Empirijsko istraživanje algoritama strojnog učenja temeljenog na informaciji

Saliu, Leonora

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Leonora Saliu

EMPIRIJSKO ISTRAŽIVANJE ALGORITAMA STROJNOG UČENJA TEMELJENOG NA INFORMACIJI

DIPLOMSKI RAD

Varaždin, 2024.

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Leonora Saliu

Matični broj: 0016130141

Studij: Informacijsko i programsko inženjerstvo

EMPIRIJSKO ISTRAŽIVANJE ALGORITAMA STROJNOG UČENJA TEMELJENOG NA INFORMACIJI

DIPLOMSKI RAD

Mentorica:

Izv. prof. dr. sc. Dijana Oreški

Varaždin, lipanj 2024.

Leonora Saliu

Izjava o izvornosti

Izjavljujem da je moj diplomski rad izvorni rezultat mojeg rada te da se u izradi istoga nisam koristio drugim izvorima osim onima koji su u njemu navedeni. Za izradu rada su korištene etički prikladne i prihvatljive metode i tehnike rada.

Autorica potvrdila prihvaćanjem odredbi u sustavu FOI-radovi

_____

Sažetak

Diplomski rad bavi se istraživanjem algoritama strojnog učenja temeljenog na informaciji, točnije algoritmima CART (eng. Classification and Regression Tree) i ID3 (eng. Iterative Dichotomiser 3). Na samom početku rada, opisano je što je to zapravo strojno učenje te kakve sve vrste strojnog učenja postoje. Zatim slijedi samo istraživanje algoritama odnosno njihova teorijska obrada te prikaz praktične primjene na odabranom skupu podataka.

Strojno učenje jedno je od najbrže rastućih područja računalne znanosti, s dalekosežnim primjenama. Glavni cilj strojnog učenja je proučavanje, projektiranje i poboljšanje matematičkih modela koji mogu biti trenirani za zaključivanje budućnosti i donošenje odluka bez potpunog poznavanja svih utjecajnih elemenata. Stablo odlučivanja jedan je od najmoćnijih alata algoritama za nadzirano učenje koji se koristi za zadatke vezane uz klasifikaciju i regresiju.

Ključne riječi: strojno učenje; nadzirano učenje; nenadzirano učenje; stablo odlučivanja; CART algoritam; ID3 algoritam; Gini nečistoća; informacijski dobitak;

Sadržaj

1. Uvod

Temeljni cilj ovog diplomskog rada je istraživanje algoritama i njihove primjene na konkretno odabranom skupu podataka. Rad se zapravo može podijeliti u dvije skupine: teorijski dio i praktični dio. U teorijski dio pripada sva teorija koja obuhvaća općenite informacije o strojnom učenju, vrstama strojnog učenja te algoritmima CART i ID3, dok se u praktični dio može svrstati konkretno računanje pojedinih metrika za algoritme (Gini nečistoća za CART algoritam te entropija i informacijska dobit za ID3 algoritam).

Nakon teorijske obrade strojnog učenja, vrsta strojnog učenja, slijedi teorijska obrada algoritama za stablo odlučivanja, odnosno CART algoritam. Nakon toga slijede poglavlja u kojima se odrađuje praktični dio ovog rada odnosno računanje Gini nečistoće kao metrike za CART algoritam primjenom podataka iz odabranog skupa podataka. Zatim se isto odrađuje za ID3 algoritam: prvo teorijska obrada te nakon toga računanje metrika i prikaz stabla odlučivanja nakon primjene metrika. Na samom kraju istaknute su neke karakteristike oba algoritma te njihove prednosti i nedostaci te prikazano je kako se kreira model stabla odlučivanja koristeći odabrani skup podataka u alatu BigML.

2. Metode i tehnike rada

Za uspješnu izradu ovog rada korišteno je mnogo literature za istraživanje teorijskog dijela rada odnosno za teorijsku obradu poglavlja o strojnom učenju te kasnije o samim algoritmima.

Što se tiče praktičnog dijela, korišten je skup podataka „Telco Customer Churn“ koji sadrži rezultate istraživanja na koojem su ispitanici odgovarali na različita pitanja vezana uz telekomunikacijske usluge te ono najvažnije pitanje: jesu li napustili telekomunikacijsku tvrtku kao klijent ili su i dalje vjerni klijent istoj. Skup podataka je dostupan na poveznici: https://www.kaggle.com/blastchar/telco-customer-churn.

Iz ovog skupa podataka koristili su se određeni podaci kako bi se istaknule metrike odabranih algoritama za istraživanje. Za izradu modela stabla odlučivanja koristio se online alat draw.io dostupan na: https://app.diagrams.net/. U posljednjem poglavlju ovog rada prikazan je model stabla odlučivanja kreiran pomoću odabranog skupa podataka u alatu BigML. Alat je dostupan na poveznici: https://bigml.com/.

3. Strojno učenje

Strojno učenje jedno je od najbrže rastućih područja računalne znanosti, s dalekosežnim primjenama. Jedan od oslonaca informacijske tehnologije postalo je strojno učenje, a time i središnji, iako obično skriveni, dio našeg života [1]. Glavni cilj strojnog učenja je proučavanje, projektiranje i poboljšanje matematičkih modela koji mogu biti trenirani za zaključivanje budućnosti i donošenje odluka bez potpunog poznavanja svih utjecajnih elemenata (vanjski faktori) [2, str. 9]. Postoje različiti pristupi strojnom učenju, a svaki od njih će detaljnije biti objašnjen u sljedećih nekoliko potpoglavlja.

3.1. Nadzirano strojno učenje

Nadzirano strojno učenje karakterizirano je konceptom učitelja odnosno nadzornika čiji je temeljni zadatak računalu (agentu) dati preciznu mjeru njegove pogreške koja se može usporediti s izlaznim vrijednostima. To se postiže na način da se agentu prezentiraju ulazni primjeri i za svaki od njih očekivani izlaz (skup parova). Polazeći od ove informacije, agent može ispraviti svoje parametre kako bi smanjio veličinu funkcije globalnog gubitka. Nakon svake iteracije može se postići povećavanje ukupne točnosti i samim time razlika između predviđene i očekivane vrijednosti postaje blizu nule [2, str. 10].

U nadziranim scenarijima cilj je osposobiti sustav koji također mora raditi i s uzorcima koji ranije nisu viđeni. Dakle, generalizacija modela vrlo je važna kako bi se izbjegao problem prekomjernog prilagođavanja čiji je glavni učinak mogućnost ispravnog predviđanja samo uzoraka koji se koriste za obuku, dok je pogreška za ostale uzorke vrlo visoka [2, str. 10].

Koraci u nadziranom učenju:

- Određivanje vrste skupa podataka za trening
- Prikupljanje označenih podataka za trening
- Podjela skupa podataka na skup podataka za trening, testiranje i za provjeru valjanosti
- Određivanje ulaznih značajki skupa podataka za trening
- Odabir algoritma za model
- Izvršavanje algoritma na skupu podataka za trening
- Ocijenjivanje točnosti modela pružanjem testnog skupa [3].

Nadzirano strojno učenje može se nadalje podijeliti u dvije kategorije: regresija i klasifikacija.

Regresijski algoritmi koriste se ako postoji odnos između ulazne varijable i izlazne varijable, za predviđanje kontinuiranih varijabli (numeričke) kao što je vremenska prognoza i slično. Poznati regresijski algoritmi su:

- Linearna regresija
- Stabla regresije
- Nelinearna regresija
- Bayesova linearna regresija
- Polinomijalna regresija [3].

Klasifikacijski algoritmi koriste se kada je izlazna varijabla kategorička, što znači da postoje klase odnosno kategorije kao što su muško – žensko, točno – netočno, da – ne i slično.

Poznati klasifikacijski algoritmi su:

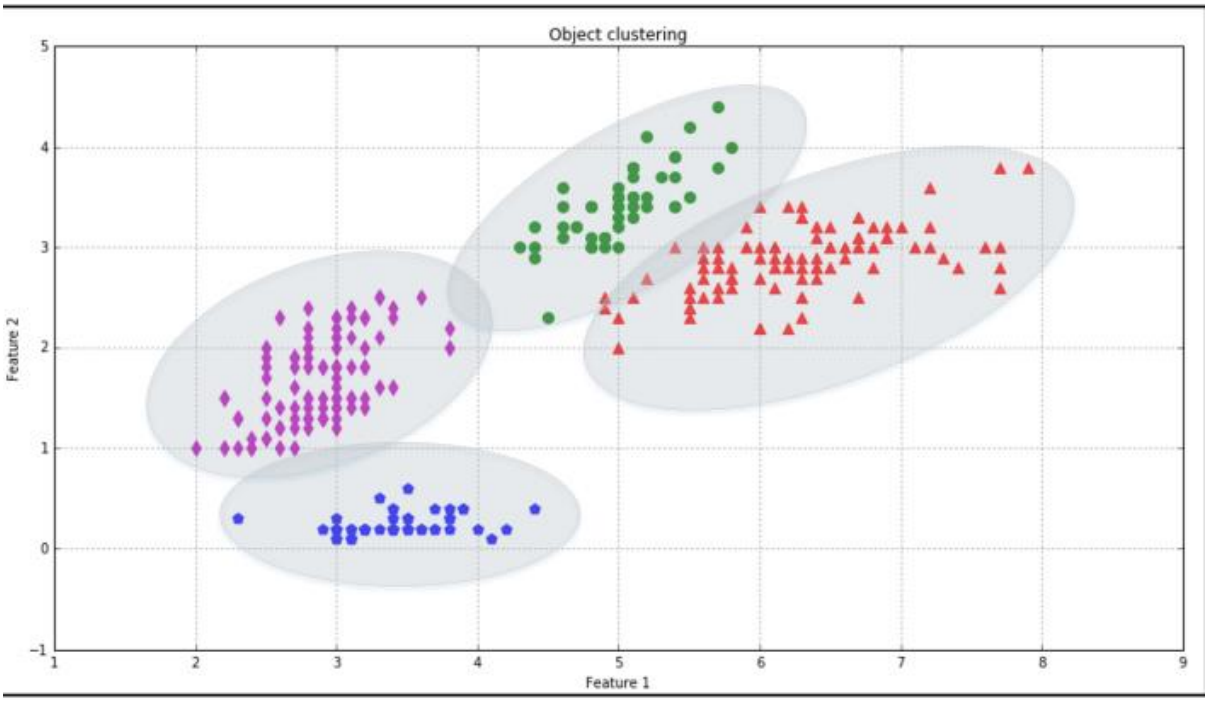
- Slučajna šuma
- Stabla odlučivanja
- K-najbližih susjeda
- Neuronske mreže
- Logistička regresija
- Vektorski strojevi za podršku [4].

3.2. Nenadzirano strojno učenje

Nenadzirano strojno učenje temelji se na nepostojanju bilo kakvog nadzornika i upravo iz tog razloga na apsolutnim mjerama pogreške [2, str. 12]. Funkcionira na način da se algoritmu učenja ne daju oznake, ostavljajući ga da sam pronađe strukturu u svom unosu. Učenje bez nadzora može biti samo sebi cilj (otkrivanje skrivenih obrazaca u podacima) ili sredstvo prema cilju (učenje značajki) [5].

