Procjena kreditnog rizika primjenom rudarenja podataka

Petra, Poljak

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Petra Poljak

PROCJENA KREDITNOG RIZIKA PRIMJENOM RUDARENJA PODATAKA

ZAVRŠNI RAD

Varaždin, 2018.

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Petra Poljak

Matični broj: 44029/15–R

Studij: Informacijski sustavi

PROCJENA KREDITNOG RIZIKA

PRIMJENOM RUDARENJA PODATAKA

ZAVRŠNI RAD

Mentorica:

doc. dr. sc. Dijana Oreški

Varaždin, rujan 2018.

Izjava o izvornosti

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

Autorica potvrdila prihvaćanjem odredbi u sustavu FOI-radovi

_____

Sažetak

Tema ovog završnog rada je poprilično dobro izražena u samom naslovu rada. S obzirom na to kako svaki pojedinac ovisi o osobnom izvoru financiranja, te kako je jedan od glavnih izvora financiranja kreditiranje, odlučila sam analizirati što to obilježava pojedince kao dobre ili loše kandidate za kreditiranje. Prilikom ispunjavanja zahtjeva za kredit, želja je svakog pojedinca da se isti odobri. Međutim, banke i slične organizacije imaju svoju strategiju poslovanja, te im je cilj naravno profit, a nikako gubitak. Zato im postaje sve važnije pronalaziti metode i tehnike za prepoznavanje svojih klijenata, kao dobrih ili loših potencijalnih kandidata za odobravanje kredita. Rudarenje podataka sve više dobiva na važnosti za rješavanje mnogo različitih problema ili za donošenje odluka u mnogim područjima, te tako postaje sve razvijenije prilikom procjenjivanja kandidata za kreditiranje te donošenje same odluke o kreditiranju od strane zajmodavca. Zato ću izabrati jednu od tehnika rudarenja podataka, kako bih sama ispitala obilježja pojedinaca koja ih dijele na dobre i loše kandidate za odobravanje kredita, uz primjenu unaprijed odabrane, pripremljene i opisane baze podataka.

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

Sadržaj

1. Uvod 1
3. Kreditni rizik i rudarenje podataka 2
3.1. Rudarenje podataka 2
3.2. Metode rudarenja podataka za procjenu kreditnog rizika 4
3.3. Procjenjivanje kreditnog rizika 6
4. Pregled prethodnih istraživanja 9
4.1. Tehnike rudarenja podataka za previđanje ponašanja vlasnika kreditnih kartica 9
4.2. Korištenje rudarenja podataka za poboljšanje procjene kreditiranja 13
4.3. Primjena hibridnog sustava za procjenu kreditnog rizika 17
5. Opis metodologije 22
6. Opis istraživanja 27
7. Rezultati istraživanja 35
8. Zaključak 42
Popis literature 43
Popis slika 44
Popis tablica 45

1. Uvod

Tema ovog završnog rada je dakle procjenjivanje kreditnog rizika primjenom rudarenja podataka. S obzirom na to kako me je u vrijeme odabira teme rada zanimao budući posao u bankarskom sektoru, temu sam odabrala motivirana i s ciljem približavanja domeni koja me zanima. Prethodno sam odlušala i položila kolegij Otkrivanje znanja u podacima, koji me upoznao s rudarenjem podataka te metodama i tehnikama koje se koriste za izradu deskriptivnih i prediktivnih modela, pa sam samo rudarenje podataka spoznala kao zanimljivo i jako korisno područje. Zbog svega toga, u dogovoru s mentoricom, nastao je spoj koji čini temu mog završnog rada.

Kreditiranje je važno gotovo za svakog pojedinca i poduzeće, koji žele pronaći izvor financiranja za neke svoje planove ili potrebe, ukoliko se nisu spremni financirati sami u trenutku potrebe. Banke i slične organizacije su pak s druge strane, kao ti koji mogu omogućiti realizaciju kredita, no imaju svoj interes i to je naravno dobitak. Loša politika kreditiranja može dovesti do propasti banaka i sličnih organizacija, zbog čega je za njih važno pronaći metode i tehnike, koje bi im pomagale razlikovati dobre i loše potencijalne kandidate za kreditiranje. Zato smatram kako je u njihovom interesu ulagati u takve metode, koje će izvoditi takvu vrstu predikcije što preciznije i uspješnije. Loše kreditiranje koje se provodi u prevelikom obujmu ili predugo, dovodi do financijskih kriza pa je sprječavanje lošeg kreditiranja itekako važno. Rudarenje podataka se naglo razvija i prepoznato je od strane mnogih područja, kao skup metoda i tehnika rada koje se mogu iskoristiti za predviđanje ponašanja i donošenje važnih odluka. Dakako, prepoznato je i u bankarskom sektoru gdje se sve više koristi za upravljanje rizikom kreditiranja. Bankama je neizbježno riskirati prilikom kreditiranja, međutim zato trebaju dobru politiku kreditiranja, kako bi rizik svele na najmanju moguću mjeru.

Za početak pisanja samog rada, bilo je potrebno odabrati skup podataka na kojemu će se provoditi ispitivanje, što sam dogovorila s mentoricom. U izradu samog rada bilo je važno uključiti neku od tehnika rudarenja podataka, kako bi se provela klasifikacija potencijalnih kandidata za kreditiranje. Također u dogovoru s mentoricom, izabrano je stablo odlučivanja kao tehnika rudarenja podataka za izradu prediktivnog modela, odnosno za donošenje poslovnih pravila koja dijele potencijalne kandidate za kreditiranje na dobre i loše.

2. Kreditni rizik i rudarenje podataka

Jedan od ključnih faktora za uspjeh vodećih organizacija koje se bave posuđivanjem novca i posebno banaka, jest procjenjivanje kreditnog rizika. Modeli za procjenjivanje kreditnog rizika su zato postali predmet interesa mnogih istraživanja, kojima je cilj unaprijediti procjenu boniteta dužnika. Predviđanje kreditnog rizika dobiva na važnosti jer može doprinijeti toku novca, osiguranju naplate kredita i smanjenju potencijalnih rizika. Najčešće je procjenjivanje kreditnog rizika tipičan klasifikacijski problem rudarenja podataka. [1]

Banke i slične organizacije koje se bave posuđivanjem novca, u potrazi su za profitom za svoje dioničare. Kao takve nude financijske usluge svojim klijentima, uz upravanje raznovrsnim rizicima. Iako je riskantno, odobravanje zajma je i dalje jedna od ključnih usluga, koja donosi prihod komercijalnim bankama i ostalim sličnim organizacijama. Do kreditnog rizika dolazi kada se donose krive odluke, u vezi odobravanja zajmova. Kriva procjena kreditnog rizika dovodi do porasta broja neplatiša te tako može dovesti i do bankrota banaka. Odluke povezane s kreditnim rizikom su kompleksne i ne mogu se baš lako riješiti. U prošlosti, kreditni se rizik procjenjivao kroz osobne odluke, od strane službenika za zajmove, na temelju njihovog iskustva i analize podataka. Donošenje takvih odluka je trošilo previše vremena i bile su netočne. Zato se javila potreba za razvojem formalnih metoda, koje bi utjecale na što točnije predviđanje kreditnog rizika. Razvile su ih banke i istraživači, kako bi klasificirali klijente s obzirom na njihov različit nivo kreditnog rizika, na temelju dostupnih informacija o kreditiranju iz prošlosti. [1]

2.1. Rudarenje podataka

Rudarenje podataka (eng. *Data Mining - DA*) je sistematski pristup koji se koristi kako bi se pronašli temeljni obrasci, trendovi i veze u podacima. Cilj rudarenja podataka je razvoj metoda i tehnika za pronalaženje korisnih informacija u podacima. Predmet interesa jesu cjelokupni procesi i otkrivanje znanja, uz čišćenje, integraciju i vizualizaciju podataka. Zbog svoje primjenjivosti prilikom donošenja ključnih poslovnih odluka, ovo je područje privuklo veliku pažnju istraživača i praktičara. [2]

