UNIVERZITET
METROPOLITAN
BEOGRAD

Univerzitet Metropolitan Beograd

FEFA

Klod Kolaro

Klasifikacioni algoritmi mašinskog učenja i njihova primena u finansijama

Doktorska disertacija

Beograd, 2024.

University Metropolitan Belgrade

FEFA

Klod Kolaro

Clasiffication machine learning algorithms and their aplication in finance

PhD Dissertation

Belgrade, 2024.

Podaci o mentoru i članovima komisije za odbranu doktorske disertacije

Mentor: Prof. dr Milan Nedeljković, redovni profesor, Univerzitet Metropolitan Beograd, FEFA Fakultet

Članovi komisije:

Prof. dr Goran Pitić, redovni profesor, Univerzitet Metropolitan Beograd, FEFA Fakultet, predsednik komisije

Prof. dr Milan Nedeljković, redovni profesor, Univerzitet Metropolitan Beograd, FEFA Fakultet

Prof. dr Branko Urošević, redovni profesor, Univerzitet Union, Računarski fakultet

Datum odbrane:

Klasifikacioni algoritmi mašinskog učenja i njihova primena u finansijama

Rezime

Prethodnu deceniju globalno poslovno okruženje karakteriše radikalan zaokret ka ubrzanoj digitalizaciji u različitim segmentima poslovanja. Sa ubrzanom digitalizacijom poslovanja došlo je do eksponencijalnog rasta količine elektronskih podataka koji postaju nova ključna vrednost. Britanski matematičar Clive Humby je još 2016 izjavio da su podaci novo gorivo (eng. *'Data is the new oil'),* čemu smo svedoci naročito u periodu nakon pandemije. Rast obima digitalnih podataka prati i razvoj tehnologije i infrastrukture, koja omogućava jednostavan pristup, prenos, transformaciju i analizu podataka različitih struktura. Time su stvoreni uslovi za široku primenu mašinskog učenja (u nastavku: *ML*) u svih segmentima poslovanja. Varian (2014) u svom preglednom radu početkom prethodne decenije opisao je nove mogućnosti primene mašinskog učenja u finansijama, u analizi nestrukturiranih podataka velikog obima i brzine nastajanja, tzv. Velikih podataka. Mullainathan i Spiess (2017) u svom seminalnom radu takođe konstatuju sve veći potencijal, ali i primenu predikcionih algoritama mašinskog učenja na problemima iz finansija i ekonomije.

Predikcija stečaja i procena kreditnih rizika predstavljaju jedan od najvećih empirijskih izazova u finansijama. Tradicionalne statističke metode, kao što su multivarijantna diskriminantna analiza i linearna-logistička regresija, dominantno su se godinama primenjivane u rešavanju ovih problema. Kako su osnovni nedostaci ovih pristupa, rigidne i često neodržive statističke pretpostavke, potencijal za primenu modela mašinskog učenja je izrazit što je dovelo do toga da je, na primer iz oblasti primene *ML* u finansijama u 2021. godini objavljeno 11 puta više stručnih radova nego od proseka u periodu 2000-2017 (Hoang i Wiegratz, 2022).

U literaturi, međutim, i delaje ne postoji konsenzus o optimalnim tipu mašinskog učenja koji se može primeniti na najčešće probleme u empirijskim finansijama. Pored toga, radovi na temu uporedne analize performansi *ML* klasifikatora najčešće minimalno opisuju primenjene algoritme. U većini slučajeva daju se konačni oblici funkcija kojima se modeli opisuju, a ne teorijski prikaz kako se do njih došlo, kao i prikaz primenjenih optimizacijskih metoda. Na ovaj način tumačenje rezultata je otežano. Iz tog razloga, postoji potreba za sveobuhvatnim, jedinstvenim teorijskim prikazom klasifikacionih *ML* modela, zajedno sa sprovedenim empirijskim istraživanjima. Istovremeno, sumiranjem do sad objavljenih komparativnih empirijskih analiza, mogli bi se preporučiti najoptimalniji algoritmi, čime bi se značajno povećala efikasnost primene mašinskog učenja na problemima predikcije stečaja i procene kreditnih rizika.

Predmet istraživanja disertacije je komparativna empirijska analiza najčešće primenjivanih klasifikacionih algoritama mašinskog učenja u finansijama. U radu se daje detaljan teorijski prikaz modela mašinskog učenja, odgovarajućih optimizacijskih metoda, kao i relevantne literature.

Opšti doprinos ove disertacije je u sistematizaciji i objedinjavanju literature iz oblasti primene mašinskog učenja u finansijama i sveobuhvatnom teorijskom prikazu klasifikacionih algoritama, po prvi put u literaturi.

Dodatan doprinos predstavlja rezultat empirijskog istraživanja, kojim je analiziran uticaj neuravnoteženih podataka (minimalno prisustvo uzoraka jedne klase) na performanse *ML* modela i kako metoda preteranog uzorkovanja manjinske klase (eng. *oversamling)* utiče na kvalitet prediktivnog modela.

U studiji slučaja, prediktivni klasifikacioni *ML* algoritmi primenjeni su u cilju procene kreditnih rizika (eng. *credit scoring*) i verovatnoće stečaja kompanije. Rešavanjem ovih binarno-klasifikacionih problema na dva nezavisna seta podataka pokazan je uticaj koji mašinsko učenje ima na tačnost predikcije i bolje razumevanje podataka.

U radu su analizirani inferencijalni klasifikacioni algoritmi koji, pored predikcije, imaju funkciju da opišu međusobnu zavisnost i uticaj koji varijable (atributi) modela imaju na rezultat predikcije. Tumačenjem ovakvih modela postiže se bolje razumevanje date pojave ili događaja. Pored inferencijalnih, istraživanjem su obuhvaćeni i tzv. *black box* algoritmi, čija je osnovna namena postizanje maksimalne tačnosti predikcije, dok zbog njihove kompleksnosti, ove prediktivne modele nije moguće tumačiti.

Klasifikacioni algoritmi *ML*, koji su predmet ove disertacije, spadaju u grupu diskriminativnih modela, kojima se na direktan način dolazi do uslovne raspodele verovatnoće zavisne promenljive, u funkciji nezavisnih promenljivih (prediktora). Funkcija koja opisuje prediktivni model određuje položaj granične hiperravni, koja na najbolji način razdvaja uzorke različitih klasa (kategorija). Pored diskriminativnih, neki od analiziranih algoritama spadaju u grupu generativnih modela, kod kojih se na indirektan način primenom *Bajesove* teoreme, na osnovu zajedničke distribucije verovatnoće promenljivih različitih klasa, dolazi do predikcije klase kojoj ispitivani uzorak pripada.

Radom su obuhvaćene osnovne optimizacijske metode (Breeden, 2020), kojima se određuju parametri modela odnosno ciljne funkcije. Pri određivanju optimalnih vrednosti *hyperparametara* primenjen je *cross validation* postupak, sa brojem podskupova (eng. *folds*) 5, što je prema Abdou (2011) najčešće primenjivana vrednost u ekonomskim istraživanjima.

U prvom poglavlju predstavljen je pojam digitalne transformacije – sveobuhvatne promene u poslovanju inspirisane novim tehnologijama, koja je neophodna kako bi kompanije očuvale svoju konkurentnost. Ova strateška odluka dovodi do značajnih promena u kompanijskom lancu vrednosti kojima se kreiraju novi poslovni modeli, povećava operativna efikasnost, a razvojem novih personalizovanih digitalnih oblika komunikacije kreira superiorno korisničko iskustvo.

Digitalni podaci i napredna analitika kojom se podaci pretvaraju u korisne informacije i znanje ključni su akceleratori i pokretači ovih promena[1]. Mogućnost jednostavnog pristupa, efikasne obrade podataka i analize primenom naprednih softverskih alata, učinili su podatke ključnim asetom kompanije. Nestrukturirane podatke, koji najviše doprinose njihovom eksponencijalnom rastu (količina digitalnih podataka se u proseku duplira na dve godine[2]), nije moguće analizirati klasičnim analitičkim alatima. Mašinsko učenje, kao podgrupa veštačke inteligencije, primenjuje se u analizi podataka svih struktura, dok efikasnost *ML* modela raste sa količinom podataka koji su predmet analize.

Drugo poglavlje daje detaljan prikaz primene *ML* u finansijama i relevantne literature. Porter i Gujarati (2009) navode da se u ekonomskim empirijskim istraživanjima, zbog velike mogućnosti interpretabilnosti, najčešće primenjuju modeli linearne i logističke regresije, kojima se objašnjava

[1] Izvor: Gartner
[2] Izvor: Gantz, J. i Reinsel, D. IDC (2012)

uticaj nezavisnih ekonomskih varijabli (x) na vrednost zavisne promenljive (y). Ova dva modela su detaljno opisana, kao i odgovarajuće metode optimizacije i regularizacije. U disertaciji navodimo i osnovne linearane statističke modele za predikciju stečaja kompanije, koji se primenjuju u ekonometrijskim analizama: *Altman Z-score* (Edward Altman, 1968), Zmijewski model (1984) i Ohlson (1980). Rezultati empirijskog istraživanja će, između ostalog, pokazati u kojoj meri se primenom *ML* može povećati tačnost predikcije u odnosu na ove klasične modele.

U poglavlju tri detaljno su analizirani najčešće primenjivani klasifikacioni algoritmi mašinskog učenja. Pokazaćemo da logistička i linearna regresija spadaju u porodicu Opštih linearnih modela – (eng. *Generalized linear models*, *GLM*, Nelder i Wedderburn, 1972), kao i da kada raspored verovatnoće zavisne promenljive (y) pripada familiji eksponencijalnih distribucija, odgovarajućom transformacijom, y postaje linearna funkcija prediktora. Analizirani su takođe i klasifikacioni modeli na bazi stabla odlučivanja (eng. *Decision tree*), kao i homogeni *ensembling* modeli *Random Forest* i *XGBoost.* Predmet istraživanja su i tzv. *distance based* neparametrički modeli, koji grupišu uzorke različitih klasa prema njihovoj sličnosti, koja se meri njihovom međusobnom razdaljinom: *K-NN* (k najbližih suseda) i *Support Vector Machine*. Iako spada u jednostavnije algoritme za primenu, empirijskom analizom obuhvaćen je generativni model *Naive Bayes*, koji u slučajevima nepostojanja multikolinearnosti prediktora postiže zadovoljavajuće rezultate.

Za empirijsku evaluaciju modela (poglavlje četiri) primenjena je Matrica konfuzije (eng. *Confusen matrix*), koja daje sumarni prikaz performansi klasifikacionih algoritama. Pored ukupne tačnosti (eng. *Accuracy*), Senzitivnosti (eng. *Sensitivity*) i Specifičnosti (eng. *Specificity*) svakog modela, usled izrazite nebalansiranosti podataka kao mera performansi klasifikatora korišćeni su i: F1 *score* – koji je funkcija Preciznosti modela (eng. *Precission)* i njegove Senzitivnosti i *Cohen's kappa* indeks – koji 'nagrađuje' klasifikator samo kada je tačnost predikcije veća od tačnosti dobijene kada bi se uvek predviđala (birala) većinska klasa. Uz pomenute indikatore, kao mera performansi klasifikacionih modela primenjena je i *Auc* – površina obuhvaćena *Roc* krivom, koja je pokazatelj sposobnosti modela da razdvoji uzorke različitih klasa (eng. *discriminatory ability*), bez obzira na usvojenu graničnu vrednost verovatnoće.

Primenjene metode procesiranja i transformacije izvornih podataka opisane su u poglavlju pet. Kako su podaci iz studije slučaja nekompletni, pristupilo se dijagnozi i imputaciji nepostojećih

vrednosti, zameni ekstremnih vrednosti odgovarajućom medijanom, kao i skaliranju podataka kako bi se sveli na isti rang vrednosti. Usled dominantnog prisustva uzoraka jedne klase (eng. *majority class*) u odnosu na manjinsku, primenom *SMOTE (*eng. *Synthetic minority over-sampling technique)* (Chawla et al., 2002) *over – sampling* tehnike generisani su novi uzorci manjinske klase na bazi sličnosti sa postojećim uzorcima (*K-NN* algoritam).

Zbog postojanja velikog broja varijabli (atributa) i njihove međuzavisnosti, opisan je postupak Analize principalnih komponenti – *PCA* (eng. *Principal Compnent analysis*), kojim je moguće smanjiti dimenzionalnost podataka i eiliminisati pojavu multikolinearnosti. Model logističke regresije razvijen je i na podacima transformisanim u principalne komponente.

Prikaz empirijskih rezultata dat je u poglavlju šest. U kontekstu procene kreditnog rizika i verovatnoće stečaja, pokazano je da se ne može govoriti o jednom univerzalnom *ML* modelu, već da je izbor optimalnog modela funkcija: veličine i uravnoteženosti podataka, prisustva ekstremnih vrednosti, broja prediktora, oblika distribucije podataka i njihove međusobne zavisnosti. Homogeni *ensembling* modeli na bazi stabla (*Random Forest i XGBoost*) pokazali su se kao najstabilniji, sa predikcijama najveće tačnosti, kako na uravnoteženim tako i na neuravnoteženim podacima. Metodom *oversamplinga* može se povećati senzitivnost modela (vrednost *TP*). Primenom *ML*, utiče se na povećanje tačnosti predikcije u odnosnu na klasične linerane aditivne modele (Altman, 1968 i Zmijewski, 1984). Poslednje poglavlja daje zaključak rada i i smernice za nastavak istraživanja i budući rad.

Struktura podataka data je u prilogu A.1. U A.2 prikazani su rezultati empirijske analize i u A.3 je dat delimičan prikaz razvijenog softverskog algoritma, primenom *R* statističkog softvera, dok je kompletan kod dostupan na kkolaro/doktorat (github.com).

Ključne reči: mašinsko učenje, klasifikacija, predikcija, finansije, kreditni rizik

SADRŽAJ

1. DIGITALNA TRANSFORMACIJA

Poslovanje kompanija u novom dinamičnom digitalnom svetu karakteriše visok stepen konkurencije, koja nudi visokopersonalizovane, inovativne proizvode i usluge, nastale primenom novih digitalnih tehnologija (Atkinson, 2005). Nove tehnologije kreiraju nove poslovne mogućnosti, ali u isto vreme predstavljaju opasnost za kompanije koje na vreme ne uoče njihov transformacijski potencijal. Nemogućnost sagledavanja uticaja eksponencijalnog razvoja digitalnih tehnologija za kompanije nastale pre početka komercijalizacije interneta (zvaćemo ih tradicionalne) znači zadržavanje *status quo*-a. Nastavak poslovanja koji je baziran na strateškim asetima i znanjima, koja u digitalnom svetu više ne mogu biti osnov konkurentske prednosti, može dovesti do njihovog nestanka sa tržišta (eng. *Digital Darwinism)* (Roger, 2016). Tradicionalne kompanije tako mogu postati taoci vlastitog uspeha, planirajući budućnost samo na osnovu prethodnog iskustva (eng. *competence trap*) (Roger, 2016)**.**

Kada u poslovnom smislu govorimo o disruptivnim promenama na tržištu prouzrokovanim digitalnim tehnologijama, potrebno je napraviti razliku između inovacije i disrupcije. Pod disruptivnom promenom smatra se promena u postojećoj industriji, gde konkurencija (eng. *challanger*) kreira znatno veću vrednost za kupca, na jedinstven način, koju tradicionalna kompanija direktno ne može ponuditi (Roger, 2016). Dok se inovacija može relativno lako iskopirati i primeniti, to nije slučaj sa disrupcijom. Kako je disrupcija po prirodi asimetrična, odnosno ne dolazi iz industrije u kojoj tradicionalna kompanija posluje, postoje značajne razlike u lancu vrednosti (eng. *value chain*) i strukturi troškova između *challanger*-a i tradicionalne kompanije. Iz tog razloga se disruptivna promena ne može jednostavno primeniti, već zahteva transformaciju kompanije.

U zavisnosti od prioriteta i prirode konkurencije, razlikujemo tri pristupa digitalnoj transformaciji:

- Eksterni, gde kompanija primenjuje nove tehnologije kao bi unapredila korisničko iskustvo kroz sve raspoložive digitalne kanale komunikacije. Fokus je pre svega na dizajniranju novog načina interakcije i kolaboracije sa klijentima, koji je baziran na

personalizovanoj ponudi proizvoda i servisa, kao i digitalnih sadržaja koji su u skladu sa potrebama, interesima i navikama kupca.

- Interni, gde je cilj povećanje efikasnosti proizvodnih operacija, poslovnog odlučivanja i organizacione strukture. Primenom robotike, veštačke inteligencije, interneta stvari (eng. *IoT*), omogućava se *m2m* (eng. *machine to machine*) komunikacija i kreiranje sajber fizičkih sistema, kojima se postiže visok stepen automatizacije i povećanje operativne efikasnosti. Velika raspoloživost digitalnih podataka, uz napredak *ICT* infrastrukture i softverskih analitičkih platformi, omogućava pametno upravljanje proizvodnjom, kao i donošenje poslovnih odluka na bazi preporuka prediktivnih i preskriptivnih (eng. *prescribing*) algoritama.

- Holistički, sveobuhvatna transformacija poslovanja, kojom se utiče na sve segmente i funkcije, te se tako dolazi do novog digitalnog poslovnog modela (Kaufman i Horton, 2015).

U doktorskoj disertaciji razmatraju se interni aspekti digitalne transformacije, čiji je prvenstveni cilj kvalitetno i efikasno odlučivanje, na osnovu analize relevantnih podataka. Sveopšta digitalizacija i razvoj tehnologija, omogućili su jednostavan pristup i brzu obradu velikih količina podataka (internih i eksternih) svih struktura. Dok se poslovnom analitikom (eng. *business intlligence*), analizira šta se i zbog čega u prošlosti desilo, primenom prediktivnih algoritama mašinskog učenja moguće je predvideti buduće događaje.

1.1. Nauka o podacima i veštačka inteligencija

Automatsko ili mašinsko učenje (eng. *machine learning – ML*) predstavlja tehnologiju kojom se kompjuterski softverski sistemi samostalno razvijaju, 'uče' na osnovu obrade velike količine ulaznih podatka, bez potrebe da se programiraju eksplicitno (Samuel, 1959). Pod samostalnim

učenjem podrazumeva se postupak optimizacije kompjuterskog programa povećanjem količine procesiranih podataka (T. Mitchell, 1998).

Razvoj i praktična primena *ML* algoritama postaje moguća sa rastom količine raspoloživih podataka na kojima se ovi algoritmi mogu razvijati, 'učiti'. Digitizacija, postupak kojim se konvertuju fizički objekti i atributi u digitalne, započeta je još krajem sedamdestih godina. Ovo je dovelo do eksponencijalnog rasta količina digitalnih podataka. Najveći uticaj na ovaj rast ima primena senzora, odnosno veliki broj *IoT* (eng. *Interent of Things*) konektovanih uređaja. Pored ovih mašinski generisanih podataka, ljudi svojim danas uobičajenim aktivnostima, a pre svega preko socijalnih mreža, značajno utiču na ovaj eksponencijalni trend rasta. Oko 80% novokreiranih podataka spada u grupu nestrukturiranih (podaci neodređenih struktura). Zbog velikog obima, brzine nastajanja i svoje raznolikosti, ovakve podatke nazivamo Velikim podacima (eng. *Big data*) (Provost & Fawcett, 2013).

Paralelno sa fenomenom Velikih podataka, sličnom brzinom razvijaju se i tehnologije koje omogućavaju njihovo procesiranje, prenos i skladištenje (čuvanje).

Performanse procesora, koji izvršavaju programske instrukcije, dupliraju se u proseku svaka 24 meseca (Moor, 1975). Razvojem kvantnih kompjutera eliminisan je osnovni ograničavajući faktor daljeg rasta procesorske snage, usled fizičkog ograničenja prostora za smeštaj tranzistora na procesorskom čipu (eng. *CPU*).

Kapaciteti za prenos podataka optičkim kablovima u proseku se dupliraju svakih devet meseci (Butter, 2001), dok se kapaciteti diskova za smeštaj podataka udvostručuju u proseku na 13 meseci (Kryder, 2015).

Velike količine podataka, kao i mogućnost efikasnog pristupa i obrade, doveli su do široke primene poslovne *data* analitike (eng. *data analytics*), koja je postala sastavni deo odlučivanja na svim nivoima upravljanja i rada. Kada se poslovna *data* analitika koristi kako bi se objasnio događaj i odgovorilo na pitanje šta se i zbog čega dogodilo, govorimo o opisnoj (eng. *describing*), dijagnostičkoj (eng. *diagnostic*) analitici (Provost i Fawcett, 2013). Kada se strukturirani i nestrukturirani podaci analiziraju sa ciljem da se predvide događaji, otkriju zavisnosti i šabloni u podacima, primenom statističkih metoda i algoritama *ML*, govorimo o naprednoj prediktivnoj analitici (Provost i Fawcett, 2013).

1.2. Mašinsko učenje

ML i duboko učenje (eng. *Deep learning*) su podgrupe veštačke inteligencije (eng. *Artificail Intelligence-AI*). Pod pojmom veštačke inteligencije podrazumevamo autonomne inteligente sisteme, koji deluju samostalno, bez ljudske intervencije. Pojam *AI* prvi put je upotrebio John McCarthy 1955. godine, a već sledeće godine održana je prva konferencija o neuronskim mrežama, koje su bile osnovni predmet istraživanja iz oblasti *AI*, sve do kraja 1970. (slika 1).

Slika 1. Razvoj veštačke inteligencije i *ML*[3]

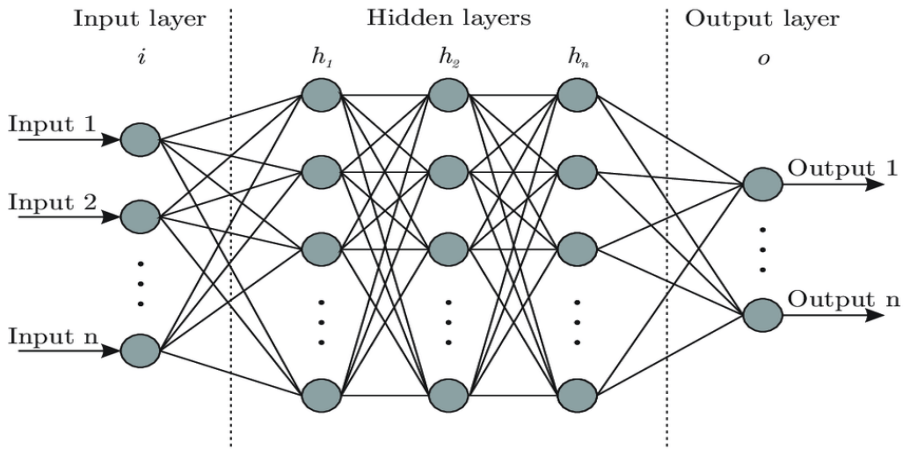
Kasnije, primat u razvoju *AI* preuzimaju algoritmi *ML*, specijalni modeli višeslojnih neuronskih mreža. Oni se zbog svog značaja i široke primene navode kao posebna podgrupa *ML* – Duboko učenje (eng. *Deep learning – DL*).

DL se primenjuje u slučajevima klasifikacije kompleksnih podataka. Razvijeni su po analogiji sa neuronskom mrežom ljudskog mozga. Ulazni podaci se iz prvog nivoa neuronske mreže (eng. *input layer*), prenose ka skrivenim nivoima (eng. *hidden layers*), gde se sukcesivno transformišu primenom nelinearnih funkcija (eng. *activation function*). Sa rastom broja nivoa *hidden* nivoa, odnosno sa svakom novom transformacijom, povećava se tačnost modela. U poslednjem, izlaznom

[3] Artificial Intelligence (AI): What it is and why it matters | SAS

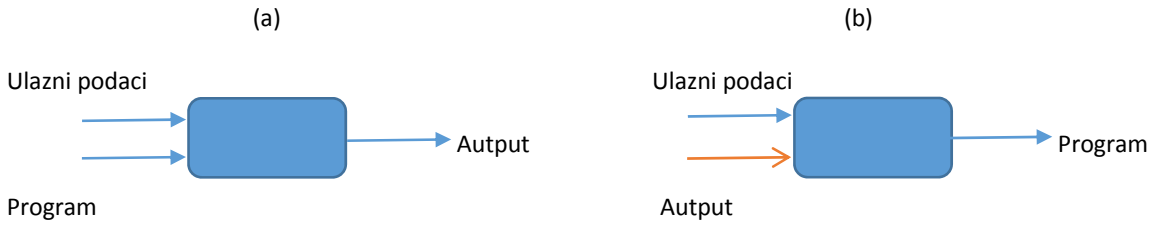
nivou mreže (eng. *output layer*), na osnovu primenjenih aktivacionih funkcija dobijamo konačnu predikciju.

Slika 2. Arhitektura neuronske mreže[4]



Inače, za razliku od tradicionalnog programiranja, gde se na osnovu ulaznih podataka i softverskog programa (logike), dobija rezultat (autput), kod mašinskog učenja se na osnovu ulaznih podataka, iterativnim postupkom učenja algoritma, samostalno (bez potrebe za eksplicitnim programiranjem) kreira *ML* model (slika 3).

Slika 3. Tradicionalno programiranje (a), Mašinsko učenje (b)



[4] Izvor: Researchgate

Ovim iterativnim postupkom optimizacije *ML* modela određuju se vrednosti njegovih parametara, za koje model daje optimalan rezultat. Kada je namena *ML* modela predikcija, konačne vrednosti parametara koji finalno određuju oblik ciljne funkcije (konačnog modela) jesu one za koje je greška predikcije najmanja (James et al., 2013). Rast kompleksnosti prediktivnog modela uobičajeno vodi ka povećanju njegove tačnosti, ali se istovremeno povećava rizik od *overfitinga*, čime se smanjuje mogućnost generalizacije modela, kao i njegovog tumačenja.

U slučaju kada je tačnost predikcije osnovni cilj, a ne razumevanje *ML* modela, kažemo da on predstavljaju crnu kutiju (eng. *black box*) (James et al., 2013). Višeslojne neuronske mreže spadaju u *black box* algoritme, koji daju veliku tačnost u regresionim i klasifikacionim problemima, ali zbog velikog broja iterativnih nelinearnih transformacija, ovaj model je nemoguće razumeti.

S druge strane, ukoliko je pored predikcije cilj istraživanja i razumevanje zavisnosti između ulaznih i izlaznih varijabli, govorimo o inferencijalnim (eng. *Inference*) *ML* modelima (James et al., 2013).

U disertaciji ćemo se baviti komparativnom analizom tačnosti najčešće primenjivanih *ML* klasifikacionih prediktivnih modela, od kojih neki spadaju u *black box*, dok su drugi inferencijalni.

ML modele razlikujemo i prema postupku kojim se dolazi do ciljne funkcije – konačnog oblika modela. Kada se njen početni oblik pretpostavi, a potom iterativnim postupkom učenja odrede optimalne vrednosti parametara, koeficijenata ciljne funkcije, govorimo o parametarskim metodama *ML* (eng. *parametric*) (Garet et al., 2014). Kako se iterativnim postupkom učenja određuju samo vrednosti parametara, a ne i oblik funkcije, ovakvi *ML* modeli su veoma efikasni – za učenje modela ne zahteva se velika količina podataka (Shwartz i Shai, 2014). Međutim, ukoliko pretpostavljena funkcija ne opisuje dobro stvarnu relaciju između promenjivih, tačnost predikcije ovih modela je niska.

Kada se oblik funkcije ne pretpostavlja, već se do njega dolazi postupkom učenja, govorimo o neparametarskim metodama *ML* (eng. *non paramtric*) (Garet et al., 2014). Ovaj potupak zahteva znatne količine trening podataka (podataka za učenje) kako bi se došlo do oblika funkcije koja najbolje opisuje zavisnost promenljivih. Neparametarski modeli imaju veću tačnost, ali su skloni *overfiting*-u. Iz tog razloga moguće je, postupkom regularizacije, ograničiti kompleksnost ovakvih modela (Garet et al., 2014).

Podaci na kojima se model razvija – uči, nazivamo trening podacima, dok su test podaci oni na kojima se proverava njegova tačnost. Trening i test podaci predstavljaju podskup tzv. univerzalnog

skupa (*US*) podataka, iz kojeg se slučajnim uzorkovanjem osigurava njihova nezavisnost, dok je verovatnoća njihovog izbora iz *US* jednaka za svaki uzorak (eng. *indipendant and indenticaly sampled*) (Garet, J. et al., 2014). Trening i test podaci postoje u dva osnovna oblika:

- Obeleženi *(*eng. *labeled)* podaci – kod kojih su poznati i ulazni (X) i izlazni (Y) podaci. Odnosno kažemo da X određuje Y ili da Y daje kontekst ili značenje podacima X. Tako se uzorak i (ili opservacija) može predstaviti u obliku (x_i, y_i). Ulazni podatak (x_i) je kolonski vektor sa m elemenata, koje nazivamo prediktorima (ili nezavisnim promenljivima), dok izlazni podatak y_i nazivamo zavisnom ili target promenljivom. Uvodimo sledeću notaciju:

$$X = \{x_i \epsilon R^m\}, \;\; i = \{1, .., n\}\,, X \epsilon R^{n\,x\,m}$$

$$Y = \{y_i \epsilon R\}, \quad i = \{1, .., n\}, \; Y \epsilon R^n$$

Y – kolonski vektor zavisne promenjive, reda n x 1

$X \epsilon R^{n\,x\,m}$ – matrica prediktora, reda n x m:

$$X = \begin{bmatrix} x_1^T \\ x_2^T \\ . \\ . \\ x_n^T \end{bmatrix}$$

$x_i - i$ uzorak, kolonski vektor reda m x 1

n – ukupan broj uzoraka

m – ukupan broj prediktora

R – skup realnih brojeva

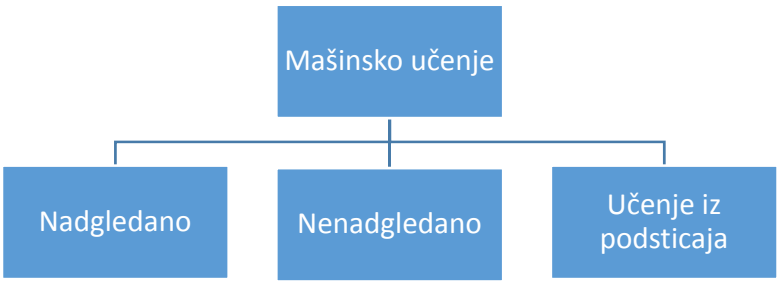
- Neobeleženi *(*eng. *unlabeled data)* podaci su oni kod kojih ne postoji target promenljiva Y, već samo ulazni podaci:

$$X = \{x_i \epsilon R^m\}, \quad i = \{1, .., n\}\ , X \epsilon R^{n\,x\,m}$$

1.2.1. Osnovni tipovi *ML*

Prema problemu koji izučavamo, kao i strukturi podataka (obeleženi ili neobeleženi), razlikujemo tri osnovna tipa mašinskog učenja i to: nadgledano (eng. *supervised learning*), nenadgledano (eng. *unsupervised learning*) i učenje iz podsticaja učenje (eng. *reinforcement learning – RL*) (slika 4) (Harrington, 2016).

Slika 4. Tipovi *ML*

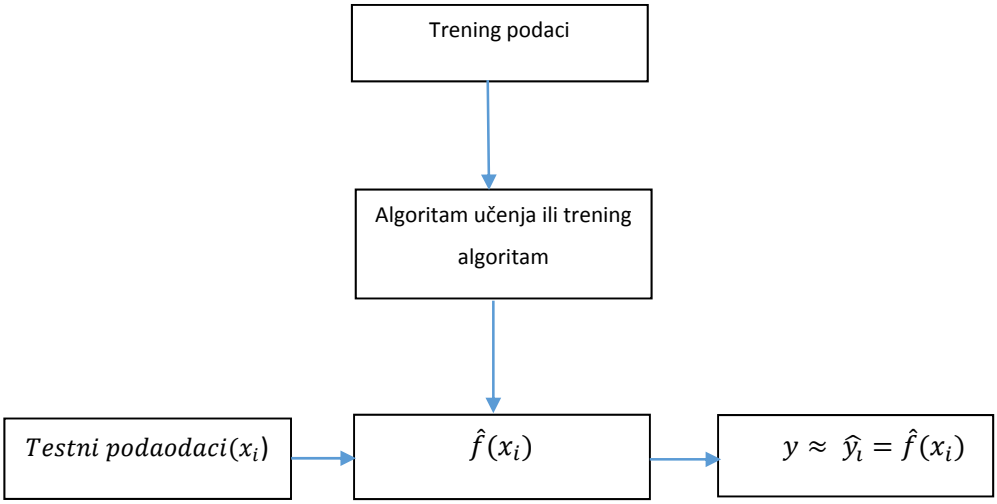


- Nadgledano učenje

  Ukoliko su podaci obeleženi, poznate su vrednosti prediktora X i zavisne promenljive Y, tada govorimo o nadgledanom učenju, kojim se razvija model ciljne funkcije $\hat{f}$, koja na osnovu uzorka prediktora x_i ima cilj da predvidi vrednost outputa $\widehat{y_i}$ što približnije njenoj stvarnoj vrednosti y (slika 5).

Ukoliko je Y numerička (kvantitativna) promenljiva, onda prediktivni *ML* algoritam nazivamo regresionim. Ukoliko je Y kategorička (može pripadati ograničenom broju kategorija ili klasa), model na osnovu vrednosti x_i svrstava uzorak u jednu od mogućih klasa zavisne promenljive. Ovakve prediktivne modele nazivamo klasifikacionim.

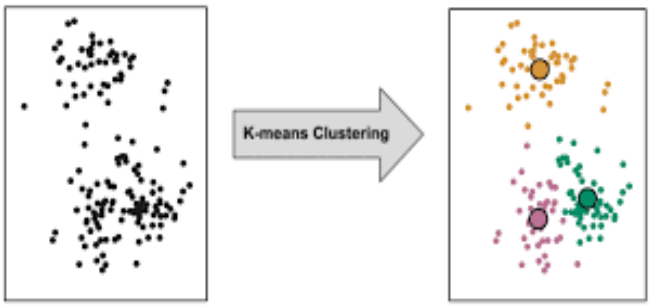
Slika 5. Algoritam nadgledanog učanja



- Nenadgledano učenje

Kada su podaci neobeleženi, govorimo o nenadgledanim algoritmima *ML*, kojima se analizira struktura podataka X, na način da se međusobno slični uzorci grupišu zajedno u klastere. Kako se svaki uzorak može prikazati kao tačka u m dimenzionalnom prostoru ($x_i \epsilon R^m$), sličnost uzoraka se najčešće meri njihovom međusobnom razdaljinom. Najbliži uzorci su najsličniji i pripadaju istom klasteru. Svaki klaster ili grupa sličnih uzoraka određeni su njihovim brojem i težištem (eng. *centroid*). Tako se novi, ispitivani uzorak, dodeljuje klasteru čijem je težištu najbliži.

Slika 6. Primer grupisanja uzoraka u tri klastera[5]



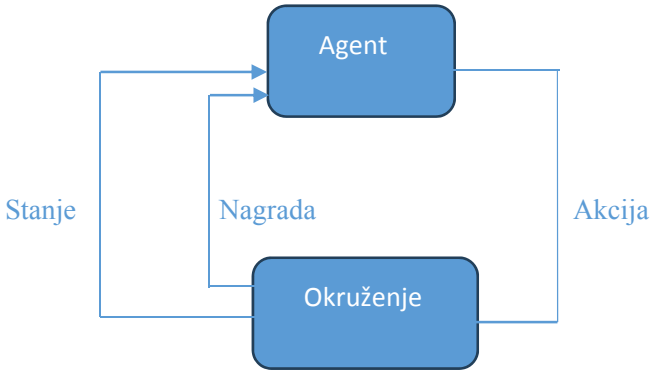
Drugu grupu algoritama nenadgledanog učenja čine oni kojima se redukuje dimenzionlnost podataka i tako dobijaju ciljne funkcije manje kompleksnosti (eng. *dimension reduction*). Podaci se projektuju u novi prostor, određen osama koje predstavljaju pravce maksimalne varijabilnosti. Ovom transformacijom eliminiše se eventualno postojanje međuzavisnosti prediktora.

- Učenje iz podsticaja

Učenje iz podsticaja (eng. *reinforcement learning – RL*) ili učenje sa podrškom, bazira se na postupku učenja algoritma kroz uzastopne pokušaje da se napravi očekivana akcija, odluka (eng. *trial and error*). Inteligentni softverski agent preduzima autonomno akciju u okruženju, bivajući nagrađen za očekivanu (pravilnu) akciju, dok u suprotnom slučaju biva sankcionisan. Na ovaj način, učeći iz vlastitog iskustva, teži se postizanju maksimalnog broja pravilnih – očekivanih akcija inteligentnog agenta (slika 7). *RL* se pre svega primenjuje u robotici, autonomnoj vožnji, kao i tzv. *recomendation* sistemima (Sutton i Barto, 2018).

[5] Izvor: K Means Clustering. In the previous story we understood… | by Ayush Kalla | DataDrivenInvestor

Slika 7. Učenje iz podsticaja



2. PRIMENA *ML* U FINANSIJAMA I PREGLED LITERATURE

Suština ekonomske nauke jeste tumačenje ekonomskih činjenica i istraživanje njihovih međusobnih relacija. Teorijski koncepti i matematički metodi često se, zbog neadekvatnih i nedovoljnih empirijskih istraživanja, ne suočavaju sa činjeničnim podacima, kako bi ekonomska teorija mogla biti proverena u praksi. Ekonometrija ima upravo ulogu da pruži empirijsku verifikaciju ekonomske teorije, koja je u osnovi bazirana na kvalitativnim istraživanjima (Gujarati i Porter, 2009). Jedan od razloga nedovoljne zastupljenosti kvantitativnih empirijskih istraživanja u ekonomiji jeste nedostatak podataka, njihov upitan kvalitet, nestandardni formati i nemogućnost njihove obrade i analize.

Sa ubrzanom digitalizacijom poslovanja došlo je do eksponencijalnog rasta količina elektronskih podataka. Nove digitalne tehnologije i infrastruktura omogućavaju danas jednostavan pristup, prenos, transformaciju i obradu podataka različitih struktura, čime su se stvorili uslovi za razvoj ekonometrije, ali i sve veću praktičnu primenu *ML* algoritama u empirijskim istraživanjima u finansijama. Tako Varian (2014), opisuje nove mogućnosti primene *ML* u finansijskoj analizi Velikih podataka, Mullainathan i Spiess (2017) konstatuju sve veću primenu *ML* predikcionih algoritama.

Kako bi se objasnio ekonomski problem ili pojava, empirijskim istraživanjem analizira se relacija i uticaj koji ekonomske varijable imaju na datu pojavu. Dominantni metod u ekonometriji, koji se zbog svoje visoke interpretabilnosti koristi, jeste linearna regresija, kojom se objašnjava pravac i statistički značaj uticaja nezavisnih ekonomskih varijabli (x), na vrednost zavisne promenjive (y). Kako je dobijeni eksplanatorni model linerarne regresije oblika $y \approx \hat{y} = \hat{f}(\beta; x)$, kažemo da se tradicionalnim ekonomskim metodama dolazi do vrednosti parametra β, koji određuje prirodu i intenzitet uticaja svake finansijske varijable pojednično na ekonomski problem koji je predmet istraživanja. Mullainathan i Spiess (2017), navode tako da se tradicionalnim ekonomskim metodama rešava β problem.

Primenom nadgledanih metoda *ML*, moguće je znatno proširiti obuhvat ovakve analize. Pored objašnjenja pojedinačnih uticaja prediktora (ekonomskih varijabli) na zavisno promenljivu, moguće je na osnovu vrednosti prediktora (x_i), predvideti (y_i), uzimajući u obzir njihovu nelinearnu (ili linearnu) zavisnost, kao i sinergetski efekat prediktora na vrednost y_i. Postupkom regularizacije *ML* modela, moguće je takođe model dodatno pojednostaviti svođenjem vrednosti linearnih koeficijenata regresije β na vrednost nula, za one prediktore koji manje utiču na varijabilnost zavisne promenljive. Ovime se dodatno može uticati na povećanje tačnosti predikcije, kao i na mogućnost generalizacije modela (Hastie et al., 2009). Henley i Hand (1997) u svom radu konstatuju da i minimalna povećanja tačnosti predikcije donose značajnu vrednost kompaniji. Tako kažemo da je primenom nadgledanih *ML* algoritama u ekonomiji moguće rešavanje $\hat{y}$ problema (Mullainathan i Spiess, 2017).

Nenadgledane algoritme *ML* u finansijama moguće je primeniti za detaljnu strukturnu analizu empirijskih podataka. Ovim postupkom slični uzorci se grupišu u klastere ili grupe na način da su razdaljine sličnih uzoraka u klasteru najmanje moguće (minimalna varijabilnost podataka u grupi), a razdaljine težišta klastera različitih grupa uzoraka najveće moguće (maksimalna varijabilnost podataka različitih grupa). Kažemo da se nenadgledanim *ML* algoritmima mogu rešavati x problemi (Mullainathan and Spiess, 2017) i da se na osnovu uočenih sličnosti i razlika mogu bolje razumeti podaci koji su predmet istraživanja.