U nekim problemima prepoznavanja uzoraka, podaci za obuku sastoje se od skupa ulaznih vektora x bez ikakvih odgovarajućih ciljnih vrijednosti. Cilj u takvim problemima učenja bez nadzora može biti otkrivanje grupa sličnih primjera unutar podataka, gdje se to naziva klasteriranje, ili određivanje kako su podaci raspoređeni u prostoru, poznato kao procjena gustoće [5].

Klasteriranje odnosno grupiranje podataka se može smatrati najvažnijim problemom učenja bez nadzora; pa se, kao i svaki drugi problem ove vrste, bavi pronalaženjem strukture u zbirci neoznačenih podataka. Labava definicija klasteriranja mogla bi biti "proces organiziranja objekata u grupe čiji su članovi na neki način slični". Klasterom se stoga naziva skup objekata koji su međusobno slični odnosno dijele neka zajednička svojstva i koji su različiti od objekata koji pripadaju drugim klasterima [5].



Slika 1: Primjer klastera [2, str. 13]

3.3. Učenje potkrijepljenjem (podrškom)

Učenje potkrepljenjem je znanost o donošenju odluka. Radi se o učenju optimalnog ponašanja u okruženju kako bi se dobila maksimalna nagrada. Ovo optimalno ponašanje uči se kroz interakciju s okolinom i promatranjem kako ona reagira, slično kao što djeca istražuju svijet oko sebe i uče radnje koje im pomažu postići cilj [6]. Također, to je autonoman sustav za samopodučavanje koji u osnovi uči metodom pokušaja i pogrešaka te izvodi akcije s ciljem maksimiziranja nagrada ili drugim riječima, uči kroz rad kako bi postigao najbolje moguće rezultate [7].

Međutim, u ovom slučaju informacija je više kvalitativna i ne pomaže agentu u određivanju točne mjere svoje pogreške, već je korisno razumjeti je li određena radnja izvršena u određenom stanju pozitivna ili ne. Ovaj se koncept temelji na ideji da racionalni agent uvijek teži ciljevima koji mogu povećati njegovo bogatstvo. Sposobnost da vide preko dalekog horizonta je karakteristika koja krasi napredne agente, dok oni kratkovidni često nisu u stanju ispravno procijeniti posljedice svojih trenutnih radnji pa su njihove strategije uvijek ispod optimalnih [2, str. 14].

Učenje s potkrepljenjem posebno je učinkovito kada okolina nije potpuno deterministička, kada je često vrlo dinamična i kada je nemoguće imati preciznu mjeru pogreške [2, str.15].

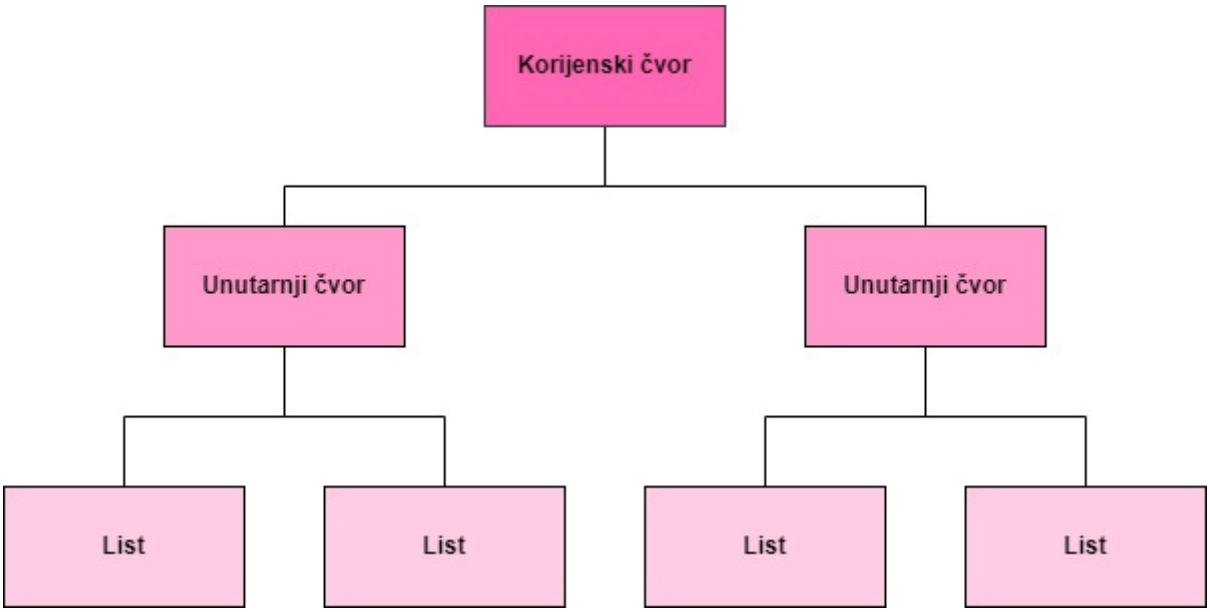
4. Strojno učenje temeljeno na informaciji

Stablo odlučivanja jedan je od najmoćnijih alata algoritama za nadzirano učenje koji se koristi za zadatke vezane uz klasifikaciju i regresiju [8]. Ima hijerarhijsku strukturu stabla stabla koja se može koristiti za podjelu velike zbirke zapisa u manje skupove klasa primjenom niza jednostavnih pravila odlučivanja, iznosi moguće ishode i puteve te na taj način donositelju odluke omogućava vizualizaciju situacije o kojoj se odlučuje [10].

Pojmovi koji se najčešće koriste kada se govori o stablu odlučivanja su: korijenski čvor, unutarnji čvor, list, grananje, grana/podstablo, roditeljski čvor, dijete, nečistoća, varijanca, dobivanje informacija, obrezivanje i pretreniranost.

Korijenski čvor je najviši čvor u stablu koji predstavlja kompletan skup podataka i ujedno je i početna točka procesa donošenja odluka. Unutarnji čvor je čvor koji simbolizira izvor povezan s ulaznom značajkom čije se grananje povezuje s lisnim čvorovima ii drugim unutarnjim čvorovima. List je čvor bez podređenih čvorova koji označava oznaku klase ili numeričku vrijednost. Grananje je proces dijeljenja čvora u dva ili više podčvorova korištenjem kriterija dijeljenja i odabrane značajke. Grana/podstablo počinje na unutarnjem čvoru i završava na lisnim čvorovima. Roditeljski čvor je čvor koji se dijeli na jedan ili više čvorova djece. Čvor dijete je čvor koji se dobije dijeljenjem roditeljskog čvora. Nečistoća je mjera homogenosti ciljane varijable u podskupu podataka, a odnosi se na stupanj slučajnosti ili nesigurnosti u nizu primjera. Varijanca mjeri koliko se predviđene i ciljne varijable razlikuju u različitim uzorcima podataka. Dobivanje informacija je mjera smanjenja nečistoće koja se postiže dijeljenjem skupa podataka na određenu značajku u stablu odlučivanja, a kriterij razdvajanja određen je značajkom koja nudi najveći dobitak informacija. Obrezivanje je postupak uklanjanja grana sa stabla odlučivanja koje ne daju dodatne informacije ili dovode do pretreniranja [8]. Pretreniranost je koncept koji se događa kada se statistički model u potpunosti točno uklapa u svoje podatke o treniranju i tada algoritam ne može točno raditi protiv neviđenih podataka i samim time se poništava njegova svrha [11].

Stablo odlučivanja sastoji se od korijenskog čvora, mnogobrojnih grana, unutarnjih čvorova i lisnih čvorova [9].



Slika 2: Stablo odlučivanja (vlastita izrada prema [9])

Na slici iznad (Slika 2) vidljiva je struktura stabla, odnosno vidljivo je da stablo počinje korijenskim čvorom koji ne sadrži ulazne grane. Odlazne grane iz korijenskog čvora zatim ulaze u nadolazeće, unutarnje čvorove stabla koji se također nazivaju i čvorovima odlučivanja. Stablo završava tzv. listovima koji predstavljaju sve moguće ishode unutar skupa podataka [9].

Tablica 1: Prednosti i nedostaci stabla odlučivanja (vlastita izrada prema [8])

Prednosti	Nedostaci
Može raditi s kategoričkim vrijednostima	Promjene u podacima uzrokuju nestabilnost
Pomaže u razmišljanju o svim mogućim ishodima problema	Problem pretreniranosti
Čišćenje podataka je manje potrebno u usporedbi s drugim algoritmima	S većim brojem oznaka klase povećava se složenost stabla

4.1. Klasifikacijsko i regresijsko stablo odlučivanja

CART (eng. Classification and Regression Tree) je prediktivni algoritam koji se koristi u strojnom učenju, a objašnjava kako se vrijednosti određene ciljne varijable mogu predvidjeti. CART su prvi producirali Leo Breiman, Jerome Friedman, Richard Olshen i Charles Stone 1984. [12].

Tablica 2: CART povijest (vlastita izrada prema [13])

CART povijest	
1974.	Leo Breiman, Charles Stone, Jerome Friedman i Richard Olsen počinju razvijati algoritam klasifikacijskog i regresijskog stabla
1977.	Prva verzija CART-a je izumljena
1984.	Službeno objavljen softver CART stabla odlučivanja

Pojam CART služi kao generički pojam za sljedeće kategorije stabala odlučivanja: klasifikacijska i regresijska stabla.

Klasifikacijskim stablima nazivaju se ona stabla koja se koriste za određivanje kojoj će klasi ciljana varijabla najvjerojatnije pripasti, a da se pritom radi o kategorijskoj varijabli. Primjer kategorijske varijable je vrsta vozila (osobno vozilo poput automobila, motocikla, teretno vozilo..).

Regresijskim stablima nazivaju se ona stabla koja se koriste za predviđanje vrijednosti kontinuirane varijable. Primjer kontinuirane varijable je cijena nekog proizvoda [12].

Što se tiče strukture, CART gradi strukturu nalik stablu koja se može sastojati od mnogobrojnih čvorova i grana pri čemu čvorovi predstavljaju različite točke odlučivanja, a grane predstavljaju moguće ishode tih odluka. Čvorovi listova u stablu sadrže predviđenu oznaku klase ili vrijednost za ciljnu varijablu.

Na svim čvorovima, CART procjenjuje sve moguće podjele i odabire onu koja najbolje smanjuje nečistoću rezultirajućih skupova. Prilikom odrađivanja zadataka klasifikacije, CART koristi Gini nečistoću (eng. Gini impurity) kao kriterij razdvajanja (što je Gini nečistoća niža, to je podskup čišći), dok za regresijske zadatke koristi rezidualnu redukciju kao kriterij

razdvajanja (što je manja rezidualna redukcija, to bolje model odgovara podacima). Gini nečistoća i rezidualna redukcija biti će pobliže objašnjene nadolazećim odlomcima [12].

