Primjena stabla odlučivanja na skupu podataka iz obrazovanja

Žitković, Bruno

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Bruno Žitković

Primjena stabla odlučivanja na skupu podataka iz obrazovanja

ZAVRŠNI RAD

Varaždin, 2020.

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Bruno Žitković

Matični broj: 0016130386

Studij: Informacijski sustavi

Primjena stabla odlučivanja na skupu podataka iz obrazovanja

ZAVRŠNI RAD

Mentor/Mentorica:

Doc. dr. sc. Dijana Oreški

Varaždin, kolovoz 2020.

Bruno Žitković

Izjava o izvornosti

Izjavljujem da je moj završni/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.

Autor/Autorica potvrdio/potvrdila prihvaćanjem odredbi u sustavu FOI-radovi

Sažetak

Ovaj završni rad sastojat će se od teorijskog dijela gdje će biti opisana prediktivna metoda stablo odlučivanja i praktičnog dijela primjene metode. U uvodu će se odgovoriti na pitanja što je to zapravo stablo odlučivanja te će se opisati pojmovi koji su nam bitni za razumijevanje teme. U nastavku rada će se obraditi algoritmi stabla odlučivanja, te definirati zahtjevi na varijable koje vode do pouzdanih i točnih prediktivnih metoda. U nastavku, razrada će nam dati odgovore na to koje su prednosti, ali i mane stabla odlučivanja u odnosu na druge prediktivne metode, a u zaključku će se navesti te u kratkim crtama prepričati najznačajniji pojmovi koji su bili obrađeni.

Također, na temelju teorijskog dijela rada, metoda stablo odlučivanja primijenit će se na javno dostupnom skupu podataka iz obrazovanja. U praktičnom dijelu dobiveni rezultati će se interpretirati i dat će se smjernice koje će služiti kao potpora odlučivanju.

Ključne riječi: stablo odlučivanja; rudarenje podataka; strojno učenje; obrazovanje

Sadržaj

1. Uvod

Živimo u svijetu gdje su podaci centralni izvor informacija i gdje svaka naša odluka utječe na buduće događaje. Vrlo je bitno pravilno upotrijebiti dane nam podatke i analizirati ih kako bismo uspjeli što jednostavnije savladati zadatak koji je pred nama. Cilj ovog rada biti će kvalitetno obraditi temu stablo odlučivanja kroz teorijsku razinu, ali isto tako i kroz praktični dio gdje ćemo na stvarnom primjeru vidjeti kako nam stablo odlučivanja može dati moguća rješenja s obzirom na zadane podatke.

Ovaj rad će donijeti detaljan opis metoda koje stablo odlučivanja koristi za prediktivnu analizu podataka kako bi poslovne odluke bile što kvalitetnije i kako bi ishod tih odluka bio što uspješniji poslovni rezultat. Objasnit ćemo metode poput ID3, C4.5 i CART, ali isto tako i kriterije podjele koje stablo odlučivanja koristi, pojam entropije i informacijske dobiti, Gini indeks-a, Chi-Squared-a, ali i što je to omjer dobiti.

U radu će također biti govora o kriterijima zaustavljanja, ali i o tome što nakon kada algoritam stabla odlučivanja završi, poznat pod pojmom obrezivanja. Prednosti i nedostaci stabla odlučivanja dati će dobru podlogu kako bi se krenulo na praktični dio iz stvarnog života te primjenu algoritma stabla odlučivanja nad podacima iz obrazovanja. Iz svega ranije navedenog će se izvesti zaključak.

2. Stablo odlučivanja

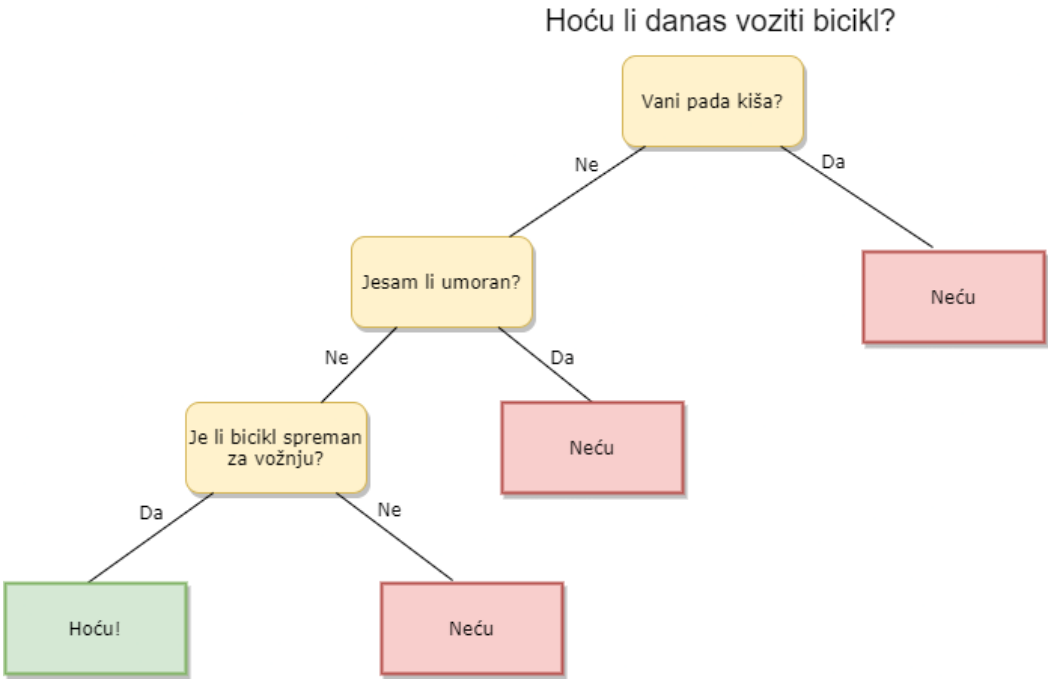
Stablo odlučivanja je grafički prikaz rješenja temeljenog na mnogobrojnim uvjetima koji su zadani. Stablo odlučivanja služi kako bi se na temelju različitih podataka donijela neka odluka. Bila to jednostavna predikcija o kvaliteti hrane u određenom restoranu, pa sve do kompleksnijih analiza poput utjecaja alkohola na mladenačku populaciju. Stabla odlučivanja nam daju različit set alata kako bi savladali sve prepreke i na kraju dobili vrlo konkretan i smislen rezultat. (Kozak, 2019) navodi kako su stabla odlučivanja vrlo bitna u primjeni na različitim skupovima podataka zbog njihove jednostavnosti i efektivnosti (u smislu računske složenosti) u procesu klasificiranja različitih objekata.

„Stablo odlučivanja se sastoji od čvorova koji u konačnici tvore stablo. Početni čvor se naziva korijen i on nema ulaznih grana. Svi ostali čvorovi imaju točno jednu ulaznu granu. Čvor koji ima izlazne grane naziva se testni čvor. Svi ostali čvorovi nazivaju se listovi (zvani i terminalni čvorovi).“ (Rokach i Maimon, 2015., str. 12).

Na temelju iznad navedene definicije možemo zaključiti kako je stablo odlučivanja u teoriji vrlo jednostavan princip predikcije koji se sastoji od početnog čvora kojeg nazivamo korijenom i gdje je sadržan cijeli skup podataka. Nadalje, korijen se dijeli na dva nova čvora, dva nova čvora na dva nova i tako dalje sve dok se ne dođe do krajnjeg rezultata. Taj novonastali čvor koji je proizašao iz čvora odluke nazivamo inačicom odluke. Bitno je za napomenuti kako svaki sljedeći čvor sadrži manji skup podataka od svog prethodnika.

Prema (Marijana Zekić-Sušac, 2017) konačno stablo odlučivanja dobivamo tako da vršimo grananje podataka s obzirom na naš početni skup, tj. učenjem na tim podacima, a to grananje temeljimo na testiranju vrijednosti određenih varijabli. To znači da ćemo taj proces učenja, tj. grananja izvoditi sve dok ne dobijemo rezultat koji daljnjim dijeljenjem više nema nikakav značajni utjecaj na rezultat koji očekujemo.

Možemo još spomenuti kako na svakom čvoru odluke biramo onu varijablu prema kojoj ćemo vršiti daljnje dijeljene koja će nam maksimizirati rezultat dijeljenja. Kako to stablo odlučivanja zna nad kojom varijablom treba vršiti podjelu skupa podataka kako bi maksimizirao očekivani rezultat? Odgovor na to glasi u raznim algoritmima koje koji se koriste kako bi se odabrala ispravna varijabla. Sam naziv algoritama i kako oni funkcioniraju bit će govora nešto kasnije u radu.



Slika 1: Jednostavni primjer stabla odlučivanja (autorski rad)

Na slici 1. možemo vidjeti jednostavan primjer stabla odlučivanja. Kao rezultat stabla odlučivanja želimo saznati koje su šanse da biciklist danas ide u vožnju biciklom. Pitamo se pada li vani kiša? Kao rezultat odluke dobivamo podjelu stabla na dva čvora. 50% početnog skupa odlazi u čvor gdje je odgovor bio da, a 50% u odgovor ne. Vidimo da ako pada kiša naš promatrani skup od 50% tu završava i taj čvor nazivamo terminalni čvor. Ostatak algoritma se nastavlja sve dok ne dođemo do jednog od mogućih rješenja, a u ovom slučaju je to da će biciklist ići u vožnju ako ne pada kiša, ako nije umoran i ako je bicikl spreman za vožnju.

U ovom uvodnom dijelu možemo još spomenuti koje su to primjene stabla odlučivanja u stvarnom životu i koliko je taj algoritam zapravo korišten. Naime, pošto je stablo odlučivanja, kao što mu i samo ime govori, bitan alat kod donošenja odluka unutar organizacija, njegova rasprostranjenost je vrlo široka.

Kao jednostavan primjer daje se korištenje stabla odlučivanja u avio prijevozu kod odabira leta vožnje. Prvobitno nas zanima postoji li let u vrijeme kada mi to trebamo, ako postoji onda nas zanima koliko dugo taj let traje, ako nam i to odgovara dalje tražimo cijenu leta i na kraju koliko ćemo taj let platiti. Sve te mogućnosti algoritam stabla odlučivanja uzima u obzir i daje nam najprikladniji rezultat.

Stablo odlučivanja primjenjuje se i u kriminalnim aktivnostima gdje se na temelju podataka koji se nalaze u bazi podataka rade analize i predikcije gdje policiju zanima kolika je vjerojatnost da će netko počiniti zločin. Atributi koji se u tom slučaju stavljaju na vrh su oni koji su važniji od ostalih, tako npr. najvažniji ili jedan od najvažnijih će sigurno biti je li čovjek već počinio zločin u prošlosti.

Devlin i Lorden (2007, str. 38) opisuju jednu zanimljivu situaciju gdje je korištenje stabla odlučivanja vrlo korisno. Uzimajući u obzir razne karakteristike putnika u avionu odredili su kolika je vjerojatnost da će ga granična policija pustiti u zemlju.

1. Putnik je starosti 20-25 godina
2. Muškarac je
3. Dolazi iz Saudijske Arabije
4. Živi u Njemačkoj
5. Student je, ali sveučilište je nepoznato
6. U zadnjih godinu dana 3 puta je ušao u zemlju
7. Zemlje koje je posjetio u zadnje 3 godine su bile Ujedinjeno Kraljevstvo i Pakistan
8. Putnik ima dozvolu pilota

Uzimajući u obzir prvih 6 točaka policajac bi najvjerojatnije pokrenuo istragu, a ostale dvije bi rezultirale konačnim rješenjem, a to je deportacija iz zemlje. Iako je algoritam koji se u stvarnosti primjenjuje mnogo složeniji i vjerojatno bi uzeo mnogo više stvari u obzir, na ovom jednostavnom primjeru možemo vidjeti kako bi vjerojatnost da je putnik terorist iznosila 29% i akcija koja bi uslijedila je da ga se ne pusti u zemlju.

Još neki primjeri korištenja stabla odlučivanja u stvarnom životu su odabir najkvalitetnijih jajašca za unutarnju oplodnju koji se provodio u Ujedinjenom Kraljevstvu 2000. godine. Isto tako primjena je i u poljoprivredi, a primjer toga je Indija gdje je soja jedan od najvažnijih izvoznih proizvoda. (Veenadhari, Mishra i Singh, 2011) navode kako usprkos povećanju obradive površine proizvodnja soje je pala. Rješenje za povećanje proizvodnje upravo je bilo korištenje stabla odlučivanja na dugoročnim meteorološkim podacima gdje se došlo do različitih analiza predvidljivosti i utjecaju različitih meteoroloških podataka na prinos soje.

2.1. Algoritmi stabla odlučivanja

Kako bi uopće mogli raspravljati o tome koliko je stablo odlučivanja korisna prediktivna metoda, moramo se prvo dotaknuti načina na koji stablo odlučivanja donosi svoje konačne rezultate. U tu svrhu, postizanja optimalnih i konačnih rješenja stablo odlučivanja koristi svojstvene algoritme. Postoje različite vrste algoritama, neki složeniji, a neki jednostavniji te prikladno svrsi u kojoj se koriste odlučuje se koji je najprikladniji za određenu situaciju.

