Skupno učenje u otkrivanju znanja iz skupa podataka „Woman, Business and Law“

Posavac, Ines

Master's thesis / Diplomski rad

2021

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Rijeka / Sveučilište u Rijeci**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:195:419927

Sveučilište u Rijeci – Odjel za informatiku

Sveučilišni diplomski studij informatike

Ines Posavac

Skupno učenje u otkrivanju znanja iz skupa podataka „Woman, Business and Law“

Diplomski rad

Mentor: prof. dr. sc. Maja Matetić

Rijeka, srpanj 2021.

ODJEL ZA INFORMATIKU
SVEUČILIŠTE U RIJECI

Rijeka, 8. ožujak 2021.

Zadatak za diplomski rad

Pristupnik: **Ines Posavac**

Naziv diplomskog rada: **Skupno učenje u otkrivanju znanja iz skupa podataka „Woman, Business and Law“**

Naziv diplomskog rada na eng. jeziku: **Ensemble learning in discovering knowledge from the data set "Woman, Business and Law"**

Sadržaj zadatka:

Otkrivanje znanja u podacima koje se temelji na postupcima strojnog učenja, temelj je za analizu i unaprjeđenje poslovanja. Zadatak diplomskog rada je analizirati podatke javnog skupa podataka o društvenom statusu žena. Zadatak je izraditi modele strojnog učenja, s naglaskom na skupno učenje, te dati usporedbu točnosti modela i zaključiti koji su najznačajniji prediktori za predviđanje klase društvenog statusa.

Mentor:
Prof. dr. sc. Maja Matetić

Voditeljica za diplomske radove:
Izv. prof. dr. sc. Ana Meštrović

Komentor:

Zadatak preuzet: 8. ožujak 2021.

(potpis pristupnika)

Sažetak

Strojno učenje u današnje vrijeme više nije novost. Napredak područja izrade sve kvalitetnijih i preciznijih predviđanja je eksponencijalan. Korak dalje u tom području je metoda skupnog učenja koja ujedinjuje predviđanja dva modela u jedan. Time se postiže model koji može kroz kombiniranje modela ispraviti grešku jednog modela snagom drugog i obrnuto. Isto je tako moguće u sklopu skupnog učenja dodati metode optimizacije kao što su *bagging* i *boosting* za postizanje boljih rezultata.

Ovaj je rad pokrio nekoliko varijacija skupnog učenja, od kombiniranja istog modela s različitim parametrima, do kombiniranja različitih modela te spomenute metode *bagging* i *boosting*[1].

Sve su metode primijenjene na set podataka koje prikazuju status žena u pojedinoj državi i regiji za svaku godinu od 1971. do 2020. Da bi se odabrale pravilne metode predviđanja u rad je uključena i vizualizacija podataka.

Rezultati su pokazali zanimljive odnose među faktorima koji su utjecali na prava i status žena kroz vrijeme te da je upravo godina izračuna ključan faktor u tome. Kao najbolja metoda skupnog učenja u ovom je slučaju bilo klasifikacijsko stablo uz korištenje *boostinga,* međutim ono što su rezultati isto tako pokazali jest da skupno učenje ima brojne varijacije koje vrijedi istražiti.

Ključne riječi

Skupno učenje, klasifikacijsko stablo, naivni Bayes, bagging, boosting, ženska prava

[1] Engl. *Bagging, boosting* – s obzirom da je teško pronaći odgovarajući prijevod na hrvatski jezik, izrazi za ove pojmove će se koristiti na engleskom jeziku kroz cijeli rad

1 Sadržaj

2 Uvod

Rad je napisan u sklopu kolegija „Otkrivanje znanja u podacima“, čiji naziv otkriva i sam cilj nadolazećeg sadržaja. Naime, na temelju određenih postupaka, može se iz prikupljenih podataka filtrirati znanje o utjecaju parametara, njihovoj zavisnosti i smjeru kretanja određenih vrijednosti. Ove i brojne druge informacije, dobivene analizom podataka, korisne su za analizu poslovanja, prema kojoj se otkrivaju inovativni načini za unaprjeđenje. Kako se prikupljeni podaci mogu analizirati, koju metodu odabrati i kako razumjeti podatke, objašnjava Shmueli u „Data Mining for Business Analytics“ [1]. Ona nudi brojne načine za manipulaciju i analizu podataka, a neki od njih bit će opisani u radu. Ova je literatura korištena kao primarna, dok je u nekim slučajevima, kako bi se izvukla paralela, uključen i izvor „Combining Pattern Classifiers“, Kuncheva [2].

U poglavlju broj 3 bit će opisana sama metoda skupnog učenja (engl. Ensemble learning). Uz sam opis, naveden će biti i pregled postupaka gdje se takav način kombiniranja može koristiti i koji će algoritmi biti korišteni. S ovim pristupom dolaze prednosti i nedostaci, a uz opis istih, naveden će biti i primjer gdje se takav pristup koristi i što je potaklo granu strojnog učenja na korištenje više od jednog pristupa.

Najprije, kako bi se postupci, ali i implementacija istog razumjeli, bit će prikazan korišteni set podataka nad kojim su korišteni postupci po uzoru na Shmueli [1]. Skup podataka sadrži podatke o društvenom statusu žena na temelju definiranih faktora i indikatora. Nakon što je sadržaj seta podataka poznat, u sljedećem će poglavlju biti prikazane vizualizacije, kako bi se odnosi među podacima mogli lakše shvatiti. Za poprilično velik set podataka kao što je ovaj, sačinjen od kategorija, potrebno je provesti agregaciju, kako bi se došlo do kvalitetnog znanja iz podataka.

Nadalje slijedi opis i implementacija postupaka klasifikacije kroz različite varijacije skupnog učenja, kao što su kombiniranje modela istog tipa, u ovom slučaju naivnog Bayesa s različitim parametrima, kombiniranje dva različita modela – naivnog Bayesa i klasifikacijskog stabla odluke, te primjena *Bagginga* i *Boostinga* kao dvije metode skupnog učenja. Svaka metoda će biti pobliže objašnjena i primijenjena, te će rezultati biti prikazani kako bi se usporedio utjecaj pristupa skupnog učenja u odnosu na korištenje samo jednog postupka.

U konačnici će u zaključku biti opisana sva zapažanja, zaključci i pregled aktualnih smjerova istraživanja u tome području.

3 Uvod u skupno učenje: kombinacija metoda i postupaka

Ensemble ili skupno učenje, kako sam pojam opisuje, označava kombinaciju ili skup različitih metoda i postupaka učenja. Shmueli [1] predstavlja metodu skupnog učenja uz modeliranje uzdizanjem (engl. Upplift modeling) kao jedan od pristupa kombiniranja kako bi se postiglo kvalitetnije predviđanje ili klasificiranje, tvoreći „super-model“. Ono što razlikuje skupno učenje od modeliranja uzdizanjem, jest što se skupnim učenjem kombiniraju modeli kako bi se dobili točniji ishodi. Modeliranje uzdizanjem, kako navodi Shmueli [1] čini jedan nasumičan eksperiment kojem je cilj direktno utjecati na ponašanje svakog modela, prema prethodno dobivenim rezulatatima predviđanja svakog pojedinačnog modela.

Skupno učenje kako Shmueli [1] navodi, prema prirodi podataka u odnosu na ostale metode za rudarenje podataka, pripada zapravo u nekoliko kategorija. Ovisno o tome koje se metode kombiniraju i u koju svrhu, metoda skupnog učenja se kategorizira prema tom kriteriju. Ono što je važno naglasiti jest da metoda skupnog učenja ne čini novu metodu predviđanja ili/i klasifikacije, već čini kombinaciju postojećih metoda.

Shmueli [1] u svojoj knjizi sadržaj dijeli na nekoliko grupa poglavlja, gdje pritom jednu grupu čine metode nadziranog učenja (engl. supervised learning methods), uključujući i skupno učenje, pa tako i sama metoda skupnog učenja pripada nadziranom učenju. Osim prema nadziranju, Shmueli [1] dijeli metode i prema odazivu, onom ponavljajućem (engl. Continuous Response) ili onom kategoričkom (engl. Categorical Response) te predviđanjima, isto tako ponavljajućima (engl. Continuous Predictors) ili kategoričkim predviđanjima (engl. Continuous Predictors). Metoda skupnog učenja se ovdje ponovno vodi za onim metodama koje kombinira, što znači da može pripadati svakoj od tih četiri skupina. Kako su te skupine raspoređene može se vidjeti u tablici broj 1 dolje.

	Nadzirano učenje	
	Ponavljajući odaziv	Kategorički odaziv
Ponavljajući prediktori	Linearna regresija[2] Neuralne mreže[3] k-NN[4] Skupno učenje	Logistička regresija[5] Neuralne mreže Analiza diskriminante[6] k-NN Skupno učenje
Kategorički prediktori	Linearna regresija Neuralne mreže Regresijska stabla[8] Skupno učenje	Neuralne mreže Klasifikacijska stabla[7] Logistička regresija Naivni Bayes[9] Skupno učenje

Tablica 1 Kategorizacija predviđanja prema Shmueli [1]

3.1 Pregled postupaka kod kojih se koristi skupno učenje

Shmueli [1] opisuje brojne metode rudarenja podataka u svrhu analize poslovanja, međutim kao dio uvodnog pregleda objašnjava i zašto postoji zapravo toliki broj različitih metoda. Po uzoru na isti pristup objašnjenje raznolikosti metoda će poslužiti kao uvod u samu svrhu korištenja skupnog učenja. Sama autorica Shmueli [1] naznačuje da postoje brojne metode za predikciju i klasifikaciju, te netko tko nema znanje o funkcionalnosti tih metoda, može se zapitati zašto uopće postoje brojne metode za istu svrhu. Naime, ako je jedna bolja od druge, uvijek bi se ista metoda odabrala. Ono što ih razlikuje je upravo razlog zašto odluka pada na određenu metodu u određenoj situaciji. Mnogi parametri ovdje odlučuju o tome koja će se metoda koristiti, pri čemu Shmueli [1] navodi veličinu seta podataka i tipove uzoraka koji postoje u setu podataka. Nadalje se postavlja i pitanje poklapa li se set podataka s nekim pretpostavkama same metode, ali u konačnici i sami cilj analize, tj. kakvi rezultati su potrebni i koja perspektiva rezultata će najviše znanja izvući iz podataka.

[2] Engl. Linear Regression
[3] Engl. Neural nets
[4] Engl. k-Nearest neighbors
[5] Engl. Logistic regression
[6] Engl. Discriminant analysis
[7] Engl. Classification trees
[8] Engl. Regression trees
[9] Engl. Naive Bayes

Metoda skupnog učenja može se koristiti za predviđanja, klasifikaciju, ono može funkcionirati i kroz jednostavan prosjek (engl. Simple Averaging) dobivenih rezultata pojedinačnih metoda zajedno, a uz to *bagging* i *boosting* su isto tako dvije metode koje spadaju u metodu skupnog učenja. Prema Shmueli [1] koja navodi ove postupke, organiziran je i ovaj rad kako bi se prikazalo djelovanje skupnog učenja na rezultate s obzirom na rezultate koji se dobiju pojedinačnim modelima. Prema setu podataka odabrane su pojedinačne metode kako bi se dobili kvalitetni rezultati. Važno je dakako odabrati one metode koje najviše odgovaraju specifičnoj situaciji.

3.2 Primjeri korištenja postupaka skupnog učenja

Primjer koji Shmueli [1] navodi za korištenje skupnog učenja jest vezano uz natjecanje Netflixa, američkog pružatelja prijenosa serija i filmova putem interneta. Naime, radilo se o nagradi od milijun dolara, a ono je započelo 2006. godine. U to je vrijeme Netflix još uvijek poslovao posudbom DVD-a i bio najveća i najuspješnija američka kompanija toga tipa na tržištu. Kako se ovdje radi o filmovima, da bi njihov posao funkcionirao što bolje, tj. da bi što više filmova bilo posuđeno, važno je bilo korisnicima ponuditi dobre preporuke filmova na temelju prijašnjeg izabranog sadržaja.

U tom je trenutku, kako izvor [1] navodi, Netflix koristio alat pod nazivom *„Cinematch“* koji je davao pristojne rezultate, međutim još uvijek rezultati nisu bili zadovoljavajući u odnosu na tadašnje tržište. Kako bi Netflix svojim korisnicima pružio bolje iskustvo gledanja i time unaprijedio poslovanje, bilo je objavljeno natjecanje među svim korisnicima. Zadano je bilo u timovima izgraditi optimalni algoritam za predikciju filmova koji bi se generirali svakom korisniku pojedinačno prema njegovom prijašnjem odabiru filmova te njegovim ocjenama za svaki pregledani film. Netflix je zbog takve prilike za poboljšanjem objavio skup podataka temeljen na anonimnim ocjenama koje su skupili tijekom godina od svojih klijenata o pregledanim filmovima.