Redoviti proces rudarenja podataka ide najprije kroz proces uvježbavanja, a cilj je pronaći model koji će najbolje oponašati zakonitosti između ulaznih varijabli, kako bi se dobila određena izlazna vrijednost. Prilikom uvježbavanja, model se poboljšava nizom primjera jer se dostavljaju različite ulazne varijabile, za koje model mora postići zadanu izlaznu vrijednost. Tako se kroz niz

iteracija, uz korištenje različitih parametara i metoda, nastoji postići precizan, točan, koristan i jednostavan model. Takav model se dalje testira s neovisnim podacima, koji nisu sudjelovali u procesu uvježbavanja, kako bi se provjerila njegova točnost. Izlazna varijabla cjelokupnog procesa je funkcija, koja opisuje model, odnosno završetak procesa je znanje koje se može primijeniti za donošenje važnih odluka u području koje je predmet samog istraživanja. Dakle, ovako nastaju prediktivni modeli koji koriste statističke metode, tehnike strojnog učenja ili umjetne inteligencije, kako bi predvidjeli izlaznu vrijednost zavisne varijable, u ovisnosti od nekoliko atributa ili nezavisnih varijabli. [2]

Dakle, rudarenje podataka je netrivijalan proces identifikacije neospornih, novih, potencijalno korisnih i razumljivih uzoraka te odnosa među podacima u skladištu podataka. Skladišta podataka su ona koja sadrže povijesne, nepromjenjive podatke koji su tolerantni na pokreške u transakcijama. Skladišta često pohranjuju podatke iz mnogih izvora, podaci se čiste, spajaju, vežu uz varijable i sažimaju na niz korisnih načina. U sustavima poslovne inteligencije, rudarenje podataka ne mora biti nužno podržano skladištima podataka, već se može jednostavno definirati kao pronalaženje zakonitosti u podacima. Poslovna inteligencija je pojam koji se može identificirati s analizom informacija i potporom sustavima odlučivanja. Samo poimanje rudarenja podataka, koje se odnosi na cjelokupni proces otkrivanja znanja iz raspoloživih podataka, uključuje sljedeće faze: određivanje cilja prema potrebama korisnika, priprema podataka (između ostalog, izgradnja skladišta podataka), rudarenje podataka, ocjena i upotreba otkrivenog znanja. Uže poimanje rudarenja podataka se pak odnosi na specifičnu fazu obrade podataka, te uključuje neku od sljedećih metoda: određivanje najbližeg susjeda, grupiranje, asocijativna pravila, stablo odlučivanja, neuronske mreže, genetski algoritmi i ostali. Kao takvo, rudarenje podataka se oslanja na visoke domete informatike, matematike i statistike, pri otkrivanju odnosa i struktura među varijablama i kreiranju novog znanja. [3]

Konkretno, istraživanja temeljena na rudarenju podataka se mogu podijeliti u dvije kategorije: metodologija i tehnologija. Metodologija se sastoji od vizualizacije podataka, strojnog učenja, statističkih tehnika i deduktivnih baza podataka. Iz ovakvog pristup nastaju aplikacije koje koriste klasifikaciju, predviđanje, klasteriranje, sažimanje, ovisnost modela, analizu povezivanjem podataka i sekvencijalnu analizu. Tehnološkom dijelu rudarenja podataka pak pripadaju statističke metode, neuronske mreže, stablo odlučivanja, genetski algoritam i neparametarske metode. Osim navedenih metodologija i tehnologija, dugo vremena se za probleme klasifikacijskog tipa koristilo proučavanje podataka prilikom donošenja poslovnih odluka, kao što

su financijsko predviđanje, detektiranje prijevara, marketinške strategije, kontrola procesa i slično. [4]

2.2. Metode rudarenja podataka za procjenu kreditnog rizika

Korištenje statističkih metoda se temelji na upotrebi teorije i testiranju hipoteza. Jedna od vodećih statističkih metoda za procjenu kreditnog rizik je linearna diskriminantna analiza (eng. *Linear Discriminant Analysis - LDA*). Također je poznata kao prva metoda, koja se koristila za izradu modela procjene kreditnog rizika. To je jednostavna metoda, no njezine su mane nedostatak točnosti, zbog pretpostavki linearnog odnosa između rezultata i nezavisnih varijabli te jednakost matrica kovarijanci za dobre i loše klase. Još jedna od statističkih metoda, koja je također često korištena, jest logistička regresija (eng. *Logistic Regression - LR*). Iako je prednost što ova metoda može proizvesti jednostavnu prediktivnu formulu klasifikacije, pri izradi takvog modela nedostatak je potreba linearne ovisnosti nezavisnih varijabli s obzirom na zavisnu varijablu, jer u regresijskom modelu poznavanje vrijednosti nezavisnih varijabli omogućava predikciju vrijednosti zavisnih varijabli. Stablo odlučivanja (eng. *Decision Tree - DT*) je također statistička metoda koja je jednostavna, ne traži stručno poznavanje domene koja se ispituje niti pretpostavke o distribuciji varijabli, te je vrlo primjenjiva za mnoga područja istraživanja. Iako se radi o vrlo fleksibilnoj metodi, nedostatak je taj što struktura stabla odlučivanja ovisi o promatranim podacima pa i mala promjena itekako utječe na strukturu modela, što onda utječe i na samu točnost modela. Kod stabla odlučivanja, informacijska dobit (eng. *information gain*) je statistička kvantitativna mjera za vrijednost atributa koja mjeri koliko dobro atribut razdvaja primjere prema njihovoj klasifikaciji. [1]

Postoji nenadgledano i nadgledano strojno učenje. Nenadgledano učenje je deskriptivno, i pronalazi prirodno grupiranje slučajeva na temelju postojećih neoznačenih podataka. Nadgledano učenje je pak prediktivno, kao što je klasifikacija koja služi za predikciju pripadnosti klasama, iz prethodno označenih (klasificiranih) slučajeva. Tehnike nadziranog učenja automatski grade odnos između skupa deskriptivnih atributa i ciljnog atributa, na temelju skupa primjera iz prošlosti. Najveći problem takvog učenja jest pronaći ravnotežu između jednostavnosti i kompleksnosti modela, podtreniranost i pretreniranost (treniranje se koristi kao termin za proces oblikovanja modela). Klasifikator najbližih susjeda (eng. *K-nearest neighbor classifiers - KNN*) jedna je od najjednostavnijih metoda strojnog učenja i temelji se na proučavanju analogije. Pripada neparametarskom i nelinearnom modelu nadziranog učenja. Ako se radi o nepoznatom

uzorku, KNN klasifikator će tom uzorku dodijeliti klasifikaciju, koju ima najviše njegovih najbližih susjeda. Ne postoji nikakav resurs koji pomaže u klasifikaciji, već samo baza primjera za učenje. Za udaljenost između podataka se mogu koristiti različite metrike, najpoznatije su pak euklidska i manhattan udaljenost. Prednost ovakvog pristupa je što ne zahtjeva stabilan prediktivni model prije klasifikacije, ali rezultat ove metode nije jednostavan i pouzdan model. Naivan Bayesov klasifikator (eng. *Naive Bayesian classifier - NB*) pripada parametarskom i linearnom modelu nadziranog učenja. Svi Bayesovi klasifikatori se temelje na Bayesovoj formuli, koja definira uvjetnu vjerojatnost događaja. Nastaje grafički model koji prikazuje vjerojatnosne ovisnosti između varijabli koje se razmatraju. Efekt vrijednosti atributa za danu klasu je neovisan spram vrijednosti ostalih atributa. Problem ovakvog pristupa je visoka pristranost, koja je definirana kao razlika između predviđenog i stvarnog modela. Koristan je kada postoji malo primjera za učenje. [1]