Stablo odlučivanja će, kao što je ranije bilo navedeno u svakom koraku dijeliti skup podataka na nove grane sve dok se ne dođe do terminalnog čvora, ili nova podjela više nema smisla. Kako stablo zna u kojem trenutku treba doći do podjele i prema kojem kriteriju se taj skup dijeli? Upravo u tom slučaju uskaču algoritmi koji nam olakšavaju rad sa stablom odlučivanja i taj sveobuhvatni posao koji je krucijalan kako bi se došlo do što preciznijeg i pouzdanijeg rješenja te da ga se svede na što jednostavniju operaciju, koliko je to u određenoj situaciji moguće.

„Kod podjele stabla odlučivanja koristi se više algoritama kako bi se donijela odluka gdje se podjela treba izvršiti i koliko ona treba iznositi. Postoji značajan rast u homogenosti kod stvaranja novih pod-čvorova. To znači da će čistoća skupa rasti s porastom ciljne varijable.“ (Sullivan, 2017., str. 32).

Sullivan govori kako porastom podjela stabla odlučivanja na nove čvorove raste i homogenost istih te samim time i čistoća samog skupa podataka nad kojim radimo. Algoritmi koji će biti obrađeni u ovome radu su:

1. ID3
2. C4.5
3. CART

Počet ćemo s najjednostavnijim, ID3 algoritmom koji se najčešće služi kao algoritam za učenje te ćemo se postepeno nadograđivati s C4.5, a zatim i CART. Spomenut ćemo isto tako koje kriterije podjele sami algoritmi koriste, a kasnije će i sami kriteriji biti podrobnije objašnjeni u svojim poglavljima zbog njihove velike važnosti za sam proces stabla odlučivanja i način na koji nam se pruža krajnji rezultat.

2.1.1. ID3

Prvi algoritam koji će biti obrađen je možda i najjednostavniji, ali nikako ne i najlošiji za primjenu nad stablom odlučivanja. ID3 znači „Iterative Dichotomiser 3“, što bi u prijevodu bilo iterativno, ponavljajuće dijeljenje skupa podataka na novi skup ili skupove prilikom svakog prolaska algoritma. ID3 algoritam je izumio Ross Quinlan na principu CLS algoritma, a danas se u velikoj većini koristi kao algoritam s kojim se počinje učenje u svijetu rudarenja podataka. Kasnije, na principu ID3 algoritma razvio se i njegov nasljednik C4.5, a o njemu će biti riječi nešto kasnije.

Kao što navodi (Sakkaf, 2020), ID3 algoritam koristi „top-down greedy“ metodu kako bi izgradio stablo odlučivanja. U suštini, to znači da stablo počinjemo graditi od vrha te se postepeno spuštamo prema dnu, a pohlepna odlika ovog postupka je ta da prilikom svakog tog spuštanja, tj. prilikom svake iteracije, odabiremo najbolji mogući atribut, koji je u tom trenutku dostupan, nad kojim ćemo vršiti daljnju podjelu. ID3 se najčešće koristi s podacima nominalnog tipa.

Kako ID3 zna prema kojem atributu je najbolje vršiti podjelu? Prema („Computer & Information Science & Engineering“, bez dat.) koristi se statistička vrijednost koja se zove informacijska dobit. Ona mjeri koliko dobro, s kojom preciznošću dani nam atribut dijeli skup nad kojim želimo vršiti podjelu na neki novi skup ili skupove. Onaj atribut koji izračunom ima najveću informacijsku dobit ima prednost nad ostalima i on se uzima kao točka podjele. Kako bi se definirala informacijska dobit koristi se teorem pod nazivom entropija. Entropija i informacijska dobit će biti detaljnije objašnjeni kasnije u radu, zasad nam je jedino bitan princip kako ID3 radi.

(Sakkaf, 2020) također opisuje 5 koraka koji se koriste prilikom primjene ID3 algoritma nad skupom podataka kod stabla odlučivanja, a to su:

1. Izračunavanje informacijske dobiti svakog atributa
2. Podjela skupa S u podskupove na temelju atributa koji je imao najveću informacijsku dobit.
3. Izrada čvora stabla odlučivanja koristeći ranije navedeni atribut.
4. Ako svi redovi pripadaju istoj klasi, taj čvor postaje list s oznakom te klase.
5. Ponavljaj sve dok se ne iscrpe svi atributi ili stablo ima sve čvorove listove.

2.1.2. C4.5

Drugi algoritam stabla odlučivanja koji će biti objašnjen je C4.5. Upravo je on, kao što je bilo ranije spomenuto, nadogradnja na ID3, a tu nadogradnju je isto tako izvršio Ross Quinlan. Poboljšana verzija u odnosu na ID3 nudi generaliziranije modele koji isto tako mogu uključivati kontinuirane brojčane podatke, a isto tako prednost C4.5 naprema ID3 je ta što C4.5 može raditi i s podacima koji nedostaju.

C4.5 koristi takozvani omjer dobitka iliti „gain ratio“ kao kriterij podjele, naspram njegovog prethodnika ID3, koji koristi informacijsku dobit. Omjer dobitka je ništa drugo nego isto tako poboljšana informacijska dobit koja smanjuje odstupanje. Rokach i Maimon (2015, str. 78) spominju kako podjela prestaje onog trenutka kada broj instanci koji će biti podijeljen je manji od određene granice. Isto tako navode kako se „prunning“ tj. modificiranje ili čišćenje stabla odvija nakon što je stablo izraslo.

(Synergy37AI ,2019) navodi kako C4.5 pretvara trenirana stabla (npr. izlaze onoga što kreira ID3 algoritam) u setove koji predstavljaju ako-onda pravila. Preciznost svakog tog pojedinačnog pravila je onda evaluirana kako bi se naposljetku mogao odrediti redoslijed prema kojem će se ta pravila poredati.

Prednosti koje C4.5 algoritam donosi i što ga to odlikuje kao veliki iskorak naspram njegovog mlađeg brata ID3, možda i najbolje opisuju Rokach i Maimon (2015., str. 78) u 3 koraka. Ta 3 koraka glase:

1. C4.5 koristi obrezivanje koje miče grane stabla koje nikako ne pridonose pouzdanosti stabla u cjelini i pretvara ih u listove.
2. C4.5 kao što je već bilo ranije navedeno dopušta da u skupu podataka postoje atributi koji nedostaju.
3. C4.5 algoritam najveću prednost nad ID3 algoritmom ima u tome što može raditi s kontinuiranim podacima tako što dijeli vrijednost atributa u 2 podskupa (tkz. Binarna podjela). Isto tako, traži najbolju moguću granicu koja maksimizira kriterij omjer dobitka. Sve one vrijednosti koje se nalaze iznad granice, pripadaju jednom skupu, a sve one koje se nalaze ispod, pripadaju drugome.

Za kraj možemo još spomenuti kako C4.5 ima isto tako svoju nadogradnju u vidu C5.0 algoritma te se on uvelike koristi u komercijalne svrhe, a pruža dosta poboljšanja gdje su samo neka od njih bolja efikasnost što se tiče memorije i složenosti.

2.1.3. CART

CART je algoritam koji je nešto složeniji od ranije opisanih ID3 i C4.5 algoritama i u prijevodu znači klasifikacijsko i regresijsko stablo. Algoritam je razvijen od strane Lea Breimana i prvi put se spominje u istoimenoj knjizi CART 1984. godine. Ono što odlikuje CART algoritam je činjenica da se pomoću njega konstruiraju binarna stabla.

(Kozak, 2019, str. 20) u svojoj knjizi Decision Tree and Ensemble Learning Based on Ant Colony Optimization govori o tome kako CART naveliko traži mjesta gdje bi podijelio skup podataka da bi minimizirao kvadratno predviđanje pogrešaka. Nadalje, dotiče se toga da je stablo odlučivanja građeno na principu višestrukog dijeljenja trening podataka na manje dijelove, a da se to najčešće odvija metodom „podijeli pa vladaj“. U kratkim crtama to znači da bi trebali biti u mogućnosti podijeliti problem, riješiti ga na što jednostavniji mogući način i na kraju to sve spojiti.

Faza rasta stabla odlučivanja korištenjem CART algoritma može se podijeliti u 3 koraka koja je opisao (Breiman, 1987), a koraci glase:

1. Pronađi za svaku kontinuirani i ordinalni atribut najbolju podjelu i sortiraj njihove vrijednosti od najmanjeg do najvećeg. Za svaku vrijednost ispitaj točku dijeljenja.
2. Pronađi najbolju točku dijeljenja, a to je ona koja maksimizira kriterij dijeljenja.
3. Podijeli čvor koristeći najbolju točku dijeljenja pronađenu pod točkom 2.

Postoje 3 kriterija podijele koji se koriste u skladu s CART metodologijom, a to su Gini, Twoing i poredani Twoing. U ovom radu detaljnije će se opisati Gini indeks kada će biti riječi o kriterijima podjele. Sada je bitno za napomenuti kako Gini indeks određuje koliko je neki čvor čist.

Kao i kod prethodno opisanih algoritama ID3 i C4.5, CART isto tako treba znati kada stati s izradom novih čvorova. Možda i najbolji opis toga daje (Brownlee, 2016) govoreći kako je najčešći razlog zaustavljanja taj da ako podjela novonastalog čvora rezultira s brojem instanci koji je manji od nekog broja koji je ranije definirao korisnik za određeni čvor, onda se taj čvor neće dijeliti i tu stablo završava te taj čvor postaje terminalni tj. završni čvor.

2.2. Kriterij podjele

U ranijim poglavljima bilo je opisano stablo odlučivanja, kako ono funkcionira i algoritme koje koristi da bi se postigli željeni rezultati. U ovom poglavlju detaljnije će se objasniti ranije spomenuti kriteriji podjele koji služe kao pomoć algoritmima u određivanju, kao što im i samo ime govori, gdje će se podjela u skupu podataka izvršiti.

Kako kriteriji podjele odabiru određeni atribut nad kojim će se izvršiti podjela ili koju granicu treba postaviti da bi ta podjela imala smisla, bit će raspravljano kroz 4 različita kriterija podjele, a to su:

1. Entropija i informacijska dobit
2. Gini indeks
3. Chi-Squared
4. Omjer dobitka

Nad tim kriterijima vidjet će se njihova razlika u određivanju gdje će se dogoditi podjela, u kojim situacijama se koji kriterij koristi, ali isto tako koji već ranije spomenuti algoritam isti upotrebljava u svrhu dobivanja konačnog rezultata.

U ovom uvodnom dijelu možemo se još dotaknuti različitih kriterija prema kojima se određuje podjela koje opisuju Rokach i Maimon (2015., str. 61), te navode sljedeće:

1. U skladu s odredištem mjerila, bilo ono iz informacijske teorije, zavisnosti ili udaljenosti.
2. U skladu sa strukturom mjerila: kriterij čistoće, normaliziran kriterij čistoće, ili bio to binarni kriterij.

2.2.1. Entropija i informacijska dobit

Kada govorimo o entropiji moramo naglasiti kako se taj pojam spominje kroz različita područja, od fizike pa sve do povijesti umjetnosti. Iako u konačnici je u svim tim područjima značenje donekle slično, za stablo odlučivanja bitna je ona informatička entropija koja prema (Enciklopedija.hr, bez dat.) glasi da je to mjera neodređenosti informacija, tj. postupak obradbe informacije gdje se naposljetku gubi dio informacije.

Ukratko možemo reći kako je entropija mjera s kojom mjerimo čistoću ili nasumičnost nekog skupa s kojim radimo. Kao primjer entropije možemo posuditi od (Molala, 2019) gdje spominje, ako uzmemo žutu loptu iz kutije žutih lopti i zamislimo da je tih lopti 100. Možemo reći kako ta kutija onda ima entropiju od 0 što implicira 0 nečistoće ili potpunu čistoću. Doduše, ako 30 tih lopti zamijenimo s crvenim, a 20 s plavim loptama i nakon toga opet izvlačimo jednu iz kutije, dobit ćemo da je vjerojatnost izvlačenja žute lopte pala s 1 na 0.5.

Na prethodnom primjeru možemo vidjeti kako se je nečistoća povećala, a s njom i sama entropija. Entropija je najčešće korištena od strane ID3 i C4.5 algoritma, a formula entropije glasi:

$$Entopija = \sum -\rho_i log_2 \rho_i$$

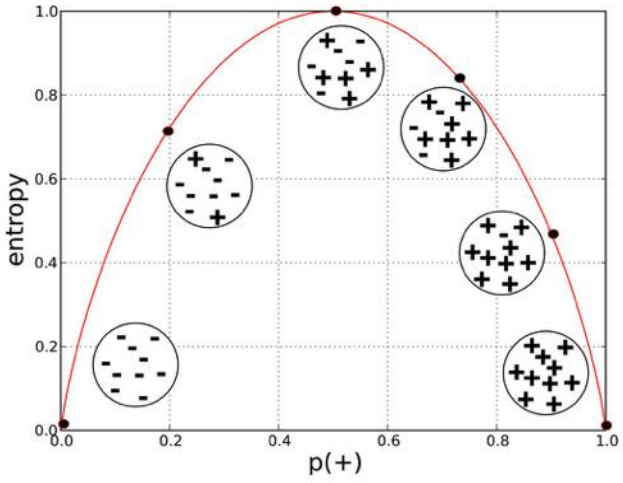
Gdje je p vjerojatnost klase izražena od 0 do 1 i log_2 je algoritam s bazom 2. Miller i Massaron (2016, str. 205) navode primjer gdje, ako postoji neki uzorak gdje se žele klasificirati dvije klase koje imaju istu vjerojatnost s distribucijom od 50/50, najveća moguća entropija onda iznosi:

$$Entropija = -0.5 * log_2(0.5) - 0.5 * log_2(0.5) = 1$$

Doduše, ako taj omjer distribucije od 50/50 padne na recimo 40/60 u skladu s time i prosječna entropija pada te ona iznosi:

$$Entropija = -0.4 * log_2(0.4) - 0.6 * log_2(0.6) = 0.97$$

Nadalje spominju kako je očigledno da je suma manje od teoretskog maksimuma koji iznosi 1 te opisuju to kao nered u podacima. Što je manje nereda, to je više reda pa je stoga lakše pogoditi ishod.