Svi su bili pozvani osnovati tim te osmisliti svoje rješenje, a rezultati svakog tima bili su predstavljeni na ljestvici koja je prikazivala trenutno stanje. Ono na što organizatori natjecanja nisu računali, jest da su se timovi počeli udruživati kako bi zajedničkim snagama došli do boljeg rješenja. Tako je pobjednički tim zapravo bio ujedinjenje dva vodeća tima koji su na natjecanje prijavili svoje kombinirano rješenje.

Shmueli [1] prenosi objašnjenje dobitnika nagrade za korišteni pristup u kojem oni opisuju da su modeli pojedinačno imali puno lošije rezultate nego kombinirano, a da je lakše bilo izračunati prosjek dva seta, nego pokušaj razvoja novog modela koji bi obuhvaćao obje metode.

3.3 Prednosti i nedostaci skupnog učenja

Prema Shmueli [1], kombiniranje rezultata iz više modela usmjereno je na generiranje preciznijih predviđanja (snižavanje varijance pogreške predviđanja). Pristup skupnog učenja je korisno kad kombinirani modeli generiraju pogreške predviđanja koje su negativno povezane, ali također može biti korisno kada je korelacija niska. Naime, ako je korelacija pogreške niska, jedan model će imati točna predviđanja za segmente u kojima je drugi pogriješio i obrnuto, čime će si zapravo modeli međusobno ispravljati greške. Skupno je učenje tako postalo glavna strategija za sudionike u natjecanja u rudarenju podataka, gdje je cilj optimizirati neku prediktivnu mjeru. Skupno učenje također pruža operativan način dobivanja rješenja s visokom prediktivnom snagom na brz način, uključivanjem više timova "kradljivaca podataka", kako ih Shmueli [1] naziva, koji rade paralelno i kombiniraju svoje rezultate.
Osim toga Shmueli [1] navodi i neke nedostatke skupnog učenja. Glavni nedostaci su pritom resursi koji su mu potrebni: računalno, kao i u smislu dostupnosti softvera i vještine analitičara i vremenskog ograničenja. Modeli skupnog učenja koji kombiniraju rezultate različitih algoritama zahtijevaju razvoj svakog modela i njihovu procjenu. Metode *bagging* i *boosting* tipovi su skupnog učenja koji ne zahtijevaju takav napor, ali zahtijevaju računalni trošak. S druge strane skupno učenje koje se oslanja na više izvora podataka zahtijeva prikupljanje i održavanje više izvora podataka. I na kraju, ono što predstavlja nedostatak jest da je to metoda "blackbox" modela, jer odnos između prediktora i varijable ishoda obično postaje netransparentan.

4 Optimizacija rezultata kroz skupno učenje

Shmueli [1] kao jedan od razloga za popularnost skupnog učenja spominje smanjenje rizika. Rizik u području predikcije autorica navodi kao ekvivalent varijaciji u pogrešci predikcije. Kako bi čitatelju to bilo jasnije, Shmueli [1] prikazuje navedeno na primjeru iz financija. Naime, u financijskom sektoru koriste se portfelji, kao dokument koji obuhvaća različite financijske segmente pojedinca, kao što su imovina, prihodi,

izdaci itd. Ako se uzme samo jedan od tih segmenata u obzir, teško je procijeniti financijsku situaciju pojedinca, jer iako jedan od segmenata može prikazivati povoljnu situaciju, ostali mogu biti u nepogodnoj situaciji za ulaganje. Rizik je u toj situaciji podosta visok, međutim ako se cijeli portfelj uzima u obzir, veća je preglednost cjelokupne financijske situacije pojedinca te je time rizik od opasnosti nepogodnog ulaganja manji.

Rizik se u predikciji ishoda i trenda podataka definira kao varijacija u greškama predikcije. Što više greške predikcije variraju, to je slabiji model. Postići model visoke točnosti kompleksan je proces gotovo uvijek, međutim ako se pojavljuju razne varijacije grešaka, teže ih je utvrditi i uzeti u obzir za daljnje istraživanje, što čini taj proces još kompleksnijim. Kako bi cijela teorija o smanjenju rizika kroz kombiniranje ulaznih faktora i njihovih rezultata bila dokazana, autorica [1] opisuje odnos modela kroz definiranu jednadžbu. Time uzima u obzir set od *n* zapisa koji koriste dva različita modela, pritom $e_{1,i}$ označava grešku predikcije za *i*-ti zapis za metodu broj 1, dok je $e_{2,i}$ greška predikcije za isti zapis za metodu broj 2. Uzima se kao pretpostavka da će svaki od navedenih modela proizvesti prosjek grešaka predikcije jednak nuli. Naime, za neke će zapise modeli precijeniti rezultate, dok će za neke podcijeniti. Ako se uzme u obzir prosjek oba slučaja, on će biti nula prema formuli:

$$E(e_{1,i}) = E(e_{2,i}) = 0.$$

Jednadžba 1 Prosjek grešaka predikcije

Nadalje Shmueli [1] objašnjava da ako se uzima za svaki zapis u setu podataka prosjek dvije predikcije, srednja vrijednost greške će ponovno biti 0 prema:

$$\begin{aligned} E\left(y_i - \overline{y_i}\right) &= E\left(y_i - \frac{\hat{y}_{1,i} + \hat{y}_{2,i}}{2}\right) \\ &= E\left(\frac{y_i - \hat{y}_{1,i}}{2} + \frac{y_i - \hat{y}_{2,i}}{2}\right) = E\left(\frac{e_{1,i} + e_{2,i}}{2}\right) = 0. \end{aligned}$$

Jednadžba 2 Srednja vrijednost greške prema prosjeku dvije predikcije

Upravo to pokazuje da se skupnim učenjem dolazi do iste srednje vrijednosti koja se dobije i od pojedinačnih modela. U nastavku autorica [1] istražuje varijancu grešaka predikcije uz skupno učenje:

$$Var\left(\frac{e_{1,i} + e_{2,i}}{2}\right) = \frac{1}{4}\left(Var(e_{1,i}) + Var(e_{2,i})\right) + \frac{1}{4} \times 2Cov\left(e_{1,i}, e_{2,i}\right).$$

Jednadžba 3 Varijanca grešaka predikcije uz skupno učenje

Varijanca, naime, u skupnom učenju prema autorici [1] može biti manja nego varijanca svakog modela pojedinačno tj. $Var(e_{1,i})$ i $Var(e_{2,i})$ u određenim okolnostima. Najvažnija komponenta je ovdje ekvivalencija ili korelacija između grešaka tih dviju pojedinačnih predikcija. U slučaju da korelacija uopće ne postoji, dolazi do povoljne situacije gdje kvantiteta postaje manja nego pojedinačne varijance, a varijanca prosječne greške predikcije će biti manja, ako su greške dviju korištenih predikcija u negativnoj korelaciji.

Shmueli [1] dolazi ovdje do zaključka da bi korištenje prosjeka dviju pojedinačnih predikcija moglo dovesti do manje varijance greške, pa tako i do boljeg rezultata i točnije predikcije. Isto tako se navodi kako je to generalno pravilo koje se može primijeniti na bilo koji model za predikciju ili klasifikaciju, što znači da je izbor korištene metode dosta slobodan i otvoren za istraživanje i testiranje, kako bi se došlo do što preciznije predikcije.

4.1 Skupno učenje uz korištenje prosjeka rezultata nekoliko modela

Za kombinaciju dviju metoda predikcija postoji nekoliko načina, međutim ona koja je po autorici [1] najjednostavnija jest upravo jednostavan prosjek. Prosjek rezultata se može koristiti za svaku kombinaciju predikcija, klasifikacija ili tendencija (engl. Combining Propensities). Shmueli [1] kao primjer navodi model linearne regresije, stablo odluke i k-NN model. Svaki od tih modela će se izraditi nad istim setom podataka i time se dobiva set za testiranje, potom se ujedinjuju sva tri seta rezultata stavivši pritom na svaki rezultat jednaku težinu. Naime, ako se koristi jednostavan prosjek, neće se rezultatima dodati zasebna težina vrijednosti, kao u težinskom prosjeku (engl. weighted average), već će ona biti podijeljena jednako, bez obzira na preciznost svakog zasebnog modela.

Međutim, kako i sama Shmueli [1] objašnjava, ovakav pristup može se primijeniti, ne samo kao prosjek rezultata različitih metoda predikcije ili klasifikacije, već i na jednu metodu samu, uzevši pritom različite parametre u svakoj predikciji. Kao primjer je navedena linearna regresija, koja se može koristiti nekoliko puta s različitim parametrima.

4.1.1 Kombiniranje predikcija

Predikcije su metode koje produciraju numeričke rezultate. Numeričkim rezultatima je vrlo jednostavno izračunati prosjek, podijelivši sumu s brojem elemenata i to je primjenjivo na svaku metodu predikcije.

Ako se kao primjer uzmu tri različita modela, npr. oni već gore navedeni (model linearne regresije, stablo odluke i k-NN model), za svaki će element u setu za testiranje biti tri različite predikcije, a ako se uzme prosjek tih tri vrijednosti, dobit će se primjer skupnog učenja. Osim prosjeka Shmueli [1] spominje i metodu uzimanja srednjeg predviđanja (engl. median prediction), na koje bi manji utjecaj imale ekstremne predikcije. Još je jedna mogućnost navedena uz to - težinski prosjek (engl. weighted average). Upravo ono što je u prethodnom poglavlju (3.2.1) kao nedostatak korištenja jednostavnog prosjeka, težinski prosjek ispravlja. Težina se pritom proporcionalno dodjeljuje kvantiteti interesa. Težina se tako može rasporediti proporcionalno prema preciznosti svakog modela pojedinačno ili prema različitim izvorima podataka koji se koriste, ako se oni razlikuju. U tom će slučaju presudni faktor biti kvaliteta samih podataka.

Skupno učenje se u izvoru [1] opisuje kao korisna metoda za vremenske prognoze. Kombinacijom nekoliko različitih predikcija vremena u budućnosti, dolazi se do točnije prognoze. Primjena skupnog učenja u tu svrhu je već danas popularna, a autorica [1] opisuje kako takve aplikacije uzimaju mnogobrojne izvore predikcija vremena u budućnosti kako bi dobile što točniju prognozu za svoje korisnike.

4.1.2 Kombiniranje klasifikacija

U slučaju klasifikacije gdje nema brojčane vrijednosti, već klase, isto se tako može koristiti dodjeljivanje veće težine određenoj predikciji tj. klasifikaciji prema točnosti i preciznosti modela ili kvaliteti podataka u setu. Na taj se način izbjegava korištenje prosjeka te se može kreirati djelotvornija kombinacija modela.

Kuncheva [2] tako opisuje optimizaciju odlučivanja (engl. Decision Optimization) i optimizaciju pokrivenosti (engl. Coverage Optimization) kroz skupno učenje sa različitim klasifikatorima. Naime, brojni su načini kako se modeli mogu kombinirati, ovisno o tome kakvi su rezultati potrebni. Optimizacija odlučivanja odnosi se na metode odabira i najboljeg načina kombinacije modela za fiksni skup klasifikatora gdje je

primarni fokus na dizajnu različitih klasifikacija za skupno učenje. Optimizacija pokrivenosti se s druge strane odnosi na metode za stvaranje različitih osnovnih klasifikatora koji predstavljaju temelj kombinacije. U ovaj način optimizacije Kuncheva [2] svrstava i korištenje različitih setova podataka ili različitih setova značajki (engl. Feature subsets).

Nadalje Kuncheva [2] razlikuje isto tako skupna učenja u kojima je za ujedinjeni model potrebno dodatno treniranje modela i one modele skupnog učenja koji ne trebaju obuku nakon što su klasifikatori trenirani pojedinačno. Primjer za one koji ne trebaju dodatno treniranje bila bi prema autorici [2] kombinacija većinskih glasova (engl. voting) ili kako je Shmueli naziva *„pristup glasovanja“*. Kombiniranje težinskim prosjekom i jednostavnim prosjekom bi bili primjeri za modele skupnog učenja koji trebaju dodatno treniranje, čak i nakon što su klasifikatori sami istrenirani.

4.1.3 Kombiniranje tendencija

Nakon klasifikacije i predikcije, Shmueli [1] spominje i tendencije(engl. Combining Propensities), koje se isto tako mogu kombinirati jednostavnim prosjekom. Međutim kombiniranje tendencija funkcionira za neke modele bolje nego za druge. Tako na primjer naivni Bayes producira pristrane tendencije gdje je teško njegove rezultate uzeti kao jednak dio prosjeka uz rezultate drugih modela. U tome bi slučaju težinski prosjek funkcionirao bolje te proizveo točniji model skupnog učenja.

4.2 Bagging

Drugi oblik skupnog učenja koji Shmueli [1] navodi, temelji se na prosjeku više nasumičnih uzoraka podataka. Dakle, više se ovdje ne radi na određenoj kombinaciji dva različita modela, već na metodi koja se nadovezuje na model, kako bi ga učinila preciznijim. *Bagging*, skraćeno od "*bootstrap agregating*", izvodi se u dva koraka. Najprije se više nasumičnih uzoraka generira, pa se oni zamjenjuju izvorne podatke. Taj se prvi dio metode naziva "uzorkovanje bootstrapa". Zatim slijedi pokretanje algoritma nad svakim uzorkom kako bi se reproducirali rezultati.