Neke od metoda koje pripadaju pristupu umjetne inteligencije su: artificijalne neuronske mreže (eng. *Artifical Neural Network - ANN*), potporni vektorski stroj (eng. *Support Vector Machine - SVM*) i evolucijske računalne tehnike (eng. *Evolutionary Computational Techniques*). Iako su neuronske mreže visoko rangirane, što se preciznosti tiče, nisu upotrebljive ako se radi o irelevantnim atributima i jako velikom broju atributa te ih je teško iskoristiti za objašnjenje rezultata. Organizirane su u slojeve: jedan ulazni sloj, jedan ili više međuslojeva i jedan izlazni sloj, a u njima se nalaze elementi obrade. Najčešće korišteni model neuronske mreže je širenje prema natrag (eng. *backpropagation*), jer ima dodatne slojeve koji dopuštaju da se rezultat jednog sloja dodatno obrađuje i uređuje te se stvara kompleksni sustav, dok je nedostatak ovakvog modela dugotrajno izvođenje i osjetljivost na početne vrijednosti. Potporni vektorski stroj se uspješno koristi u većini studija, za koje su izabrani kao tehnika rada te se primjenjuju, na primjer, kod aplikacija za prepoznavanje lica i kategorizaciju teksta. Međutim, nedostatak je dugo vrijeme implementacije i nemogućnost primjenjivosti rezultata van kruga koji je predmet istraživanja te raznovrsni podaci. Neke od evolucionarnih računalnih tehnika su genetski algoritam (eng. *Genetic Algorithms - GA*) i genetičko programiranje (eng. *Genetic Programming - GP*). Genetski algoritam se može kombinirati s bilo kojom klasifikacijskom tehnikom, kako bi se povećala preciznost, dok su nedostaci veliki troškovi i otežano razumijevanje. Genetičko programiranje se pak pokazalo preciznijom metodom od recimo neuronskih mreža ili logističke regresije, no nedostatak je otežana interpretacija rezultata za financijske i poslovne analize. [2]

S obzirom na to kako ne postoje pouzdani zaključci o tome je li bolje koristiti statističke metode ili metode koje pripadaju umjetnoj inteligenciji za procjenjivanje kreditnog rizika, u

posljednje se vrijeme razvija hibridni pristup za unaprjeđenje aplikacija koje su nastale na temelju tehnika iz područja umjetne inteligencije. Tako su nastali hibridni sistemi koji su povezani i kompaktni te se sastoje od dvije ili više jednostavnih tehnika. Takav sistem može nadići nedostatke jedne od tehnika koja se koristi i postići prednost ispred ostalih tehnika. Sljedeći pristup ovakvog tipa se naziva ansambl metode i predstavlja labavo spojen sistem, koji koristi više učenika za rješavanje istog problema, a svaki učenik predstavlja zasebnu jedinicu. Iako je takav pristup precizniji od bilo koje zasebne metode umjetne inteligencije, nedostatak su povećana potreba za pohranom podataka i klasifikacijom objekata te nedostatak transparentnosti, što može dovesti do ograničene upotrebe modela. [2]

2.3. Procjenjivanje kreditnog rizika

Staro bankarsko pravilo je kako nad kreditom moraju bdijeti i onaj koji uzima kredit i onaj koji ga pozajmljuje, te kako se novac pozajmljuje samo onome kome je potreban. Ipak, svakoj financijskoj krizi prethodila je neodrživa i nezdrava ekspanzija kredita koje su odobravale banke i slične ustanove. Ako su krediti i kreditne kartice dostupni, teško je očekivati suzdržavanje od stanovništva. Svakako, prije ulaska u kredit, treba razmisliti o osobnoj sposobnosti otplaćivanja (budući dohodak/poslovna stabilnost), dužničkom karakteru te dopunskim izvorima financiranja kredita (pokretna i nepokretna imovina). No, uvijek treba imati na umu kako je moguće doći do promjene ekonomskih uvjeta, što može bitno utjecati na mogućnost otplate kreditnih obveza. Smatra se kako je dobar kredit ulaganje u obrazovanje, nekretnine ili mali posao. Takav kredit ima nižu kamatnu stopu i obično se odbija od poreza. Što je pak kredibilitet klijenata niži, viša je kamatna stopa. Postoje fiksna i promjenjiva kamatna stopa, koje se pribrajaju glavnici kredita prilikom otplaćivanja. Prednost fiksne kamatne stope je što otplate ostaju iste i troškovi kredita se neće povećavati, dok je nedostatak što kamata ostaje ista, čak i ako se kamatne stope smanje te da početna stopa bude viša od promjenjive stope. Promjenjiva kamatna stopa ima prednost što je početna stopa niža od fiksne stope i ukupni troškovi su niži, ako kamatna stopa padne, dok su nedostaci ti ako se kamatna stopa poveća, jer se povećavaju i ukupni toškovi, te nije uvijek dostupna. Preporučuje se izbjegavati potrošačke kredite, zbog relativno visokih kamatnih stopa, a i stvari koje se kupuju tim novcem (automobil, odjeća, tehnički aparati i slično) s vremenom gube na vrijednosti. [5]

Kreditiranje poduzeća, država i pojedinaca jedna je od najvažnijih usluga koje banke i njihovi konkurenti daju, te je također među najrizičnijima. Vrste kredita koje banke mogu pružiti su: krediti za nekretnine, krediti financijskim institucijama, poljoprivredni krediti, komercijalni i

industrijski krediti, krediti pojedincima, mješoviti krediti, potraživanja i financijski lizing. Jedan od ključnih čimbenika oblikovanja kreditnog portfelja pojedinačnog zajmodavca je profil karakteristika tržišnog prostora koji banka opslužuje. Također je važna veličina banke. Veće banke su tipično veliki zajmodavci, dok manje banke tipično pružaju male kredite. Iskustvo i stručnost rukovodstva u odobravanju različitih vrsta kredita, također oblikuje sastav kreditnog portfelja, kao i kreditna politika kreditne institucije. Samo područje predviđanja kreditnog rizika je predstavljeno 1940–ih godina, a kroz godine je dobilo na važnosti i znatno se razvilo. 1960-ih, s pojavom kreditnih kartica, banke i slične institucije su otkrile važnost predviđanja kreditnog rizika prilikom procesa odobravanja kredita. 1980-ih se predviđanje kreditnog rizika koristilo kao pomoć pri donošenju odluka za odobravanje osobnih zahtjeva za kredit. U posljednjih se pak nekoliko godina, ovakvi prediktivni modeli koriste za stambene kredite, kredite za male tvrtke te za prijavu i obnovu osiguranja. [6]

Kreditni rizik je varijacija mogućih povrata, koji bi se mogli zaraditi na financijskoj transakciji, zbog zakašnjelog ili nepotpunog plaćanja glavnice i/ili kamate. Kao takav, kreditni rizik je zapravo širi pojam od rizika vezanog samo za kredite, jer se odnosi na svaku financijsku transakciju banke koja producira potraživanje za glavnicu i kamate. U najužem smislu riječi, kreditni rizik se odnosi na vjerojatnost neplateži, po ugovoru o novčanom kreditu. Djelatnost banaka je upravljanje rizicima, a ne njihovo izbjegavanje. Uspjeh banke leži u njenoj sposobnosti da predvidi i kvantificira ukupni rizik te je zato procjena kreditne sposobnosti jedan od najvažnijih prethodnih informacijskih procesa. Sam proces utvrđivanja kreditne sposobnosti se svodi na procjenu dužnikove volje i poslovne sposobnosti da ugovorenom dinamikom vraća kredit. [7]

Za izradu modela pri predviđanju kreditnog rizika, od statističkih metoda se najčešće koriste linearna diskriminantna analiza i logistička regresija. Međutim, LDA je kritizirana zbog potrebe isključivo kategorijskog tipa podataka i jednakosti matrica kovarijance za različite klase pa je alternativa postala LR, koja je efikasnija. No, njezin je nedostatak homogenost varijabli, što ograničava upotrebu aplikacija prilikom rukovanja s problemima koji se odnose na kreditni rizik. Neuronske mreže su alternativa za obje metode te se zbog velike primjenjivosti najčešće koriste, pogotovo ako ovisne i neovisne varijable rezultiraju kompleksnim nelinearnim odnosom. Međutim, važnost ulaznih podataka i otežano interpretiranje rezultata ograničavaju mogućnost korištenja aplikacija, koje su nastale tehnikom neuronskih mreža. Modeli koji nastaju korištenjem stabla odlučivanja ili hibridnih modela koji koriste stablo odlučivanja, ističu se transparentnošću koja je itekako važna. [1]

Sposobnost procjene kreditnog rizika svakako ovisi o strukturi i kvaliteti podataka na temelju kojih se provodi istraživanje, te je teško pronaći najbolju metodu koja bi se mogla koristiti za svaki slučaj istraživanja. Za što preciznije i točnije modele potrebno je nastaviti istraživanja u ovom smjeru, preporučuje se koristiti baze podatake iz više bankarskih sektora i da se podaci baziraju na važnim varijablama, te se orijentirati na transparentnost modela. Također bi bilo poželjno proširiti studije na ostala slična područja, kao što su analiza potrošačke košarice ili profit klijenata. [2]

3. Pregled prethodnih istraživanja

Ovaj dio rada je posvećen ranijim istraživanjima, koja su provedena i javno objavljena, na istu ili sličnu temu, koja je predmet ovog rada. Proučit ću kakve su se tehnike rada koristile u tim istraživanjima i kakvi su rezultati, što se tiče procjenjivanja kreditnog rizika.