Slika 2. Prikaz entropije na grafu (https://hr.sciencewal.com/, bez dat.)

Na slici 2. možemo vidjeti graf koji prikazuje entropiju te iz njega zaključujemo kako najmanje što entropija može iznositi je 0, a to znači da vjerojatnost ishoda mora biti 0 ili 1, dok je entropija najveća kada je vjerojatnost ishoda 0.5.

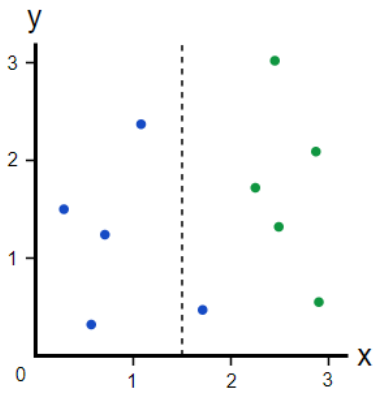
Sada kada smo objasnili što je to entropija i kako se ona računa, možemo se dotaknuti informacijske dobiti. (Brownlee, 2019) govori o informacijskoj dobiti kao mjeri koja mjeri smanjenje entropije, tako što dijeli skup podataka s obzirom na vrijednost dane nam slučajne varijable. Dok recimo (Guanga, 2019) definira informacijsku dobit kao razliku entropije prije i poslije nekog događaja. Nadalje, da nam je cilj dobiti najznačajniju informacijsku dobit, a samim time i najmanju entropiju.

Informacijska dobit je korištena od strane ID3 algoritma i formula po kojoj se ona izračunava glasi:

$$Informacijska\ dobit\ (S,A) = Entropija(S) - \sum \frac{|S_v|}{|S|}\ Entropija(S_v),$$

Gdje je (S_v) podskup skupa S za koji atribut A ima vrijednost v. Iz navedene formule izračunavamo informacijsku dobit i onaj atribut s najvećom dobiti se odabire kao kriterij podjele, prema (Yang, Li, i Song , 2007). Isti autori navode kako jedna od najvećih mana ID3 algoritma je upravo korištenje informacijske dobiti kao kriterij podjele zato što informacijska dobit preferira atribute s prepoznatljivim vrijednostima kojih ima mnogo. Taj problem je donekle riješen upotrebom C4.5 algoritma i korištenjem omjera dobiti kao kriterij podjele, a o njemu će biti govora nešto kasnije.

Primjer informacijske dobiti temeljen na (Zhou 2019) koji se sastoji od plavih i zelenih loptica nakon što se dogodila podjela.



Slika 3. Prikaz ID na grafu (https://victorzhou.com/blog/information-gain/, 2019)

Na slici 3. možemo vidjeti kako su na lijevoj strani podjele raspoređene 4 plave loptice, dok su na desnoj 1 plava i 5 zelenih. Prije ove podijele bilo je 5 zelenih i 5 plavih i entropija je iznosila 1. Izračunavajući entropiju lijeve strane dobijemo da ona iznosi 0 zato što što su sve loptice iste boje. Entropija desne strane iznosi 0.65. Nadalje, izračunavamo entropiju cjelokupne podjele:

$$E_{podjele} = 0.4 * 0 + 0.6 * 0.65 = 0.39$$

Sada možemo izračunati informacijsku dobit:

$$ID = 1 - 0.39 \ = 0.61$$

Vidimo kako oduzimanjem entropije podjele od entropije prije podjele dobivamo da informacijska dobit iznosi 0.61. Kada bi se dogodila savršena podjela gdje bi sa svake strane bile loptice iste boje onda bi entropija iznosila 0.

2.2.2. Gini nečistoća i gini indeks

Kada govorimo o pojmovima gini nečistoće i gini indeksa moramo spomenuti kako su oni vrlo slični informacijskoj dobiti i entropiji. Gini nečistoća odnosi se na nesrazmjer unutar čvora. Ayyadevara (2018, str. 79) opisuje gini nečistoću kroz primjer, gdje, ako čvor ima 50% elemenata jedne klase, a 50% elemenata neke druge klase to je onda najnečistiji oblik koji čvor može imati. Nadalje, gini nečistoću definira kroz formulu koja glasi:

$$GI = 1 - (\rho^2 + \ q^2)$$

Gdje su p i q vjerojatnosti koje se vežu za pojedinu klasu. Sada kada smo definirali što je to gini nečistoća, možemo se posvetiti gini indeksu. Gini indeks nije ništa drugo nego kriterij za određivanje nečistoće koji mjeri razliku između distribucije vjerojatnosti zadanih atributa određene klase, kako navode Rokach i Maimon (2015., str. 62).

Gini indeks je ponderirana suma gini nečistoće i on kombinira šum u kategoriji kako bi se iz toga izvukao šum atributa, isto tako (Bui , bez dat.) daje formulu za gini indeks:

$$I_{Gini} = \sum_{k \in M} P_{k,a} * Gini(k)$$

Gdje je a svojstvo, M lista svih kategorija u svojstvu a, dok je $P_{k,a}$ frakcija kategorije k u svojstvu a.

2.2.3. Chi-Squared

Chi-Squared je kriterij podjele koji se zasniva na pronalaženju statističke značajnosti u odstupanju dvaju čvora. Točnija definicija bi glasila:

„Chi-Squared je korišten kako bi se pronašla statistička značajnost odstupanja između čvora roditelja i čvora djeteta. Računa se kao suma svih kvadrata svih odstupanja između očekivane frekvencije ciljane varijable i promatrane frekvencije ciljane varijable*." (Sullivan, 2017., str. 34)*

Ono što odlikuje Chi-Squared i što ga izdvaja kao poseban kriterij podjele od ranije navedenih, prema (Jain, 2017) je sljedeće:

1. Radi s kategoričkim varijablama „uspjeh" ili „neuspjeh".
2. Može napraviti dvije ili više podjela.
3. Što je veća vrijednost Chi-Square-a veći je statistički značaj razlike između čvora roditelja i čvora djeteta.
4. Chi-Square svakog čvora se računa prema sljedećoj formuli:

$$ChiSquare = \left(\frac{(promatrana\ frek.var. - očekivana\ frek.var.)^2}{očekivana\ frek.var.}\right)^{1/2}$$

5. Generira se stablo koje nazivamo CHAID (engl. Chi-Square Automatic Interaction Detector).

(Sullivan, 2017, str. 34) spominje 2 nužna koraka koja se moraju proći u izračunavanju Chi-Square-a, a to su:

1. Izračunaj Chi-Square za individualni čvor tako što izračunaš odstupanje i za „uspjeh" i „neuspjeh".
2. Izračunaj Chi-Square podjelu koristeći sumu svih Chi-Square-a „uspjeha" i „neuspjeha" za svaki čvor podjele.

2.2.4. Omjer dobitka

Omjer dobitka je u suštini poboljšana verzija informacijske dobiti, odlikuje ga to što za razliku od informacijske dobiti koju koristi ID3 algoritam, omjer dobitka smanjuje odstupanje kada se radi s velikom količinom atributa.

(Quinlan, 1993) daje formulu po kojoj se izračunava omjer dobitka:

$$OmjerDobitka = \frac{Informacijska\ dobit\ (a_i, S)}{Entropija\ (a_i, S)}$$

Ono što Rokach i Maimon (2015., str. 64) ističu je to da ovaj omjer ne vrijedi ako je entropija 0. Nadalje, naglašavaju važnost činjenice da omjer dobitka favorizira atribute čiji je nazivnik vrlo malen, a kao rješenje ističu sljedeće:

1. Informacijska dobit se izračunava za svaki atribut.
2. Odabire se atribut koji je postigao najbolji omjer dobitka.

Quinlan (kao što navode Rokach i Maimon) naglašava kako omjer dobitka ima uveliko bolje rezultate od jednostavne informacijske dobiti i u preciznosti, a i u klasifikacijskoj kompleksnosti. Bitno je još za napomenuti kako C4.5 algoritam kao kriterij podjele koristi upravo omjer dobitka.

2.3. Kriterij zaustavljanja

Do sada smo već mogli zaključiti što je to stablo odlučivanja i na koji on način rad, kako se stablo gradi primjenom različitih algoritama i koje kriterije podjele koristi kako bi ishod podjele bio što kvalitetniji. Međutim, kako stablo ne bi raslo u nedogled i kako taj rast ponekad ne bi bio nepotreban primjenjuju se određeni kriteriji zaustavljanja.

Neki algoritmi, poput ID3-a, kako navodi (J. Kent Martin, 1997) uključuju kriterij zaustavljanja kada se inkrementalno napredovanje svede na beznačajne podjele i taj tip zaustavljanja kolektivno se zove kriterij „pred obrezivanja“. Što je to obrezivanje i kako on o funkcionira bit će objašnjeno kasnije. Nadalje, naglašava kako se ostali algoritmi koriste tkz. rekurzivnom procedurom za „post obrezivanje“ što uključuje zamjenu daljnje podjele terminalnim čvorom. Postoje još i ostali kriteriji gdje neki algoritmi zamjenjuju podjelu nekom drugom podjelom, a primjer toga je C4.5 o kojem je bilo riječi malo ranije.

Postoje određena pravila zaustavljanja koja su limit za zaustavljanje ekspanzije stabla i ta pravila rade uzimajući u obzir 3 aspekta:

1. Inicijalna veličina skupa
2. Krajnja veličina skupa
3. Informacijska dobit nakon izvršene podjele

Kako bismo se dotakli tih pravila treba uzeti u obzir da se ona aktiviraju u određenim trenucima prilikom generiranja stabla odlučivanja. Kada se zadovolje određeni uvjeti i ranije navedena pravila, aktivirat će se zaustavljanje i stablo će prestati rasti. Ta pravila najbolje daju Rokach i Maimon (2015, str. 30):

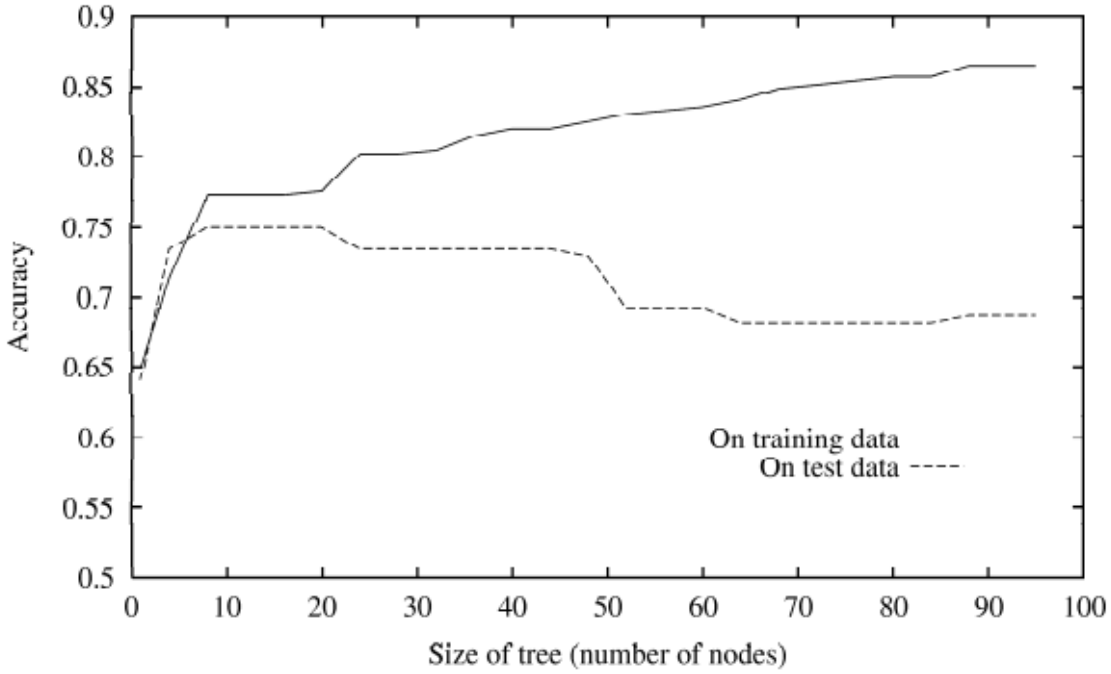
1. Sve instance u setu trening podataka pripadaju jedinstvenoj vrijednosti y.
2. Dostignuta je maksimalna dubina stabla.
3. Broj slučajeva u terminalnom čvoru manji je od minimuma slučajeva koji mora biti u čvoru roditelj.
4. Ako se čvor podijeli, broj slučajeva u čvoru dijete bit će manji od broja slučajeva definiranih za čvor dijete.
5. Najbolji kriterij podjele nije veći od određene granice.

2.4. Pretreniranost i obrezivanje

Ono po čemu se stablo odlučivanja ističe je rekurzivni algoritam koji će u konačnici stalno dijeliti podatke i samim time neprestano stvarati nove čvorove dok ne postigne savršenstvo, tj. dok u svakom čvoru list ne bude 1 podatak. Taj problem naziva se pretreniranost, te ukratko rečeno javlja se kada je model koji generiramo pre prilagođen testnim podacima s kojima isti testiramo.