Ova metoda poboljšava stabilnost performansi modela i pomaže izbjeći pretjeranog prilagođavanja modela (engl. overfitting). Zasebno se oblikuju različiti uzorci podataka, a zatim se rezultati kombiniraju. Stoga Shmueli [1] zaključuje da je ovaj pristup posebno koristan za stabla odluke i neuronske mreže.

4.3 *Boosting*

Boosting ne samo da ima drugačije korake, već ima i drugačiji način poboljšanja određenog modela. Ovdje je cilj, prema Shmueli [1], izravno poboljšati segmente podataka u kojima određeni model griješi, navodeći time model da obrati više pozornosti na te zapise koji su pogrešni.

U *Boostingu* se tako model najprije prilagodi podacima, zatim se predstavi uzorak podataka koji navodi model da krivo klasificirane elemente skupa prije uzme u obzir, jer su oni onaj dio podataka koji treba biti ispravljen. Model se na kraju prilagodi novom uzorku i sve se zajedno zatim ponovi nekoliko puta.

4.4 Odabrani modeli koji će se obraditi u radu

Nakon predstavljenih metoda *bagginga* i *boostinga* koji se koriste za skupno učenje, važno je opisati i glavne modele koji će se koristiti. Naime, s obzirom na set podataka i da će u ovom radu se predstaviti metode klasifikacije, odabrani su i modeli.

Uvest će se u tu svrhu naivni Bayes (engl. Naive Bayes), koji se može primijeniti na podatke s kategoričkim prediktorima. Shmueli [1] opisuje ovaj model kao jednostavan te objašnjava da za svaki zapis koji treba klasificirati treba pronaći sve ostale zapise s istim prediktorskim profilom, odrediti kojim klasama zapisi pripadaju i koja je klasa najviše utjecajna pa se ta klasa dodijeljuje novom zapisu. Međutim, ono što ovaj model čini „naivnim“ jest činjenica da model radi tako da kreće od pretpostavke da prisutnost određene značajke za neku grupu ili klasu nije povezana s prisutnošću bilo koje druge značajke. Da bi se klasificirao zapis, izračunava se njegova vjerojatnost pripadnosti svakoj od klasa. Pa se tako klasificira zapis u razred koji ima najveću vjerojatnost. Model sam po sebi nije kompliciran za implementaciju, a može biti podosta koristan za predviđanje klase nad velikim skupom podataka. Ono što Shmueli [1] napominje, a dokazat će se i u daljnjem tekstu jest da se ovdje radi o modelu koji radi samo s kategoričkim prediktorima. Ako se koristi skup numeričkih prediktora, onda je vrlo mala vjerojatnost da će više zapisa imati identične vrijednosti na tim numeričkim prediktorima. Stoga se numerički prediktori moraju pretvoriti u kategoričke prediktore, što je u ovom radu učinjeno pretvaranjem WBL indeksa u četiri klase prema rasponu vrijednosti. Unatoč nekim nedostacima koje ovaj model ima, kao što je potreba da sve klase koje će

se predviđati obavezno se moraju naći u setu za testiranje, inače je model nikako neće predvidjeti, on je dobar kandidat za potrebu ovog seta podataka.

Za razliku od modela naivnog Bayesa, drugi model koji će se koristiti poprilično je fleksibilan. Stabla se mogu koristiti jednako učinkovito za klasifikaciju kao i predviđanje (stablo regresije). Prema Shmuel [1] stabla su najtransparentnija i najlakša za tumačenje. Stabla se temelje na odvajanju zapisa u podskupine stvaranjem podjela na prediktore. Podjele koje stvaraju logička pravila su transparentna i lako razumljiva, na primjer, „AKO dob < 55 godina I obrazovanje > 12 ZATIM razred = 1“. Nastale podskupine trebale bi biti homogenije u pogledu ishoda varijable, čime se stvaraju korisna pravila predviđanja ili klasifikacije. Ono što stabla razlikuje od naivnog Bayesa jest da stabla zahtijevaju veći opseg seta podataka, no njihova prednost leži u tome da jako dobro vladaju s podacima koji nedostaju, dok naivni Bayes neće prepoznati nešto čime se nije susreo. Da bi se klasificirao novi zapis, on se "spušta" niz stablo. Kada padne sve do terminalnog čvora (engl. terminal node), može mu se dodijeliti njegova klasa jednostavnim uzimanjem "glasovanja" sa svim podacima koji su pripadali putanji do terminalnog čvora kada je drvo je izgrađeno. Razred s najvećim brojem glasova uzima novi zapis. Ono što Shmueli [1] napominje kao prednost korištenja klasifikacijskih stabla i što ih čini veoma popularnima, jest što pružaju lako razumljiva pravila klasifikacije i jednako dobro mogu savladati klase kao i brojčane vrijednosti, što će biti dokazano pri implementaciji ovog modela u poglavlju broj 6.

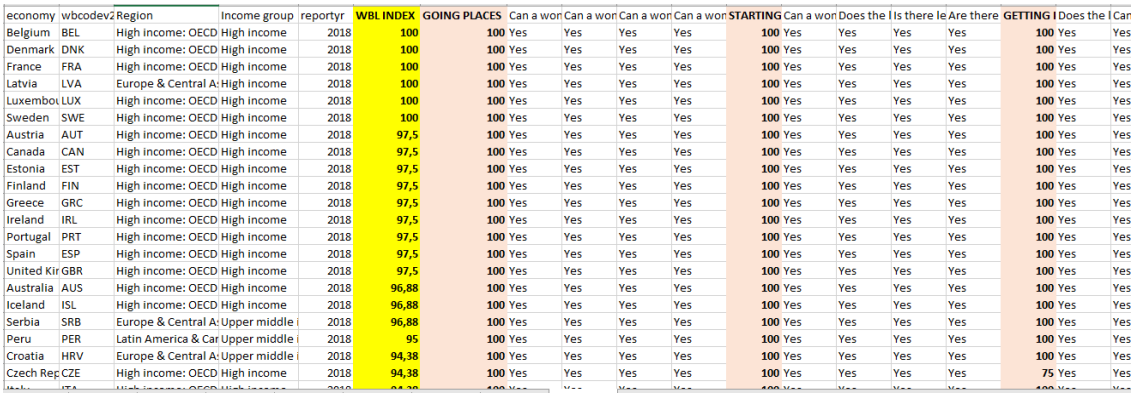
Kako Kuncheva [2] opisuje, važno je znati koliko dobro funkcionira korišteni klasifikator. Izvedba klasifikatora je spoj njegovih karakteristika, čija je najvažnija komponenta točnost klasifikacije. Kad bi se mogli isprobati klasifikatori na svim mogućim ulaznim objektima, točno bi se znalo koliko je svaki zapis točno predviđen. Nažalost, teško da je ovaj scenarij moguć, pa se umjesto toga mora upotrijebiti procjena točnosti. Za svrhu procjene točnosti navedenih modela, koristit će se konfuzijska matrica koja pokazuje nekoliko značajki predviđanja, a one će biti opisane kasnije u radu. Kuncheva [2] opisuje konfuzijsku matricu kao način da bi se saznalo kako se pogreške raspoređuju po razredima. Konstruira se konfuzijska matrica pomoću skupa podataka za testiranje, Z_{ts}. Zapis a_{ij} takve matrice označava broj elemenata iz Z_{ts} čija je prava klasa ω_i, a koje dodjeljuje D razredu ω_j.

5 Korišteni set podataka

Prema izvoru [2] set podataka čini skup informacija za kreiranje klasifikatora. Set podataka čini $Z = \{z_1, ..., z_n\}$, $z_j \sum R^n$. Označena klasa z_j je tako obilježena s $l(z_j) \sum \Omega, j = 1, ..., N$. Autorica objašnjava kako je često teško odrediti značajke u setu podataka tj. kakav će utjecaj imati jedan faktor klasifikacije na drugi. Ako se ponovno uzme primjer iz financija, teško je analitičaru bez iskustva unaprijed pretpostaviti koliko će točno utjecaja broj djece imati na financijsku stabilnost i je li lakše izračunati financijsko stanje pojedinca uzevši u obzir broj djece ili iznos rate kredita. Međutim, pregledom skupa podataka i određivanjem koji su faktori ključni i koliko koji od njih ima utjecaja na druge čini dobru podlogu za klasifikaciju. Tako je u ovom poglavlju proveden pregled seta podataka s kojim će se raditi. Oni podaci koji nisu bili u pogodnom formatu za izradu modela klasifikacije su transformirani, a drugi podaci su grupirani kako bi se promotrile klase, a ne zasebne vrijednosti za lakšu preglednost.

Podaci koji su korišteni u ovom radu, preuzeti su u *csv* formatu s web stranice *The world bank* [3], koja nudi slobodno korištenje podataka koji su rezultat njihova istraživanja. Organizacija se bavi nejednakošću spolova u poslovnoj i pravnoj sferi, pritom prema [3] slave napredak koji je učinjen od početka istraživanja do sad, međutim smatraju da situacija još uvijek nije na zadovoljavajućoj razini. Konkretno set podataka čini prikaz demografskih čimbenika, koji su uzeti u obzir za kalkulaciju tzv. „WBL indeksa“. Informacije su unesene za svaku godinu od 1971. do 2020. godine, uz koje dolazi i sumirani prikaz. 190 zemalja je prikazano na popisu, pri čemu su podijeljene u regije kako bi se na temelju većih kategorija lakše mogli donositi zaključci. Regije se dijele na Europu i centralnu Aziju, Subsaharsku Afriku, Južnu Ameriku, istočnu Aziju i Pacifik, Srednji Istok i sjevernu Aziju, južnu Aziju i Grupu OECD (prema izvoru [5] Organizacija za ekonomsku suradnju i razvoj[10]) koju čine Austrija, Belgija, Češka Republika, Danska, Estonija, Francuska, Njemačka i druge države, a cilj im je postići ekonomsko dobrostanje članova grupe.

[10] Engl. Organisation for Economic Co-operation and Development



economy	wbcodev2	Region	Income group	reportyr	WBL INDEX	GOING PLACES	Can a wor	Can a wor	Can a wor	Can a wor	STARTING	Can a wor	Does the l	Is there le	Are there	GETTING I	Does the l	Can
Belgium	BEL	High income: OECD	High income	2018	100	100	Yes	Yes	Yes	Yes	100	Yes	Yes	Yes	Yes	100	Yes	Yes
Denmark	DNK	High income: OECD	High income	2018	100	100	Yes	Yes	Yes	Yes	100	Yes	Yes	Yes	Yes	100	Yes	Yes
France	FRA	High income: OECD	High income	2018	100	100	Yes	Yes	Yes	Yes	100	Yes	Yes	Yes	Yes	100	Yes	Yes
Latvia	LVA	Europe & Central A	High income	2018	100	100	Yes	Yes	Yes	Yes	100	Yes	Yes	Yes	Yes	100	Yes	Yes
Luxembou	LUX	High income: OECD	High income	2018	100	100	Yes	Yes	Yes	Yes	100	Yes	Yes	Yes	Yes	100	Yes	Yes
Sweden	SWE	High income: OECD	High income	2018	100	100	Yes	Yes	Yes	Yes	100	Yes	Yes	Yes	Yes	100	Yes	Yes
Austria	AUT	High income: OECD	High income	2018	97,5	100	Yes	Yes	Yes	Yes	100	Yes	Yes	Yes	Yes	100	Yes	Yes
Canada	CAN	High income: OECD	High income	2018	97,5	100	Yes	Yes	Yes	Yes	100	Yes	Yes	Yes	Yes	100	Yes	Yes
Estonia	EST	High income: OECD	High income	2018	97,5	100	Yes	Yes	Yes	Yes	100	Yes	Yes	Yes	Yes	100	Yes	Yes
Finland	FIN	High income: OECD	High income	2018	97,5	100	Yes	Yes	Yes	Yes	100	Yes	Yes	Yes	Yes	100	Yes	Yes
Greece	GRC	High income: OECD	High income	2018	97,5	100	Yes	Yes	Yes	Yes	100	Yes	Yes	Yes	Yes	100	Yes	Yes
Ireland	IRL	High income: OECD	High income	2018	97,5	100	Yes	Yes	Yes	Yes	100	Yes	Yes	Yes	Yes	100	Yes	Yes
Portugal	PRT	High income: OECD	High income	2018	97,5	100	Yes	Yes	Yes	Yes	100	Yes	Yes	Yes	Yes	100	Yes	Yes
Spain	ESP	High income: OECD	High income	2018	97,5	100	Yes	Yes	Yes	Yes	100	Yes	Yes	Yes	Yes	100	Yes	Yes
United Kir	GBR	High income: OECD	High income	2018	97,5	100	Yes	Yes	Yes	Yes	100	Yes	Yes	Yes	Yes	100	Yes	Yes
Australia	AUS	High income: OECD	High income	2018	96,88	100	Yes	Yes	Yes	Yes	100	Yes	Yes	Yes	Yes	100	Yes	Yes
Iceland	ISL	High income: OECD	High income	2018	96,88	100	Yes	Yes	Yes	Yes	100	Yes	Yes	Yes	Yes	100	Yes	Yes
Serbia	SRB	Europe & Central A	Upper middle i	2018	96,88	100	Yes	Yes	Yes	Yes	100	Yes	Yes	Yes	Yes	100	Yes	Yes
Peru	PER	Latin America & Car	Upper middle i	2018	95	100	Yes	Yes	Yes	Yes	100	Yes	Yes	Yes	Yes	100	Yes	Yes
Croatia	HRV	Europe & Central A	Upper middle i	2018	94,38	100	Yes	Yes	Yes	Yes	100	Yes	Yes	Yes	Yes	100	Yes	Yes
Czech Rep	CZE	High income: OECD	High income	2018	94,38	100	Yes	Yes	Yes	Yes	100	Yes	Yes	Yes	Yes	75	Yes	Yes

