# Tehnike rudarenja podataka za procjenu kreditnog rizika

**Pakšec, Matea**

**SVEUČILIŠTE U ZAGREBU**

**FAKULTET ORGANIZACIJE I INFORMATIKE**

**V A R A Ž D I N**

**Matea Pakšec**

# TEHNIKE RUDARENJA PODATAKA ZA PROCJENU KREDITNOG RIZIKA

**ZAVRŠNI RAD**

**Varaždin, 2019.**

**SVEUČILIŠTE U ZAGREBU**

**FAKULTET ORGANIZACIJE I INFORMATIKE**

**V A R A Ž D I N**

**Matea Pakšec**

**Matični broj: 44606/16-R**

**Studij: Ekonomika poduzetništva**

# TEHNIKE RUDARENJA PODATAKA ZA PROCJENU KREDITNOG RIZIKA

**ZAVRŠNI RAD**

**Mentor/Mentorica:**

Doc. dr. sc. Dijana Oreški

**Varaždin, rujan 2019**

*Matea Pakšec*

**Izjava o izvornosti**

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

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

**Sažetak**

Tema ovog završnog rada dobro je opisana samim naslovom. Danas, u urbanom svijetu, potrebni su krediti kako bi si pojedinci i poduzetnici ostvarili ciljeve koji su nemogući. Zbog same velike potražnje za kreditima i kreditiranjem, banke i ostale financijske institucije dolaze u pitanje kreditnog rizika. Svaka banka ima svoju strategiju prilikom odobravanja kredita za pojedinca ili poduzetnika. Pojedinci i poduzetnici podliježu provjerama banaka kako bi se ustanovila njihova kreditna sposobnosti, a ujedno banke zaštitile od loših klijenata. Ovakvi procesi zahtijevaju razne postupke koji mogu dugo trajati, pa čak i biti nepouzdani. Ovdje ulazi rudarenje podataka kao najnovija i najučinkovitija metoda sprječavanja kreditnog rizika. U ovom radu korištena je jedna od metoda rudarenja podataka pomoću koje ću pokušati riješiti određeni problem.

**Ključne riječi:** kreditni rizik; rudarenje podataka; prediktivni model; stablo odlučivanja

# Sadržaj

# 1. Uvod

Tema rada je procjena kreditnog rizika pomoću tehnika rudarenja podataka. Veliku zainteresiranost za ovu temu dobila sam upisavši kolegij Otkrivanja znanja u podacima koji sam položila i pomoću njega saznala i naučila tehnike i princip rada rudarenja podataka. Također, kreditni rizik i sami rizici banaka jako su me zanimala, te sam ih proučavala i učila tokom studiranja. Spoj ovih dva područja činila su mi se veoma zanimljivim, te sam se odlučila za ovu temu kako bih si proširila znanje i dobila neki uvid na koji način banke mogu smanjiti kreditni rizik, a ujedno pozabaviti se nekim od tehnika rudarenja podataka.

Davanje kredita jedan je od najčešćih procesa u današnje vrijeme. Svaka osoba i poduzeća mogu dobiti kredit za vlastite ili pak poslovne potrebe. Tim načinom dolazi se do novaca kojih trenutno nemamo i otvaraj nam vrata onome što prije nikad ne bismo uspjeli. Iako krediti uspijevaju dati novčana sredstva kako bi ispunili ciljeve, ono može biti veoma rizično za banke. Kao što svi znamo krediti su novčana sredstva banke koja ta novčana sredstva daje zajmoprimcu na korištenje i za to korištenje banka sebi zauzvrat uzima kamatnu stopu. Zbog kamatne stope sam kredit dobiva veću vrijednost te neki klijenti nisu u mogućnosti ispuniti svoj zadatak. Time se dovodi u pitanje procjena kreditne sposobnosti klijenta, odnosno smanjenje kreditnog rizika banaka. Ovaj rizik jedan je od najstarijih rizika banaka, te je dobila i najveću pozornost. U današnje vrijeme postoje razne tehnike otklanjanja ovakvih rizika. Najnovije tehnike koje su razvijene prijašnjih godina su tehnike rudarenja podataka. One su se pokazale kao najutjecajnije, najbrže i najbolje metode za procjenu kreditnog rizika. Također, nisu samo važne tehnike kojima će se procijeniti rizik, važno je da banka ima dobru politiku i da dobro ustraje u tome.

Kako bih krenula s radom, bili su mi potrebni podaci na kojima ću provesti analizu i dati primjere loših i dobrih klijenata za kreditiranje. Podatke koje sam pronašla predložila sam mentorici te ih je kasnije, uz pregled, odobrila. Promatranjem podataka zaključile smo da bi najbolja tehnika za ove podatke bilo stablo odlučivanja. Njime bih dobila dobro klasificirane podatke i vizualizaciju iz kojeg ću kasnije izvući podatke koji će mi dati informacije za učenje ovog problema.

# 2. Kreditni rizik i rudarenje podataka

Glavni čimbenik uspjeha financijskih organizacija je procjenjivanje kreditnog rizika. Jedan od ključnih temelja za procjenjivanje kreditnog rizika su modeli za procjenu, odnosno bodovanje kreditnog rizika. Ovi modeli su zanimljivi i jedni su od interesnih grana kojima se bave istraživači. Predviđanje kreditnog rizika dobiva na važnosti jer može doprinijeti toku novca, osiguranju naplate kredita i smanjenju potencijalnih rizika, stoga je ovaj rizik tipičan klasifikacijski problem rudarenja podataka. [1]

Rudarenje podataka, prije nekoliko godina, koristilo se samo u znanstvenim krugovima. Tek je nedavno uvedeno u poduzeća, kada je postalo jasno da je ova disciplina neizbježna za stjecanje komparativne prednosti poduzeća. Kao i ostala poduzeća banke su se okrenule istoj metodi kako bi smanjile sve svoje rizike. Banke dnevno bilježe ogromne količine podataka koje dobivaju od svojih klijenata što uvelike koristi rudarenju podataka. Pomoću podataka iz prošlosti i sofisticiranih metoda rudarenje podataka omogućava uspješno provođenje istraživanja i smanjenje kreditnog rizika kao i ostalih rizika koji idu uz bankarstvo. [2]

## 2.1. Rudarenje podataka

Rudarenje podataka ili otkrivanje znanja iz baze podataka je sustavni pristup pronalaženja osnovnih obrazaca, trenda i odnosa ukopanih podataka. Rudarenje podataka privuklo je pažnju i istraživače, pa čak i praktičare zbog široke primjene u ključnim poslovnim odlukama. [3]

Rudarenje podataka više se smatra vještinom nego znanošću. Iako smatramo da svako strojno učenje ide prema pravilima i ono je tako kako je napisano, rudarenje podataka nema takav slučaj. Rudarenje podataka nema recept za uspješno otkrivanje znanja iz podataka koje će sigurno rezultirati pronalaženje vrijednih informacija. Mogućnost uspjeha dolazi u koracima procesa koji pomažu pri pronalasku mogućeg pravog rješenja. [5]

Razvoj tehnologije ima važnu ulogu u kreiranju i održavanju kvalitete baze podataka. Dobro razvijena informatička infrastruktura omogućava spremanje i održavanje podataka koje je potrebno pohraniti, obraditi i analizirati. Analiza podataka moguća je samo uz dobro napravljene računalne programe koji omogućuju strojne analize podataka koji su spremljeni u bazi. Nakon analize programi dostavljaju, odnosno pronalaze informacije koje kasnije pomažu pri samom odlučivanju i koordiniranju poduzeća. Cilj rudarenja podataka je riješiti dva glavna zadatka, a to su usmjereni i neusmjereni zadaci. Usmjereni zadaci su klasificiranje,

procjenjivanje i predviđanje. Cilj ovih zadataka je pronalazak vrijednosti određene, odnosno ciljne varijable koja za izvršenje zadatka ne može jednako primjenjivati sve tehnike rudarenja podataka. Neusmjereni podaci su grupiranje prema afinitetu i klasteriranje, te prikazivanje i profiliranje. [4]

Klasificiranje je ispitivanje karakteristika novih objekata i njihove raspodjele u klase koje su prethodno određene. Ovaj zadatak izvršava radnje kao što su izrada modela i samo uvježbavanje modela na ranije klasificiranim primjerima. Koristi neuronske mreže, stablo odlučivanja i analizu veza. Drugi zadatak rudarenja podataka je procjenjivanje koje može odrediti vrijednost nepoznate varijable. Za procjenjivanje se koriste regresijski modeli, neuronske mreže i analize opstanka kako bi se individualni zapisi dobiveni procjenom mogli rangirati na temelju dobivenih informacija. Predviđanje i klasificiranje dva su različita zadatka rudarenja podataka. Jedina razlika između ova dva zadatka je to što predviđanje koristi povijesne podatke kako bi se klasificiralo buduće moguće ponašanje ili buduća vrijednost. Za predviđanje se koriste neuronske mreže, stablo odlučivanja te klasteri pomoću kojih možemo grupirati prema afinitetu, a nadalje utvrditi koje proizvode i usluge možemo kombinirati. Klasteriranje omogućava segmentacija heterogene skupine na više homogenih klastera gdje klase nisu unaprijed definirane, već se zapisi grupiraju na temelju međusobne povezanosti. Zadnji neusmjereni zadatak je profiliranje. Profiliranje se vrši na osnovi preferirane kupovine gdje se koriste stablo odlučivanja, asocijativna pravila i klasteriranje kako bi se došlo do rezultata. [4]

Dakle, proces rudarenja podataka najprije prolazi kroz proces uvježbavanja gdje je cilj pronaći najbolju moguću zakonitu ulaznu varijablu iz koje će proizaći određena izlazna varijabla. Uvježbavanje se radi i poboljšava nizom primjera gdje se pomoću različitih varijabli dolazi do zadane izlazne vrijednosti. Pomoću niza iteracija, parametara i metoda, dolazi se do preciznog, točnog, korisnog, reprezentativnog i jednostavnog modela. U model se kasnije dodaju neovisni podaci kako bi se provjerila točnost modela i vrijednosti. Nakon toga dolazimo do zaključnog stanja gdje dobivamo opis modela, odnosno funkciju koja nam daje znanje. Ovaj proces je prediktivni dio rudarenja podataka koji koristi statističke metode, tehnike strojnog učenja ili umjetne inteligencije kako bi se previdjela izlazna vrijednost zavisne varijable uz ovisnost nekoliko atributa ili nezavisnih varijabli. [3]