„Koncept pretreniranosti je vrlo važan u rudarenju podataka. Odnosi se na situaciju gdje indukcijski algoritam generira klasifikator koji savršeno pripada trening podacima, ali je izgubio kapacitet generaliziranja instanci koje nisu bile u trening podacima. Drugim riječima, umjesto da model uči, on samo pamti instance treninga.“ (Rokach i Maimon, 2015, str. 57)

Oslanjajući se na definiciju od Rokacha i Maimona možemo zaključiti kako će model koji generiramo raditi samo nad podacima koje stablo još nije vidjelo, dok će točnost stabla znatno padati nad podacima koji su stablu novi.



Slika 4. Točnost stabla u odnosu na njegovu veličinu (Mitchell, 1997, str. 67)

Ranije navedeno možemo vidjeti na slici 4. gdje točnost stabla raste nad trening podacima dok ista ta točnost prvo raste, ali se s vremenom smanjuje nad podacima za testiranje s obzirom na veličinu stabla, tj. broj čvorova.

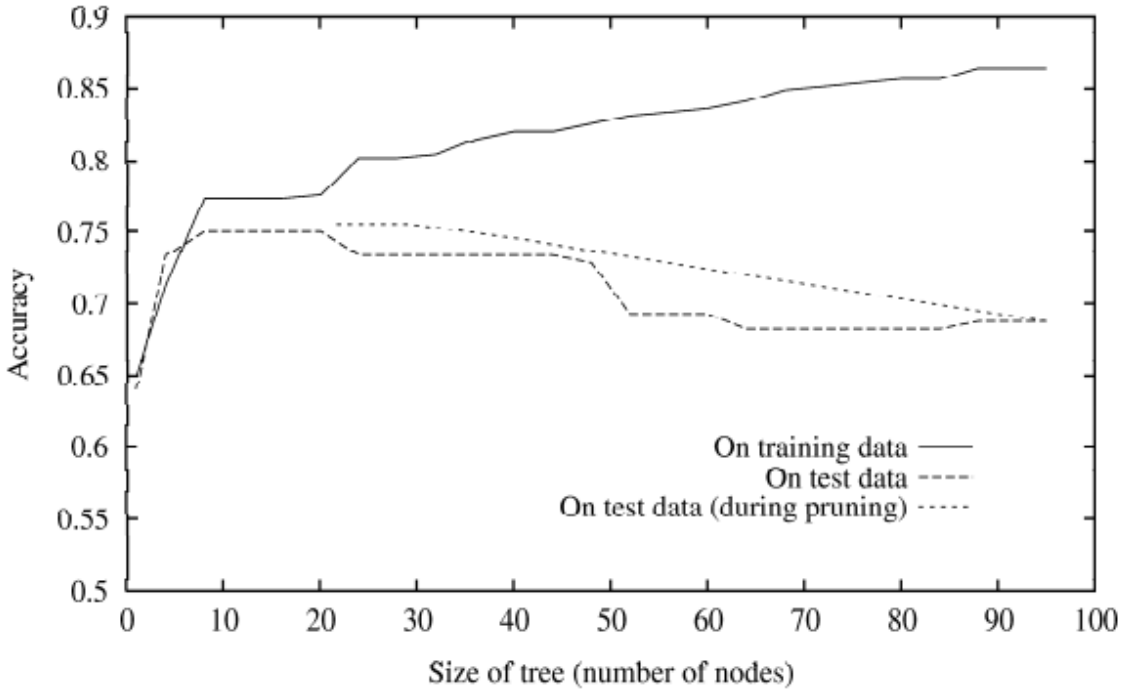
Nakon što se shvati što je to pretreniranost stabla, javlja se pitanje kako to riješiti? Odgovor leži u korištenju tkz. pruning-a što bi značilo obrezivanje stabla. Postoje dvije vrste obrezivanja, a to su pred obrezivanje i post obrezivanje.

Pred obrezivanje funkcionira tako da tijekom izrade, tj. rasta stabla odlučivanja, provjeravamo jeli stablo pretrenirano. Cilj pred obrezivanja je zaustaviti stablo prije nego što postigne veliku količinu listova s malim brojem instanci. Problem koji se javlja kod upotrebe pred obrezivanja je taj što pohlepni algoritmi koji se najčešće upotrebljavaju kod izrade stabla odlučivanja ne znaju kad točno zaustaviti izradu novih čvorova. (Hoare, bez dat.) piše kako tijekom svake podjele stabla provjerava se krosvalidacijska greška i ako se ona značajno ne smanji, stablo zaustavlja svoj rast. Isto tako, navodi kako ova metoda može i stati prerano što uzrokuje greške u točnosti modela.

Post obrezivanje se odvija kada u potpunosti razvijemo stablo te na njemu vršimo obrezivanje tako da mičemo grane koje ne rade dobro s testnim podacima kako bismo postigli što kvalitetniji model i smanjili pretreniranost. Postoje različite procedure post obrezivanja, a jedna od njih je obrezivanje sa smanjenjem greške (engl. Reduced error pruning). Ta procedura razvijena je od strane Quinlan-a te ju koristi C4.5 algoritam.

„Kako prolazi kroz čvorove od dna prema vrhu stabla, REP procedura provjerava svaki čvor te ga zamjenjuje s najznačajnijom klasom koja ne smanjuje točnost stabla. U tom slučaju čvor je obrezan. Procedura se nastavlja sve dok bilo kakvo daljnje obrezivanje više ne bi imalo smisla.“ (Patel i Aluvalu, 2014, str. 10)

(Patel i Aluvalu, 2014) govore o tome kako se obrezivanje vrši sve dok je to potrebno, a da obrezano stablo ima smisla. REP isto tako nasumično dijeli podatke u dva skupa. Jedan od tih skupa je set trening podataka, a drugi skup je validacijski skup, nad kojim se i odvija obrezivanje. Nakon što se napravi obrezivanje, novokreirano, tj. obrezano stablo provjerava se s prvim skupom koji tvore trening podaci.



Slika 5. Točnost stabla u odnosu na njegovu veličinu nakon obrezivanja (Mitchell, 1997, str. 70)

Na slici 5. možemo vidjeti koliko se točnost stabla povećala nad testnim podacima korištenjem obrezivanja naspram slike 4. gdje se obrezivanje stabla nije koristilo.

Naposljetku, možemo navesti pravila koja se koriste tijekom post obrezivanja, a koja je naveo (Mitchell, 1997, str. 71) te su u nekom obliku korištena u C4.5 algoritmu. Pravila su sljedeća:

1. Spoznati kada trening podaci daju najbolji omjer točnosti, a da se ne dozvoli pretreniranost.
2. Konvertiraj naučeno stablo u set ekvivalentnih pravila tako da kreiraš jedno pravilo za svaku putanju od korijena pa sve do čvora list.
3. Generaliziraj svako pravilo, tako da makneš sve preduvjete koji rezultiraju poboljšanjem procijenjene točnosti.
4. Sortiraj obrezana pravila po njihovoj procijenjenoj točnosti te ih uzmi u obzir kod klasificiranja sljedećih instanci.

3. Prednosti i nedostaci

Polazeći od pretpostavke kako su stabla odlučivanja upravo to, virtualna stabla koja nama ljudima pomažu kod donošenja poslovnih odluka, onda možemo zaključiti kako, kao i sve ostale prediktivne metode, imaju svoje prednosti i svoje nedostatke.

Do sada smo obradili što su to stabla odlučivanja i kako ona funkcioniraju te pomoću kojih algoritama dolaze do konačnih prediktivnih modela, te kako ti isti algoritmi koriste različite kriterije podjele i kriterije zaustavljanja, naposljetku smo se dotakli pretreniranosti i kako je riješiti obrezivanjem stabla.

Iz svega ranije navedenog, možemo zaključiti kako stabla odlučivanja kao svoju glavnu prednost ističu vrijeme koje je potrebno za pripremu podataka, prije nego što ti isti podaci krenu u obradu. To vrijeme u odnosu na ostale algoritme je značajno kraće.

Stabla odlučivanja ne zahtijevaju normalizaciju podataka, što znači da se mogu pojavljivati duplikati, a da stablo i dalje funkcionira relativno dobro i da krajnji rezultat, tj. prediktivni model bude prihvatljiv.

„Stabla odlučivanja su samo objašnjiva i kada se prezentiraju vrlo ih je lako shvatiti. Uzimajući to u obzir, ako stabla odlučivanja imaju određen broj listova mogu ih shvatiti čak i neprofesionalni korisnici. Nadalje, pošto se stabla odlučivanja mogu pretvoriti u set pravila, takva reprezentacija smatra se razumljivom. „ (Rokach i Maimon, 2015, str. 81)

Rokach i Maimon žele reći kako, iako stabla odlučivanja jesu jedan moćan alat u rukama korisnika i mogu se koristiti za ozbiljne analize, vrlo su lako shvatljivi te je vrlo jednostavno uroniti u svijet rudarenja podataka korištenjem stabla odlučivanja.

Stabla odlučivanja su vizualno vrlo prihvatljiva i vrlo ih je lagano primijeniti u svakodnevnoj upotrebi. Jedna od bitnih značajki je ta da mogu raditi i s kontinuiranim i s kategorijskim varijablama što je vrlo velika i snažna prednost nad ostalim algoritmima.

„Vjeruje se kako su stabla odlučivanja jedan od najbržih, ako ne i najbrži algoritam za identifikaciju *najznačajnijih varijabli, isto tako i za pronalaženje veze između dvije ili više varijabli. Stabla odlučivanja pomažu korisnicima u kreiranju novih varijabli kao i značajki. Te nove značajke će imati bolje predispozicije za predviđanje ciljne varijable. Isto tako stabla su korisna u istraživanju podataka.“ (Sullivan, 2017., str. 29).*

Sullivan naglašava brzinu stabla odlučivanja i daje još primjer gdje u slučaju dvije varijable stablo odlučivanja će vrlo lako i brzo identificirati koja od njih je najznačajnija.

Kao što navodi (Woodruff, 2019), stablo odlučivanja je jako svestran algoritam koji je korišten u sve većem broju kod poslovnih problema gdje se taj problem analizira i rješava korištenjem stabala. Također, spominje ih i u smislu gdje se stablo odlučivanja koristi kao tehnika zdravog razuma za pronalaženje najboljeg rješenja za problem nesigurnosti i kao primjer daje: trebam li uzeti kišobran danas na posao?

Kada govorimo o nedostacima s kojima se susrećemo kod korištenja stabla odlučivanja kao prediktivne metode, kao jednu od ključnih mana spominju se matematički izračuni koji u velikoj većini slučajeva iziskuju ogromnu količinu memorije kako bi se izračunali.

Ista tako, kao što treba mnogo memorije za izračun, potrebno je i dosta vremena kako bi se oni izveli, a stablo odlučivanja je također vrlo osjetljivo na male promjene u podacima što u konačnici može uzrokovati skroz drugačije rezultate nego što je to možda bilo prije promjene koja je nastala.

Ranije navedeni problem spominju i Rokach i Maimon (2015, str. 82) te još navode nedostatak gdje algoritmi mogu, uzimajući u obzir malu varijaciju u setu trening podataka, odabrati atribut koji u konačnici i nije najbolji koji je mogao biti izabran od ponuđenih u setu trening podataka.

„Iako je sposobnost rada bez nedostajućih podataka smatrana kao prednost, veliki napor koji je potrebno uložiti smatra se kao nedostatak. Prava grana koja će se uzeti smatra se nepoznatom, ako značajka koja se testira nedostaje, a algoritam mora upotrijebiti poseban mehanizam kako bi radio s podacima koji nedostaju. U svrhu smanjenja testova nad podacima koji nedostaju, C4.5 penalizira informacijsku dobit proporcijski prema nepoznatim instancama i zatim dijeli te instance u pod-stabla, dok CART koristi kompleksniju shemu poput surogat obilježja.“ (Rokach i Maimon, 2015, str. 83)

Prethodno navođenje također potvrđuje činjenicu kako su stabla odlučivanja skup algoritam u smislu korištenja memorije i brzine koja je potrebna kako bi se dobilo kvalitetno konačno rješenje u obliku dobrog modela.

Za kraj se još možemo dotaknuti problema korištenja stabla odlučivanja kod kompleksnijih problema gdje je jedno stablo vrlo nestabilno i neprilagodljivo zahtjevnijim izračunima koji se mogu pojaviti. Doduše, tom problemu se pristupa upotrebom slučajne šume (eng. Random forest) koja je u principu skup više stabala. Slučajna šuma je kompleksan algoritam i nije tema ovog rada, ali ju je vrlo bitno spomenuti jer rješava dosta problema koje jedno stablo odlučivanja ima.

4. Ranija istraživanja

Prije praktičnog rada obradit ćemo ranija istraživanja čija je tema bila obrazovanje i analiza podataka korištenjem stabla odlučivanja gdje ćemo naglasiti bitne odrednice tih radova i koji su bili njihovi zaključci kako bismo nakon praktičnog rada imali s čime usporediti koliko su ishodi slični, a koliko različiti.

