes in electron microscopy images**”, Advances in neural information processing systems 2012: 2843–2851
4. Urban G. et al. **Multi-modal brain tumor segmentation using deep convolutional neural networks**. MICCAI Multimodal Brain Tumor Segmentation Challenge (BraTS) 2014:31–35.
5. Zikic D. et al. **Segmentation of brain tumor tissues with convolutional neural networks**. MICCAI Multimodal Brain Tumor Segmentation Challenge (BraTS) 2014:36–39
6. Pereira S, Pinto A, Alves V, Silva C A. **Brain tumor segmentation using convolutional neural networks in MRI images**. IEEE Trans Med Imaging 2016;35(5):1240–1251.
7. Ronneberger O, Fischer P, Brox T (2015) **U-net: convolutional networks for biomedical image segmentation**. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 234–241
8. He, K.; Zhang, X.; Ren, S.; Sun, J. **Deep Residual Learning for Image Recognition.** In Proceedings of the 2016 IEEE Conference onComputer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; IEEE Computer Society: Los Alamitos,CA, USA, 2016; pp. 770–778, doi:10.1109/CVPR.2016.90
9. Ioffe, S.; Szegedy, C. **Batch normalization: Accelerating deep network training by reducing internalcovariate shift.** Proceedings of the 32nd International Conference on Machine Learning, PMLR 37:448-456, 2015.
10. Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. **Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations**. In Deep learning in medical image analysis and multimodal learning for clinical decision support, pages 240–248. Springer, 2017.
11. L.G. Nyúl, J.K. Udupa, X. Zhang. **New variants of a method of MRI scale standardization**. IEEE Trans. Med. Imaging, 19 (2000), pp. 143-150
12. Link sa kog se može preuzeti skup podataka: https://www.kaggle.com/datasets/mateuszbuda/lgg-mri-segmentation