Istraživanja temeljena na rudarenju podataka mogu se podijeliti u dvije kategorije, a to su metodologija i tehnologija. Metodologija koristi vizualizaciju podataka, strojno učenje, statističke tehnike te neku bazu podataka koji služe za nastanak aplikacije u kojima se koriste klasifikacija, predviđanje, klasteriranje, sažimanje, ovisnost modela, analizu povezivanja podataka i sekvencijalnu analizu. Tehnološki dio predstavlja statističke metode, neuronske

mreže, stablo odlučivanja, genetske algoritme i neparametarske metode. Metode koje služe za otkrivanje znanja iz podataka su neuronske mreže, grupiranje podataka ili klaster analiza, genetski algoritmi, stablo odlučivanja, metoda potpornih vektora. Neuronske mreže se sastoje od umjetnih neurona i čvorova koji pomažu pri rješavanju problema umjetne inteligencije. Koriste se za prediktivno modeliranje te upravljanje aplikacija iz skupa podataka. Klaster analiza je zadatak grupiranja skupa objekata. Služi za iskopavanje podataka i uobičajena je tehnika za statističku analizu na poljima strojnog učenja, prepoznavanja uzoraka, analiza slika, pronalaženje podataka, bioinformatika, kompresija podataka i računalna grafika. Stablo odlučivanja sadrži ciljane varijable koje imaju diskretni skup vrijednosti gdje lišće prikazuje klase oznake, a grane predstavljaju veze značajki koje vode do tih oznaka klasa. Svaki unutarnji čvor odgovara ulaznoj varijabli, a čvor prikazuje vrijednost ciljane varijable. Stablo odlučivanja koristi se kao vizualno i eksplicitno predstavljanje odluka i odlučivanja. [7]

## 2.2. Metode rudarenja podataka za predviđanje kreditnog rizika

Statistička metoda temelji se na upotrebi teorije i testiranja hipoteza. Iako, se najviše upotrebljava u procjeni kreditnog rizika svaka od metoda ima i nekih mana. Mane linearne diskriminantne metode (eng. *Linear Discriminant Analysis* – LDA) jest ta da ima nedostatak točnosti, zbog toga što se predviđa linearni odnos između rezultata i nezavisnih varijabli te jednakost kovarijanci za dobre i loše klase. LDA je jedna od prvih metoda koja se je koristila za otklanjanje kreditnog rizika zbog svoje jednostavnosti i brzog postupka. Logistička regresijska metoda (eng*. Logistic Regression* – LR) pak daje jednostavnu prediktivnu formulu klasifikacije, ali nedostatak je potreba linearne ovisnosti nezavisnih varijabli s obzirom na zavisnu jer poznavanjem vrijednosti nezavisnih varijabli omogućava predikciju zavisnih. Treća metoda koja se koristi za procjenu kreditnog rizika je stablo odlučivanja. Stablo odlučivanja je, također, statistička metoda pomoću koje lako i jednostavno pomoću vizualizacije možemo doći do rješenja. Fleksibilnosti i jednostavnost ove metode pogoduje za mnoga područja istraživanja jer nije potrebno stručno znanje niti poznavanje domene koja se ispituje. Nedostatak ove metode je ta da ima velike oscilacije pri promjeni strukture modela što utječe na točnost modela i krajnjeg rezultata. [1]

U strojnom učenju postoje dvije vrste učenja: nenadgledano i nadgledano. Nenadgledano učenje je deskriptivno te na taj način grupira sve slučajeve na temelju postojećih podataka. Nadgledano učenje se svrstava u prediktivno učenje, što znači da ono predviđa na temelju nekih podataka. Klasifikacija je jedna od tehnika nadgledanog učenja koja

služi za predviđanje pripadnosti klasa iz prethodno klasificiranih podataka. Pomoću skupa primjera i podataka iz prošlosti tehnike nadgledanog učenja grade odnos između skupa deskriptivnih atributa i ciljanog atributa. U ovim učenjima dolazi do velikog problema pronalaska mjerljivih, reprezentativnih i jednostavnih modela. Također, može doći do podtreniranosti gdje model nije dovoljno oblikovan, nije prošao dovoljno primjera. Drugi slučaj pretreniranost je proces u kojem model podliježe previše modificiranja, odnosno oblikovanja samog modela. Zbog ovih problema postoje manje metode koje smanjuju probleme koji su mogući. Jedan od najjednostavnijih metoda je klasifikator najbližih susjeda (eng. *K-nearest neighbor classifiers* – KNN). KNN klasifikator se temelji na proučavanju analogije, odnosno srodnošću dvaju slučajeva i spada u nadgledano strojno učenje. Ako je uzorak nepoznat, KNN klasifikator, tom uzorku dodjeljuje klasifikaciju koja ima najviše njegovih bližnjih susjeda. Za ovakav tip učenja ne postoje nikakvi resursi koji pomažu klasifikaciji, već se klasifikacija provodi na temelju povijesnih podataka, odnosno primjera iz prošlosti koji su napisani u bazi podataka. Zbog udaljenosti između podataka, koriste se različite metrike, a kao najpoznatije i najučestalije koriste se euklidska i manhattan udaljenost kojima nije potreban stabilan predviđeni model prije klasifikacije, ali rezultata na kraju nije jednostavan i pouzdan. Drugi klasifikator koji definira uvjetnu vjerojatnost događaja je Naivan Bayesov klasifikator (eng. *Naive Bayesian classifier* – NB). NB pripada parametarskom i linearnom modelu nadgledanog strojnog učenja. Pomoću njega nastaje grafički model koji prikazuje vjerojatnosne ovisnosti između varijabli koje smo zadali. NB prikazuje efekt vrijednosti atributa za klasu koji je neovisan o samoj vrijednosti ostalih atributa. Jedini problem kod ovog klasifikatora je velika pristranost između predviđenog i stvarnog modela pa se samo koristi kad imamo malo primjera za učenje. [1]

Procjenjivanje kreditnog rizika jedan je od najvažnijih problema financijskih institucija te je u ovom području najveća posvećenost stručnjaka. Tehnike diskiminantne analize i logističke regresije najzastupljenije su pri obavljanju ovakvih problema, ali i kritizirane do strane stručnjaka zbog svojih snažnih pretpostavki o modelu. Zbog kritiziranja, okreće se prema umjetnim neuronskim mrežama, koje postaju alternativa za dosadašnje tehnike. Zbog izvanredne sposobnosti bodovanja kreditne sposobnosti, sposobnosti generalizacije te pridružene karakteristike memorije, umjetne neuronske mreže prikladniji su način pronalaska rješenja. Najčešći korišteni model neuronskih mreža je širenje prema natrag (eng. *backpropagation*) zbog slojeva koji dopuštaju da se rezultat jednog sloja dodatno obrađuje i uređuje gdje se stvara kompleksni sustav. Mane ove metode nalazi se u nemogućnosti prepoznavanja relativne važnosti ulaznih varijabli i određene poteškoće u tumačenju rezultata. Drugi pristup umjetnoj inteligenciji su metode potpornih vektora (eng. *Support Vector Machine* – SVM). Ovaj pristup smatra se novom i obećavajućom tehnikom za zadatke klasifikacije

podataka. Uspješno se primjenjuje na strojno učenje i prepoznavanje uzoraka poput prepoznavanja lica, kategorizacija teksta i tekstura klasifikacijskih problema. Problem ovog pristupa je vrijeme koje je potrebno za provedbu zadatka, odnosno proces stavljanja podataka u potporne vektorske strojeve traje predugo. Treći pristup umjetnoj inteligenciji su evolucionarne računalne tehnike (eng. *Evolutionary Computational Techniques*) u koje spadaju genetski algoritmi (eng. *Genetic Programming* – GP). Genetski algoritmi koriste se za povećanje preciznosti u kombinaciji s bilo kojom tehnikom klasificiranja. Iako se pokazala kao preciznija metoda od linearne regresije ili neuronske mreže, nedostatak je u kompliciranoj interpretaciji rezultata poslovne i financijske analize. [3]

## 2.3. Kreditni rizik

Banke su jedne od najvažnijih institucija na tržištu novca. Poslovi banaka vezane su uz funkcioniranje financijskog tržišta koji je vezan uz nacionalno gospodarstvo pa je zbog toga jako bitan element svake države. Tri su funkcije banaka: mjenjačka funkcija, funkcija primanja depozita i funkcija kreditiranja. Financijske institucije prikupljaju novčana sredstva te ih plasiraju u obliku kredita. Pomoću plasiranja kredita zauzvrat dobivaju aktivnu kamatu kojom pune svoje blagajne. Krediti banaka danas su značajne za poduzeća, ali i za pojedince kako bi zadržali istu razinu kontrole nad poduzećem, ostvarili niže novčane izdatke ili pak nemaju dovoljno količine novaca, pa je kredit jedini način financiranja. [8]

Banke su izložene višestrukim rizicima kao što su financijski, operativni, poslovni rizici i rizik događaja. Financijski rizici sastoje se od dvije vrste rizika: osnovni rizici i špekulativni. Osnovni rizici uključuju rizik likvidnosti, kreditni rizik i rizik solventnosti, a špekulativni kamatni, devizni i cjenovni rizik. Nadzor banaka osniva se na kontinuiranoj analitičkoj provjeri banaka i ona je jedna od ključnih čimbenika za ostvarenje stabilnosti i povjerenja. Kreditiranje je jedna od najvažnijih usluga banaka, te se svrstava u najrizičnije usluge. Banke mogu pružiti različite vrste kredita kao što su krediti za nekretnine, krediti financijskim institucijama, poljoprivredni krediti, komercijalni i industrijski krediti, krediti pojedincima, mješoviti krediti, potraživanja i financijski lizing. Kako bi se formirao dobar kreditni portfelj potreban je profil karakteristika tržišnog prostora koji neka banka opslužuje, ono također ovisi o veličini banaka gdje su veće banke veliki zajmodavci, a male pružaju male kredite. Kasnih 1980-ih godina počela su se razvijati nova tržišta na kojima je omogućeno stvaranje novih proizvoda i usluga. Samim time razvilo se je međunarodno financijsko tržište. Razvitkom financijskog tržišta dovelo je do povećanja financijskih instrumenata banaka što je omogućilo veći pristup izvorima financiranja. Povećanjem izvora financiranja dovodi do povećanja rizika banaka što je uvjetovalo

promjenom pristupa u uređivanju i nadzoru banaka. 60-ih godina 20-og stoljeća uvele su se bankovne kartice, te i prvi pravi krediti. Tih godina otkrivena je važnost procjene kreditnog rizika prilikom odobravanja kredita, što je kasnije, 1980-ih godina, predviđanje rizika koristilo kao pomoć pri donošenju odluka za odobravanje kredita. Danas je situacija puno drugačija, bankarski sustav se razvio, te su potrebne druge metode kojima bi se mogao predvidjeti rizik, no prediktivne metode iz prošlosti danas se koriste samo za stambene kredite, kredite male tvrtke te za prijavu i obnovu osiguranja. [9]