Kako bi se spriječilo prekomjerno uklapanje podataka (eng. Overfitting) kada stablo dostigne punu strukturu, koristi se tehnika obrezivanja (eng. Prunning) kojom se uklanjaju čvorovi čiji je doprinos točnosti modela malen. Prekomjerno uklapanje podataka događa se kada model objašnjava nasumične elemente podataka za obuku za koje postoji mala vjerojatnost da će biti značajke veće populacije podataka te kao rezultat pretreniran model može imati dobre rezultate na podacima za treniranje, ali loše na novim, neviđenim podacima [14]. Obrezivanje složenosti troškova i obrezivanje prirasta informacija dvije su popularne tehnike obrezivanja stabla. Obrezivanje složenosti troškova podrazumijeva računanje troška svakog čvora stabla i uklanjanje čvorova čiji je trošak negativan, dok obrezivanje prirasta informacija podrazumijeva računanje prirasta informacija svakog čvora i uklanjanje čvorova čiji je prirast informacija nizak [12].

Osnovni koraci pri korištenju CART algoritma

1. Korijenski čvor stabla trebao bi biti kompletan skup podataka za obuku (treniranje).
2. Određivanje nečistoće podataka na temelju svake značajke prisutne u skupu podataka. Nečistoća se može mjeriti pomoću metrika kao što je Gini indeks nečistoće.
3. Odabir značajke koja rezultira najvećim dobitkom informacija ili smanjenjem nečistoća prilikom dijeljenja podataka.
4. Za svaku moguću vrijednost odabrane značajke, dijeli se skup podataka u dva podskupova (lijevi i desni), pri čemu u jednom podskupu značajka poprima vrijednost, a u drugom je nema. Važno je da se ovom podjelom dobiju podskupovi koji su čišći u odnosu na ciljnu varijablu.
5. Određivanje nečistoće svakog rezultirajućeg podskupa na temelju ciljne varijable.
6. Iterativno ponavljanje koraka 2 – 5 za svaki podskup dok se ne ispuni uvjet zaustavljanja, a uvjet zaustavljanja može biti najveća dubina stabla, minimalni broj uzoraka potrebnih za cijepanje ili minimalni prag nečistoće.
7. Dodijeljivanje oznake većinske klase za zadatke klasifikacije ili srednju vrijednost za zadatke regresije za svaki terminalni čvor (lisni čvor) u stablu [8].

4.1.1. CART za klasifikaciju

Kao što je u ranijim poglavljima već spomenuto, klasifikacijsko stablo je prediktivni algoritam gdje je ciljna varijabla kategorička. Algoritam se koristi za identifikaciju klase unutar koje će ciljna varijabla najvjerojatnije pasti. Klasifikacijska stabla koriste se kada skup podataka treba podijeliti u klase koje pripadaju varijabli odgovora (npr. da ili ne). Predviđene oznake klase prisutne su na svakom čvoru lista stabla [12].

CART za klasifikaciju funkcionira na način da rekurzivno dijeli podatke o treniranju u skupove koji postaju sve manji i manji na temelju određenih kriterija. Cilj ovih podjela je podijeliti podatke na način koji minimalizira nečistoću unutar svakog podskupa, a nečistoću u ovom sučaju označava pomiješanost podataka u određenom podskupu [12].

4.1.2. CART za regresiju

Regresijsko stablo je algoritam gdje je ciljna varijabla kontinuirana (na primjer dnevna temperatura), a koristi se za predviđanje njezine vrijednosti.

CART za regresiju je metoda učenja stabla odlučivanja koja stvara strukturu nalik stablu za predviđanje kontinuiranih ciljnih varijabli. Stablo se sastoji od čvorova koji predstavljaju različite točke odlučivanja i grana koje predstavljaju moguće ishode tih odluka. Predviđene vrijednosti za ciljnu varijablu pohranjene su u svakom čvoru lista stabla.

CART za regresiju radi tako da podatke o obuci rekurzivno dijeli na manje podskupove na temelju specifičnih kriterija. Cilj je podijeliti podatke na način koji smanjuje rezidualno smanjenje u svakom podskupu.

- **Smanjenje reziduala -** Smanjenje reziduala je mjera koliko je prosječna kvadratna razlika između predviđenih vrijednosti i stvarnih vrijednosti za ciljnu varijablu smanjena dijeljenjem podskupa. Što je manja rezidualna redukcija, to bolje model odgovara podacima.

- **Kriteriji dijeljenja -** CART procjenjuje svako moguće dijeljenje na svakom čvoru i odabire ono koje rezultira najvećim smanjenjem preostale pogreške u rezultirajućim podskupovima. Ovaj se proces ponavlja sve dok se ne ispuni kriterij zaustavljanja, kao što je postizanje maksimalne dubine stabla ili premalo instanci u čvoru lista [12].

4.1.3. Gini nečistoća

Ginijeva mjera nečistoće jedna je od metoda koja se koristi u algoritmima stabla odlučivanja za odlučivanje o optimalnom dijeljenju iz korijenskog čvora i naknadnim dijeljenjima [15]. Ginijeva nečistoća poprima vrijednosti u intervalu [0, 1]. Ako je vrijednost ginija blizu 0, može se reći da je vjerojatnost pogrešne klasifikacije tog određenog podatka niska, drugim riječima, ova podatkovna točka vjerojatno će biti ispravno klasificirana. Ako je blizu 1, tumačenje je upravo suprotno – ova podatkovna točka vjerojatno će biti pogrešno klasificirana [16].
U sljedećim odlomcima biti će prikazano računanje gini nečistoće na primjeru prvih deset redaka iz skupa podataka [18]. Formula za računanje gini indeksa izgleda ovako:

$$Ginx = p_1 \cdot (1 - p_1) + p_2 \cdot (1 - p_2)$$

Iz čega slijedi

$$Ginx = 2 \cdot p_1 p_2$$

Jednadžba iznad dati će mjeru Gini nečistoće za potpodjelu (samo jednu stranu dobivenu podjelom iz čvora), ali ako se želi doznati Gini mjera nečistoće za cijelu podjelu (jer se nakon dijeljenja iz čvora podaci podijele lijevo i desno) tada je potrebno koristiti formulu pomoću koje se računa „izvagana“ Gini mjera nečistoće koja uzima u obzir i lijevu i desnu podjelu iz čvora.

$$WeightedGinx = p_L \cdot (2 \cdot p_{L1} p_{L2}) + p_R \cdot (2 \cdot p_{R1} p_{R2})$$

U formuli iznad p_L predstavlja udio podjele koja se račva na lijevu stranu, dok p_R predstavlja dio podjele koji se račva na desnu stranu [16].

Skup podataka koji će se koristiti je javno dostupan skup podataka [18] po imenu „Telco Customer Churn“. Skup sadrži različite podatke sakupljene od 7043 različitih ljudi (svaki čovjek predstavlja jednu instancu) koji koriste različite telekomunikacijske usluge i uvjete i cijene pod kojima se te usluge koriste. Jedan od zanimljivijih atributa u skupu podataka je podatak koji govori o tome je li korisnik napustio telekomunikacijsku tvrtku kod koje je bio klijent ili ne.

Podaci će u sljedećih nekoliko odlomaka biti opisani, prikazani te će se koristiti za prikaz postupka računanja Gini nečistoće te analize rezultata.

Skup se sastoji od dvadeset i jednog stupca s podacima odnosno dvadeset i jednog atributa različitog tipa podatka (numerički, tekstualni i kategorijski), a to su:

- ID korisnika,
- Spol
- Je li korisnik stariji građanin
- Ima li korisnik partnera
- Ima li korisnik ljude o kojima brine
- Je li korisnik stalno zaposlen
- Telefonska usluga
- Mnogobrojne linije
- Internetske usluge
- Online sigurnost
- Online sigurnosna kopija
- Zaštita uređaja
- Tehnička podrška
- Streamanje TV
- Streamanje filmova
- Ugovorna obveza
- Elektronički račun
- Način plaćanja
- Mjesečan iznos
- Ukupan iznos
- Churn (odlazak korisnika).

Detaljniji prikaz podataka biti će prikazan u sljedećoj tablici na sljedećoj stranici:

Tablica 3: Prikaz prvih 10 instanci iz skupa podataka [18] – 1.dio

Customer ID	**Gender**	**Senior citizen**	**Partner**	**Dependents**	**Tenure**	**Phone Service**	**Multiple lines**	**Internet service**	**Online security**
7590-VHVEG	Žensko	0	Da	Ne	1	Ne	Ne	DSL	Ne
5575-GNVDE	Muško	0	Ne	Ne	34	Da	Ne	DSL	Da
3668-QPYBK	Muško	0	Ne	Ne	2	Da	Ne	DSL	Da
7795-CFOCW	Muško	0	Ne	Ne	45	Ne	Ne	DSL	Da
9237-HQITU	Žensko	0	Ne	Ne	2	Da	Ne	Optička vlakna	Ne
9305-CDSKC	Žensko	0	Ne	Ne	8	Da	Da	Optička vlakna	Ne
1452-KIOVK	Muško	0	Ne	Da	22	Da	Da	Optička vlakna	Ne
6713-OKOMC	Žensko	0	Ne	Ne	10	Ne	Ne	DSL	Da
7892-POOKP	Žensko	0	Da	Ne	28	Da	Da	Optička vlakna	Ne
6388-TABGU	Muško	0	Ne	Da	62	Da	Ne	DSL	Da

Tablica 4: Prikaz prvih 10 instanci iz skupa podataka [18] – 2.dio

Online backup	**Device protection**	**Tech support**	**Streaming TV**	**Streaming movies**	**Contract**	**Paperless billing**	**Payment method**	**Monthly charges**	**Total charges**	**Churn**
Da	Ne	Ne	Ne	Ne	Mjesečni	Da	El.račun	29,85	29,85	Ne
Ne	Da	Ne	Ne	Ne	1 godina	Ne	Račun	56,95	1889,5	Ne
Da	Ne	Ne	Ne	Ne	Mjesečni	Da	Račun	53,85	108,15	Da
Ne	Da	Da	Ne	Ne	1 godina	Ne	Transakcija	42,3	1840,75	Ne
Ne	Ne	Ne	Ne	Ne	Mjesečni	Da	El.račun	70,7	151,65	Da
Ne	Da	Ne	Da	Da	Mjesečni	Da	El.račun	99,65	820,5	Da
Da	Ne	Ne	Da	Ne	Mjesečni	Da	Kreditna kartica	89,1	1949,4	Ne
Ne	Ne	Ne	Ne	Ne	Mjesečni	Ne	Račun	29,75	301,9	Ne
Ne	Da	Da	Da	Da	Mjesečni	Da	El.račun	104,8	3046,05	Da
Da	Ne	Ne	Ne	Ne	1 godina	Ne	Transakcija	56,15	3487,95	Ne

Na temelju podataka u prethodno prikazanim tablicama, računati će se Gini nečistoća. Atributi koji će se uzeti u obzir prilikom računanja su: spol (gender), telefonska usluge (phone service), tehnička podrška (tech support), vrsta ugovorne obveze (contract), mjesečni iznos troškova (monthly charges) te ukupan iznos troškova (total charges).