	A	B	C	D	E	F	G	H
	ID	eco0my	wbcodev2	Region	Income gr	reportyr	WBL INDEX	WBL INDEX GROUP
	AFG1971	Afghanistan	AFG	South Asia	1	1971	26,3	2
	AFG1972	Afghanistan	AFG	South Asia	1	1972	26,3	2
	AFG1973	Afghanistan	AFG	South Asia	1	1973	26,3	2
	AFG1974	Afghanistan	AFG	South Asia	1	1974	26,3	2
	AFG1975	Afghanistan	AFG	South Asia	1	1975	26,3	2
	AFG1976	Afghanistan	AFG	South Asia	1	1976	26,3	2
	AFG1977	Afghanistan	AFG	South Asia	1	1977	26,3	2
	AFG1978	Afghanistan	AFG	South Asia	1	1978	26,3	2
	AFG1979	Afghanistan	AFG	South Asia	1	1979	26,3	2
	AFG1980	Afghanistan	AFG	South Asia	1	1980	26,3	2
	AFG1981	Afghanistan	AFG	South Asia	1	1981	26,3	2
	AFG1982	Afghanistan	AFG	South Asia	1	1982	26,3	2
	AFG1983	Afghanistan	AFG	South Asia	1	1983	26,3	2
	AFG1984	Afghanistan	AFG	South Asia	1	1984	26,3	2
	AFG1985	Afghanistan	AFG	South Asia	1	1985	26,3	2

Slika 1 WBL set podataka s pobliže prikazanim faktorima

5.1 Indikatori WBL indeksa

Indikatori WBL indeksa čine kategorije vezane uz njena ekonomska, politička i socijalna prava u državi. Kategorije koje su navedene u setu podataka su: mobilnost van države, mogućnost zaposlenja, stanje i status na radnom mjestu, uloga u braku i njena prava u bračnoj zajednici, prava po pitanju djece, ženska administrativna prava i prava upravljanja financijama, dokumentacijom i poslovanjem te u konačnici prava u segmentu mirovine. Cijeli popis nalazi se u tablici koja je preuzeta sa izvora [3] sa setom podataka te se može vidjeti na slici 2 dolje.

GOING PLACES
Can a woman apply for a passport in the same way as a man?
Can a woman legally travel outside the country in the same way as a man?
Can a woman legally travel outside her home in the same way as a man?
Can a woman legally choose where to live in the same way as a man?
STARTING A JOB
Can a woman get a job or pursue a trade or profession in the same way as a man?
Does the law mandate nondiscrimination based on gender in employment?
Is there legislation on sexual harassment in employment?
Are there criminal penalties or civil remedies for sexual harassment in employment?
GETTING PAID
Does the law mandate equal remuneration for work of equal value?
Can women work the same night hours as men?
Can women work in jobs deemed hazardous, arduous or morally inappropriate in the same way as men?
Are women able to work in the same industries as men?
GETTING MARRIED
Is a married woman not legally required to obey her husband?
Can a woman be head of household or head of family in the same way as a man?
Is there domestic violence legislation?
Can a woman obtain a judgment of divorce in the same way as a man?
Do women have the same rights to remarry as men?
HAVING CHILDREN
Is there paid leave of at least 14 weeks available to women?
Does the government pay 100% of maternity leave benefits, or parental leave benefits (where maternity
Is there paid paternity leave?
Is there paid parental leave?
Is dismissal of pregnant workers prohibited?
RUNNING A BUSINESS
Can a woman legally sign a contract in the same way as a man?
Can a woman legally register a business in the same way as a man?
Can a woman legally open a bank account in the same way as a man?
Does the law prohibit discrimination by creditors on the basis of sex or gender?
MANAGING ASSETS
Do men and married women have equal ownership rights to property?
Do sons and daughters have equal rights to inherit assets from their parents?
Do female and male surviving spouses have equal rights to inherit assets?
Does the law grant spouses equal administrative authority over assets during marriage?
Does the law provide for the valuation of nonmonetary contributions?
GETTING A PENSION
Are the ages at which men and women can retire with full pension benefits equal?
Are the ages at which men and women can retire with partial pension benefits equal?
Is the mandatory retirement age for men and women equal?
Does the law establish explicit pension care credits for periods of childcare?

Slika 2 Popis indikatora WBL indeksa [3]

Pitanja su oblikovana tako da je odgovor isključivo „da“ ili „ne“, stoga je njihova vrijednost tipa *boolean,* a svaka kategorija čini postotak potvrdno odgovorenih pitanja unutar nje same. Zbroj svih postotaka predstavlja WBL indeks.

Za potrebu predviđanja klase kreiran je dodatni stupac WBL_index_group. Dodatni stupac sadrži zapis klase koja je dobivena tako da je uzeta vrijednost polja WBL indeksa, određena u kojem se rasponu nalazi ta vrijednost i prema tome je određeno kojoj će klasi pripasti. Tako WBL grupa jedan sadrži sve zapise koji imaju WBL indeks manji od 25, WBL grupa dva čini zapise između 25 i 50 posto, WBL grupa tri čini zapise između 50 i 75 posto, dok WBL grupa ćetiri čini sve one iznad 75 posto. Podaci su

grupirani tako, kako bi predviđanje klase bilo djelotvornije s obzirom na to da sam WBL indeks ima brojne vrijednosti.

Faktor koji je također utjecajan za analizu je kategorija prihoda, koja je grupirana na četiri razine - niska razina prihoda, niža srednja razina, viša srednja razina te visoka razina. Kako bi analiza bila jednostavnija, razine su iz tekstualnog oblika oblikovane u brojčane vrijednosti prema razini od jedan do četiri.

```
WBL[WBL=='Low income']<-1
WBL[WBL=='Lower middle income']<-2
WBL[WBL=='Upper middle income']<-3
WBL[WBL=='High income']<-4
```

Slika 3 Grupacije prema prihodu

Iz istih je razloga svaka vrijednost „yes“ u tablici, korištena kao potvrdna oznaka, zamijenjena jedinicom, a „no“ nulom. Dodane su i donje crte umjesto razmaka te prvi stupac koji sadrži redni broj svakog retka za lakše snalaženje. U svrhu analize, izlučeni su podskupovi na temelju prve (1971.) i zadnje (2020.) godine provedenog istraživanja obilježenima u stupcu *reportyr*. Tako se vizualizacijom mogu uočiti promijene koje su se dogodile u navedenom periodu, što je cilj samog istraživanja - pratiti demografske promjene iz segmenta prava žena koja doprinose njenom poslovnom uspjehu.

```
#subsetovi za najraniji i najkasniji set podataka (prema godini)
WBL1971 <- WBL[ which(WBL$reportyr=='1971'), ]
WBL2020 <- WBL[ which(WBL$reportyr=='2020'), ]
```

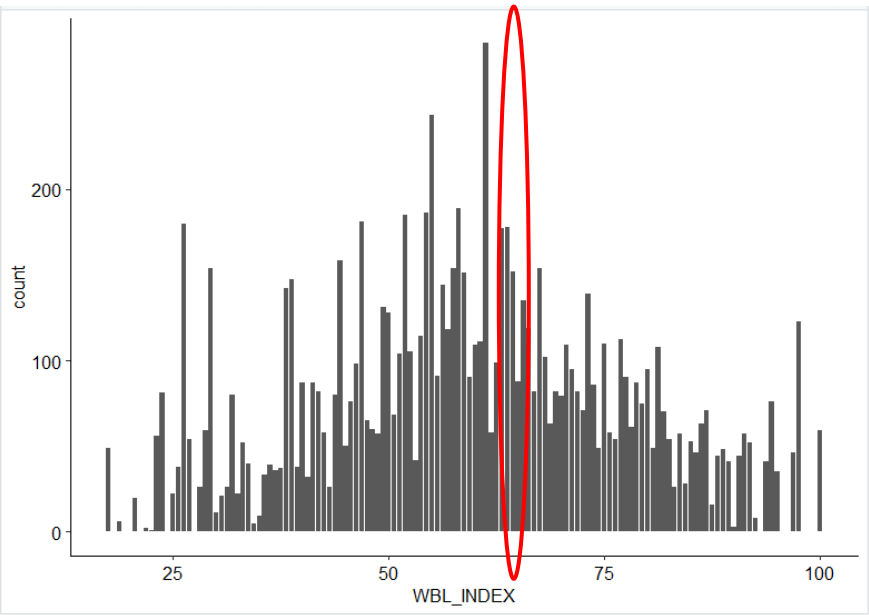
Slika 4 Stvaranje podsetova za 1971. i 2020. godinu

5.2 Vizualni prikaz seta podataka

Kako bi se izgradio dobar model ili čak odabrala metoda koja će se koristiti, važno je razumjeti podatke. Međutim, na setu podataka koji sadrži nekoliko tisuća zapisa, teško je znati kakvi se točno podaci nalaze u njemu i kako im pristupiti. Za bolji pregled podataka često se koriste vizualizacije. Programski jezik *R* sadrži brojne biblioteke koje se mogu koristiti za vizualizacije. U svrhu ovog rada korištena je biblioteka *ggplot2*.

Shmueli [1] iznosi brojne primjere dijagrama, a najjednostavniji od njih je stupčasti. Njime će biti prikazano stanje WBL indeksa za cijeli period, a potom i za prvu i

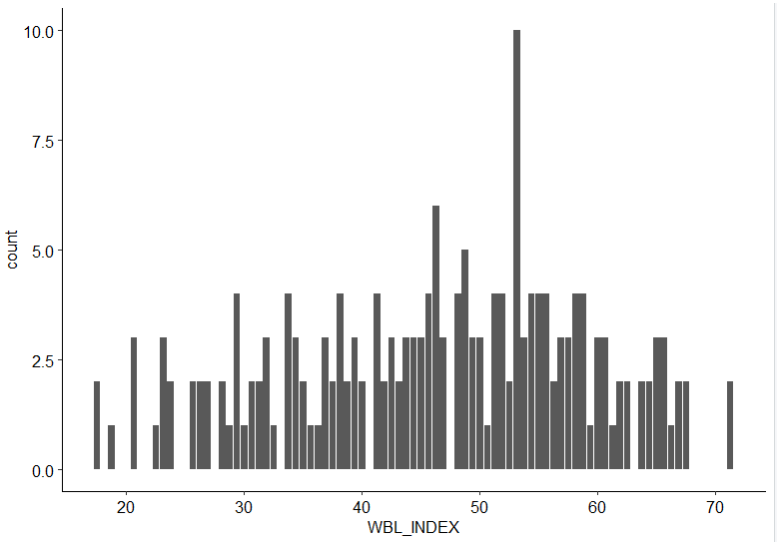
zadnju godinu istraživanja, kako bi se uočio napredak. WBL indeks je mjera koja čini ukupnu kalkulaciju svih indikatora (spomenuto u poglavlju 5.1) iz svih kategorija u obliku postotka. Na slici 5 vidi se zbroj vrijednosti prema vrijednosti WBL indeksa od 1971. do 2020. godine. Dakle, prema rezultatu se vidi da u navedenom periodu broj zemalja po visini WBL indeksa prema godini i državnoj ekonomiji raste do otprilike indeksa 60. Ono što se isto tako ovdje može zaključiti da je u tom periodu ukupno status žena u navedenih 190 zemlja bio najvećim dijelom oko srednje vrijednosti.Broj segmenata koji ima određeni WBL indeks varira, pa se vidi nagli skok na vrijednosti 25, koji uvelike pada u sljedećem koraku, međutim graf prati određeni trend. Tako se najveći broj zemalja nalazi u srednjem dijelu, što znači da ženski rod ima prosječna prava u najvećem broju zemalja prema godini, dok loše i odlično stanje, kako graf ide prema ekstremima, prikazuje znatno niže vrijednosti. Najveći broj čini označeni stupac, koji je blizu vrijednosti od 60, što znači da je velik broj svjetskih ekonomija prema prosjeku kroz 49 godina, u dobroj poziciji s pravima žena u društvu.