Treća moguća primena *ML* algoritama jeste analiza nestrukturiranih, nekonvencionalnih podataka, kao što su tekst, video ili slika, koji u klasičnim finansijskim analizama nisu uzimani u razmatranje – nije ih bilo moguće analizirati.

Tabela 1. Razlika između tradicionalne ekonomske analize i dva tipa *ML*: nadgledanog i nenadgledanog. Ekonomska analiza objašnjava ekonomski fenomen, dok *ML* omogućava tačniju predikciju i bolje razumevanje strukture podataka.

Pristup	Podaci	Metod	Rezultat	Svrha
Tradicionalna ekonometrija	Obeleženi (x_i, y_i)	Linerana regresija	Model zavisnosti i statističkog značaja promenljivih	$\underline{\beta \text{ problem}}$
Nadgledano učenje	Obeleženi (x_i, y_i)	*ML*, Nadgledano učenje	Predikcija	$\underline{\hat{y} \text{ problem}}$
Nenadgledano učenje	Neobeleženi (x_i)	*ML*, Nenadgledano	Sličnosti, šabloni podataka	$\underline{x \text{ problem}}$

Predikcija stečaja i procena kreditnih rizika jedan su od najvećih izazova moderne ekonomije i finansijskih istraživanja. Tradicionalne statističke tehnike, kao što su multivarijantna diskriminantna analiza i logistička regresija, dominantno su se primenjivale u rešavanju ovih problema. Kako su osnovni nedostaci ovog pristupa rigidne i često neodržive statističke pretpostavke, kao što je linearnost aditivnog modela, nezavisnost prediktora, otpočelo se sa većom primenom *ML*. Interesantno je da je u 2021. stručnih radova iz oblasti primene *ML* u finansijama objavljeno 11 puta više od proseka u periodu 2000–2017. (Hoang i Wiegratz, 2022). Isti autori, kao i Breeden (2020), dalje navode da u slučajevima kada su podaci visoke dimenzionalnosti (veliki broj prediktora), kada zavisnost promenljivih nije linearna i postoji visok stepen multikolinearnosti, treba dati prednost *ML* u odnosu na statističke metode. Pregled nekih od objavljenih radova na temu primene prediktivnih *ML* algoritama na problemima kreditnog rizika i verovatnoće stečaja kompanija dat je u tabeli 2.

Tabela 2. Pregled značajnijih radova iz oblasti primene *ML* u svrhu predikcije stečaja *(BP)* i proceni kreditnog rizika *(CS)*

ML Modeli	Algoritmi	Kreditni rejting - *CS*	Predikcija stečaja - *BP*
Regresioni	Log regresija, Lasso, Ridge, Elastic Net	Addo, P.M. et al. (2018) Baesens, B. et al. (2015) Stelzer, A. (2019)	Alak, H.A. et al. (2018), Lombardo, G. et al. (2022),Chen,M. (2011)
	LDA	Brown, I. i Mues, C. (2012) Baesens, B. et al. (2015) Stelzer, A. (2019)	Alak, H.A. et al. (2018), Baesens, B. et al. (2010), Chen,M. (2011)
	QDA	Brown, I. i Mues, C. (2012) Baesens, B. et al. (2015) Stelzer, A. (2019)	Baesens, B. et al. (2010)
Modeli na bazi stabla	Decision tree (CRAT)	Brown, I. i Mues, C. (2012), Stelzer, A. (2019) , Yeh, I. i Lien, C. (2009)	Alak, H. A. et al. (2018), Chen, M. (2011)
	Random Forest	Addo, P.M. et al. (2018) Brown, I. i Mues, C. (2012) Stelzer, A. (2019)	Lombardo, G. et al. (2022), Narvekar, A. et al. (2021), Barboza, F. et al. (2017)
	Gradient tree boosting	Addo, P.M. et al. (2018) Brown, I. i Mues, C. (2012) Stelzer, A. (2019)	Lombardo, G. et al. (2022), Narvekar, A. et al. (2021), Barboza, F. et al. (2017)
Naive Bayes	NB	Stelzer, A. (2019) Yeh, I. i Lien, C. (2009) Baesnes, B. et al. (2015)	Aghaie, A. et al. (2009), Patel, P. et al. (2019)
K najbližih suseda	K-NN	Brown, I. i Mues, C. (2012) Stelzer, A. (2019) Yeh, I. i Lien, C. (2009)	Lendasse, A. et al. (2009)
Support vector machine	SVM, Linear	Brown, I. i Mues, C. (2012) Baesens, B. et al. (2015) Stelzer, A. (2019	Alak, H.A. et al. (2018), Lombardo, G. et al. (2022) , Narvekar, A. et al. (2021), Baesens, B. et al. (2010)
	SVM, Radial kernel	Stelzer, A. (2019) Baesens et al. (2003)	Min, J. H. et al. (2005), Barboza, F. et al. (2017)
	SVM, Polinomial kernel	Putri, N. et al. (2021) Baesens et al. (2003)	Min, J. H. et al. (2005)

U radu su opisani osnovni *ML* modeli i njihova primena u svrhu razvoja *credit scoring*-a *(CS)* i predikcije stečaja *(BP)* na dve nezavisne grupe podataka *(„Default of credit card clients dataset"* i *„Polish companies bankruptcy data set"*). Sprovedena empirijska analiza performansi ovih algoritama nema za cilj da predloži najbolji, najoptimalniji model, već da pokaže da se njihovom primenom može dobiti zadovoljavajući nivo tačnosti predikcije. Tako i Baesens, B. et al., (2003), u analizi klasifikacionih *ML* algoritama (41 model, 8 data setova) u razvoju *CS* modela, konstatuju da ne postoji jedan najbolji *ML* model i da u zavisnosti od strukture i količine podataka, svaki model pokazuje određene prednosti i nedostatke. Ovu tezu potvrđuju i Abdou, H. i Pointon, J. (2011) koji, na osnovu 214 stručnih radova iz oblasti primene statističkih i *ML* metoda na problem *CS* i *BP*, zaključuju da ne postoji jedinstven stav u pitanjima: izbora indikatora – promenljivih prediktivnog modela; određivanja granične vrednosti verovatnoće; validacione metode (odnosa trening i test podataka); veličine uzorka, te da se ne može govoriti o postojanju jednog univerzalnog modela.

Kao osnovne nedostatke u primeni *ML* u ekonomiji, Lessmann et al., (2015) navode da je količina podataka za analizu često nedovoljna (samo 5 od 48 naučnih radova na temu *CS* imalo je više od 10.000 uzoraka), kao i da prioritet postizanja veće tačnosti predikcije dovodi do kompleksnih algoritama, koje je nemoguće tumačiti.

2.1. Linearna regresija, *LR*

Regresioni modeli spadaju u grupu osnovnih algoritama, najčešće primenjivanih u ekonometrijskim analizama. Njima se objašnjava zavisnost jedne promenljive (y), od jedne ili više nezavisnih promenljivih – prediktora (x). Radi se o analizi zavisnosti u jednom smeru, uticaja prediktora na vrednost zavisne promenljive (y). Kada je cilj da se regresionim modelom opiše zavisnost varijabli modela, kažemo da se *LR* primenjuje u svrhu inferencijalne analize (Gareth et al., 2017). Regresioni algoritmi se takođe primenjuju i u predikciji vrednosti zavisne promenljive, na osnovu poznatih vrednosti prediktora.

Regresioni model dat je funkcijom $f(x_i)$, koja daje očekivanu vrednost zavisne promenljive, uslovno od vrednosti prediktora – $E[y|x_i]$:

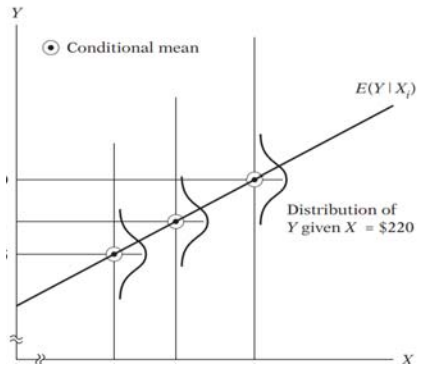
$$E[y|x_i] = f(x_i)$$

Postupkom učenja ili fitovanja modela, na osnovu raspoloživih trening podataka, određuju se optimalne vrednosti regresionih parametara funkcije $f(x_i)$. U slučaju jednog prediktora, ova funkcija ima sledeći oblik:

$$E[y|x_i] = f(x_i) = \beta_0 + \beta_1 x_i \quad (1)$$

Izraz (1) nazivamo deterministčkim modelom regresije (Mendenhall et al., 2013). U slučaju jednog prediktora, funkcija $f(x)$, predstavlja regresionu pravu liniju. Na osnovu zakona o univerzalnoj regresiji (Galton, 1885), vrednosti slučajne (stohastičke) promenljive y, za date fiksne vrednosti x, regresiraju ka svojoj srednjoj vrednosti. Tako možemo pretpostaviti da optimalni regresioni model prolazi kroz srednju vrednost, očekivanu vrednost zavisne promenljive y. Kako se radi o slučajnoj promenljivoj, pretpostavlja se da je njen raspored verovatnoće Normalan (Gujarati i Porter, 2009).

Slika 8. Normalna distribucija zavisne promenljive za fiksne vrednosti x[6]



Regresioni model (1) ne opisuje relaciju zavisne promenljive i prediktora na idealan način – postoje odstupanja očekivanih vrednosti zavisne pormenljive $E[y|x_i]$ od njenih stvarnih vrednosti y_i. Zato

[6] Izvor: Porter i Gujarati (2009). *Basic Econometrics,* s. 37

se ovaj model prikazuje u *probabalistic* obliku (Mendenhall i dr. 2013), koji nazivamo i regresionom funkcijom populacije (eng. *Population regression function – PRF*) (Gujarati i Porter, 2009):

$$y_i = E[y|x_i] + u_i = f(x_i) + u_i = \beta_0 + \beta_1 x_i + u_i \quad (2)$$

$u -$ slučajna greška (eng. *random error*), koju još nazivamo rezidualom, jeste mera odstupanja stvarne vrednosti zavisne promenljive od očekivane. Očekivana vrednost izraza (2) jednaka je:

$$E[y_i] = E[E[y|x_i]] + E[u_i]$$

Kako je $E[y_i] = E[E[y|x_i]]$, sledi da je $E[u_i] = 0$.

Pretpostavlja se da je vrednost prediktora x fiksna i nezavisna od reziduala. Tako je njihova kovarijansa, mera njihove linearne zavisnosti, jednaka nuli, $cov(x_i, u_i) = 0$. Kako je varijansa reziduala jednaka:

$$Var(u_i) = E[u_i - E[u_i|x_i]]^2$$

a kako smo pokazali da je $E[u_i] = 0$, sledi i da je $E[u_i|x_i] = 0$, pa dobijamo da je varijansa reziduala konstantna i jednaka:

$$Var(u_i) = E(u_i)^2 = \sigma^2$$

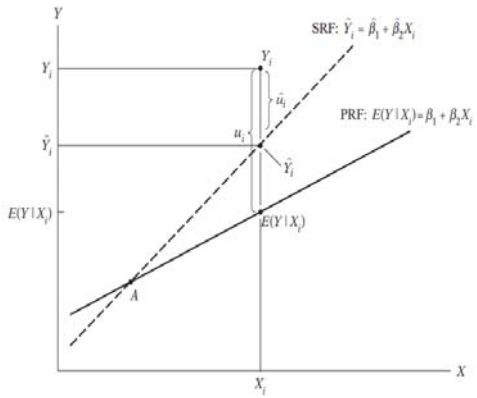
Na osnovu pretpostavke da regresiona prava prolazi kroz srednje vrednosti simetrične distribucije zavisne promenljive, dobili smo da je $E[u_i] = 0$, dok je varijansa reziduala konstantna i jednaka σ^2. Pretpostavimo još da reziduali imaju normalnu distribuciju, $u \in N(0, \sigma^2)$.

Kakao nam podaci iz cele populacije najčešće nisu poznati, cilj je da primenom statističkih metoda na slučajno uzorkovanim podacima dođemo do najboljeg modela (eng. *best fitting line*) $\hat{f}(x)$, linije srednjih očekivanih vrednosti zavisne promenljive y, koja najbolje aproksimira funkciju $f(x)$. Model opisan funkcijom $\hat{f}(x)$ nazivamo regresionom funkcijom uzorka (eng. *Sample regression function – SRF*) (Gujarati i Porter, 2009):

$$y_i = \hat{f}(x_i) + \hat{u}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i$$

$$\hat{u}_i = y_i - \hat{y}_i$$

Slika 9. Regresiona funkcija uzorka (*SRF*) i regresiona funkcija populacije (*PRF*)[7]



Vrednosti regresionih parametara $\hat{\beta}$, koji će biti najbolja moguća (optimalna) aproksimacija parametara β, regresione funkcije populacije (*PRF*), određujemo primenom dve metode (Gujarati i Porter, 2009):

[7] Izvor: Porter i Gujarati (2009). *Basic Econometrics,* s. 45

- Najmanjih kvadrata (eng. *Ordinary least squares – OLS*)
- Maksimalne verodostojnosti (eng. *Maximum Likelihood – MaxL*)

Metoda najmanjih kvadrata, *OLS*

Na osnovu pretpostavke da je srednja vrednost reziduala nula, pretpostavljeni oblik multivarijantne regresione funkcije *SRF*, $\hat{f}(x)$, ima oblik:

$$\hat{y} = \hat{f}(x;\hat{\beta}) = \hat{\beta}_o + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots . \hat{\beta}_m x_m \quad (3)$$

Regresioni model podrazumeva samo linearnu zavisnost promenljive y od regresionih parametara, dok je dozvoljena nelinearna zavisnost od prediktora. Sledi da ukoliko zavisnost između y i prediktora nije linearna, potrebno je transformisati prediktore, najčešće stepenovanjem. Tada govorimo o polinominalnoj regresiji. Pored stepenovanja, često se prediktori transformišu i logaritmovanjem. Moguće je i primeniti postupak tzv. *spline* regresije, kada se regresioni model sastoji od više međusobno povezanih regresionih modela, u prelomnim tačkama (eng. *knots*), najčešće određenim kvartilnim vrednostima prediktora (Q_1, Q_2, Q_3), (Alboukadel K., 2017).

Izraz (3) koji definiše *SRF* funkciju može se pojednostaviti:

$$\hat{y} = \hat{f}(x;\hat{\beta}) = \sum_{i=0}^{m} \hat{\beta}_i x_i \quad (4)$$

pretpostavili smo da je $x_0 = 1$.

U matričnom obliku izraz (4) postaje:

$$\hat{y} = \hat{f}(x; \hat{\beta}) = x^T \hat{\beta} \tag{5}$$

gde su:

$x = \begin{bmatrix} x_0 \\ x_1 \\ . \\ . \\ x_m \end{bmatrix}$ – vektor prediktora; $\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ . \\ . \\ \hat{\beta}_m \end{bmatrix}$ – vektor parametara linearnog regresionog modela

Kako jednačina (3), definiše hiperravan u *m* dimenzionlanom prostoru, njen optimalan položaj se određuje tako da suma kvadrata razlika rastojanja između $\hat{y}(x)$ i stvarnih vrednosti y bude minimalna. Funkcija sume kvadrata reziduala naziva se *cost* ili *objective* funkcija $J(\hat{\beta})$ linearnog modela (Alboukadel, 2017) i ima oblik:

$$J(\hat{\beta}) = \frac{1}{2} \sum_{i=1}^{n} (\hat{y}(x_i) - y_i)^2 \tag{6}$$

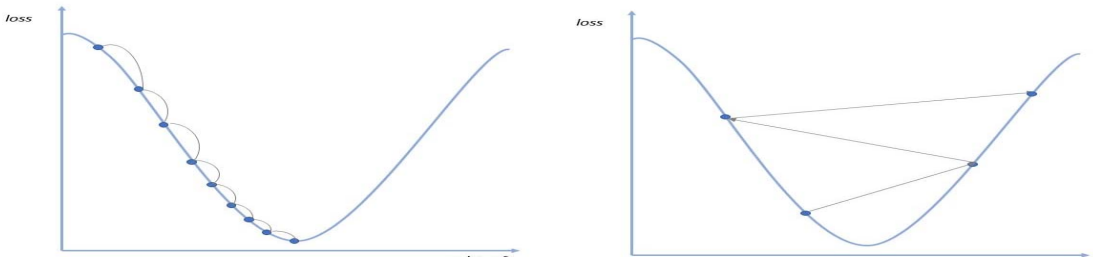
Do traženih vrednosti vektora parametara $\hat{\beta}$, za koju *cost* funkcija $J(\hat{\beta})$ ima minimalnu vrednost, dolazi se iterativnim postupkom. Početna, pretpostavljena vrednost parametara $\hat{\beta}$ smanjuje se dok ne započne da konvergira ka minimalnoj vrednosti optimizacione funkcije. Kako je cilj da se u što manjem broju iteracija dođe do ove optimalne vrednosti vektora parametra $(\hat{\beta})$, a kako gradijent vektor (vektorsko polje) određuje pravce najbrže promene funkcije $J(\hat{\beta})$, primenjuje se optimizacijski algoritam (eng. *gradient descen)* (Porter i Gujarati, 2009), koji možemo prikazati kao:

$$\hat{\beta}_j := \hat{\beta}_j - \alpha \frac{\partial}{\partial \hat{\beta}_j} J(\hat{\beta}) \qquad j = 0,1,.....m$$

Promene $\hat{\beta}$ dešavaju se istovremeno za sve vrednosti j. Koeficijent učenja α, povećava intenzitet gradijent vektora i tako utiče na veličinu svakog inkrementa promene vrednosti parametra. Male vrednosti koeficijenta učenja povećavaju tačnost pronalaženja minimalne vrednosti *cost* funkcije jer se ka njoj dolazi u malim koracima, iteracijama promene parametra $\widehat{\beta}$. Treba imati u vidu da je posledica male vrednosti koeficijenta produženo vreme procesiranja. Veća vrednost ovog koeficijenta, s druge strane, povećava performanse modela, ali može dovesti do toga da se netačno odredi minimalna vrednost *cost* funkcije (slika 10).

Slika 10. *Gradient descent* sa manjim i većim koeficijentom učenja α

Mala vrednost α Veća vrednost α



U slučaju jednog uzorka (x_i, y_i), parcijalni izvod *cost* funkcije po parametru $\widehat{\beta}$, jednak je:

$$\frac{\partial}{\partial \hat{\beta}_j} J(\hat{\beta}) = \frac{\partial}{\partial \hat{\beta}_j} \frac{1}{2} (\hat{y}(x) - y)^2$$

$$= 2\frac{1}{2} (\hat{y}(x) - y) \frac{\partial}{\partial \hat{\beta}_j} (\hat{y}(x) - y)$$

$$= (\hat{y}(x) - y) \frac{\partial}{\partial \hat{\beta}_j} (\sum_{i=0}^{m} \hat{\beta}_i x_i - y)$$

$$= (\hat{y}(x) - y) x_j$$

Dobijeni izraz nazivamo inkrement promene parametra (eng. *update rule*) (Gareth et al., 2017). Za uzorak (x_i, y_i) dalje važi:

$$\hat{\beta}_j := \hat{\beta}_j - \alpha(\hat{y}(x) - y)x_j \qquad j = 0,1,.....m \tag{7}$$

Izraz (7) se naziva Widrow Huff pravilo učenja (Gareth, 2017). Vidimo da je inkrement promene $\hat{\beta}_j$, proporcionalan vrednosti reziduala $(\hat{y}(x) - y)$. Inkrement promene $\hat{\beta}_j$ je tako veći u slučaju većih vrednosti reziduala odnosno manji za manje vrednosti $(\hat{y}(x) - y)$.

Kada primenimo izraz (7) na sve uzorke trening podataka dobijamo:

$$\hat{\beta}_j := \hat{\beta}_j - \alpha \sum_{i=1}^{n} (\hat{y}(x_i) - y_i)x_j \qquad j = 0,1,\ldots.m \tag{8}$$

Ovakav postupak dobijanja optimalne vrednosti za $\hat{\beta}$ nazivamo još i *batch gradient descent* (Gareth, 2017). Kako se pri svakoj iteraciji razlika reziduala računa za sve trening podatke, ovaj postupak je veoma zahtevan (vremenski i resursno).

Alternativa je *stohastic gradient descent*, gde se svaki inkrement promene vrednosti parametara modela bazira na rezidualu jednog trening uzorka (L´eon Bottou, 2010). Tako algoritam prolazi kroz trening podatke samo jednom, a ne onoliko puta koliko ima inkrementa do postizanja minimalne vrednosti $J(\hat{\beta})$:

Za svako *i* od 1:n {

$$\hat{\beta}_j := \hat{\beta}_j - \alpha\left(\hat{y}\left(x^{(i)}\right) - y^{(i)}\right)x_j \tag{9}$$

} , za svako *j*

Stohastički pristup je brži, manje resursno zahtevan, ali i manje tačan. Preporučuje se u slučajevima velikog broja uzoraka.

Kako je $J(\hat{\beta})$ konveksna kvadratna funkcija, ona ima jedinstvenu minimalnu vrednost. Tako će postupkom *gradient descent* minimalna vrednosti $J(\hat{\beta})$ biti jednoznačno određena.

U slučaju linearne regresije, do parametara $\hat{\beta}$ možemo doći na jednostavniji način od postupka *gradient discent*, direktno, bez potrebe za prethodno opisanim iterativnim postupkom.

Ako vrednosti reziduala napišemo u matričnom obliku:

$$X\,\hat{\beta} - \overrightarrow{y} = \begin{bmatrix} x_1^T\,\hat{\beta} \\ x_2^T\,\hat{\beta} \\ . \\ . \\ x_n^T\,\hat{\beta} \end{bmatrix} - \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} \hat{y}(x_1) - y_1 \\ \hat{y}(x_2) - y_2 \\ . \\ . \\ \hat{y}(n) - y_n \end{bmatrix}$$

X, matrica prediktora (često se u literaturi naziva *design* matricom).

Imajući u vidu da za matrice oblika kolonskog vektora, npr. matricu Z, važi da je $Z^T Z = \sum_{i=1} Z_i^2$, sledi da *cost* funkciju možemo napisati kao:

$$J(\hat{\beta}) = \frac{1}{2}\left(X\,\hat{\beta} - \overrightarrow{y}\right)^T \left(X\,\hat{\beta} - \overrightarrow{y}\right) = \frac{1}{2}\sum_{i=1}^{n}(\hat{y}(x_i) - y_i)^2 \qquad (10)$$

$\nabla_\beta J(\hat{\beta})$ je gradijent vektor, koji određuje pravac tangente na *cost* funkcije $J(\hat{\beta})$. Funkcija optimizacije (*cost* funkcija) ima minimalnu vrednosti, za vrednost parametra $\widehat{\beta}$, za koju je izvod funkcije po $\hat{\beta}$ jednak nuli:

$$\nabla_\beta J(\hat{\beta}) = \frac{1}{2}\nabla_\beta\left(\left(X\,\hat{\beta}\right)^T X\,\hat{\beta} - \left(X\,\hat{\beta}\right)^T \overrightarrow{y} - \vec{y}^T X\,\hat{\beta} + \vec{y}^T\,\overrightarrow{y}\right)$$

$$= \frac{1}{2}\nabla_\beta\left(\hat{\beta}^T (X^T X)\hat{\beta} - \vec{y}^T X\,\hat{\beta} - \vec{y}^T X\,\hat{\beta}\right)$$

$$= \frac{1}{2}\nabla_{\beta}(\hat{\beta}^{T}(X^{T}X)\hat{\beta} - 2(X^{T}\overrightarrow{y})^{T}\hat{\beta}$$

Kako je je $\nabla_x X^T AX = 2AX$ i $\nabla_x A^T X = A$, sledi:

$$= \frac{1}{2}\left(2X^{T}X\hat{\beta} - 2X^{T}\overrightarrow{y}\right) = X^{T}X\hat{\beta} - X^{T}\overrightarrow{y} = 0$$

$$\hat{\beta} = (X^{T}X)^{-1}X^{T}\overrightarrow{y} \qquad (11)$$

(podrazumeva se da je matrica $X^T X$ invertibilna)

Na ovaj, direktan način, u jednom koraku, za prediktivne probleme koji su opisani modelom linearne regresije, možemo doći do vrednosti parametara $\hat{\beta}$.

Metod maksimalne verodostojnosti, *MaxL*

Pored metode najmanjih kvadrata, do vrednosti parametara $\hat{\beta}$ možemo doći probablističkom metodom, Maksimalne verodostojnosti (eng. *Maximum Likelihood*). Ovaj metod se koristi za optimizaciju velikog broja različitih *ML* algoritama.

Verodostojnost u statistici je uslovna verovatnoća da poznati slučajni uzorci $(y_1, y_2, .. y_n)$, dolaze iz populacije ili nekog nadskupa podataka, čiji je raspored verovatnoće određen vektorskim ili skalarnim parametrom (α). Verodostojnost u opštem obliku možemo napisati kao $L(\alpha|y_1, y_2, .. y_n)$. S druge strane, verovatnoća da slučajni uzorci dolaze iz nadskupa podataka, čiji je parametar distribucije (α) poznat, označavamo sa $Pr(y_1, y_2, .. y_n|\alpha)$. U oba slučaja radi se dakle o uslovnoj verovatnoći.

Ako postoje nezavisni događaji $y = \{y_1, y_2, .. y_n\}$, prema multiplikativnoj teoriji, verovatnoća zajedničkog dešavanja nezavisnih događaja jednaka je proizvodu njihovih verovatnoća, tako sledi da je:

$$P_r\ (y_1\ , y_2, .. y_n|\ \alpha) = \prod_{i=1}^{n} P_r(y_i|\alpha)$$

Ako su vrednosti $(y_1\ , y_2, .. y_n)$ poznate, nepoznati parametar distribucije α određujemo primenom metode maksimalne verodostojnosti (eng. *Maximum Likelihood)*. Logaritmovanjem ovog izraza dobijamo funkciju verodostojnosti (eng. *Likelihood function – LF*):

$$LF(\alpha) = lo\ g\ \ P_r(y_1\ , y_2, .. y_n|\ \alpha) = log \prod_{i=1}^{n} P_r(y_i|\alpha) = \sum_{i=1}^{n} \log\ P_r(y_i|\alpha)$$

Tražena vrednost α je ona za koju funkcija $LF(\alpha)$ ima maksimalnu vrednost i dobijamo je iz uslova $\frac{\partial}{\partial \alpha} LF(\alpha) = 0$.

Za primenu metode maksimalne verodostojnosti, u slučaju linearne regresije date izrazom $y_i = \ x_i^T \hat{\beta} + \ u_i$, uslovnu zajedničku verovatnoću zavisne promenljive možemo napisati u obliku:

$$P_r\ \left(\vec{y}\right|\ X; \hat{\beta}\ ) = \prod_{i=1}^{n}\ P_r\ \left(y_i\right|\ x_i; \hat{\beta}\ )$$

Na osnovu pretpostavljene normalne distribucije reziduala, $u_i \approx N(0, \sigma^2)$, sledi da je gustina raspodele (eng. *Cumulative density function – CDF*) reziduala jednaka:

$$Pr(u_i) = (2\pi\sigma^2)^{-\frac{1}{2}}\ e^{-\frac{{u_i}^2}{2\sigma^2}}$$

Kako je $y_i \approx N\left(x_i^T \hat{\beta}, \sigma^2\right)$, iz ovog dalje proizilazi:

$$P_r\ \left(\vec{y}\right|\ X; \hat{\beta}\ ) = \ \prod_{i=1}^{n}\ \ (2\pi\sigma^2)^{-\frac{1}{2}}\ e^{-\frac{1}{2\sigma^2}(y_i - x_i^T \hat{\beta}\ )^2}$$

Vrednosti (y_i, x_i) su poznate, dok su parametari distribucije nepoznati $\hat{\beta}, \sigma^2$. Za njihovo određivanje primenjujemo metod maksimalne verodostojnosti. Funkcija $LF(\hat{\beta}, \sigma^2)$ je jedanka:

$$LF(\widehat{\beta},\ \sigma^2) = \log(P_r\ (\vec{y}|\ X; \hat{\beta}\ )) = log \prod_{i=1}^{n}\ (2\pi\sigma^2)^{-\frac{1}{2}}\ e^{-\frac{1}{2\sigma^2}(y_i - x_i^T \widehat{\beta}\ )^2}$$

$$= \sum_{i=1}^{n} log(2\pi\sigma^2)^{-\frac{1}{2}}\ e^{-\frac{1}{2\sigma^2}(y_i - x_i^T \widehat{\beta}\ )^2}$$

$$= n\ log(2\pi\sigma^2)^{-\frac{1}{2}}\ \ - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - x_i^T \hat{\beta}\ )^2$$

$$= \ n\ log(2\pi\sigma^2)^{-\frac{1}{2}}\ \ - \frac{1}{2\sigma^2}\ \left(X\ \hat{\beta}\ -\ \overrightarrow{y}\right)^T (X\ \hat{\beta}\ -\ \overrightarrow{y\ })$$

Vrednost $\hat{\beta}$, za koju LF ima maksimalnu vrednost, ekvivalentna je uslovu dobijanja minimalne vrednosti izraza $\frac{1}{2}\left(X\ \hat{\beta}\ -\ \overrightarrow{y}\right)^T (X\ \hat{\beta}\ -\ \overrightarrow{y\ })$. Vidimo da je ovaj uslov identičan uslovu postizanjem minimalne vrednosti *cost* funkcije metodom najmanjih kvadrata jednačina (10). Iz ovog sledi da je u slučaju linearne regresije metoda najmanjih kvadrata, specijalni slučaj metode maksimalne verodostojnosti i da oba pristupa daju istu vrednost parametra $\hat{\beta},$ datu jednačinom (11) .

Dobijene vrednosti $\hat{\beta},$ primenom *OLS* i *MaxL* metoda, imaju svojstva slučajne promenljive normalne distribucije, što je posledica pretpostavke da reziduali imaju normalnu distribuciju i da je $u = f(\beta)$. Tako na osnovu *central limit* teoreme, očekivana vrednost parametra $\hat{\beta}$ približno je jednaka je parametru β *PRF* modela, tj. $E[\hat{\beta}] \approx \beta$. Tačnost ove predikcije merimo standardnom greškom (SE), te možemo na osnovu njene vrednosti definisati i interval u kojem sa velikom sigurnošću možemo očekivati da se nalazi vrednost β (eng. *Interval estimate).*

$$\Pr(\hat{\beta} - SE \leq \beta \leq \hat{\beta} + SE) = (1 - \alpha)_{:}$$

$(1 - \alpha)$ nazivamo intervalom sigurnosti (eng. $confidence\ interval$). Najčešće je to vrednost 95%, verovatnoća da se β nalazi u intervalu $\hat{\beta} \pm SE$.

2. 1. 1. Regularizacija regresionih modela

Određivanje regresionih parametara $\hat{\beta}$ metodom *OLS* često dovodi do pojave *overfitinga* modela, što uzrokuje nizak nivo generalizacije. U slučajevima velikog broja prediktora, a specijalno kada je $m > n$, regresioni modeli nisu dovoljno stabilni i tačnost predikcije je niska. Iz tog razloga se pristupa regularizaciji, kojom se vrednost regresionih parametara, za prediktore koji minimalno utiču na zavisno promenljivu, smanjuju do vrednosti približnoj ili jednakoj nuli. Tako se ovim postupkom, 'sankcioniše' regresioni model, usled velike kompleksnosti (velikog broja prediktora u modelu).

Najčešće primenjivane metode regularizacije regresionih modela su:

1. *Ridge* regresioni model (Hoerl i Kennard, 1970), primenjuje L2 regularizaciju. Cilj je da se za prediktore koji manje utiču na vrednost zavisne promenljive njihovi regresioni koeficijenti svedu na vrednosti približno jednake nuli. Regresioni parametri ovog modela dobijaju se iz uslova postizanja minimalne vrednosti *cost* funkcije sledećeg oblika:

$$J_{ridge}(\beta) = \sum_{i=1}^{n}(\hat{y}(x_i) - y_i)^2 + \lambda \sum_{j=1}^{m} \beta^2$$

Za $\lambda = 0$, *Ridge* regularizacija daje iste vrednosti regresionih parametara kao one koje dobijamo metodom najmanjih kvadrata (*OLS*). Kada je $\lambda \neq 0$, jedini način da se vrednost *cost* funkcije smanji je smanjenjem vrednosti regresionih parametara.

2. *Lasso* regresija (Tibshirani, 1996), primenjuje L1 formu regularizacije. Parametri ovog modela određuju se iz sličnog uslova kao kod *Ridge* regresije, sa tom razlikom da tzv. *penalty term* (drugi sabirak u izrazu) nije jednak zbiru kvadrata regresionih parametara, već zbiru njihovih apsolutnih vrednosti:

$$J_{lasso}(\beta) = \sum_{i=1}^{n}(\hat{y}(x_i) - y_i)^2 + \lambda \sum_{j=1}^{m} |\beta|$$

Osnovna razlika između ove dve metode regularizacije (*Ridge i Lasoo*) je u tome što u slučaju *Ridge* regresije vrednosti regresionih parametara ne mogu imati vrednost nula, tako da svi prediktori ostaju u modelu. U slučaju *Lasoo* regresije, β parametri mogu imati vrednost nula, pa se tako u finalnom modelu prediktori sa najmanje uticaja na vrednost zavisne promenljive izostavljaju iz modela.

Parametar $\boldsymbol{\lambda}$ je *shrinkige* koeficijent, koji može imati vrednosti u rangu $[0, \infty]$. Do optimalne vrednosti $\boldsymbol{\lambda}$ dolazi se najčešće postupkom *cross* validacije (CV). Ukoliko je vrednost ovog parametra velika, može doći do značajnog pojednostavljenja modela i pojave *undrefiting*-a.

3. *Elastic Net* regresija je kombinacija prethodna dva modela.

2.2. Logistička regresija, *LogR*

Logistička regresija je parametarski klasifikacioni diskriminativni metod. Primenjuje se u slučajevima kada je potrebno predvideti vrednost zavisne kategoričke promenljive, koja može imati dve ili više kategorija (klasa). *LogR*, kao probabalistički klasifikator, daje uslovnu raspodelu verovatnoće oblika $P_r(y|x)$. Kako se u našem primeru (studiji slučaja), radi o binarnoj zavisno promenljivoj, govorimo o primeni logističke regresije u rešavanju binarno klasifikacionog problema. Tako, kategorička zavisna promenljiva može pripadati npr. jednoj od dve klase, i to:

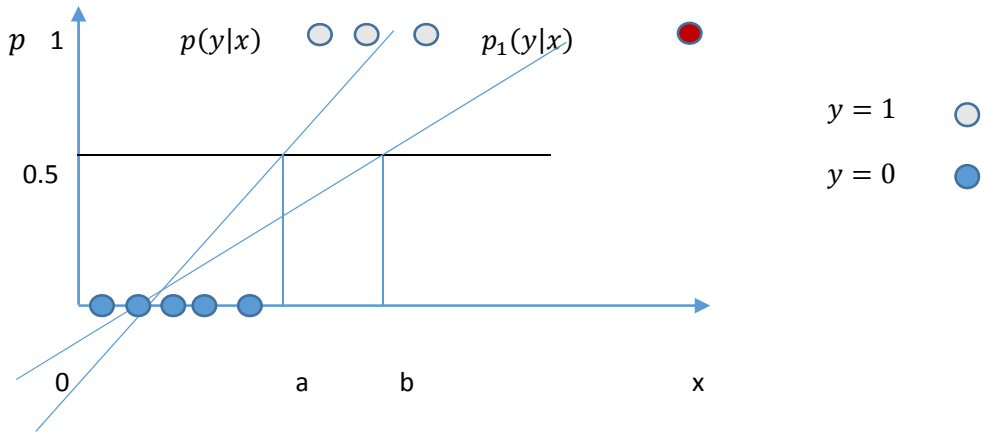
$$Y = \begin{cases} 0, & \text{za kompanije koje nisu u stečaju} \\ 1, & \text{za kompanije u stečaju} \end{cases}$$

Model uslovne raspodele verovatnoće, prediktivni model, $P_r(y = 1|x)$, teorijski je moguće definisati kao linearnu funkciju prediktora, na sličan način kao u *LR*:

$$P_r(y = 1|x) = p(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m = \beta^T X \qquad (1)$$

Model linearne regresije (1) daje vrednosti u granicama $[-\infty, +\infty]$. Kako verovatnoća ima pozitivne vrednost u rasponu $[0,1]$, očigledno je da se linearna regresiona funkcija ne može primeniti kao model raspodele verovatnoće. Pored ovog očiglednog razloga, problem sa linearnom regresijom je i u tome da je ona izuzetno podložna uticaju ekstremnih vrednosti (eng. *outliers*). Pod pretpostavkom da je granična vrednost verovatnoće 0.5, tako de je za $p > 0.5 => y = 1$, na slici 11 možemo videti kako pojava ekstremnih vrednosti utiče na promenu modela i njegovu tačnost.

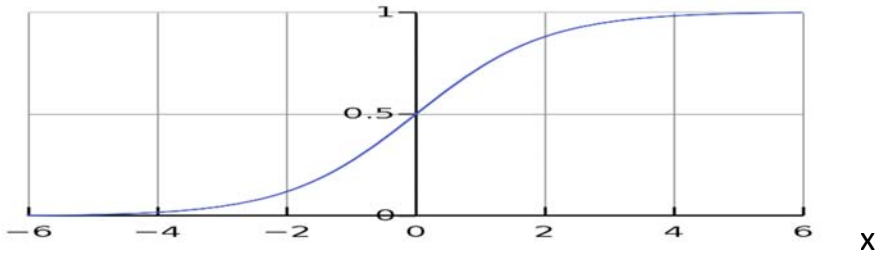
Slika 11. Model linearne regresije, binarno klasifikacionog problema. Regresioni model $p(y|x)$ za $x < a$, klasifikuje uzorak $y = 0$, a za vrednosti $x > a$, $y = 1$. Broj pogrešno klasifikovanih uzoraka je nula. Ukoliko postoji *outlier* (●), regresioni model postaje $p_1(y|x)$, koji klasifikuje uzorke kao $y = 1$, za vrednosti $x > b$ i $y = 0$ za $x < b$. Broj netačno klasifikovanih uzoraka je 2.



Kako bismo ograničili vrednosti $p(x)$ modela linearne regresije, primenićemo transformaciju logističkom funkcijom $g(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$, koja daje vrednosti u očekivanim granicama $[0,1]$ (James et al., 2013).

Slika 12. Logistička funkcija

$$g(x) = \frac{e^x}{1+e^x}$$



Promenom oblika funkcije $p(x)$ iz linearne u logističku dobijamo:

$$P_r(y = 1|x;\ \beta) = p(x) = g(\beta^T X) = \frac{e^{\beta^T X}}{1+e^{\beta^T X}} = \frac{1}{1+e^{-\beta^T X}} \quad (2)$$

Sledi da je verovatnoća da uzorak x pripada klasi $y = 0$ jednaka:

$$P_r(y = 0|x;\ \beta) = 1 - p(x) = \frac{1}{1+e^{\beta^T X}} \quad (3)$$

Izrazi (2) i (3), napisani u jedinstvenom obliku, predstavljaju *Bernoulli* diskretnu distribuciju verovatnoće zavisne binarne kategoričke promenljive y:

$$P_r(y|x;\ \beta) = p(x)^y (1 - p(x))^{1-y} \quad (4)$$

Daljom transformacijom izraza (2) sledi:

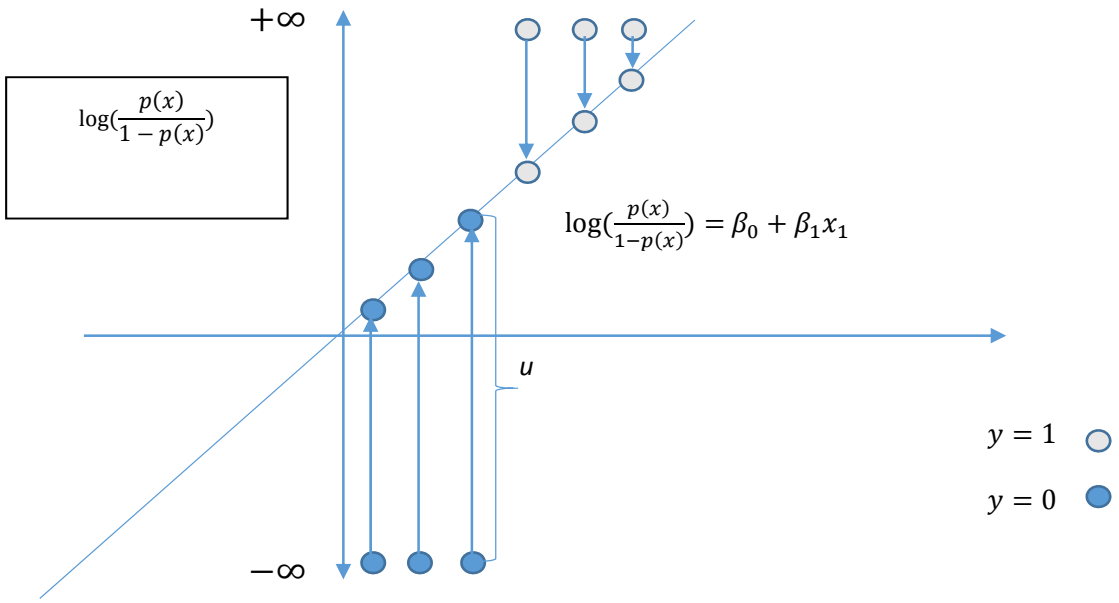
$$\frac{p(x)}{1-p(x)} = e^{\beta^T x} \tag{5}$$

Količnik $\frac{p(x)}{1-p(x)}$ predstavlja izglednost (eng. *odds*) događaja. Odnosno koliko je puta jedan događaj više izgledan od drugog. Logaritmovanjem izraza (5) postaje:

$$\log(\frac{p(x)}{1-p(x)}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m \tag{6}$$

Leva strana (6) je *logit* funkcija od $p(x)$, koja je sad linearno zavisna od x. Na ovaj način se problem predikcije, u slučaju klasifikacionih problema, ponovo svodi na model identičnog oblika, kao u slučaju linearne regresije (slika 13). Ovaj model tumačimo na način da se povećanjem vrednosti prediktora x_i za jediničnu vrednost (pretpostavljajući da su ostali prediktori konstantni), logaritam izglednosti događaja da je $y = 1$ povećava približno za β_i.