U radu „CRISP-DM Process model in educational setting“, autori Oreški, Pihir i Konecki (2017) obrađuju temu akademske uspješnosti kroz CRISP-DM metodologiju. Bitne odrednice kojih se dotiču i koje su im zanimljive za što bolju prediktivnu analizu su sljedeće:

1. Socio-demografske značajke
2. Prijašnje obrazovanje
3. Motivacija i stil učenja

Analiza je provedena nad studentima Fakulteta organizacije i informatike u Varaždinu s ciljem pronalaska korelacije između studentskih karakteristika i akademske uspješnosti. Anketa se sastojala od 36 pitanja, a na kraju je odabrano 15 najvažnijih varijabli koje su se koristile u daljnjoj obradi.

U fazi modeliranja, primjenom stabla odlučivanja spoznaju kako studenti s najmanjom prosječnom ocjenom su oni studenti koji su u srednjoj školi imali prosječnu ocjenu manju od 4.0, koji nisu bili u top 30% na prijemnom ispitu i koji su teže savladavali zahtjevnije gradivo. Studenti s najboljom prosječnom ocjenom su oni studenti s odličnom ocjenom u srednjoj školi i oni koji su dobro napisali prijemni ispit. Isto tako, Oreški, Pihir i Konecki (2017) dolaze do zaključka kako varijable koje su najviše utjecale na taj ishod su: rezultat prijemnog ispita, prosječna ocjena u srednjoj školi i prisustvo na nastavi.

Drugo istraživanje, „Performance Prediction of Engineering Students using Decision Trees“ provedeno od strane Kabra i Bichkar (2011) stavlja fokus na uspješnost studenata inženjerstva koristeći stabla odlučivanja.

Podaci nad kojima su vršili analizu prikupljeni su iz S. G. R. Education Foundation’s College of Engineering and Management u Indiji. Anketa je provedena nad 346 studenata i bitno je za naglasiti kako su svi ispitanici bili s prve godine studija.

Dobiveni model dao je točnost od 60.46% što bi značilo da od 346 instanci, 209 ih je bilo točno. Analizirajući model utvrdili su kako najznačajnija varijabla je ocjena dobivena prilikom prijemnog testa, a varijable poput zanimanja roditelja, mjesto stanovanja, spol i drugo

ne pojavljuju se kao bitne odrednice modela. Kabra i Bichkar zaključuju kako od 205 studenata 186 će ih pasti te da vjerojatnost pada iznosi 0.907.

Treći rad koji ćemo spomenuti je „Educational Data Driven Decision Making: Early Identification of Students at Risk by Means of Machine Learning“ i napisan je od strane Kovač i Oreški (2018). Rad je baziran na rudarenju podataka u obrazovanju, a algoritmi učenja koji su korišteni su neuronske mreže, stablo odlučivanja, potporni vektori i logistička regresija.

Podaci su bazirani na studentima Fakulteta organizacije i informatike u Varaždinu, a prikupljeni su u sustavu za upravljanje učenjem, Moodle. Model je baziran na 235 prikupljenih zapisa. U početku skup se sastojao od 13 varijabli, a 11 ih je u konačnici korišteno u analizi podataka.

Primjenom ranije spomenutih algoritama, Kovač i Oreški (2018) su dobili sljedeće pouzdanosti za određeni model:

1. Logistička krivulja: 92.85%
2. Neuronske mreže: 94.28%
3. Stablo odlučivanja: 97.14%
4. Potporni vektori: 95.7%

Vidimo kako je primjenom stabla odlučivanja dobivena najveća točnost modela te Kovač i Oreški (2018) zaključuju kako je to zbog sposobnosti stabla odlučivanja da uzima u obzir samo one atribute koji su bitni za klasifikaciju.

„Decision trees for predicting the academic success of students“ je rad napisan od strane Mesarić i Šebalj (2016), a govori o uspjehu studenata nakon završene prve godine studija gdje ih se svrstava u dvije skupine ovisno o njihovoj uspješnosti.

Podaci se temelje na studentima Fakulteta Ekonomije u Osijeku, a prikupljeni su tijekom tri generacije (2016/16, 2014/15 i 2013/14) te se temelje na ocjenama iz ISVU sustava i ocjenama mature. U obzir su uzeti samo studenti koji su uspješno upisali drugu godinu studija. Završni skup sastojao se od 665 studenata i korišteno je 8 varijabli za kreiranje modela stabla odlučivanja. Autori su koristili različite algoritme stabla odlučivanja, poput J48, random forest, random tree i REPTTree, a najbolji rezultat dao je REPTTree algoritam te je pouzdanost stabla iznosila 79.35%.

Mesarić i Šebalj (2016) zaključuju kako atributi koji su najznačajnije utjecali na model su: ukupna ocjena, ocjena iz srednje škole, ocjena iz hrvatskog jezika na maturi i status, ali kako model, iako dobro predviđa ishod studenata čiji je prosjek manji od 3.5, daje loše rezultate za studente prosjeka većeg od 3.5.

Nadalje, spominju manjkavost rada jer studenti mogu upisati drugu godinu bez da su položili sve predmete s prve što mijenja ponderiranu aritmetičku sredinu ocjene, ali naglašavaju da to ne bi trebala biti velika promjena.

Zadnje istraživanje koje će biti spomenuto „Role of Personal Factors in Academic Success and Dropout of IT Students: Evidence From Students and Alumni“ autora Oreški, Hajdin i Kliček (2016). Autori su rad podijelili u 3 glavne odrednice:

1. Identificirati faktore koji vode do uspješnosti studenata.
2. Istražiti razlike između percepcije alumni studenata i trenutnih studenata.
3. Istražiti razlike između spola.

Podaci koji su korišteni u analizi su prikupljeni u akademskoj godini 2014/2015 putem ankete na jednom od fakulteta koji je sastavnica Sveučilišta u Zagrebu. U konačnici autori su prikupili 516 instanci, od kojih su 354 trenutni studenti, a 162 alumni. Također, od 516 instanci, 41.67% su žene, a 58.33 su muškarci. Isto tako 37.38% ispitanika su studenti informatike, 23.36% ekonomije, a 20.79% studenata s diplomom, dok je prosječna starost iznosila 25 godina.

U svojom radu Oreški, Hajdin i Kliček (2016) koriste neuronske mreže i t-test, a kao odgovore na ranije postavljena pitanja zaključuju kako faktori koji vode do uspješnosti studenata su studijski program, godina, status i rang studenta prilikom upisa programa. Kao glavnu razliku kod alumni studenata i trenutnih studenata ističu akademsku neuspješnost te napominju kako taj dio nije dovoljno istražen, a i nedostaju podaci za usporedbu. Kod razlike između spola, dolaze do zaključka kako muški pristupnici brinu više o upravljanju vremenom i o motivaciji, dok ženske pristupnice više pažnje pridaju zdravlju.

Autor/i	Godina	Uzorak	Metodologija	Broj ulaznih varijabli	Rezultat	Najznačajnija varijabla
Oreški, Pihir, Konecki	2017	Nepoznat	Stablo odlučivanja	15	Stablo odlučivanja: 52%	Rezultat prijemnog ispita, prosječna ocjena u srednjoj školi, prisustvo na nastavi
Kabra, Bichkar	2011	346	Stablo odlučivanja	17	Stablo odlučivanja: 60.46%	Ocjena na prijemnom ispitu
Kovač, Oreški	2018	235	Logistička krivulja, neuronske mreže, stablo odlučivanja. Potporni vektori	11	Logistička krivulja: 92.85%, neuronske mreže: 94.28%, stablo odlučivanja:97.14%, potporni vektori: 95.7%	Prisustvo na seminarima, prisustvo na predavanjima
Mesarić, Šebalj	2016	665	Stablo odlučivanja	8	Stablo odlučivanja:79.35%	Ukupna ocjena, ocjena iz srednje škole, ocjena iz hrv. Na maturi i status
Oreški, Hajdin, Kliček	2016	516	Neuronske mreže, t-test	36	-	Studijski program, godina, status i rang

Tablica 1: Pregled prijašnjih istraživanja (autorski rad)

5. Primjena na podacima iz obrazovanja

Nakon obrađene teorije koja stoji iza stabla odlučivanja i pregleda istraživanja mnogobrojnih autora na temu obrazovanja, naš fokus će sada biti primjena metode stabla odlučivanja. Cilj primjene na podacima iz obrazovanja je donijeti zaključak koje to varijable najviše utječu na pripadnost određenoj klasi ocjene za pojedinog učenika. Nakon izrade i interpretiranja modela napravit će se usporedba s ranije navedenim istraživanja i na kraju će se izvesti zaključak.

5.1. Opis podataka

Skup podataka koji će biti korišteni u ovom radu prikupljeni su na Sveučilištu Jordan u Ammanu od strane Amrieh, Hamtini i Aljarah (2016). Slično kao što su Kovač i Oreški (2018) svoje podatke za analizu prikupili sa sustava za upravljanje učenjem Moodle, Amrieh i Hamtini, Aljarah su ih prikupili s Kalboard-a 360, također sustava za upravljanje učenjem.

Skup podataka se sastoji od 480 zapisa aktivnosti učenika. Od toga 305 je muškaraca, a 175 žena. Skup se također sastoji od više različitih nacionalnosti od kojih su sljedeće:

- Kuvajt – 179
- Jordan – 172
- Palestina – 28
- Irak – 22
- Libanon – 17
- Tunis – 12
- Saudijska Arabija – 11
- Egipat – 9
- Sirija – 7
- Sjedinjene Američke Države, Iran, Libija – 6
- Maroko – 4
- Venezuela - 1

Podaci su prikupljeni tijekom jedne godine, 245 zapisa tijekom prvog semestra, a 235 tijekom drugog. Skup se sastoji od 17 atributa, od kojih su 13 kategorički, a 4 numerička.

Atributi su podijeljeni u 3 velike kategorije. Demografska obilježja, poput spola i nacionalnosti, akademske pozadine, kao što je ocjena i trenutno stanje studija, i atribute ponašanja u koje spada dizanje ruke na nastavi, otvaranje resursa za učenje itd.

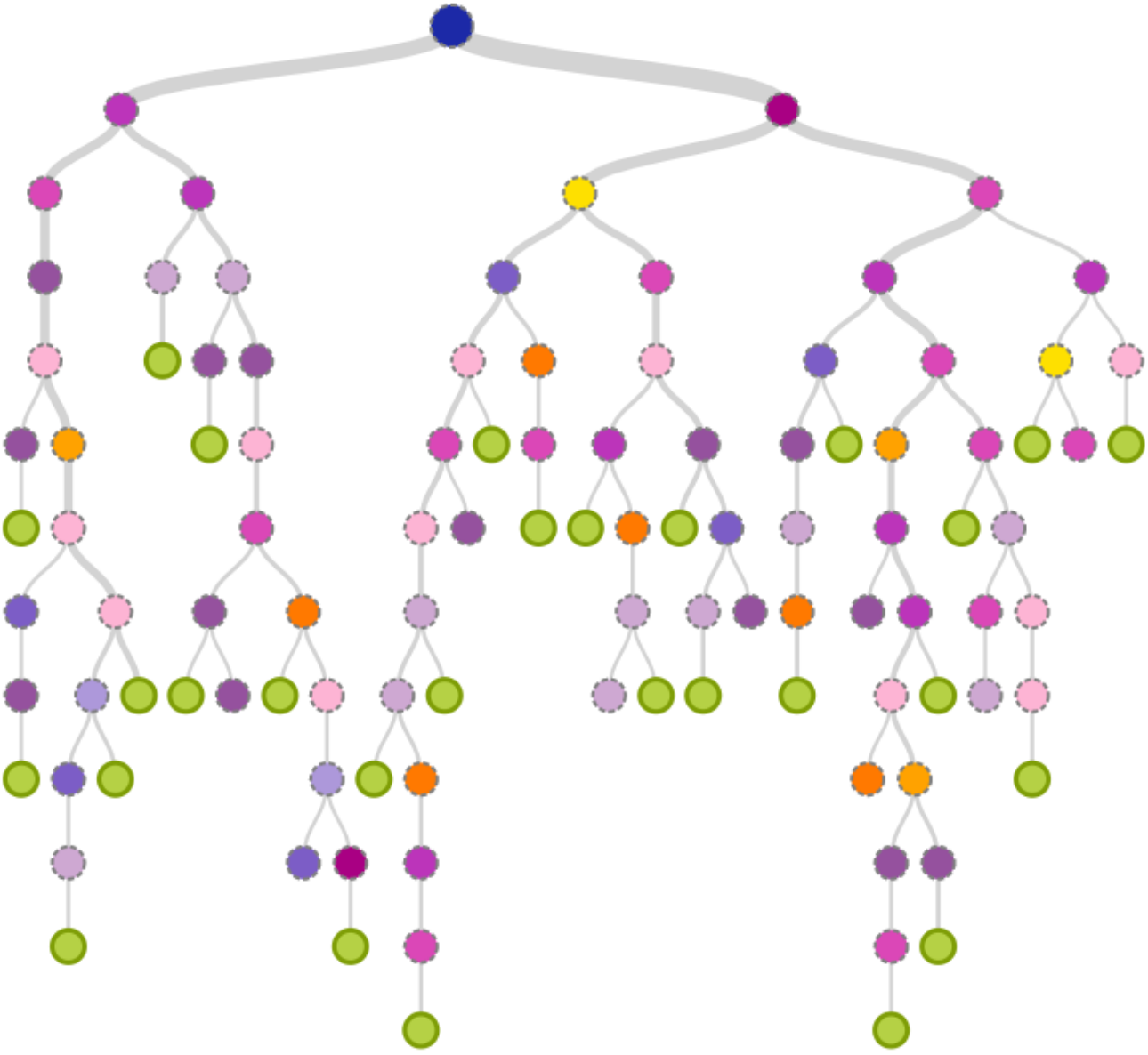
Atribut	Opis	Vrsta
Spol	Spol učenika („Muškarac“ ili „Žena“)	Kategorički
Nacionalnost	Nacionalnost učenika („Kuvajt“, „Jordan“, „Palestina“, „Irak“, „Libanon“, „Tunis“, „Saudijska Arabija“, „Egipat“, „Sirija“, „SAD“, „Maroko“, „Venezuela“)	Kategorički
Mjesto rođenja	Mjesto rođenja učenika („Kuvajt“, „Jordan“, „Palestina“, „Irak“, „Libanon“, „Tunis“, „Saudijska Arabija“, „Egipat“, „Sirija“, „SAD“, „Maroko“, „Venezuela“)	Kategorički
Stupanj edukacije	Stupanj edukacije kojem učenik pripada („Predškola“, „Osnovna škola“, „Srednja škola“)	Kategorički
Razred	Razred kojem učenik pripada („G-01“, „G-02“, „G-03“, „G-04“, „G-05“, „G-06“, „G-07“, „G-08“, „G-09“, „G-10“, „G-11“, „G-12“)	Kategorički
ID sekcije	Učionica kojoj učenik pripada („A“, „B“, „C“)	Kategorički