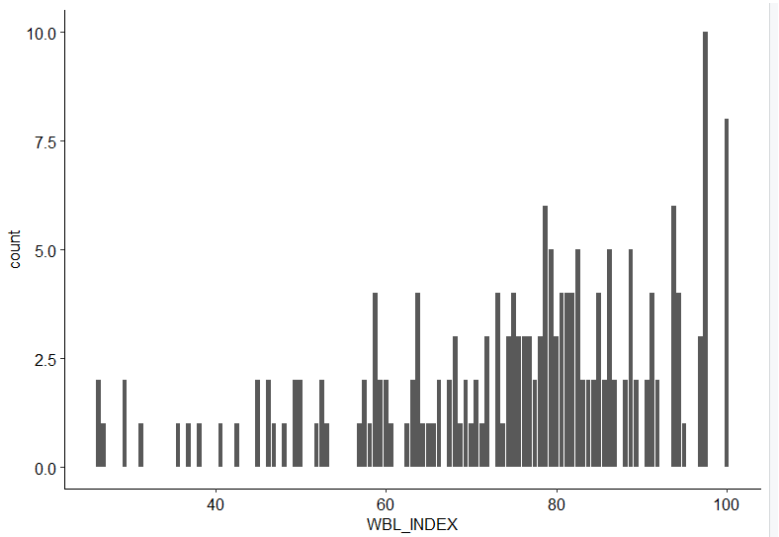
Slika 5 WBL indeks od 1971. do 2020. godine

Kako je očekivani trend da će se svakom sljedećom godinom poboljšati stanje žena u društvu i poslovnom svijetu, izveden je dijagram za 1971. i 2020. godinu kako bi se takva situacija provjerila. Na slikama 6 i 7 se vidi usporedba navedenih informacija. Tako se na slici 6 vidi da je za 1971. godinu zbroj država sa određenim WBL indeksom nema toliko drastične izmjene. Izuzetak je zbroj država obilježen s WBL indeksom nešto

iznad 50. Iz toga se može zaključiti da je 1971. godine najveći broj zemalja pružio ženama tek prosječan status. Vrijednosti za 2020. godinu na slici 7 su raspoređene većim dijelom na nešto veće vrijednosti, počevši od indeksa 80. U konačnici se može primijetiti i da su oni unosi s niskim WBL indeksom smanjeni, što znači da je WBL indeks porastao za određene ekonomske situacije i da se status žena zaista popravio tijekom perioda od 49 godina.



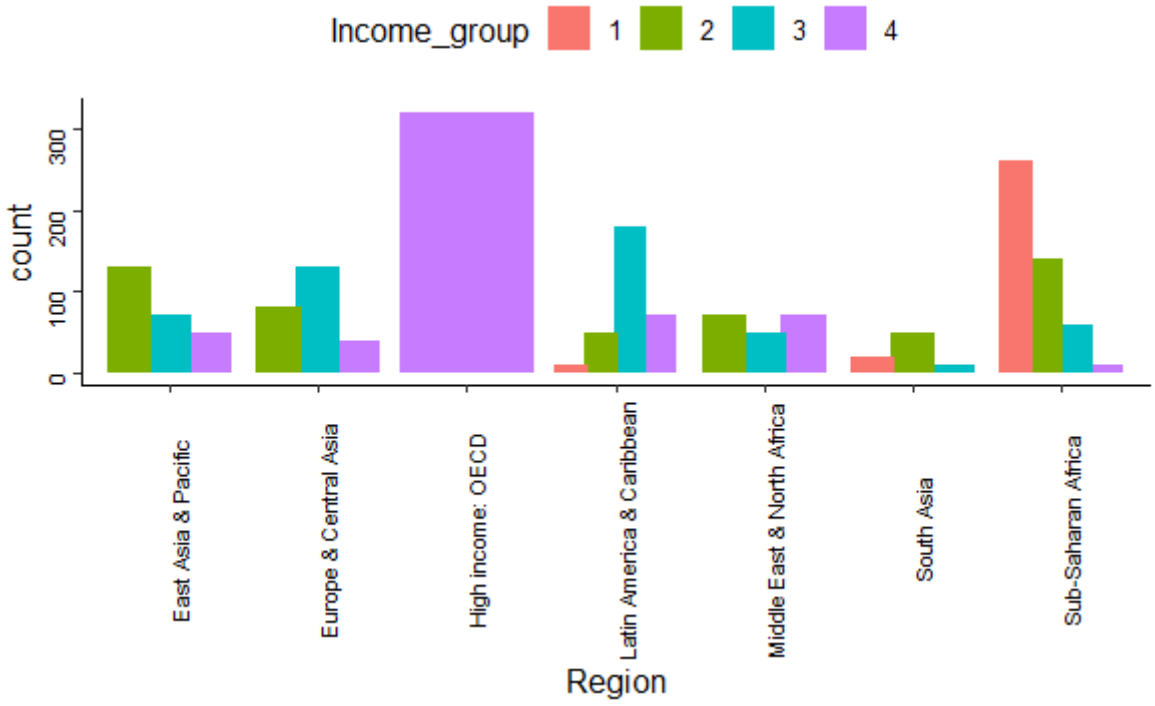
Slika 6 WBL indeks za 1971. godinu



Slika 7 WBL indeks za 2020. godinu

Osim WBL indeksa, za dobar pregled podataka uzet je u obzir i faktor prihoda koji je prethodno numeriran prema visini prihoda. Na slici 8 prikazan je stupčasti dijagram koji opisuje koliko država iz pojedine regije pripada kojoj grupaciji prihoda. Subsaharska Afrika je regija koja posjeduje daleko najviše država u grupi s najnižim

prihodom. Zanimljiva je situacija u Južnoj Americi, gdje je pripadnost trećoj skupini mnogo veća od prethodne i sljedeće. Znači da je mali broj niže srednje skupine i visoke, međutim broj skupine s višim srednjim prihodom je izrazito visoka u odnosu na njih.



Slika 8 Razina prihoda prema regiji

5.3 Sumiranje podataka za lakši pregled

Za velike količine podataka teško je izvući znanje iz podataka bez da ih se na neki način organizira. Shmueli [1] opisuje organizaciju podataka kroz agregaciju, računanje prosjeka ili sumiranje zbrajanjem. U slučaju ovog seta podataka izrađene su dvije agregacije - prosječna vrijednost WBL indeksa prema zbroju svih država prema regiji kojoj pripadaju, kao što je prikazano na slici 9, te prosječna vrijednost WBL indeksa prema sumi svake skupine prihoda na slici 10. Očekivan je ishod da će države OECD-a [5] imati najviši WBL indeks s obzirom na to da je njihova primarna motivacija pobrinuti se za balansiranu i kvalitetnu ekonomsku situaciju kako bi sama država imati bolju reputaciju.

Ponovnom agregacijom ove dvije agregacije može utvrditi korelacija među prikazanim grupama. Naime gore na slici 8 je prikazano npr. da Subsaharska Afrika nema

dobrostojeću ekonomsku situaciju te da velik broj država pripada u nižu srednju razinu prihoda. Kako razina prihoda jasno utječe na WBL indeks (slika 10), tako udruživanjem toga može se razumjeti i zašto Subsaharska Afrika ima najmanji postotak WBL indeksa na slici 9.

	Group.1	x
1	East Asia & Pacific	70.72640
2	Europe & Central Asia	84.70320
3	High income: OECD	93.53687
4	Latin America & Caribbean	79.09516
5	Middle East & North Africa	47.37105
6	South Asia	58.36250
7	Sub-Saharan Africa	69.63064

Slika 9 WBL index prema regiji

	Group.1	x
1	1	67.56828
2	2	68.74058
3	3	75.92760
4	4	82.85875

Slika 10 WBL indeks prema grupi prihoda (1- niska razina prihoda, 2- niža srednja razina, 3- viša srednja razina, 4- visoka razina)

6 Primjena skupnog učenja u R-u uz korištenje prosjeka rezultata dva modela

Ovo poglavlje prikazuje samu implementaciju skupnog učenja. Prvi dio poglavlja čini istraživanje veza među zapisima kako bi se ustanovilo koji model odabrati, dok u nastavku slijedi prikaz korištenja jednog modela s različitim parametrima kroz model naivnog Bayesa. U zadnjem dijelu poglavlja prikazano je korištenje dva različita modela: naivni Bayes i klasifikacijsko stablo, čija će se pojedinačna predviđanja ujediniti kako bi se postigli bolji rezultati.

6.1 Istraživanje veza među zapisima

Kako bi sama analiza bila potpuna, osim agregacije i grupiranja, važno je gledati skup podataka i kao skup zavisnih segmenata, a kako bi se njihova zavisnost odredila, mogu se izlučiti pravila na temelju kojih oni ovise međusobno. Iako je u prethodnom poglavlju predstavljen set podataka, dodatne provjera odnosa među poljima pridaje lakšem odabiru algoritma koji će se koristiti.

U ovom se setu podataka nalaze polja koja su direktno utjecala na primarni WBL indeks, a oni su binarnog tipa, stoga će se oni istražiti kako bi se upravo to i dokazalo. Drugi dio skupa je zanimljiviji i neizvjesniji jer se u opsnim poljima nalaze podaci o državi i regiji, godini istraživanja te financijskom stanju države za koje je dobiven primarni WBL indeks, međutim, ta polja nisu ulazila u direktnu kalkulaciju indeksa.

6.1.1 Istraživanje veza među binarnim poljima koji čine WBL indeks

U ovom skupu podataka bilo je prema pregledu seta podataka sigurno da svaki od segmenata koji ulazi u kalkulaciju primarnog WBL indeksa utječe na sami krajnji izračun. Naime, kao što je u poglavlju opisa seta objašnjeno, WBL indeks čini prosjek svih vrijednosti demografskih kategorija (mobilnost van države, mogućnost zaposlenja, stanje i status na radnom mjestu, ulogu u braku i njena prava u bračnoj zajednici, prava po pitanju djece, ženska administrativna prava i prava upravljanja financijama, dokumentacijom i poslovanjem te u konačnici prava u segmentu mirovine). Ono što na prvi pogled nije bilo sasvim jasno jest utjecaj navedenih kategorija međusobno jedne na drugu. Osim polja koje čine dio kalkulacije WBL indeksa, u setu se nalaze i pripadna opisna polja koja pobliže opisuju WBL indeks. U ovom su slučaju ta polja uklonjena iz seta, koji se pretvorio u matricu, kao što je prikazano na slici 11.

Ako se radi o skupu podataka koji sadrži 0 i 1 kao vrijednosti, kao u ovom slučaju, svaki zapis se može tretirati kao transakcija, koja se nadalje koristi za generiranje pravila. Zbog toga se set podataka pretvori u transakcijsku matricu (prikazano na slici 11). Prema tome se onda može vidjeti koji segmenti najčešće dolaze zajedno i kojim je pravilom došlo do tog zaključka. Taj princip koriste npr. velike web trgovine kako bi nudili kupcu proizvode koje će najvjerojatnije kupiti na temelju onoga što je do sada kupio.

Nakon što su podaci učitani u transakcijsku klasu unutar R-a, na nju se primjenjuje *apriori* metoda, pripadnica *"arules"* biblioteke, koja generira pravila zavisnosti. Odrediti se može *support* (hrv. potpora) i *confidence* (hrv. pouzdanost), gdje *support* označuje minimalni broj koji se želi uzeti u obzir kao zadanu popularnost tj. pojavljivanje neke kombinacije zavisnih elemenata, a računa se prema broju zapisa koje sadrže određenu vrijednost na prema ukupnom broju zapisa. S druge strane je parametar

confidence, koji se odnosi na vjerojatnost pojavljivanja jednog elementa E_1, ako se pojavi drugi element E_2. Za izračun navedenog, potrebno je pronaći sve retke tj. zapise gdje se E_1 i E_2 oboje pojavljuju, pa taj zbroj podijeliti s brojem pojavljivanja elementa E_1. Tako će se dobiti postotak u odnosu na koliko se element E_1 pojavio u skupu i koliko se element E_2 pojavio uz njega. Ako je potrebna navedena vjerojatnost za element E_2, obrnuti će se postupak primijeniti te će se tada provjeravati zbroj kombinacija elemenata E_1 i E_2 nad brojem pojavljivanja elementa E_2.