Procjena kreditne sposobnosti, u današnje vrijeme, važan je zadatak kreditnih institucija. Zbog velike konkurencije između financijskih institucija svaka od njih želi što više klijenta, što znači i veći kreditni rizik, a s time i brže metode za rješavanje ovakvih situacija. Razvitkom tehnologije, problem kreditnog rizika, a i ostalih rizika, ubrzano se rješava pomoću rudarenja podataka. Jedna od metoda, koja se je pokazala kao dobra alternativa, jesu umjetne neuronske mreže zbog svojih pridruženih karakteristika memorije i mogućnosti brze generalizacije. No, kritiziranjem ove metode zbog relativne važnosti potencijalnih ulaznih varijabli usporava proces obuke i stvara poteškoće u tumačenju. Najčešće tehnike rudarenja podataka za predviđanje kreditnog rizika su klasifikacijsko i regresijsko stablo (eng. *Classification and Regression Tree* - CART) i multivarijantna adaptivna regresijska krivulja (eng. *Multivariate Adaptive Regression Splines* – MARS). Ove tehnike pokazale su veliku učinkovitost predviđanja kreditnog rizika, te su premašile i tradicionalne tehnike kao što su diskriminantna analiza, logistička regresija, neuronske mreže, i vektorski sustav podrške. CART i MARS pristup pokazuje sposobnost modeliranja složenog odnosa između varijabli bez snažnih pretpostavki modela. Obje su sposobne prepoznati važne neovisne varijable kroz funkcije izgrađenog stabla i baze podataka, te ne trebaju dugi proces obuke što štedi vrijeme za modeliranje velikog skupa podataka. Također, u odnosu na druge tehnike klasifikacije, CART i MARS pružaju lako čitljive i jednostavne modele. Učinkovitost i točnost ove alternative pokazuju uspješne rezultate u provedbi zadatka predviđanja rizika. [3]

# 3. Prethodna istraživanja

U ovom dijelu rada objasnit ću neke od metoda za rješavanje istih ili sličnih tema koje su povezane s temom ovog rada. Metode ću objasniti pomoću ranijih istraživanja koja su javno provedena i objavljena na Internetu. Metode će mi pomoći u sagledavanju koje se sve tehnike mogu koristiti za procjenjivanje kreditnog rizika.

## 3.1. Tehnike rudarenja podataka za predviđanje točnosti podataka klijentovih kreditnih kartica

Istraživanje se bazira na slučaj neispunjavanja obveza kupaca na Tajvanu. Uspoređuje se šest metoda rudarenja podataka gdje je rezultat prediktivne točnosti procjene vjerojatnosti neplaćanja vrjednija od binarnog rezultata klasifikacije koja se dijeli na kredibilne i nekredibilne klijente. U posljednjih nekoliko godina izdavatelji kreditnih kartica na Tajvanu suočili su se sa krizom gdje potrošači sve više koriste kreditne kartice za pretjeranu potrošnju koja produbljuje dugovanja na karticama. Produbljenje dugova izazvala je veliku nesigurnost izdavateljima kreditnih kartica kao i samim potrošačima koji su izgubili povjerenje banaka. Kako bi se financijski sustav dobro razvijao potrebno je predviđanje rizika koji je jedan od postupaka upravljanja rizicima. Za predviđanje rizika potrebna su financijska izvješća o poslovanju, evidencija transakcija i otplata dugova koji su tada bili potrebni za danje statističke metode kao što su diskriminacijska analiza, logistička regresija i Bayesov klasifikator za razvoj modela predviđanja kreditnog rizika. Kasnije, s evolucijom strojnog učenja, umjetne neuronske mreže i klasifikacije, poboljšano je i olakšano previđanje kreditnog rizika kao i rješavanje vjerojatnosti kašnjenja otplate odobrenog kredita. Cilj ovog istraživanja bio je pronalazak najbolje metode od njih šest (diskriminacijska analiza, logistička regresija, Bayrsov klasifikator, najbliži susjed, umjetne neuronske mreže i klasifikacija stabla) kao i točnost rezultata procjene vjerojatnosti neispunjavanja obveza. [3]

Podaci iz istraživanja preuzeti su 2005. godine iz važne banke Tajvana gdje se baziralo samo na korisnike koji su imali kreditne kartice navedene banke. Ukupni broj uzorka bio je 25.000 od kojih su 5.529 bili vlasnici kartica sa zadanim plaćanjem, pa se koristila binarna varijabla (da=1, ne=0). Također, koristile su se i sljedeće varijable:

X1: iznos odobrenog kredita,

X2: spol,

X3: obrazovanje,

X4: bračni status,

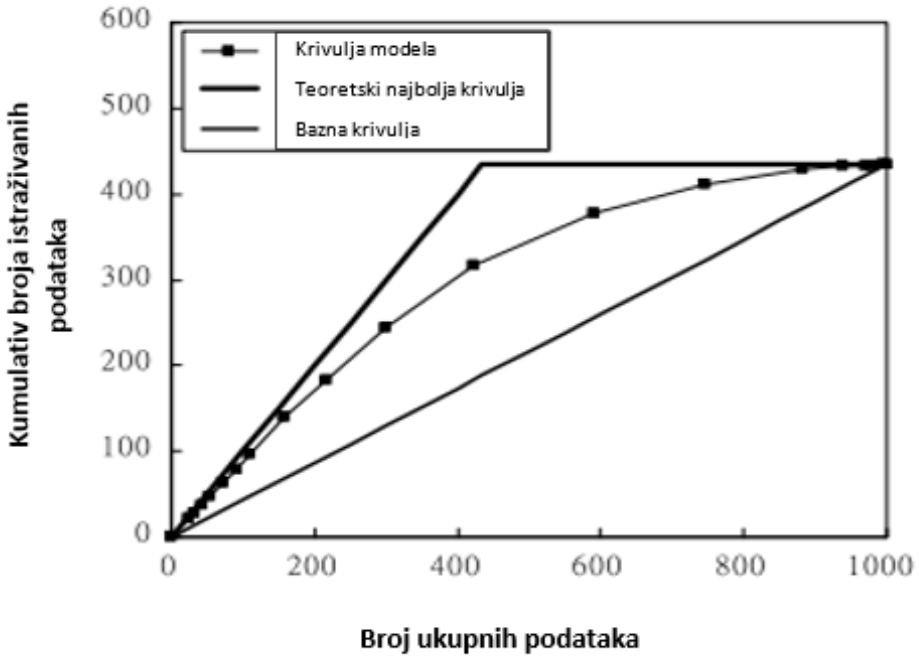
X5: dob,

X6 – X11: povijest prethodnih plaćanja (status otplate od travnja do rujna),

X12 – X17: iznos računa (iznos za prošlih 6 mjeseci),

X18 – X23: iznos prošlih plaćanja (iznos plaćanja za prošlih 6 mjeseci) (3)

Podaci su nasumično podijeljeni u dvije skupine, model obuke koja je koristila za uvježbavanje modela i drugi za provjeru ili validaciju modela. Za mjerenje točnosti klasifikacije modela koristi se stopa pogreške, ali u ovom istraživanju ona nije bila potrebna jer su podaci o klijentovim kreditnim karticama nerizični, stoga stopa pogreške nije osjetljiva na točnost klasifikacije modela, odnosno nema nikakav utjecaj. Kako stopa pogreške nema nikakav utjecaj na točnost klasifikacije modela, istražitelji su koristili omjer površine. Na grafikonima koje su prikazali horizontalna os predstavlja sveukupni broj podataka, a vertikalna os prikazuje kumulativni niz ciljanih podataka. Na grafikonu su prikazane tri krivulje (krivulja modela, teoretski najbolja krivulja i krivulja osnovica) iz kojih zaključuju što je veća površina između krivulje modela i krivulje osnovice, to je model bolji. Omjer površine se računa kao odnos površine između krivulje modela i krivulje osnovice zajedno i površina između teoretski najbolje krivulje i krivulje osnovice. [3]




Slika 1 Grafikon i krivulje (prevedeno, Izvor: I-Cheng, Che-hui, 2009)

Pomoću dijagrama dizanja željeli su utvrditi koja od metoda ima najbolje performanse. Validacijskim procesom na kraju su zaključili da je umjetna neuronska mreža najbolja metoda jer postiže najveći omjer površine od 0,54 i relativno nisku stope pogreške od 0,17.(3)

Kako bi utvrdili, odnosno dobili stvarnu vjerojatnost neplaćanja ispitivači su najprije poredali vrednovane podatke od minimuma prema maksimumu, s obzirom na procijenjenu vjerojatnost. Nakon poretka slijedi metoda razvrstavanja čija formula glasi:

$$P_i = \frac{Y_{i-n} + Y_{i-n+1} + \cdots + Y_{i-1} + Y_i + Y_{i+1} + \cdots + Y_{i+n-1}}{2n+1},$$

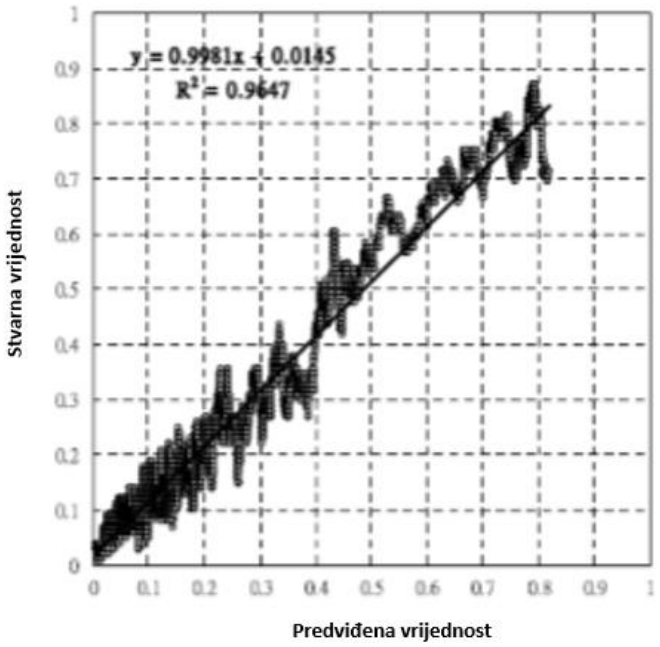
gdje je $Pi$ procijenjena vjerojatnost, redoslijeda kao i rangirani podatci, $Yi$ je binarna varijabla sa stvarnim određenim rizikom, u redoslijedu kao i rangirani podaci ($Yi$ = 1 , znači dogodilo se; $Yi$ = 0, znači nije se dogodilo), dok je $n$ broj podataka za razvrstavanje.(3)

Nakon što su utvrdili stvarnu vjerojatnost plaćanja, vjerojatnosti se prikazuju dijagramom raspršenja gdje vodoravna os predstavlja procijenjene vjerojatnosti neplaćanja, a okomita os označava procijenjenu stvarnu vjerojatnost neispunjavanja obveza. Iz dijagrama raspršenja dobivamo linearnu regresiju, odnosno koeficijent determinacije (R2) koji služi sa predviđanje buduće izlaze na temelju relevantnih informacija. Formula linearne regresije:

$Y = A + BX$,

gdje je $Y$ zavisna varijabla, a $X$ nezavisna varijabla.