Predmet	Predmet koji se predaje („Engleski“ „Španjolski“, „Francuski“, „Arapski“, „IT“, „Matematika“, „Kemija“, „Biologija“, „Znanost“, „Povijest“, „Kuran“, „Geologija“	kategorički
Semestar	Semestar u godini („Prvi“, „Drugi“)	Kategorički
Roditelj	Roditelj odgovaran za učenika („Majka“, „Otac“)	Kategorički
Dizanje ruke	Broj dizanja ruke na nastavi (0-100)	Numerički
Otvaranje resursa	Broj posjeta resursu s nastave (0-100)	Numerički
Pregled objava	Broj pregleda novih objava (0-100)	Numerički
Diskusijske grupe	Broj pristupanja diskusijama (0-100)	Numerički
Zadovoljstvo roditelja	Zadovoljstvo roditelja školom („Da“, „Ne“)	Kategorički
Odsustvo s nastave	Broj izostanaka s nastave („Više od 7“, „Manje od 7“)	Kategorički
Roditelj ispunio anketu	Je li roditelj ispunio anketu („Da“, „Ne“)	Kategorički
Klasa ocjene	Klasa ocjene kojoj učenik pripada s obzirom na završnu ocjenu (L → 0-69, M → 70-89, H → 90-100)	Kategorički

Tablica 2: Pregled atributa (autorski rad)

5.2. Izrada i analiza modela

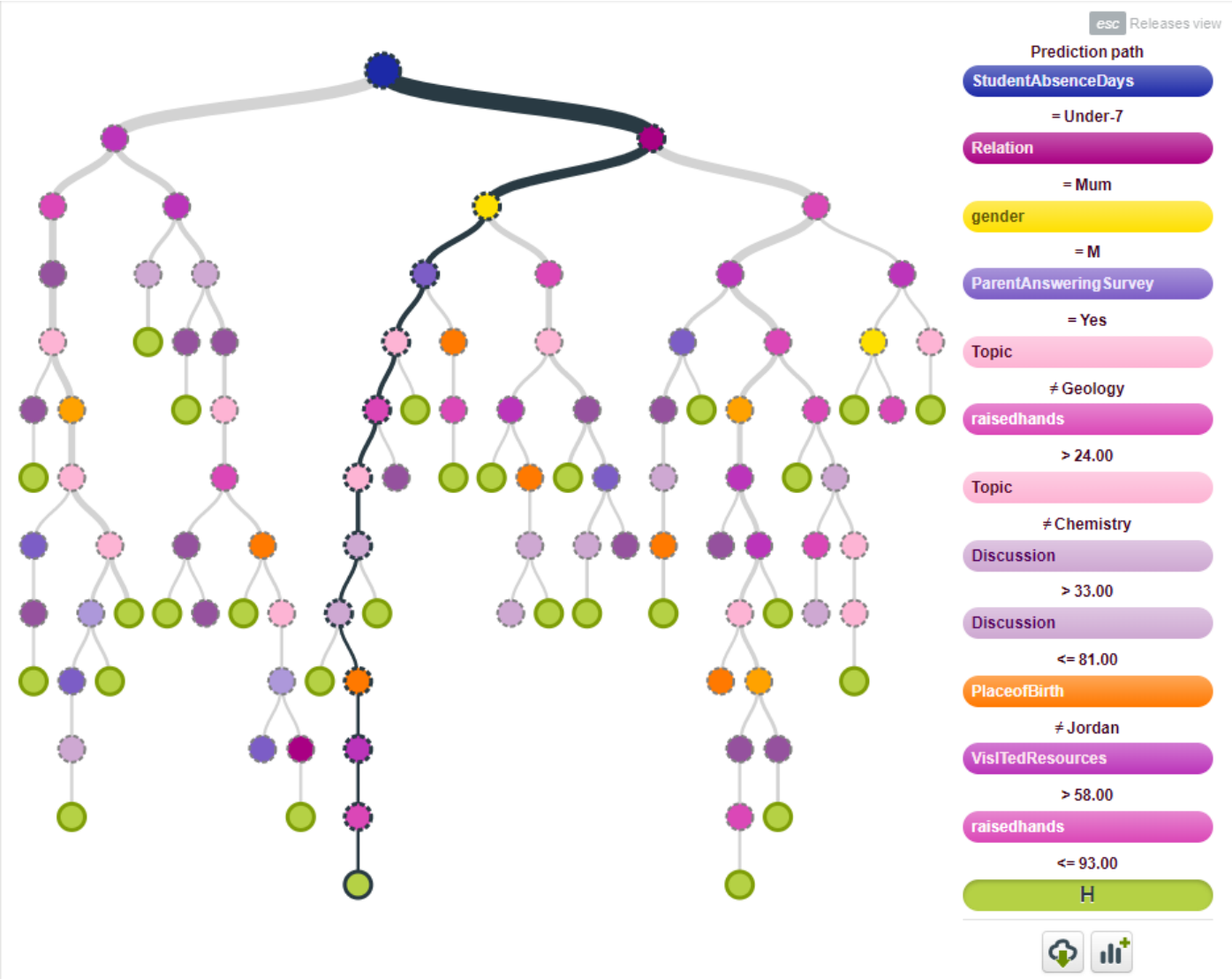
Za izradu modela stabla odlučivanja koristit ćemo online alat BigML koji će uz pomoć algoritma modelirati stablo s ciljem predviđanja kojoj skupini ocjene pripada učenik i koji to faktori, tj. atributi najviše utječu na to. Za izradu modela koristit ćemo sve ranije navedene atribute. U postavkama ćemo odabrati mogućnost obrezivanja i granicu ćemo postaviti na 240 što je polovica svih instanci. Nakon što smo u aplikaciju ubacili naš skup podataka i odabrali da želimo koristiti metodu stabla odlučivanja, aplikacija je kreirala stablo prikazano na slici 6.



Slika 6: Model stabla odlučivanja (autorski rad)

Možemo još spomenuti kako BigML za osnovu svog algoritma koristi modificirani C4.5 algoritam, dok kao kriterij podjele služi se informacijskom dobiti, a za kriterij zaustavljanja primjenjuje ranije definirani broj mogućih čvorova ili kada stablo više nema smisla dijeliti. Obrezivanje je također bazirano na način kao što to radi C4.5 (BigML, bez dat.).

Generirani model koji se nalazi na slici 6 daje pouzdanost od 39.58%, a kao što ćemo kasnije vidjeti, prva varijabla, tj. atribut koji je odabran za dijeljenje je broj izostanaka učenika s nastave.



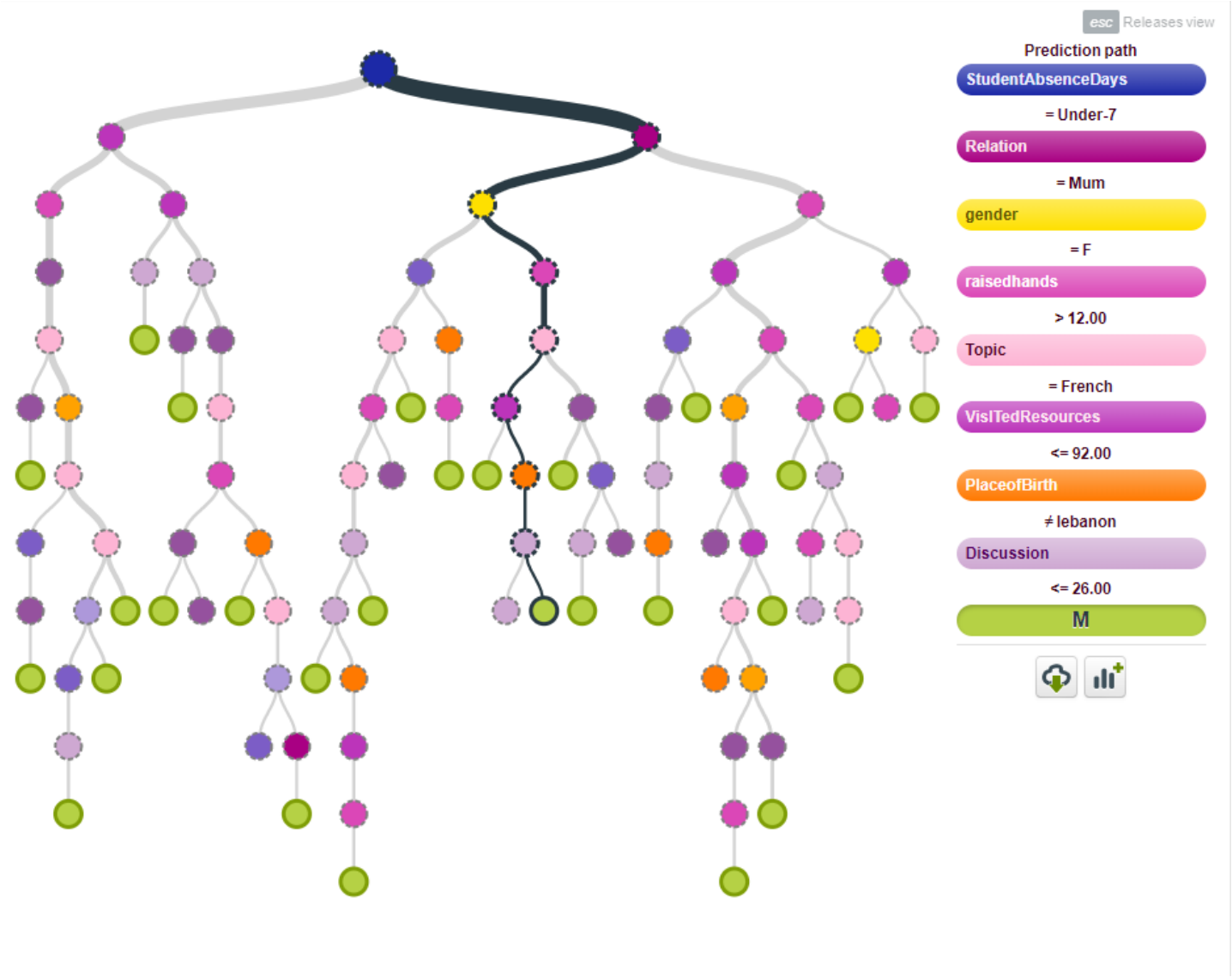
Slika 7: Jedna od mogućih putanja na modelu (autorski rad)

Slika 7. prikazuje jednu od mogućih putanja koje vode do terminalnog čvora, tj. do konačnog odgovora koji glasi da, ako su sva pitanja na ovom putu zadovoljena, učenik će pripasti klasi ocjene H.

Atribut	Vrijednost
Odsustvo s nastave	Manje od 7
Odgovorna osoba	Majka
Spol	Muškarac
Roditelj ispunio anketu	Da
Predmet	Različito od geologije
Dizanje ruke	Više od 24 puta
Predmet	Različito od kemije
Diskusijske grupe	Više od 33
Diskusijske grupe	Manje ili jednako od 81
Mjesto rođenja	Različito od Jordana
Otvaranje resursa	Više od 58
Dizanje ruke	Manje ili jednako od 93

Tablica 3: Prikaz pravila sa slike 7. (autorski rad)

Tablica 3. prikazuje pravila koja vode do terminalnog čvora koji sadrži 9 instanci koje pripadaju klasi ocjene H što je 1.88% od ukupnog broja učenika i visokom pouzdanošću od 70.08%.