Slika 13. Logit funkcija, linerano zavisna od prediktora x_1 (model sa jednim prediktorom)



Sada slično kao u linearnoj regresiji, iterativnim postupkom, primenom odgovarajućeg optimizacijskog algoritma, dolazimo do optimalnih vrednosti parametara β, a time i optimalnog položaja granične hiperravni, koja na najbolji način razdvaja uzorke različitih klasa.

Optimalni pravac hiperravni u slučaju linearne regresije određen je primenom *OLS* optimizacijskog algoritma, tako da suma kvadrata reziduala bude minimalna. U slučaju logističke regresije, ovaj algoritam se ne može koristiti jer za svaku vrednost parametra β, reziduali imaju beskonačnu vrednost (od $\pm\infty$ do projekcije na regresionu pravu (slika 13). Zato se u slučaju logističke regresije primenjuje algoritam Maksimalne verodostojnosti (Hastie et al., 2008). Imajući u vidu *Bernoulli,* binarnu raspodelu verovatnoća zavisne promenljive, funkcija verodostojnosti ima oblik:

$$LF(\beta) = \prod_{i=1}^{n} p\left(x^{(i)}\right)^{y^{(i)}} (1 - p\left(x^{(i)}\right))^{1-y^{(i)}} \tag{7}$$

Logaritmovanjem ovog izraza dobijamo:

$$log\, LF(\beta) = \sum_{i=1}^{n} y^{(i)} log(p\left(x^{(i)}\right)) + \left(1 - y^{(i)}\right) log(1 - p(x^{(i)})) \tag{8}$$

Što je veća vrednost $log\, LF(\beta)$, to je bolja, tačnija predikcija, zbog čega se i postavlja uslov postizanja njene maksimalne vrednosti, koju dobijamo iz uslova $\frac{\partial}{\partial\, \beta_j} LF(\beta) = 0$

U slučaju jednog uzorka:

$$\frac{\partial}{\partial\, \beta_j} \log LF\,(\beta) = (y\, \frac{1}{p(x)} - (1 - y)\, \frac{1}{1 - p(x)})\, \frac{\partial}{\partial\, \beta_j}\, p(x)$$

$$= (y \frac{1}{g(\beta^T x)} - (1 - y) \frac{1}{1 - g(\beta^T x)})\, \frac{\partial}{\partial\, \beta_j} g(\beta^T x) *$$

$$= (y \frac{1}{g(\beta^T x)} - (1 - y) \frac{1}{1 - g(\beta^T x)})\, g(\beta^T x)(1 - g(\beta^T x))\, \frac{\partial}{\partial\, \beta_j} \beta^T x$$

$$= (y\big(1 - g(\beta^T x)\big) - (1 - y)\, g(\beta^T x))x_j$$

$$= (y - \ g(\beta^T x))x_j = (y - p(x))x_j \qquad (9)$$

* Prvi izvod proizvoljne *sigmoid* funkcije $g(x) = \frac{1}{1+e^{-x}}$ jednak je:

$$\frac{d}{dx} g(x) = \frac{e^{-x}}{(1+e^{-x})^2}$$

$$= \frac{1}{(1+e^{-x})} - \frac{1}{(1+e^{-x})^2} = \frac{1}{(1+e^{-x})}\left(1 - \frac{1}{(1+e^{-x})}\right)$$

$$= g(x)(1 - g(x))$$

Primećujemo da je *update rule* (9), ima sličan oblik kao u slučaju linearne regresije. Ovo proizilazi iz činjenice da linerana i logistička regresija pripadaju istoj familiji algoritama – Generalnih linearnih modela, *GLM* (Hastie et al., 2008).

Za razliku od linearne regresije, gde smo tražili minimalnu vrednost *cost* funkciije (*OLS*) metodom *gradient descent,* ovde tražimo maksimalnu vrednost funkcije Maksimalne verodostojnosti, postupkom *gradient ascent.* Tako vrednost parametra β svakom iteracijom raste, dok ne počne da konvergira ka vrednosti β, za koju $LF(\beta)$ ima maksimalnu vrednost:

$$\beta_j \coloneqq \beta_j + \alpha \frac{\partial}{\partial \beta_j} \log LF(\beta) \qquad j = (0,1,\ldots\ldots m) \qquad (10)$$

Odnosno

$$\beta_j \coloneqq \beta_j + \alpha \sum_{i=1}^{n} (y^{(i)} - p\big(x^{(i)}\big)) x_j{}^{(i)} \qquad (11)$$

Na osnovu dobijenih optimalnih vrednosti parametra β, iz jednačine (2), dobijamo konačno verovatnoću *p(x),* da je $y = 1$.

Osnovne pretpostavke modela logističke regresije su da je logit funkcija linearno zavisna od prediktora, da su prediktori međusobno nezavisni (da ne postoji multikolinearnost), da u podacima nema ekstremnih vrednosti. Kada su ovi preduslovi ispunjeni, ovim modelom se postiže zavidan tačnost predikcije u binarno klasifikacionim problemima (Porter i Gujarati, 2009).

Na sličan način kao i kod linearne regresije, kako bi se izbegla mogućnost overfitinga i nestabilnost modela usled manjeg broja uzoraka u odnosu na broj prediktora, primenjuje se postupak regularizacije (*Ridge, Lasso* ili *Elastic Ne*t).

2.3. Opšti linerni modeli, *GLM*

Modeli linerane i logističke regresije su specijalni slučajevi Opštih linearnih modela – *GLM* (*Generalized linear models*), čiji su autori Nelder i Wedderburn (1972). Primenom *GLM* generalizujemo primenu modela kojim se zavisna promenljiva prikazuje kao linearna funkcija prediktora:

$$\hat{y} = f(x;\beta) = x^T\beta \tag{1}$$

Prediktivni model dat jednačinom (1), nije isključivo ograničen na slučajeve linearne i logističke regresije, već se može primeniti uvek kada raspodela verovatnoće zavisne promenljive pripada nekoj od eksponencijalnih distribucija (McCullagh i Nelder, 1983). Transformacijom očekivanih vrednosti zavisne promenljive, primenom tzv. *link* funkcije, dobijamo prediktivni model koji postaje linearna funkcija prediktora.

Familija eksponencijalnih distribucija može se prikazati u opštem obliku (McCullagh i Nelder, 1983):

$$Pr(y;\ \eta)=b(y)\exp(\ \eta \quad T(y)-a(\eta)) \tag{2}$$

$b(y) - eng.$ base measure

$a(\eta) - eng.$ log partition funkcija

$T\,(y) - eng.$ sufficient statistics

$\eta \;-\; eng.$ natural (canonical) parameter, prirodni parameter distribucije

Svaka distribucija koja pripada eksponencijalnoj familiji jeste ona koja može biti prikazana izrazom (2).

Kako smo prethodno naveli, $Bernoulli$ i Normalna distribucija pripadaju familiji eksponencijalnih distribucija.

Kako $Bernoulli$ funkcija raspodela verovatnoće ima oblik:

$$Pr(y, p) = p^y(1-p)^{1-y}$$

transformacijom, ovaj se izraz može svesti na eksponencijalni oblik:

$$= \exp(\log(p^y(1-p)^{1-y})$$

$$= \exp(\log p^y + \log(1-p)^{1-y})$$

$$= \exp(y \log p + (1-y)\log(1-p))$$

$$= \exp(y \log p + \log(1-p) - y\, log(1-p))$$

$$= \exp(log\, \frac{p}{(1-p)} y + \log(1-p)$$

Poređenjem ovog izraza sa izrazom koji definiše eksponencijalnu familiju distribucija (2), dobijamo da je:

b(y) = 1

T(y) = y

$\eta = log\ \frac{p}{1-p}$, odnosno da je $p = \frac{1}{1+e^{-\eta}}$

a(η) = $-log(1-p) = -log\left(1-\frac{1}{1+e^{-\eta}}\right) = log(1+e^{\eta})$

Vrednosti koeficijenata *b,T* i a su karakteristične za $Bernoulli$ familiju distribucija, dok vrednost prirodnog parametra η odgovara konkretnoj distribuciji iz $Bernoulli$ familije, za koju je verovatnoća pozitivnog događaja jednaka p (McCullagh i Nelder, 1983).

Normalna distribucija zavisne promenljive y data je sledećim izrazom:

$$Pr(y,\beta) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\ (y-\mu)^2\right), \quad \text{pretpostavimo } \sigma^2 = 1$$

Transformacijom ovog izraza možemo pokazati da i normalna distribucija pripada eksponencijalnoj familiji. Odnosno gornji izraz jednak je:

$$= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}y^2\right)\exp(y\mu - \frac{\mu^2}{2})$$

Sledi da je:

$b(y) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}y^2\right)$

$T(y) = y$

$\eta = \mu$ (srednja vrednost)

$\mathrm{a}(\eta) = \frac{\mu^2}{2} = \frac{\eta^2}{2}$

Dobijene vrednosti koeficijenata *b,T* i *a* su karakteristične za familiju Normalnih distribucija, dok je konkretna distribucija koja pripada ovoj familiji jednoznačno određena svojom srednjom vrednošću μ (McCullagh i Nelder, 1983).

Pretpostavke koje treba da budu ispunjene u slučaju primene prediktivnog *GLM* modela su sledeće (McCullagh i Nelder, 1983):

1. $(y|x;\beta) \sim Exponential\ familly(\eta)$, za date vrednosti prediktora x i parametra β, zavisno promenljiva y ima eksponencijalnu distribuciju, određenu parametrom η.

2. Linearni prediktivni model $f(x;\beta)$ daje očekivanu vrednost zavisne promenljive

$$E[(y|x;\beta)] = f\ (x;\ \beta)$$

3. Prirodni parameter η i prediktori su linearno zavisni

$$\eta = \beta^T x$$

Primenom *GLM*, u slučaju predikcije vrednosti kvantitativne zavisne promenljive, normalne raspodele verovatnoće, za koju važi $(y|x;\beta) \sim N(\mu, \sigma^2)$, zaključujemo da je na osnovu pretpostavke (2):

$$f(x;\beta) = E[(y|x;\beta)] = \mu$$

$\mu -$ srednja vrednost

Kako normalna distribucija pripada eksponencijalnoj familiji, gde je $\eta = \mu$, a kako iz pretpostavke (3) sledi da je $\eta = \beta^T x$, vidimo da smo primenom *GLM* algoritma dobili model identičan modelu linearne regresije:

$$E[(y|x;\beta)] = \mu = \beta^T x$$

Slično kao primenom *GLM*, za slučaj klasifikacionih problema, gde je očekivana vrednost jednaka verovatnoći pozitivnog ishoda slučajnog događaja, na osnovu druge pretpostavke dobijamo da je:

$$f(x;\beta) = E[(y|x;\beta)] = p$$

Kako $Bernoulli$ distribucija pripada eksponencijanoj familiji, za nju važi da je $\eta = log\frac{p}{1-p}$, odnosno da je $p = \frac{1}{1+e^{-\eta}}$. Na osnovu pretpostavke (3), $\eta = \beta^T x$, sledi da je:

$$log\frac{p}{1-p} = \beta^T x$$

odnosno:

$$E[(y|x;\beta)] = p = \frac{1}{1+e^{-\beta^T x}}$$

Primenom *GLM* za slučaj binarnog klasifikacionog problema, dobili smo isti prediktivni model kao u slučaju logističke regresije. Sledi da izbor *Sigmoid* funkcije u logističkoj regresiji, kojom smo osigurali da vrednost predikcije bude u rangu 0 do 1, nije bio proizvoljan, već rezultat činjenice da logistička regresija spada u *GLM*.

Funkciju transformacije očekivanih vrednosti *GLM* modela nazivamo *link* funkcijom, g(μ), gde je μ očekivana vrednost svake od eksponencijalnih distribucija (McCullagh i Nelder, 1983). U tabeli 3 vidimo prikaz link funkcija za najznačajnije eksponencijalne distribucije.

Tabela 3. *Link* funkcije, za najčešće primenjivane eksponencijalne distribucije

Distribucija	Link funkcija $g(\mu) = \beta^T x, \quad \mu(x) = E[(y\|x;\beta)]$
Normalna	$\mu = \beta^T x$
Bernoulli	$log\frac{\mu}{1-\mu} = \beta^T x$
Poisson	$\log(\mu) = \beta^T x$
Exponential	$-\mu^{-1} = \beta^T x$
Gamma	$-\mu^{-1} = \beta^T x$

Opšti oblik *GLM* modela možemo napisati kao

$$g(\mu) = \eta(x) = \beta^T x$$

$\mu(x) = E[(y|x;\beta)]$, očekivana vrednost odgovarajuće distribucije

Primenom *GLM* generalizuje se linearni model, tako da se može primeniti u slučajevima kada zavisna promenljiva ima neku od eksponencijalnih distribucija, primenom odgovarajuće *link* funkcije, kojom se očekivana vrednost zavisne promenljive transformiše.

2.4. *Gausova* diskriminantna analiza, *GDA*

Gausova diskriminantna analiza, *GDA,* jeste generativni klasifikacioni algoritam koji se primenjuje u klasifikacionim problemima kada zavisna kategorička promenljiva ima više od dve klase. Za razliku od logističke regresije, gde smo direktno modelovali uslovnu raspodelu verovatnoće da zavisna promenljiva pripada klasi k, $Pr(y = k|x)$, u slučaju *GDA* do ove verovatnoće dolazimo na indirektan način. Kako se radi o generativnom algoritmu, modeluje se distribucija prediktora za

različite klase zavisne promenljive, $Pr(x|y=k)$, pa primenom *Bajesove* teoreme dolazimo do tražene $Pr(y=k|x)$:

$$Pr(y=k|x) = \frac{Pr(x|y=k)\Pr(y=k)}{\sum_{l=1}^{k} Pr(x|y=l)\Pr(y=l)} \quad (1)$$

k, klasa promenljive y

Do vrednosti prethodne (eng. *prior*) verovatnoće $\Pr(y=k)$, koja predstavlja verovatnoću da slučajni uzorak pripada klasi *k*, dolazimo na jednostavan način iz trening podataka:

$$Pr(y=k) = \frac{n_k}{n}$$

n_k, broj uzorka klase k

Kako bismo došli do distribucije verovatnoće $Pr(x|y=k)$, pretpostavljamo multivarijantnu normalnu distribuciju prediktora ($X = \{x_i \epsilon R^m\}$) za svaku klasu zavisne promenljive. Tako se podrazumeva da svaki prediktor pojedinačno ima normalnu raspodelu, dok između parova prediktora postoji neki stepen korelacije. Parametri *Gaussian* multivarijantne distribucije su vektor srednjih vrednosti prediktora $\mu_k \epsilon R^m$ i kovarijantna matrica $\varepsilon_k \epsilon R^{mxm}$. Dijagonalni elementi ove matrice su varijanse, a vandijagonalni elementi kovarijanse prediktora.

GDA algoritam postoji u dva oblika, za koje važe različite pretpostavke za kovarijantnu matricu ε_k. Kada podrazumevamo da je ova matrica različita za uzorke različitih klasa (manje restriktivna pretpostavka), govorimo o Kvadratnoj diskriminatornoj analizi (eng: *Quadratic dicriminante*

analysis – *QDA*, Gareth et al,, 2014). Funkcija raspodele verovatnoće prediktora u slučaju normalne multivarijantne distribucije tada ima oblik:

$$Pr(x|y=k;\mu_k\varepsilon_k) = \frac{1}{(2\pi)^{\frac{m}{2}}|\varepsilon_k|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu_k)^T{\varepsilon_k}^{-1}(x-\mu_k)\right) \quad (2)$$

Kada pretpostavimo da je kovarijantna matrica prediktora ista za uzorke različitih klasa, govorimo o Linearnoj diskriminatornoj analizi (eng: *Linear discriminatne analysis – LDA*, Gaber et al, 2017). Ako je zavisna promenljiva binarna, kategorija 0 i 1, na osnovu pretpostavke o istim kovarijantnim matricama, imamo sledeća dva oblika multivarijantne distribucije:

$$\Pr(x|y=0;\mu_0,\varepsilon) = \frac{1}{(2\pi)^{\frac{m}{2}}|\epsilon|^{\frac{1}{2}}}\exp(-\frac{1}{2}(x-\mu_0)^T\varepsilon^{-1}(x-\mu_0))$$

$$\Pr(x|y=1;\mu_1,\varepsilon) = \frac{1}{(2\pi)^{\frac{m}{2}}|\epsilon|^{\frac{1}{2}}}\exp(-\frac{1}{2}(x-\mu_1)^T\varepsilon^{-1}(x-\mu_1))$$

$\mu_0 -$ vektor srednjih vrednosti prediktora klase y $= 0$, $\mu_0 \epsilon R^m$

$\mu_1 -$ vektor srednjih vrednosti prediktora klase y $= 1$, $\mu_1 \epsilon R^m$

$\varepsilon \quad -$ jedinstvena kovarijantna matrica prediktora R^{mxm}

Novi testni uzorak biće dodeljen klasi za koju je brojilac iz *Bajesove* formule (1), odnosno proizvod $Pr(x|y=k)Pr(y=k)$,najveći.Imenilac ovog količnika je konstantan. Tražimo dakle maksimalu vrednosti sledećeg izraza za različite klase k:

$$\frac{1}{(2\pi)^{\frac{m}{2}}|\varepsilon|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu_k)^T\varepsilon^{-1}(x-\mu_k)\right) Pr(y=k) \quad (3)$$

Kako $\frac{1}{(2\pi)^{\frac{m}{2}}|\varepsilon|^{\frac{1}{2}}}$ ne zavisi od *k*, zamenićemo ovaj količnik konstantom *C*

$$= C \exp\left(-\frac{1}{2}(x-\mu_k)^T\varepsilon^{-1}(x-\mu_k)\right) Pr(y=k)$$

logaritmovanjem dobijamo

$$= logC - \frac{1}{2}(x-\mu_k)^T\varepsilon^{-1}(x-\mu_k) + \log\big(Pr(y=k)\big)$$

kako je $logC$ konstanta, možemo je izostaviti

$$= -\frac{1}{2}[x^T\varepsilon^{-1}x - {\mu_k}^T\varepsilon^{-1}x - x^T\varepsilon^{-1}\mu_k + {\mu_k}^T\varepsilon^{-1}\mu_k] + log\big(Pr(y=k)\big)$$

Iz ${\mu_k}^T x = x^T\mu_k$, dalje sledi

$$= log\big(Pr(y=k)\big) - \frac{1}{2}[x^T\varepsilon^{-1}x + {\mu_k}^T\varepsilon^{-1}\mu_k] + x^T\varepsilon^{-1}\mu_k$$

$x^T\varepsilon^{-1}x$ ne zavisi od *k*

$$LDF_k = log(Pr(y=k)) + x^T\varepsilon^{-1}\mu_k - \frac{1}{2}\mu_k^T\varepsilon^{-1}\mu_k \quad (4)$$

Ovaj izraz nazivamo linearnom diskriminantnom funkcijom, LDF_k (Gareth et al., 2014). Testni ili novi uzorak pripadaće klasi za koju LDF_k ima veću vrednost.

U slučaju *QDA* algoritma, uzorak će biti dodeljen klasi za koju *quadratic dicriminant function QDF* ima maksimalnu vrednost. Po analogiji sa *LDA,* imajući u vidu različite kovarijantne matrice uzoraka različitih klasa, tražimo maksimalnu vrednost izraza:

$$C|\varepsilon_k|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu_k)^T {\varepsilon_k}^{-1}(x-\mu_k)\right) Pr(y=k)$$

$C = \frac{1}{(2\pi)^{\frac{m}{2}}}$

Izostavljanjem konstante C i logaritmovanjem, prethodni izraz postaje:

$$QDF_k = -\frac{1}{2}(x-\mu_k)^T {\varepsilon_k}^{-1}(x-\mu_k) - \frac{1}{2} log|\varepsilon_k| + \log\big(Pr(y=k)\big)$$

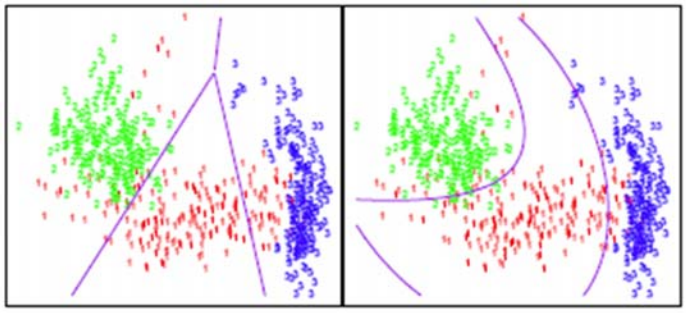
Odnosno

$$QDF_k = -\frac{1}{2}x^T\varepsilon_k^{-1}x + x^T\varepsilon_k^{-1}\mu_k - \frac{1}{2}{\mu_k}^T\varepsilon_k^{-1}\mu_k - \frac{1}{2}log|\varepsilon_k| + \log(Pr(y=k) \quad (5)$$

Testni uzorak biće dodeljen klasi za koju QDF_k im najveću vrednost (Gareth et al., 2014).

QDA, kao podgrupa *GDA,* smatra se fleksibilnijom metodom u odnosu na *LDA*, iz razloga da pretpostavka jednakih kovarijantnih matrica prediktora za različite klase ne važi. Granične linije između klasa, određene funkcijom QDF_k nisu prave linije (kao u slučaju *LDA*) (slika 14). Ove granične linije se još i nazivaju Bayes granicama odlučivanja (James et al., 2013).

Slika 14. Granične linije za *LDA* i *QDA*[8]



LDA *QDA*

U slučaju manjeg broja uzoraka, bolja tačnost predviđanja postiže se primenom *LDA*, jer je za 'učenje' ovog modela potrebna manja količina podataka u odnosu na *QDA*.

Za primenu *LDA* i *QDA* modela potrebno je da broj prediktora bude znatno manji od broja uzoraka. Ukoliko taj uslov nije ispunjen, smanjuje se tačnost predikcije. Iskustveno pravilo je da se *LDA* i *QDA* primenjuju u slučajevima kada je ispunjen uslov $n \geq 5\,m$ [9].

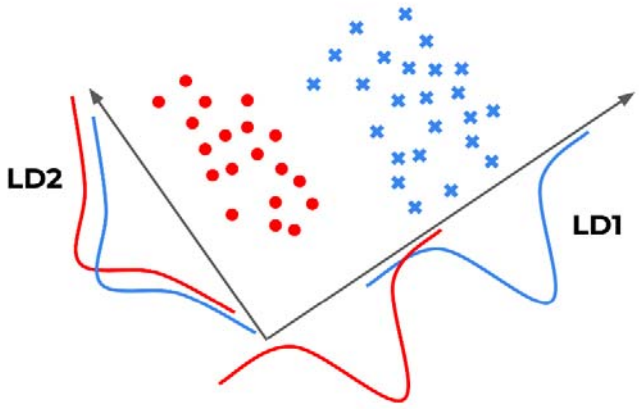
U poređenju sa logističkom regresijom, oba modela su stabilnija u slučajevima kada je moguća separacija uzoraka različitih klasa. Ovi generativni algoritmi koriste se uvek u klasifikacionim problemima, kada je broj klasa zavisne kategoričke promenljive veći od dva (eng. *multiclass problem*).

Pored rešavanja klasifikacionih problema, *LDA* se koristi u cilju smanjivanja dimenzionalnosti prediktivnog modela. Prediktori, kao tačke u originalnom *m* dimenzionalnom prostoru ($x \in R^m$), transformišu se u novi tzv. *LD* prostor, određen koordinantnim osama $LD_1, .., LD_m$, (eng. *linear desriptors – LD*), tako da se maksimalna separacija uzoraka različitih klasa poststiže u pravcu LD_1, a najmanja u pravcu LD_m (slika 15) (Gaber et al., 2017). Konačan broj *linearnih deskriptora* ograničava se samo na one koji određuju pravce najveće separacije podataka različitih klasa

[8] Izvor: James et al., 2013.
[9] Izvor: Linear & Quadratic Discriminant Analysis UC Business Analytics R Programming Guide (uc-r.github.io)

(pravce minimalne disperzije uzoraka unutar klase i maksimalne između klasa), dok se *LD* pravci minimalne separacije mogu zanemariti.

Slika 15. Uzorci u originalnom R^2prostoru određeni su koordinatama prediktora x_1, x_2. Novi *LD* prostor određen je osama maksimalne – $LD_1 = f(x_1, x_2)$ i minimalne – $LD_2 = f(x_1, x_2)$ separacije uzoraka. Prediktivni model tako može biti predstavljen kao $\hat{y}_i = \hat{f}(LD_1)$, čime je smanjena dimenzionalnost prediktivnog modela za jednu varijablu.



Pravci linearnih deskriptora određuju se u tri koraka.

Prvo se računa separabilnost uzoraka različitih klasa (eng. *between class variance* – Sb) merenjem razdaljine srednjih vrednosti uzoraka svake klase, od srednje vrednosti svih uzoraka (Gaber et al., 2017). Što je veća vrednost Sb, mogućnost separacije uzoraka različitih klasa je bolja.

$$\text{Sb} = \sum_{i=1}^{k} n_i(\mu_i - \mu)(\mu_i - \mu)^{\text{T}}$$

$\mu_i \in R^m$, vektor srednjih vrednosti uzoraka klase i

$\mu \in R^m$, vektor srednjih vrednosti svih uzoraka

k – broj klasa

n_i – broj uzoraka klase *i*

Sb – matrica $m\ x\ m$

U drugom koraku računamo varijansu uzoraka svake klase (eng. *within class variance* – Sw) (Gaber, et al., 2017):

$$\text{Sw} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \mu_i)\,(x_{ij} - \mu_i)^{\text{T}}$$

$n_i - broj\ uzoraka\ klase\ i$

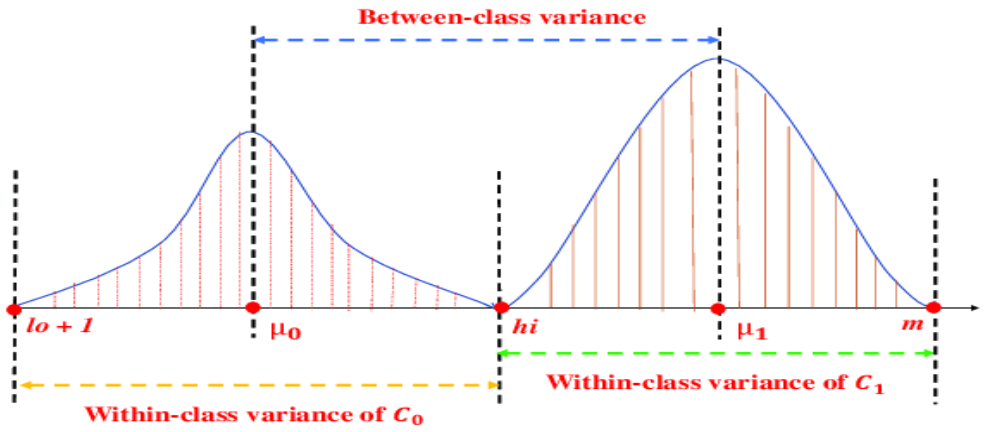
$\mu_i \in R^m$, vektor srednjih vrednost uzoraka klase i

$x_{ij} \in R^m$, j uzorak klase i

Sw - matrica $m\ x\ m$

LD pravci se određuju iz uslova postizanje maksimalne vrednosti Sb i minimalne Sw. Nakon projekcije uzoraka na tako određene pravce, postiže se maksimalno razgraničavanje uzoraka različitih klasa.

Slika 16. Sw i Sb, za uzorke koje pripadaju različitim klasama C_0 i C_1 [10]



Ako sa *V* označimo matricu transformacije, kojom se vektori uzoraka projektuju u novi *LD* prostor, onda je projekcija vektora uzoraka i vektora srednjih vrednosti jednaka $V^T x_{ij}$ i $V^T \mu_i$. Iz ovog dalje sledi da $(x_{ij} - \mu_i)^2$, posle projekcije u novi *LD* prostor, postaje:

$$\left(V^T(x_{ij} - \mu_i)\right)^T \left(V^T(x_{ij} - \mu_i)\right) = \left((x_{ij} - \mu_i)^T V\right)^T \left((x_{ij} - \mu_i)^T V\right)$$
$$= V^T(x_{ij} - \mu_i)(x_{ij} - \mu_i)^T V$$

Sledi da je *Sw*, posle projekcija vektora uzoraka x_i i srednjih vrednosti μ_i, na *LD* jednak:

$$\widehat{Sw} = V^T SwV$$

Po istoj analogiji dobijamo da je

[10] Scientific Diagram (researchgate.net)

$$\widehat{Sb} = V^T SbV$$

Transformacijsku matricu *V,* koja određuje *LD* pravce*,* dobijamo iz uslova koji se naziva *Fisher* kriterijum:

$$J(V) = \arg max_V \frac{\widehat{Sb}}{\widehat{Sw}} = \frac{V^T SbV}{V^T SwV} \tag{6}$$

Iz uslova da je izvod $\frac{d}{dV}[J(V) = 0]$ sledi:

$$(V^T S_w V)2S_b V - (V^T S_b V)2S_w V = 0$$

$$SbV = \lambda \mathrm{Sw}V, \text{ gde je } \lambda = \frac{V^T S_b V}{V^T S_w V}$$

$$Sw^{-1}SbV = \lambda V \tag{7}$$

Jednačina (7) predstavlja tzv. problem sopstvenih vrednosti (eng. *eigenvalue*) (Gaber et al., 2017), čija su rešenja sopstvene vrednosti $\lambda = \{\lambda_1, \lambda_2, .., \lambda_k, .., \lambda_m\}$ i sopstveni vektori $V = \{v_1, \dots, v_k, \dots, v_m\}$ (eng. *eigenvectors*), matrice $S_w^{-1} S_b$ (Gaber, 2017). Iz (7) dalje sledi:

$$(Sw^{-1}Sb - \lambda \mathrm{I})V = 0$$

Iz ovog uslova dobijamo sopstvene vrednosti λ, jediničnih vektora:

$$det(Sw^{-1}Sb - \lambda \mathrm{I}) = 0$$

Sa poznatim vrednostima λ, iz (7) dobijamo sopstvene vektore V.

Sopstveni vektori određuju pravce novih osa u novom *LDA* prostoru. Pravac LD_1 određen je sopstvenim vektorom $v_{k,}$, čija je odgovarajuća sopstvena vrednost najveća, $\lambda_k = \max\{\lambda_1, \lambda_2, \ldots, \lambda_m\}$. Preostali sopstveni vektori $\{v_{k+1,} \ldots, v_m\}$ se zanemaruju, odnosno samo sopstveni vektori $V_k = \{v_1 \ldots, v_k\}$, $V_k \epsilon R^{m\,x\,k}$ se koriste pri određivanju novog *LDA* prostora. Na ovaj način smo početni prostor $X \epsilon R^{n\,x\,m}$ transformisali u novi *LDA* manje dimenzionalnosti $R^{n\,x\,k}$, gde je k<m. Sad se prediktivni model može predstaviti kao funkcija linearnih deskriptora $LD_1, \ldots, LD_{k.}$

2.5. Klasične statističke metode

Najviše primenjivane statističke metode u problemima procena kreditnih rizika i verovatnoće stečaja su multivarijantna dsikriminantna analiza (*MDA*) i logistički regresioni modeli (Balcaen i Ooghe, 2006). Chen (2011), upoređuje rezultate predikcije stečaja *accounting-based* modelima: Altman – MDA (1968), Zmijewski – probit (1984) i Ohlson – logit (1980), sa kako ih autor zove 'inteligentnim' *ML* tehnikama. Zaključuje da je, pored činjenice da se primenom *ML* može postići veća tačnost predikcije, značaj statističkih metoda u njihovom jednostavnom tumačenju i mogućnosti primene kada se raspolaže manjim količinama podataka.

2.5.1. *Altman* Z-score

NYU Stern, profesor finansija Edward Altman je primenom diskriminantne analize razvio model kojim se opisuje uticaj finansijskih indikatora poslovanja na verovatnoću stečaja kompanije, u periodu od dve godine. Nastao je 1967, kao rezultat analize finansijskih izveštaja 66 kompanija, od kojih je polovina bila u stečaju. Radi se o proizvodnim kompanijama, čija vrednost aktive nije manja od 1mil $. Do 1975. godine analizirano je dodatnih 70 kompanija, a u periodu od 1996. do 1999. još njih 120, kada je ovaj linearni model dobio svoj konačni oblik.

$$\text{Altman Z-Score} = 1.2A + 1.4B + 3.3C + 0.6D + 1.0E$$

A = *working capital / total assets*

B = *retained earnings / total assets*

C = *earnings before interest and tax / total assets*

D = *market value of equity / total liabilities*

E = *sales / total assets*

U zavisnosti od dobijene Z- Score vrednosti, može se utvrditi izglednost stečaja (tabela 4)

Tabela 4. Granične Z- Score vrednosti

Z Score		
Z<1.8	1.8<Z<3	Z>3
Izgledan stačaj	'siva' zona	Ne postoji rizik od stečaja

U 2012. godini novi Z- Score model razvijen je za neproizvodne i kompanije sa tržišta u razvoju:

Z-Score za neproizvodne kompanije:

$$\text{Altman Z-Score} = 6.56A + 3.26B + 6.72C + 1.05D$$

Z-Score kompanija sa tržišta u razvoju:

$$\text{Altman Z-Score} = 3.25 + 6.56A + 3.26B + 6.72C + 1.05D$$

D = *book value of equity / total liabilities*

Tabela 5. Granične Z- Score vrednosti za neproizvodne firme i tržišta u razvoju

Z- Score		
Z<1.1	1.1<Z<2.6	Z>2.6
izgledan stečaj	'siva' zona	ne postoji rizik od stečaja

Prema Frydman et al., (1985), Altman MDA je najoptimalniji model za predikciju stečaja jer uključuje najznačajnije finansijske indikatore poslovanja.

2. 5. 2. *Zmijewski* model

Takođe je baziran na finansijskim indikatorima poslovanja. Nastao je 1984. godine, analizom finansijskih izveštaja o poslovanju 840 USA kompanija, od kojih je 40 bilo u stečaju.

$$Zm = -4.3 - 4.5X_1 + 5.7X_2 + 0.004X_3$$

$$X_1 = \frac{net\ revenue}{total\ assets}$$

$$X_2 = \frac{total\ debt}{total\ assets}$$

$$X_3 = \frac{current\ assets}{current\ liabilities}$$

$$\Pr(stečaja) = \frac{1}{1 + e^{-zm}}$$

Za vrednost $\Pr(stečaja)$ veće od 0.5 predviđa se stečaj.

2.5.3 *Ohlson* O-Score model

NYU profesor James Ohlson je 1980. na osnovu analize 105 kompanija u stečaju i 2. 058 koje to nisu, došao do sledećeg modela:

$$T = -1.32 - 0.407\log(\frac{TA_t}{GNP}) + 6.03\frac{TL_t}{TA_t} - 1.43\frac{WC_T}{TA_T} + 0.0757\frac{CL_t}{CA_t}$$
$$-1.72X - 2.37\frac{NI_t}{TA_t} - 1.83\frac{FFO_t}{TL_t} + 0.285Y - 0.521\frac{NI_t - NI_{t-1}}{|NI_t| + |NI_{t-1}|}$$

TA = *total assets*

WC = *working capital*

FFO = *funds from operations*

CA = *current assets*

CL = *current liabilities*

TL = *total liabilities*

GNP = *Gross product price index US*

X=1 , if TL>TA else X=0

Y=1, ako je net prihod za poslednje dve godine bio negative, ako ne 0

NI = net prihod

$$\Pr(stečaja) = \frac{1}{1 + e^{-T}}$$

Za vrednost $\Pr(stečaja)$ veće od 0.5, predviđa se stečaj.

3. MODELI MAŠINSKOG UČENJA

U disertaciji su analizirani najčešće primenjivani parametarski/neparametarski, linearni/nelinearni *ML* klasifikatori. Prema načinu kako se određuje vrednost target kategoričke promenljive, odnosno uslovne verovatnoće da ispitivani uzorak pripada jednoj od klasa zavisne promenljive, klasifikatore možemo još svrstati u diskriminativne i generativne.

Diskriminativni (eng. *Discriminative*) klasifikacioni algoritmi su oni kod kojih se na direktan način dolazi do uslovne verovatnoće $Pr(y|x)$, verovatnoće da uzorak x pripada jednoj od klasa zavisne promenljive. Ne pretpostavlja se distribucija verovatnoće promenljive x za različite klase y ($Pr(x|y)$). Kod ovih klasifikacionih algoritama iterativnim postupkom se pretpostavljaju različite granične vrednosti x^* (eng. *threshold value*), koje određuju položaj hiperravni u prostoru R^m, koja na najbolji način razdvaja uzorke različitih klasa. Optimalna vrednost x^* je ona za koju je broj pogrešno klasifikovanih uzoraka najmanja. Ispitivani uzorak se tako dodeljuje klasi u odnosu na njegov položaj prema graničnoj hiperravni.

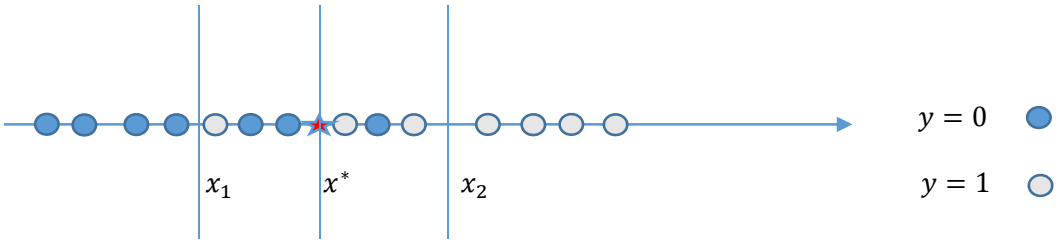
U primeru zavisne kategoričke promenljive sa dve klase, $y = \{0,1\}$ i jednim prediktorom, ovaj se uslov može napisati na sledeći način:

$$Pr(y = 0|x) > Pr(y = 1|x) \ za\ x < x^*$$

$$Pr(y = 1|x) > Pr(y = 0|x) \ za\ x > x^*$$

Slika 17. Optimalni položaj granične tačke određen je sa x^*, za koju je broj misklasifikovanih uzoraka najmanji. Za vrednost x_1 broj pogrešno klasifikovanih uzoraka je 3, za x_2 je 3, a za x^* je 2.

.



U slučaju generativnih (eng. *Generative*) algoritama, uzorak se dodeljuje klasi za koju je zajednička verovatnoća $Pr(x, y)$ najveća. Kako su x i y zavisne, na osnovu multiplikativne teorije verovatnoće sledi da je:

$$argmax_y \Pr(x, y) = argmax_y Pr(x|y) \Pr(y)$$

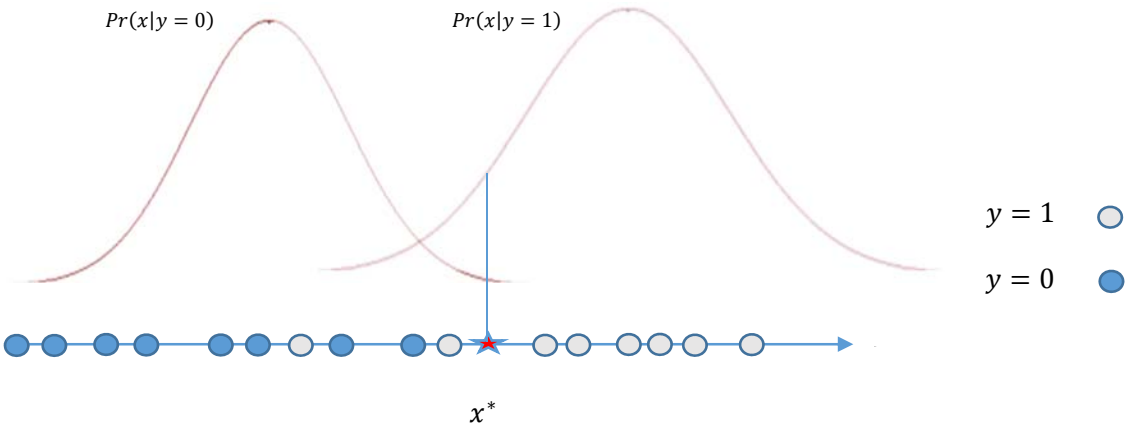
U slučaju dve klase imamo da je:

$$\Pr(x, y = 0) = \Pr(x|y = 0)\, \Pr(y = 0)$$

$$\Pr(x, y = 1) = \Pr(x|y = 1) \Pr(y = 1)$$

Kako je za novi uzorak x^*, $Pr(x^*|y = 1) > \Pr(x^*|y = 0)$, dok je $\Pr(y = 0) = \Pr(y = 1)$, sledi da će uzorak x^* biti klasifikovan kao $y = 1$ (slika 18).