Za prethodno nabrojane atribute pomoću Gini nečistoće pokušati će se odrediti utjecaj na najzanimljiviji atribut u skupu, a to je odlazak korisnika iz telekomunikacijske tvrtke (churn). Odnosno, računati će se koji od atributa je optimalniji prilikom dijeljenja iz čvora.

Tablica 5: Podaci o spolu i odlasku [18]

Gender	**Churn**
Žensko	Ne
Muško	Ne
Muško	Da
Muško	Ne
Žensko	Da
Žensko	Da
Muško	Ne
Žensko	Ne
Žensko	Da
Muško	Ne

$$Gini_{female} = 2 \cdot \frac{3}{5} \cdot \frac{2}{5} = 0{,}48$$

$$Gini_{male} = 2 \cdot \frac{1}{5} \cdot \frac{4}{5} = 0{,}32$$

Zatim „izvagana“ Gini nečistoća:

$$WeightedGini_{gender} = \frac{5}{10} \cdot Gini_{female} + \frac{5}{10} \cdot Gini_{male}$$

$$= \frac{5}{10} \cdot 0{,}48 + \frac{5}{10} \cdot 0{,}32$$

$$= 0{,}40$$

Na temelju Gini mjere za muški spol i Gini mjere za ženski spol, izračunata je izvagana Gini nečistoća koja iznosi 0,40. S obzirom na to da je rezultat bliži 0, nego što je 1, može se zaključiti da su veće šanse da je vjerojatnost ove klasifikacije točna, odnosno da su manje šanse da je pogrešna.

Tablica 6: Podaci o telefonskoj usluzi i odlasku [18]

Phone Service	**Churn**
Ne	Ne
Da	Ne
Da	Da
Ne	Ne
Da	Da
Da	Da
Da	Ne
Ne	Ne
Da	Da
Da	Ne

$$Gini_{yesPhoneService} = 2 \cdot \frac{4}{7} \cdot \frac{3}{7} = 0{,}49$$

$$Gini_{noPhoneService} = 2 \cdot \frac{3}{3} \cdot 0 = 0$$

Zatim „izvagana“ Gini nečistoća:

$$WeightedGini_{phoneService} =$$

$$= \frac{7}{10} \cdot Gini_{yesPhoneService} + \frac{3}{10} \cdot Gini_{noPhoneService}$$

$$= \frac{7}{10} \cdot 0{,}49 + \frac{3}{10} \cdot 0$$

$$= 0{,}343$$

Na temelju Gini mjere za telefonske usluge - da i Gini mjere za telefonske usluge - ne, izračunata je izvagana Gini nečistoća koja iznosi 0,343. S obzirom na to da je rezultat bliži 0, nego što je 1, može se zaključiti da su veće šanse da je vjerojatnost ove klasifikacije točna, odnosno da su manje šanse da je pogrešna.

Tablica 7: Podaci o tehničkoj podršci i odlasku [18]

Tech support	**Churn**
Ne	Ne
Ne	Ne
Ne	Da
Da	Ne
Ne	Da
Ne	Da
Ne	Ne
Ne	Ne
Da	Da
Ne	Ne

$$Gini_{yesTechSupport} = 2 \cdot \frac{1}{2} \cdot \frac{1}{2} = 0{,}50$$

$$Gini_{noTechSupport} = 2 \cdot \frac{5}{8} \cdot \frac{3}{8} = 0{,}469$$

Zatim „izvagana“ Gini nečistoća:

$$WeightedGini_{techSupport} = \\ = \frac{2}{10} \cdot Gini_{yesTechSupport} + \frac{8}{10} \cdot Gini_{noTechSupport}$$

$$= \frac{2}{10} \cdot 0{,}50 + \frac{8}{10} \cdot 0{,}469$$

$$= 0{,}475$$

Na temelju Gini mjere za korisničku podršku - da i Gini mjere za korisničku podršku - ne, izračunata je izvagana Gini nečistoća koja iznosi 0,475. S obzirom na to da je rezultat bliži 0, nego što je 1, može se zaključiti da su veće šanse da je vjerojatnost ove klasifikacije točna, odnosno da su manje šanse da je pogrešna.

Tablica 8: Podaci o vrsti ugovorne obveze i odlasku [18]

Contract	**Churn**
Mjesečno	Ne
1 godina	Ne
Mjesečno	Da
1 godina	Ne
Mjesečno	Da
Mjesečno	Da
Mjesečno	Ne
Mjesečno	Ne
Mjesečno	Da
1 godina	Ne

$$Gini_{monthly} = 2 \cdot \frac{3}{7} \cdot \frac{4}{7} = 0{,}4898$$

$$Gini_{oneYear} = 2 \cdot \frac{3}{3} \cdot 0 = 0$$

Zatim „izvagana“ Gini nečistoća:

$$WeightedGini_{contract} = \\ = \frac{7}{10} \cdot Gini_{monthly} + \frac{3}{10} \cdot Gini_{oneYear}$$

$$= \frac{7}{10} \cdot 0{,}4898 + \frac{3}{10} \cdot 0$$

$$= 0{,}34$$

Na temelju Gini mjere za vrstu ugovorne obveze - mjesečno i Gini mjere za vrstu ugovorne obveze – jedna godina, izračunata je izvagana Gini nečistoća koja iznosi 0,34. S obzirom na to da je rezultat bliži 0, nego što je 1, može se zaključiti da su veće šanse da je vjerojatnost ove klasifikacije točna, odnosno da su manje šanse da je pogrešna.

S obzirom na to da su u ovom slučaju podaci o mjesečnom iznosu troškova brojčano izraženi, za potrebe računanja Gini mjere, napraviti će se granica za mjesečne troškove u iznosu od pedeset dolara, odnosno gledati će se je li mjesečan iznos troškova veći od 50 dolara.

Tablica 9: Podaci o mjesečnom iznosu troškova i odlasku [18]

Monthly charges	**Churn**
29.85 => Ne	Ne
56.95 => Da	Ne
53.85 => Da	Da
42.3 => Ne	Ne
70.7 => Da	Da
99.65 => Da	Da
89.1 => Da	Ne
29.75 => Ne	Ne
104.8 => Da	Da
56.15 => Da	Ne

$$Gini_{moreThan50} = 2 \cdot \frac{3}{7} \cdot \frac{4}{7} = 0{,}4898$$

$$Gini_{lessThan50} = 2 \cdot \frac{3}{3} \cdot 0 = 0$$

Zatim „izvagana“ Gini nečistoća:

$$WeightedGini_{monthlyCharges} = \\ = \frac{7}{10} \cdot Gini_{moreThan50} + \frac{3}{10} \cdot Gini_{lessThan50}$$

$$= \frac{7}{10} \cdot 0{,}4898 + \frac{3}{10} \cdot 0$$

$$= 0{,}34$$

Na temelju Gini mjere za mjesečni iznos troškova veći od pedeset - da i Gini mjere za mjesečni iznos troškova veći od pedeset - ne, izračunata je izvagana Gini nečistoća koja iznosi 0,34. S obzirom na to da je rezultat bliži 0, nego što je 1, može se zaključiti da su veće šanse da je vjerojatnost ove klasifikacije točna, odnosno da su manje šanse da je pogrešna.

S obzirom na to da su u ovom slučaju podaci o ukupnom iznosu troškova brojčano izraženi, za potrebe računanja Gini mjere, napraviti će se granica za ukupne troškove u iznosu od tisuću dolara, odnosno gledati će se je li ukupan iznos troškova veći od 1000 dolara.

Tablica 10: Podaci o ukupnom iznosu troškova i odlasku [18]

Total charges	**Churn**
29.85 => Ne	Ne
1889.5 => Da	Ne
108.15 => Ne	Da
1840.75 => Da	Ne
151.65 => Ne	Da
820.5 => Ne	Da
1949.4 => Da	Ne
301.9 => Ne	Ne
3046.05 => Da	Da
3487.95 => Da	Ne

$$Gini_{moreThan1000} = 2 \cdot \frac{4}{5} \cdot \frac{1}{5} = 0{,}32$$

$$Gini_{lessThan1000} = 2 \cdot \frac{2}{5} \cdot \frac{3}{5} = 0{,}48$$

Zatim „izvagana“ Gini nečistoća:

$$WeightedGini_{totalCharges} = \\ = \frac{5}{10} \cdot Gini_{moreThan1000} + \frac{5}{10} \cdot Gini_{lessThan1000}$$

$$= \frac{5}{10} \cdot 0{,}32 + \frac{5}{10} \cdot 0{,}48$$

$$= 0{,}40$$

Na temelju Gini mjere za ukupan iznos troškova veći od tisuću - da i Gini mjere za ukupan iznos troškova veći od tisuću - ne, izračunata je izvagana Gini nečistoća koja iznosi 0,40. S obzirom na to da je rezultat bliži 0, nego što je 1, može se zaključiti da su veće šanse da je vjerojatnost ove klasifikacije točna, odnosno da su manje šanse da je pogrešna.

Tablica 11: Prikaz rezultata Gini nečistoće za pojedine atribute, vlastita izrada

Atributi	Gender	Phone service	Tech support	Contract	Monthly charges	Total charges
Gini nečistoća	0,40	0,336	0,475	0,34	0,34	0,40

Tablica iznad sadrži podatke o prethodno izračunatim iznosima Gini nečistoće za pojedine atribute. Kao atribut s najvećom Gini nečistoćom istaknuo se Tech support odnosno tehnička podrška (0,475). Kao atributi s najmanjim iznosom Gini nečistoće istaknuli su se Phone service odnosno telefonska usluga, zatim Contract odnosno vrsta ugovorne obveze i Monthly charges odnosno iznos mjesečnih troškova koje klijent plaća.

Prema tome, može se reći da su atributi s podacima o telefonskoj usluzi, vrsti ugovorne obveze i iznosu mjesečnih troškova koje klijent plaća optimalniji prilikom dijeljenja iz čvora jer imaju manju mogućnost pogrešne klasifikacije opažanja.

4.2. ID3 (Iterative dichotomiser 3) algoritam

ID3 algoritam je algoritam indukcije stabla odlučivanja koji konstruira stabla odlučivanja koristeći pristup odozgo prema dolje [19]. Dihotomizacija (iz imena) znači dijeljenje na dvije potpuno suprotne stvari. Zbog toga algoritam iterativno dijeli atribute u dvije skupine koje su najdominantniji atributi i ostale za konstruiranje stabla [22]. Razvijen je 1986. godine, a razvio ga je Ross Quinlan. Algoritam funkcionira na način da gradi stablo odlučivanja rekurzivnim dijeljenjem skupa podataka u sve manje i manje podskupove dok sve podatkovne točke u svakom podskupu ne pripadaju istoj klasi [20].

Može se reći da je ID3 algoritam zapravo metoda za odabir optimalnog atributa, a sami odabir atributa temelji se na minimiziranju informacijske entropije čvorova [21].

Osnovni koraci pri izvođenju ID3 algoritma

1. Uzimanje cijelog skupa podataka kao ulaz
2. Računanje entropije ciljane varijable i prediktivnih atributa
3. Računanje informacijskog dobitka svih atributa
4. Biranje atributa s najvećom informacijskom dobiti kao korijenskog čvora
5. Ponavljanje istog postupka na svakoj pojedinoj grani dokle god se ne finalizira čvor odluke svake grane [20].