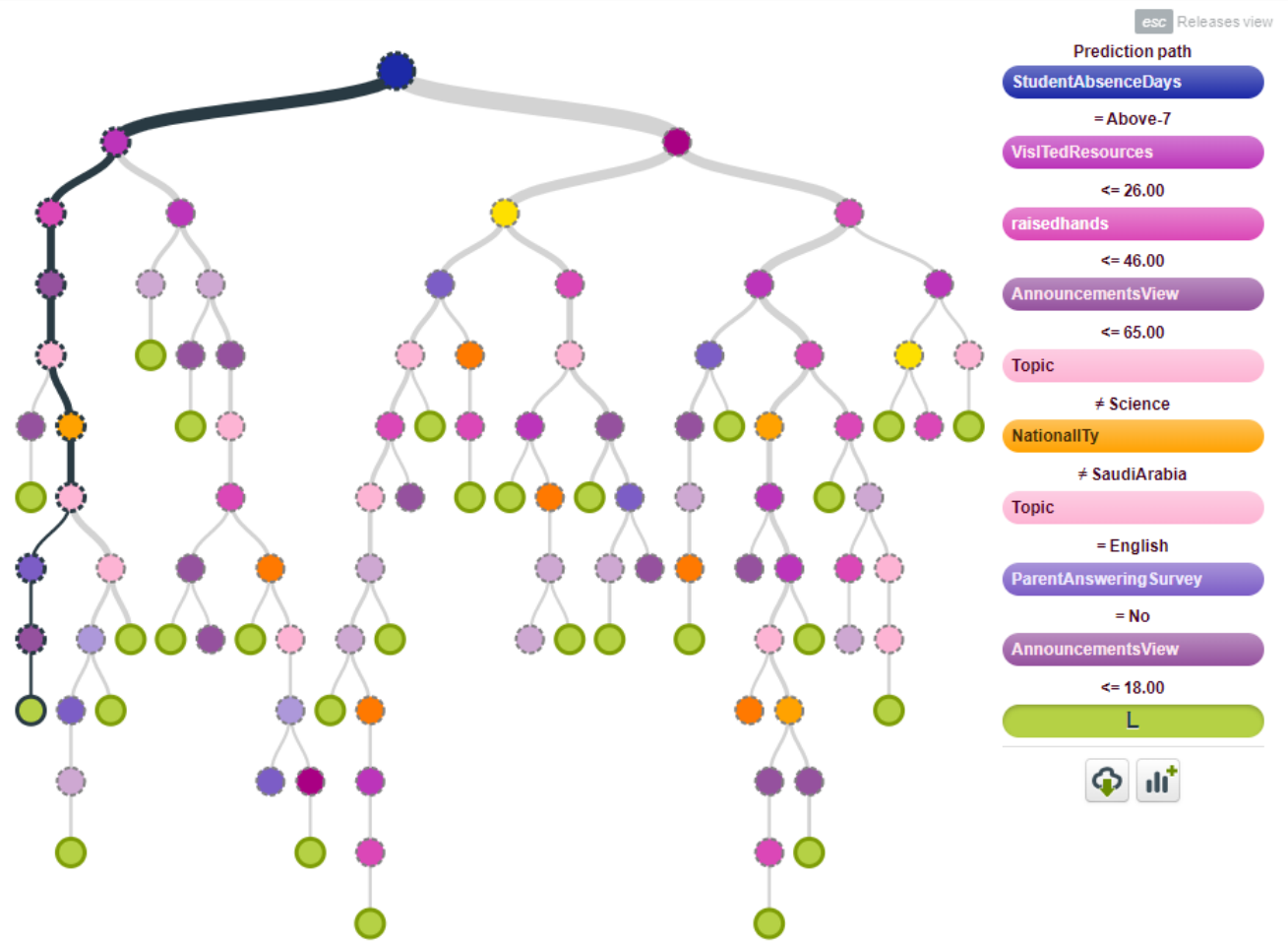
Slika 8: Jedna od mogućih putanja na modelu (autorski rad)

Atribut	Vrijednost
Odsustvo s nastave	Manje od 7
Odgovorna osoba	Majka
Spol	Žena
Dizanje ruke	Više od 12

Predmet	Francuski
Otvaranje resursa	Manje ili jednako od 92
Mjesto rođenja	Različito od Libanona
Diskusijske grupe	Manje ili jednako od 26

Tablica 4: Prikaz pravila sa slike 8. (autorski rad)

Tablica 4. prikazuje pravila s putanje na slici 8 koje vode do terminalnog čvora u kojem se nalazi 5 instanci koje pripadaju klasi ocjene M, što je 1.04% od ukupnog broja učenika te pouzdanost iznosi 56.55%.



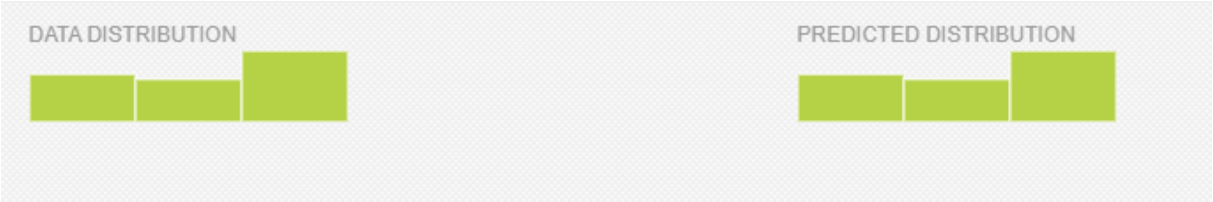
Slika 9: Jedna od mogućih putanja na modelu (autorski rad)

Atribut	Vrijednost
Odsustvo s nastave	Više od 7
Otvaranje resursa	Manje ili jednako od 26
Dizanje ruke	Manje ili jednako od 46
Pregled objava	Manje ili jednako od 65
Predmet	Različito od znanosti
Nacionalnost	Različito od Saudijska Arabija
Predmet	Engleski
Roditelj ispunio anketu	Ne
Otvaranje resursa	Manje ili jednako 18

Tablica 5: Prikaz pravila sa slike 9. (autorski rad)

Tablica 5. prikazuje pravila s putanje na slici 9 koje vode do terminalnog čvora u kojem se nalazi 6 instanci koje pripadaju klasi ocjene L, što je 1.25% od ukupnog broja učenika te pouzdanost iznosi 60.97%.

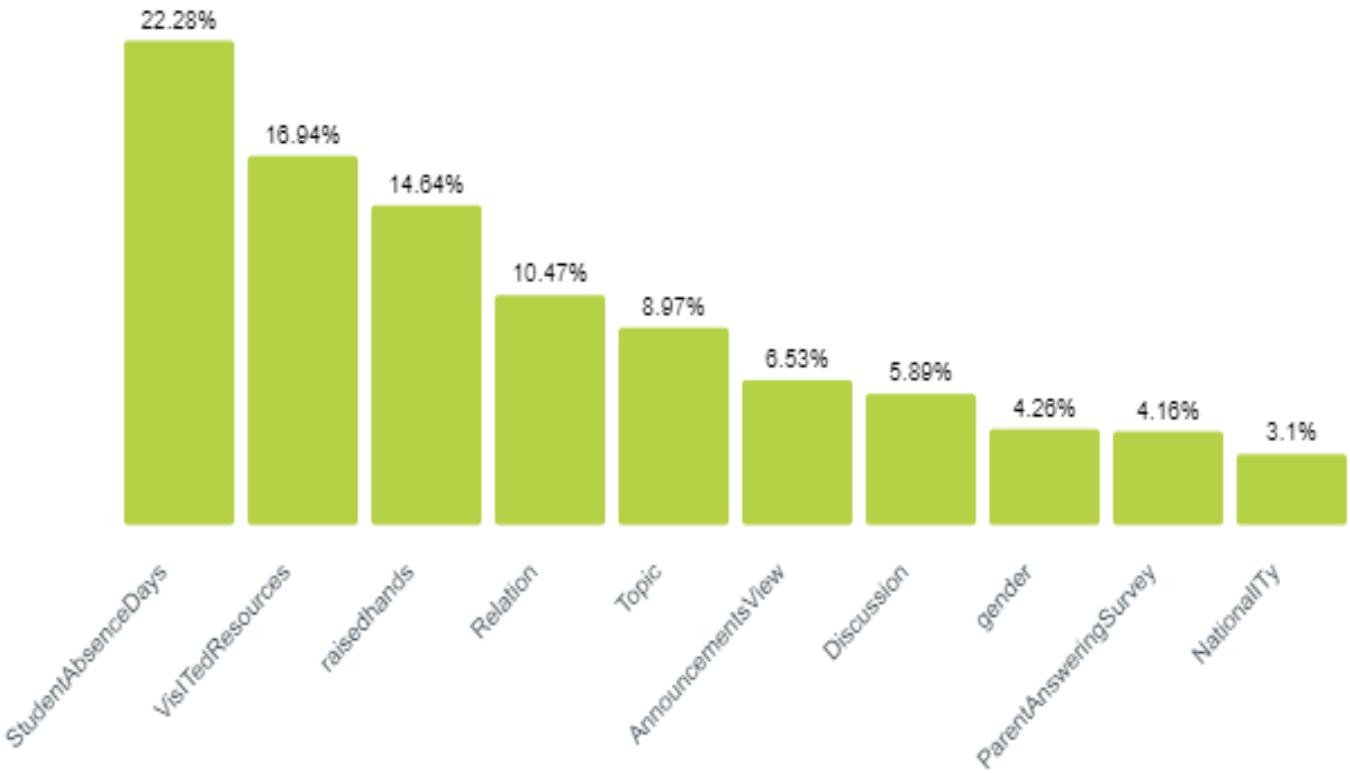
Na prijašnjim putanjama prikazan je jedan od mogućih puteva za svaku od klasa ocjena kojoj određeni učenik može pripadati (Klasa H, M, L) i pravila za svaki od tih puteva koja moraju biti zadovoljena kako bi se došlo do terminalnog čvora i zadovoljio izrađeni model.



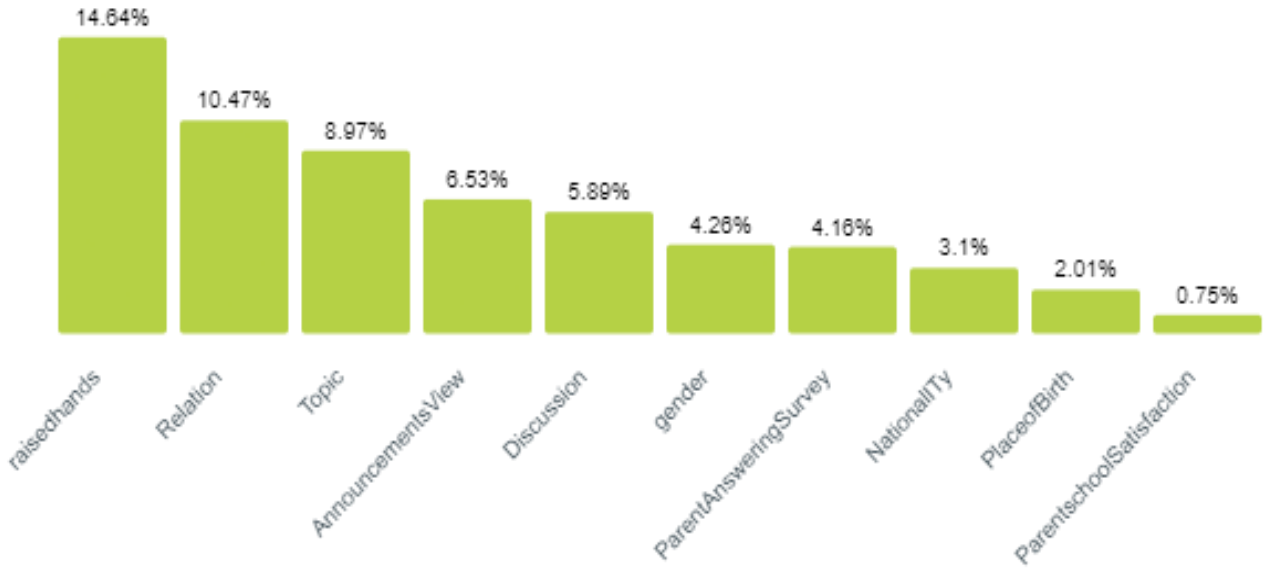
Slika 10: Raspodjela podataka (autorski rad)

Na slici 10 možemo vidjeti raspodjelu podataka prije izrađenog modela i onoga što je izrađeni model predvidio. Obje distribucije daju jednake podjele, a ona je sljedeća:

- H: 29.58% (142 instance)
- L: 26.46% (127 instance)
- M: 43.96% (211 instance)



Slika 11: Značaj atributa (autorski rad)