Slika 18. Normalna distribucija uzoraka x, klase 0 i 1. Novi uzorak x^* dodeljuje se klasi za koju je $Pr(x^*|y)$ najveće



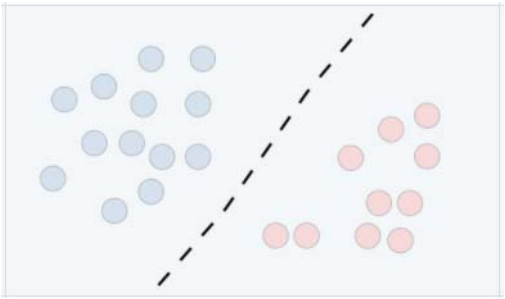
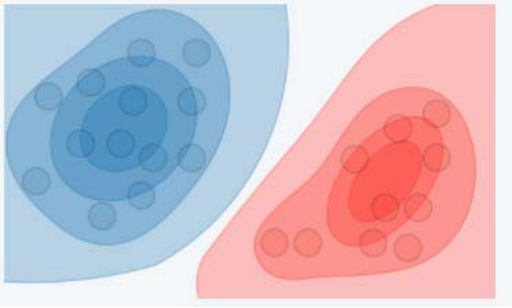
Do verovatnoće $Pr(y|x)$ možemo doći i primenom *Bayes* teoreme (T. Bayes, 1702-1761) na indirektan način:

$$Pr(y|x) = \frac{Pr(x|y)Pr(y)}{Pr(x)}$$

Sledi da će x takođe biti dodeljen klasi za koju $Pr(x|y)Pr(y))$ ima najveću vrednost:

$$argmax_y \, Pr(x|y)Pr(y)$$

Tabela 6. Diskriminativni i generativni *ML* modeli

	Diskriminativni model	Generativni model
Pristup	Direktan, Pr$(y\|x)$	Indirektan Pr$(x\|y)$, potom sledi Pr$(y\|x)$
Postupkom učenja *ML* modela dobijamo	Graničnu liniju (hiperravan)	Distribuciju verovatnoće podataka po klasama
Prikaz		
ML algoritmi	Logistička regresija Stablo odlučivanja K-NN Random Forest SVM	Naive Bayes Gaussian discriminant analysis

3.1. Naive Bayes, *NB*

NB je generativni *probabilistic* klasifikacioni algoritam, koji spada u manje kompleksne *ML* algoritme, zbog čega ima široku primenu. Njegova relativna jednostavnost proizilazi iz pretpostavke da su prediktori modela međusobno nezavisni. Iz razloga da ovaj uslov (*naivan*) u najvećem broju slučajeva nije ispunjen, kao i činjenice da je baziran na *Bajesovoj* teoremi, ovaj algoritam je nazvan *Naive Bayes.*

Iz ove osnovne pretpostavke sledi da je uslovna verovatnoća da uzorak pripada klasi *k*, na osnovu multiplikativne teorije verovatnoće, jednaka (Shai, 2014):

$$\Pr(x|y = k) = \prod_{i=1}^{m} \Pr(x_i|y = k) \quad (1)$$

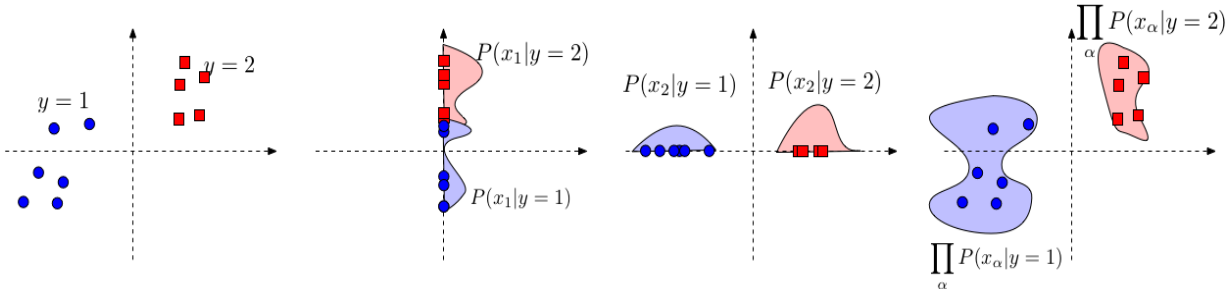
$x = (x_1, x_2, \dots x_m)$ – vektor prediktora

U slučaju dva prediktora x_1 i x_2, njihova zajednička distribucija verovatnoće jednaka je proizvodu pojedinačnih. Primenom izraza (1) dobijamo:

$$\Pr(x_1, x_2|y = k) = \Pr(x_1|y = k)\Pr(x_2|y = k)$$

k – klasa zavisne promenljive y

Slika 19. Zajednička distribucija prediktora x_1, x_2 jednaka je proizvodu njihovih pojedinačnih distribucija (Alboukadel, 2017*)*



NB kao generativni algoritam baziran je na *Bajesovoj* teoremi:

$$Pr(y = k|\, x_1, x_2, \ldots, x_m) = \frac{Pr(x_1, x_2, \ldots, x_m|y = k)Pr(y=k)}{Pr(x_1, x_2, \ldots, x_m)} \tag{2}$$

Brojilac jednačine (2) predstavlja zajedničku verovatnoću međusobno zavisnih događaja $\Pr(x_1, x_2, \ldots, x_m, y = k)$:

$$\begin{aligned}\Pr(x_1, x_2, \ldots, x_m, y = k) &= \Pr(x_1 I\, x_2, \ldots, x_m, y = k)\Pr\big(x_2, \ldots, x_m, y = k\big) \\ &= \Pr(x_1 I\, x_2, \ldots, x_m, y = k)\Pr(x_2 I\, x_3, .., x_m, y = k)\Pr(x_3, \ldots, x_m, y = k) \\ &= \Pr(x_1 I\, x_2, \ldots, x_m), y = k)\Pr((x_2 I\, x_3, .., x_m), y = k)\Pr\big(x_3 I x_4, .. x_m, y = k\big) \\ &\quad \Pr(x_4, \ldots, x_m, y = k)\big) \\ &= \Pr(x_1 I\, x_2, \ldots, x_m, y = k)\Pr(x_2 I\, x_3, .., x_m), y = k)\Pr(x_3 I x_4, .. x_m, y = k\,) \ldots. \\ &\quad \Pr(x_{m-1}\, I\, x_m, y = k)\Pr(x_m I\, y = k)\Pr(y = k)\end{aligned} \tag{3}$$

Na osnovu *NB* pretpostavke o nezavisnosti prediktora sledi de je:

$$\Pr\big(x_j\big|x_{j+1}, x_{j+2}, \ldots, x_m, y = k\big) = \Pr(x_j|y = k) \tag{4}$$

Iz uslova (4), izraz (3) je jednak:

$$\Pr(x_1, x_2, \dots, x_m, y = k) = \Pr(y = k) \prod_{j=1}^{m} \Pr(x_j | y = k) \quad (5)$$

Tako iz (5) jednačina (2) dobija svoj konačan oblik:

$$Pr(y = k | \, x_1, x_2, \dots, x_m) = \frac{\Pr(y=k) \prod_{j=1}^{m} \Pr(x_j | y=k)}{Pr(x_1, x_2, \dots, x_m)} \quad (6)$$

Novi uzorak će biti dodeljen klasi za koju će $Pr(y = k \mid x_1, \dots, x_m)$ imati maksimalnu vrednost. Ovaj uslov je ispunjen za maksimalnu vrednost brojioca jednačine (6):

$$argmax_y \ \Pr(y = k) \prod_{j=1}^{m} Pr(x_j | y = k) =$$

$$argmax_y \ \sum_{j=1}^{m} log \Pr(x_j | y = k) + log \Pr(y = k)$$

Do vrednosti $Pr(y = k$) dolazimo na jednostavan način:

$$\Pr(y = k) = \frac{\sum_{i=1}^{n} I(y_i)}{n} \ , \ 1 < k < \text{ukupan broj klasa}$$

$I(y_i) -$ ima vrednost jedan za uzorke klase k

Vrednost $Pr(x_j | y = k)$, odnosno distribuciju verovatnoća prediktora za različite klase zavisne promenljive, određujemo u zavisnosti od tipa prediktora, pa prema tome razlikujemo tri *NB* modela (Advait, 2020):

Gaussina Naive Bayes, u slučaju kada su prediktori kvantitativne promenljive, pretpostavlja se njihova normalna distribucija:

$$Pr(x|y=k) = \frac{1}{\sqrt{2\pi\sigma_k{}^2}}\, e^{-\frac{(x-\mu_k)^2}{2\sigma_k{}^2}}$$

$\mu_{k,}$ srednja vrednost prediktora klase *k*

σ_k, varijansa prediktora klase *k*

Bernoulli Naive bayes prediktori su binarna kategoričke promenljiva, njena diskretna raspodela verovatnoće jednaka je:

$$Pr(x|y=k) = \prod_{i=1}^{m} p_{ki}^{x_i}(1-p_{ki})^{1-x_i}$$

Ovaj model se najčešće primenjuje pri klasifikaciji dokumenata prema sadržaju. Binarna vrednost prediktora x_i (npr. da–ne) označava da li u dokumentu postoji *i*-ta reč, sa liste od *m* elemenata – reči (eng. *bag of words*), dok je p_{ki} verovatnoća da klasa dokumenata *k* sadrži datu reč. Unapred definisana lista od *m* reči sastavljena je tako da najbolje karakteriše različite vrste dokumenata (Singh et al., 2019)

Dokument će tako biti dodeljen klasi dokumenta za koju je zajednička verovatnoća najveća:

$$argmax_y \ \Pr(y=k) \prod_{i=1}^{m} p_{ki}^{x_i}(1-p_{ki})^{1-x_i}$$

Multinominalni Naive Bayes, za razliku od *Bernoulli NB*, vrednost kategoričke promenljive nije binarna, već prediktori pokazuju učestalost pojavljivanja nezavisnog događaja (reči u dokumentu). Ovaj algoritam se takođe najčešće primenjuje u problemima klasifikacije dokumenata. Tako svaki uzorak predstavlja histogram, u kojem vrednost, npr. x_i, predstavlja frekvenciju pojavljivanja događaja *i* (*i*-te reči sa liste od *m*) u dokumentu:

$$Pr(x|y=k) = \frac{(\sum_{i=1}^{m} x_i)!}{\prod_{i=1}^{m} x_i\,!} \prod_{i=1}^{m} p_{ki}^{x_i}$$

Dokument se dodeljuje klasi za koju je:

$$argmax_y \ \Pr(y=k) \frac{(\sum_{i=1}^{m} x_i)!}{\prod_{i=1}^{m} x_i!} \prod_{i=1}^{m} p_{ki}^{x_i}$$

Kada su prediktori kategoričke promenljive, može se pojaviti slučaj nazvan '*Zero frequency problem*' (Alboukadel, 2017*),* kada za neku klasu zavisne promenljive ne postoji nijedan uzorak sa jednom od klasa prediktora.

Tabela 7. Frekvencija pojavljivanja prediktora $\mathrm{x_j} = \{\mathrm{yes, no}\}$ i zavisne promenjive $\mathrm{y} = \{0,1\}$. Sledi da je $Pr(x_j = no|y = 1) = 0$, što je malo verovatan slučaj u populaciji iz koje su ovi podaci uzorkovani

	y =1	*y*=0
$x_j = yes$	5	4
$x_j = no$	0	1

U ovakvim slučajevima najjednostavnije je povećati broj pojavljivanja klasa prediktora i klasa zavisne promenljive za jedan (tabela 8).

Tabela 8.

	y=1	y=0
$x_j = yes$	5 (+1)	4 (+1)
$x_j = no$	0 (+1)	1 (+1)

Primenom *NB* bolji rezultati se postižu u slučaju kategoričkih prediktora u odnosu na numeričke. Razlog tome je što pretpostavka u normalnoj distribuciji numeričkih prediktora najčešće nije ispunjena (Advait, 2020). *NB* algoritam ne treba primenjivati u slučajevima kada uzorci različitih klasa nisu jasno razdvojeni.

3.2. K – Najbližih suseda, *K-NN*

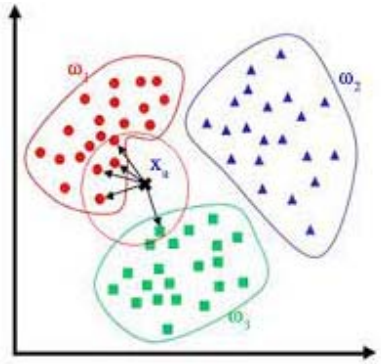
*K – najbližih suseda (*eng. *K- nearest neighbor*) je diskriminativni nadgledani *ML* algoritam, koji se može primenjivati u slučaju klasifikacionih i regresionih problema. Spada u grupu neparametarskih (eng. *non parametric)* modela, tako da ne pretpostavlja oblik funkcije zavisnosti prediktora i target promenljive, $y \approx \hat{y} = f(x)$.

Kada se *K-NN* primenjuje u slučaju klasifikacionih problema, odlučivanje je bazirano na sličnosti uzoraka (Alboukadel, 2017). Algoritam bira *K* broj uzoraka najsličnijih novom uzorku, za koji želimo predvideti kojoj klasi zavisne promenljive pripada. Većinska klasa kojoj *K* broj najsličnijih uzoraka pripada jeste klasa novog uzorka. Ovaj način klasifikacije naziva se većinsko odlučivanje (eng. *majority voting*) (Harrington, 2106).

U slučaju regresionih problema, predviđena vrednost zavisne promenljive novog uzorka jednaka je srednjoj vrednosti zavisne promenljive, *K* najsličnijih uzoraka (Alboukadel, 2017).

Kako se uzorak može predstaviti tačkom u *m* dimenzionalnom prostoru (*m* je broj prediktora), sličnost uzoraka se određuje na osnovu njihovog međusobnog rastojanja. Uzorci se smatraju sličnim ako je njihovo rastojanje malo.

Slika 20. Novi uzorak x_a se dodeljuje klasi kojoj većina od K=5 suseda pripada, klasi w_1



Izbor *K* najbližih, a time i najsličnijih suseda, najčešće se određuje primenom *Euclidean* formule (Harrington, 2016). Kako je svaki uzorak određen vektorom u *m* dimenzionalnom prostoru $x_i \epsilon R^m$, tada je *Euclidean* rastojanje *d*, prava linija koja spaja tačke x_i i x_j jednako:

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^{m}(x_{i,l} - x_{j,l})^2}$$

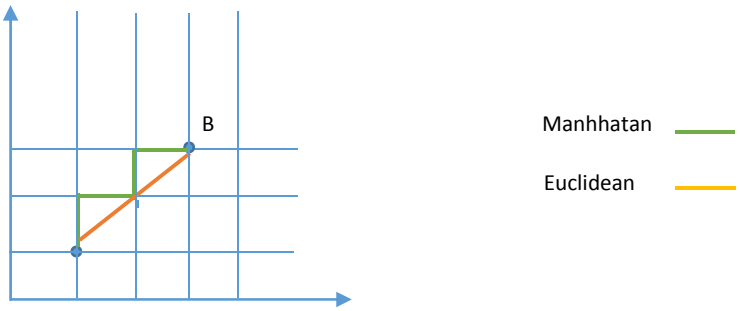
K-NN je u slučaju primene euklidske razdaljine vrlo osetljiv na postojanje ekstremnih vrednosti, specijalno u slučajevima male vrednosti *K*. Takođe, kada se radi o većoj količini podataka, sa velikim brojem prediktora, računanje razdaljine na ovaj način je resursno izuzetno zahtevno.

Alternativa primeni *Euclidean* je *Manhhatan* formula za izračunavanje razdaljine (Altman, 1992):

$$d(x_i, x_j) = \sum_{l=1}^{m}\left|x_{i,l} - x_{j,l}\right|$$

Ovakav pristup daje najkraću razdaljinu između dve tačke, u pravcima koji su paralelni koordinatnim osama prostora u kojem se tačke nalaze.

Slika 21. Primer *Manhhatan* i *Euclidean* razdaljine tačaka A i B u dvodimenzionalnom prostoru



Kada je broj prediktora veliki, kao i u slučaju postojanja ekstremnih vrednosti, *K-NN* je stabilniji kad se primenjuje *Manhhatan* formula (Harrington, 2016).

Kako se sličnost uzoraka meri međusobnim rastojanjem, algoritam je osetljiv na postojanje razlika u rangu vrednosti prediktora, te ih je iz tog razloga potrebno transformisati, kako bi njihove vrednosti bile uporedive (eng. *feature scaling*).

Ukoliko prediktori imaju normalnu distribuciju, predlaže se standardizacija (*z* Normalizacija) podataka (Gareth, 2017), kojom se srednja vrednost svodi na nulu, a standardna devijacija postaje jednaka jedinici.

Pored razdaljine uzoraka, poređenjem pravaca vektora koji te uzorke određuju, može se takođe odrediti njihova sličnost. Kada su vektori dva uzorka približno upravni, oni se smatraju različitim. Kada se pravci poklapaju, uzorci su slični. Metodom *Cosine* meri se kosinus ugla koji dva vektora zahvataju:

$$cosine(x_i, x_j) = \frac{x_i^T x_j}{\|x_i\|\|x_j\|}$$

$x_i, x_j -$ vektori uzoraka i i j

$0 \leq cosine(x_i, x_j) \leq 1,$ za vrednosti $\approx 1,$ sledi da su uzorci slični

Cosine se često primenjuje u tekst analitici, kada se na osnovu učestalosti pojavljivanja iste grupe reči određuje sličnost dokumenata.

Sličnost dva uzorka, gde je uzorak određen skupom svojih prediktora, može se predstaviti i kao količnik broja elemenata koji su u oba skupa isti (njihov presek) i broja različitih elemenata dva skupa (njihova unija). Ovakav način merenja sličnosti naziva se *Jaccard (*Alboukadel, 2017).

$$J(x_i, x_j) = \frac{|x_i \cap x_j|}{|x_i \cup x_j|} \ , \qquad 0\% \leq J(x_i, x_j) \leq 100\%$$

Veći broj istih elemenata dva skupa znači i veću sličnost uzoraka. Kada su uzorci isti, *Jaccard* sličnost je 100%.

Ovaj način merenja sličnosti se takođe često koristi u analizi dokumenata. Za razliku od *Cosine*, ključne reči se ne predstavljaju vektorom, već skupom. Primenjuje se kao i *Cosine*, u slučaju velikih količina podataka za analizu.

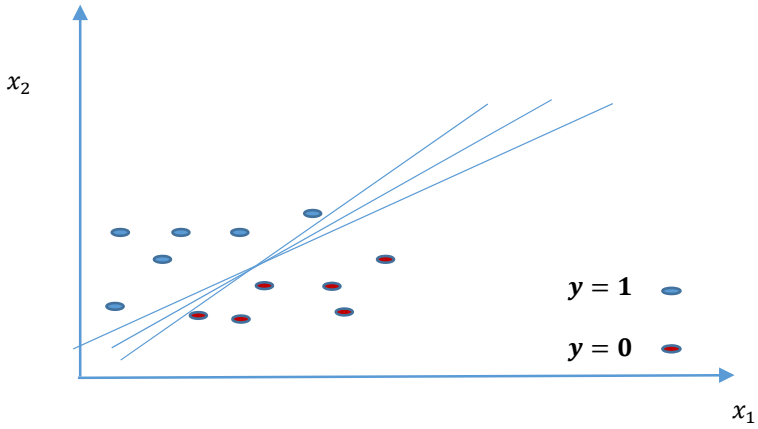
Potrebno je još napomenuti da veći broj susednih uzoraka (*K*), daje veću tačnost *K-NN* algoritma, ali je proces učenja algoritma znatno duži. Do optimalnog broja suseda uobičajeno dolazimo postupkom *cross validation*, odnosno postizanjem minimalne *cross validation* greške.

K-NN je jednostavan algoritam, jedini hiperparametar koji je potrebno odrediti je *K*.

3.3. Support Vector Machine – *SVM*

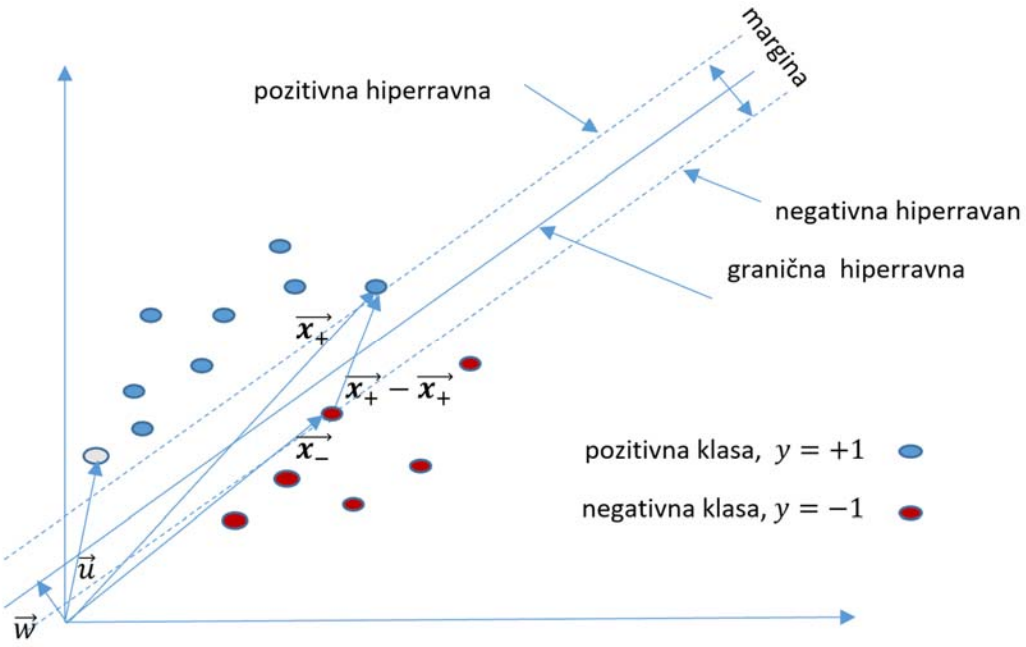
SVM pripada grupi diskriminativnih, nadgledanih algoritama, koji se najčešće primenjuje u klasifikacionim, a manje u regresionim problemima. Razvio ge je Vapnik (1995), te tako spada u novije algoritme. Primenom *SVM* , kao i kod ostalih diskriminativnih algoritama, dolazimo do optimalne, granične hiperravni (eng. *hiperplane*), koja na najbolji način razdvaja uzorke (u multidimenzionalnom prostoru) različitih klasa. Kako se uzorak može predstaviti tačkom u prostoru $x_i \epsilon R^m$, on se klasifikuje prema svom položaju u odnosu na graničnu hiperravan. Broj mogućih graničnih hiperravni je neograničen.

Slika 22. Primer dvodimenzionalnog prostora, $x_i \epsilon R^2$, u kojem je granična hiperravan prava linija



Za razliku od modela logističke regresije, gde se iz skupa mogućih graničnih hiperravni, optimalna određuje metodom maksimalne verodostojnosti (eng. *maximum likelihood – ML*), kojom se ne uzima u obzir razdaljina tačaka (uzoraka) od hiperravni, u *SVM* optimalna granična hiperravan je ona za koju je rastojanje (margina) između međusobno najbližih tačaka različitih klasa najveće. Položaji ovih tačaka određeni su vektorima $\overrightarrow{x_+}$ **i** $\overrightarrow{x_-}$, koji se nazivaju *support* vektorima. Optimalna hiperravan prolazi sredinom margine i naziva se hiperravan maksimalne margine (eng. *maximum margine hiperplane*). Pouzdanost, tačnost SVM modela je utoliko veća ukoliko je margina šira, čime se verovatnoća pogrešne klasifikacije uzoraka smanjuje. Tačke određene *support* vektorima jedine utiču na položaj granične hiperravni. Ovo je jedan od razloga zašto *SVM* ima dobru mogućnost generalizacije.

Slika 23. Granična heiperravan je prava linija u dvodimenzionalnom prostoru, koja razdvaja uzorke dve klase



3.3.1. Određivanje položaja granične hiperravni

Neka svaki trening uzorak određen vektorom $\overrightarrow{x_i} \in R^m$ pripada jednoj od dve klase binarne promenljive $Y = \{y_i \epsilon [+1, -1]\}$. Ako je vektor $\vec{w}$ upravan na hiperravan, a vektor $\vec{u}$ određuje položaj tačke testnog uzorka (koji želimo da klasifikujemo), prvi potreban uslov da uzorak npr. pripada pozitivnoj klasi (slika 23) dat je izrazom (Winston, 2010):

$$\vec{w}\ \vec{u}\ \geq b \tag{1}$$

Ukoliko je skalarni proizvod ova dva vektora (projekcija vektora $\vec{u}$ na $\vec{w}$) veći i jednak od neke konstantne vrednosti b, uzorak će biti klasifikovan kao pozitivan. Odnosno, što je vrednost

skalarnog proizvoda ova dva vektora veća, uzorak će biti udaljeniji (mereno u pravcu vektora $\vec{w}$) od koordinatnog početka, te je verovatnoća da je uzorak pozitivan veća.

Ovaj izraz možemo transformisati

$$\vec{w}\ \vec{u} - b \geq 0 \tag{2}$$

Izraz (2) nazivamo osnovnim pravilom odlučivanja u *SVM* (Winston, 2010). Kako bismo došli do vrednosti nepoznatih $\vec{w}$ i b, postavljamo još dva dodatna uslova (Winston, 2010):

$$\vec{w}\ \overrightarrow{x_{\iota+}} - b \geq 1, \text{ za uzorke koji pripadaju pozitivnoj klasi} \tag{3}$$

$$\vec{w}\ \overrightarrow{x_{\iota-}} - b \leq -1, \text{ za uzorke koji pripadaju negativnoj klasi}$$

$\overrightarrow{x_{\iota+}}$ – vektor uzorka pozitivne klase (osim pozitivnog uzorka određenog *support* vektorom)

$\overrightarrow{x_{\iota-}}$ – vektor uzorka negativne klase (osim negativnog uzoraka određenog *support* vekotorom)

Za različite vrednosti $y_i = \{+1, -1\}$ izrazi (3) se mogu napisati u sledećem obliku:

$$y_i(\vec{x}_i\vec{w} - b) - 1 \geq 0 \tag{4}$$

Dalje ćemo za uzorke koji se nalaze na granicama margina (pozitivna ili negativna hiperravan) i koji su određeni *support* vektorima, pretpostaviti da je (Winston, 2010):

$$y_i(\vec{x}_i\vec{w} - b) - 1 = 0 \tag{5}$$

Kako je već rečeno, položaj granične hiperravni određuje se tako da je širina margine (koju ova hiperravan polovi) najveća. Margina je jednaka skalarnom proizvodu, razlike *support* vektora $(\overrightarrow{x_+} - \overrightarrow{x_-})$ i jediničnog vektora upravnog na hiperravan $\frac{\vec{w}}{|w|}$,

$$\text{Širina margine} = (\overrightarrow{x_+} - \overrightarrow{x_-})\frac{\vec{w}}{|w|} \tag{6}$$

Za uzorke na granicama margine (određene *support* vektorima) iz (5) sledi da za različite vrednosti y_i važi:

$$y_i = \ 1 \ \rightarrow \ \overrightarrow{x_+}\vec{w} = 1 + b$$
$$y_i = -1 \ \rightarrow \ -\overrightarrow{x_-}\vec{w} = 1 - b$$

Sada jednačinu (6) možemo napisati u obliku

$$\text{Širina margine} = \frac{1+b+1-b}{|w|} = \frac{2}{|w|} \tag{7}$$

Iz (7) sledi da se uslov maksimalne širine margine, koju smo definisali jednačinom (6), dobija za minimalne vrednosti intenziteta vektora $\vec{w}$, tj. minimalnu vrednost $|w|$. Tako se rešenje osnovnog *SVM* problema svodi na dobijanje minimalne vrednosti $\frac{1}{2}|w|^2$ (ovako definsan uslov za min. vrednost $|w|$ je uveden kako bi se pojednostavilo računanje parcijalnih izvoda u daljem postupku), uz ispunjenje uslova datog jednačinom (4). Potrebno je dakle rešiti ovaj optimizacijski problem, koji možemo definisati *Lagrangian* funkcijom oblika:

$$L(w, b, \alpha) = \frac{1}{2}|w|^2 - \sum_{i=1}^{n} \alpha_i [y_i(\overrightarrow{x_i}\vec{w} + b) - 1] \tag{8}$$

Rešavanje ovog problema moguće je u dva koraka. Prvo za pretpostavljenu konstantnu vrednost *Lagrangian multiplier* α_i tražimo minimalnu vrednost *Lagrangian* funkcije, odnosno $min_{(w,b)}L(w,b,\alpha)$, iz uslova da su sledeći parcijalni izvodi jednaki nuli:

$$\frac{\partial L}{\partial \vec{w}} = 0$$

$$\frac{\partial L}{\partial b} = 0$$

Dobijamo da je:

$$\vec{w} = \sum_{i=1}^{n} \alpha_i y_i \vec{x_i} \tag{9a}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{9b}$$

Kada (9a i 9b) zamenimo u jednačinu (8), sledi:

$$L(\alpha_1 \dots \alpha_n) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i \, y_j \, \alpha_i \alpha_j \vec{x_i} \, \vec{x_j} \tag{10}$$

Ovo je novi optimizacijski problem, u kojem tražimo maksimalnu vrednost *Lagrangian* funkcije, u kojoj je sad nepoznata jedino α_i, uz ispunjenost uslova da je $\sum_{i=1}^{n} \alpha_i y_i = 0$.

Ovaj postupak rešavanja optimizacijskog problema u dva koraka naziva se *dual Lagrangian* i možemo ga napisati kao:

$$max_{\alpha_i} \left[min_{(w,b)} L(w,b,\alpha), \right]$$

Kada smo dobili α_i, iz (9a) dobijamo $\vec{\omega}$.

Kako je $y_i^{-1} = y_i,\ y_i \epsilon \{-1, 1\}$, iz (5) dobijamo vrednost preostale nepoznate *b*:

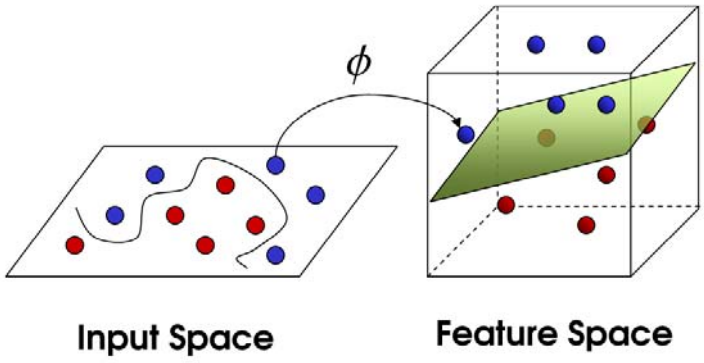
$$b = \vec{x_i}\vec{\omega} - y_i \tag{11}$$

Sa poznatim b i $\vec{w}$ rešen je *SVM* problem, tako da primenom osnovnog pravila odlučivanja (2) možemo utvrditi da li testni uzorak određen vektorom $\vec{u}$, pripada pozitivnoj klasi ili ne.

U slučajevima kada je $m \gg n$, kada je broj prediktora daleko veći od broja uzoraka, za rešavanje osnovnog *SVM* problema, odnosno dobijanje minimalne vrednosti $\frac{1}{2}|w|^2$ uz ispunjenje uslova (4), potrebno je procesirati $n * m$ podataka. Transformacijom osnovnog *SVM* problema u dual formu (10), podaci su dati u obliku skalarnog proizvod vektora $\vec{x_i}\,\vec{x_j}$, tako da je količina podataka za procesiranje jednaka n^2. Kako je $n^2 \ll n * m$, sledi da se transformacijom uslova u dual formu znatno smanjuje vreme učenja *SVM* algoritma.

Ukoliko nije moguće napraviti linearnu separaciju uzoraka dve (ili više) klase kao u gornjem primeru (kada govorimo o eng. *linear support vector machine*), pristupa se transformaciji vektora prediktora *x* iz početnog m dimenzionalnog prostora (tzv. *input space*), u prostor *p* dimenzija, gde je *p>m* (eng. *feature space*). Odnosno, potrebno je transformisati originalne podatke trening seta primenom neke funkcije $\Phi(x_i)$, tako da u novom, *p* dimenzionalnom prostoru, separacija uzoraka različitih klasa lineranom hiperravni postane moguća.

Slika 24. Transfomacija prediktora funkcijom Φ iz originalnog prostora u prostor veće dimenzionalnosti[11]



Kako bi ovakva transformacija trening podataka bila veoma zahtevna zbog skalarnog proizvoda vektora iz izraza (10), u p dimenzionalnom prostoru, primenjuje se postupak nazvan *Kernel trick* (Alboukadel, 2017). Postoji *kernel* funkcija $K(\overrightarrow{x_i}, \overrightarrow{x_j})$, čiji su ulazni atributi, vektori prediktora originalnog *input* prostora $\overrightarrow{x_i}, \overrightarrow{x_j} \in R^m$, koja daje output jednak skalarnom proizvodu transformisanih vektora ($\Phi(\overrightarrow{x_i})$ i $\Phi(\overrightarrow{x_j})$) u novom prostoru više dimenzije (*feature space*):

$$K\left(\overrightarrow{x_i}, \overrightarrow{x_j}\right) = \Phi(\overrightarrow{x_i})\Phi(\overrightarrow{x_j}) \tag{12}$$

Lagrangian funkcija (10), koja je linearno zavisna od skalarnog proizvoda vektora $\overrightarrow{x_i}$ i $\overrightarrow{x_j}$, u novom *feature* prostoru primenom *kernel* trika može se napisati kao:

$$L=\sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} y_i\, y_j\, \alpha_i\alpha_j \Phi(\overrightarrow{x_i})\Phi(\overrightarrow{x_j}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} y_i\, y_j\, \alpha_i\alpha_j K(\overrightarrow{x_i}, \overrightarrow{x_i}) \tag{13}$$

Najčešće korišćene *kernel* funkcije su (Alboukadel, 2017):

[11] Izvor: The Kernel Trick in Support Vector Classification | by Drew Wilimitis | Towards Data Science

- Polinomalna funkcija stepena *d*

$$K(\overrightarrow{x_i}, \overrightarrow{x_i}) = (\overrightarrow{x_i}\,\overrightarrow{x_j} + 1)^d$$

- *Radial basis* funkcija

$$K(\overrightarrow{x_i}, \overrightarrow{x_i}) = \exp(-\gamma|x_i - x_j\big|^2)$$

Iako je razvijeno dosta teorijskih metoda za izbor odgovarajuće k*ernel* funkcije, do optimalne najčešće dolazimo postupcima *bootstraping*-a i *cross validation*.

U poređenju sa ostalim klasifikacionim algoritmima, *SVM* se ne preporučuje u slučajevima velikih setova podataka i kada nije moguća dobra separacija uzoraka različitih klasa. Uobičajeno se primenjuje u slučajevima visoke dimenzionalnosti podataka, specijalno kada je broj prediktora veći od broja uzoraka (npr. klasifikacija slika). Ovaj algoritam ima dobru mogućnost generalizacije.

3.4. Klasifikacioni modeli na bazi stabla odlučivanja

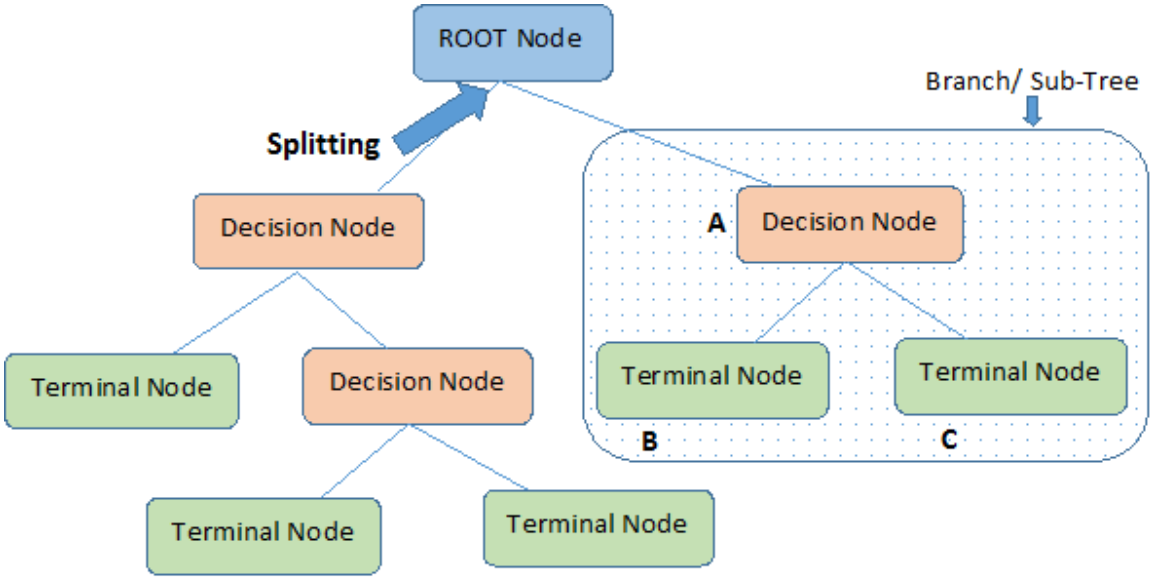
3.4.1. Stablo odlučivanja, *DT*

Stablo odlučivanja (*decission tree – DT*) spada u grupu nadgledanih neparametarskih klasifikacionih diskriminantnih modela *ML*, koji se uspešno može primenjivati i u regresionim problemima. Model 'Klasifikaciona i regresiona stabla' (eng. *Clasification and Regression Trees-CART*), prvi je predstavio Breiman et al., (1984).

DT se razvija sukcesivnim particionisanjem podataka (eng. *data splitting*), na osnovu graničnih vrednosti prediktora. Iz polaznog noda ili čvora (*root node*), koji predstavlja grafički prikaz početnog skupa trening podataka, isti se deli u manje podgrupe (particije). Svaka nova particija je novi čvor odlučivanja (eng. *decision node*). Postupak grananja stabla završava se kad se postigne zadovoljavajuća homogenost uzoraka u krajnjim (terminalnim) čvorovima stabla (u slučaju

klasifikacionih problema), postigne unapred projektovana dubina stabla ili maksimalni broj nodova. Tako kreirano stablo postaje model odlučivanja.

Slika 25. Rekursivno particionisanje stabla[12]



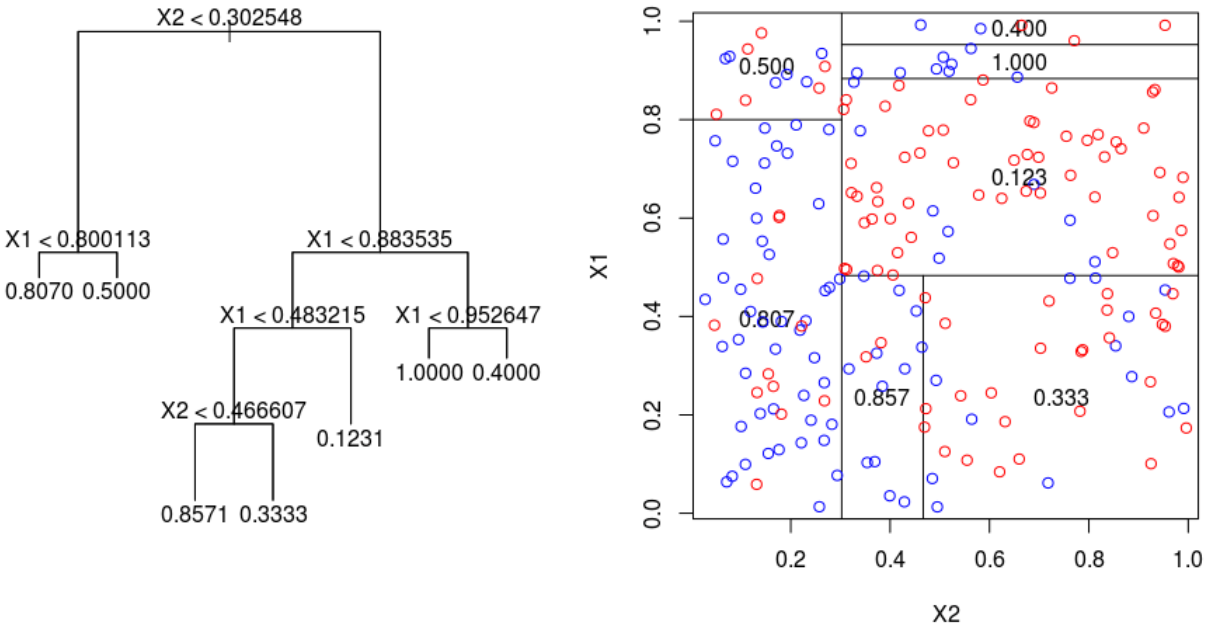
Novi uzorak, prolaskom kroz model stabla, ispunjavajući pravila particionisanja, dolazi do odgovarajućeg krajnjeg (terminalnog) noda. U slučaju klasifikacionih problema, uzorak će biti dodeljen većinskoj klasi uzoraka u tom nodu. Kada je problem regresioni, zavisna promenljva novog uzorka, jednaka je srednjoj vrednosti zavisne promenjive trening uzoraka terminalnog noda.