Koeficijent determinacije je proporcija varijabilnosti u skupu varijabli iz statističkog modela, ako je on blizu 1, koeficijent $A$ blizu 0 i koeficijent regresije $B$ blizu 1, možemo zaključiti da je prediktivni model reprezentativan. Najviši nivo vjerojatnosti ovim pristupom pokazale su neuronske mreže, kao i u prethodnom istraživanju gdje je koeficijent determinacije bio najviši ($R_2$ = 0.9647) i samo je $B$ bio blizu vrijednosti 1, a $A$ blizu vrijednosti 0. [3]




Slika 2 Dijagram raspršivanje neuronske mreže (prevedeno, Izvor: I-Cheng, Che-hui, 2009)

U ovom dijelu rada istraživalo se šest metoda rudarenja podataka gdje se uspoređivala izvedba, klasifikacija i točnost predviđanja na danim podacima. U klasifikaciji točnosti između šest tehnika rudarenja podataka rezultati pokazuju da postoje male razlike u stopama pogreške, ali zato postoje velike u omjeru površine između tih tehnika. Vidimo da je omjer površine osjetljiviji od same stope pogreške te je on prikladan kriterij za mjerenje točnosti klasifikacije modela. Umjetne neuronske mreže točnije i bolje obavljaju klasifikaciju nego ostale tehnike te ih ispitivači preporučaju kako bi se procijenio kreditni rizik. [3]

## 3.2. Tehnike rudarenja podataka za procjenu kreditnog rizika

Procjena kreditnog rizika važna je operacija u bankarskim sustava kako bi se osiguralo da klijenti plaćaju kredite prema rasporedu. Cilj rada je identificirati čimbenike koji su potrebni ruralnoj banci *(*eng. *Bank Perkreditan Rakyat* - BPR) kako bi procijenila zahtjeve klijenata za kredit. Također, cilj je smanjenje broja nenaplativih zajmova. Model se primjenjuje na PT BPR X na Baliju koji je imao 1082 klijenata koji su imali loše kredite i identificirani su kao slučajevi loših kredita, zbog toga se PT BPR X smatra lošom izvedbom. [10]

Istraživanje je napravljeno na modelu stabla odlučivanja kako bi se procijenio kreditni rizik te ukazao može li se zahtjev klijenta klasificirati kao rizični ili loši zajmovi. Koristeći metodologiju C 5.0, generira se novi model stabla odlučivanja, ovaj model sugerira nove kriterije u analizi zahtjeva za kredit. Rezultati evaluacije pokazuju da, ako se primjenjuje ovaj model, PT BPR X može smanjiti nenaplative kreditne na manje od 5%, a banka se može klasificirati kao banka koja dobro posluje. [10]

Dobra i zdrava banka trebala bi imati BPR kao sljedeće stavke (Bank Indonesia, 2011.):

1. Omjer adekvatnosti kapitala – >8%
2. Produktivnost nenaplativih kredita - <5%
3. Omjer gotovine - >4%
4. Povrat na imovinu (ROA) - >1,3%
5. Omjer zajma i depozita – 80-90% (10)

Pokazatelj loših kredita (NPL) postaje ključni pokazatelj za BPR. Studija slučaja koristi se u BPR-u na Baliju koja je služila svojim klijentima s različitim proizvodima kao što su štednja, oročeni depoziti, usluga zapadnog sindikata i uzajamni zajmovi. Ova banka ima pokazatelje loših kredita od 11,99% što je znatno više od raspona koji su navedeni. Studija želi pomoći i usmjerena je na smanjenje ovog pokazatelja. [10]

Kreditna procjena je studija kojom se utvrđuje izvedivost zahtjeva za kredit, koristi kako bi se procijenilo ima li potencijalni klijent poslovne aktivnosti koje su izvodljive, profitabilne i hoće li se plaćati na vrijeme. Ova procjena izvodi se kako bi se analizirali svi čimbenici uključeni u zahtjev za kredit, kao što je financijski rezultat poslovanja i kreditni rejting klijenta. Za ovu analizu potrebna je analiza šest značajnih pokazatelja ( karakter, kapital, kapacitet, kolateralnost, stanje ekonomije, ograničenja). [10]

Proces kreditne procjene u PT BPR pratio je postupak podnošenja zahtjeva, provjeru podataka, ocjenu kreditne sposobnosti i odobrenje ili neodobravanje kredita. Podaci koji su korišteni za PT BPR X su :

1. Spol
2. Dob
3. Iznos kredita
4. Mjesečni prihod
5. Troškovi svaki mjesec
6. Tekuće plaćanje mjesečno
7. Štednja (plaćanje troškova dohotka)
8. Vrste kolaterala

9. Vrijednost osiguranja
10. Razdoblje zajma
11. Vrsta poslovanja
12. Izvor financiranja
13. Prethodni kreditni status/ocjena (10)

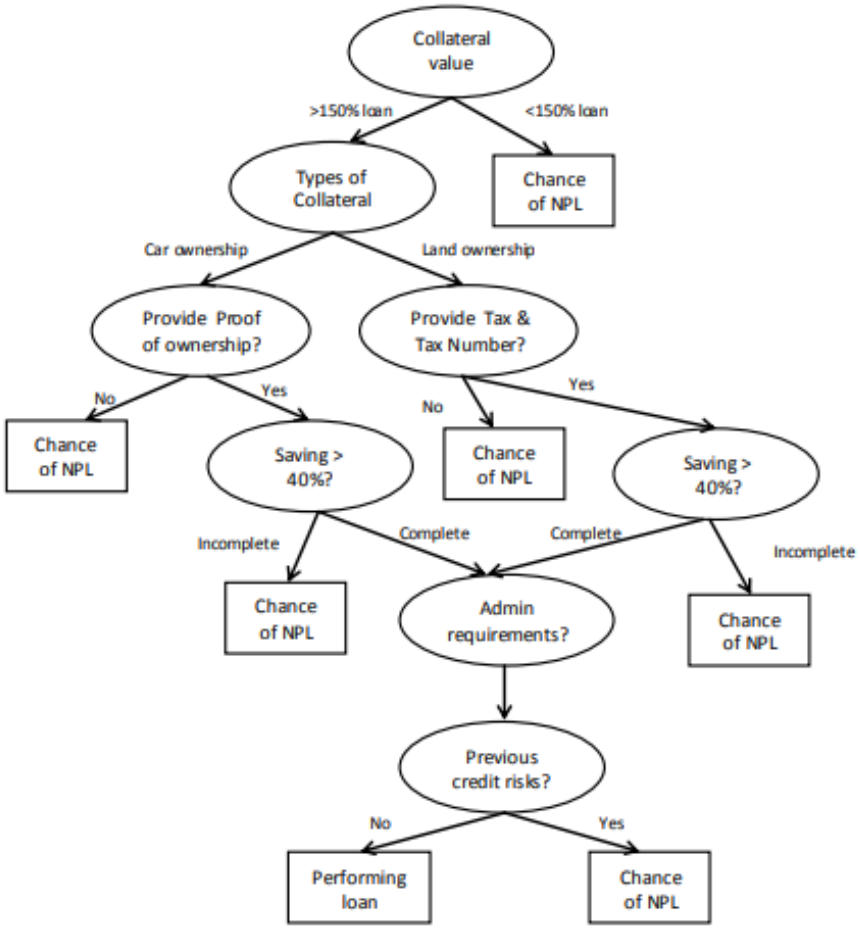


Figure 3: Credit assessment criteria

Slika 3 Model za procjenu kreditnog rizika (Izvor: Mandala, Nawangpalupi, Praktikto, 2012)

Početni model za kreditnu procjenu ima kriterij za procjenu kreditne sposobnosti kao vrijednost kolaterala koji je korijenski čvor, a završni čvor je prethodni kreditni rejting. Ostali čvorovi predstavljaju vrste kolaterala, kompletnost potrebnih dokumenta i spremanje. Ovaj model nije koristio sve atribute pa će se nadalje koristiti algoritam C5.0 za kreiranje novog modela procjene. [10]

U početnom stanju postoji pet diskretnih, odnosno nekontinuiranih varijabli koje su spol, vrste kolaterala, vrsta poslovnih aktivnosti, izvor financiranja, kreditni status i korištenje zajma. Nakon modeliranja dobiva se 8 kontinuiranih varijabli, a to su dob, mjesečni prihodi, troškovi svakog mjeseca, tekuće plaćanje mjesečno, štednje, vrijednosti osiguranja i rok posudbe. Ovaj model pokazuje da su kolateralne vrijednosti korijenski čvor, a slijedeći čvor list kao razdoblje posudbe i vrijednosti kolaterala. Završni čvor je vrsta kolaterala i vrsta poslovnih aktivnosti. Nakon modeliranja model procjenjuje da stopa NPL-a postaje 3%, što je niže od tražene stope i mnogo niše od trenutne postojeće. [10]

Kao najvažniji kriterij, u izgrađenom modelu, su vrijednosti kolaterala, odnosno korijen čvor. Iako postoje mnoge druge varijable u kriterijima za ocjenu kreditnog rejtinga, model pokazuje da nije mnogo njih te su varijable relevantne za kriterij kreditne procjene. Model je pokazao dobar rezultat te smanjio loše kredite i preporučuje se kao ocjenjivač kredita za PT BPR X. U slijedećoj tablici prikazani su rezultati iz trenutnih i predloženih modela za PT BPR X. [10]

Tablica 1 Rezultati iz trenutnih i predloženih modela za PT BPR X (prevedeno, Izvor: Mandala, Nawangpalupi, Praktikto, 2012)

| | Dosadašnji iznos (u tisućama rupia) | % | Predloženi model (u tisućama rupia) | % | Test (u tisućama rupia) | % |
|---|---|---|---|---|---|---|
| Obavljeni krediti | 36.996.600 | 88,01 | 36.531.700 | 94,14 | 10.519.900 | 96,75 |
| Neobavljeni krediti | 5.038.800 | 11,29 | 2.275.600 | 5,86 | 353.700 | 3,25 |
| Ukupno | 2.035.400 | 100,00 | 38.807.300 | 100,00 | 10.873.600 | 100,00 |

# 4. Opis istraživanja

Nakon opisa najznačajnijih i najviše korištenih metoda rudarenja podataka za predviđanje kreditnog rizika krećem sa svojim istraživanjem na njemačkim podacima. Konzultacijom s mentoricom odredile smo metodu stabla odlučivanja kao najbolje rješenje za određeni slučaj. Koristi se kao metoda za potrebe klasificiranja, predviđanja, procjenjivanja, klasteriranja, opisivanja i vizualizaciju podataka. Nabrojene stavke daju nam uvid da ova metoda ima široku primjenu, ali i jednostavni i razumljiv krajnji rezultat. Kako bi točno znali što je i kako funkcionira stablo odlučivanja prvo ću ga opisati, odnosno iznijeti teorijski dio stabla odlučivanja.