Kako bi se istražili spomenuti odnosi, parametri koji su korišteni jesu već spomenuti *support,* kojem je dodijeljena minimalna vrijednosti pojavljivanja u skupu od 20 posto i *confidence* kojem je dodijeljena minimalna vrijednost od 50 posto pojavljivanja elementa ako se pojavi drugi element.

```
###1. provjera kolumna s boolean vrijednostima

library("arules")

# uklanjanje ne boolean kolumne iz seta podataka
WBL_num_rules <-as.matrix(WBL[,  -c(1,2,3,4,5,6,7,8,9,14,19,24,30,32,35,37,38,39,41,46,52)])

#transformacija binarne matrice u transakcijsku bazu
fp.trans <- as(WBL_rules, "transactions")
inspect(fp.trans)
```

Slika 11Kreiranje transakcijske matrice podataka

Na slici 16 se vidi da je generirano 18 pravila uz unesene parametre, a na slici 17 dolje se mogu vidjeti prvih 6 od njih. Lhs označuje „ako ovi segmenti“, a rhs „onda ovi segmenti“. S desne se strane vide *support*, *confidence* i novi pojam *lift*, koji je indikator zavisnosti tj. ako je veći od nule, segmenti su zavisni, a ako je manji od nule, prvi segment ima negativan utjecaj na drugi segment.

```
> rules <- apriori(fp.trans, parameter = list(supp = 0.2, conf = 0.5, target = "rules"))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target   ext
        0.5    0.1    1 none FALSE            TRUE       5     0.2      1     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 2

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[6 item(s), 10 transaction(s)] done [0.00s].
sorting and recoding items ... [5 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 done [0.00s].
writing ... [18 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
>
```

Slika 12Prikaz reneriranih pravila

```
> inspect(head(sort(rules, by = "lift"), n = 6))
    lhs                                    rhs                                support confidence     lift count
[1] {apply.for.a.passport,
     legally.travel.outside.the.country} => {partial.pension.benefits.equal}         0.2        0.5 2.500000     2
[2] {partial.pension.benefits.equal}     => {apply.for.a.passport}                   0.2        1.0 1.666667     2
[3] {legally.travel.outside.the.country,
     partial.pension.benefits.equal}     => {apply.for.a.passport}                   0.2        1.0 1.666667     2
[4] {equal.ownership.rights.to.property} => {legally.travel.outside.the.country}     0.2        1.0 1.428571     2
[5] {partial.pension.benefits.equal}     => {legally.travel.outside.the.country}     0.2        1.0 1.428571     2
[6] {apply.for.a.passport,
     partial.pension.benefits.equal}     => {legally.travel.outside.the.country}     0.2        1.0 1.428571     2
>
```

Slika 13 Prvih 6 apriori pravila

Ono što se iz tih pravila može iščitati jest, da ne samo da o njima ovisi primarni indeks, već i oni međusobno ovise jedan o drugome. Ako pogledamo parove ili trojke, vidi se da su neki od njih u očitoj korelaciji, kao na primjer trojka broj [6] na slici 17. Vidi se da „prijava za putovnicu“ ima utjecaja na „legalno putovanje van države“. Kako bi osoba mogla legalno putovati van države, često je potrebna putovnica, pa ako je dopušteno ženskoj osobi prema ovom setu podataka, prijaviti se za putovnicu, očekuje se da će imati dozvolu je upotrijebiti. Ono što je s druge strane manje očekivano u istoj trojci broj [6] na slici 17 jest da osim segmenta „prijava za putovnicu“, uz njega ide i segment „jednake prednosti djelomične mirovine“ koji u kombinaciji utječu na segment „legalno putovanje van države“. Može se primjetiti da to nisu segmenti iste kategorije, no oni ipak često dolaze zajedno.

6.1.2 Istraživanje veza među binarnim poljima koji čine WBL indeks

S druge strane set podataka sadrži i kategoričke podatke, također opisane u prethodnom poglavlju broj 5. U te kategorije spadaju kategorija prihoda, koja je grupirana na četiri razine - niska razina prihoda, niža srednja razina, viša srednja razina te visoka razina, isto tako država i regija u kojoj se ta država nalazi i za koju vrijedi uneseni WBL indeks. Osim toga, upisana je i godina za koju se odnose navedeni podaci korišteni za izračunavanje WBL indeksa.

Za ove se podatke koristila jednostavna funkcija za izračun korelacije na temelju danog seta podataka. (vidi sliku 14).

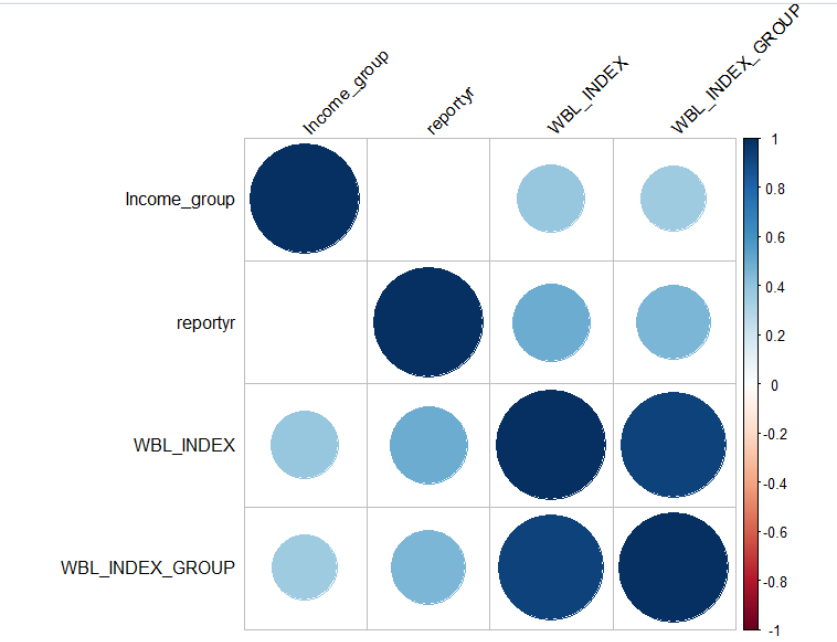
```
###2. provjera kolumna s opisnim/kategoričkim vrijednostima

#uzimanje samo polja kategoričkih vrijednosti
WBL_cat_rules <-WBL[,c(5,6,7,8)]

#provjera korelacije
corelation <-cor(WBL_cat_rules)
round(corelation, 2)
```

Slika 14 Provjera koorelacije između kategoričkih polja

Prema navedenim se podacima dobila korelacija među tim kategorijama, decimala se zaokružila na dva i izradila se matrica korelacije. Na slici 15 se vidi kako matrica ima vrijednost od sto posto za korelaciju kategorije same sa sobom. Isto tako se vide i ostale očekivane vrijednosti, kao što je visoka vrijednost korelacije WBL indeksa s WBL indeks grupom, s obzirom na to da je WBL indeks grupa izvedena iz samog indeksa. Zanimljivo je vidjeti i kako godina za koju je napravljen izvještaj pedeset posto utječe na WBL indeks, dok grupacija prema dohotku i regije za koju je izvještaj napravljen čak 12 posto manje, dakle 38 posto. Ista matrica je za bolji pregled, oblikovana u graf (slika 16) za koji se koristila *corrplot* biblioteka, namijenjena za prikaz korelacije. Na vizualizaciji se vidi kako su puni krugovi koji prikazuju korelaciju tamniji i veći za one kategorije s većim postotkom te manje i svjetlije s obzirom na manji postotak.

```
> round(corelation, 2)
                Income_group reportyr WBL_INDEX WBL_INDEX_GROUP
Income_group            1.00     0.00      0.38            0.36
reportyr                0.00     1.00      0.50            0.46
WBL_INDEX               0.38     0.50      1.00            0.92
WBL_INDEX_GROUP         0.36     0.46      0.92            1.00
```

Slika 15 Pregled koorelacije kategoričkih polja



Slika 16 Pregled koorelacije kategoričkih polja u grafu

6.2 Skupno učenje naivnim Bayesom s različitim parametrima

Skupno se učenje, kao što je u teorijskom dijelu objašnjeno, može koristiti kroz nekoliko pristupa. Ono čini kombinaciju rezultata u svrhu dobivanja čim bolje predikcije. Jedan od načina za postizanje istoga jest korištenje samo jednog modela, međutim kombiniranje njegovih rezultata dobivenih uz pomoć različitih parametara.

Za potrebu različite primjene skupnog učenja, ovdje će se primijeniti naivni Bayes. Kako točno model funkcionira je objašnjeno u teorijskom dijelu rada, a unatoč tome što je ova metoda jednostavna, često se koristi baš zbog pretpostavke teorema da su sve značajke neke klase nezavisne jedna od druge. Tako je u ovom slučaju u prethodnom poglavlju (6.1) objašnjena ovisnost i korelacija s jedne strane podataka koji su se koristili za sami izračun WBL indeksa, a s druge strane opisne faktore poput države i godine koji zapravo proizveli takvu situaciju i odredili parametre koji su doveli od određenog indeksa. S obzirom na to da je njihov utjecaj neizvjesniji, upravo će se ti podaci iz seta koristiti za klasifikaciju.

Biblioteka koja je potrebna za implementaciju naivnog Bayesa naziva se *e1071*. Ovaj model će se koristiti nekoliko načina. Nakon što se istražio set podataka, odlučeno je da će se u jednom slučaju koristiti regija zemlje iz koje potječe indeks kao glavni parametar, a zatim će se koristiti godina za koju taj indeks vrijedi kao glavni parametar. U oba slučaja, osim primarni parametar, korišten je i WBL indeks, grupacija WBL indeksa koja čini kategoriju u kojoj se određeni indeks nalazi prema određenom rasponu, a osim toga i financijsko stanje tj. financijska grupacija određene regije. Grupa WBL indeksa u bit će klasa koju model treba odrediti, a upravo su se zato WBL indeksi grupirali, kako model ne bi imao previše klasa s obzirom na broj različitih vrijednosti WBL indeksa.

6.2.1 Izrada modela naivnog Bayesa s regijom kao primarnim parametrom u setu

U prvom slučaju dodatni primarni parametar je regija. Uzima se cijeli set podataka i odvajaju se samo one kolumne koje su u ovom slučaju potrebne. Na slici dolje je prikazano (Slika 17) odvajanje polja, kao i seta indeksa za trening, pa zatim i kreiranje seta za treniranje i seta za validaciju. Ispod toga može se vidjeti da je kreiran Bayesov model koji predviđa WBL indeks grupaciju prema rasponu od jedan do četiri.

```
#naivni Bayes sa regijama
region_selected.var <- c(5, 6, 8, 9)
train.index <- sample(c(1:dim(WBL)[1]),dim(WBL)[1]*0.8)
region_train.df <- WBL[region_train.index, region_selected.var]
region_valid.df <- WBL[-train.index, region_selected.var]

region_bayes <- naiveBayes(WBL_INDEX_GROUP~., data=region_train.df)
```

Slika 17 Izrada modela naivnog Bayesa - regija

Nakon što je kreiran model, uz pomoć dobivenog trening seta, ppotrebno je odrediti koliko će model dobre predikcije producirati. Svi oni elementi koji su se nalazili u setu za treniranje, isključeni su iz seta za testiranje koji se koristi za validaciju modela. Cilj je postići dovoljno velik set za treniranje, jer što više podataka model dobije u setu za treniranje, prikupit će više iskustva o danim podacima i time ih moći primijeniti u koraku klasifikacije. U ovom slučaju je korišteno osamdeset posto seta za treniranje, dok je preostalih dvadeset posto korišteno za testiranje. Za pregled točnosti samog modela i njegovih rezultata koristi se konfuzijska matrica. Ona prikazuje stvarne vrijednosti i predviđene vrijednosti, ali isto tako i postotak točno predviđenih vrijednosti. Ovdje su to klase od jedan do četiri. Za Bayesov model predikcije za ovaj set podataka, dobiveno je 92 posto točnosti, kao što se vidi na slici dolje (Slika 18). Isto tako se na slici vidi matrica s vrijednostima WBL grupe (od jedan do četiri) koje su pogođene i količina krivo klasificirani grupacija WBL indeksa. Tako na primjer ako se gleda WBL indeks grupaciju jedan, koja sadrži sve zapise kojima je WBL indeks ispod 25, može se vidjeti da je u devedeset slučajeva točno predviđena WBL grupa, dok se može isto tako vidjeti da je za sedam zapisa krivo predviđena WBL grupa broj dva, dok je u pitanju bila WBL grupa 1. Najveći brojevi grešaka se javljaju kod srednjih WBL grupa – grupe dva i grupe tri. WBL grupa dva čini raspon od 25 do 50 WBL indeksa i ima 945 točno predviđenih zapisa, dok ona najveća, WBL grupa tri ima čak 1819 točno predviđenih zapisa, dok je nekoliko bilo krivo predviđeno kao WBL grupa dva ili WBL grupa četiri. Ono što potvrđuje da je model napravio dobro predviđanje jest da niti jedna od krivo predviđenih WBL grupa nije dobila suprotnu vrijednost, dakle nema WBL grupe jedan predviđene kao WBL grupa četiri ili obrnuto. Dodatne se informacije mogu vidjeti za svaku klasu posebno, desno na slici 18.

```
Confusion Matrix and Statistics

          1    2    3    4
     1   90   36    0    0
     2    7  945   82    0
     3    0   72 1819   56
     4    0    0   56  713

Overall Statistics

               Accuracy : 0.9203
                 95% CI : (0.9113, 0.9286)
    No Information Rate : 0.5049
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8742

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity          0.92784   0.8974   0.9295   0.9272
Specificity          0.99047   0.9685   0.9333   0.9820
Pos Pred Value       0.71429   0.9139   0.9343   0.9272
Neg Pred Value       0.99813   0.9620   0.9285   0.9820
Prevalence           0.02503   0.2717   0.5049   0.1984
Detection Rate       0.02322   0.2438   0.4693   0.1840
Detection Prevalence 0.03251   0.2668   0.5023   0.1984
Balanced Accuracy    0.95915   0.9330   0.9314   0.9546
```