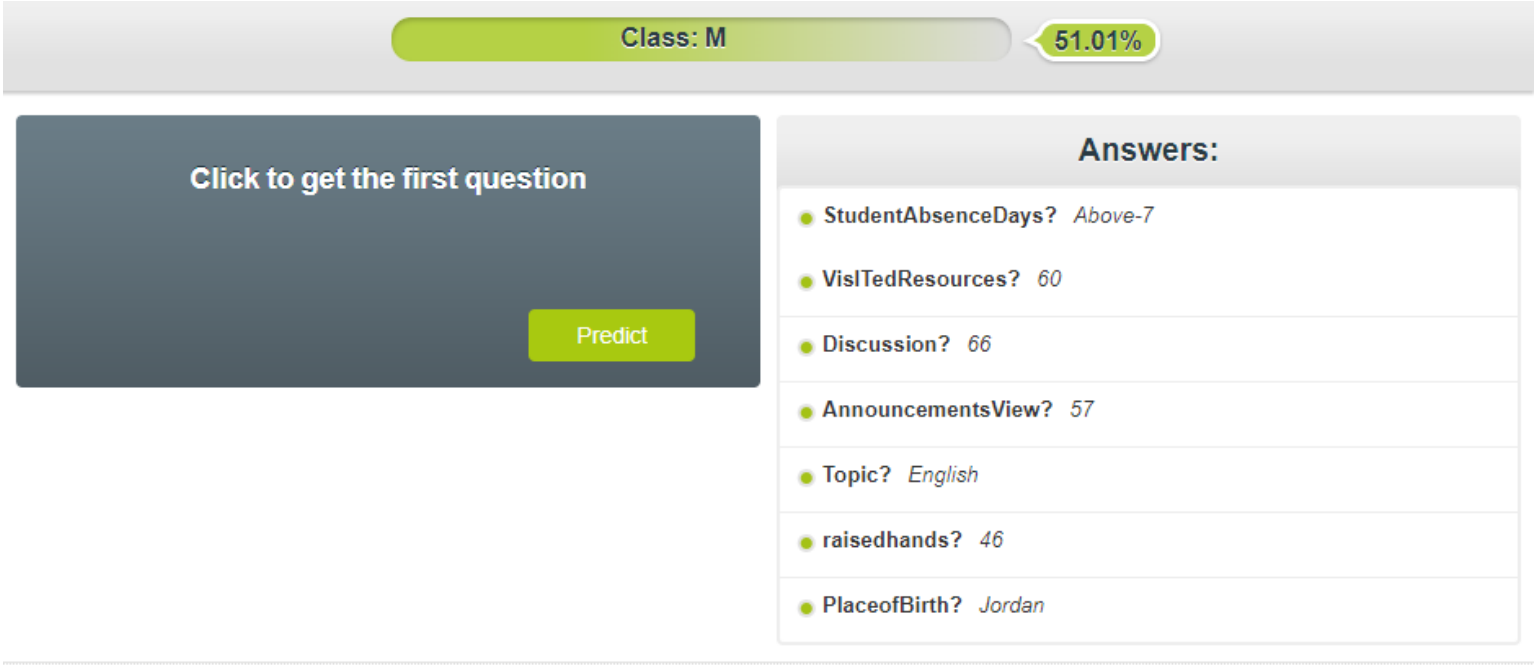


Slika 12: Značaj atributa (autorski rad)

Slika 11 i slika 12 prikazuju koji to atributi utječu na izradu modela i koji je njihov značaj prilikom, tj. udio tijekom podjele na nove čvorove. Poredani po značaju, atributi glase:

1. Odsustvo s nastave: 22.28%
2. Otvaranje resursa: 16.94%
3. Dizanje ruke: 14.64%
4. Odgovorna osoba: 10.47%
5. Predmet: 8.97%
6. Pregled objava: 6.53%
7. Diskusijske grupe: 5.89%
8. Spol: 4.26%
9. Roditelj ispunio anketu: 4.16%
10. Nacionalnost: 3.10%
11. Mjesto rođenja: 2.01%
12. Zadovoljstvo roditelja : 0.75%

Jedna od mogućnosti BigML alata je ta da samostalno možemo unijeti odgovore na pitanja prema kojima se izgrađuje stablo i prema tome donijeti zaključak kojoj klasi ocjene pripada učenik s tim vrijednostima.



Slika 13: Previđanje prema pitanjima (autorski rad)

Tako prema slici 13 možemo zaključiti kako učenik s brojem izostanaka većim od 7, sa 60 posjećenih resursa, 66 pristupa diskusiji, 57 pregleda objava, gdje je predmet engleski, ima 46 dizanja ruku i mjesto rođenja mu je Jordan, da će pripasti skupini ocjene M s 51.01% pouzdanosti.

5.3. Rezultati primjene

Cilj primjene stabla odlučivanja nad podacima iz obrazovanja je bilo vidjeti koji su to najznačajniji faktori koji utječu na pripadnost učenika određenoj klasi ocjene. U skladu s prijašnjim istraživanjima gdje su najvažniji atributi bili prisustvo na nastavi, prosječna ocjena, ocjena na prijemnom ispitu i slično, dolazimo do zaključka kako niti ovdje nije ništa drugačije.

Kao glavni faktor ističe se odsustvo s nastave s čak 22.28% udjela dok je odmah iza njega otvaranje resursa koji se tiču predmeta kojeg učenik pohađa. Slijedi dizanje ruke, odgovorna osoba za učenika, predmet itd.

Samim time možemo vidjeti kako učenikovo prisustvo na nastavi i pripremljenost za predmet uistinu utječu na njegovu ocjenu i vrlo su bitan čimbenik u njegovom daljnjem obrazovanju. Završni model je dao pouzdanost od 39.58% dok je ta ista pouzdanost u završnim čvorovima dostizala i preko 80%.

Prema ranije viđenom, distribucija podataka je bila jednaka u originalnoj podjeli i u predviđenoj distribuciji od stabla odlučivanja. Tako je od 480 instanci, 142 svrstano u skupinu ocjene H, 127 u L, a 211 u M, što i ima smisla jer M predstavlja prosjek te je najveći broj učenika upravo svrstano u tu skupinu.

6. Zaključak

Ovaj rad se bavi stablom odlučivanja i njegovom primjenom nad stvarnim podacima iz obrazovanja. Cilj je bio objasniti teorijsku razinu stabla odlučivanja, a zatim napraviti istraživanje koristeći programski alat BigML i podatke iz obrazovanja. Prikazani su algoritmi stabla odlučivanja poput ID3, C4.5 i CART, kriteriji podjele korišteni od stabla odlučivanja kod stvaranja novih čvorova, ali isto tako i kriteriji zaustavljanja kako bi stablo znalo kada stati s daljnjom podjelom. Na kraju teorijskog dijela bilo je riječi o pretreniranosti stabla i kako se to rješava obrezivanjem, navedene su prednosti i mane, a pod temom ranija istraživanja obradili su se znanstveni radovi koji su vezani uz temu obrazovanja.

Praktični dio bila je primjena stabla odlučivanja nad podacima koji su prikupljeni sa Sveučilišta u Jordanu putem njihovog sustava za upravljanje učenjem. Na temelju 17 atributa koji su korišteni u analizi, konačni model stabla odlučivanja dao je pouzdanost od 39.58%, dok je pouzdanost u nekim završnim čvorovima rasla i preko 80%. Kao najvažniji atribut istaknulo se odsustvo s nastave s visokih 22.28% udjela, prati ga otvaranje resursa s 16.94%, dizanje ruke 14.64% itd. Cilj ove primjene bilo je saznati kojoj skupini ocjene određeni učenik pripada. Od 480 instanci, 142 učenika pripalo je skupini H, što je i najbolja skupina, 211 skupini M, a 127 u L. Kao što je bilo i za očekivati, najveći broj učenika pripao je srednjoj skupini ocjene H. Važna odrednica koju je potrebno spomenuti je i raspodjela podataka prije izrađenog modela i nakon, gdje je ona u potpunosti jednaka.

Na kraju možemo zaključiti kako je stablo odlučivanja vrlo korisna prediktivna metoda koja uz minimalni trud donosi vrlo kvalitetne rezultate koji su isto tako vizualno prihvatljivi i lako shvatljivi običnim korisnicima kao što smo mogli vidjeti u ovome radu i primjenom nad podacima iz obrazovanja.

Popis literature

Rokach, M., Maimon, O. (2015). Data mining with decision trees: theory and applications. Singapur: World Scientific

Sullivan, W. (2017). Machine Learning For Beginners Guide Algorithms: Supervised & Unsupervsied Learning. Decision Tree & Random Forest Introduction. Healthy Pragmatic Solutions Inc.

Kozak, J. (2019). Decision Tree and Ensemble Learning Based on Ant Colony Optimization. Springer International Publishing.

Ross Quinlan, J. (1993). C4.5: programs for machine learning. SAD: Morgan Kaufmann Publishers.

Yang, N., Li, T., Song, J. (2007). Construction of Decision Trees based Entropy and Rough Sets under Tolerance Relation. International Journal of Computational Intelligence Systems, 10.2991/iske.2007.258.

Kent Martin, J. (1997). An Exact Probability Metric for Decision Tree Splitting and Stopping. Nizozemska: Kluwer Academic Publishers.

Aluvalu, R., Patel, R. (2014). A Reduced Error Pruning Technique for Improving Accuracy of Decision Tree Learning. International Journal of Engineering and Advanced Technology.

Amrieh, E. A., Hamtini, T., Aljarah, I. (2016). Mining Educational Data to Predict Student's academic Performance using Ensemble Methods. International Journal of Database Theory and Application, 9(8), 119-136.

Amrieh, E. A., Hamtini, T., Aljarah, I. (2015). Preprocessing and analyzing educational data set using X-API for improving student's performance. In Applied Electrical Engineering and Computing Technologies (AEECT), 2015 IEEE Jordan Conference on (pp. 1-5). IEEE.

Breiman, L. (1987). Classification and regression trees. WADSWORTH INTERNATIONAL Group.

Ayyadevara, V., K. (2018). Pro Machine Learning Algorithms: A Hands-On Approach to Implementing Algorithms in Python and R. Apress.

Mitchell, T., M. (1997). Machine learning. McGraw Hill.

Devlin, K., Lorden, G. (2007). The numbers behind NUMB3RS: solving crime with mathematics. New York: Plume.

Mueller, J., P., Massaron, L. (2016). Machine Learning for dummies. New Jersey: John Wiley & Sons.

Oreški, D., Pihir, I., Konecki, M. (2017). CRISP-DM PROCESS MODEL IN EDUCATIONAL SETTING. 20th International Scientific Conference „Economic and Social Development“, Prag.

Oreški, D., Kovač, R. (2018). Educational Data Driven Decision Making: Early Identification of Students at Risk by Means of Machine Learning. 29th CECIIS, Varaždin.

Oreški, D., Hajdin, G., Kliček, B. (2016). Role of Personal Factors in Academic Success and Dropout of IT Students: Evidence From Students and Alumni. TEM Journal, Volume 5, Issue 3, Pages 371-378.

Mesarić, J., Šebalj, D. (2016). Decision trees for predicting the academic success of students. Croatian Operational Research Review, CRORR 7, 367–388.

Kabra, K., K., Bichkar, R., S. (2011). Performance Prediction of Engineering Students using Decision Trees. International Journal of Computer Applications (0975 – 8887), Volume 36– No.11.

Marijana Zekić-Sušac (2017). Stabla odlučivanja. Preuzeto 18.8.2020. s http://www.efos.unios.hr/sustavi-poslovne-inteligencije/wp-content/uploads/sites/192/2017/10/P4_Stabla-odlucivanja-2017.pdf.

Veenadhari, S., Mishra, B., Singh, CD. (2011). Soybean Productivity Modelling using Decision Tree Algorithms. Preuzeto 18.8.2020. s https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.259.2088&rep=rep1&type=pdf

Sakkaf, Y. (2020). Decision Trees: ID3 Algorithm Explained. Preuzeto 18.8.2020. s https://towardsdatascience.com/decision-trees-for-classification-id3-algorithm-explained-89df76e72df1

Synergy37AI (2019). Tree algorithms: ID3, C4.5, C5.0 and CART. Preuzeto 18.8.2020. s https://medium.com/datadriveninvestor/tree-algorithms-id3-c4-5-c5-0-and-cart-413387342164

Brownlee, J. (2016). Classification And Regression Trees for Machine Learning. Preuzeto 18.8.2020. s https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/

Molala, R. (2019). Entropy, Information gain, and Gini Index; the crux of a Decision Tree. Preuzeto 18.8.2020. s https://blog.clairvoyantsoft.com/entropy-information-gain-and-gini-index-the-crux-of-a-decision-tree-99d0cdc699f4

Brownlee, J. (2019). Information Gain and Mutual Information for Machine Learning. Preuzeto 18.8.2020. s https://machinelearningmastery.com/information-gain-and-mutual-information/

Enciklopedija.hr (bez dat.). Entropija. Preuzeto 18.8.2020. s https://www.enciklopedija.hr/natuknica.aspx?id=18042

Guanga, A. (2019). Machine Learning Series Day 7 (Decision Tree Classifier). Preuzeto 18.8.2020. s https://becominghuman.ai/machine-learning-series-day-7-decision-tree-classifier-40e4352806f1

Zhou, V. (2019). A Simple Explanation of Information Gain and Entropy. Preuzeto 18.8.2020. s https://victorzhou.com/blog/information-gain/

BigMl (bez dat.). What kind of algorithm does BigML use to build decision tree models and how does it work? Preuzeto 18.8.2020. s https://support.bigml.com/hc/en-us/articles/206616279-What-kind-of-algorithm-does-BigML-use-to-build-decision-tree-models-and-how-does-it-work-

Jain, R. (2017). Decision Tree. It begins here. Preuzeto 18.8.2020. s https://medium.com/@rishabhjain_22692/decision-trees-it-begins-here-93ff54ef134

Hoare, J. (bez dat.). Machine Learning: Pruning Decision Trees. Preuzeto 18.8.2020. s https://www.displayr.com/machine-learning-pruning-decision-trees/

Woodruff, J. (2019). What Are the Advantages of Decision Trees? Preuzeto 18.8.2020. s https://smallbusiness.chron.com/advantages-decision-trees-75226.html

Computer & Information Science & Engineering (bez dat.). The ID3 Algorithm. Preuzeto 19.8.2020. s https://www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/2.htm

Bui, H. (bez dat.), Decision Tree Fundamentals. Preuzeto 19.8.2020. s https://towardsdatascience.com/decision-tree-fundamentals-388f57a60d2a

Popis slika

Popis slika treba biti izrađen po uzoru na indeksirani sadržaj, te upućivati na broj stranice na kojoj se slika može pronaći.

Popis tablica

Popis tablica treba biti izrađen po uzoru na indeksirani sadržaj, te upućivati na broj stranice na kojoj se tablica može pronaći.