## 4.1. Metodologija stabla odlučivanja

Stablo odlučivanja jedna je od najčešće korištenih tehnika rudarenja podataka. Kao što je navedeno maloprije ona se koristi za klasificiranje, predviđanje, procjenjivanje, klasteriranje, opisivanje i vizualizaciju podataka. Sadrži pravila kojima se stabla brzo i jednostavno konstruiraju što na kraju daje jednostavan i razumljiv rezultat. Stablo je nastalo na bazi statističkih metoda raspoznavanja uzorka te je dobila značaj u rješavanju prediktivnih problema, odnosno predviđanje ciljne značajke u budućnosti. Kako bi se predviđala vrijednost ciljne značajke potrebne su ulazne varijable koje se klasteriraju u klase kako bi odgovarale ciljnoj značajki. Postoji više kriterija po kojima se dijeli stablo odlučivanja. Prema kriteriju ciljne značajke stablo odlučivanja dijelimo na regresijska stabla, koja služe kada imamo više značajki, a ciljna varijabla je realni broj, i na stabla za razvrstavanje, kada je ciljna varijabla diskretan skup vrijednosti. [11]

Stablo odlučivanja, kao klasifikacijski algoritam, u obliku je stablaste strukture. Stablo odlučivanja sadrži dva tipa čvora koje su povezane granama. Kranji čvor (eng. *leafe node*) predstavlja klasu kojoj pripadaju primjeri koji zadovoljavaju uvjete na nekoj grani stabla. Čvor odluke (eng. *decision node*) definira određeni uvjet u obliku određenog atributa, a iz njega izlaze grane koje odgovaraju i zadovoljavaju određene vrijednosti tog atributa. Osnovni preduvjeti korištenja tehnike stabla odlučivanja su: da podaci moraju biti opisani u obliku konačnog broja, moraju imati definiran broj klasa koje moraju biti diskretne, odnosno svaki primjer mora odgovarati samo jednoj postojećoj klasi i mora imati značaj broj primjera za kasniju generalizaciju. [12]

Iako postoji čitav niz algoritama kojima se može konstruirati stablo odlučivanja, jedan od najboljih i najčešće korištenih algoritama je C4.5. Prethodnik ovog algoritma razvijen je prije četrdesetak godina gdje ga je J. Ross Quinlan analizirao i predstavio u svojoj knjizi *Machine Learning* pod nazivom ID3. C4.5 omogućuje pronalazak određenog atributa u skupu podataka koji odvaja primjere određene klase, pomoću jednog određenog atributa. Proces se ponavlja sve dok se ne dobije atribut s najvećom informacijskom dobiti koja će kasnije postati onaj koji donosi odluku. Kako bi se precizirala informacijska dobit služimo se entropijom. Entropija je mjera homogenosti nekog skupa podataka. Na primjer, ako je zadan skup S i sadrži dvije klase, pozitivne i negativne primjere, entropija takve binarne klasifikacije je definirana sljedećim izrazom:

$$S = \ p_p \ \log 2 \ p_p \ p_n \log 2 \, p_n,$$

gdje $p_p$ označava postotak pozitivnih primjera u *S*, a $p_n$ postotak negativnih primjera u skupu *S*. Ako su svi primjeri u skupu *S* pozitivni tada vrijedi da je $p_p = 1$, odnsno $p_n = 0$, a entropija -1 i obrnuto. Ukoliko u skupu podataka postoji nejednak broj pozitivnih i negativnih primjera entropija poprima vrijednost u intervalu od 0 do 1. Ako ciljni atribut poprima više od dvije vrijednosti, tada je entropija u skupu *S*, u odnosu na klasifikaciju, definirana s:

$$S = \sum_{i=1}^{0} - p_i log_2 p_i,$$

gdje je $p_i$ postotak klase u skupu *S*. Ako ciljni atribut poprima *c* različitih vrijednosti, maksimalna entropija iznosi $\log_2 C$. Informacijska dobit, koja predstavlja očekivanu redukciju entropije uzrokovanu razdvajanjem primjera na osnovu tog atributa. Točnije, informacijska dobit atributa, u odnosu na skup, je definirana kao:

$$(S, A) = S - \sum_{v \varepsilon Values(A)} \frac{|S_v|}{|S|} (S_v)_1$$

gdje je *Values A* skup svih mogućih vrijednosti atributa *A*. Prvi član u jednadžbi za informacijsku dobit je entropija originalnog skupa *S*, dok je drugi član očekivana vrijednost entropije, nakon što je skup *S* razdvojen korištenjem atributa *A.* Informacijska dobit (*S, A*) je očekivana redukcija entropije uzrokovana poznavanjem vrijednosti atributa *A*, odnosno informacija o vrijednosti ciljnog atributa, uz poznate vrijednosti atributa *A*. Informacijska dobit bi trebala biti veći od 0.2 za varijablu koja se smatra prikladnom za uključivanje u uzorke. Iznos manji od 0.1 se smatra slabom varijablom, manji od 0.3 srednje jakom varijablom, a manji od 0.5 jaka varijabla. Ako je iznos informacijske dobiti veći od 0.5, obilježje takve varijable može biti previše predvidljivo, odnosno može se smatrati kako je na neki način trivijalno povezana s dobrom ili lošom informacijom. [12]

Iako je ova metoda jedna od najboljih, u nekim situacijama podliježe problemima kompleksnosti i efikasnosti stabla i atributa. Ovaj problem naziva se „over-fitting“. „Over-fitting“ je značajna poteškoća kod stabla odlučivanja, ali i drugih metoda za modeliranje podataka. Stablo odlučivanja sadrži algoritme koji mogu klasificirati sve primjere iz skupa podataka za učenje, iz kompleksnih podataka. Ovakav način uzrokuje i dodatne probleme kao što su šumovi u podacima ili pak ima nedovoljno velike uzorke podataka koji bi trebao reprezentirati populaciju za određeni klasifikacijski problem. Neovisno o kojem se slučaju radi, jednostavni algoritmi u poziciji su generalizacije stabla, što dovodi da stablo postaje pretjerano dobro („over-fitting“). Postoje razne mogućnosti za rješavanje ovog problema. Rješenja mogu biti grupirana kao rješenje koje zaustavlja proces rasta stabla prije postignuća savršene klasifikacije primjera ili pak kao rješenje u kojima se najprije generira stablo koje savršeno klasificira primjere, a kasnije se određene grane skraćuju prema nekom definiranom kriteriju. Općenito, glavno pitanje je kako ćemo odrediti optimalnu kompleksnost, tj. veličinu stabla za neki problem. Prvi pristup rješenju je korištenje posebnog skupa primjera koji je različit od onog korištenog za generiranje stabla kako bi se ocijenila uspješnost skraćivanja stabla. Ovaj pristup je najčešće korišten te je sastavljen od dva skupa. Skup za učenje koji se koristi za generiranje stabla i skup za provjeru, odnosno validaciju, koji se koristi za provjeru učinkovitosti. Drugi pristup za rješavanje problema je korištenje posebnog statističkog testa na čvorovima stabla koji bi bili potencijalni kandidati za skraćivanje. Treći pristup koristi eksplicitne mjere kompleksnosti kodiranja primjera stablom odlučivanja koja zaustavlja rast stabla kada je taj kriterij zadovoljen. [12]

Kao i svake druge metode koje postoje, stablo odlučivanja, također, ima svoje prednosti i mane. Prednosti ove metode su da ima sposobnost generiranja razumljivih modela, odnosno rezultati na kraju procesa lako se uče. Stablo odlučivanja omogućava korištenje svih tipova atributa, kategorički i numerički, te jasno odražava važnost pojedinih atributa za konkretni klasifikacijski ili predikcijski problem. Velika prednost ove metode je da ne treba puno vremena niti puno memorije kako bi se provela, što uvelike znači samim istraživačima, bilo znanstvenicima ili amaterima. Slabe strane metode stabla odlučivanja jesu te da su manje prikladne za probleme kod kojih se traži predikcija kontinuiranih vrijednosti ciljnog atributa, te je sklona greškama u višeklasnim problemima koje imaju relativno mali broj primjera za učenje modela. Također, u nekim situacijama generiranje stabla može biti zahtjevan računalni problem. Neki od situacija kada postaje zahtjevan problem je kada se sortiraju kandidati na čvorove, te sama metoda skraćivanja stabla. Stabla odlučivanja nisu dobro rješenje za klasifikacijske probleme kod kojih su regije određenih klasa omeđene nelinearnim krivuljama u višedimenzionalnim atributnom prostoru. [12]

## 4.2. Istraživanje

Originalni podaci, na kojima ću raditi, sadrže 20 kategorijskih i tekstualnih atributa koje je pripremio profesor Hofmann. Svaki ulaz predstavlja osobu koja je uzela kredit u nekoj njemačkoj banci, te se oni dijele na dobre i loše klijente ovisno o vrsti atributa. U praksi podaci koji se koriste za istraživanja nikada nisu idealna te je potrebna selekcija. Pomoću selekcije odabiru se podaci koji su najkorisniji za ispitivanje. Selekcija označava odabir pravih atributa odnosno podskupova. Nakon selekcije, pomoću predobrade podataka podaci koji su selektirani se obrađuju u prikladnije oblike, a ti oblici transformiraju se kasnije iz npr., više tablica u jednu glavnu. Nakon toga dobivamo reprezentativan model. Takvi modeli mogu biti jednako kvalitetni kao i cjelokupan skup podataka. Zbog velike kompleksnosti originalnih podataka i nemogućnosti razumijevanja cijelog sistema atributi su smanjeni i jednako reprezentativni pa su opisani kao:

1. Dob (numerički)
2. Spol (muško, žensko)
3. Posao (0 – nekvalificirani i nerezidentni, 1 – nekvalificirani i rezidentni, 2- kvalificirani, 3 – visoko kvalificirani)
4. Stanovanje (vlastiti, najam ili besplatno)
5. Štednja (malo, umjereno, prilično bogato, bogato)
6. Provjera računa (numerički, u Deutsch Mark)
7. Iznos kredita (numerički, u Deutsch Mark)
8. Trajanje kredita (numerički, u mjesecima)
9. Namjena kredita (automobil, namještaj, oprema, radio, TV, kućanski aparati, popravci, obrazovanje, posao, odmor, ostalo)

Podaci sadrže četiri numerička atributa i šest kategorijalna atributa. Podatke nije bilo potrebno pretvarati u Excel format jer je postojeći dokument već bio pripremljen u tom formatu. Nakon što sam podatke skinula sa stranice provjerila sam ih, te ih smatram reprezentativnim i prikladnim za ispitivanje. Ovim postupkom završila sam s pripremom podataka.