Na slici 26 dat je primer stabla odlučivanja, na bazi dva prediktora, sa sedam terminalnih čvorova. Svakom terminalnom čvoru odgovara jedna particija, određena graničnim vrednostima prediktora. Generalizacijom ovog pristupa na veći broj prediktora (m), zaključujemo da je svaka particija ograničen m dimenzionalni prostor.

[12] Izvor: Decision Tree Algorithm, Explained - KDnuggets

Slika 26. (a) Stablo odlučivanja (b) Uzorci u R^2 prostoru. Svakom terminalnom čvoru odgovara jedna particija, subprostor, određena graničnim vrednostima prediktora. Zavisna promenljiva je binominalna.

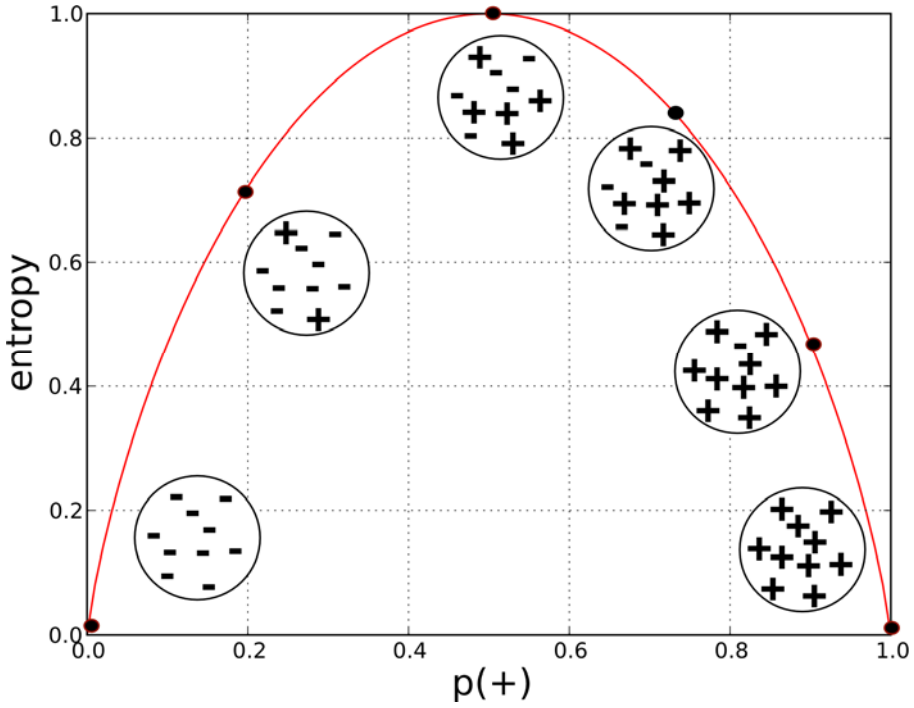
(a) (b)



Redosled prediktora, na osnovu kojih se podaci particionišu, kao i njihove granične vrednosti, određuju se na osnovu njihove informativnosti (eng. *information gain*).

U slučaju klasifikacionih problema, informativnost prediktora je mera povećanja homogenosti (eng. *purity,* čistoća) uzoraka u nodu, nastalih particionisanjem na bazi njegove granične vrednosti. Pod homogenosti uzoraka noda podrazumeva se nivo uređenosti – stepen prisustva uzoraka iste klase. Mera homogenosti noda ili skupa naziva se entropijom (eng. *entropy*) (Provost i Fawcett, 2013), čije su vrednosti u rangu od 0 do 1. Nodovi kod kojih svi uzorci pripadaju istoj klasi, potpuno su homogeni i imaju entropiju jednaku nuli. Maksimalnu entropiju imaju heterogeni skupovi, u kojima je broj uzoraka različitih klasa jednak.

Slika 27. Entropija kao mera uređenosti skupa uzoraka dve klase (+ i -). p(+), verovatnoća da uzorak pripada pozitivnoj klasi[13]



Entropija skupa, jednaka je:

$$entropy = -\sum_{i=1}^{k} p_i log_2(p_i)$$

p_i, verovatnoća da uzorak pripada klasi i

k , broj klasa zavisne promenljive

Entropija noda koji particionisanjem nastaje (eng. *child node*) na osnovu granične vrednosti prediktora, ne daje podatak o ukupnoj informativnosti tog prediktora (Provost i Fawcett, 2013).

[13] Izvor: Foster Provost i Tom Fawcett, 2013. *Data Science for Business*. s. 52

Pokazatelj za koliko se entropija noda (eng. *parent nod*) smanjila nakon particionisanja, osnovna je mera informativnosti prediktora i naziva se *Information Gain – IG* (Gareth et al., 2014). Tako za prediktor na osnovu kojeg se nakon particionisanja uzoraka *parent* noda postigne najveća uređenost *child* nodova, kažemo da ima najveći *IG*.

$$IG(P, x) = entropy(P) - \sum_{i=1}^{m} \frac{n_i}{n} \, entropy(C_i)$$

$x -$ prediktor na osnovu koga se $parent$ nod particioniše

$P -$ *parent* nod

$C_i -$ *child* nod

$n_i -$ broj uzoraka u C_i

$n -$ ukupan broj uzoraka parent noda P

$m -$ broj *child* nodova koji nastaje particionisanjem

Prediktor sa kojim se postiže maksimalna vrednost *IG* je onaj sa sa kojim particionisanje – grananje stabla započinje (Provost i Fawcett, 2013). Dalje particionisanje nastavlja se prediktorima prema opadajućim vrednostima *IG*, koji se njima postiže.

Pored prethodno opisanog postupka, informativnost prediktora u binarnim klasifikacionim problemima moguće je odrediti na osnovu vrednosti *Gini impurity* indeksa (Provost i Fawcett, 2013). Ovaj indeks je mera neuređenosti skupa. Za uzorke *child* noda jednak je:

$$Gini_{child} = 1 - \sum_{i=1}^{k} P_{r_i}^2$$

k, klase zavisne promenjive

Pr_i, verovatnoća da uzorci pripadaju klasi *i*

Minimalna vrednost *Gini* indeksa je 0, u slučaju kada svi uzorci pripadaju istoj klasi. Najveća moguća vrednost je 0.5 , kada je broj uzoraka dve klase jednak.

Vrednost *Gini* indeksa *parent* noda jednaka je ponderisanoj vrednosti sume *Gini* indeksa *child* nodova:

$$Gini_{parent} = \sum_{i=1}^{m} \frac{n_i}{n} Gini_i$$

$m -$ broj *child* nodova koji nastaje particionisanjem

$n_i -$ broj uzoraka *child* noda *i*

$n -$ ukupan broj uzoraka parent noda *P*

Prediktor sa najmanjom vrednošću $Gini_{parent}$ indeksa je najinformativniji.

Postupak sukcesivnog particionisanja, razvoj stabla, nastavljamo sve dok je vrednost *IG* značajna (Gareth et al., 2014). Ukoliko je dobitak u *IG* sa novim particionisanjem minimalan, razvoj stabla se zaustavlja. Postupak ograničavanja rasta stabla odlučivanja naziva se postupkom regularizacije. Tako prema Zhang (2016), potreba za daljim particionisanjem prestaje kada se ispune neki od sledećih uslova:

- Svi terminalni nodovi su potpuno homogeni
- Postignuta je predefinisana vrednost minimalnog broja uzoraka u terminalnom nodu
- Dostignuta je maksimalna dubina ili maksimalan broj nodova u stablu
- Ograničavanje broja prediktora na osnovu kojih je moguće particionisati

U slučaju regresionog stabala odlučivanja, izbor prediktora i graničnih vrednosti bazira se na vrednosti sume kvadrata reziduala. Odnosno, sumi kvadrata razlika vrednosti zavisne promenljive uzoraka *child* noda i njene srednje vrednosti (Gareth et al., 2014)

$$RSS = \sum_{i=1}^{c} \sum_{j=1}^{l} \left(y_{ij} - \overline{y}_{\iota}\right)^2$$

RSS – suma kvadrata rezidualnih vrednosti

c, broj *child* nodova koji particionisanjem nastaje

l, broj uzoraka u *child* nodu

$\overline{y}_{\iota}$, srednja vrednost zavisne promenljive u nodu *i*

y_{ij}, vrednost zavisne promenljive u nodu i

Najinformativniji prediktor je onaj za koji je vrednost *RSS* najmanja.

Jasno je da što je dubina stabla veća, ono je kompleksnije i teže za interpretaciju. Do optimalne dubine dolazimo uvođenjem *cost complexity* parametra α, kojim se penalizuje vrednost *RSS* (Bruce P. i Bruce A., 2017).

Tako uslov postizanja minimalne vrednosti *RSS* postaje:

$$minimize\{RSS + \alpha T\}$$

T, broj terminalnih nodova stabla

Mala vrednost α proizvodi kompleksnija stabla, dok veće vrednosti daju stabla manje dubine.

Iz ovog sledi da stablo treba razvijati sve dok je smanjenje *RSS*, koje se postiže rastom dubine stabla, veće od vrednosti za koji se model penalizuje ($\alpha\ T$). Do optimalne vrednosti α dolazimo postupkom *cross* validacije (Gareth et al., 2014).

U slučaju regresionog stabla, informativnost prediktora se može odrediti i na osnovu varijanse zavisne promenljive u terminalnim čvorovima (Gareth et al., 2014):

$$Var = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}$$

n broj uzoraka u nodu

Za izabrani prediktor (x_i) računa se varijansa zavisne promenljive svakog *child* noda, koji nastaje račvanjem na osnovu njegove granične vrednosti. Varijansa *parent* noda jednaka je ponderisanoj vrednosti zbira varijansi svakog noda:

$$Var(x_i)_P = \frac{n_1}{n} Var_{c1} + \frac{n_2}{n} Var_{c2} \qquad * \ primer\ sa\ dva\ noda$$

$n_1, n_2,$ broj uzoraka prvog i drugog noda

$n = n_1 + n_2,$ ukupan broj uzoraka

$Var_{c1} -$ varijansa zavisne promenljive uzoraka u prvom nodu

$Var_{c2} -$ varijansa zavisne promenljive uzoraka u drugom nodu

$Var(x_i)_P -$ varijansa zavisne promenljive *parent* noda, za prediktor x_i

Redosled prediktora na bazi kojih se stablo grana, određuje se prema rastućoj vrednosti varijanse. Odnosno, najinformativniji je prediktor za koji se dobija najmanja vrednost varijanse (Bruce P. i Bruce A., 2017).

Regresioni i klasifikacioni *DT* algoritmi mogu se dobro interpretirati i to je jedna od njihovih osnovnih prednosti. Međutim, kako *DT* model ima samo jedno stablo odlučivanja, on nije stabilan, podložan je overfitingu (u slučaju kompleksnog stabla), a time i slaboj generalizaciji.

DT se primenjuje u slučajevima kada je relacija zavisne promenljive i prediktora kompleksna i kad se ne može aproksimovati linearnom funkcijom. Ovaj algoritam ne zahteva transformaciju podataka, što dodatno pojednostavljuje njegovu primenu.

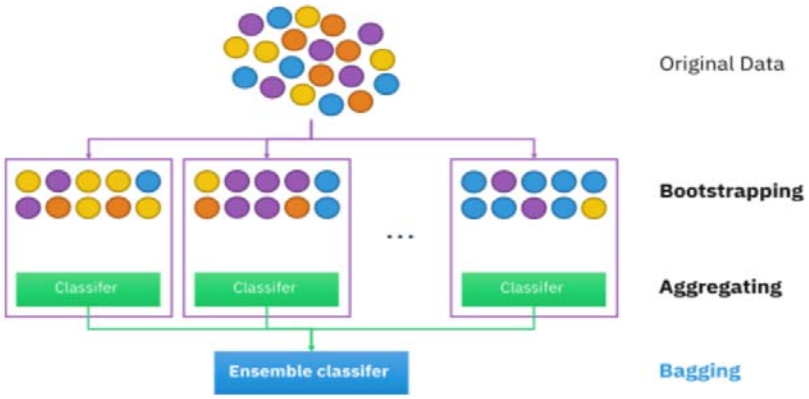
3.4.2. Random Forest, *RF*

RF je sličan *DT*, sa razlikom da se ovim modelom kreira više stabala odlučivanja. Na taj način se rešava osnovni problem nestabilnosti predikcije *DT*, uzrokovan postojanjem samo jednog stabla odlučivanja (Bruce P. i Bruce A., 2017).

Svako pojedinačno stablo, kao element *RF*, razvija se na skupovima podataka istog broja uzoraka, koji su podskupovi originalnog trening *data* seta i nazivaju se *bootstrap*. Uzorci *bootstrap*-a biraju se postupkom uzorkovanja sa zamenom *(*eng. *with replacement)*, tako da se svaki uzorak može jednom ili više puta pojaviti u istom *bootstrap*-u. Uzorci koji su 'izostavljeni' i ne pripadaju niti jednom *bootstrap*-u, sačinjavaju tzv. *out of bag dataset – OBD*, koji se koristi za testiranje i proveru tačnosti prediktivnog *RF* modela.

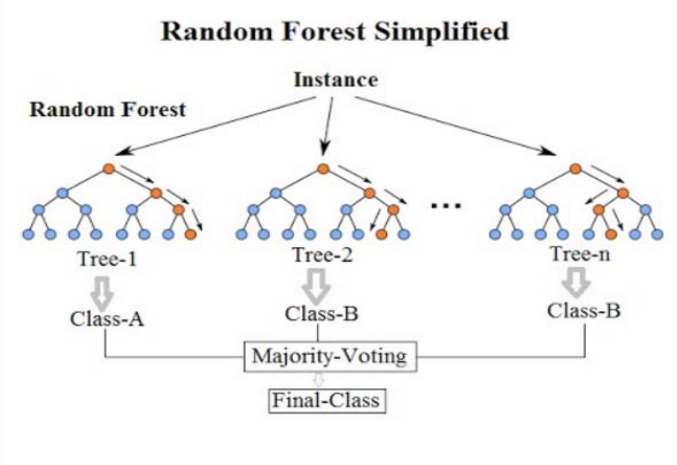
Postupak kojim se agregiraju rezultati dobijeni primenom pojedinačnih modela, u slučaju *RF* su to pojedinačna stabla, naziva se *bagging,* što je skraćenica od *bootstrap aggregating* (Leo Breiman, 1994). Ovim postupkom postiže se veća stabilnost modela i smanjuje mogućnost overfitinga (Breiman, 1994). *RF* spada u grupu *bagging* algoritama, gde svako kreirano stablo predstavlja zaseban klasifikator (u slučaju klasifikacionih problema), dok je finalna predikcija rezultat agregacije predikcija svakog pojedinačnog stabla.

Slika 28. *Bootstrap aggregating*[14]



U slučaju regresionih problema, kao i kod *DT,* konačni rezultat predikcije jednak je srednjoj vrednosti zavisne promenljive u odgovarajućim terminalnim nodovima svakog pojedinačnog stabla. Kod klasifikacionih problema, konačna predikcija modela jednaka je preovlađujućoj klasi odgovarajućih terminalnih nodova (eng. *majority voating)* svakog stabla.

Slika 29. *RF* klasifikacioni model[15]



Kada *RF* kreira novo stablo, koristi se samo ograničen broj proizvoljno izabranih prediktora. Ako je *m* njihov ukupan broj, u slučaju klasifikacionih problema, za razvoj pojedinačnog stabla preporučuje se $\sqrt{m}$, a u slučaju regresionih problema *m*/3, proizvoljno izabranih prediktora

[14] Izvor: (Wikipedia, Bagging)
[15] Izvor: Edureca: Random forest in R

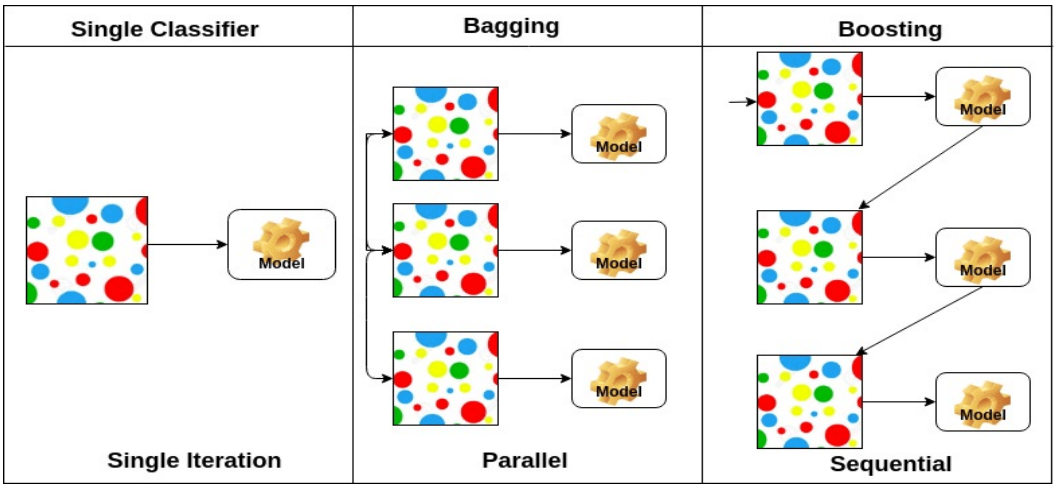
(Breiman i Cutler, 2018). Kako je ovaj broj prediktora znatno manji od *m*, model svakog stabla je različit, a stabla međusobno nezavisna. Korišćenje samo ograničenog broja prediktora i velikog broja stabala razlog je veće tačnosti predikcije i manje mogućnosti overfitinga u odnosu *DT*. S druge strane, za 'učenje 'ovog modela potreban je veliki broj uzoraka, imajući u vidu da je uobičajen broj stabala *RF* veći od 500 (Breiman and Cutler, 2018).

Za razliku od standardnih pristupa u rešavanju problema nedostajućih podataka, kao što su maksimalna verodostojnost i regresija, *RF* algoritam rekonstruiše ili procenjuje podatke koji nedostaju (eng. *not vailable – NA*) primenom *proximity* matrice (kvadratne matrice, reda koji odgovara broju uzoraka). Elementi ove matrice predstavljaju meru međusobne sličnosti uzoraka. Pod sličnim uzorcima smatraju se oni koji pripadaju istom terminalnom nodu nakon 'prolaska' uzoraka kroz *RF* stablo. Vrednosti *proximity* matrice kreću se u rangu od 0 za različite, do 1 za slične uzorke. Maksimalno slični uzorci u svakom stablu *RF* pripadaju istom terminalnom nodu. Suprotno, uzorci koji se u svim stablima nalaze u različitim terminalnim nodovima različiti su. Kako se *NA* rekonstruiše na osnovu postojećih uzoraka, njihov uticaj će biti srazmeran sličnosti sa uzorkom čiji se nedostajući podatak rekonstruiše. Tako će najsličniji uzorci imati najveći uticaj na rekonstruisanu vrednost.

RF spada u jedan od najtačnijih prediktivnih algoritama, kako za klasifikacione tako i regresione probleme. Algoritam nije osetljiv na ekstremne vrednosti. Kako su stabla ovog modela bazirana na manjem, ograničenom broju prediktora, *RF* se primenjuje u slučajevima kada je broj prediktora velik. Osnovni nedostatak *RF* je njegova kompleksnost, odnosno nemogućnost tumačenja.

U *RF* svako stablo odlučivanja ima isti značaj (eng. *weight*) u odlučivanju. Stabla su međusobno nezavisna i kreirana istovremeno. U slučaju kada se kreiraju sukcesivno i kada su zavisna, tako da svako novo stablo uzima u obzir grešku predikcije načinjenu prethodnim, govorimo o *Boosting* klasifikacionim algoritmima.

Slika 30. *Bagging* i *Bossting* klasifikacioni algoritmi[16]



Postupkom *boosting-a*, sukcesivno kreirani tzv. slabi klasifikatori (eng. *weak classifiers – WC)*, niske tačnosti predikcije, kombinuju se u jake (eng. *strong classifiers – SC*), koji klasifikuju uzorak na osnovu rezultata dobijenih primenom više *WC* (Bottou, 2010).

Ako za binarni klasifikacioni problem, gde je $y(x)\ \epsilon\ [+1, -1]$, slab klasifikator označimo sa $\hat{y}(x)$, a jak sa $\hat{Y}$, sledi da je:

$$\hat{Y}(x) = sign(\alpha_1\ \hat{y}_1(x) + \alpha_2\hat{y}_2(x) + \alpha_3\hat{y}_3(x) \ldots + \alpha_b\hat{y}_b(x)) \qquad (1)$$

b, ukupan broj slabih klasifikatora *WC*

α, koeficijent koji određuje značaj svakog *WC* pri odlučivanju

Boosting algoritmi se uspešno primenjuju u rešavanju binarnih klasifikacionih problema. U slučaj u kada je zavisna promenljiva kategorička, sa više od dve klase, problem se svodi na višestruko rešavanje binarnih problema (Freund i Schapire, 1997; Schapire i Singer, 1999).

[16] Izvor: AdaBoost Classifier Algorithms

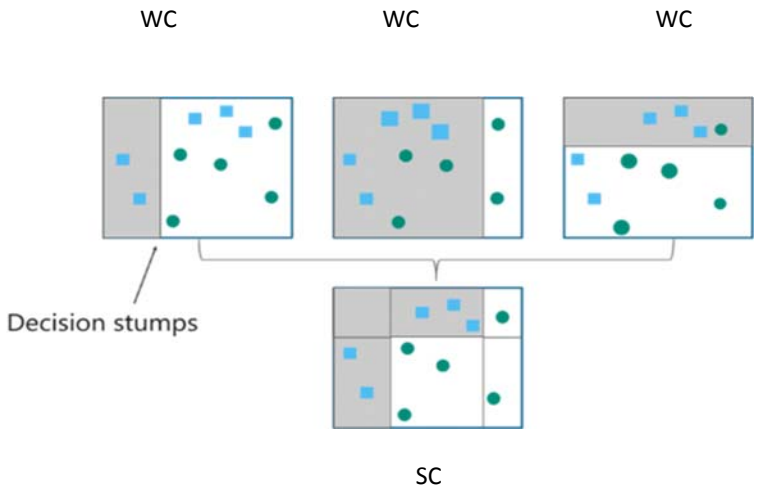
U zavisnosti od strukture *WC* i načina na koji se određuje njihov uticaj (factor α), razlikujemo sledeće *bosting* algoritme: *Adaboost, Gradinet tree boost* i *Extreme gradient boost*.

3.4.3. Adaboost – Binarni klasifikator

Adaboost algoritam primenjuje se u binarno klasifikacionim problemima. Za razliku of *RF* model a, stabla koja se kreiraju imaju samo jednu granu odlučivanja (eng. *decision stump)* (Freund i Schapire, 1997). Pri kreiranju *stump-a* koristi se samo jedan prediktor, tako da njihov broj odgovara broju prediktora. *Stump* koji postigne najveću tačnost predikcije, jeste *WC* najvećeg značaja.

Za uzorke u dvodimenzionalnom prostoru granične vrednosti prediktora određuju položaj prave koja razdvaja uzorke različitih klasa. Značaj (eng. *weight*) svakog pojedinačnog uzorka koji je pogrešno klasifikovan, primenom *WC* povećava se u narednoj iteraciji. Tako se sa svakim novim *WC* povećava verovatnoća da uzorci sa većim značajem budu pravilno klasifikovani. Za razliku od *RF*, gde se stabla kreiraju nezavisno, u *Adaboost*-u svako novo stablo uzima u obzir grešku nastalu primenom prethodnog *WC*. Kombinovanjem svih *WC,* dobijamo konačni *SC*, kojim se postiže bolja tačnost predikcije.

Slika 31. Tri *WC* binarna klasifikatora, kombinovani u SC[17].



Ako sa w_i^t označimo značaj (*'težinu'*) *i* uzorka, u trenutku *t* = 0, u prvoj iteraciji ona će biti ista za sve uzorke:

$$w_i^t = \frac{1}{n}$$

n, ukupan broj uzoraka

Postavićemo dalje uslov da suma 'težina' svih uzoraka, u svakoj iteraciji (za svako *t*), mora biti jedan:

$$\sum_{i=1}^{n} w_i^t = 1$$

Ukupna greška klasifikacije (eng. *total error*) slabog klasifikatora *WC,* u trenutku *t* = 0, jednaka je:

[17] Izvor: Boosting-machine-learning

$$\varepsilon^t = \sum_{i=1}^{j} \frac{1}{n} = \sum_{i=1}^{j} w_i^t$$

j, broj pogrešno klasifikovanih uzoraka

U svakoj iteraciji slabi klasifikatori *(WC)* se kreiraju za svaki prediktor. Onaj za koji je broj pogrešno klasifikovanih uzoraka najmanji, odnosno ε^t ima najnižu vrednost, postaje izabrani *WC* u datoj iteraciji,odnosno *WC* najvećeg značaja.

Primenom tako izabranog *WC*, dobijamo prvu prediktivnu vrednost $\hat{y}_t$.

Težine uzoraka u narednoj iteraciji w_i^{t+1} funkcija su prethodnih težina i tačnosti predikcije u trenutku *t* (Winston, 2010):

$$w_i^{t+1} = \frac{w_i^t}{z} e^{-\alpha_t \hat{y}_t(x) y(x)} \quad (1)$$

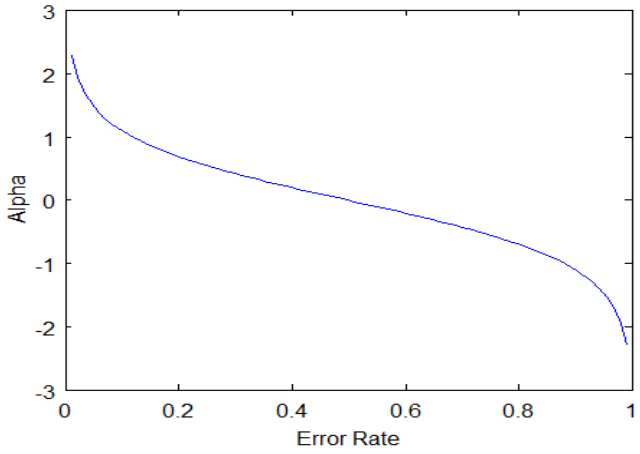
$z \in R$ je tzv. Normalizacioni faktor, kojim se obezbeđuje ispunjavanje uslova da je zbir težina u svakoj iteraciji jednak jedan

α_t, je faktor značaja slabog klasifikatora (eng. *amount of say)* (Winston, 2010):

$$\alpha_t = \frac{1}{2} ln \frac{1-\varepsilon^t}{\varepsilon^t}$$

Iz grafičkog prikaza funkcije α_t vidimo da manjim vrednostima ukupne greške ε_t odgovaraju veće vrednosti α_t.

Slika 32. Rastom ukupne greške klasifikacije ε_t, značaj slabog klasifikatora α_t se smanjuje (Masashi & Sugiyama, 2016)



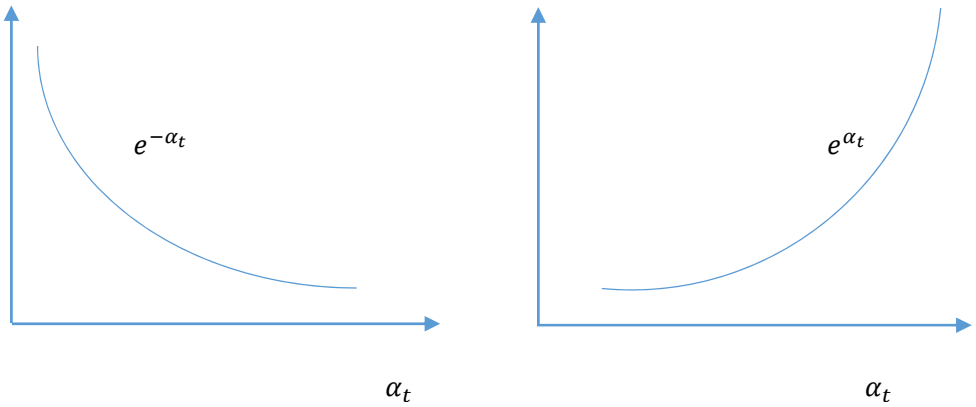
Proizvod $\hat{y}_t(x)y(x)$ ima vrednost + 1, kada je predikcija sa *WC* tačna i -1, kad je netačna ($y(x)\ \epsilon\ [+1, -1]$). Sledi da (1) možemo napisati u sledećem obliku:

$$w_i^{t+1} = \frac{w_i^t}{z} \times\ e^{-\alpha_t} \quad - \text{ za pravilno klasifikovane uzorke} \tag{2a}$$

$$w_i^{t+1} = \frac{w_i^t}{z} \times e^{\alpha_t} \quad - \text{ za pogrešno klasifikovane uzorke} \tag{2b}$$

Eksponencijalne funkcije date izrazima (2a i 2b) pokazuju da se za pravilno klasifikovane uzorke rastom α_t smanjuje njihov značaj, w_i^{t+1}. Za pogrešno klasifikovane, vrednost w_i^{t+1} raste sa rastom α_t.

Slika 33. (a) Eksponencijalna funkcija pravilno klasifikovanih uzoraka (b) eksponencijalna funkcija za nepravilno klasifikovane uzorke



Jednačinama (2a i 2b) daljom transformacijom postaju:

$$w_i^{t+1} = \frac{w_i^t}{z} \times \sqrt{\frac{\varepsilon^t}{1-\varepsilon^t}} \quad - \text{ za pravilno klasifikovane uzorke} \tag{3a}$$

$$w_i^{t+1} = \frac{w_i^t}{z} \times \sqrt{\frac{1-\varepsilon^t}{\varepsilon^t}} \quad - \text{ za pogrešno klasifikovane uzorke} \tag{3b}$$

Kako je zbir ’težina’ svih uzoraka jednak 1, sledi:

$$\sum_{i=1}^{k} \frac{w_i^t}{z} \times \sqrt{\frac{\varepsilon^t}{1-\varepsilon^t}} + \sum_{i=1}^{j} \frac{w_i^t}{z} \times \sqrt{\frac{1-\varepsilon^t}{\varepsilon^t}} = 1$$

Odnosno:

$$\sqrt{\frac{\varepsilon^t}{1-\varepsilon^t}} \sum_{i=1}^{k} w_i^t + \sqrt{\frac{1-\varepsilon^t}{\varepsilon^t}} \sum_{i=1}^{j} w_i^t = \mathrm{z}$$

k, broj pravilno klasifikovanih uzoraka

j, broj pogrešno klasifikovanih uzoraka

Kako je $\sum_{i=1}^{j} w_i^t = \varepsilon^t$, a $\sum_{i=1}^{k} w_i^t = 1 - \varepsilon^t$, sledi da je normalizacioni faktor jednak:

$$z = 2\sqrt{(1-\varepsilon^t)\varepsilon^t} \tag{4}$$

Kada vrednost *z* uvrstimo u 3a i 3b, dobijamo finalno (Winston, 2010):

$w_i^{t+1} = \frac{w_i^t}{2}\frac{1}{1-\varepsilon^t}$ – za pravilno klasifikovane uzorke (5a)

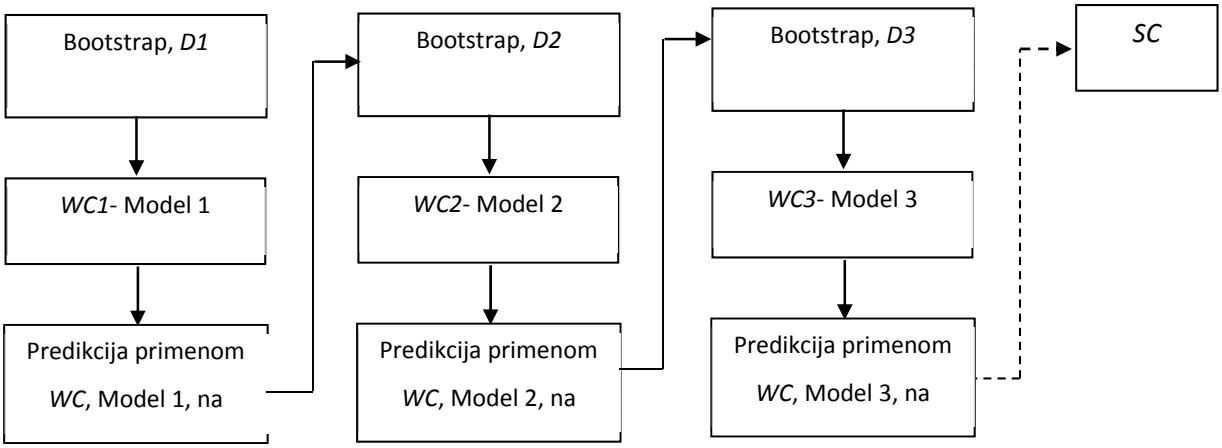
$w_i^{t+1} = \frac{w_i^t}{2}\frac{1}{\varepsilon^t}$ – za pogrešno klasifikovane uzorke (5b)

Na slici 34 prikazan je model *Adaboost* algoritma. U prvom *bootstrap-u* – u *D1*, svi uzorci imaju isti značaj ω_1 . Model *M1*, prvi *WC* na podacima *D1*, slab je klasifikator najvećeg značaja (za koji faktor α ima najveću vrednost u prvoj iteraciji). Uzorci koji su primenom *M1* pogrešno klasifikovani, u sledećoj iteraciji dobijaju veći značaj od uzoraka koji su pravilno klasifikovani. Na osnovu novih vrednosti $\omega_{2,}$ formira se novi *bootstrap* (*D2*), u kojem veću zastupljenost imaju pogrešno klasifikovani uzorci iz prethodne iteracije. Postupak se dalje nastavlja kreiranjem novog *WC*, model *M2*. Na osnovu rezultata klasifikacije ovim modelom, računaju se nove vrednosti ω_3, a na osnovu njih formira *bootstrap D3* itd. Ovaj postupak kreiranja *WC* nastavlja se dok se ne postigne njihov unapred definisani maksimalni broj ili minimalna projektovana dozvoljena greška predviđanja, uz uslov da je broj kreiranih *WC* neparan. Konačna predikcija *SC* rezultat je kombinovanja svih kreiranih *WC* i jednaka je većinskoj klasi iz skupa predikcija, svakog *WC* pojedinačno:

$$SC(x) = sign(\alpha_1 WC_1(x) + \alpha_2 WC_2(x) + \alpha_3 WC_3(x) \ldots + \alpha_t WC_t(x))$$

t, ukupan broj *WC*

Slika 34. Adaboost model



Adaboosting je relativno jednostavan za implementaciju i nije podložan overfitingu. Treba imati u vidu da je dosta osetljiv na ekstremne vrednosti prediktora.

3.4.4. Gradient tree boosting, *GTB*

GTB takođe spada u grupu *boosting* algoritama, kod kojeg se kreira veći broj stabala odlučivanja – *WC*, pri čemu svako stablo uzima u obzir grešku predikcije prethodnog. Za razliku od *AdBoost*-a, gde stabla odlučivanja imaju strukturu *stump-a*, kod *GTB* algoritama stablo se razvija na osnovu graničnih vrednosti više prediktora, tako da je veće dubine. Takođe, predviđena vrednost zavisne promenljive u prvoj iteraciji, koja je kod *Adaboost* algoritma rezultat primene prvog *WC* , kod *GTB* modela je, u slučaju regresionih problema, srednja vrednost zavisne promenjive *bootstrap*-a. U slučaju kategoričke zavisne promenljive, njena predviđena vrednost, kojom se inicijalizuje *GTB* model, $\hat{Y}_0(x)$ jednaka je:

$$\hat{Y}_0(x) = \log(\theta) \tag{1}$$

θ, izglednost pozitivnog događaja (eng. *odds*) ($y = 1$)

$\hat{Y}_0(x)$ se takođe može prikazati i kao verovatnoća pozitivnog događaja (p)

$$\hat{Y}_0(x) = p = \frac{e^{\log(\theta)}}{1+e^{\log(\theta)}} \quad (2)$$

Ako usvojimo da je granična vrednost verovatnoće 0.5, svaki novi uzorak biće klasifikovan kao pozitivan događaj ako je vrednost $p > 0.5$.

Može se pokazati da je inicijalna vrednost *GTB* modela, $\hat{Y}_0(x)$, data jednačinama (1) ili (2), optimalna (Friedman, 2001).

Kako skup mogućih vrednosti zavisne promenljive $y \in \{0,1\}$ predstavlja skup nezavisnih događaja, koji imaju binomnu distribuciju, ona je određena brojem uzoraka i parametrom p, verovatnoćom pozitivnog događaja, $y \sim B(n,p)$.

Zajednička distribucija binarne zavisne promenljive jednaka je:

$$P_r(y_1, y_2, .., y_n; p) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i} .$$

Logaritmovanjem ovog izraza (log verodostojnosti) dobijamo:

$$Log(P_r(y_1, y_2, .., y_n; p)) = \sum_{i=1}^{n}(y_i \log p + (1-y_i)\log(1-p)) \quad (3)$$

Funkcija gubitaka *LF* (eng. *loss function*), koja se koristi u *GTB*, a čiju vrednost želimo minimizirati kako bi predikcija modelom imala najveću tačnost, u slučaju binarno klasifikacionih problema jednaka je negativnoj vrednosti izraza (3).

$$LF = -(\sum_{i=1}^{n} y_i \log(p) + (1 - y_i)\log(1 - p)) \quad (4)$$

Iz razloga jednostavnijeg računanja parcijalnih izvoda, LF se prikazuje u funkciji $\log(\theta)$. Kako je $\theta = \frac{p}{1-p}$, odnosno $\log\theta = \log(\frac{p}{1-p})$ i $p = \frac{e^{\log\theta}}{1+e^{\log\theta}}$, izraz (4) se može napisati kao:

$$LF(y_{i,} \log(\theta)) = \sum_{i=1}^{n} -y_i \log(\theta) + \log(1 + e^{\log(\theta)}) \quad (5)$$

Potrebno ja tako naći vrednost $\log(\theta)$, odnosno p, za koju funkcija gubitaka ima minimalnu vrednost (Friedman, 2001):

$$\underset{\log(\theta)}{\operatorname{argmin}} \sum_{i=1}^{n} LF(y_i, \log(\theta))$$

Iz uslova $\frac{\partial LF}{\partial \log(\theta)} = 0$ dobijamo:

$$\frac{\partial LF}{\partial \log(\theta)} = \sum_{i=1}^{n} -y_i + \frac{e^{\log(\theta)}}{1+e^{\log(\theta)}} = \sum_{i=1}^{n} -y_i + p = 0 \quad (6)$$

Iz (6) se dobija de je:

$$p = \frac{n_p}{n}$$

n_p, broj pozitvnih događaja $(y = 1)$

n_n, broj negativnih događaja $(y = 0)$

$n = n_p + n_n$

Tako se minimalna vrednost funkcije gubitaka postiže za $p = \frac{n_p}{n}$, odnosno kako je $\theta = \frac{p}{1-p}$, sledi $\theta = \frac{n_p}{n_n}$, tj. $\log(\theta) = \log\left(\frac{n_p}{n_n}\right)$. Ovime je dokazano da je (1), najbolja aproksimacije za predikciju zavisne promenljive u prvoj iteraciji.