3.1. Tehnike rudarenja podataka za predviđanje ponašanja vlasnika kreditnih kartica

Iz perspektive upravljanja rizicima, rezultat prediktivne analize će biti vrijedniji, nego rezultat binarne klasifikacije koja klijente dijeli na kredibilne i nekredibilne. U dobro razvijenom financijskom sustavu, kriza menadžmenta je u padu, a predviđanje rizika u porastu. Glavna svrha predviđanja rizika je koristiti financijske informacije, kao što su poslovni financijski izvještaji, transakcije klijenata ili zapisi o otplatama te slično, kako bi se predvidjeli poslovni događaji ili individualni kreditni rizici te kako bi se smanjila šteta i nepreciznost. Za predviđanje kreditnih rizika su se koristile mnoge statističke metode (linearna diskriminantna analiza, logistička regresija i ostale), a s razvojem umjetne inteligencije i strojnog učenja, u predviđanje kreditnog rizika se uključuju neuronske mreže i stablo odlučivanja. Kreditni rizik znači mogućnost odgode u otplaćivanju odobrenog kredita, pa je cilj ovog rada bio pronaći najprecizniju metodu od uključenih šest (linearna diskriminantna analiza, logistička regresija, naivni Bayesov klasifikator, klasifikator najbližih susjeda, neuronske mreže i stablo odlučivanja) za predviđanje kreditnog rizika i odgovoriti na pitanje jesu li rezultati takvih pristupa doista pouzdani? Istraživanje je provedeno na skupu podataka iz Tajvana. [8]

U današnje vrijeme vlada eksplozija informacija, te tvrtke proizvode i prikupljaju golemu količinu podataka svakodnevno. Otkrivanje korisnog znanja iz baze podatka i pretvaranje informacija u korisne rezultate jest veliki izazov i potreba. Rudarenje podataka je proces istraživanja i analiziranja velike količine podataka, kako bi se otkrili važni obrasci i pravila. Trenutno je rudarenje podataka neophodno kao potpora u sistemima za donošenje odluka i ima važnu ulogu u marketingu, službama za potrošače, detekcijama prijevara, vrednovanjima i ostalim područjima. Predmet interesa za procjenu kreditnog rizika je odlučiti treba li produžiti krediti, za koliko vremena produžiti kredit, što je to ugrožavajuće ponašanje i kako na njega reagirati. Procjenjivanje kreditnog rizika je zapravo termin koji se koristi za opis formalnih statističkih metoda, koje se koriste za klasificiranje podnositelja zahtjeva za kreditiranje na dobre i loše rizične

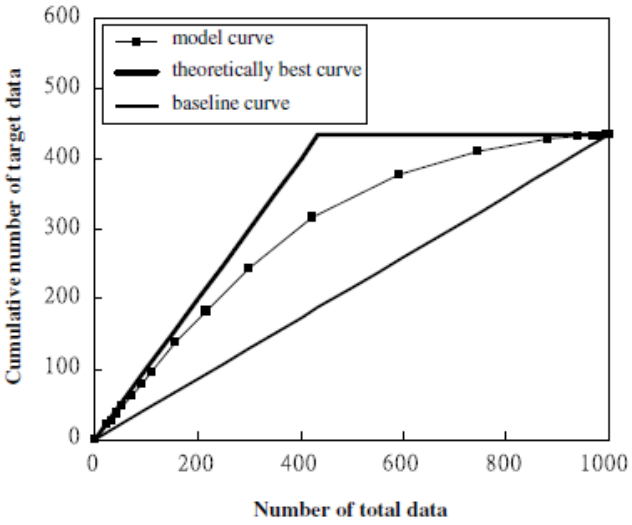
klase. Procjenjivanje takvog rizika je sve precizinje ako se koriste hibridne metode, koje kao takve nadilaze tradicionalne metode kao što su linearna diskriminantna analiza i logistička regresija. [8]

Ovo je istraživanje preuzelo podatke o plaćanjima iz listopada 2005. godine iz važne banka Tajvana, a provedeno je na podacima vlasnika kreditnih kartica te banke. Radilo se 25 000 uzoraka, a 5 529 ih je bilo vlasnika kreditnih kartica sa zadanim datumom uplate, pa je uvedena binarna varijabla (da = 1, ne = 0). [8]

Slijede preostale korištene varijable u istraživanju:

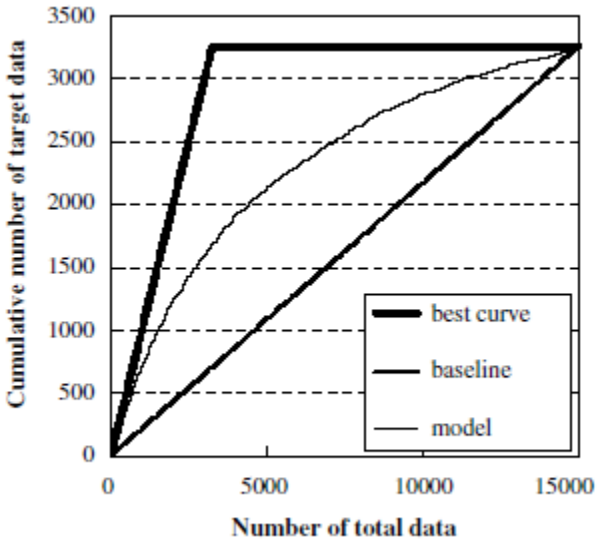
- X1: iznos odobrenog kredita,
- X2: spol,
- X3: edukacija,
- X4: bračni status,
- X5: dob,
- X6 – X11: prošlost plaćanja (status otplate za prošlih 6 mjeseci),
- X12 – X17: iznos računa (iznos za prošlih 6 mjeseci),
- X18 – X23: iznos prošlih plaćanja (iznos plaćanja za prošlih 6 mjeseci). [8]

Podaci su bili podijeljeni u dvije skupine, jedna za uvježbavanje modela i druga za validiranje modela. Većina podataka iz baze podataka nije bila rizična pa izračun pogreške nije bio važan za klasifikaciju preciznosti modela. Umjesto toga, istraživači su koristili omjer površine, kako bi izračunali preciznost klasifikacije između šest odabranih metoda. Grafikoni koje su koristili, imaju horizontalnu os koja predstavlja sveukupni broj podataka i vertikalnu os koja prikazuje kumulativni niz ciljanih podataka. Tri su krivulje: krivulja modela, teoretski najbolja krivulja i krivulja osnovica. 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, odnosno to je brojnik, dok je nazivnik površina između teoretski najbolje krivulje i krivulje osnovice. [8]



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

Prilikom validacije podataka, istraživači su zaključili kako su neuronske mreže imale najviši omjer površine (=0.54) i relativno malen iznos pogreške (=0.17). Stoga su zaključili kako su neuronske mreže najbolji model, između šest ispitanih metoda. Sveukupno, u skladu s rezultatima ovog ispitivanja, rang metoda je: neuronske mreže, stablo odlučivanja, naivni Bayesov klasifikator, klasifikator najbližih susjeda, logistička regresija i linearna diskriminantna analiza. [8]



Slika 2 Grafikon neuronske mreže (Izvor: I-Cheng, Che-hui, 2009)

Nadalje, ispitivači su računali stvarnu vjerojatnost modela, za što su najprije poredali vrednovane podatke od minimuma prema maksimumu, s obzirom na prediktivnu vjerojatnost. Nakon toga su koristili metodu razvrstavanja (eng. *Sorting Smoothing Method - SSM*), koja ima formulu:

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

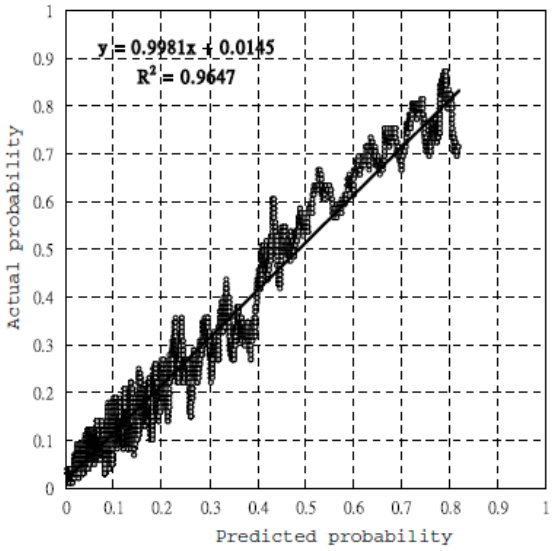
gdje je P_i procijenjena vjerojatnost, redoslijeda kao i rangirani podatci, Y_i je binarna varijabla sa stvarnim određenim rizikom, u redoslijedu kao i rangirani podaci (Y_i = 1 , znači dogodilo se; Y_i = 0, znači nije se dogodilo), dok je n broj podataka za razvrstavanje. [8]

Na temelju takvog pristupa, napravljeni su dijagrami raspršivanja, gdje horizontalna os predstavlja prediktivnu vjerojatnost, a vertikalna os procijenjenu stvarnu vjerojatnost. Korisno je napraviti nizove dijagrama raspršenosti jer mogu na jednostavan način sumirati odnose nekoliko parova atributa. Svaka točka na dijagramu raspršenosti je jedan par vrijednosti iz sloga podataka. Dakle, dijagram raspršenosti ukazuje na oblik odnosa između dvije varijable. Sljedeći kriterij za procjenjivanje vjerojatnosti modela je bio izračun linearne regresije. Linija regresije nastaje iz dijagrama raspršivanja i izračuna se koeficijent determinacije, čija je glavna uloga predviđati buduće izlaze na temelju relevantnih informacija. Formula za izračun linearne regresije je:

$$Y = A + BX,$$

gdje je Y zavisna varijabla, a X nezavisna varijabla. [8]

Koeficijent determinacije je proporcija varijabilnosti u skupu varijabli iz statističkog modela te ako je on blizu 1, koeficijent A blizu 0 i koeficijent regresije B blizu 1, tada se može zaključiti kako je prediktivni model, koji je nastao rudarenjem podataka, reprezentativan. Najviši nivo vjerojatnosti ovim pristupom, opet su pokazale neuronske mreže, kod kojih je koeficijent determinacije bio najviši ($R^2 = 0.9647$) i samo je u njihovom slučaju B bio blizu vrijednosti 1, a A blizu vrijednosti 0. [8]



Slika 3 Dijagram raspršivanja neuronske mreže (Izvor: I-Cheng, Che-hui, 2009)

Dakle, istraživanje je provjerovalo šest odabranih metoda rudarenja podataka, što se tiče klasifikacije preciznosti i vjerojatnosti modela. SSM metoda je pak korištena po prvi put, kako bi procijenila stvarna vjerojatnost modela. Istražitelji su zaključili kako su male razlike u iznosu pogrešaka između metoda, ali su velike razlike ako se računa omjer površine, koji je zato uzet kao valjano mjerilo za klasifikaciju preciznosti modela. Neuronske mreže su se pokazale najboljim izborom tehnike za izradu ovakvog modela, koje zato istražitelji preporučuju za korištenje u slučaju potrebe procjenjivanja kreditnog rizika. [8]

3.2. Korištenje rudarenja podataka za poboljšanje procjene kreditiranja

Cilj modela za predviđanje kreditnog rizika je podjela rizika kreditiranja na grupu dobrog rizika, koja je izgledna za otplaćivanje financijskih obveza i na grupu lošeg rizika, koja ima dobre izglede zakazati s otplaćivanjem dugovanja. Naravno, izgradnja takvih modela zahtjeva korištenje tehnika rudarenja podataka. Korištenjem podataka iz prošlosti o otplaćivanjima, demografskih karakteristika i statističkih tehnika, modeli za procjenu kreditnog rizika mogu pomoći pri identifikaciji važnih demografskih karakteristika, povezanih s kreditnim rizikom i napraviti procjenu za svakog klijenta. Dakle, rudarenje podataka podrazumijeva ekstrakciju obrazaca i pravila iz baze podataka. Uključuje identifikaciju poslovnog problema i samog cilja rudarenja podataka, te

korištenje tehnika za izradu modela, koji trebaju pomoći prilikom donošenja važnih strategijskih odluka. [9]

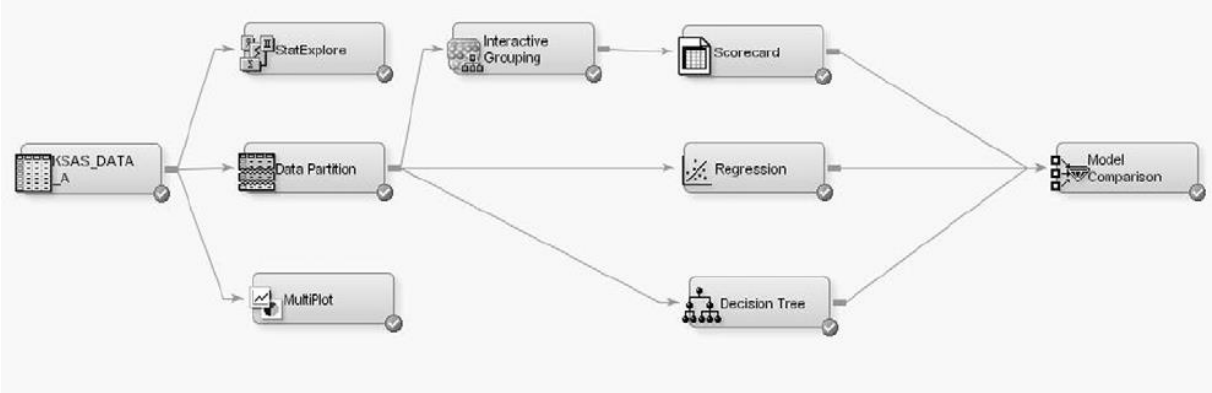
S obzirom na zaštitu privatnosti i nemogućnosti pribavljanja stvarnih financijskih podataka iz banaka, ovo istraživanje je bilo provedeno na temelju podataka o plaćanju članova jednog rekreacijskog kluba. Klub ima problem s povećanjem broja neplatiša, što se njihovih mjesečnih obveza tiče. Menadžment kluba bi volio imati model koji bi im pomogao prepoznavati potencijalne neplatiše. Diskusija istraživača s menadžmentom kluba je otkrila kako oni sami koriste osobne i subjektivne tehnike, pri procjenjivanju platiša i neplatiša. Kao i za ostale rekreacijske klubove, većina prihoda kluba dolazi od mjesečnih članarina. Visok broj neplatiša zato može rezultirati smanjenjem prihoda kluba, što će utjecati na financijsko planiranje klupskih aktivnosti, a menadžement se može suočiti i s bankrotom. Cilj je rada istraživača bio otkriti bonitet klijenata za predviđanje mogućih kašnjenja pri plaćanju obveza i za ranu intervenciju u tom slučaju, koja bi tako mogla smanjiti financijski gubitak. [9]

Istražitelji su odlučili koristiti model kartica bodova (eng. *Scorecard*), logističku regresiju i stablo odlučivanja. Kartice bodova su matematički model koji nastoji pružiti kvantitativnu procjenu vjerojatnosti o ponašanju klijenta. Na temelju tog modela, nastaju procjene o kreditiranju neke homogene populacije, a proučavaju se podaci klijenata koji su uzeli kredit i onih koji nisu uzeli kredit. Istražitelji ističu kako se generalno koriste linearna diskriminantna analiza i logistička regresija, kao dvije popularne statističke metode za izradu ovakvih modela. Ipak, s razvojem informacijske i računalne tehnologije, pojavljuju se nove tehnike pod imenom tehnike rudarenja podataka. Iako su metode za procjenu kreditnog rizika široko korištene u financijskim institucijama i bankama, mogu se koristiti i za ostale organizacije kao što su osiguranja, tržište nekretnina, telekomunikacijski i rekreativni klubovi, kako bi predviđale kašnjenja pri plaćanjima. [9]