Nakon pripreme podataka, atribute koje sam pripremila, povezala sam ih s alatom BigML pomoću kojih ću dobiti informacije i vizualizaciju samog procesa. BigML služi za istraživanje podataka i pronalaženje korisnih podataka, odnosno učenje iz podataka koje smo sami povezali s ovim alatom. Nastao je kako bi se približilo strojno učenje raznim ljudima, odnosno većoj populaciji koji su zainteresirani za ovo područje. Ovaj alat je vrlo jednostavan i kvalitetan za rad što omogućava brzo i lako svladavanje samog alata, a kasnije i jednostavan i brz pronalazak informacija. U nastavku slijede slike svakog atributa.

Tablica 2 Opis numeričkih atributa

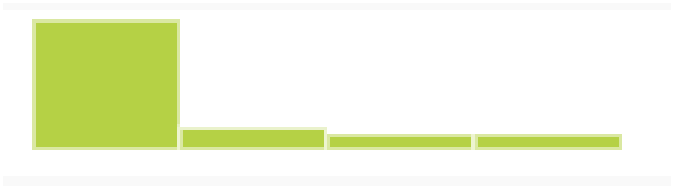
| | MINIMUM | ARTMETIČKA SREDINA | MEDIJAN | STANDARDNA DEVIJACIJA | MAKSIMUM |
|---|---|---|---|---|---|
| DOB | 19.00 | 35.55 | 33.00 | 11.38 | 75.00 |
| POSAO | 0.00 | 1.90 | 2.00 | 0.65 | 3.00 |
| IZNOS KRE | 250.00 | 3271.26 | 2319.50 | 2822.74 | 18424.00 |
| TRAJANJE | 4.00 | 20.90 | 18.00 | 12.06 | 72.00 |

Slika 4 Namjena kredita u ovisnosti s obzirom na broj godina klijenta, numerički atribut – uniformna distribucija iskrivljena udesno

Slika 5 Namjena kredita u ovisnosti sa spolom klijenta, kategorijski atribut – multimodalna distribucija

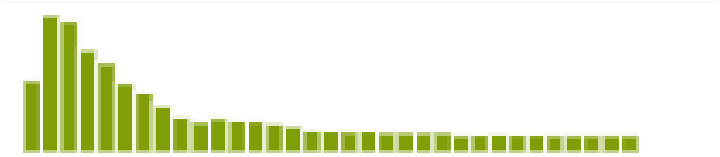
Slika 6 Namjena kredita u ovisnosti na posao klijenta, numerički atribut – uniformna distribucija

Slika 7 Namjena kredita u ovisnosti na stanovanje, kategorijski atribut – multimodalna distribucija

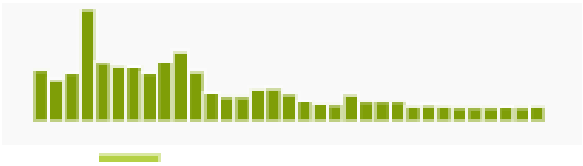


Slika 8 Namjena kredita u ovisnosti na štednju klijenta, kategorijski atribut – uniformna distribucija iskrivljena udesno

Slika 9 Namjena kredita u ovisnosti na provjeru računa u banci, kategorijski atribut – uniformna distribucija



Slika 10 Namjena kredita u ovisnosti na iznos kredita, numerički atribut – uniformna distribucija iskrivljen udesno

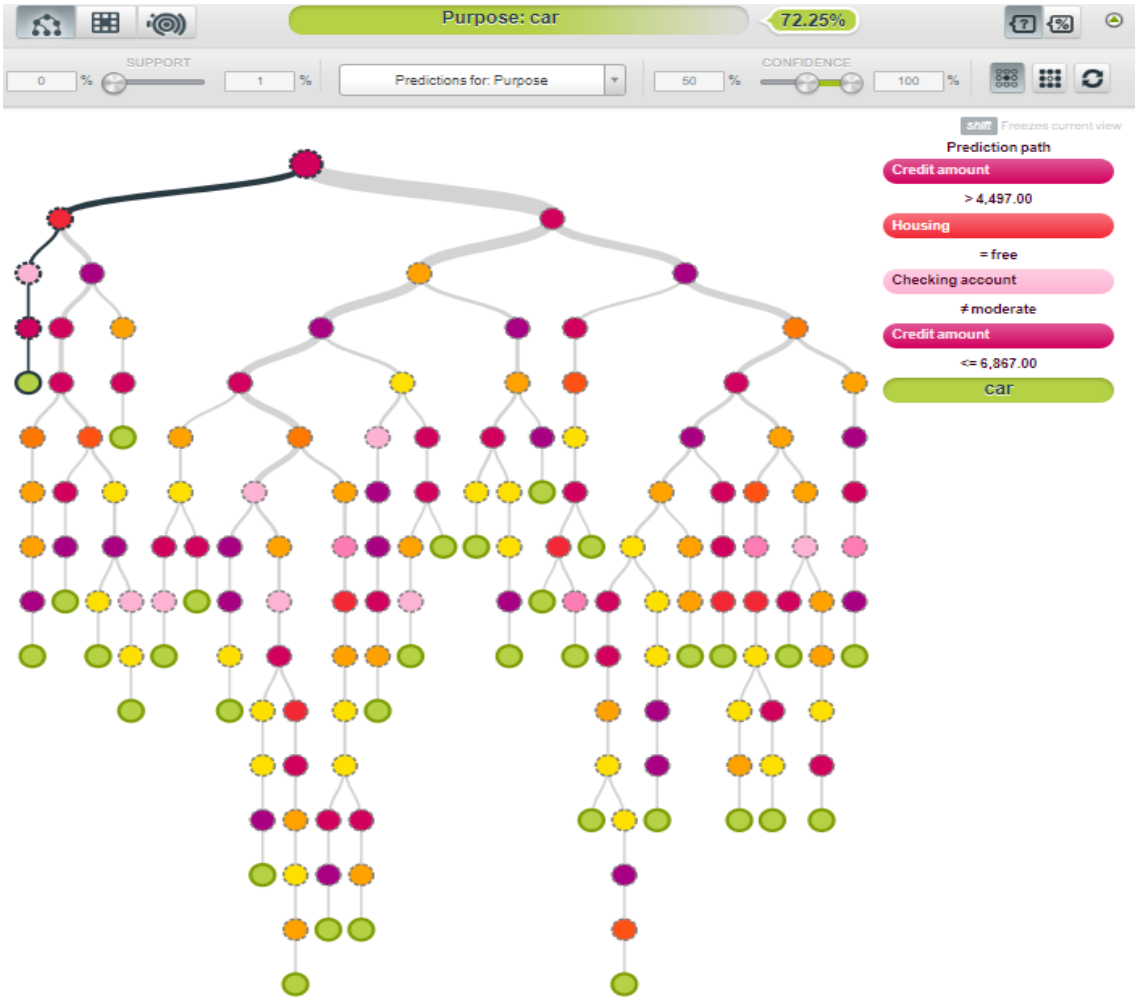


Slika 11 Namjena kredita u ovisnosti na trajanje kredita, numerički atribut – uniformna distribucija

Slika 12 Namjena kredita u ovisnosti na namjenu kredita, kategorijski atribut – multimodalna distribucija

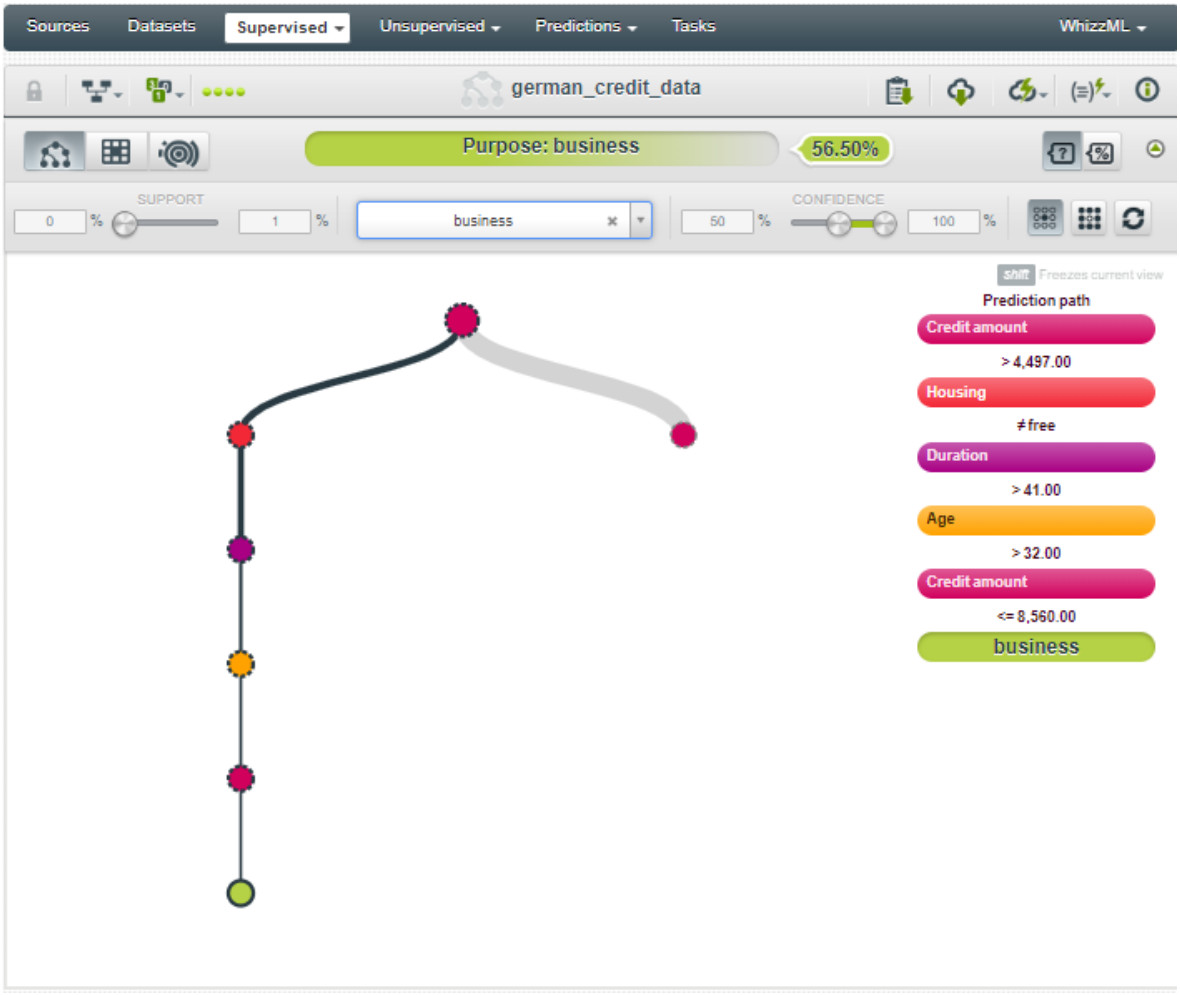
## 4.3. Rezultati istraživanja

Analiziranje podataka i provedba istraživanja radit će se pomoću alata BigML pomoću kojeg ću napraviti stablo odlučivanja ne temelju danih podataka koje sam obradila. U nastavku ću staviti slike te objasniti svaku od njih, odnosno izvući ću informacije.