Model *GTB* dalje se razvija dodavanjem stabala odlučivanja (*WC*). Za razliku od *Adaboost* modela, gde se stabla razvijaju, nad podacima strukture (x_i, y_i), kod *GTB* se formira nova struktura podataka, u kojoj se zavisna promenljiva zamenjuje rezidualnim vrednostima prethodne iteracije. Odnosno, struktura podataka postaje (x_i, r_{im}), gde je *m* broj iteracije u modelu. Ove rezidualne vrednosti dobijamo iz uslova (Friedman, 2001):

$$r_{i,m} = -\left[\frac{\partial \mathrm{LF}(y_i, \hat{y}(x_i))}{\partial \hat{Y}_m(x_i)}\right], \hat{Y}_m(x_i) = \hat{Y}_{m-1}(x_i)$$

Tako posle nulte iteracije, $r_{i,0} = -(y + \frac{e^{\log(\theta)}}{1+e^{\log(\theta)}}) = y - p$

Sledi da je

$$r_{i,m} = y_i - p_{i,m} \tag{7}$$

Kako se svako naredno stablo razvija na podacima oblika (x_i, r_{im}), rezidualne vrednosti koje pripadaju istom terminalnom nodu potrebno je transformisati u jedinstvenu vrednost (*R*). U slučaju regresionih problema, do vrednosti *R* dolazi se jednostavnim uzimanjem srednje vrednosti reziduala u terminalnim nodovima. U slučaju klasifikacionih problema, ova vrednost $R_{j,m}$ određuje se iz uslova minimalne vrednosti funkcije gubitaka (Fredman, 2001):

$$\gamma_{j,m} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{i,j}} LF(y_i, \hat{Y}_{m-1}(x_i) + \gamma) \tag{8}$$

$j = 1 \ldots \ldots J_m$, ukupan broj terminalnih nodova stabla m

$x_i \in R_{i,j}$, uzorci koji pripadaju terminalnom nodu *j*

$\gamma_{j,m} = R_{j,m}$

$LF,$ funkcija gubitaka

Na osnovu funkcije gubitaka za *GTB* (5), uslov (8) postaje:

$$\underset{\gamma}{\operatorname{argmin}}\, LF = \sum_{x_i \in R_{i,j}} -y_i \left(\hat{Y}_{m-1}(x_i) + \gamma\right) + \log(1 + e^{\hat{Y}_{m-1}(x_i)+\gamma})$$

Iz uslova $\frac{\partial L}{\partial \gamma} = 0$, dobijamo

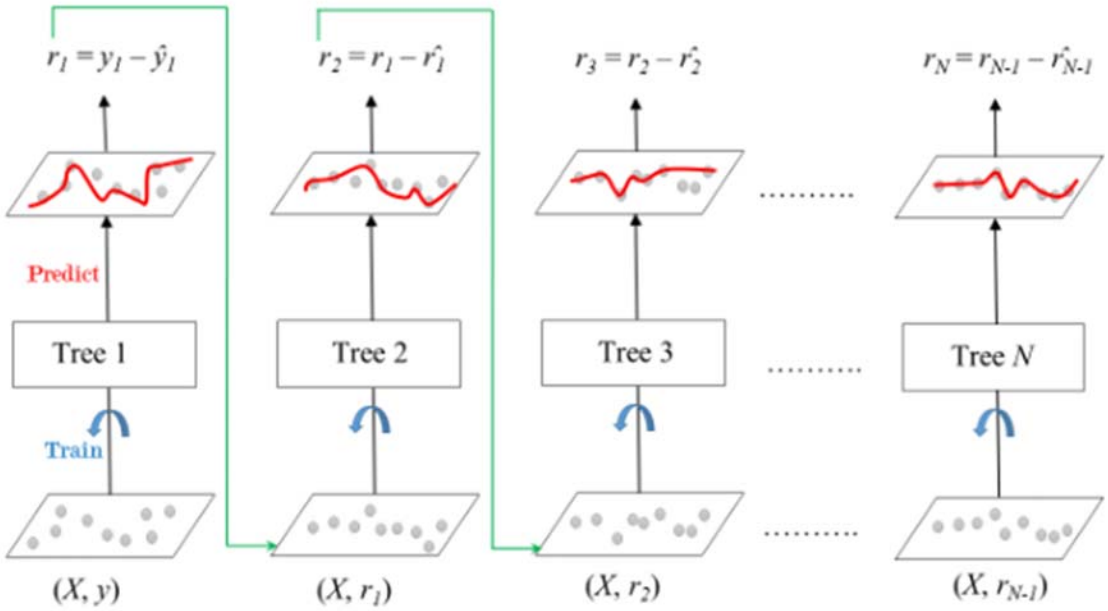
$$R_{j,m} = \frac{\sum_{i \epsilon j,m} r_i}{\sum_{i \epsilon j,m} p_{i,}(1-p_i)}$$

r_i – rezidualna vrednost za uzorak *i* , u terminalnom čvoru *j,* stabla *m*

p_i– prediktivna verovatnoća prethodne iteracije, uzorka *i*, terminalnog čvora *j*

Postupak kreiranja stabala nastavljamo dok ne postignemo njihov maksimalni predviđeni broj ili kad se sa novim stablom tačnost predikcije samo minimalno poveća (male rezidualne vrednosti $r_{i,m}$).

Slika 35. Razvoj *GTB* modela[18]



Tako da finalno prediktivni model, odnosno *SC*, ima oblik:

$$\hat{Y}(x_i) = \hat{Y}_o + \delta\,\hat{Y}_1(x_i) + \delta\,\hat{Y}_2(x_i) + \cdots + \delta\,\hat{Y}_N(x_i)$$

N, maksimalni broj stabala *GTB* modela

x_i, uzorak koji želimo da klasifikujemo primenom *SC*

δ , parametar učenja

Broj stabala (N) i parametar učenja (δ) su hiperparametri *GTB* algoritma, koji značajno mogu uticati na tačnost prediktivnog modela. Veći broj stabala može dovesti do veće tačnosti, ali i do pojave *overfitinga*. Ukoliko se odlučimo za veće vrednosti parametra δ, potreban je manji broj stabala u modelu. Pravilnim izborom ovih parametara (najčešće postupkom eng. *cross validation*),

[18] Izvor: Ml-gradient-boosting

možemo doći do modela zadovoljavajuće tačnosti i dobre sposobnosti generalizacije (Friedman, 2001).

Pored *Adaboost* i *GTB*, često primenjivani *boosting* klasifikacioni algoritam je i *XGBoost* (eng. *Extreme gradient boosting*). Sličan je *GTB*, ali kako omogućava paraleleno procesiranje, ima bolje performanse. *XGBoost,* za razliku od *GDB,* u prvoj iteraciji ima fiksnu vrednost inicijalne prediktivne verovatnoće jednaku 0.5. Stabla odlučivanja se isto granaju, na osnovu graničnih vrednosti prediktora, za koje se maksimalno poveća homogenost uzoraka u nodovima koji grananjem nastaju. Jedinstvena vrednost, output vrednosti *R*, u terminalnim nodovima regresionog stabla kod *XGBoost modela* jednaka je:

$$R_{j,m} = \frac{\sum_{i \epsilon j,m} r_i}{\sum_{i \epsilon j,m} p_{i,}(1 - p_i) + \vartheta}$$

$\vartheta -$ regularizacioni parametar

Sa rastom vrednosti parametre ϑ povećava se stabilnost modela i može se dodatno uticati na njegovu mogućnost generalizacije.

Primenom *boosting* metoda dodatno se povećava tačnost *DT* prediktivnih modela. Ove algoritme preporučeno je primenjivati u slučajevima velikog broja uzoraka (više od 1.000) i broja prediktora (do 100) (Kassambra, 2017).

4. EVALUACIJA PERFORMANSI KLASIFIKACIONIH MODELA

U ovom delu opisani su osnovni kriterijumi koji se koriste za evaluaciju performansi klasifikacionih algoritama(Baesens, 2015):

- Matrica konfuzije

Tačnost klasifikacionih algoritama možemo meriti primenom matrice konfuzije (eng. *confusion matrix*), koja prikazuje broj tačno i netačno klasifikovanih uzoraka za svaku klasu zavisne promenljive (Gareth et al., 2014). U slučaju binarno klasifikacionih problema (pozitivna i negativna klasa), matrica je kvadratna drugog reda. Broj uzoraka pozitivne klase koje prediktivni model tačno klasifikuje, nazivamo tačnim pozitivima (eng. *true positives*, *TP*). Netačni pozitivi (eng. *false positives*, *FP*) su uzorci negativne klase, klasifikovani kao pozitivni. Uzorci negativne klase koji su tačno klasifikovani su tačni negativi (eng. *true negatives*, *TN*). Broj netačnih negativa (eng. *false negatives*, *FN*) su pozitivni uzorci, pogrešno klasifikovani kao negativni. Ukupan broj pozitivnih uzoraka jednak je $P = TP + FN,$ a negativnih $N = TN + FP$.

Tabela 9. Matrica konfuzije

		Tačna klasa	
		Pozitivna	Negativna
Predviđena klasa	Pozitivna	TP	FP
	Negativna	FN	TN

Performanse klasifikacionih modela primenom ove matrice možemo meriti na više načina (tabela 10).

Tabela 10. Mere performansi binarnog klasifikacionog modela

Mere performansi klasifikacionog modela	
Tačnost (*Accuracy*)	$\frac{TP+TN}{P+N}$
Greška (*Error rate*)	1-Tačnost
Stopa tačnih pozitiva *TP,* (*Sensitivity*)	$\frac{TP}{P}$
Stopa netačnih pozitiva, *FP*	$\frac{FP}{N}$
Preciznost (*Precision*)	$\frac{TP}{TP+FP}$
Specifičnost (*Specificity*)	$\frac{TN}{N}$

U zavisnosti od predmeta istraživanja, cilj može biti postizanje maksimalne senzitivnosti modela. U studiji slučaja, koja je predmet ove disertacije (kompanija u stečaju i slučaj credit *default*-a su pozitivne klase), prioritet je postizanje maksimalnog broja pravilno klasifikovanih uzoraka pozitivne klase (*TP*). Promenom granične vrednosti verovatnoće, čija je podrazumevana vrednost 0.5, može se uticati na senzitivnost prediktivnog modela. Smanjivanjem granične vrednosti raste broj *TP* (smanjuje *FN*), a time i senzitivnost modela. Istovremeno, ukupna tačnost modela se smanjuje, usled smanjenja specifičnosti, odnosno broja *TN (*raste *FP).*

Optimalnu vrednost granične verovatnoće, za koju bi stopa *TP* bila optimalna, dobijamo primenom metode *ROC* krive (eng. *Receiver operating charactheristic curve*), koja predstavlja grafičku simulaciju vrednosti stopa *TP* i *FP* za različite granične vrednosti verovatnoće (Kassambara, 2017). U početnom delu *ROC* kriva je velikog nagiba, kada za minimalne vrednosti rasta *FP* značajno raste senzitivnost modela. U njenom drugom delu, dolazi do zaravnjenja *ROC* krive; tako za minimalne dobitke, odnosno rast stope *TP*, značajno raste *FP*. Optimalna vrednost granične verovatnoće modela je ona za koju je zbir vrednost senzitivnosti i specifičnosti modela najveći (Kassambara, 2017).

Slika 36. *ROC* kriva prikazuje kako se menja stopa *TP* (senzitivnost) i stopa $FP = 1 - \frac{TN}{N}$ za različite vrednosti granične verovatnoće[19]

- *AUC* (eng. *Area under the curve, AUC*)

je površina obuhvaćena *ROC* krivom. Ima maksimalnu vrednost jedan i predstavlja procenat pravilno klasifikovanih uzoraka (Kassambara, 2017). Klasifikacioni model sa najvećom vrednošću *AUC* je najoptimalniji model, nezavisno od izabrane granične vrednosti verovatnoće. *AUC* je pokazatelj sposobnosti modela da razdvoji uzorke različitih klasa (eng. *discriminatory ability)*. Nebalansiranost podataka, usled znatno većeg prisustva uzoraka jedne klase, ne utiče na *AUC* vrednost (Fawcett, 2006). Statistički značaj razlika u vrednostima *AUC*, različitih *ML* modela, može se utvrditi primenom *Delong* testa (DeLong, 1998).

- *Kappa* indeks (Cohen, 1960)

Iz matrice konfuzije možemo dobiti još jedan važan pokazatelj tačnosti klasifikacionog algoritma, nazvan *Kappa* indeks (Cohen, 1960). Dok Tačnost (*Accuracy*), ne pravi razliku između tačne klasifikacije, koja je rezultat postojanja preovlađujuće klase (većinske) i tačnosti dobijene

[19] Izvor: https://commons.wikimedia.org/wiki/File:Roc_curve.svg

primenom prediktivnog modela, *Kappa* nagrađuje klasifikator samo kada je tačnost predikcije veća od tačnosti dobijene kada bi se uvek predviđala (birala) većinska klasa. Vrednost ovog indeksa je u rangu [0,1], veća vrednost označava bolji model.

- *F1 score*

*F1 j*e funkcija preciznosti modela (procenat *TP* uzoraka u odnosu na broj predviđenih pozitiva) i njegove senzitivnosti (*Recall*) – mogućnost modela da identifikuje *TP* od ukupnog broja stvarno pozitivnih uzoraka. Tako je u slučaju nejednake distribucije uzoraka različitih klasa (nebalansirani podaci) tačnost (*Accuarcy*) modela očekivano visoka (model tačno predviđa uzorke većinske klase) i ona nije dobar pokazatelj performansi, već je to *F1 score,* koji pokazuje sposobnost modela da tačno predvidi *TP* (uzorke manjinske klase).

$$F1\ score = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = \frac{2 * (Precision * Recall)}{(Precision + Recall)}$$

ili

$$F_\beta = (1 + \beta^2) * \frac{(Precision * Recall)}{(\beta^2 * Precision) + Recall}$$

Beta indeks označava koliko puta je senzitivnost modela važnija od njegove preciznosti.

Veća vrednost *F1 score* označava bolji model.

- *Brier score* (Glenn W. Brier, 1950)

je mera tačnosti probabalističkih *ML* modela. Jednak je srednjoj vrednosti kvadrata reziduala. Primenjuje se u svrhu procene tačnosti predikcije pozitivne klase. Rang mogućih vrednosti je od 0 do 1. Perfektni klasifikacioni modeli imaju vrednost ovog indeksa nula.

5. PRETPROCESIRANJE I TRANSFORMACIJA PODATAKA

U zavisnosti od vrste klasifikacionog algoritma, strukture i kvaliteta podataka, primenjuju se različite metode pretprocesiranja i transformacija, kojima se podaci dovode u oblik koji će omogućiti pouzdanu predikciju. U nastavku ovog poglavlja prikazane su osnovne metode za pretprocesiranje i transformaciju podataka.

5.1. *SCALING* Podataka

Kada podaci, kao u našem slučaju, imaju različite jedinice mere, kao i raspone vrednosti, podaci se transformišu primenom s*caling* metoda (skaliranje), kojima se svode na isti rang vrednosti ili na istu srednju vrednost (jednaku nuli). Ovim postupkom uticaj svake promenljive u modelu postaje nezavisan od njene apsolutne vrednosti. Ukoliko bi se ova transformacija izostavila, *ML* algoritam bi veći značaj u modelu (npr. veća vrednost odgovarajućeg regresionog koeficijenta) dodelio promenljivoj koja ima veću vrednost. Najčešće primenjivani *scaling* postupci su:

- *Normalizacija*, kojom se numeričke vrednosti prediktora svode na isti rang, dok se oblik distribucije ne menja. Najčešće se koristi *min-max* normalizacija, kojom se vrednosti promenljive svode na rang vrednosti $[0,1]$.

$$\frac{x - x_{min}}{x_{max} - x_{min}}$$

Primenjujemo je u slučajevima nepoznate distribucije prediktora (ili kad ta distribucija nije *Gaussian*), za algoritme koji su zavisni od rastojanja između uzoraka (*SVM, K-NN*).

- *Standardizacija* ili *z* Normalizacija je *scaling* metod kojim se srednja vrednost promenljive svodi na 0, a standardna devijacija na 1. Preporučuje se u *ML* modelima u kojima se kao optimizacijski algoritam koristi *gradient descent* (Regresioni modeli, *LDA*).

5.2. Nedostajući podaci

Prisustvo *NA* (eng. *no available*), nedostajućih podataka, utiče negativno na efikasnost *ML* algoritma, a može prouzrokovati i algoritamsku pristrasnost (eng. *algorithmic bias*). Na osnovu raspoloživih podataka, koji sadrže veći broj uzoraka sa *NA*, *ML* algoritam u postupku učenja raspoznaje strukturu i relacije između podataka, koju generalizuje na nove podatke. Ovakva predikcija će biti pristrasna kada novi (testni) podaci nemaju ili imaju znatno manje prisustvo *NA*.

U zavisnosti od toga da li je postojanje *NA* uslovljeno vrednostima drugih promenjivih, razlikujemo sledeće slučajeve (Allison, 2001; Schafer et al., 2002):

- *MCAR* (eng. *Missing clompletely at random*), kada ne postoji zavisnost između vrednosti promenljive x_i (koja sadrži *NA*) i ostalih promenljivih, uključujući i x_i
- *MAR* (eng. *Missing at random*), kada na pojavljivanje *NA* promenljive x_i utiču druge promenljive, ali ne preostale vrednosti x_i
- *NIM* (eng. *non ignorable missing*), kada pojavljivanje *NA* promenljive x_i samo zavisi od preostalih vrednosti x_i
- *MNAR* (eng. *Missing not at random*), kada postoji relacija između *NA,* promenljive x_i i drugih promenjivih, uključujući i x_i

Kako u našoj studiji slučajeva u podacima postoji znatno prisustvo *NA*, pristupilo se imputaciji nedostajućih podataka.

U slučaju *MCAR* nedostajući podaci se mogu zameniti srednjom vrednošću te promenljive. Na ovaj način srednja vrednost promenljive x_i ostaje ista. Njena varijansa se međutim menja, usled promene distribucije x_i, kao posledica veće koncentracije podataka oko srednje vrednosti. Little et al. (1989) tako smatraju da je bolje uzorke sa nedostajućim vrednostima izostaviti nego ih zameniti srednjim vrednostima.

U slučajevima *MAR*, *NIM* i *MNAR*, kada je pojava *NA* zavisna od drugih promenljivih, rekonstrukciji nedostajućih podataka najčešće se pristupa drugim tehnikama, i to pre svega metodama maksimalne verodostojnosti – *MaxL* i regresije (Little i Rubin, 1987). Primenom metode *MaxL* dolazimo do najverovatnijih vrednosti parametara distribucije promenljive x_i, koja sadrži *NA*. Kad nam je distribucija verovatnoće x_i poznata, *NA* zamenjujemo vrednošću koja ima najveću verovatnoću pojavljivanja. Kada je količina podataka za analizu velika, za rešavanje problema *NA* preporučuje se *MaxL* metod (Schafer i Graham, 2002). Primenom regresije, zavisna promenljiva postaje promenljiva sa *NA* (x_i). Tako regresioni model predviđa vrednosti *NA* na osnovu poznatih podataka preostalih promenljivih.

U slučaju klasifikacionih problema, praktikuje se i primena *Random Forest* algoritma u rekonstrukciji *NA*. Ovaj postupak je primenjen u studiji slučaja ove doktorske disertacije.

5.3. Ekstremne vrednosti

Prisustvo ekstremnih vrednosti može značajno uticati na stabilnost i tačnost prediktivnih modela, i to pre svega: Linerane i Logističke regresije, *SVM*, *KNN*, *PCA*. Autlajerima (eng. *outliers*) se nazivaju ekstremne vrednosti zavisne promenljive, dok u slučaju ekstremnih vrednosti nezavisne promenljive govorimo o podacima sa velikim uticajem (eng. *high leverage*). Uzorci sa ekstremnim vrednostima se najčešće isključuju iz analize, ili se te vrednosti zamenjuju svojom srednjom vrednošću (ili medijanom). Postoji više metoda kojima se utvrđuje postojanje ekstremnih vrednosti:

- *Box and whisker* grafikon prikazuje raspodelu podataka po kvartilima, tako se ekstremne vrednosti mogu lako uočiti. Interkvartilna razlika jednaka je:

$$IQR = Q_3 - Q_1$$

Q_1 je vrednost veća od 25% podataka te promenljive

Q_3 je vrednost veća od 75% podataka te promenljive

Vrednosti manje od $Q_1 - 1.5\ IQR$ smatraju se ekstremno malim, dok se vrednosti veće od $Q_3 + 1.5\ IQR$ smatraju ekstremno velikim.

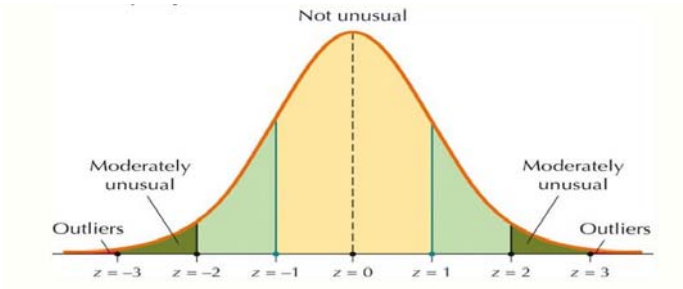
- U slučaju *ML* modela, koji pretpostavljaju normalnu distribuciju prediktora, postojanje ekstremnih vrednosti se može utvrditi primenom *Z- scores* metode. Vrednosti prediktora x za koje je $z > 3$ mogu se smatrati ekstremnim

$$z = \frac{x - \mu}{\sigma}$$

μ, srednja vrednost x

σ, standardna devijacija x

Slika 37. Određivanje ekstremno velikih/malih vrednosti podataka primenom z- Scores[20]



- U slučaju regresionih modela uticajne ekstremne vrednosti određujemo primenom *Cook* rastojanja (eng. *Cook's distance*). Vrednost rastojanja (D_i) pokazuje za koliko bi se promenio regresioni model kada bi uzorak i bio isključen iz modela:

$$D_i = \frac{\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_{ji})^2}{m * MSE}$$

m, broj prediktora

n, ukupan broj uzoraka

$\hat{y}_{ji}$, predikcija modela, bez uzorka *i*

$MSE = \frac{\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}{n-m}$, (eng. *mean square error*)

Uobičajeno je da se vrednosti za koje je $D_i > \frac{4}{n}$ smatraju ekstremnim (eng. *rule of thumb*).

[20] **Izvor:** : http://slideplayer.com/slide/6394283/

5.4. Multikolinearnost

Postojanje korelacije između prediktora takođe negativno utiče na regresione modele tako što se smanjuje njihova stabilnost i tačnost predikcije. Kako parametri regresionih modela pokazuju uticaj prediktora na vrednost zavisne promenljive, zbog postojanja korelacije sa drugim prediktorima, ovaj uticaj se ne može posmatrati izolovano. Postojanje korelacije može se utvrditi računanjem *Pearson* koeficijenta korelacije , koji ima vrednost -1 i +1. Kada koeficijent ima vrednost nula, korelacija ne postoji. Problem multikolinearnosti (kada se jedan prediktor može prikazati kao funkcija više drugih prediktora), može se rešiti regularizacijom regresionih modela – *Lasso* i *Ridge*, kada se smanjuje uticaj zavisnog prediktora ili se isti isključuje iz modela. Pored ovog pristupa, moguće je problem multikolinearnosti rešiti kreiranjem novih promenljivih, koje su međusobno nezavisne, a izražene su u funkciji originalnih prediktora. Ove promenljive se nazivaju principalne komponente (*PC*), a postupak Analiza principalnih komponenti – *PCA* (eng. *Principal Compnent analysis*).

- *PCA* je nenadgledani, neparametarski metod mašinskog učenja, kojim je moguće redukovati dimenzionalnost podataka, kako bi prediktivni regresioni i klasifikacioni modeli bili funkcija manjeg broja novoformiranih promenljivih, kojima se na dovoljno dobar način može objasniti varijabilnost podataka. Primenom *PCA* originalni podaci (prediktori) se linearnom transformacijom mapiraju u prostor definisan novim koordinatnim osama – Principalnim komponentama (eng. *principal components, PC*), koje određuju pravce njihove maksimalne varijacije (disperzije). Redukcija dimenzionalnosti se postiže tako što se prediktivni model 'uči' samo na onim *PC* koje u najvećoj meri objašnjavaju varijabilnost podataka.

  PCA primenjujemo u slučajevima kada zbog velikog broja prediktora postoji mogućnost pojave *overfitinga*. Lineranom *PCA* transformacijom prediktora rešavamo i problem nestabilnosti modela, prouzrokovanu postojanjem korelacije između prediktora.

Metoda *PCA* daje važne informacije o varijabilnosti podataka, utvrđivanjem pravaca u multidimenzionalnom prostoru u kojima podaci imaju najveću disperziju. Pravci principalnih komponenti u kojima je varijabilnost podataka najveća, smatraju se najinformativnijim, najvažnijim, eng. *'the most principal'* (Jolliffe, 2002). Linearnu transformaciju kojom se uzorci iz originalnog *m* dimenzionalnog prostora, u kojem je svaki uzorak jednoznačno određen vektorom $x_i \epsilon R^m$, mapiraju u novi prostor, određen pravcima principalnih komponenti $PC_1 \dots PC_m$, možemo prikazati kao

$$PC_1 = \beta_{1,1}x_1 + \beta_{2,1}x_2 + \ldots + \beta_{m,1}x_m \quad (1)$$

....

$$PC_m = \beta_{1,m}x_1 + \beta_{2,m}x_2 + \cdots + \beta_{m,m}x_m$$

Koeficijenti ove transformacije (β), koji se nazivaju i *loading* koeficijenti principalnih komponenti (Gareth et al., 2014), jesu jedinični sopstveni vektori (eng. *eigenvectors*), kovarijantne matrice prethodno transformisanih podataka (srednja vrednost svih prediktora mora biti jednaka nuli).

Kovarijantna matrica ima oblik:

$$COV(X) = E[(X - E[X])(X - E[X])^T]$$

COV (*X*), kovarijantna matrica podataka *X*

Dijagonalni elementi ove matrice pokazuju disperziju prediktora, njihovu varijansu:

$$Var(X) = E[(X - E[X])^2]$$

Vandijagonalni elementi predstavljaju kovarijansu između prediktora. Kada je ona pozitivna, prediktori su pozitivno linearno zavisni, kada je negativna, prediktori su negativno linearno zavisni. Sopstveni vektori (*eignevectors*) ove matrice određuju pravce najveće disperzije podataka. Odgovarajuće sopstvene vrednosti (*eigenvalues*) predstavljaju magnitudu jediničnih vektora, odnosno intenzitet disperzije.

Ako označimo sa $V_1, .., V_m$ sopstvene vektore kovarijantne matrice *COV(X),* a njihove sopstvene vrednosti sa $\lambda_1, \lambda_2, .., \lambda_m$, sledi:

$$COV(X)\, V = \lambda\, V \tag{2}$$

Odnosno:

$$(\, COV - \lambda I)V = 0$$

Kako *V* nije nula vektor, sledi da je:

$$\det(COV - \lambda I) = 0 \tag{3}$$

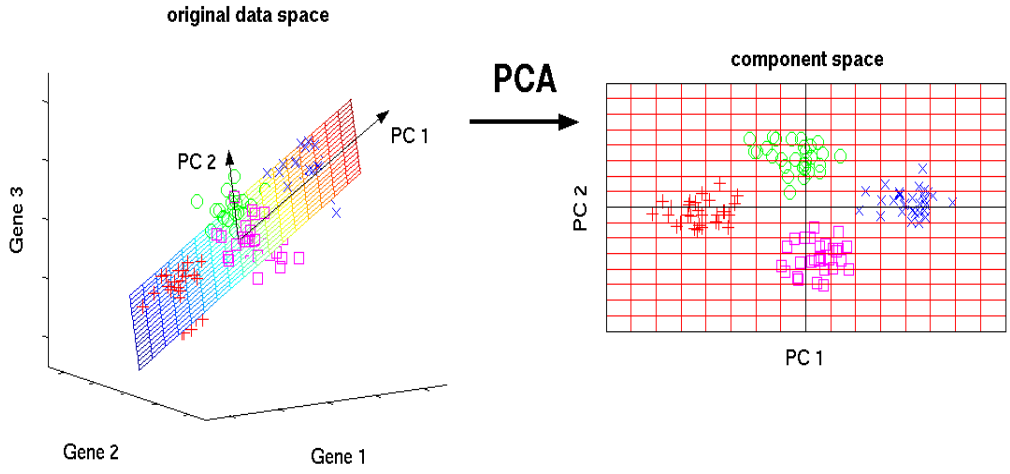
I, jedinična matrica

Iz ovog uslova (3) dobijamo vrednosti za $\lambda_1, \ldots \lambda_m$, a potom iz (2), $V_1, .., V_m$. Ako dobijene sopstvene vrednosti uredimo prema opadajućim vrednostima, tako da je $\lambda_1 < \lambda_2 < .. < \lambda_m$, tada λ_1 predstavlja intenzitet odgovarajućeg sopstvenog vektora V_1, koji određuje pravac maksimalne disperzije uzoraka, odnosno pravac prve principalne komponente *PC1*. Sopstvena vrednost λ_2 je intenzitet sopstvenog vektora V_2, koji određuje pravac principalne komponente *PC2* (pravac manje disperzije od one definisane sa PC_1). Pravac PC_m je pravac najmanje disperzije podataka. Jedinične

vrednosti sopstvenih vektora ($\frac{V_i}{|V_i|}$) predstavljaju keficijente linerne transformacije $\beta_i = (\beta_{1i,}\beta_{2i,..,}\beta_{mi})^T$.

Kako su sopstvene vrednosti λ mera varijabilnosti podataka, one se koriste i kako bi se odredio konačni broj *PC* u transformisanom prostoru, manje dimenzionalnosti (Kaiser, 1961). Tako samo one *PC* za koje su vrednosti odgovarajućih sopstvenih vrednosti veće od jedan (to su pravci veće disperzije podataka od disperzije u pravcima osa originalnog prostora), definišu novi prostor definisan osama $PC_1, PC_2, .. PC_p$, gde je $p < m$.

Slika 38. Primer transformacije podataka iz trodimenzionalnog prostora (m=3, $x_1 = Gene\ 1, x_2 = Gene\ 2, x_3 = Gene\ 3$), u dvodimenzionalan prostor određen osama PC_1 i PC_2 ($p=2$). PC_1 pravac maksimalne disperzije podataka ($\lambda_1 > \lambda_2$, $\lambda_1\ i\ \lambda_2 > 1$)[21]



Kako je svaka *PC* upravna na prethodnu ($PC_i \perp PC_{i-1}$), *PC* su međusobno nezavisne. Tako *PCA,* pored smanjivanja dimenzionalnosti, eliminiše i problem multikolinearnosti prediktora, koji u

[21] Izvor: Nlpca.org

regresionim modelima uzrokuje netačnu parametarizaciju modela, kao i moguće izostavljanje iz modela važnih prediktora (usled niskog statističkog značaja) (Graham, 2003).

Primenom *PCA* postižemo i smanjenje uticaja autlajera na prediktivne modele. Kako je PC_1 pravac najveće disperzije, to je ujedno i pravac sa najmanje autlajera. Svaka naredna principalna komponenta određuje pravac manje disperzije. Tada je mogućnost pojave autlajera veća, ali je i uticaj takve *PC* na model manji.
Važno je napomenuti da linearnom transformacijom prediktora u *PC* i redukcijom dimenzionalnosti iz originalnog prostora R^m u prostor manje dimenzionalnosti R^p, neće doći do gubitka originalnih podataka jer je svaka *PC* linearna funkcija svih raspoloživih prediktora i njihovih vrednosti.

5.5. Nebalansirani podaci

Originalna struktura podataka koji su predmet empirijskog istraživanja ima izrazito nebalansiranu distribuciju (oko 1% uzoraka pripada manjinskoj klasi). Ovo može dovesti do pristrasnog (eng. *biased*) prediktivnog modela, usled činjenice da je model treniran na uzorcima koji većinski pripadaju jednoj klasi. Prema Wiess i Provost (2003), procenat manjinske klase u trening podacima treba biti u granicama 50%–90 %, kako bi se povećala tačnost predikcije na nebalansiranim podacima. Iz ovog razloga pristupa se *over-sampling*- u manjinske klase, najčešće tehnikom *Rwo* (Zhang i Li, 2014), *SMOTE* i *ROSE*:

- *RWO* (eng. *Random walk over-sampling approach)* je tehnika kojom se novi uzorci manjinske klase generišu na način da se ne promeni varijansa i srednja vrednost uzoraka te klase.
- *SMOTE (*eng. *Synthetic minority over-sampling technique)* (Chawla et. al., 2002), tehnika kojom se primenom K- NN na *boostrap* uzorcima manjinske klase generišu novi.

- *ROSE (*eng. *Random over sampling)* (Menardi, Torelli, 2014), metoda kojom se slučajno izabrani uzorci manjinske klase dodaju u trening data set (*bootstraping*), dok se slučajno izabrani uzorci većinske klase uklanjaju iz trening podataka. Primenjuje se u binarno klasifikacionim problemima.

6. REZULTAT EMPIRIJSKOG ISTRAŽIVANJA

Performanse različitih modela analizirane su na dva nezavisna skupa podataka. Cilj analize je istraži ponašanje algoritama na skupovima podataka koji predstavljaju dva najčešća tipa podataka koja se koristi u predviđanju nenaplativih potraživanja – potraživanja prema kompanijama i prema pojedincima. Dalje u radu referisaćemo se na *„Default of credit card clents Data",* kao prvi i *„Polish companies bankruptcy Data Set''* kao drugi set podataka. Podaci su preuzeti sa *UCI Machine Learning Repository*[22] (osnivač *Universiy of California, Irvine*) je kolekcija baza i generatora podataka koji se koriste za empirijske analize algoritama *ML*. Od osnivanja 1987, *UCI* predstavlja osnovni izvor podataka za studente, predavače, istraživače, koji ih primenjuju u *ML*. Citiran je kao izvor podataka preko 1. 000 puta, što prema broju citata spada u top 100 radova iz oblasti *ICT*. Detaljan opis podataka dat je u Aneksu 1 rada. U naredna dve sekcije date su osnovne informacije o uzorku i izvršenoj pripremi podataka.

6.1. Opis uzorka

Polish companies bankruptcy Data Set[23]. Autor je Sebastian Tomczak, *Department of Operations Research, Wroclaw University of Science and Technology*, koji je kao izvor koristio *EMIS - Emerging Markets Information Service*[24]. EMIS sakuplja i analizira podatke sa preko 125 tržišta (eng. *emerging markets*).

Podaci predstavljaju skup finansijskih indikatora poslovanja poljskih kompanija koje su otišle u stečaj u periodu od 2000. do 2012. godine, kao i onih koje su nastavile nesmetano da posluju do 2013. Podaci su strukturirani, numerički. Istraživanjem je obuhvaćeno 64 različita finansijska indikatora koji utiču na verovatnoću stečaja kompanija. Zavisna promenljiva je binarna kategorička, sa vrednostima 0 (za kompanije koje nisu u stečaju) i 1 (za kompanije u stečaju). Primenjena je sledeća notacija, kojom se podaci predstavljaju u matričnom obliku:

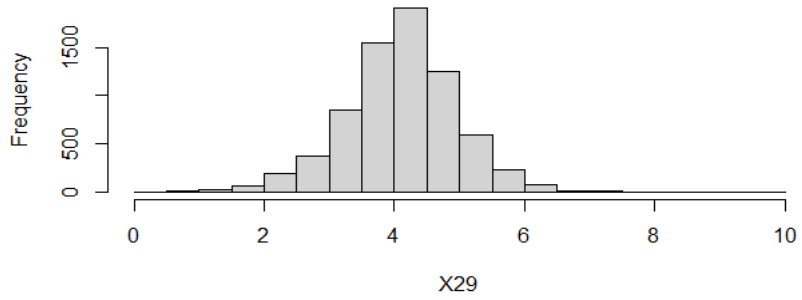
[22] Izvor: UCI Machine Learning Repository: About
[23] Izvor: https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data
[24] Izvor: EMIS

Prediktori (*Attr*)	$X = \{x_i \epsilon R^{64}\}, \quad i = \{1,..,9792\}$
Zavisna (target) promenjiva (*Classs*)	$Y = \{y_i \epsilon\ [0,1]\}, \quad i = \{1,..,9792\}$

Prilog A.1, tabela 12

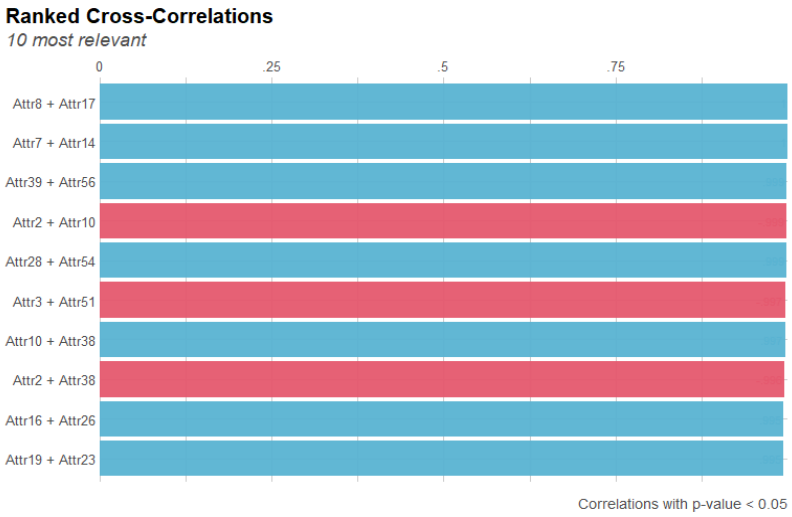
Ukupan broj jedinica postmatranja u uzorku je 9.792, sa 65 varijabli, od kojih su 64 numerički prediktori, dok je zavisna promenljiva binarna kategorička, sa vrednostima 0 i 1. Izračunate su srednje i središnje vrednosti (medijana), kvartili Q1 i Q3, kao i vrednosti trećeg i četvrtog momenta empirijskog rasporeda verovatnoće (eng. *skewness* i *kurtosis*) za sve numeričke varijable. Proveren je oblik rasporeda verovatnoće svih promenljivih. Visoke negativne i pozitivne vrednosti *skewness* ukazuju na nesimetričnu distribuciju većine varijabli. Takođe, visoke vrednosti *kurtosis*-a ukazuju na *Leptokurtic* raspored verovatnoće, za koju je karakteristično da su vrhovi rasporeda tanji i visoki, a repovi zadebljani. Samo promenljiva Attr29, ima vrednosti *skewness* između -1 i 1, a *kurtosis* -3 and 3, i može se smatrati da ima približno normalnu distribuciju (logaritam vrednosti kompanijskih aseta).

Slika 39. Histogram promenljive Attr29



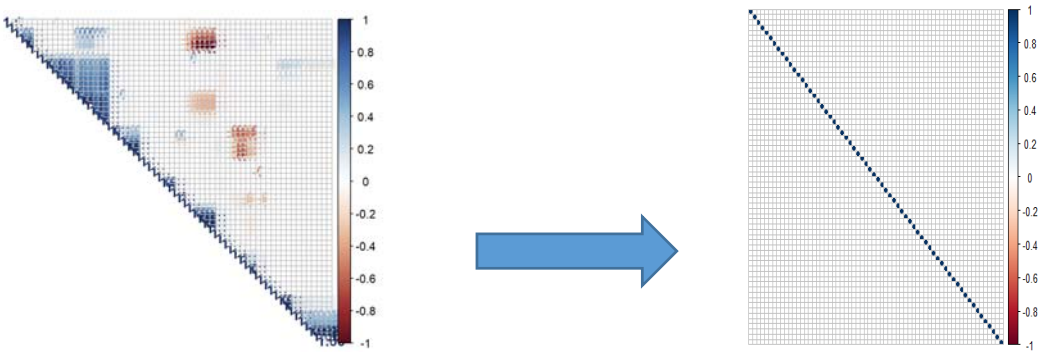
Statistički značajna (*p-values* < 0.05) korelacija, postoji između većine varijabli. Ovo ukazuje na postojanje multikolinearnosti.

Slika 40. Parovi varijabli sa pozitivnom (plava boja) i negativnom korelacijom



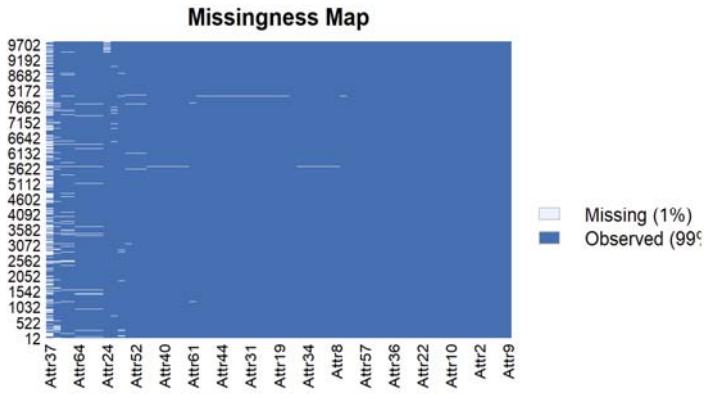
Podaci su transformisani u principalne komponente (*PC)*, čime je ujedno omogućeno smanjivanje dimenzionalnosti modela. U algoritmima koji su osetljivi na postojanje linearne zavisnosti varijabli, kao u slučaju logističke regresije, prediktivni model je razvijen i na osnovu *PC*.

Slika 41. Kovarijantna matrica (64x64), posle transformacije verijabli u principalne komponente. Tamnoplava boja po dijagonali označava varijansu svakog prediktora (vrednost 1), dok je kovarijansa prediktora 0.



Nedostajućih podataka (*NA*) ima ukupno 8,776 za sve jedinice posmatranje i posmatrane promenljive, i to najviše za promenjivu Attr37 (*current assets - inventories) / long-term liabilities*). Imputacija nedostajućih vrednosti realizovana je primenom *RF* algoritma, a na osnovu poznatih vrednosti preostalih prediktora. Zavisna promenjiva je kompletna, bez nedostajućih vrednosti. Na nivou uzorka 4,769 jedinica posmatranje nema nedostajuće podatke ni za jednu promenljivu

Slika 42. Mapa nedostajućih vrednosti



U pripremi podataka izvršena je transformacija ekstremnih vrednosti, kako bi se eleminisao njihov negativan efekat na prediktivne performanse modela. Ekstremne vrednosti su identifikovane kao one koje se nalaze van granica $Q_1 - 1.5\,IQR$ i $Q_3 + 1.5\,IQR$. Ove vrednosti zamenjene su odgovarajućim medijanama.