Slika 18 Konfuzijska matrica za naivnog Bayesa - prema regiji

6.2.2 Izrada modela naivnog Bayesa s godinom kao primarnim parametrom u setu

Zatim, kako bi se dobila srednja vrijednost predviđanja istog modela s različitim parametrima, bilo je potrebno promijeniti parametre. U drugom slučaju izrade modela izmijenjeno je polje regije koje se koristilo u prošlom modelu, s poljem koje prikazuje godinu za koju vrijedi izračunati WBL indeks, pa tako i WBL grupa indeksa. Na slici 19 dolje se vidi da se ponovio isti postupak, samo su uvršteni drugačiji parametri u prvom retku. Važno je pritom primijetiti da je set za treniranje, kao i set za testiranje ostao isti, kako bi se jasno vidjela razlika u predviđanju s isključivo samo različitim ulaznim parametrima koji će se za predviđanje koristiti.

```
#naivni Bayes sa godinama
year_selected.var <- c(6, 7, 8, 9)

year_train.df <- WBL[year_train.index, year_selected.var]
year_valid.df <- WBL[-train.index, year_selected.var]

year_bayes <- naiveBayes(WBL_INDEX_GROUP~., data=year_train.df)
```

Slika 19 Izrada modela naivnog Bayesa - godina

Ono što je ključno provjeriti je konfuzijska matrica za nove parametre. Na slici 20 je prikazan rezultat predviđanja za drugi set parametara. Preciznost modela je čak bolja od prethodnog – veća od 95 posto. Matricu točno i krivo predviđenih vrijednosti se isto tako može pregledati za novi model, kao i statistiku prema klasi.

```
Confusion Matrix and Statistics

          1   2   3   4
  1  26   0   0   0
  2   7 556  11   0
  3   0  17 876  17
  4   0   0  34 394

Overall Statistics

               Accuracy : 0.9556
                 95% CI : (0.9455, 0.9644)
    No Information Rate : 0.4752
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9309

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity           0.78788   0.9703   0.9511   0.9586
Specificity           1.00000   0.9868   0.9666   0.9777
Pos Pred Value        1.00000   0.9686   0.9626   0.9206
Neg Pred Value        0.99634   0.9875   0.9562   0.9887
Prevalence            0.01703   0.2957   0.4752   0.2121
Detection Rate        0.01342   0.2869   0.4520   0.2033
Detection Prevalence  0.01342   0.2962   0.4696   0.2208
Balanced Accuracy     0.89394   0.9786   0.9589   0.9682
```

Slika 20 Konfuzijska matrica za naivnog Bayesa - prema godini

6.2.3 Primjena jednostavnog prosjeka na kombinaciju dva predviđanja

Sada se dolazi do dijela primjene skupnog učenja. Naime, nakon što su se predvidjele klase za oba seta parametara koristeći isti set za treniranje i validaciju s istim modelom za predviđanje, slijedi pokušaj stvaranja kombiniranog modela koji će imati bolje rezultate. Bez dodatnog treniranja uzeti su svi parametri koji su korišteni u prethodnim predviđanjima klasa i isti set za treniranje te se izračunao prosjek prethodna dva predviđanja WBL grupe indeksa. Korišten je pritom jednostavan prosjek dijeljenja na dva dijela i zaokruživanja broja na višu vrijednost s funkcijom *round* kao što se može vidjeti na slici 21. Naime, u većini slučajeva je decimala bila polovica tj. broj 5, što je dovelo do dileme zaokruživanja na nižu WBL grupu ili višu WBL grupu. Provjerena su oba slučaja (za dohvaćanje niže vrijednosti je korištena funkcija *floor*), međutim zaokruživanje na višu grupu pokazalo se kao djelotvornija solucija.

```
##kombiniranje dvije predikcije prosjekom
comb_selected.var <- c( 5, 6, 7, 8, 9)
comb_valid.df <- WBL[-train.index, comb_selected.var]

comb_pred<-round((region_num+year_num)/2)
```

Slika 21 Izrada kombiniranog modela primjenom prosjeka

Rezultati konfuzijske matrice skupnim učenjem bili su neočekivano lošiji od rezultata pojedinačnih modela (Slika 22).

```
Confusion Matrix and Statistics

          1    2    3    4
  1   29    8    0    0
  2   68  860  430   30
  3    0  168 1127  165
  4    0   17  400  574

Overall Statistics

               Accuracy : 0.6682
                 95% CI : (0.6531, 0.683)
    No Information Rate : 0.5049
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4985

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity          0.298969   0.8167   0.5759   0.7464
Specificity          0.997883   0.8130   0.8265   0.8658
Pos Pred Value       0.783784   0.6196   0.7719   0.5792
Neg Pred Value       0.982287   0.9224   0.6565   0.9324
Prevalence           0.025026   0.2717   0.5049   0.1984
Detection Rate       0.007482   0.2219   0.2908   0.1481
Detection Prevalence 0.009546   0.3581   0.3767   0.2557
Balanced Accuracy    0.648426   0.8148   0.7012   0.8061
```

Slika 22 Konfuzna matrica kombiniranog modela

Zbog lošijih rezultata ideja je bila ponovno treniranje modela s cijelim setom podataka i svim korištenim parametrima kao što je prikazano na slici 23.

```
##ponovno treniranje kombiniranjem oba subseta podataka
comb_selected.var <- c(5, 6, 7, 8, 9)

comb_train.df <- WBL[train.index, comb_selected.var]
comb_valid.df <- WBL[-train.index, comb_selected.var]

comb_bayes <- naiveBayes(WBL_INDEX_GROUP~., data=comb_train.df)

# validation
comb_pred.class <- predict(comb_bayes, newdata = comb_valid.df, type = "class")
confusionMatrix(table(as.factor(comb_pred.class), comb_valid.df$WBL_INDEX_GROUP))
```

Slika 23 Ponovno treniranje kombiniranog modela

Kombinirajući tako cijeli set podataka i sve korištene parametre dobiven je ponovno dosta dobar rezultat. Međutim ako se pogleda konfuzijska matrica na slici broj 24, jasno se vidi da je rezultat preciznosti između rezultata prvog i drugog modela. U ovom se slučaju skupno učenje nije pokazalo kao bolja solucija, s obzirom na to da nije dobivena veća preciznost modela niti povećanjem broja parametara. Najtočniji model je ipak onaj koji uzima samo godinu za koji vrijedi WBL indeks kao primarni parametar.

```
Confusion Matrix and Statistics

          1   2   3   4
  1  30  18   0   0
  2   3 513  35   0
  3   0  42 859  17
  4   0   0  27 394

Overall Statistics

               Accuracy : 0.9267
                 95% CI : (0.9142, 0.9379)
    No Information Rate : 0.4752
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8863

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity           0.90909   0.8953   0.9327   0.9586
Specificity           0.99055   0.9722   0.9420   0.9823
Pos Pred Value        0.62500   0.9310   0.9357   0.9359
Neg Pred Value        0.99841   0.9567   0.9392   0.9888
Prevalence            0.01703   0.2957   0.4752   0.2121
Detection Rate        0.01548   0.2647   0.4432   0.2033
Detection Prevalence  0.02477   0.2843   0.4737   0.2172
Balanced Accuracy     0.94982   0.9337   0.9373   0.9705
```

Slika 24 Konfuzna matrica ponovno treniranog modela

6.3 Skupno učenje kombiniranjem naivnog Bayesa i klasifikacijskog stabla

Druga vrsta skupnog učenja je redanje dva različita modela tj. njihovih rezultata jednog na drugog. Spajanjem dva različita modela u jedan, mogu se prikriti nedostaci jednog od njih. U ovome poglavlju će za to poslužiti već spomenuti naivni Bayes i klasifikacijsko stablo. Pritom će se koristiti isti set podataka i isti parametri za oba modela. Odabir parametara je u ovom poglavlju nešto drugačiji. Naime, oni čine kombinaciju opisnih podataka i *boolean* vrijednosti koje su doprinijele izračunu WBL indeksa, pa uključuju osim godine, regije i financijske kategorije države i informacije o radnom statusu žene, plaći i njenoj mogućnosti da putuje van države.

6.3.1 Izrada modela klasifikacijskog stabla

Ovaj je model zajedno s modelom naivnog Bayesa, objašnjen u teorijskom dijelu. No, u ovom dijelu će riječ biti o njegovoj primjeni. Za izradu klasifikacijskog stabla, korištena je biblioteka *rpart.* Već navedena je i biblioteka *caret,* zadužena za izradu dolje prikazanu konfuzijsku matricu (Slika 25). Set parametara koji se koristio u ovom dijelu je promijenjen zbog ovog modela. Koristeći samo set parametara iz prethodnog poglavlja, klasifikacijsko stablo je dalo lošije rezultate, dok je dodavanjem novih parametara, preciznost postala znatno bolja.

```
#stablo odluke
library(rpart)
library(caret)
tr <- rpart(WBL_INDEX_GROUP~ ., data = train.df)
tree_pred <- predict(tr, valid.df, type = "class")
confusionMatrix(table(tree_pred, valid.df$WBL_INDEX_GROUP))
```

Slika 25 Izrada klasifikacijskog stabla

Konfuzijska je matrica pokazala dobre rezultate za točnost procjene i predviđanja WBL grupe indeksa. Ono što je zanimljivo jest da stablo nije uspjelo točno predvidjeti niti jedan zapis u WBL grupi jedan, već je sve njih svrstalo u WBL grupu dva, kako se vidi na slici 26.

```
Confusion Matrix and Statistics

tree_pred    1    2    3    4
        1    0    0    0    0
        2   92  994  142    0
        3    0   99 1716  175
        4    0    0   36  622

Overall Statistics

               Accuracy : 0.8596
                 95% CI : (0.8483, 0.8704)
    No Information Rate : 0.4886
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7754

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity           0.00000   0.9094   0.9060   0.7804
Specificity           1.00000   0.9159   0.8618   0.9883
Pos Pred Value            NaN   0.8094   0.8623   0.9453
Neg Pred Value        0.97626   0.9626   0.9056   0.9456
Prevalence            0.02374   0.2820   0.4886   0.2056
Detection Rate        0.00000   0.2564   0.4427   0.1605
Detection Prevalence  0.00000   0.3168   0.5134   0.1698
Balanced Accuracy     0.50000   0.9127   0.8839   0.8844
```

Slika 26 Konfuzna matrica klasifikacijskog stabla

6.3.2 Izrada modela naivnog Bayesa

Naivni Bayesov model objašnjen je već u prethodnom poglavlju te je u ovom slučaju izrađen model na isti način. Jednako je tako korištena biblioteka *e1071,* a klasa koja se trebala predvidjeti je WBL grupa indeksa kao što se vidi na slici 27.

```r
#naivni Bayes
library(e1071)

bayes <- naiveBayes(WBL_INDEX_GROUP~., data=train.df)

# validation
library(caret)

bayes_pred <- predict(bayes, newdata = valid.df, type = "class")
confusionMatrix(table(bayes_pred, valid.df$WBL_INDEX_GROUP))
```

Slika 27 Izrada modela naivnog Bayesa (za kombiniranje sa stablom)

Zanimljivo je pritom, ako se pogleda konfuzijska matrica na slici 28 kako je model, koji je u ovom slučaju izgrađen s više parametara nego u prethodnom poglavlju (6.2), proizvedo manje precizno predviđanje. Ovdje se dobro vidi koliko je bitan izbor parametara i poznavanje seta podataka, ali i funkcioniranje samog modela, kako bi se izradio precizan model klasifikacije.

```
Confusion Matrix and Statistics

bayes_pred    1    2    3    4
         1   76   77    0    0
         2   22  898  131    0
         3    0  116 1641   94
         4    0    0  119  702

Overall Statistics

               Accuracy : 0.8558
                 95% CI : (0.8443, 0.8667)
    No Information Rate : 0.4879
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7768

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity           0.77551   0.8231   0.8678   0.8819
Specificity           0.97962   0.9451   0.8942   0.9614
Pos Pred Value        0.49673   0.8544   0.8865   0.8551
Neg Pred Value        0.99409   0.9317   0.8765   0.9692
Prevalence            0.02528   0.2815   0.4879   0.2054
Detection Rate        0.01961   0.2317   0.4234   0.1811
Detection Prevalence  0.03947   0.2712   0.4776   0.2118
Balanced Accuracy     0.87756   0.8841   0.8810   0.9216
```

Slika 28 Konfuzna matrica modela naivnog Bayesa

6.3.3 Kombiniranje modela klasifikacijskog stabla s naivnim Bayesom prosjekom

Kako se moglo primijetiti, klasifikacijsko stablo i naivni Bayes podržavaju različite vrste parametara. Tako je naivni Bayes dobio bolje predviđanje bez binarnih polja, iako je to činilo više ulaznih informacija za model, međutim klasifikacijsko stablo je zahtijevalo dodavanje istih kako bi predviđanje WBL grupe bilo uspješnije.