4.2.1. Informacijski dobitak i odabir atributa

ID3 algoritam koristi mjeru nečistoće kao što je entropija ili Gini nečistoća, za računanje informacijskog dobitka svakog atributa. Entropija je mjera nereda u skupu podataka. Skup podataka s visokom entropijom je skup podataka u kojem su podatkovne točke ravnomjerno raspoređene u različitim kategorijama. Skup podataka s niskom entropijom je skup podataka u kojem su podatkovne točke koncentrirane u jednoj ili nekoliko kategorija [20].

Može se reći da se entropija koristi za procjenu kvalitete podjele. Kada je entropija nula, uzorak je potpuno homogen, što znači da svaka instanca pripada istoj klasi, a kada je entropija jedan, tada je uzorak jednako podijeljen između različitih klasa [19].

Za skup podataka S, s klasama {c1, c2, …, cn}, entropija se računa pomoću formule:

$$Entropy\ (S) = \sum_{i=1}^{n} p_i\ log_2(p_i)$$

Gdje p_i predstavlja udio instanci klase c_i u skupu podataka [23].

Mjera koliko dobro određena kvaliteta smanjuje nesigurnost naziva se dobitak informacija. ID3 dijeli podatke u svakoj fazi, odabirući svojstvo koje maksimizira dobitak informacija. Da bi se izračunao konkretan dobitak informacija potrebno izračunati razliku između iznosa entropije prije i nakon podjele.

Dobitak informacija mjeri učinkovitost atributa A u smanjenju nesigurnosti u skupu S, a računa se pomoću formule:

$$Information\ Gain\ (A, S) = Entropy\ (S) - \sum_v \frac{|S_v|}{|S|} \cdot H\ (S_v)$$

- gdje $|S|$ predstavlja ukupan broj instanci u skupu
- $|S_v|$ predstavlja broj instanci u skupu koje za atribut A imaju vrijednost v [23].

Na temelju podataka u prethodno prikazanim tablicama (Tablica 3 i Tablica 4), računati će se entropija i informacijski dobitak. Atributi koji će se uzeti u obzir prilikom računanja su: spol (gender), telefonska usluge (phone service), tehnička podrška (tech support).

Koristeći prethodno nabrojane atribute odnosno pomoću njihove entropije i informacijskog dobitka pokušati će se kreirati stablo odlučivanja. Prvi korak je dakle računanje entropije, zatim informacijskog dobitka kako bi se dobio korijenski čvor za stablo, zatim se postupak ponavlja kako bi se dobile grane i preostali čvorovi u stablu.

Tablica 12: Podaci o spolu, tel. usluzi, tehničkoj podršci te odlasku [18]

Gender	Phone service	Tech support	Churn
Žensko	Ne	Ne	Ne
Muško	Da	Ne	Ne
Muško	Da	Ne	Da
Muško	Ne	Da	Ne
Žensko	Da	Ne	Da
Žensko	Da	Ne	Da
Muško	Da	Ne	Ne
Žensko	Ne	Ne	Ne
Žensko	Da	Da	Da
Muško	Da	Ne	Ne

Prvi korak je računanje entropije za atribut Churn. Iz tablice se mogu izvući podaci o odlasku: 6 x Ne i 4 x Da pa prema tome entropija za Churn iznosi:

$$E(Churn) = -\frac{6}{10} \cdot log_2 \frac{6}{10} - \frac{4}{10} \cdot log_2 \frac{4}{10} = 0{,}97$$

Sljedeći korak je računanje informacijskog dobitka za svaki atribut. Kreće se s atributom Gender. Pogleda li se stupac Gender, vidljivo je da se iz njega može izdvojiti 5 redaka žensko i 5 redaka muško. Od redaka u kojima stoji žensko, ciljana varijabla (Churn) je 3 puta da i 2 puta ne, dok za redke u kojima stoji muško, ciljana varijabla je 4 puta ne i 1 da.

|S| = ukupan broj redaka = 10, v = žensko, $|S_v|$ = 5

$$E(S_v) = -\frac{3}{5} \cdot log_2 \frac{3}{5} - \frac{2}{5} \cdot log_2 \frac{2}{5} = 0{,}74$$

v = muško, $|S_v|$ = 5

$$E(S_v) = -\frac{1}{5} \cdot log_2 \frac{1}{5} - \frac{4}{5} \cdot log_2 \frac{4}{5} = 0{,}72$$

Informacijski dobitak zatim iznosi:

$$IG(S, Gender) = 0{,}97 - \frac{5}{10} \cdot 0{,}74 - \frac{5}{10} \cdot 0{,}72 = 0{,}24$$

Tablica 13: Podaci o tel. usluzi i odlasku [18]

Phone Service	**Churn**
Ne	Ne
Da	Ne
Da	Da
Ne	Ne
Da	Da
Da	Da
Da	Ne
Ne	Ne
Da	Da
Da	Ne

Pogleda li se stupac Phone service, vidljivo je da se iz njega može izdvojiti 7 redaka Da i 3 redaka Ne. Od redaka u kojima stoji Da, ciljana varijabla (Churn) je 4 puta da i 3 puta ne, dok za redke u kojima stoji Ne, ciljana varijabla je 3 puta ne i nijednom da.

|S| = ukupan broj redaka = 10, v = Da, $|S_v|$ = 7

$$E(S_v) = -\frac{3}{7} \cdot log_2\frac{3}{7} - \frac{4}{7} \cdot log_2\frac{4}{7} = 0{,}98$$

v = Ne, $|S_v|$ = 3

$$E(S_v) = -\frac{3}{3} \cdot log_2\frac{3}{3} - 0 = 0$$

Informacijski dobitak zatim iznosi:

$$IG(S, Phone\ service) = 0{,}97 - \frac{7}{10} \cdot 0{,}98 - \frac{3}{10} \cdot 0 = 0{,}28$$

Tablica 14: Podaci o tehničkoj podršci i odlasku [18]

Tech support	**Churn**
Ne	Ne
Ne	Ne
Ne	Da
Da	Ne
Ne	Da
Ne	Da
Ne	Ne
Ne	Ne
Da	Da
Ne	Ne

Isti postupak provodi se za atribut Tech support. Pogleda li se stupac Tech support, vidljivo je da se iz njega može izdvojiti 2 redaka Da i 8 redaka Ne. Od redaka u kojima stoji Da, ciljana varijabla (Churn) je 1 puta da i 1 puta ne, dok za redke u kojima stoji Ne, ciljana varijabla je 5 puta ne i 3 puta da.

|S| = ukupan broj redaka = 10, v = Da, $|S_v|$ = 2

$$E(S_v) = -\frac{1}{2} \cdot log_2 \frac{1}{2} - \frac{1}{2} \cdot log_2 \frac{1}{2} = 1$$

v = Ne, $|S_v|$ = 8

$$E(S_v) = -\frac{3}{8} \cdot log_2 \frac{3}{8} - \frac{5}{8} \cdot log_2 \frac{5}{8} = 0{,}95$$

Informacijski dobitak zatim iznosi:

$$IG(S, Tech\ support) = 0{,}97 - \frac{2}{10} \cdot 1 - \frac{8}{10} \cdot 0{,}95 = 0{,}01$$

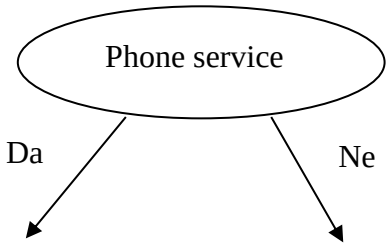
Na temelju izračunatih informacijskih dobiti dobiva se sljedeća tablica:

Tablica 15: Prikaz podataka o informacijskoj dobiti

Naziv atributa	**Informacijska dobit**
Gender	0,24
Phone service	0,28
Tech support	0,01

Na temelju rezultata iz tablice iznad, vidljivo je da najveću informacijsku dobit ima atribut Phone service odnosno atribut koji govori o tome imaju li klijenti telefonsku uslugu ili ne. Prema tome, atribut Phone service je odabran kao korijenski čvor stabla koje se gradi.

Stablo odlučivanja trenutno izgleda ovako:



Slika 3: Stablo odlučivanja prema ID3 algoritmu – 1.dio [samostalna izrada]

Sljedeći korak je biranje koji od preostalih atributa (Tech support i Gender) bolje odgovara poziciji lijevog grananja odnosno koji atribut treba doći na mjesto Phone service -> Da ->.
Da bi se ta informacija dobila, potrebno je kreirati „podskup“ pomoću kojeg će se računati ponovno informacijske dobiti, te na temelju najveće odrediti sljedeći korak grananja.

U novi podskup će se uzeti u obzir redci iz tablice 12 u kojima je atribut Phone service = Da te se dobiva sljedeća tablica:

Tablica 16: Novi podskup za računanje informacijske dobiti dobiven iz tablice 12

Gender	Phone service	Tech support	Churn
Muško	Da	Ne	Ne
Muško	Da	Ne	Da
Žensko	Da	Ne	Da
Žensko	Da	Ne	Da
Muško	Da	Ne	Ne
Žensko	Da	Da	Da
Muško	Da	Ne	Ne

Sljedeći korak je računanje entropije ciljnog atributa (Churn) za trenutno promatran podskup.

Iz tablice se mogu izvući podaci o odlasku: 3 x Ne i 4 x Da pa prema tome entropija za Churn iznosi:

$$E(Churn) = -\frac{3}{7} \cdot log_2\frac{3}{7} - \frac{4}{7} \cdot log_2\frac{3}{7} = 0{,}98$$

Zatim slijedi računanje informacijskog dobitka za svaki preostali atribut. Kreće se s atributom Gender. Pogleda li se stupac Gender, vidljivo je da se iz njega može izdvojiti 3 redaka žensko i 4 redaka muško. Od redaka u kojima stoji žensko, ciljana varijabla (Churn) je 3 puta da, dok za redke u kojima stoji muško, ciljana varijabla je 3 puta ne i 1 da.

|S| = ukupan broj redaka = 7, v = žensko, $|S_v|$ = 3

$$E(S_v) = -\frac{3}{3} \cdot log_2\frac{3}{3} - 0 = 0$$

v = muško, $|S_v|$ = 4

$$E(S_v) = -\frac{1}{4} \cdot log_2\frac{1}{4} - \frac{3}{4} \cdot log_2\frac{3}{4} = 0{,}81$$

Informacijski dobitak zatim iznosi:

$$IG(S, Gender) = 0{,}98 - \frac{3}{7} \cdot 0 - \frac{4}{7} \cdot 0{,}81 = 0{,}51$$

Tablica 17: Podaci o tehničkoj podršci i odlasku [18]

Tech support	**Churn**
Ne	Ne
Ne	Da
Ne	Da
Ne	Da
Ne	Ne
Da	Da
Ne	Ne

Isti postupak provodi se za atribut Tech support. Pogleda li se stupac Tech support, vidljivo je da se iz njega može izdvojiti 1 redak Da i 6 redaka Ne. Od redaka u kojima stoji Da, ciljana varijabla (Churn) Da, dok za redke u kojima stoji Ne, ciljana varijabla je 3 puta ne i 3 puta da.