Kako je Attr55 (*working capital*) dat u apsolutnom, a preostali finansijski indikatori u relativnom iznosu, postupkom standardizacije (zNormalizacija), podaci su svedeni na uporedive rangove vrednosti.

Podaci su podeljeni u trening i testni (eng. *out of sample*) podskup u odnosu 70:30. Kako je prisustvo uzoraka manjinske klase samo 5.3% ($y = 1,$ kompanije koje su u stečaju), pristupilo se *upscaling* – u, primenom SMOTE metode. Algoritmi koji su postigli najbolje performanse (Rf i GTB), testirani su i na balansiranim podacima.

Default of credit card clients Data[25].Autor je I-Cheng Yeh sa Department of Information Management, Chung Hua University, Taiwan. Ovaj skup podataka se odnosi na analizu rizika nevraćanja kredita od strane pojedinaca.

Primenjena je sledeća notacija, kojom se podaci predstavljaju u matričnom obliku:

Prediktori	$X = \{x_i \epsilon R^{24}\},$	$i = \{1,..,30{,}000\}$
Zavisna (target) promenljiva	$Y = \{y_i \epsilon\ [0,1]\},$	$i = \{1,..,30{,}000\}$

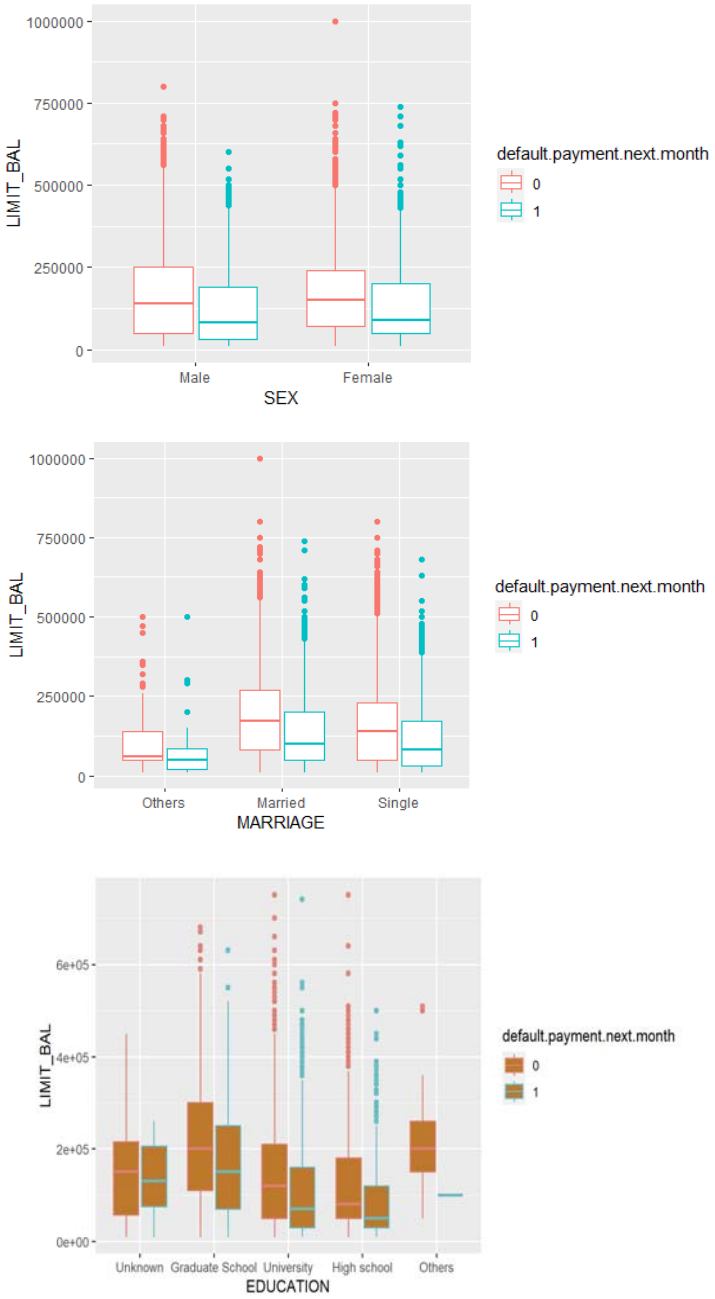
Prilog A.1, tabela 13

Ovaj uzorak ima 30,000 jedinica posmatranja, sa 25 varijabli. Zavisna promenjiva, *default.payment.next.month* ima vrednosti y = 0, za klijente koji su kredit vratili i y= 1, za one koji nisu. Za svakog klijenta postoje podaci za vrednost uzetog kredita, pol, obrazovanje, bračni status, istoriju prethodnih plaćanja po mesecima (za period april-septembar 2005 godine), zaduženja za isti period, kao i prethodna plaćanja. Broj klijenata koji su kredit vratili, prema broju onih koji nisu (balansiranost podataka) je 0.77/ 0.23. Podaci su kompletni, ondnosno nema nedostajućih vrednosti.

Prediktori su kombinacija numeričkih i kategoričkih promenljivih (*MARRIAGE, SEX* i *EDUCATION*). Kategoričke promenljive imaju minimalan uticaj na y (*default.payment.next.month*). Vrednost *Tetrachoric* korelacionog koeficijenta (Pearson, 1900), kao mere međusobne zavisnosti binarnih kategoričkih promenljivih, u slučaju *SEX* i *default.payment.next.month* je zanemarljiva (-0.05). *Cramer's V* (Harald Cramer, 1946) koeficijent, kao mera zvisnosti nominalnih kategoričkih promenljivih sa dve i više klasa, u slučaju *EDUCATION* i *default.payment.next.month* ima vrednost 0.07, a za *MERRIAGE* i *default.payment.next.month* samo 0.03.

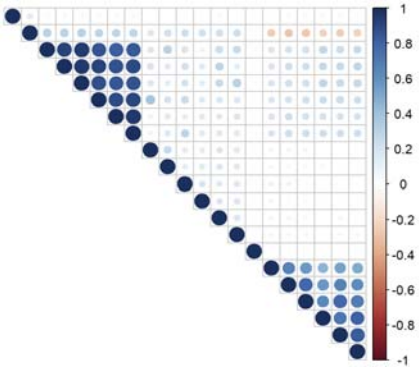
[25] UCI Machine Learning Repository: default of credit card clients Data Set

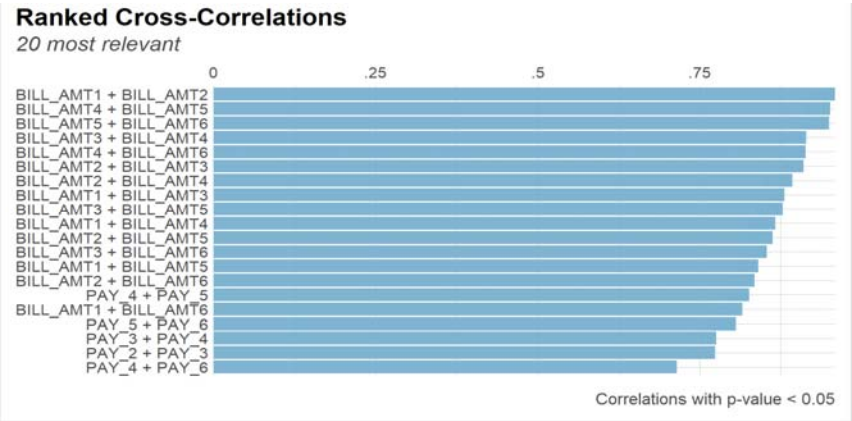
Slika 43. Uticaj LIMIT_BAL i kategoričkih promenjivih na y (default.payment.next.month)



Između numeričkih prediktora postoji značajna međusobna zavisnost.

Slika 44. Kovarijantna matrica numeričkih prediktora i prikaz prediktora sa najvećom međusobnom zavisnošću





Numerički prediktori transformisani su u PC. Podaci su ponovo podeljeni u trening i test (eng. *out of sample*) podskup u odnosu 70:30.

6.2. Regresioni modeli

Iako spadaju u grupu jednostavnijih klasifikacionih algoritama, regresioni linearni modeli su na oba *data* seta postigli dobre performanse, uprkos činjenici da pretpostavke o postojanju linearne

zavisnosti prediktora u modelu sa $\log(\frac{p}{1-p})$ i nezavisnosti prediktora, nisu ispunjene. Na performanse ovih modela nebalansiranost podataka je imala minimalan uticaj, što je u skladu sa (Baesens et al., 2003).

Pokazano je da se modelom logističke regresije, na podacima transformisanim u međusobno nezavisne *PC*, kojima se može objasniti 95% kumulativne varijabilnosti prediktora, postigla približno ista vrednost *AUC* u odnosu na originalni, početni model logističke regresije (*LR*). Tako je model *LR* sa 24 varijable prvog *data* seta sveden na model *LRpc* sa 14 *PC*, kojim je postignuta vrednost *AUC* 0.741, identična vrednosti originalnog modela logističke regresije. Model *LR* sa 64 prediktora drugog seta podataka sveden je na 38 *PC*. Postignuta *AUC* vrednost ovog modela je 0.763, neznatno manja od vrednosti *AUC* modela logističke regresije *LR* od 0.773.

Značajnije smanjivanje dimenzionalnsti podataka moguće je korišćenjem samo *PC* za koje su odgovarajuće sopstvene vrednosti (*eigenvalues)* veće od jedan. U primeru prvog data seta, kada je broj *PC* je sveden na tri, postignuta je vrednost *AUC* od 0.732 , što je samo za 1.2% manje u odnosu na originalni model *LR*. U slučaju drugog data seta, sa 64 prediktora, broj *PC* komponenti sa sopstvenim vrednostima većim od jedan je 13, i postignut *AUC* takvog modela je 0.75, što je 2.8% manja od vrednosti *AUC* modela *LR*.

Kako su *PC* međusobno nezavisne, rezultati su pokazali da se eliminacijom problema multikolinearnosti, malo utiče na tačnost klasifikacionih algoritama, a više ne vrednosti regresionih koeficijenata (Bruce, 2017).

Lasso regularizacijom (*Lasso LR*) primenom *L1 penalty term*-a , iz modela su eliminisane 2 promenljive prvog data seta i 33 drugog, svođenjem odgovarajućih regresionih koeficijenata na vrednost nule. Vidimo da su *Ridge* regresijom (*Ridge LR)*, primenom *L2 penalty term*-a, dakle uz prisustvo svih prediktora u modelu, i *Elastic Net*, metodom kojim se kombinuje *L1* i *L2 penalty term,* postignute skoro identične performanse sa *Lassomo* modelom, a samo neznatno slabije performanse u odnosu na model *LR*.

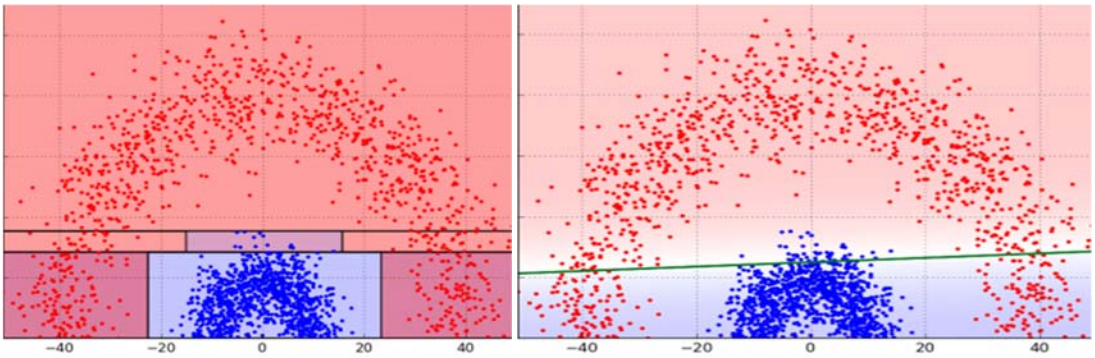
Algoritmi *LDA* i *QDA*, koji se pre svega primenjuju u klasifikacionim problemima sa više od dve klase (kategorije), imaju nešto manju tačnost od *LR,* kada se primenjuju u binarnim klasifikacionim problemima (Hastie et al., 2009). Linearne diskriminante (*LD*) *LDA* modela su linearne funkcije *X*, i predstavljaju prave kojima se razgraničavaju uzorci različitih klasa, a kojima se aproksimiraju

Bayes granične linije. U slučaju manjeg broja uzoraka, *LDA* je manje fleksibilan model, sa nižom varijansom, i postiže bolje prediktivne performanse u odnosnu na *QDA* (Hastie et al., 2009), što pokazuju i rezultati primene ovog algoritma na oba seta podataka. *QDA,* isto kao i *LDA,* pretpostavlja multivarijantnu normalnu distribuciju uzoraka različitih klasa, ali za razliku od *LDA*, kovarijantna matrica je za svaku klasu različita. Granične linije su krive, koje u nekim slučajevima mogu biti bolje aproksimacija *Bayes* graničnih linija. Kada je pretpostavka o jednakim kovarijantnim matricama neodrživa, bolje performanse moguće je postići primenom *QDA*.

6.3. Modeli na bazi stabla odlučivanja

Algoritmi na bazi stabla odlučivanja postigli su ukupno najveću tačnost na oba seta podataka. Očekivano, u odnosu na regresione algoritme, imajući u vidu da se ovim modelima ne pretpostavlja uslov linearnosti sa transformisanom zavisnom promenljivom, kao i da na njih ne utiče multikolinearnost, kao i nebalansiranost podataka. Rezultati su dalje pokazali da u slučaju većeg broja promenljivih, veće dimenzionalnosti podataka, ovi algoritmi postižu veću tačnost u odnosu na regresione (Zhang, 2016). Tako su modeli na bazi stabla na drugom setu podataka, sa 64 promenljive, postigli u proseku bolju vrednost *AUC* za 12%, dok u slučaju manjeg, prvog seta podataka, sa 24 promenljive, ovi modeli su postigli bolju vrednost *AUC* u odnosu na regresione, za približno 5%. Razlog boljih performansi je u činjenici da u slučaju kada se uzorci različitih klasa ne mogu razdvojiti linearnom graničnom linijom (ili sa hiperravni u višedimenzionalnom prostoru), algoritmi na bazi stabla bolje prepoznaju složene šablone između podataka i razgraničavaju podatke različitih klasa u više sekcija.

Slika 45. Razgraničavanje podataka različitih klasa u slučaju stabla odlučivanja i logističke regresije [26]



Iz dobijenih rezultata vidimo da *RF*, sa kojim se razvija više stabala, sa različitim grupama slučajno izabranih prediktora, postiže veću tačnost u odnosu na *DT,* na oba seta podataka. *DT* je zbog postojanja samo jednog stabla, kreiranog primenom svih raspoloživih prediktora, sklon overfitingu. Iz tog razloga, *DT* ima lošiju mogućnost generalizacije u odnosu na *RF* (Cutler, 2007).

Međutim, za razliku od *RF*, kojeg je zbog velikog broja stabala (u našem slučaju 500), praktično nemoguće interpretirati, *DT* model se jednostavno tumači, što je i njegova osnovna prednost. Empirijska analiza je dalje pokazala da je primenom *RF* postignuta znatno veća senzitivnost. U slučaju nebalansiranih podataka, primenom *DT,* broj tačno klasifikovanih pozitivnih uzoraka manjinske klase (*TP)* je izrazito nizak.

Gradient boosting algoritam (*GTB)*, takođe razvija više stabala, ali za razliku od *RF*, stabla su zavisna. Odnosno, svako novo stablo uči se na grešci u predikciji načinjenoj prethodnim. Kako ovaj model kombinuje veći broj jednostavnih *stump* stabala (eng. *weak learners*), ima visok *bias* i nisku varijabilnost, a time bolju generalizaciju od *RF* (Friedman, 2001). Ovo je razlog zašto *GTB* ima bolje performanse od *RF*. Na prvom data setu tačnost je 0.790, vrednost *AUC* 0.778, na drugom setu podataka tačnost je 0.*834* i AUC 0.*876*. Na oba data seta, *GTB* je imao minimalnu vrednost *BS*, 0.137 i 0.035. Zbog mogućnosti paralelnog procesiranja, *GTB* algoritmu je potrebno znatno manje vremena za 'učenje' od *RF*.

[26] Izvor: Logistic Regression versus Decision Trees – The Official Blog of BigML.com

Balansiranjem podataka, povećavanjem broja uzoraka manjinske klase, modelima na bazi stabla povećana je tačnost predikcije manjinske klase, naročito u slučaju *DT*, dok je vrednost *AUC,* modela *RF* i *GTB*, ostala skoro nepromenjena.

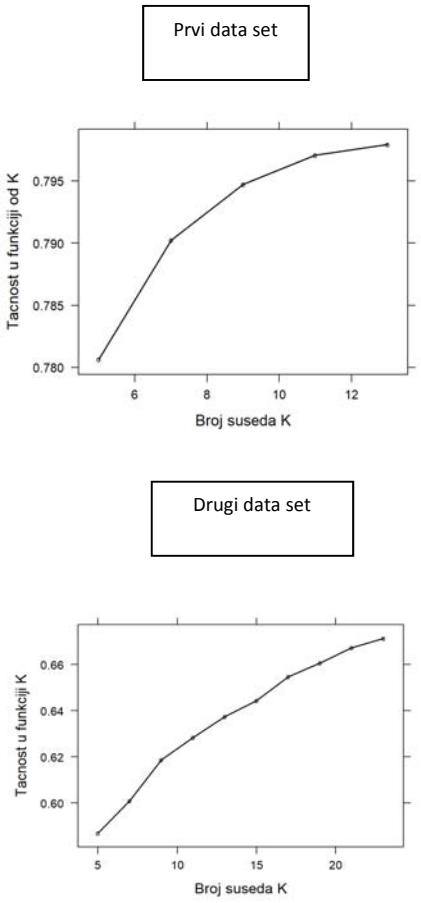
6.4. Naive Bayes

Imajući u vidu da osnovna pretpostavka ovog modela o međusobnoj nezavisnosti prediktora date klase nije ispunjena, ovim modelom su postignute iznenađujuće dobre *AUC* vrednosti 0.732 i 0.746. Zbog svoje jednostavne primene, preporučuje se u početnim fazama istraživanja, kako bi se bolje razumeli podaci i došlo do prvih predikcija, a specijalno u slučajevima kada je količina trening podataka mala i kada je većina varijabli kategorička.

6.5. K-NN

K-NN, koji spada u jednostavnije algoritme, bolju tačnost postiže sa podacima manje dimenzionalnosti i sa brojem uzoraka ne većim od 50.000 (Altman, 1992). Osetljiv je na prisustvo ekstremnih vrednosti, koje su u postupku transformacije podataka zamenjene svojim središnjim vrednostima. Model nije moguće tumačiti, već isključivo koristiti u svrhu predikcije. Za *K-NN* prvog seta podataka broj suseda *K* je 13, dok je za drugi *data* set optimalan broj suseda je 23. Tako se prema sličnosti ispitivanog uzorka sa K suseda, određuje kojoj klasi dati uzorak pripada.

Slika 46. Tačnost predikcije K-NN u funkciji broja suseda



Ovaj algoritam nije dobar izbor u slučaju velike dimenzionalnosti podataka (Hastie, 2008). Tako u prvom data setu, sa manje varijabli, *AUC* modela je zadovoljavajućih 0.738, dok u drugom *data* setu, sa znatno većim brojem varijabli, vrednost *AUC* je 0.668.

6.6. SVM

SVM modeli se najviše primenjuju za analize podataka koji su nestrukturirani. Mogu se koristiti i u slučaju klasifikacionih problema sa strukturiranim podacima, ali je potrebno imati u vidu da se

vreme učenja ovih algoritama značajno povećava sa rastom količine podataka i njihove dimenzionalnosti (Vapnik i Cortes, 1995). *SVM* sa polinominalnim *kernelom* bio je najzahtevniji kada je u pitanju potrebno vreme za učenje modela. Očekivano, *SVM* sa linearnim *kernelom* ima najmanju tačnost, kao i vrednost *AUC,* iz razloga što se uzorci različitih klasa ne mogu razgraničiti linearnom hiperravni. *SVM* sa polinominalnom *kernel* funkcijom na drugom data setu ima vrednost *AUC* 0.759 i najveća je od *SVM* algoritama, dok je na prvom data setu to *SVM Radial,* sa vrednosti *AUC* jednakoj 0.725. Pokazano je da je u slučaju nebalansiranih podataka, senzitivnost *SVM* modela dosta niska (Baruwita et al., 2013), a koja se postupkom *over-sampling*-a povećava.

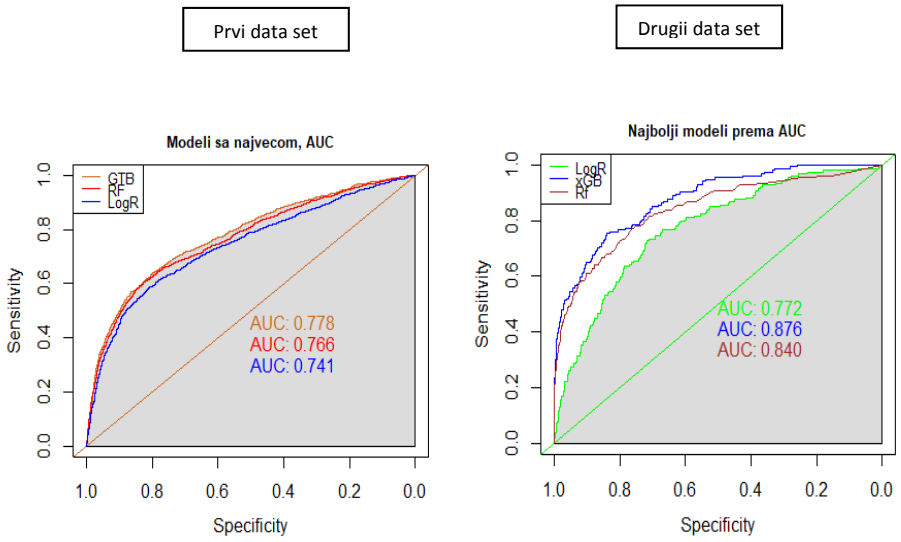
6.7. Optimalni *ML* modeli

Iako se ne može govoriti o jednom najboljem modelu za predikciju stečaja ili procenu kreditnih rizika, već je izbor optimalnog *ML* modela potrebno sagledati u funkciji veličine, uravnoteženosti i oblika distribucije podataka, prisustva ekstremnih vrednosti, broja prediktora i postojanja multikolinearnosti, empirijska analiza je pokazala je da su isti modeli postigli najbolje performanse na oba seta podataka, i to:

	AUC (prvi data set)	AUC (drugi data set)
1. GTB	0.778	0.876
2. RF	0.776	0.840
3. LR	0.741	0.772

Imajući u vidu da je većina klasičnih finansijskih modela bazirana na linearnom modelu regresije, kao i da na modele *ML* na bazi stabala ne utiče postojanje multikolinearnosti, prisutne u oba data seta, i kako su ovi modeli nezavisni od oblika distribucije podataka, dobijeni rezultati se mogu smatrati očekivanim.

Slika 47. Prikaz vrednosti AUC najboljih modela na oba *data* seta



6.8. Uporedna analiza performansi modela *ML* i Altman *Z-score*

Na kreiranom podskupu podataka, sa samo pet nezavisnih promenljivih, koje odgovaraju Altman *Z-score* modelima (sa prediktorima *Attr3, Attr6, Atte7, Attr8 i Attr9,* drugog *data* seta), proverena je tačnost predikcije *GTB* algoritama u odnosu na Altman *Z-score* model. Dobijeni rezultati potvrđuju da se primenom *ML* u ekonomiji može povećati tačnost predikcije.

Tabela 11. Performanse *Altman Z-score* i *Gradient boosting* modela

Model	Accuracy	Sensitivity	Specificity
Z score	0. 713	0. 506	0.725
GTB	0.834	0.753	0.838

7. ZAKLJUČAK

Ova disertacija predstavlja komparativnu analiza najčešće primenjivanih klasifikacionih algoritama mašinskog učenja u finansijama. U radu je, po prvi put u literaturi, dat detaljan teorijski prikaz najčešće korišćenih modela mašinskog učenja u finansijama, odgovarajućih optimizacijskih metoda, kao i sistematizacija relevantne literature. U empirijskom delu istraživanja prediktivni klasifikacioni algoritmi mašinskog učenja primenjeni su u cilju procene kreditnih rizika (eng. *credit scoring*) i verovatnoće stečaja kompanije. Rešavanjem ovih binarno-klasifikacionih problema na dva nezavisna skupa podataka pokazan je uticaj koji mašinsko učenje ima na tačnost predikcije i bolje razumevanje podataka. Dodatno je analiziran uticaj neuravnoteženih podataka (minimalno prisustvo uzoraka jedne klase) na performanse *ML* modela i kako metoda preteranog uzorkovanja manjinske klase utiče na kvalitet prediktivnog modela.

Finansijsku industriju, odlikuje intenzivna primena veštačke inteligencije, koja predstavlja jedan od osnovnih izvora konkurentske prednosti kompanija. Raspoloživost velikih količina digitalnih podataka, njihova raznovrsnost, kao i postojanje standarda za razmenu i semantiku podataka, koji važe na globalnom finansijskom tržištu, doveli su do značajne primene veštačke inteligencije u ovom sektoru.

Pored primene *ML* u predikciji stečaja i procene kreditnih rizika, rasprostranjena je primena veštačke inteligencije i u automatizaciji rutinskih procesa u finansijama, čime se povećava produktivnost, smanjuju operativni troškovi i rizici u radu.

Mogućnost obrade velikih količina podataka, omogućava razvoj specijalizovanih, individualnih usluga za klijente, koje odgovaraju njihovim navikama i potrebama. Primenom ovakvih algoritama, pored razvoja personalizovanih usluga, omogućava se i otkrivanje transakcija, koje odstupaju od uobičajenog, standardnog ponašanja klijenata, čime je moguće sprečiti finansijske zloupotrebe (eng. *fraud detection*).

Algoritamska trgovina (eng. *algorithmic trading*), predstavlja automatizaciju trgovine finansijskih proizvoda, primenom mašinsko učenje. Modeli *ML*, koji opisuju zavisnost vrednosti finansijskog proizvoda (akcije, obveznice, izvedeni finansijski instrumenti), u funkciji velikog broja atributa koji na tu vrednost utiču, autonomno donose odluke o trgovini. Na ovaj način, sprovodi se više

hiljada transakcija dnevno, bez angažovanja profesionalnog trgovca, brokera. Odluke koje algoritam *ML* donosi, nisu pristrasne i nezavise su od raspoloženja i ličnih afiniteta brokera.

ML se primenjuje i za razvoj optimalnog investicionog portfolija (eng. *portfolio management*), kojim se postiže maksimalna vrednost očekivane dobiti, za zadati nivo rizika, koji je za investitora prihvatljiv i u skladu sa njegovom averzijom prema riziku.

Potrebno je imati u vidu i rizike nekontrolisane primene *ML*. Kako se modeli mašinskog učenja, razvijaju na bazi istorijskih podataka, oni ne mogu predvideti situacije koje su karakteristične za predikciju ekstremnih tržišnih uslova i/ili događaja. Nedostatak transparentnosti primenjenih *AI* rešenja, može dovesti do narušavanja finansijske regulative. Stabilnost i likvidnost finansijskog tržišta, može biti narušena u slučaju masovne primene istog *ML* modela. Kako su za razvoj kompleksnih modela, potrebne značajne investicije, koje mogu priuštiti samo velike finansijske institucije, može doći do smanjenja konkurencije na tržištu.

Imajući u vidu rezultate istraživanja u ovoj disertaciji, kao i trendove primene modela mašinskog učenja u finansijama, u budućim istraživanjima baviću se primenom heterogenih *ML ensemble* modela i analizom uticaja ovakvih modela na tačnost predikcije. Takođe, u oblasti portfolio menadžmenta, nastaviću sa istraživanjem primene modela mašinskog učenja sa ciljem povećanja performansi investicionog portfolija prema berzanskom indeksu, kao pokazatelju opštih privrednih kretanja i obima prometa (eng. *alpha factor*).

8. LITERATURA

1. Abdi, H. i Williams, J. (2010) *Principal Component Analysis.* WIREs Cumputational Statistics, 2(4), s. 433–459.
2. Abdou, H. i Pointon, J. (2011) *Credit scoring, statistical techniques and evaluation criteria: A review of the literature*. University Salford Manchester. Wiley-Blackwel.
3. Addo, P. M. et al. (2018) *Credit Risk Analysis Using Machine and Deep Learning Models.* MDP, Switzerland.
4. Advait, J. (2020) *Naïve Bayes Classifier Advanced concept*. Technical Publications. Safari, O'Reilly Media Company.
5. Aghaie, A. et al. (2009) *Using Bayesian Networks for Bankruptcy Prediction: Empirical Evidence from Iranian Companies*. IEEE Xplore.
6. Alak, H. A. et al. (2018) *Systematic review of bankruptcy prediction models: Towards a framework for tool selection*. Expert Systems with Applications, 94, s. 164–184.
7. Alboukadel, K. (2018) *Cox proportional-hazards model***.** Statistical tools for high-throughput data analysis-STHDA.
8. Alboukadel, K. (2017) *Machine learning essentails, practical guide in R*. Statistical tools for high-throughput data analysis-STHDA.
9. Altman, E. I. (1968) *Financial Ratios, Discriminant analysis and the prediction of corporate bankruptcy*. The Journal of Finance, 23(4), s. 589–609.
10. Arminger, G. et al. (1997) *Analyzing credit risk data: A comparison of logistic discrimination,* classification tree analysis, and feed forward networks. Computational Statistics, 12 , s. 293–310.
11. Baesens, B. (2003) *Benchmarking state-of-the-art classification algorithms for credit scoring.* Journal of the Operational Research Society, 54(6), s. 627–635.
12. Baesens, B. et al. (2015) *Benchmarking state of the art classification algorithms for credit scoring. An update of the research*. Journal of the Society.
13. Baesens, B. et al. (2010) *From linear to non linear karnel based classifiers for bankruptcy prediction*. Neurocomputing. Elsevier.
14. Balcean, S. i Ooghe, H. (2005) *35 Years of Studies on Business Failure: An Overview of the Classic Statistical Methodologies and Their Related Problems*. The British Accounting Review, 38(1), s. 63–93.
15. Barboza, F. i Kimura, H. (2017) *Machine learning models and bankruptcy prediction*. Expert System with Applications, 83, s. 405–417.

16. Barnard, J. i Meng, X. L. (1999) *Applications of multiple imputation in medical studies: From AIDS to NHANES.* Statistical Methods in Medical Research, 8(1).
17. Batuwita, R. i Palade, V. (2013) *Class Imbalance Learning Methods for Support Vector Machines*. MIT Alliance for Research and Technology Centre. University of Oxford.
18. Berger, J. O. (2013) *Statistical decision theory and Bayesian analysis*. Springer Series in Statistics. Drugo izdanje.
19. Bishop, C. M. (2006) *Pattern Recognition and Machine Learning*. Springer Link.
20. Breiman, L. (1994) *Bagging Predictors.* Department of Statistics, University of California. Tekst dostupan na: https://www.stat.berkeley.edu/~breiman/bagging.pdf. [pristupljeno: 18. decembra 2022].
21. Breeden, J. L. (2020) *A Survey of Machine Learning in Credit Risk.* Tekst dostupan na*: https://www.researchgate.net/publication/341804274_A_Survey_of_Machine_Learning_in_Credit_Risk* [pristupljeno: 21. januara 2023].
22. Brown, I. i Mues, C. (2012) *An experimental comparison of classification algorithms for imbalances credit scoring data sets*. Expert Systems with Applications. 39(3), s. 3446–3453. Tekst dostupan na: https://core.ac.uk/reader/82354140 [pristupljeno: 11. oktobra 2021].
23. Butte,G. (2000) *Butter's Law of Photonics*. Tekst dostupan na: https://en.wikipedia.org/wiki/Moore%27s_law#:~:text=Butters%27%20law%20says%20that%20the,by%20half%20every%20nine%20months. [pristupljeno: 3. juna 2022].
24. Chawla, N. V. et al. (2002) *Smote: Synthetic minority over-sampling technique*. Journal of Artificial Intelligence Research, 16, s. 321–357.
25. Chen, M. (2011) *Bankruptcy prediction in firms with statistical and intelligent techniques and a comparison of evolutionary computation approaches.* Computers and Mathematics with Applications, 62(12), s. 4514–4524**.** Tekst dostupan na: https://www.sciencedirect.com/science/article/pii/S0898122111008947?via%3Dihub [pristupljeno: 21. februara 2022].
26. Chih, F. T. et al. (2015) *A comparative study of classifier ensembles for bankruptcy prediction.* Applied Soft Computing, 24, s. 977–984.
27. Cox, D. R. (1972) *Regression models and life tables*. Journal of the Royal Statistical Society, 34, s. 187–220.
28. Crook, J. N. et al. (2007). *Recent developments in consumers credit risk assessment*. European Journal of Operational Research, 183(3), s. 1447–1465.
29. Crook, J. N. (1996) *Credit scoring: An overview*. European Journal of Operational Research, 95, s. 24–37.

30. Cutler, R. et al. (2007) *Random Forests for Classification and Regression.* Ecological Society of America, 88(11), s. 2783–2792.
31. DeLong, E. R. i DeLong, D. M. (1998) *Comparing the area under two or more correlated ROC curves: A nonparametric Approach.* Biometrics, 44, s. 837–845.
32. Desai, V. S. et al. (1996) *A Comparison of Neural Networks and Linear Scoring Models in the Credit Union Environment*. European Journal of Operational Research 95(1), s. 24–37.
33. Draelos, R. (2020) *Comparing AUCs of Machine Learning Models with DeLong's Test*. Machine Learning and medicine.
34. Fawcett, T. (2006) *An introduction to ROC analysis*. Pattern Recognition Letters, 27, s. 861–874.
35. Friedman, J. (2001) *A Gradient Boosting Machine.* The Annals of Statistics, 29(5), s. 1189–1232.
36. Friedman, J. (2001). *Greedy function approximation: A gradient boosting machine*. The Annals of Statistics, 29(5), s. 1189–1232.
37. Gaber, T. et al. (2017) *Liner discriminant Analysis: A detailed tutorial*. University of Salford. Tekst dostupan na: https://usir.salford.ac.uk/id/eprint/52074/1/AI_Com_LDA_Tarek.pdf [pristupljeno: 3. juna 2022].
38. Gantz, J. i Reinsel, D. (2012) *The Digital Universe in 2020: Big data, Bigger Digital Shadows, and Biggest Growth in the Far East*. IDC.
41. Gareth, J. (2017) *An introduction to statistical learning, with application in R.* Springer texts in Statistics. NY.
42. Graham, M. (2003) *Confronting multicolinearity in ecological multiple regression*. Ecology, 84(11), s. 2809–2815.
43. Hand, D. J. i Henley, W. E. (1997) *Statistical Classification Methods in Consumer Credit Scoring: A Review*. Journal of the Royal Statistical Society: Series A (Statistics in Society), 160, s. 523–541.
44. Hansen, L. i Salamon, P. (1990) *Neural network ensembles.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 12, s. 993–1001.
45. Harrell, F. (2019) *Road map for choosing between statistical modeling and machine learning.* Tekst dostupan na: https://www.fharrell.com/post/stat-ml [pristupljeno: 3. maja 2022].
46. Hastie, T. et al. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Drugo izdanje. Springer Science & Business Media, Luxemburg.
47. Henley, W. i Hand, D. (1997) *Construction of a K-nearest-neighbour credit-scoring system*. IMA Journal of Management Mathematics, 8(4), s. 305–321.
48. Hoang, D. i Wiegratz, K. (2021) *Machine Learning Methods in Finance: Recent Applications and Prospects*. Tekst dostupan na: https://finance.fbv.kit.edu/rd_download/Machine%20Learning%20Methods%20in%20Finance.pdf [pristupljeno: 20. juna 2022].

49. Huang, A. H. et al. (2014) *Evidence on the Information Content of Text in Analyst Reports*. The Accounting Review, 89, s. 2151–2180.
50. Huang, C. et al. (2007) *Credit scoring with a data mining approach based on support vector machines*. Expert Systems with Applications, 33 (4), s. 847–856.
51. Huaxiang, Z. i Minfang, L. (2014.) *RWO Sampling: A random walk over sampling approach to imbalanced data classification.* Information Fusion, 20, s. 99–116.
52. Hyeongjun, K. et al. (2020) *Corporate default predictions: Literature review.* MDPI, Basel, Switzerland. Tekst dostupan na: https://www.mdpi.com/2071-1050/12/16/6325/xml [pristupljeno: 15. marta 2022].
53. Jaggi, M. (2013) *An Equivalence between the Lasso and Support Vector Machines.* ETH Zurich, Switzerland.
54. Jardn, P. (2016) *A two-stage classification technique for bankruptcy prediction.* European Journal of Operational Research, 254(1), s. 236–252.
55. Jolliffe, I. T. (2002) *Principal Component Analysis*. Drugo izdanje. Springer–Verlag, New York.
56. Kaufman, I. i Horton, C. (2015) *Digital Transformation: Leveraging Digital Technology with Core Values to Achieve Sustainable Business Goals*. The European Financial Review (December–January), s. 63–67.
57. Kelly, B. et al. (2020) *Empirical Asset Pricing via Machine Learning.* The Review of Financial Studies (SFS), 33(5).
58. Kelly, B. et al. (2020) *The Virtue of Complexity in return prediction*. Swiss Finance Institute Research Paper, s. 21–90.
59. Kleinert, M. K. (2014) *Comparison of accounting-based bankruptcy prediction models of Altman (1968), Ohlson (1980), and Zmijewski (1984) to German and Belgian listed companies during 2008 – 2013.* University of Twente, Twente.
60. Kumar, R. i Ravi, V. (2007) *Bankruptcy Predictions in banks and firms via statistical and intelligent techniques: A review*. European Journal of Operational Research, 180(1), s. 1–28.
61. Kryder, M. (2005). *Kryder's Law*. Tekst dostupan na: https://www.techtarget.com/searchstorage/definition/KrydersLaw#:~:text=Kryder%27s%20Law%20is%20the%20assumption,improves%2C%20storage%20will%20become%20cheaper. [pristupljeno: 3. juna 2022].
62. Lanzolla, G. i Anderson, J. (2008) *Digital Transformation*. Business Strategy Review, 19(2), s.72–76.
63. Lendasse, A. i Yu, Q. (2009) *Ensemble KNNs for Bankruptcy Prediction*. Researchgate. Tekst dostupan na:

https://www.researchgate.net/publication/255669581_Ensemble_KNNs_for_Bankruptcy_Prediction [pristupljeno: 4. aprila 2022].

64. L´eon Bottou (2010) *Large-Scale Machine Learning with Stochastic Gradient Descent*. Princeton, NJ, NEC Labs America.
65. Little, R. i Rubin, D. (1989). *The analysis of social science data with missing values*. Sociological Method and Research, SAGE Journals, 18(2–3).
66. Lombardo, G. et al. (2022) *Machine Learning for Bankruptcy Prediction in the American Stock Market: Dataset and Benchmarks***.** MDPI, Basel, Switzerland.
67. McCullagh, P. i Nelder, J. A. (1983) *Generalized Liner Models*. Monographs on Statistics and Applied Probability. Chapman and Hall, London.
68. Min, J. H. i Lee, YC. (2005) *Bakruptcy Prediction using support vector machinewith optimal choice of kernel function parameters*. Expert Systems with Applications, 28, s. 603–614. Tekst dostupan na: https://www.researchgate.net/publication/222580945_Bankruptcy_prediction_using_support_vector_machine_with_optimal_choice_of_kernel_function_parameters [pristupljeno: 3. aprila 2021].
69. Mitchell, T. (1997) *Machine Learning*, McGraw Hill, New York.
72. Moore, G. (1975) *Moore's Law* .Tekst dostupan na https://en.wikipedia.org/wiki/Moore%27s_law [pristupljeno: 3. juna 2022].
73. Mullainathan, S. i Spiess, J. (2017) *Machine Learning: An Applied Econometric Approach*. Journal of Economic Perspectives, 31, s. 87–106.
74. Narvekar, A. i Guha, D. (2021) *Bankruptcy prediction using machine learning and an application to the case of the COVID-19 recession*. Data Science in Finance and Economics, 1(2), s. 180–195.
75. Ohlson, J. (1980) *Financials ratios and probablaistic prdiction of Bankruptcy.* Journal of Accounting research, 18(1). Tekst dostupan na: https://math.ryerson.ca/ramlab/projects/crd/ohlson1980.pdf [pristupljeno: 23. juna 2022].
76. Olson, D. L. et al. (2012) *Comparative analysis of data mining methods for bankruptcy prediction.* Decision Support Systems, 52, s. 464–473. Tekst dostupan na: https://www.researchgate.net/publication/220196365_Comparative_analysis_of_data_mining_methods_for_bankruptcy_prediction [pristupljeno: 23. juna 2021].
77. Patel, P. et al. (2019) *Bankruptcy Prediction Model Using Naïve Bayes Algorithms*. International Journal of innovative trends in engineering (IJITE), 83(59).
78. Pawitan, Y. (2001) *In all likelihood: statistical modelling and inference using likelihood*. Oxford University Press.