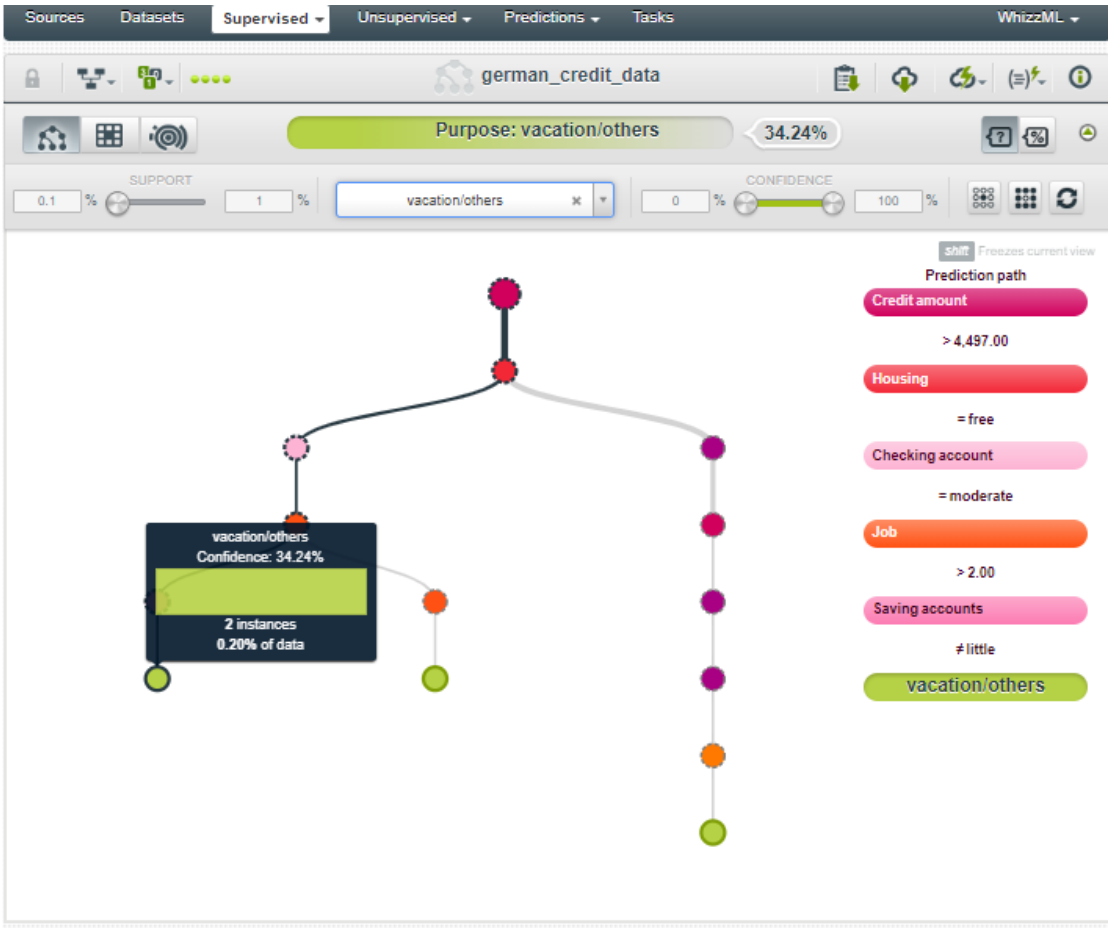

Slika 13 Stablo odlučivanja – dobar klijent

Prethodna slika predstavlja točnost od 72,25 %, što ovaj primjer čini dobrim primjerom. Primjer sadrži 10 instanci  iz skupa podataka na temelju čega klijent ima besplatno stanovanje i nema umjeren račun u banci. Ovaj kredit koristi kako bi kupio automobil, te ima mogućnost dobiti više od ili jednako 6.867 novčanih jedinica.



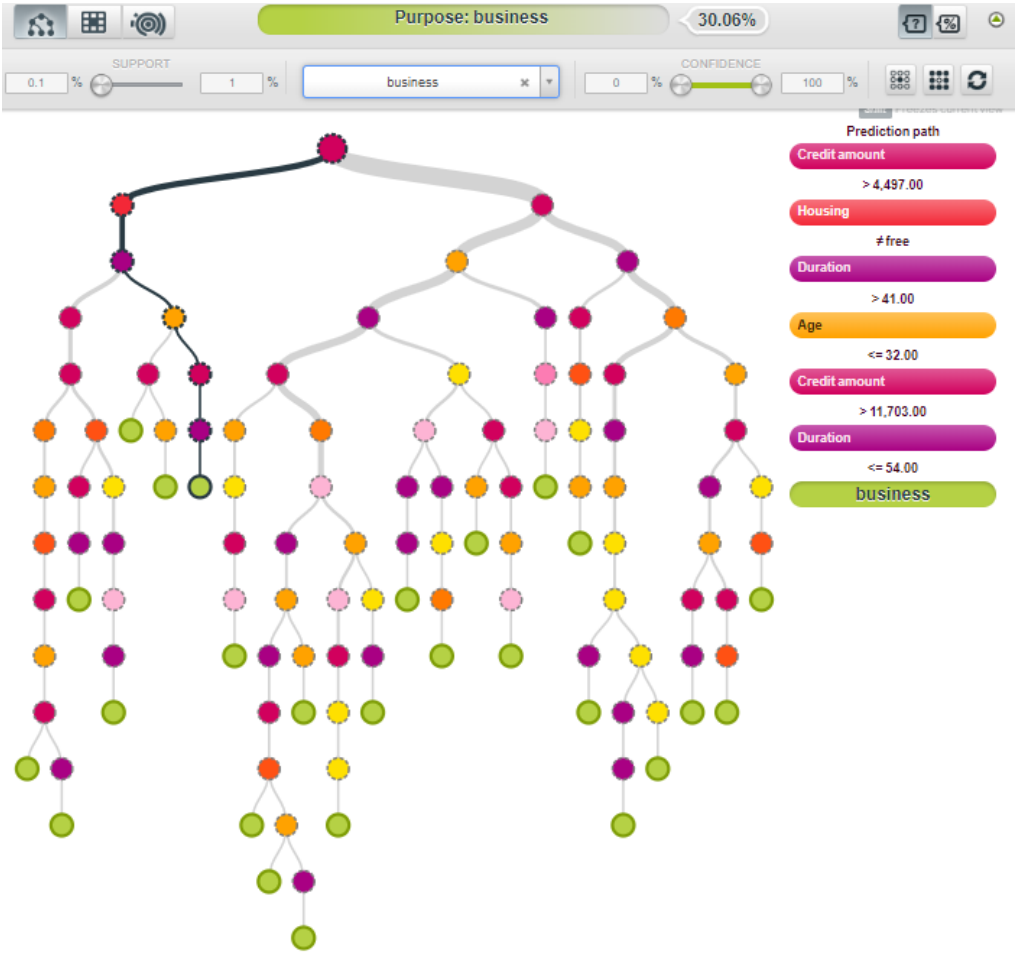

Slika 14 Odsječak stabla odlučivanja

Ovaj primjer prikazuje točnost od 56,50%. Klijent nema besplatno stanovanje, te ima manje od 32 godine. Svoj kredit želi iskoristiti u svrhu poslovanja, od čega bi dobio više ili jednako 8.560 novčanih jedinica u trajanju 41 jedinice.



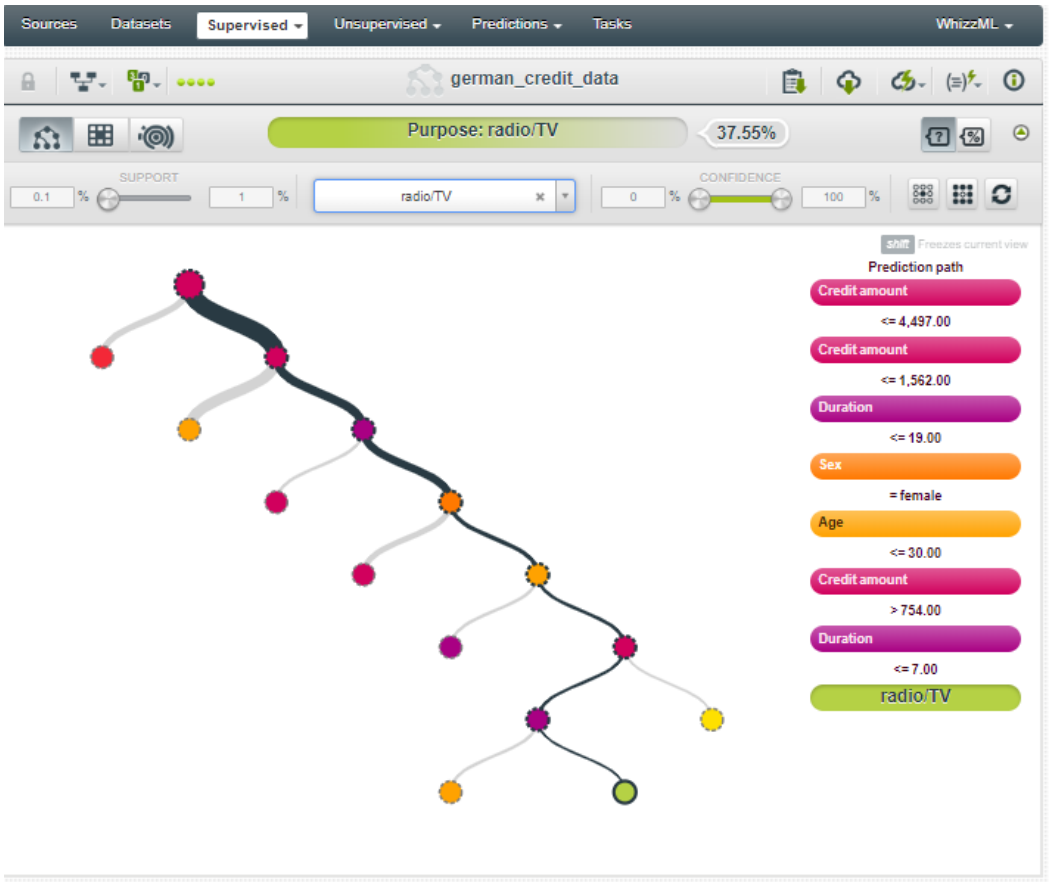

Slika 15 Odsječak stabla odlučivanja

Ovaj primjer prikazuje točnost od 32,24% što je jako nisko, te možemo reći da nije sigurno, odnosno da klijent nije pouzdan. Iznos kredita manje od 4,497 novčanih jedinica klijent bi koristio za godišnji odmor ili drugo. Klijent ima besplatno stanovanje, te se provjerom računa dokazalo da na računu ima umjerenu količinu novaca. Također, klijent radi manje od dva posla, te ne štedi.