|S| = ukupan broj redaka = 7, v = Da, $|S_v|$ = 1

$$E(S_v) = -\frac{1}{1} \cdot log_2 \frac{1}{1} - 0 = 0$$

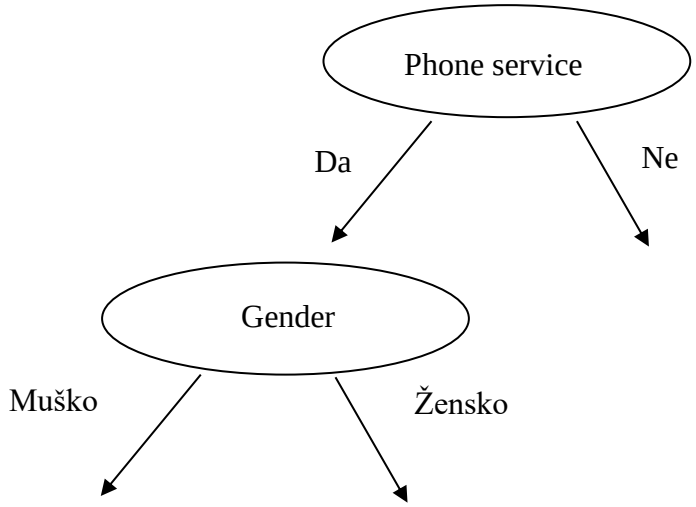
v = Ne, $|S_v|$ = 6

$$E(S_v) = -\frac{3}{6} \cdot log_2 \frac{3}{6} - \frac{3}{6} \cdot log_2 \frac{3}{3} = 1$$

Informacijski dobitak zatim iznosi:

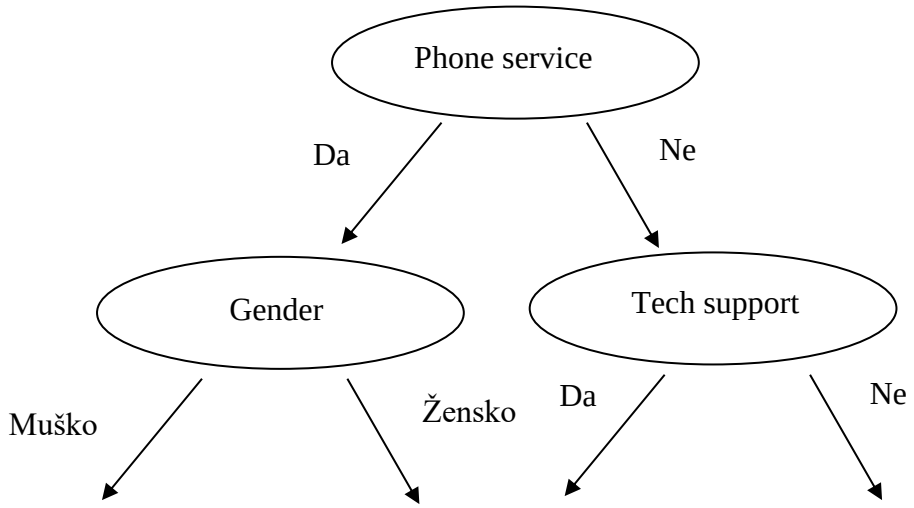
$$IG(S, Tech\ support) = 0{,}98 - \frac{1}{7} \cdot 0 - \frac{6}{7} \cdot 1 = 0{,}12$$

S obzirom na to da je informacijska dobit atributa Gender veća od informacijske dobiti atributa Tech support, kao sljedeći korak grananja uzima se atribut Gender te stablo odlučivanja sada izgleda ovako:



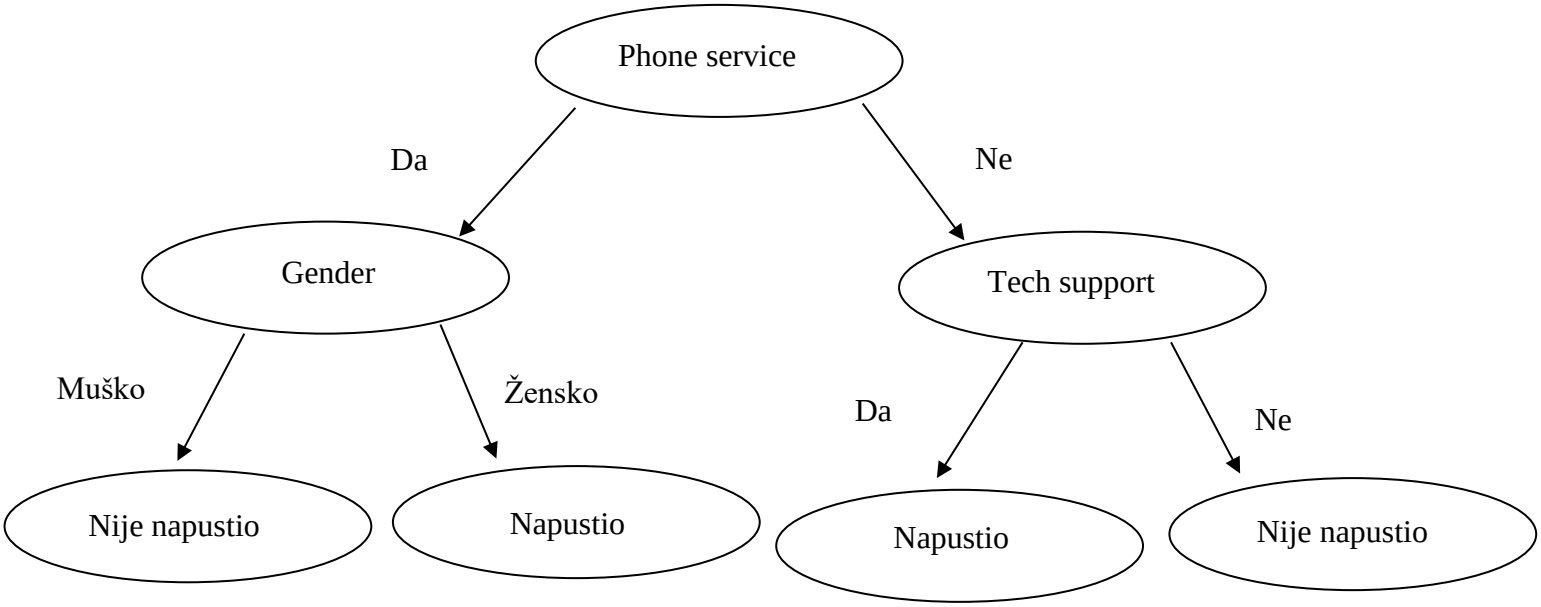
Slika 4: Stablo odlučivanja prema ID3 algoritmu – 2.dio [samostalna izrada]

S obzirom na to da je preostao samo jedan atribut (Tech support), njega je potrebno smjestiti na desnu stranu grananja iz korijenskog čvora, odnosno: Phone service -> Ne -> Tech support. Stablo odlučivanja sada izgleda ovako:



Slika 5: Stablo odlučivanja prema ID3 algoritmu – 3.dio [samostalna izrada]

Zadnji korak je dodavanje lišća na stablo odlučivanja, s obzirom na to da za Phone service - Da -> Gender -> muško većina redaka ciljnog atributa stoji Ne, kao list na lijevoj grani stavlja se Nije napustio. Zatim, za Phone service -> Da -> žensko -> Napustio. Za desnu stranu grananja (Phone service -> Ne) gleda se prvotni skup podataka u kojem Phone service redovi sadrže i Ne pa prema tome sljedeći list je Phone service ->Tech support -> Da -> Napustio i Phone service -> Tech support -> Ne -> Nije napustio. Potpuno izgrađeno stablo odlučivanja sada izgleda ovako:



Slika 6: Stablo odlučivanja prema ID3 algoritmu – 4.dio [samostalna izrada]

4.3. CART i ID3 – usporedba

Kao i sve na svijetu, tako i ovi algoritmi imaju svoje prednosti i nedostatke:

Tablica 18: Prednosti i nedostaci algoritama prema [20] i [12]

	CART	**ID3**
Prednosti	• Jednostavni rezultati • Minimalan nadzor	• Lako razumljiv • Zahtijeva malo podataka za obuku
Nedostaci	• Pretreniranost • Visoka varijanca	• Pretreniranost • Možda neće biti učinkovit kod skupova s mnogo atributa

Oba algoritma mogu se koristiti u različitim područjima pa se tako CART algoritam često koristi u slučajevima kada je potreban brzi uvid u podatke, u klasifikaciji darivatelja krvi, u području ekologije te u financijskom sektoru [12]. ID3 algoritam koristi se za procese otkrivanja prijevara, prilikom određivanja medicinskih dijagnoza, za segmentaciju kupaca, za procjenu

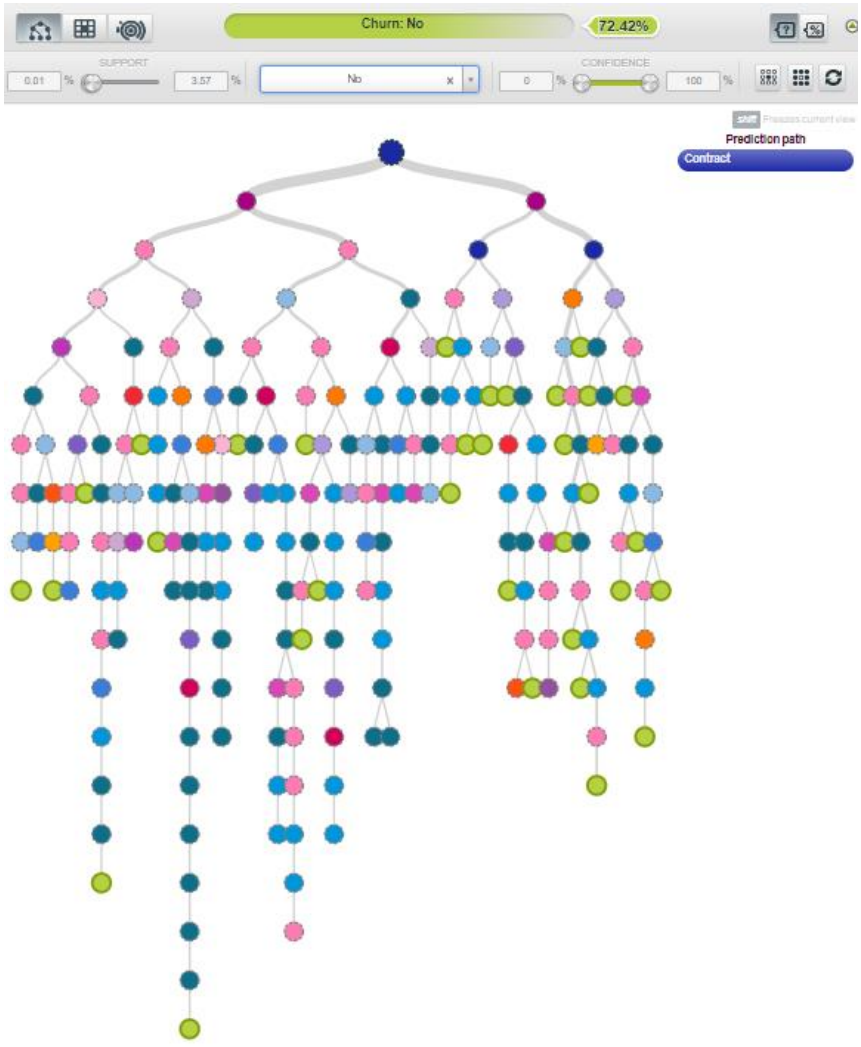
rizika (financije, osiguranje..) te za sustave preporuka za različite proizvode i usluge na temelju preferencija iz prošlosti [20].