79. Piri, S. et al. (2018) *A synthetic informative minority over sampling (SIMO) algorithm levaraging support vector machine to enhance learning from impalances datasets.* Decission Support Systems, 106, s. 15–29.
80. Porter, D. i Gujarati, D. (2009). *Basic Econometrics.* NY,The McGraw-Hill Series Economics.
81. Provost, F. i Fawcett, T. (2013) *Data Science for Business*. New York, O'Reilly Media Inc.
82. Putri, N. H. et al. (2021) *Credit risk analysis using SVM*. Journal of Physics. Tekst dostupan na: https://www.researchgate.net/publication/350337379_Credit_risk_analysis_using_support_vector_machines_algorithm[pristupljeno: [pristupljeno: 10. decembra 2022].
83. Roger, D. (2016) *The Digital transformation playbook.* Columbia Business School, USA.
84. Rosenberg, E. i Gleit, A. (1994) *Quantittative methods in Credit Management: Survey. Operations Research*, 42(4), s. 589–613. Tekst dostupan na: https://www.researchgate.net/publication/242932275_Quantitative_Methods_in_Credit_Management_A_Survey [pristupljeno: 2. novembra 2022].
85. Rundo, F. et al. (2019) *Machine Learning for Quantitative Finance Applications: A Survey*. Tekst dostupan na: https://www.mdpi.com/2076-3417/9/24/5574 [pristupljeno 3. aprila 2022].
86. Sarker, I. H. (2021) *Machine Learning: Algorithms, Real-World Applications and Research Directions.* SN Computer Science, 160(2). Tekst dostupan na: https://link.springer.com/article/10.1007/s42979-021-00592-x [pristupljeno: 11. septembra 2022].
87. Samuel, A. (1959) *Some studies in Machine learning using the game of checkers*. IBM Journal, 3(3). Tekst dostupan na: http://people.csail.mit.edu/brooks/idocs/Samuel.pdf [pristupljeno: 3. avgusta 2022].
88. Schafer, J. L. i Graham, J. W. (2002) *Missing data: Our view of the state of the art*. Psychological Methods, 7(2), s. 147–177. US, National Library of Medicine.
89. Shwartz, S. i Shai Ben David (2014) *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press.
90. Singh, G. (2019). *Comparison between Multinomial and Bernoulli for Naive Bayes for Text Classification.* London, IEEE. Tekst dostupan na: https://www.researchgate.net/publication/334765873_Comparison_between_Multinomial_and_Bernoulli_Naive_Bayes_for_Text_Classification [pristupljeno 19. jula 2022].
91. Stelzer, A. (2019) *Predicting credit default probabilities using ML techniques in the face of unequal class distribution*. Vienna University of Economics and Business. Tekst dostupan na: https://www.researchgate.net/publication/334785564_Predicting_credit_default_probabilities_using_machine_learning_techniques_in_the_face_of_unequal_class_distributions [pristupljeno 7. februara 2022].
92. Steenackers, A. i Goovaerts, M. J. (1989) *A credit scoring model for personal loans***.** Insurance: Mathematics and Economics, 8(1), s. 31–34.

93. Sugiyama, M. (2015) *Introduction to Statistical Machine Learning*. Prvo izdanje. USA, MK.
94. Tobback, E. et al. (2017) *Bankruptcy prediction for SMEs using relational data*. Decission Support Systems, 102, s. 69–81. USA. Tekst dostupan na: https://www.researchgate.net/publication/298205240_Bankruptcy_prediction_for_SMEs_using_relational_data [pristupljeno: 25. maja 2022].
95. Tzong, H. L. (2009) *A cross model study of corporate financial distress prediction in Taiwan: Multiple discriminant analysis, logit, probit and neural networks models*. Neurocomputing, 72 (1618), s. 3507–3516.
96. Vapnik, V. i Cortes, C. (1995) *Support vector networks.* Machine learning, 20, s. 273–297. Tekst dostupan na: http://image.diku.dk/imagecanon/material/cortes_vapnik95.pdf [pristupljeno 13. juna 2022].
97. Vapnik, V. (1995) *The Nature of Statistical Learning Theory*. Drugo izdanje. Statistics for Engineering and Information Science. Springer, NY.
98. Varian, H. R. (2014) *Big Data: New Tricks for Econometrics*. Journal of Economic Perspectives, 28, s. 3–28.
99. Zhang, H. i Li, M. R. (2014) *RWO-Sampling: A random walk over-sampling approach to imbalanced data classification*. Information Fusion, 20, s. 99–116.
100. Zhou, L. (2012) *Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods*. Knowladge based systems. 41. s. 16–25.
101. Zmijewski, M. E. (1984) *Methodological issues related to the estimation of financial distress prediction models.* Journal of Accounting research, 22, s. 59–82.
102. Weiss, G. M. i Provost, F. (2003) *Learning when training data are costly: The effect of class distribution on tree induction.* Journal of Artificial Intelligence Research, 19, s. 315–354.
103. Winston, P. (2010) *Support Vector Machine*. Massachusetts Institute of Technology. Tekst dostupan na: https://ocw.mit.edu/courses/6-034-artificial-intelligence-fall-2010/ [pristupljeno: 3. juna 2022].
104. Winston, P. (2010) *Boosting*. Massachusetts Institute of Technology. Tekst dostupan na: http://ocw.mit.edu/6-034F10 [pristupljeno: 3. jula 2022].
105. Yang, Y. (2007) *Adaptive credit scoring with kernel learning methods*. European Journal of Operational Research, 183 (3), s. 1521–1536.
106. Yeh, I. i Lien, C. (2009), *The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients*. Expert Systems with Applications 36(2), 2473–2480.

Prilog A.1. Struktura podataka

Tabela 12. *Polish companies bankruptcy Data Set* (prvi data set)

Prediktori	Značenje
Attr1	net profit / total assets
Attr2	total liabilities / total assets
Attr3	working capital / total assets
Attr4	current assets / short-term liabilities
Attr5	[(cash + short-term securities + receivables - short-term liabilities) / (operating expenses - depreciation)] * 365
Attr6	retained earnings / total assets
Attr7	EBIT / total assets
Attr8	book value of equity / total liabilities
Attr9	sales / total assets
Attr10	equity / total assets
Attr11	(gross profit + extraordinary items + financial expenses) / total assets
Attr12	gross profit / short-term liabilities
Attr 13	(gross profit + depreciation) / sales
Attr 14	(gross profit + interest) / total assets
Attr15	(total liabilities * 365) / (gross profit + depreciation)
Attr16	(gross profit + depreciation) / total liabilities
Attr17	total assets / total liabilities
Attr18	gross profit / total assets
Attr19	gross profit / sales
Attr20	(inventory * 365) / sales
Attr21	sales (n) / sales (n-1)
Attr22	profit on operating activities / total assets
Attr23	net profit / sales
Attr24	gross profit (in 3 years) / total assets
Attr25	(equity - share capital) / total assets
Attr26	(net profit + depreciation) / total liabilities
Attr27	profit on operating activities / financial expenses
Attr28	working capital / fixed assets
Attr29	logarithm of total assets
Attr30	(total liabilities - cash) / sales
Attr31	(gross profit + interest) / sales
Attr32	(current liabilities * 365) / cost of products sold
Attr33	operating expenses / short-term liabilities
Attr34	operating expenses / total liabilities
Attr35	profit on sales / total assets
Attr36	total sales / total assets
Attr37	(current assets - inventories) / long-term liabilities
Attr38	constant capital / total assets
Attr39	profit on sales / sales
Attr40	(current assets - inventory - receivables) / short-term liabilities
Attr41	total liabilities / ((profit on operating activities + depreciation) * (12/365))
Attr42	profit on operating activities / sales

Attr43	rotation receivables + inventory turnover in days
Attr44	(receivables * 365) / sales
Attr45	net profit / inventory
Attr46	(current assets - inventory) / short-term liabilities
Attr47	(inventory * 365) / cost of products sold
Attr48	EBITDA (profit on operating activities - depreciation) / total assets
Attr49	EBITDA (profit on operating activities - depreciation) / sales
Attr50	current assets / total liabilities
Attr51	short-term liabilities / total assets
Attr52	(short-term liabilities * 365) / cost of products sold)
Attr53	equity / fixed assets
Attr54	constant capital / fixed assets
Attr55	working capital
Attr56	(sales - cost of products sold) / sales
Attr57	(current assets - inventory - short-term liabilities) / (sales - gross profit - depreciation)
Attr58	total costs /total sales
Attr59	long-term liabilities / equity
Attr60	sales / inventory
Attr61	sales / receivables
Attr62	(short-term liabilities *365) / sales
Attr63	sales / short-term liabilities
Attr64	sales / fixed assets
Target promenjiva	**Značenje**
Class	0 kompanija nije u stečaju, 1 kompanija u stečaju

Tabela 13. *Default of credit card clients Data* (drugi data set)

Prediktori	**Značenje**
LIMIT_BAL	Amount of the given credit
SEX	Gender (1 = male; 2 = female).
EDUCATION	Education (1 = graduate school; 2 = university; 3 = high school; 4 = others)
MARRIAGE	Marital status (1 = married; 2 = single; 3 = others).
AGE	Age
PAY_0	History of past payment, Sept 2005*
PAY_2	History of past payment, Avgust 2005*
PAY_3	History of past payment, July 2005*
PAY_4	History of past payment, June 2005*
PAY_5	History of past payment, May 2005*
PAY_6	History of past payment, April 2005*
BILL_AMT1	Amount of bill statement, Sept 2005
BILL_AMT2	Amount of bill statement, Avgust 2005
BILL_AMT3	Amount of bill statement, July 2005
BILL_AMT4	Amount of bill statement, June 2005)
BILL_AMT5	Amount of bill statement, May 2005
BILL_AMT6	Amount of bill statement, April 2005
PAY_AMT1	Amount paind in Sept 2005

PAY_AMT2	Amount paind in Avgust 2005
PAY_AMT3	Amount paind in July 2005
PAY_AMT4	Amount paind in June 2005
PAY_AMT5	Amount paind in May 2005
PAY_AMT6	Amount paind in April 2005
Target promenjiva	**Značenje**
default.payment.next.month	default payment (Yes = 1, No = 0)

*) -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above

Prilog A.2. Rezultat empirijske analize

Tabela 14. Rezultati empirijske evaluacije *ML* modela na '*Default of credit card clients data'* nebalansiranim podacima

Struktura podataka	Tip modela	Model	Tačnost	Senzitivnost	Specifičnost	F1	Kappa	BS	AUC
Nebalansirana	Regresioni	**Logistic regression (LR)**	**0.759**	**0.585**	**0.811**	**0.526**	**0.367**	**0.147**	**0.741**
		Logistic regression sa PC[27] (LRpc)	0.755	0.594	0.802	0.525	0.363	0.147	0.741
		Lasso LR	0.758	0.589	0.808	0.526	0.366	0.147	0.7395
		Ridge LR	0.759	0.586	0.810	0.526	0.367	0.148	0.7399
		Elastic Net	0.760	0.584	0.812	0.527	0.368	0.147	0.7394
		LDA	0.746	0.607	0.787	0.522	0.354	0.147	0.740
		QDA	0.703	0.680	0.709	0.511	0.316	0.342	0.735
	Na bazi stabla	Decision tree	0.802	0.479	0.898	0.525	0.401	0.143	0.696
		RF	**0.780**	**0.593**	**0.835**	**0.551**	**0.407**	**0.140**	**0.765**
		GTB	**0.790**	**0.593**	**0.849**	**0.563**	**0.425**	**0.137**	**0.778**
	Naïve Bayes	Naïve Bayes	0.758	0.590	0.808	0.527	0.367	0.222	0.732
	K NN	K NN	0.740	0.587	0.785	0.507	0.334	0.148	0.738
	Support Vector machine	SVM Lin	0.765	0.545	0.830	0.514	0.360	0.153	0.704
		SVM Radial	0.748	0.563	0.802	0.504	0.337	0.145	0.725
		SVM Poli	0.773	0.530	0.844	0.515	0.367	0.148	0.707

RF(package h2o), AUC 0.766

GTB(package h2o), AUC 0.778

GLM(package h2o), AUC 0.724

Heterogeni ensembling (rf+glm+gbm) , AUC 0.745

Tabela 15. Rezultati empirijske evaluacije *ML* modela, na '*Default of credit card clients Data'* balansiranim podacima – SMOTE, modela na bazi stabla

Struktura podataka	Tip modela	Model	Tačnost	Senzitivnost	Specifičnost	F1	Kappa	BS	AUC
Balansirana	Na bazi stabla	Decision tree	0.775	0.541	0.843	0.523	0.375	0.206	0.692
		RF	0.750	0.625	0.787	0.532	0.367	0.153	0.763
		GTB	0.768	0.586	0.822	0.535	0.382	0.147	0.760

[27] Model sa *PC* kojim sa objasnjava 95% kumulativne varijabilnosti prediktora

Tabela 16. Rezultati empirijske evaluacije *ML* modela na '*Polish companies bankruptcy Data Set*' nebalansiranim podacima

Struktura podataka	Tip modela	Model	Tačnost	Senzitivnost	Specifičnost	F1	Kappa	BS	AUC
Nebalansirana	Regresioni	**Logistic regression (LR)**	**0.719**	**0.727**	**0.718**	**0.213**	**0.136**	**0.047**	**0.772**
		Logistic regression sa PC[28] (LRpc)	0.654	0.805	0.645	0.196	0.114	0.047	0.763
		Lasso LR	0.714	0.734	0.713	0.212	0.135	0.047	0.764
		Ridge LR	0.715	0.701	0.716	0.205	0.127	0.047	0.763
		Elastic Net	0.710	0.734	0.709	0.210	0.132	0.047	0.763
		LDA	0.681	0.721	0.679	0.192	0.111	0.048	0.742
		QDA	0.750	0.617	0.758	0.206	0.130	0.266	0.740
	Na bazi stabla	Decision tree	0.840	0.377	0.866	0.198	0.131	0.045	0.630
		RF	**0.754**	**0.779**	**0.753**	**0.249**	**0.177**	**0.039**	**0.840**
		GTB	**0.834**	**0.753**	**0.838**	**0.322**	**0.261**	**0.035**	**0.876**
	Naïve Bayes	Naïve Bayes	0.693	0.721	0.691	0.198	0.118	0.224	0.746
	K NN	K NN	0.759	0.532	0.771	0.188	0.111	0.048	0.668
	Support Vector machine	SVM Lin	0.596	0.610	0.595	0.137	0.048	0.049	0.606
		SVM Radial	0.790	0.610	0.800	0.234	0.163	0.047	0.751
		SVM Poli	0.806	0.578	0.818	0.238	0.168	0.047	0.759

RF(package h2o), AUC 0.821

GTB(package h2o), AUC 0.866

GLM(package h2o), AUC 0.770

Heterogeni ensembling (rf+glm+gbm) , AUC 0.792

Tabela 17. Rezultati empirijske evaluacije *ML* modela, na *'Polish companies bankruptcy Data Set*' balansiranim podacima – SMOTE, modela na bazi stabla

Struktura podataka	Tip modela	Model	Tačnost	Senzitivnost	Specifičnost	F1	Kappa	BS	AUC
Balansirana	Na bazi stabla	Decision tree	0.758	0.675	0.762	0.226	0.152	0.147	0.740
		RF	0.699	0.864	0.690	0.231	0.155	0.06	0.842
		GTB	0.784	0.766	0.785	0.271	0.202	0.049	0.846

[28] Model sa *PC* kojim sa objasnjava 95% kumulativne varijabilnosti prediktora

Prilog A.3. Delimičan prikaz razvijenog softverskog algoritma

cred_card_fin.r (Predikcija stečaja kompanija, Polish companies bankruptcy)

(kod dostupan na kkolaro/doktorat (github.com))

 # PREDIKTIVNI MODELI

 # LOGISTICKA REGRESIJA

Kreiranje modela

model_lr<-glm(default.payment.next.month ~ ., data=train_norm, family = binomial)

Log reg sa PC

model_lrpc<-glm(default.payment.next.month ~ ., data=train_pc, family = binomial)#model sa PC

PENALIZED LOGISTIC REGRESSION, regularizacija modela

#Lasso regresija

Kreiranje modela

model_lrlasso<- glmnet(x_train,y_output, alpha = 1, family = "binomial",lambda = lasso_param$lambda.min)

 # Ridge regresija

Kreiranje modela

model_ridge<- glmnet(x_train,y_output, alpha = 0, family = "binomial",lambda = ridge_param$lambda.min)

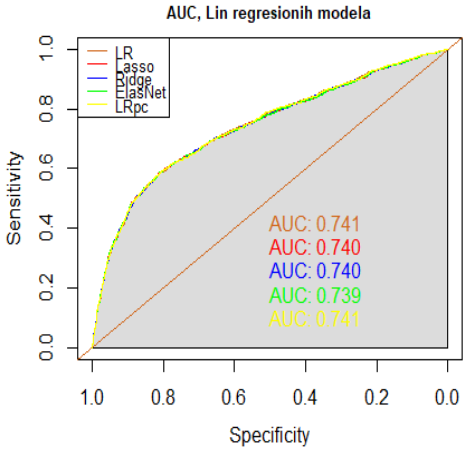
 # Elastic net regresija

Kreiranje modela

model_elasnet<- glmnet(x_train,y_output, alpha = elas$bestTune$alpha, family = "binomial",lambda = elasticnet_param$lambda.min)

Pokazatelji performansi regresionih modela

	Model	Tacnost	Sensitivity	Specificity	F1	Kappa	AUC	BS
1	LinReg	0.759	0.585	0.811	0.526	0.367	0.7413	0.147
2	LinRegPc	0.755	0.594	0.802	0.525	0.363	0.741	0.147
3	Lasso	0.758	0.589	0.808	0.526	0.366	0.7395	0.147
4	Ridge	0.759	0.586	0.81	0.526	0.367	0.7399	0.148
5	ElasNet	0.76	0.584	0.812	0.527	0.368	0.7394	0.147



 # LDA

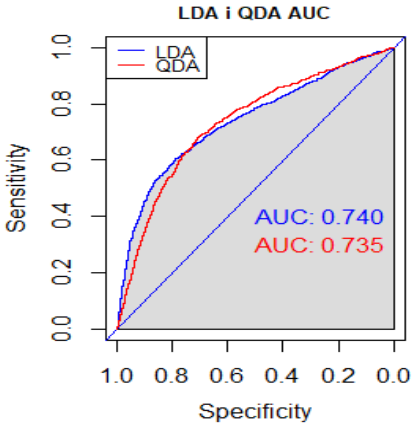
model_lda<-lda(default.payment.next.month ~ ., data=train_norm)

 # QDA

model_qda<-qda(default.payment.next.month ~ .,data=train_norm)

Pokazatelji performansi LDA i QDA modela

	Model	Tacnost	Sensitivity	Specificity	F1	Kappa	AUC	BS
1	LDA	0.746	0.607	0.787	0.522	0.354	0.74	0.147
2	QDA	0.703	0.68	0.709	0.511	0.316	0.735	0.342



 # STABLA ODLUCIVANJA

trControl=trainControl("cv",number=5) # sampling = "smote", Balansiranje trening podataka ML modela stabla sa sampaling opcijama smote

Kreiranje modela

model_tree<-train(default.payment.next.month~., data=train,method="rpart",trControl= trControl, tuneLength=10)# tuneLength daje moguce cp vrednosti

 # RANDOM FOREST

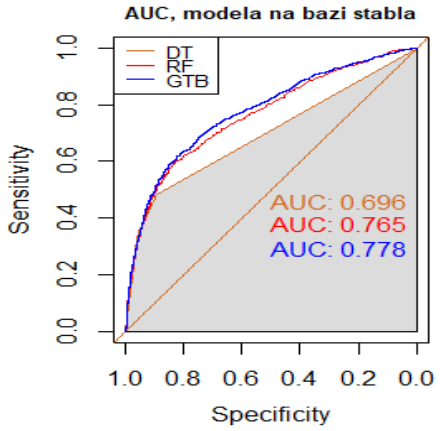
model_rf<-train(default.payment.next.month~., data=train,method="rf",trControl= trControl,tuneLength=5)# smanjana tunelength na 5, zbog performansi

BOOSTING MODEL, xgboost sa xgbTree. Koristi sve corove procesora, paralelno procesiranje

model_xgb<-train(default.payment.next.month~., data=train,method="xgbTree",trControl= trControl)

Pokazatelji performansi modela na bazi stabla

	Model	Tacnost	Sensitivity	Specificity	F1	Kappa	AUC	BS
1	DT	0.802	0.479	0.898	0.525	0.401	0.696	0.143
2	RF	0.78	0.593	0.835	0.551	0.407	0.765	0.14
3	GTB	0.79	0.593	0.849	0.563	0.425	0.778	0.137



Pokazatelji performansi modela na bazi stabla, balansirani podaci, SMOTE

	Model	Tacnost	Sensitivity	Specificity	F1	Kappa	AUC	BS
1	DT	0.775	0.541	0.843	0.523	0.375	0.692	0.206
2	RF	0.75	0.625	0.787	0.532	0.367	0.763	0.153
3	GTB	0.768	0.586	0.822	0.535	0.382	0.76	0.147

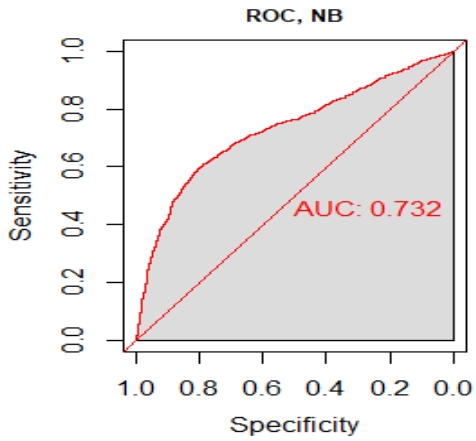
 # NB

Kreiranje modela

model_nb<-naiveBayes(train$default.payment.next.month~., data=train)

Pokazatelji performansi NB modela

	Model	Tacnost	Sensitivity	Specificity	F1	Kappa	AUC	BS
Accuracy	NB	0.758	0.59	0.808	0.527	0.367	0.732	0.222



 # KNN model

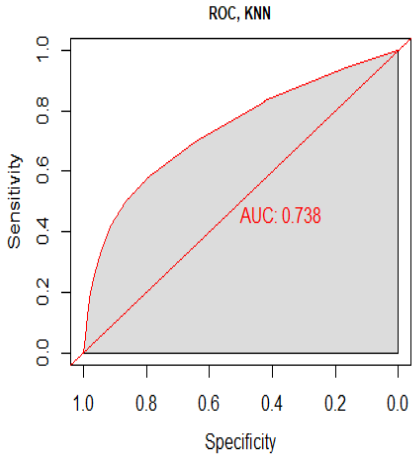
Kreiranje modela

model_knn<-train(Stecaj~., data = train, method='knn',tuneLength=10,

trControl=trControl,metric="ROC")

Pokazatelji performansi K NN modela

	Model	Tacnost	Sensitivity	Specificity	F1	Kappa	AUC	BS
Accuracy	KN	0.74	0.587	0.785	0.507	0.334	0.738	0.148



 # SUPORT VECTOR MACHINE

 # Linear SVM

model_svmlin <- train(default.payment.next.month ~ ., data = train, method = "svmLinear", trControl=trControl)

 # SVM Radial. Kreiranje modela, SVM sa nelinearnom kernel funkcijom (Radial)

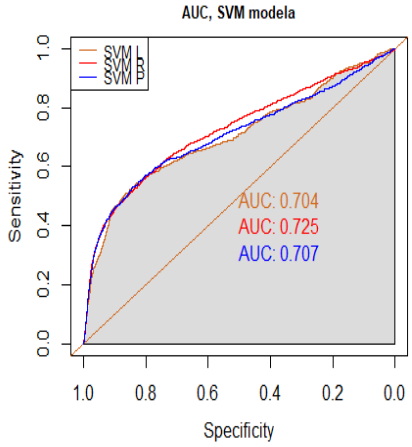
model_svmradial <- train(default.payment.next.month ~., data = train, method = "svmRadial",trControl=trControl)

 # SVM Poli

model_svmpoli <- train(default.payment.next.month ~ ., data = train, method = "svmPoly",trControl=trControl)

Pokazatelji performansi SVM modela

	Model	Tacnost	Sensitivity	Specificity	F1	Kappa	AUC	BS
1	SVM L	0.765	0.545	0.83	0.514	0.36	0.704	0.153
2	SVM R	0.748	0.563	0.802	0.504	0.337	0.725	0.145
3	SVM P	0.773	0.53	0.844	0.515	0.367	0.707	0.148



 # Heterogeni modeli

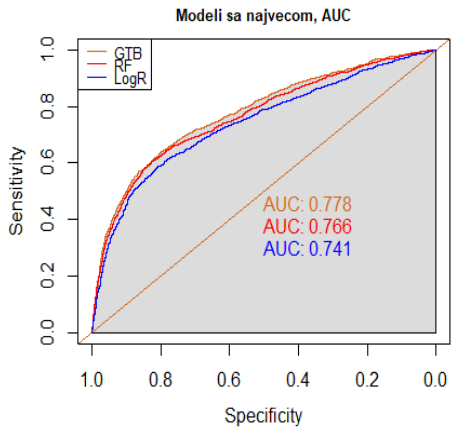
```
# GBM
prvi_gbm <- h2o.gbm(x = x,
                    y = y,
                    training_frame = train_df_h2o,
                    nfolds = 5,
                    keep_cross_validation_predictions = TRUE,
                    seed = 5)
# GLM
drugi_glm <- h2o.glm(x = x,
                     y = y,
                     training_frame = train_df_h2o,
                     nfolds = 5,
                     keep_cross_validation_predictions = TRUE,
                     seed = 5)
# RF
treci_rf <- h2o.randomForest(x = x,
                         y = y,
```

```
                    training_frame = train_df_h2o,
                    nfolds = 5,
                    keep_cross_validation_predictions = TRUE,
                    seed = 5)
# Objedinjavanje pojedinacnih predikcija, stacking
ensemble <- h2o.stackedEnsemble(x = x,
                      y = y,
                      metalearner_algorithm="drf",
                      training_frame = train_df_h2o,
                      base_models = list(prvi_gbm, drugi_glm,treci_rf))
```

Pokazatelji performansi modela, primenom H2o

	Model	Tacnost	Sensitivity	Specificity	F1	Kappa	AUC
1	RF_h2o	0.783	0.593	0.839	0.554	0.411	0.766
2	GLM_h2o	0.771	0.553	0.835	0.524	0.373	0.724
3	GBM_h2o	0.763	0.639	0.8	0.552	0.395	0.778
4	Ensemb	0.758	0.568	0.815	0.815	0.358	0.745

Modeli sa najvecim AUC



Kraj

Ml_2years.r (*Default prediction of credit card clients in two years period****)***

(kod dostupan na kkolaro/doktorat (github.com))

 # PREDIKTIVNI MODELI

 # LOGISTICKA REGRESIJA

Kreiranje modela

model_lr<-glm(default.payment.next.month ~ ., data=train_norm, family = binomial)

Log reg sa PC

model_lrpc<-glm(default.payment.next.month ~ ., data=train_pc, family = binomial)#model sa PC

PENALIZED LOGISTIC REGRESSION, regularizacija modela

#Lasso regresija

Kreiranje modela

model_lrlasso<- glmnet(x_train,y_output, alpha = 1, family = "binomial",lambda = lasso_param$lambda.min)

 # Ridge regresija

Kreiranje modela

model_ridge<- glmnet(x_train,y_output, alpha = 0, family = "binomial",lambda = ridge_param$lambda.min)

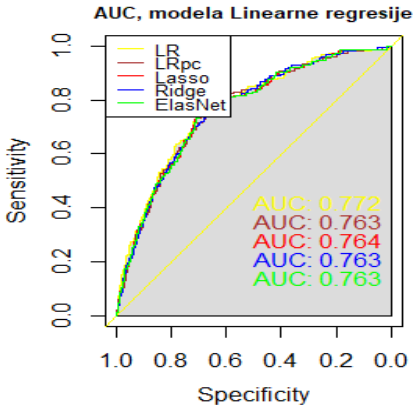
 # Elastic net regresija

Kreiranje modela

model_elasnet<- glmnet(x_train,y_output, alpha = elas$bestTune$alpha, family = "binomial",lambda = elasticnet_param$lambda.min)

Pokazatelji performansi regresionih modela

	Model	Tacnost	Sensitivity	Specificity	F1	Kappa	AUC	BS
1	LinReg	0.719	0.727	0.718	0.213	0.136	0.772	0.047
2	LinReg_PC	0.654	0.805	0.645	0.196	0.114	0.763	0.047
3	Lasso	0.714	0.734	0.713	0.212	0.135	0.764	0.047
4	Ridge	0.715	0.701	0.716	0.205	0.127	0.763	0.047
5	ElasNet	0.71	0.734	0.709	0.21	0.132	0.763	0.047



 # LDA

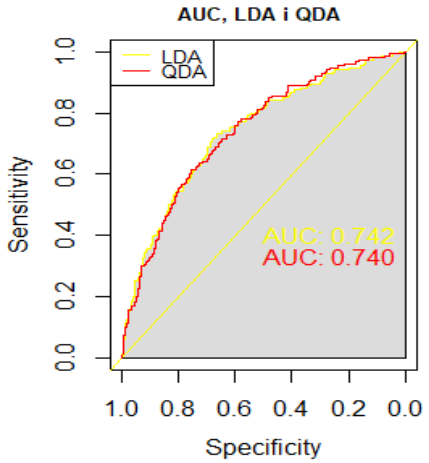
model_lda<-lda(default.payment.next.month ~ ., data=train_norm)

 # QDA

model_qda<-qda(default.payment.next.month ~ .,data=train_norm)

Pokazatelji performansi LDA i QDA modela

	Model	Tacnost	Sensitivity	Specificity	F1	Kappa	AUC	BS
1	LDA	0.681	0.721	0.679	0.192	0.111	0.742	0.048
2	QDA	0.75	0.617	0.758	0.206	0.13	0.74	0.266



 # STABLA ODLUCIVANJA

trControl=trainControl("cv",number=5) # sampling = "smote",Balansiranje trening podataka ML modela stabla sa sampaling opcijama smote

Kreiranje modela

model_tree<-train(default.payment.next.month~., data=train,method="rpart",trControl= trControl, tuneLength=10)# tuneLength daje moguce cp vrednosti

 # RANDOM FOREST

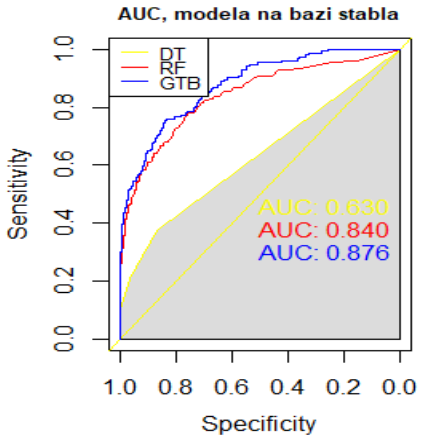
model_rf<-train(default.payment.next.month~., data=train,method="rf",trControl= trControl,tuneLength=5)# smanjana tunelength na 5, zbog performansi

 # BOOSTING MODEL, xgboost sa xgbTree. Koristi sve corove procesora, paralelno procesiranje

model_xgb<-train(default.payment.next.month~., data=train,method="xgbTree",trControl= trControl)

Pokazatelji performansi modela na bazi stabla

	Model	Tacnost	Sensitivity	Specificity	F1	Kappa	AUC	BS
1	DT	0.84	0.377	0.866	0.198	0.131	0.63	0.045
2	RF	0.754	0.779	0.753	0.249	0.177	0.84	0.039
3	GTB	0.834	0.753	0.838	0.322	0.261	0.876	0.035



Pokazatelji performansi modela na bazi stabla, balansirani podaci sa SMOTE

	Model	Tacnost	Sensitivity	Specificity	F1	Kappa	AUC	BS
1	DT	0.758	0.675	0.762	0.226	0.152	0.74	0.147
2	RF	0.699	0.864	0.69	0.231	0.155	0.842	0.06
3	GTB	0.784	0.766	0.785	0.271	0.202	0.846	0.049

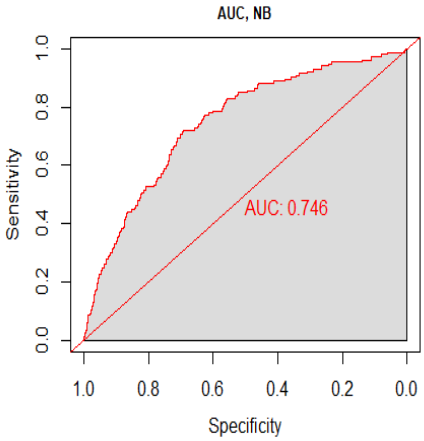
 # NB

Kreiranje modela

model_nb<-naiveBayes(train$default.payment.next.month~., data=train)

Pokazatelji performansi NB modela

	Model	Tacnost	Sensitivity	Specificity	F1	Kappa	AUC	BS
Accuracy	NB	0.693	0.721	0.691	0.198	0.118	0.746	0.224



 # KNN model

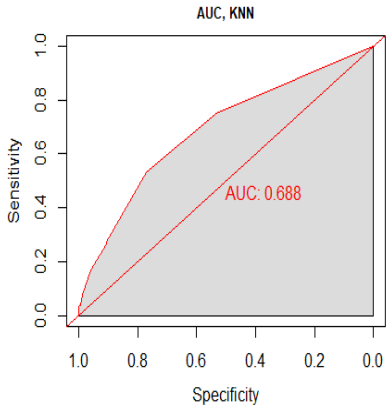
Kreiranje modela

model_knn<-train(Stecaj~., data = train, method='knn',tuneLength=10,

trControl=trControl,metric="ROC")

Pokazatelji performansi K-NN modela

	Model	Tacnost	Sensitivity	Specificity	F1	Kappa	AUC	BS
Accuracy	K-NN	0.759	0.532	0.771	0.188	0.111	0.688	0.048



 # SUPORT VECTOR MACHINE

 # Linear SVM

model_svmlin <- train(default.payment.next.month ~ ., data = train, method = "svmLinear", trControl=trControl)

SVM Radial. Kreiranje modela, SVM sa nelinearnom kernel funkcijom (Radial)

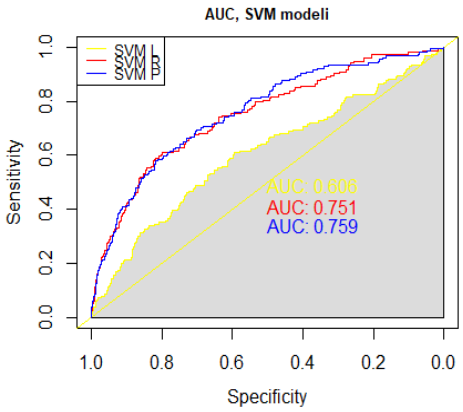
model_svmradial <- train(default.payment.next.month ~., data = train, method = "svmRadial",trControl=trControl)

SVM Poli

model_svmpoli <- train(default.payment.next.month ~ ., data = train, method = "svmPoly",trControl=trControl)

Pokazatelji performansi SVM modela

	Model	Tacnost	Sensitivity	Specificity	F1	Kappa	AUC	BS
1	SVM L	0.596	0.61	0.595	0.137	0.048	0.606	0.049
2	SVM R	0.79	0.61	0.8	0.234	0.163	0.751	0.047
3	SVM P	0.806	0.578	0.818	0.238	0.168	0.759	0.047



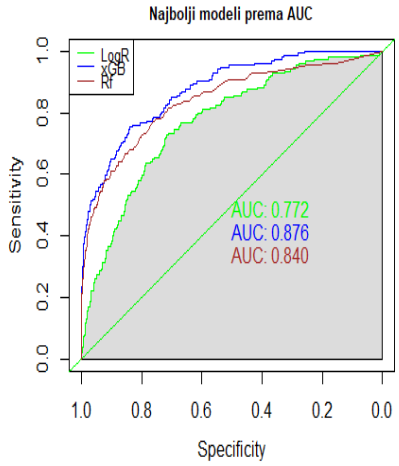
```
# Heterogeni modeli

# GBM

prvi_gbm <- h2o.gbm(x = x,

                    y = y,

                    training_frame = train_df_h2o,

                    nfolds = 5,

                    keep_cross_validation_predictions = TRUE,

                    seed = 5)

# GLM

drugi_glm <- h2o.glm(x = x,

                     y = y,

                     training_frame = train_df_h2o,

                     nfolds = 5,

                     keep_cross_validation_predictions = TRUE,

                     seed = 5)

# RF

treci_rf <- h2o.randomForest(x = x,

                         y = y,

                         training_frame = train_df_h2o,

                         nfolds = 5,

                         keep_cross_validation_predictions = TRUE,

                         seed = 5)

# Objedinjavanje pojedinacnih predikcija, stacking

ensemble <- h2o.stackedEnsemble(x = x,

                            y = y,

                            metalearner_algorithm="drf",

                            training_frame = train_df_h2o,
```

base_models = list(prvi_gbm, drugi_glm,treci_rf))

Pokazatelji performansi primenom h2o paketa: GLM, RF,GBM i heterogeni ensembling model

	Model	Tacnost	Sensitivity	Specificity	F1	Kappa	AUC
1	RF_h2o	0.755	0.76	0.755	0.246	0.173	0.821
2	GLM_h2o	0.695	0.747	0.692	0.204	0.125	0.77
3	GBM_h2o	0.71	0.864	0.701	0.238	0.162	0.866
4	Ensemb	0.785	0.753	0.787	0.787	0.2	0.792

Modeli sa najvecim AUC



Kraj

IZJAVA O AUTORSTVU (PRILOG 3.)

Ime i prezime autora Klod Kolaro

Broj indeksa 2018/3007

Izjavljujem

da je doktorska disertacija pod naslovom

„Klasifikacioni algoritmi mašinskog učenja i njihova primena u finansijama“

• rezultat sopstvenog istraživačkog rada;

• da disertacija u celini ni u delovima nije bila predložena za sticanje druge diplome prema studijskim programima drugih visokoškolskih ustanova;

• da su rezultati korektno navedeni i

• da nisam kršio/la autorska prava i koristio/la intelektualnu svojinu drugih lica.

U Beogradu, Januar 2024

Potpis autora

Klod Kolaro

IZJAVA O ISTOVETNOSTI ŠTAMPANE I ELEKTRONSKE VERZIJE DOKTORSKOG RADA (PRILOG 4.)

Ime i prezime autora Klod Kolaro

Broj indeksa 2018/3007

Studijski program Ekonomija

Naslov rada „Klasifikacioni algoritmi mašinskog učenja i njihova primena u finansijama“

Mentor prof. dr Milan Nedeljković, redovni profesor

Izjavljujem da je štampana verzija mog doktorskog rada istovetna elektronskoj verziji koju sam predao/la radi pohranjenja u Digitalnom repozitorijumu Univerziteta u Beogradu.

Dozvoljavam da se objave moji lični podaci vezani za dobijanje akademskog naziva doktora nauka, kao što su ime i prezime, godina i mesto rođenja i datum odbrane rada.

Ovi lični podaci mogu se objaviti na mrežnim stranicama digitalne biblioteke, u elektronskom katalogu i u publikacijama Univerziteta u Beogradu.

U Beogradu, Januar 2024

Potpis autora

Klod Kolaro

IZJAVA O KORIŠĆENJU (PRILOG 5.)

Ovlašćujem Univerzitet Metropolitan da u Digitalni repozitorijum Univerziteta u Beogradu unese moju doktorsku disertaciju pod naslovom:

„Klasifikacioni algoritmi mašinskog učenja i njihova primena u finansijama“

koja je moje auorsko delo.

Disertaciju sa svim prilozima predao/la sam u elektronskom formatu pogodnom za trajno arhiviranje.

Moju doktorsku disertaciju pohranjenu u Digitalnom repozitorijumu Univerziteta u Beogradu i dostupnu u otvorenom pristupu mogu da koriste svi koji poštuju odredbe sadržane u odabranom tipu licence Kreativne zajednice (Creative Commons) za koju sam se odlučio/la.

1. Autorstvo (CC BY)
2. Autorstvo – nekomercijalno (CC BY-NC)
3. Autorstvo – nekomercijalno – bez prerada (CC BY-NC-ND)
4. Autorstvo – nekomercijalno – deliti pod istim uslovima (CC BY-NC-SA)
5. Autorstvo – bez prerada (CC BY-ND)
6. Autorstvo – deliti pod istim uslovima (CC BY-SA)

(Molimo da zaokružite samo jednu od šest ponuđenih licenci. Kratak opis licenci je sastavni deo ove izjave).

U Beogradu, Januar 2024

Potpis Autora

Klod Kolaro