Za ovo istraživanje su se koristile varijable: spol, dob, adresa, zanimanje, rasa, bračni status, broj uzdržavanih osoba, broj automobila, platiša/neplatiša, radni sektor. Za izradu modela, istraživači su koristili softver SAS Enterprise Miner 5.3, koji nudi tradicionalne metode, ali i nove prediktivne te klasifikacijske tehnike. Za izradu modela, moraju se definirati ulazna ili nezavisna varijabla te ciljna varijabla koja je zavisna, kao i nivo mjerenja. Ciljna varijabla je status plaćanja, to je binarna varijabla s dvije kategorije: platiša ili neplatiša. Neplatiše su oni koji nisu platili svoje mjesečne obveze tri uzastopna mjeseca. Sljedeće čime su se istraživači bavili je bila priprema podataka. Podatke, koji su bili teksualnog formata, su prebacili u Excel format te su čistili podatke na način da su odstranjivali ekstreme, redundantne podatke i nadomještali nedostajuće vrijednosti koristeći srednju vrijednost. Nakon toga su Excel format prebacili u SAS dokument, koristeći SAS Enterprise Guide. Sljedeći je korak bilo samo modeliranje. Podaci su se sastojali od 977 (35%)

neplatiša i 1 788 (65%) platiša. Zavisna varijabla, status plaćanja je imao vrijednost 1 za neplatiše i 0 za platiše. [9]

Korišteni alat se služi čvorovima, a oni formiraju dijagram koji strukturira i dokumentira tok analitičkih aktivnosti. Čvor uzorka podataka je potrebno spojiti sa čvorom koji dijeli podatke, kako bi se podaci podijelili na podatke za modeliranje i one za vrednovanje. Uzorak podataka se podijelio na 70% podataka za modeliranje i 30% podataka za vrednovanje. Nadalje su se sa čvorom za podjelu podataka povezali *scorecard* model, čvor logističke regresije i čvor stabla odlučivanja. Nakon toga je slijedila usporedba preciznosti modela, korištenjem čvora za usporedbu modela. Na temelju vrijednosti informacijske dobiti, izabrane su četiri varijable koje su imale vrijednost 0.1 ili više, jer se takve smatraju reprezentativnima. To su varijable: broj uzdržavanih osoba, broj automobila, adresa i dob, dok su ostale varijable bile odbačene. [9]



Slika 4 Dijagram toka (Izvor: Wah Yap, Huat Ong, Huain 2011)

Relativan rizik atributa se naziva njegovim težištem (eng. *Weight of Evidence - WOE*), a to je vrijednost koja ovisi o vrijednosti ciljne binarne varijable. U ovom slučaju, dobar atribut je platiša, a loš atribut neplatiša. Težište atributa je logaritam omjera proporcije dobrih atributa, spram proporcije loših atributa. Visoka negativna WOE vrijednost atributa odgovara visokom riziku, a visoka pozitivna vrijednost niskom riziku. U slučaju ovog istraživanja, težište atributa za dob, svrstava ljude mlađe od 32 godine u najrizičniji skupinu za neplaćanje, slijede osobe stare 32 – 37 godina, dok osobe iznad 53 godine predstavljaju najmanje rizičnu skupinu. Članovi klupa koji nemaju prijavljenu uzdržavanu osobu ili uzdržavaju samo jednu osobu, su najrizičniji za neplaćanje. Rezultati su još pokazali kako su ljudi koji posjeduju više automobila manje rizični za neplaćanje, dok veće izglede za neplaćanje imaju oni koji žive u Kuala Lumpuru. [9]

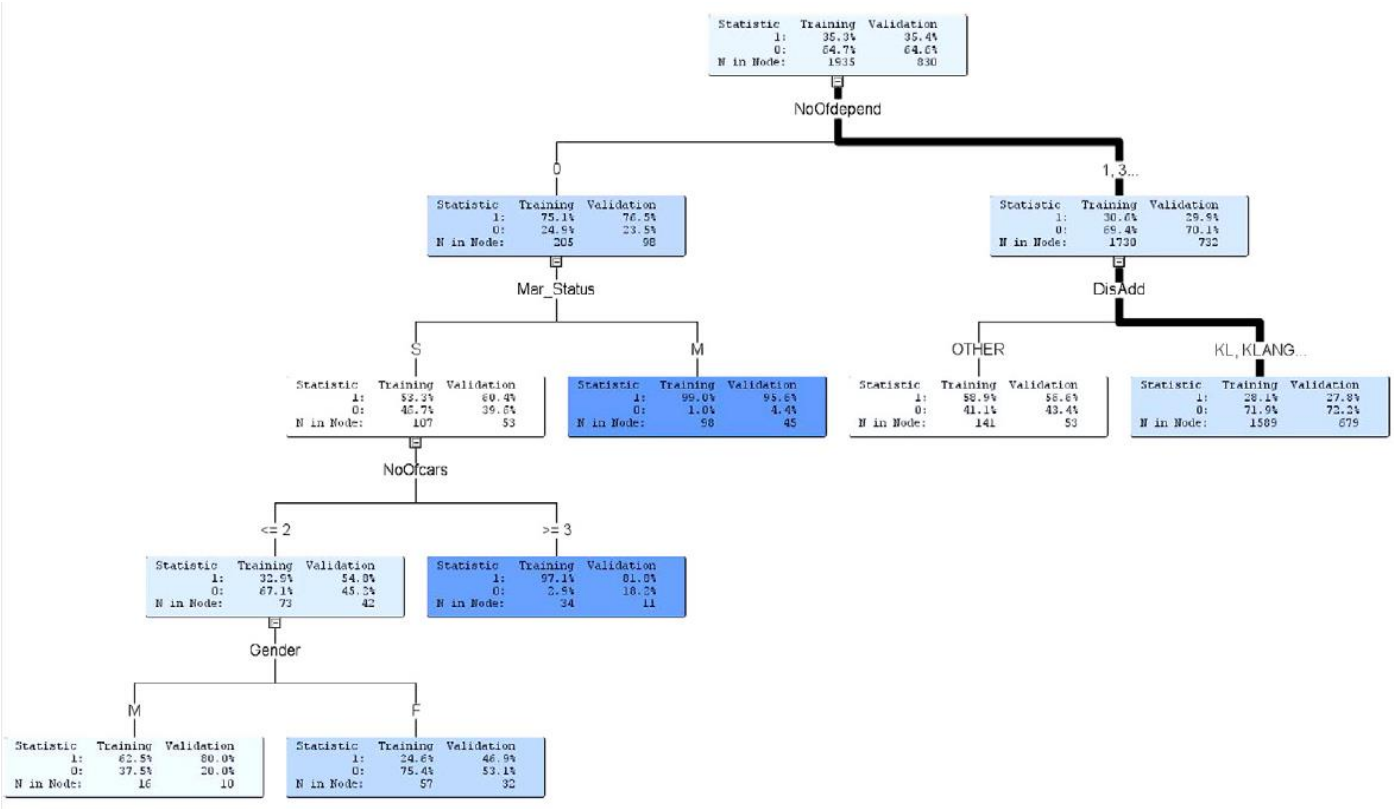
Što se interpetacije stabla odlučivanja tiče, istražitelji su izveli pravila navedena ispod.

- Ako su članovi oženjeni i broj uzdržavanih osoba je 0, imaju tendenciju ne plaćati.

- Također, tendenciju neplaćanja imaju osobe koje su kao adresu prijavile drugo od svega što je bilo ponuđeno, a broj uzdržavanih osoba je veći od 0.
- Ako je adresa člana u Kuala Lumpuru, Klangu ili Petalingu Jayiu i broj uzdržavanih osoba je veći od 0, onda članovi nemaju tendenciju ne plaćati.
- Ako članovi imaju više od 2 automobila, slobodni su i broj uzdržavanih osoba je 0, imaju tendenciju ne plaćati.
- Ako su članovi muškarci, imaju 1 ili 2 automobila, slobodni su i broj uzdržavanih osoba je 0, imaju tendenciju ne plaćati.
- Ako su članovi žene, imaju 1 ili 2 automobila, slobodne su i broj uzdržavanih osoba je 0, imaju manju tendenciju ne plaćati.