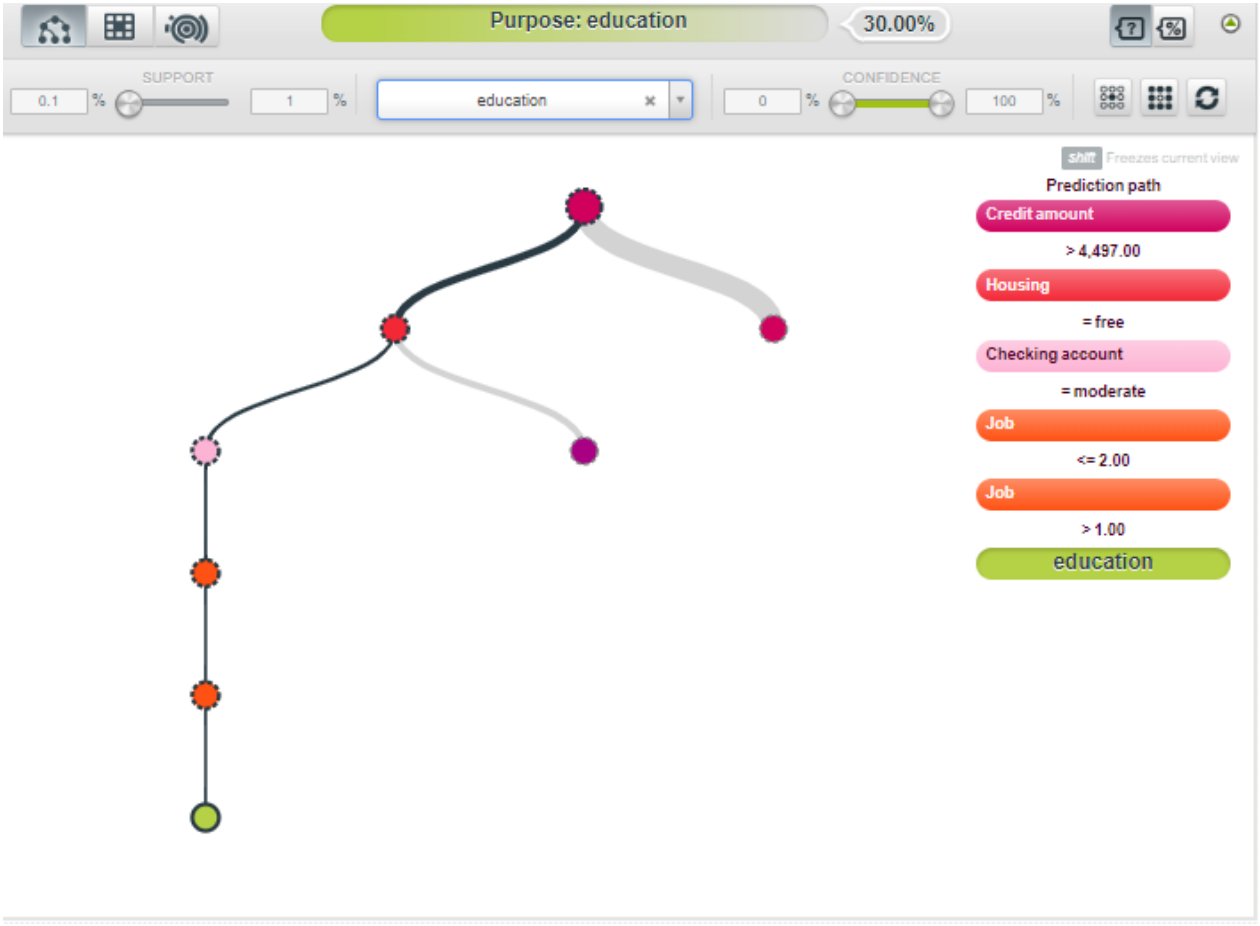

Slika 16 Stablo odlučivanja – loš klijent

U ovom primjeru klijent nema besplatno stanovanje i u dobi je više od ili jednako 32 godine. U mogućnosti je dobiti iznos manji od 11.703,00 novčanih jedinica u trajanju više od 54 jedinica kako bi unaprijedio ili započeo poslovanje. Ovaj primjer sadrži 30,06% točnosti što znači da je nepouzdan, te ovakav klijent je loš klijent.




Slika 17 Odsječak stabla odlučivanja

Primjer sadrži 5 instanci koje prikazuju 37,55% točnosti, što dokazuje da je ovaj primjer nepouzdan. Ovakav klijent se opisuje kao žena od 30 ili više godina koja je u mogućnosti dobiti kredit od 1562,00 novčanih jedinica na više od 19 jedinica, a za radio ili TV može dobiti manje od 754,00 novčanih jedinica u razdoblju većem od 7 jedinica.




Slika 18 Odsječak stabla odlučivanja - loš klijent

Primjer je prikazan pomoću 6 instanci koje prikazuju točnost od 30,00%, što ukazuje na nepouzdanost primjera, odnosno klijenta. Klijent je osoba koja ima besplatno stanovanje, račun u banci je umjeren, te ima jedan ili više od dva posla. Svoj iznos, koji je manji od 4.497,00 novčanih jedinica koristio bi za edukaciju.

# 5. Zaključak

Davanje kredita, odnosno kreditiranje, jedan je od najčešćih poslova banaka. Ovaj posao na sebe nosi razne rizike koje banka treba svladati. Također, ovakvi rizici ne utječu samo na banke već na cijelo financijsko tržište. Ono može dovesti do financijske krize koje će osjetiti sve države svijeta. Kreditiranje mene kao pojedinca, također, čeka u budućnosti. Stoga smatram da bi svaka osoba trebala realno sagledati stvari i sama sebi odrediti mogućnosti. Iako je kreditiranje postalo jedini način izvora financiranja za pojedince ili poduzeća, ono se ne smije shvatiti kao nešto što svatko može priuštiti. Također, same banke trebale bi bolje upoznati svoje klijente o kreditima i njihovim posljedicama na financijskom tržištu. Banke, kao odgovorna poduzeća, trebaju ulagati u politiku kreditiranja i ne samo sagledati na profit već uložiti napore u edukaciju klijenata i razvitak alata koji će im pomoći u rješavanju problema svih rizika.

Pripremom podataka i izradom stabla odlučivanja, upoznala sam se sa mogućim situacijama koje bi mogle utjecati na kreditni rizik. Preko atributa i vizualizacijom stabla odlučivanja pronašla sam dobre i loše klijente za kreditne institucije. Loši klijenti većinom imaju besplatno stanovanje, a na računu kod banke imaju umjereno novaca, odnosno neku srednju vrijednost, također ovakvi klijenti imaju oko 30 godina te imaju 1 ili 2 posla. To su jedni od primjera koji sadrže par atributa, uobičajeno postoje više faktora koji utječu na loš primjer klijenta. Važno je dobro se upoznati sa karakteristikama svog klijenta kao što su osobne potrebe i njihove mogućnosti kako se ne bi doveli u financijske probleme, a ujedno i probleme za kreditne institucije.

# Popis literature


[1] N. Mohammed, S. Mohammed i M. Taha, "Credit Scoring using Data Mining techniques with particular reference to Sudanese Banks", International Conference on Computer, Electrical and Electronics Engineering (ICCEEE)

[2] Mirjana Pejić Bach, "Rudarenje podataka u bankarstvu" (siječanj 2005.) [Na internetu]. Dostupno: https://www.researchgate.net/publication/27210032_Rudarenje_podataka_u_bankarstvu [pristupano 20.08.2019].

[3] Chi-Jie Lu, Chih-Chou Chiu, Tian-Shyug Lee, Yu-Chao Chou, "Mining the customer credit using classification and regression tree and multivariate adaprive regression splines“, Computional Statistics & Data Analysis, pro. 2004. [Na internetu]. Dostupno: ScienceDirect, https://www.sciencedirect.com [pristupano 20.08.2019].

[4] Damir Dobrinić, Iva Gregurec, FOI, Varaždin, (2016). Integrirani marketing

[5] Željko Panjan, Katarina Ćurko, Zagreb,(2010). Poslovni informacijski sustavi

[6] Elizabeth Mays, "Credit Risk Modeling" (1998). [Na internetu]. Dostupno: https://books.google.hr/books?id=k0zB89BSNH8C&pg=PA130&dq=data+mining+techniques+for+credit+risk&hl=hr&sa=X&ved=0ahUKEwiK0t2BpP3jAhVp-SoKHcC3D10Q6AEIKjAA#v=onepage&q&f=true [pristupano 17.08.2019]

[7] "Rudarenje podataka" (09.09.2013.) [Na internetu]. Dostupno: https://hr.wikipedia.org/wiki/Rudarenje_podataka [pristupano 16.08.2019].

[8] doc.dr.sc. Marina Klačmer Čalopa, prof.dr.sc. Marijan Cingula,Varaždin,foi (2009). Financijske institucije i tržište kapitala

[9] P. S. Rose i S. C. Hudgins: Upravljanje bankama i financijske usluge. Zagreb: Mate d.o.o. (2015).

[10] I. Gusti Ngurah Narindra Mandala, Catharina Badra Nawangpalupi, Fransiscus Rian Praktikto, "Assessing Credit Risk: An Application of Data Mining in a Rural Bank" (2012). [Na internetu]. Dostupno: https://www.sciencedirect.com/science/article/pii/S2212567112003553 [pristupano 20.08.2019].

[11] "Stablo odlučivanja" (14.05.2002.) [Na internetu]. Dostupno: http://www.skladistenje.com/stabla-odlucivanja/ [pristupano 17.08.2019]

[12] "Stablo odlučivanja" (2001). [Na internetu]. Dostupno: http://dms.irb.hr/tutorial/hr_tut_dtrees.php [pristupano 17.07.2019]

# Popis slika

# Popis tablica