Što se tiče sličnosti i razlika između ova dva algoritma, mogu se izdvojiti različita svojstva. CART algoritam može raditi s kontinuiranim varijablama jednako kao i s kategorijskim, dok za razliku od njega, ID3 algoritam radi s kategorijskim atributima. Nedostajuće vrijednosti u skupovima podataka vrlo su čest problem u različitim situacijama i kod različitih algoritama pa isto tako i kod ID3 algoritma, dok za CART algoritam nedostajuće vrijednosti ne predstavljaju problem. Kao metrike ID3 algoritam koristi entropiju i informacijski dobitak, dok CART algoritam najčešće koristi mjeru Gini nečistoće . Također, pomoću CART algoritma brže se može doći do željenih informacija, nego što se to može korištenjem ID3 algoritma [10].

5. Stablo odlučivanja u BigML

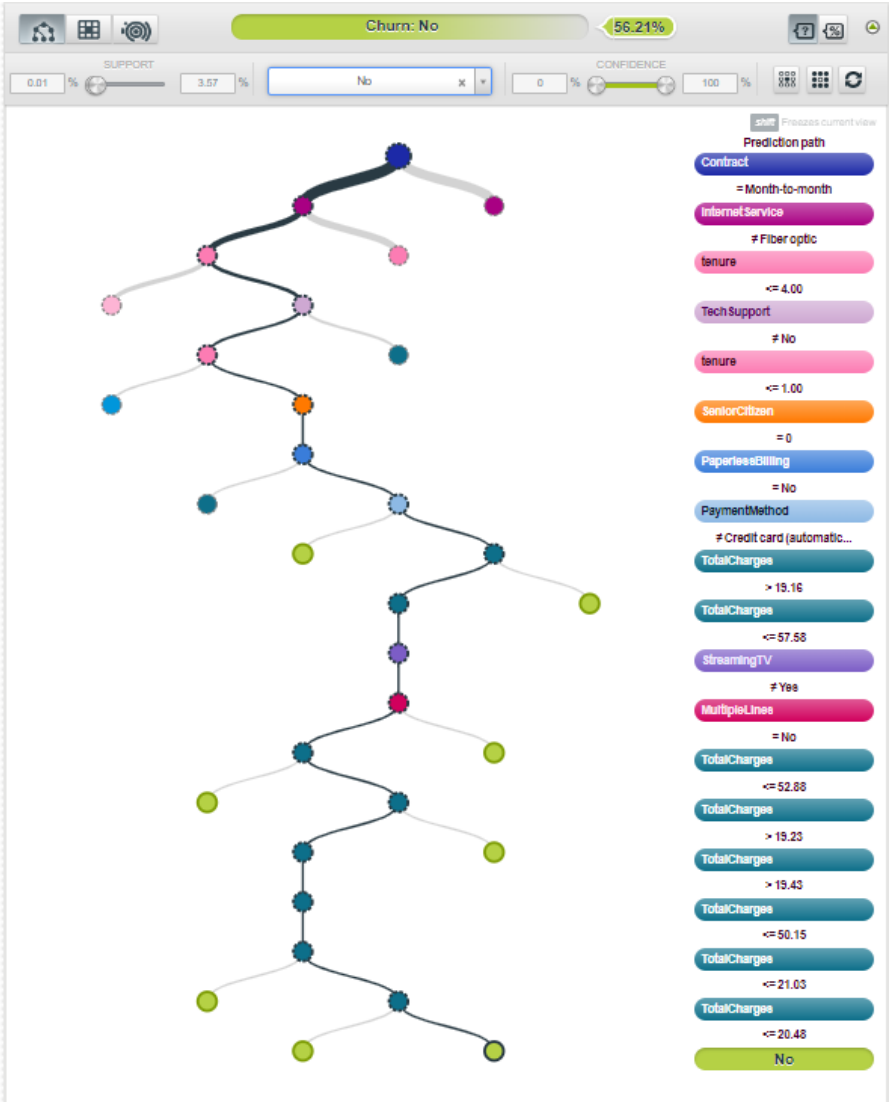
Za kreiranje stabla odlučivanja u BigML alatu, potrebno je prvo imati izvor (eng. source, a to je u ovom slučaju odabrani skup podataka [18]. Nakon toga slijedi kreiranje tzv.skupa podataka (eng. dataset) iz kojeg će se kreirati model stabla odlučivanja. Pri izradi stabla, kao zavisni atribut koristi se atribut Churn koji predstavlja odlazak odnosno ostanak klijenata u telekomunikacijskoj tvrtki.

Veličina stabla odlučivanja očituje se kroz njegovo grananje, dok se njegova dubina očituje kroz duljinu najdužeg puta u stablu (od korijenskog čvora do krajnjeg lista). U ovom slučaju, veličina stabla jednaka je broju grananja koje iznosi 45. Pri vrhu stabla, kao početni atribut stoji atribut Contract koji daje informacije o tome koliko je korisnik pod ugovorom s telekomunikacijskom tvrtkom zatim se odvijaju različiti scenariji grananja s ostalim atributima.



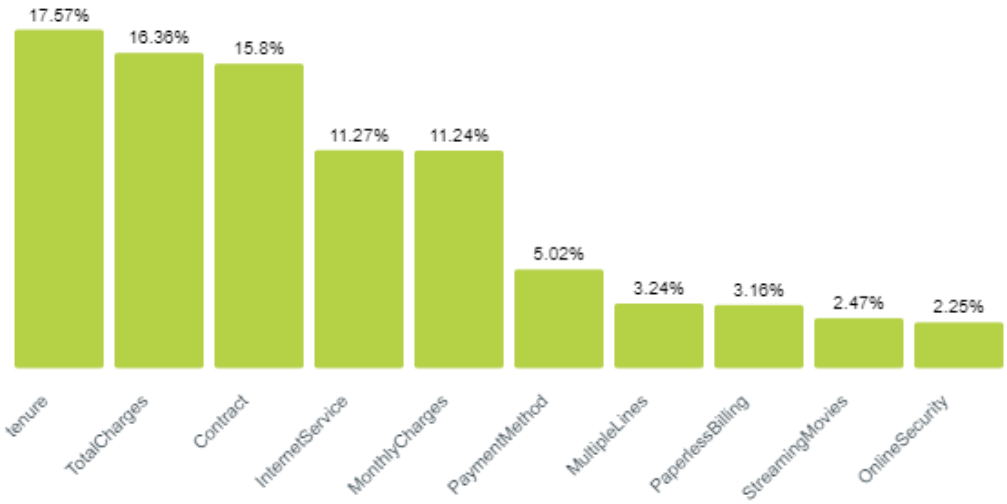
Slika 7: Stablo odlučivanja kreirano u alatu BigML

Što se tiče dubine stabla, prikazan je na sljedećoj slici, a on iznosi 18, odnosno da se od korijenskog čvora do krajnjeg lista na putu prikazanom na slici potrebno je proći kroz 18 čvorova da bi se došlo do krjnjeg lista. Pritom korijenski čvor ima dubinu 0, dok se dubina svakog sljedećeg povećava za 1.

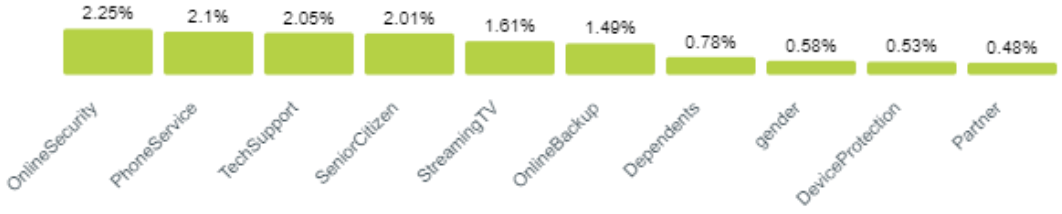


Slika 8: Primjer najveće dubine

Prilikom kreiranja stabla odlučivanja mogući je prikaz atributa po njihovoj važnosti. Konkretni postoci važnosti atributa za prethodno spomenuto stablo odlučivanja biti će prikazani u sljedeće dvije slike.