Prema napisanom, profil neplatiša je:

- oženjen(a) i bez djece,
- slobodan(a) s više od 2 automobila,
- muškarac, slobodan s 1 ili 2 automobila,
- članovi koji žive izvan Selangora. [9]



Slika 5 Model stabla odlučivanja (Izvor: Wah Yap, Huat Ong, Huain 2011)

Za uspoređivanje kvalitete modela, istražitelji su koristili pogrešku tipa I (klijent dobrog kreditnog rizika je pogrešno klasificiran kao klijent lošeg kreditnog rizika) i pogrešku tipa II (klijent lošeg kreditnog rizika je pogrešno klasificiran kao klijent dobrog kreditnog rizika). Troškovi pogrešne procjene su viši, ako se radi o pogrešci tipa II, nego o pogrešci tipa I. Rezultati istražitelja su pokazali kako je logistička regresija najosjetljivija metoda, i ima najmanji iznos za pogrešku tipa II. Stablo odlučivanja je ispao najgori model jer ima najveći iznos za pogrešku tipa II i najmanje je osjetljiva metoda. Za sve tri metode je zapravo nizak iznos pogreške tipa I, a visok iznos pogreške tipa II. [9]

Istražitelji zaključuju kako je ograničavajući faktor za izradu ovakvih modela, dostupnost podataka i uzorkovanje. Dobar model se ne može napraviti ako podaci nisu primjereni, nastaju pogreške pri bilježenju podataka ili ima previše nedostajućih vrijednosti. Zaključuju kako nema najbolje metode za izradu ovakvih modela jer ovise o strukturi podataka, kvaliteti podataka i cilju klasificiranja. No, usprkos razvoju novih tehnologija, logistička regresija i stablo odlučivanja su za istraživače najprihvatljivije tehnike, s obzirom na to kako je relativno lako identificirati važne ulazne varijable, interpetirati rezultate i implementirati sam model. [9]

3.3. Primjena hibridnog sustava za procjenu kreditnog rizika

Predmet istraživanja ovog rada je bilo ispitivanje podataka, u vlasništvu hrvatske banke, koji predstavljaju dobru bazu za predviđanje sposobnosti dužnika za vraćanje duga na vrijeme. Autori ističu važnost odabira tehnike, za pronalazak optimalnog podskupa značajki, koji povećava točnost klasifikacije. Istraživači su odlučili koristiti hibridni sistem sastavljen od genetskog algoritma, za koji su zaključili kako je kompetitivan i može se koristiti kao tehnika u budućnosti, da bi se otkrile važne značajke za utvrđivanje rizika kreditiranja klijenata. [10]

Kriza koja je započela u srpnju 2007. godine, potresla je financijsko tržište, potkopala potrošačko i ulagačko povjerenje, podigla je ozbiljnu zabrinutost i strah financijskih institucija oko stabilnosti, te je prijetila ekonomiji cijelog svijeta. Dok je uzroka za krizu bilo mnogo, autori smatraju kako su banke, vlade i ostale institucije mogle napraviti puno više, kako bi se ovakvi problemi spriječili, i tome svakako mogu pomoći razni prediktivni modeli. [10]

U članku se navodi kako je pojedinačni klasifikator prvi, kombinacija klasifikatora drugi, a ulazni podaci su treći važan faktor za studiju koja ispituje kreditni rizik. Također se ističe važnost selekcije varijabli, što je uvelike zanemareno u sličnim ispitivanjima, zbog otežanog pristupa podacima ovog tipa. Međutim, autori smatraju kako je to itekako važan i izazovan problem, kojeg treba riješiti prije samog istraživanja. Različite tehnike daju različite rezultate na istom skupu

podataka. Cilj istraživača je razviti hibridni sistem genetskog algoritma i neuronskih mreža (kraće: GA-NN), za postizanje optimalnog podskupa značajki pri procjenjivanu kreditnog rizika, koji povećava preciznost klasifikacije neuronskih mreža. Ispituju se različiti ulazni podaci, kako bi se otkrio njihov utjecaj na ispravnu klasifikaciju zahtjeva za kredite, s obzirom na kreditni rizik. Na preciznost predviđanja dobrih i loših klijenata, s obzirom na kreditni rizik, može se utjecati dobrim odabirom ulaznih podataka, korištenjem najboljih metoda klasifikacije i kombiniranjem rezultata različitih klasifikacijskih metoda. Prema istraživanju autora po prethodnim sličnim studijama, zaključeno je kako se uglavnom koristi 6 do maksimalno 81 nezavisna varijabla, povezane s demografijom i financijama, te da ne postoji studija koja obuhvaća sve tipove podataka klijenata. Autori su još zaključili kako su se u većini ostalih istraživanja neuronske mreže pokazale najtočnijima, najviše prilagodljivima i robustnima, što se tehnika ispitivanja tiče. [10]

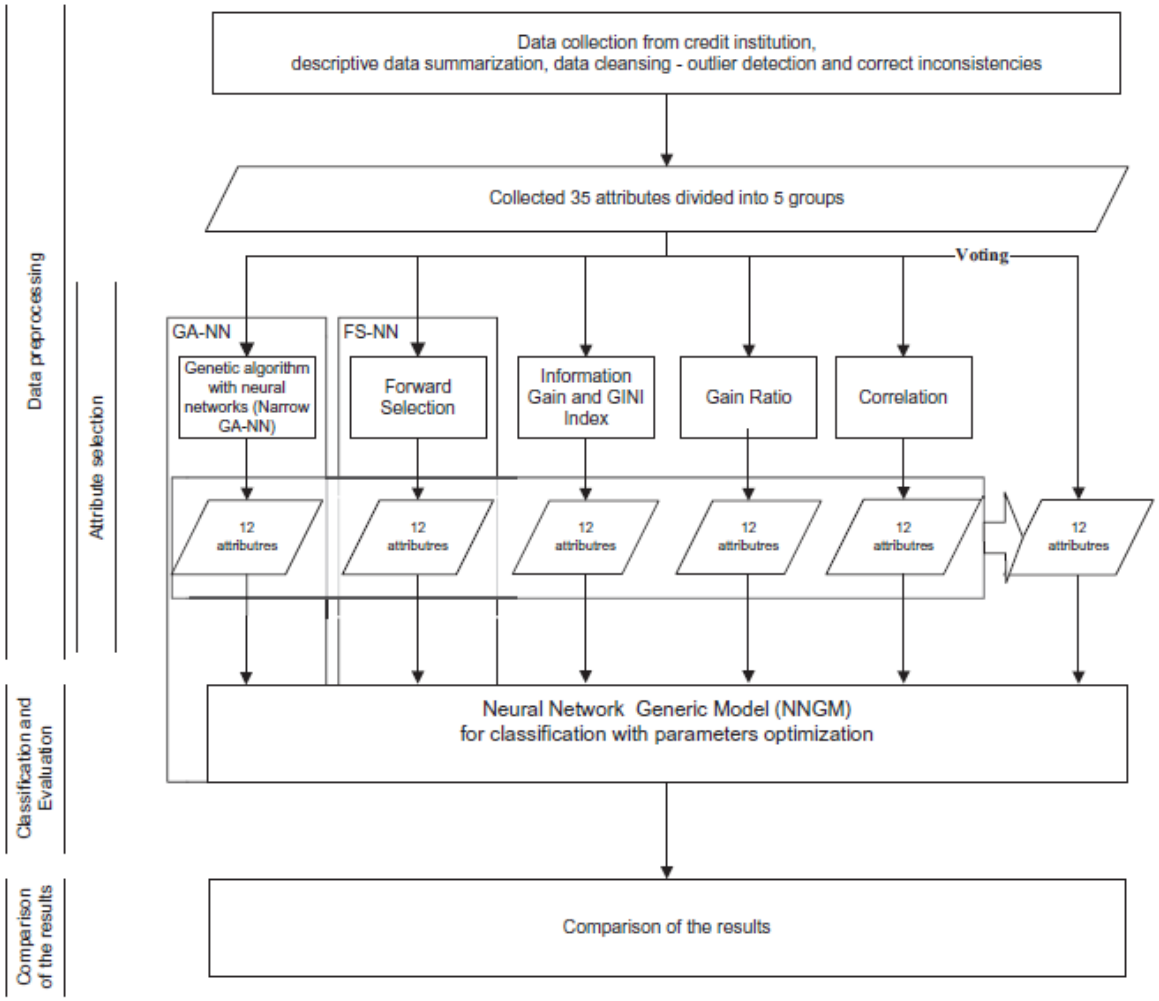
Kako bi provjerili jesu li podaci neke banke zaista dobra baza za predviđanje sposobnosti dužnika oko vraćanja duga, istraživači su postavili dvije hipoteze. [10]

H1: Iz postojećih podataka klijenata banke, može se odabrati takav skup podataka koji pruža dobru bazu za prognoziranje kreditnog rizika dužnika. Za skup podataka, baza koja se smatra dobrom za predviđanje kreditnog rizika dužnika, bit će ona koja temelji točnost predviđanja na nivou iznad 80%.