Kao što se vidi na slici 29, modeli su kombinirani bez dodatnog treniranja. Predviđanje oba modela je uključeno kako bi se izračunao njihov prosjek. Pritom je za zaokruživanje dobivene klase korištena funkcija *floor*, s obzirom na to da se moglo vidjeti da klasifikacijsko stablo nije proizvelo niti jedno točno predviđanje za WBL grupu jedan, pa bi u slučaju gdje je naivni Bayes predvidio WBL grupu jedan, dok je stablo predvidjelo WBL grupu dva, dobio se rezultat 3/2, što kad bi se zaokružilo na veći cijeli broj bilo dva, pa kombinirano predviđanje ne bi sadržalo niti jedno predviđanje za WBL grupu jedan.

```
#Predicting the probabilities
pred_bayes_prob<-predict(object = bayes, valid.df, type='class')
pred_tree_prob<-predict(object = tr, valid.df ,type='class')

pred_avg<-(as.numeric(pred_tree_prob)+as.numeric(pred_bayes_prob))/2

pred_avg <- as.factor(floor(pred_avg))

confusionMatrix(table(pred_avg, valid.df$WBL_INDEX_GROUP))
```

Slika 29 Izrada kombiniranog modela klasifikacijskog stabla i naivnog Bayesa primjenom prosjeka

Ujedinjenjem ova dva predviđanja dobiven je, kao što se vidi na slici 30, bolji rezultat nego što su imali modeli s istim setom podataka i parametara pojedinačno. Preciznost predviđanja je došlo do 92 posto, što iako ima još mjesta za napredak, nije loš rezultat. Time se vidjelo da je skupno učenje u ovom slučaju optimiziralo model i modeli su popravili nedostatke drugog.

```
Confusion Matrix and Statistics

pred_avg    1    2    3    4
       1   79   26    0    0
       2   14  962   88    0
       3    0   78 1810   31
       4    0    0   57  731

Overall Statistics

               Accuracy : 0.9241
                 95% CI : (0.9154, 0.9323)
    No Information Rate : 0.5044
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8804

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity           0.84946   0.9024   0.9258   0.9593
Specificity           0.99313   0.9637   0.9433   0.9817
Pos Pred Value        0.75238   0.9041   0.9432   0.9277
Neg Pred Value        0.99629   0.9630   0.9259   0.9900
Prevalence            0.02399   0.2750   0.5044   0.1966
Detection Rate        0.02038   0.2482   0.4670   0.1886
Detection Prevalence  0.02709   0.2745   0.4951   0.2033
Balanced Accuracy     0.92129   0.9331   0.9345   0.9705
```

Slika 30 Konfuzna matrica kombiniranog modela

7 Primjena bagging i boosting metode u R-u

Shmueli u literaturi [1] navodi razne postupke za strojno učenje, a u poglavlju broj 13 objašnjava kako ih kombinirano koristiti za još bolje rezultate. Takav složeni model Shmueli naziva „Super-model“ i objašnjava kako kroz vrednovanje klasifikacije odabrati najbolji model. Međutim ovo poglavlje opisuje implementaciju metoda koje se tek nadovezuju na jedan postojeći model, dok te metode nisu modeli koji se zasebno koriste. Navedene se metode nazivaju *bagging i boosting*, a one će se primijeniti nad modelom klasifikacijskog stabla. *Bagging* prikuplja nasumične uzorke i gradi rezultate na temelju algoritma koji pokreće za svakog od njih, dok s druge strane *boosting* nastoji razriješiti greške, pa je metoda usmjerena na one krivo predviđene uzorke, kako bi se baš oni riješili.

7.1 Implementacija *bagging* i *boosting* metode nad klasifikacijskim stablom

Same metode *bagging* i *boosting*, kao i model klasifikacijskog stabla opisane su u teorijskom dijelu rada. Ovdje će zato biti prikazana njihova implementacija nad jednostavnim stablom. Kako se već u prethodnom poglavlju (6) koristilo klasifikacijsko stablo, u ovom je poglavlju klasa koja će se predviđati zamijenjena. Naime prethodno je cilj bio izgraditi stablo koje će predvidjeti WBL grupu indeksa, dok će ovdje, kako bi se istražili novi odnosi, istraživati utjecaj WBL grupe na predviđanje financijske grupe

regije. Kao što se vidi na slici dolje (31), implementirano je jednostavno stablo i *bagging* i *boosting* modeli uz isti set podataka.

```
# single tree
tr <- rpart(Income_group ~ ., data = train.df)
pred <- predict(tr, valid.df, type = "class")
confusionMatrix(table(pred, valid.df$Income_group))

# bagging
bag <- bagging(Income_group ~ ., data = train.df)
pred <- predict(bag, valid.df, type = "class")
confusionMatrix(table(as.factor(pred), valid.df$Income_group))

# boosting
boost <- boosting(Income_group ~ ., data = train.df)
pred <- predict(boost, valid.df, type = "class")
confusionMatrix(as.factor(pred$class), valid.df$Income_group)
```

Slika 31 Izrada modela stabla, bagging i boosting

Kako bi se pratio napredak, on se može provjeriti s jednom od metoda za validaciju predviđanja, kako je gore navedeno. Na slici 32 dolje se nalazi konfuzijska matrica koja prikazuje statistiku za model stabla. Kao željeni ishod odabran je *Income_group* tj. financijska grupa regije s obzirom na to da je već podijeljen u kategorije, a prethodna analiza je utvrdila i da je u korelaciji s WBL indeksom i WBL grupom indeksa, stoga je dobar kandidat za navedeno. Točnost ovog modela je 77,8% kao što se vidi na slici 32.

```
Confusion Matrix and Statistics


pred  1  2  3  4
   1 45 13  1  0
   2  5 69  9  1
   3  5 21 91 23
   4  0  5  0 86

Overall Statistics

               Accuracy : 0.7781
                 95% CI : (0.7325, 0.8192)
    No Information Rate : 0.2941
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.6998

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity            0.8182   0.6389   0.9010   0.7818
Specificity            0.9561   0.9436   0.8205   0.9811
Pos Pred Value         0.7627   0.8214   0.6500   0.9451
Neg Pred Value         0.9683   0.8655   0.9573   0.9152
Prevalence             0.1471   0.2888   0.2701   0.2941
Detection Rate         0.1203   0.1845   0.2433   0.2299
Detection Prevalence   0.1578   0.2246   0.3743   0.2433
Balanced Accuracy      0.8871   0.7912   0.8608   0.8814
```

Slika 32 Konfuzijska matrica za stablo

Točnost modela već sada nije loša, međutim ona se može poboljšati upravo metodom skupnog učenja. Na slikama 33 i 34 vidi se rezultat nakon dodanih metoda - najprije *bagging*, pa onda i *boosting*. Primjenom *bagginga*, na slici 14, vidi se velik napredak u točnosti modela, a on je postao još bolji nakon primjene *boostinga*, što je dovelo do zadovoljavajućeg rezultata od 98,6% točnosti. Za ovu se kombinaciju može reći da je postigla najbolje rezultate od svih modela skupnog učenja u ovome radu.

```
Confusion Matrix and Statistics


     1   2   3   4
  1 57   2   0   0
  2  0  89   0   0
  3  0   3 108   2
  4  0   1   4 108

Overall Statistics

               Accuracy : 0.9679
                 95% CI : (0.9446, 0.9833)
    No Information Rate : 0.2995
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9564

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity            1.0000   0.9368   0.9643   0.9818
Specificity            0.9937   1.0000   0.9809   0.9811
Pos Pred Value         0.9661   1.0000   0.9558   0.9558
Neg Pred Value         1.0000   0.9789   0.9847   0.9923
Prevalence             0.1524   0.2540   0.2995   0.2941
Detection Rate         0.1524   0.2380   0.2888   0.2888
Detection Prevalence   0.1578   0.2380   0.3021   0.3021
Balanced Accuracy      0.9968   0.9684   0.9726   0.9814
```

Slika 33 Konfuzijska matrica za bagging

```
Confusion Matrix and Statistics


     1   2   3   4
  1 55   0   0   0
  2  0 107   0   0
  3  0   1 101   4
  4  0   0   0 106

Overall Statistics

               Accuracy : 0.9866
                 95% CI : (0.9691, 0.9956)
    No Information Rate : 0.2941
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.9818

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity            1.0000   0.9907   1.0000   0.9636
Specificity            1.0000   1.0000   0.9817   1.0000
Pos Pred Value         1.0000   1.0000   0.9528   1.0000
Neg Pred Value         1.0000   0.9963   1.0000   0.9851
Prevalence             0.1471   0.2888   0.2701   0.2941
Detection Rate         0.1471   0.2861   0.2701   0.2834
Detection Prevalence   0.1471   0.2861   0.2834   0.2834
Balanced Accuracy      1.0000   0.9954   0.9908   0.9818
```

Slika 34 Konfuzijska matrica za boosting

8 Zaključak

Sama ideja skupnog učenja seže iz davne prošlosti, što se vidi kroz priču statističara iz 19. stoljeća o natjecanju u pogađanju težine vola. Samostalne predikcije na natječaju bile su u vrlo širokom rasponu i varirale u većim razmacima, ali srednja vrijednost svih procjena bila je iznenađujuće točna – unutar jedan posto stvarne težine vola. Tako su se greške iz višestrukih nagađanja međusobno poništavale. U prosjeku zajedno višestruka nagađanja dat će precizniji odgovor od velike većine pojedinačnih nagađanja. Iako zvuči poprilično banalno, ta se ideja preslikava i skupno učenje u strojnom učenju. Sukladno tome dokazan je takav pristup na klasifikaciji nad danim podacima u ovom radu. Model naivnog Bayesa je dao dobre rezultate, kao i model klasifikacijskog stabla, međutim kombiniranjem njihovih predviđenih klasifikacija, postigao se model koji je proizveo još bolje rezultate. Ono što se moglo primijetiti, da unatoč dosta dobroj klasifikaciji stabla, ono nije klasificiralo niti jedan zapis unutar prve klase (WBL grupa jedan). Ta se greška riješila kombiniranjem klasifikacije naivnog Bayesa, kako bi i WBL grupa jedan bila zastupljena u setu. Time su se za one zapise, koji su se nalazili negdje između dviju klasa, smanjio i kako je vidljivo na konfuznoj matrici, sveukupna preciznost modela popravila. Zanimljivo bi bilo istražiti kakvi bi se rezultati dobili korištenjem težinskog prosjeka za kombiniranje ova dva modela. Veća bi se težina mogla pripisati naivnom Bayesu s obzirom na njegove rezultate, gdje bi klasifikacijsko stablo bilo tek mala pomoć za postizanje boljeg modela. Isto tako bi bilo interesantno vidjeti kako bi treći model utjecao na rezultat ova dva modela. Naime, skupno učenje ne mora se sastojati samo od dva modela. Dozvoljen je *n* broj modela, dok god njihovo redanje doprinosti postizanju boljih rezultata.

S druge strane anegdota isto tako prikazuje da je u toj situaciji bilo nekoliko (sretnih) pojedinaca koji su postigli bolji rezultat od prosjeka. Procjena skupnog učenja stoga neće uvijek biti preciznija od svih pojedinačnih procjena u svim slučajevima. Ta je izjava dokazana pri implementaciji kombinirane metode, gdje se koristio isti model (naivni Bayes) s različitim parametrima. Prema rezultatima se vidi da je godina izračuna faktor koji više doprinosi predviđanju klase, nego regija za koju rezultat vrijedi. Čak i korištenjem oba parametra zajedno, predviđanje nije točnije nego ono gdje se koristila samo godina kao primarni parametar.

Najbolje je rezultate imalo klasifikacijsko stablo s primijenjenom metodom *boostinga,* koja, prisjetimo se, izravno se trudi poboljšati segmente podataka u kojima određeni model griješi. Boostingom su tako riješena nedostajuća predviđanja klase jedan, pa je već dobar model klasifikacijskog stabla, postao još bolji i dostigao preciznost predviđanja od preko 98 posto.

Skupno učenje se počinje sve više primjenjivati, a očekivan je rastući trend s obzirom na to da sve mogućnosti koje dolaze s tom metodom i brojne načine koji su predloženi za optimizaciju modela, a kojih će u budućnosti biti samo više.

9 Izvori

[1] Shmueli, Galit; Bruce, Peter C.; Patel, Nitin R.(2018.): Data Mining for Business Analytics: Concepts, Techniques, and Applications in R; Wiley: SAD

[2] Kuncheva, Ludmila I.(2004): Combining Pattern Classifiers Methods and Algorithms; Wiley: SAD

[3] The world bank(veljača, 2021.): Women, Business and the Law; https://wbl.worldbank.org/ (3.7.2021.)

[4] thebalance.com; Amadeo, Kimberly (8.4.2019.): The OECD and How It Can Help You; https://www.thebalance.com/organization-economic-cooperation-development-3305871 (13.6.2021.)

10 Prilozi tablica

11 Grafički prilozi

12 Prilozi jednadžbi