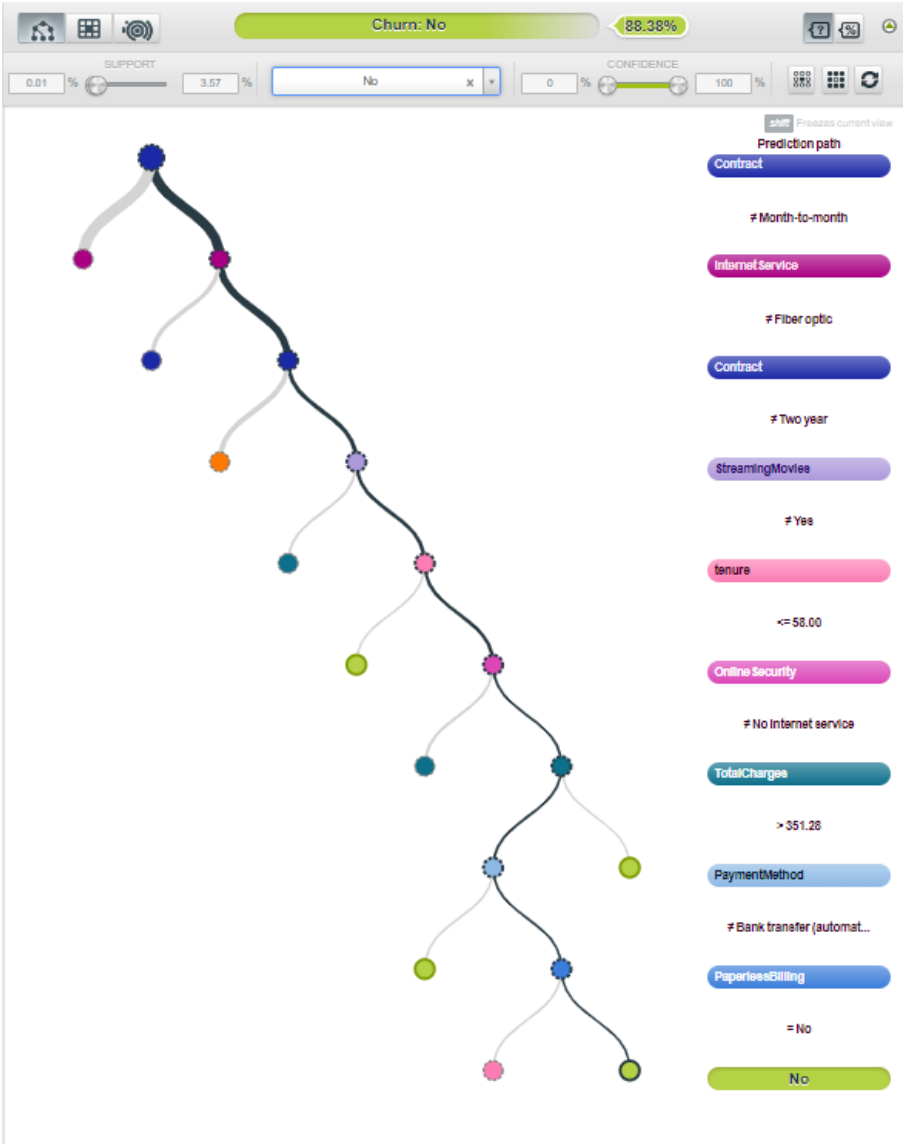
Slika 9: Važnost atributa - 1.dio



Slika 10: Važnost atributa - 2.dio

Prethodne dvije slike prikazuju važnost atributa iz skupa podataka po postocima. Jasno je vidljivo da prvih šest atributa koji su prikazani sa svojim postotkom odskaču od ostatka atributa na prvoj slici te svih ostalih atributa na drugoj slici. Bez imalo sumnje, pogledom na dijagram odmah je vidljivo da je najvažniji atribut u skupu tenure (s važnošću od 21.37%) koji govori koliko je klijent već dugo korisnik telekomunikacijskih usluga u tvrtki. Nakon tenura, sljedeći po

važnosti je atribut Contract (s važnošću od 16.17%) koji govori pod kakvom je ugovornom obvezom klijent, odnosno koliko dugo je pod ugovornom obvezom. Nakon toga slijede atributi InternetService, TotalCharges i MonthlyCharges (s važnostima od 11.53%, 10.19% i 9.84%) koji govore koristi li korisnik internetsku uslugu, koliki su ukupni i mjesečni troškovi. Vidljivo je da su ta tri atributa po važnosti moglo bi se reći slični jer ne odstupaju toliko jedan od drugog po postotku važnosti. Zadnji atribut koji odstupa od ostatka po postotku važnosti je PaymentMethod (s vrijednošću od 5.5%) koji govori o načinu na koji klijent plaća uslugu. Ostalim atributima pada postotak važnosti od 2.92% do 0.48%.



Slika 11: Prikaz pravila odabirom pojedine grupe

Nakon kreiranja stabla odlučivanja, postoji mogućnost pregleda određenog pravila. Pregled pravila dobije se na način da se odabere pojedina grupa, što znači da od početka stabla, na

svakom grananju, odabire se jedna grana odnosno jedan put kojim se ide i tako na svakom sljedećem grananju sve dok se ne dođe do kraja stabla.

Pouzdanost pravila prikazanog na prethodnoj slici:
Contract != Month-to-month and InternetService !=Fiber optic and Contract != Two year and StreamingMovies != Yes and tenure <=58.00 and OnlineSecurity !=No Internet service and TotalCharges > 351.28 and PaymentMethod != Bank transfer and PaperlessBilling = No and No.

Odnosno ako ugovorna obveza nije od-mjeseca-do-mjeseca (što znači da klijent potpisuje ugovor na duže vremensko razdoblje) i ako internetska usluga nije putem optičkih vlakana i ako ugovorna obveza nije na dvije godine i ako klijent ne koristi uslugu streaming filmova i ako klijent koristi usluge tvrtke manje od 58 mjeseci i koristi internetsku zaštitu i ako su ukupni troškovi veći od 351.28 dolara i ako kao oblik plaćanja ne koristi bankovne transakcije i ako ne koristi plaćanje bez računa tada je odgovor za prebacivanje na korištenje usluga druge tvrtke „No“.

6. Zaključak

Strojno učenje jedno je od najbrže rastućih područja računalne znanosti, s dalekosežnim primjenama. Glavni cilj strojnog učenja je proučavanje, projektiranje i poboljšanje matematičkih modela koji mogu biti trenirani za zaključivanje budućnosti i donošenje odluka bez potpunog poznavanja svih utjecajnih elemenat**a.** Stablo odlučivanja jedan je od najmoćnijih alata algoritama za nadzirano učenje koji se koristi za zadatke vezane uz klasifikaciju i regresiju**.**

Nakon teorijske i praktične obrade oba algoritma, može se zaključiti da algoritmi međusobno dijele mnogo sličnosti, no isto tako imaju i mnogo razlika. Oba algoritma razvijena su otprilike u isto vrijeme. CART algoritam pokazao se kao bolji izbor kada je potreban brži pristup informacijama, dok se ID3 algoritam može koristiti i kada nije dostupno mnogo podataka. Kao glavna metrika vrijedna isticanja kod CART algoritma koristi se Gini nečistoća za odlučivanje o optimalnom dijeljenju iz korijenskog čvora i naknadnim dijeljenjima u stablu. ID3 s druge strane za kreiranje stabla odlučivanja odnosno za odabir korijenskog i preostalih čvorova u grananju koristi entropiju i informacijsku dobit.

Popis literature

[1] J.E. Akinsola, *Supervised Machine Learning Algorithms: Classification and Comparison.* Srpanj 2017, Dostupno: https://www.researchgate.net/profile/J-E-T-Akinsola/publication/318338750_Supervised_Machine_Learning_Algorithms_Classification_and_Comparison/links/596481dd0f7e9b819497e265/Supervised-Machine-Learning-Algorithms-Classification-and-Comparison.pdf [pristupano 20.09.2023.]

[2] G. Bonaccorso, Machine Learning Algorithms. Birmingham, UK: Packt Publishing. 2017.

[3] Javatpoint, „Supervised Machine Learning“ [na internetu], dostupno: https://www.javatpoint.com/supervised-machine-learning [pristupano 25.09.2023.]

[4] Kili, Edouard d'Archimbaud, „Supervised Machine Learning: What Is It?“ [na internetu], dostupno: https://kili-technology.com/data-labeling/machine-learning#14 [pristupano 26.09.2023.]

[5] Medium, Sanatan Mishra, „Unsupervised Learning and Data Clustering“, svibanj 2017., [na internetu], dostupno: https://towardsdatascience.com/unsupervised-learning-and-data-clustering-eeecb78b422a [pristupano 05.10.2023.]

[6] Synopsys, „What is reinforcement learning?“ [na internetu], dostupno: https://www.synopsys.com/ai/what-is-reinforcement-learning.html [pristupano 05.10.2023.]

[7] GeeksForGeeks, „Reinforcement learning“, [na internetu], dostupno: https://www.geeksforgeeks.org/what-is-reinforcement-learning/ [pristupano 06.10.2023.]

[8] GeeksForGeeks, „Decision Tree“, [na internetu], dostupno: https://www.geeksforgeeks.org/decision-tree/ [pristupano 09.01.2024.]

[9] IBM, „Decision Trees“, [na internetu], dostupno: https://www.ibm.com/topics/decision-trees [pristupano 09.01.2024.]

[10] Medium, Sandeep Sharma, „Decision Tree and its types“, studeni 2021., [na internetu], dostupno: https://sid-sharma1990.medium.com/decision-tree-and-its-types-76db66644622 [pristupano 09.01.2024.]

[11] IBM, „What is overfitting?“, [na internetu], dostupno: https://www.ibm.com/topics/overfitting [pristupano 09.01.2024.]

[12] GeeksForGeeks, „CART(Classification and Regression Tree) in Machine Learning“, [na internetu], dostupno: https://www.geeksforgeeks.org/cart-classification-and-regression-tree-in-machine-learning/ [pristupano 10.01.2024.]

[13] Explorium, „The Complete Guide to Decision Tree Analysis“, Kolovoz 2023., [na internetu], dostupno: https://www.explorium.ai/blog/machine-learning/the-complete-guide-to-decision-trees/#:~:text=1974%3A%20Statistics%20professors%20Leo%20Breiman,invented%20the%20first%20CART%20version [pristupano 10.01.2024.]

[14] Morris, „Classification & Regression Trees (Decision & Regression Trees), [na internetu], dostupno: https://mnstats.morris.umn.edu/multivariatestatistics/cart.html , [pristupano 15.01.2024.]

[15] Medium, Steven Loaiza, „Gini Impurity Measure – a simple explanation using python“, ožujak 2020., [na internetu], dostupno: https://towardsdatascience.com/gini-impurity-measure-dbd3878ead33 [pristupano 20.02.2024.]

[16] Medium, Hidir Yesiltepe, „What is Gini impurity?“, ožujak 2021., [na internetu], dostupno: https://hidir-yesiltepe.medium.com/what-is-gini-impurity-b821dfb63b6e [pristupano 05.03.2024.]

[17] Victor Zhou, „A simple Explanation of Gini Impurity“, ožujak 2019., [na internetu] dostupno: https://victorzhou.com/blog/gini-impurity/ [pristupano 06.03.2024.]

[18] „Telco Customer Churn“ (bez datuma) [na internetu] dostupno: https://www.kaggle.com/blastchar/telco-customer-churn [pristupano 21.04.2024.]

[19] Medium, Ashirbad Pradhan, „Decision Tree ID3 Algorithm| Machine Learning“, 11.lipanj 2023., [na internetu], dostupno: https://medium.com/@ashirbadpradhan8115/decision-tree-id3-algorithm-machine-learning-4120d8ba013b [pristupano 15.05.2024.]

[20] GeeksForGeeks, „Sklearn | Iterative Dichotomiser 3 (ID3) Algorithms“, 22.svibnja 2024., [na internetu], dostupno: https://www.geeksforgeeks.org/sklearn-iterative-dichotomiser-3-id3-algorithms/ [pristupano 25.05.2024.]

[21] Journal of Clinical Images and Medical Case Reports, Zeye Liu, Xiangbin Pan, „Comparison and analysis of applications of ID3, CART decision tree models and neural network model in medical diagnosis and prognosis evaluation“, 04.svibnja 2021., [na internetu], dostupno: https://jcimcr.org/pdfs/JCIMCR-v2-1101.pdf [pristupano 27.05.2024.]

[22] Sefik Ilkin Serengil, „A Step by Step ID3 Decision Tree Example“, 20.studeni 2017., [na internetu], dostupno: https://sefiks.com/2017/11/20/a-step-by-step-id3-decision-tree-example/ [pristupano 27.05.2024.]

[23] GeeksForGeeks, „Iterative Dichotomiser 3 (ID3) Algorithm From Scratch“, 02.siječnja 2024., [na internetu], dostupno: https://www.geeksforgeeks.org/iterative-dichotomiser-3-id3-algorithm-from-scratch/ [pristupano 01.06.2024.]

Popis slika

Popis tablica