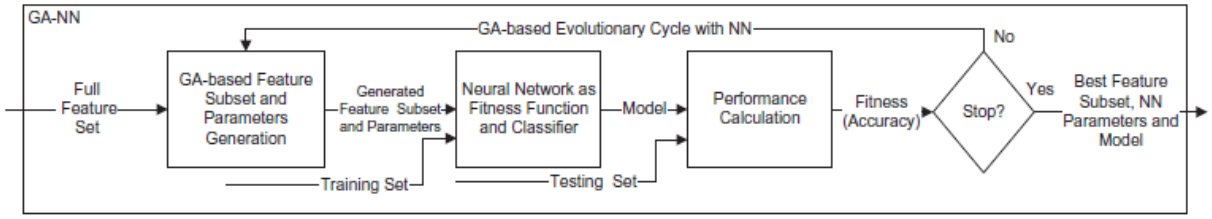
H2: Genetski algoritam i neuronske mreže, kao tehnika korištena u ovom ispitivanju, je statistički točnija 95%, što se pouzdanosti tiče, od ostalih metoda koje se često koriste za ovu svrhu, kao što su informacijska dobit, omjer dobiti, indeks dobiti, korelacija. [10]

Prema autorima, razvoj modela se sastoji od pripreme podataka koja uključuje selekciju atributa, klasifikaciju i procjenjivanje te usporedu rezultata. Nakon što su prikupili podatke, napravljeno je njihovo sažimanje. Sažimanje je analitička priprema podataka za njihovu predobradu. Bazične statističke mjere za sažimanje podataka uključuju standardnu devijaciju i rangiranje, kao korisne vrijednosti za mjerenje raspršenja podataka. Slijedi čišćenje podataka koje podrazumijeva nadopunjavanje nedostajućih vrijednosti, identificiranje stršećih vrijednosti i ispravljanje nekonzistentnosti podataka. Čišćenje podataka je iterativni proces koji se sastoji od detekcije raskoraka i transformacije podataka. Preobrada podataka je uključivala selekciju atributa, kako bi se utjecalo na efikasnost klasifikacijskog sistema, s aspekta preciznosti, brzine i skalabilnosti. Optimalni skup značajki ne mora biti jedinstven, jer je moguće postići preciznost i korištenjem drugačijeg skupa. Nakon tog koraka, slijedi klasifikacija i evaluacija podataka koji su prikupljeni iz hrvatske banke, u periodu od rujna 2004. godine do rujna 2011. godine. Od 32 000 potencijalnih podnositelja zahtjeva, ispitivači su odabrali klijente koji su tražili kredite iznosa manjeg ili jednakog 100 000 kuna, te one koji su imali otvoreni račun u banci najmanje 15 mjeseci

prije datuma odobravanja kredita. Iz tog skupa je nadalje nasumično odabrano 1 000 slučajeva, uključujući njih 750 koji su na vrijeme otplaćivali svoje obveze i 250 njih koji su kasnili s otplaćivanjem. Inicijalno je svaki klijent bio opisan s 37 varijabli, no nakon što se otkrilo kako neke varijable imaju identične vrijednosti u svim slučajevima i ekstremno visoku korelaciju, finalni broj varijabli je sveden na 33 regularne i 2 specijalizirane (id, oznaka). Naime, korelacija je mjera linearne ovisnosti između objekata. Kako bi se izračunala korelacija, standardiziraju se podatkovni objekti i računa se njihov vektorski produkt. Nadalje, autori su varijable podijelili u 5 glavnih grupa: bazične karakteristike, povijest plaćanja (mjesečni prosjeci), financijski uvjeti, delinkvencijska prošlost i prethodna iskustva s kreditima. Na podacima se koristilo sedam selekcijskih tehnika: genetski algoritam s neuronskim mrežama, unaprijeđena selekcija s neuronskim mrežama, informacijska dobit, omjer dobitka, indeks dobitka, korelacija i metoda glasanjem. Omjer dobiti je omjer informacijske dobiti i entropije atributa te se njegovim korištenjem izbjegava favoriziranje atributa s više vrijednosti. Indeks dobiti pak odražava vjerojatnost ako dvije slučajno odabrane instance, pripadaju različitoj klasi. Korištenjem tehnika je odabrano 12 značajki, jer je procijenjena preciznost padala nakon reduciranja ispod brojke 12. Nakon što su odabrane značajke za svaku tehniku, one koje se pojavljuju u više od pola metoda, jesu one koje se odabiru za metodu glasanjem. Kako bi se procijenila efikasnost tehnika, istraživači su uspoređivali tehničku preciznost i trošak klasifikacije, kako bi se pronašao model koji maksimalno reducira troškove za banke. [10]



Slika 7 Dijagram rada za procjenu kreditnog rizika (Izvor: Oreški, Oreški, Oreški 2012)



Slika 6 Dijagram rada za GA-NN tehniku (Izvor: Oreški, Oreški, Oreški 2012)

Na temelju ispitivanja, za svaku usporedbu kada su promijenjeni uvjeti i u kojima se tražilo najbolje predstavljanje od izabranih atributa, jedino je GA-NN tehnika ponudila mogućnost dobivanja novog skupa značajki, koji bi najbolje odgovarao zahtjevima optimizacije. Nedostatak ove tehnike je pak dugo vrijeme obrade jer se radi o računalno zahtjevnoj tehnici. Ipak, zaključak

autora kako je GA-NN model značajno bolji izbor za izradu modela klasifikacije, u usporedbi s ostalim provedenim metodama, a to povrđuje hipotezu H2. Također je dokazano kako se na temelju podataka s kojima banka raspolaže, može napraviti klasifikacija klijenta temeljem rizičnosti njihova kreditiranja, s maksimalnom točnosti iznad 80%, što dokazuje hipotezu H1. Savjet je autora da daljnje studije nastavljaju uspoređivati ostale metode klasifikacije na ovakvom skupu podataka, ali i na proširenom skupu podataka, koji su na primjer prikupljeni od strane ureda za kreditiranje. [10]

4. Opis metodologije

Stablo odlučivanja je moderna i popularna tehnika za klasifikacijske i predikcijske probleme. Jednostavnost primjene stabla odlučivanja je što se model podataka može pročitati u obliku pravila. Ta se pravila mogu direktno interpretirati ili se mogu koristiti u nekom od programskih jezika za rad s bazama podataka, pa se određeni primjeri iz baze mogu izdvojiti korištenjem pravila koja su generirana iz stabla odlučivanja. [11]

Postoji čitav niz različitih algoritama koji se koriste za konstruiranje stabla odlučivanja, a napoznatiji i najčešće korišten algoritam je C4.5, odnosno njegova poboljšana komercijalna verzija See5/C5.0. Na svakom čvoru, C4.5 odabire jedan atribut, koji najučinkovitije razdvaja skup podataka na podskupove, koji se pridodaju jednoj ili drugoj klasi. Kriterij je normalizirana vrijednost informacijske dobiti, što dolazi od izbora nekog atributa za razdvajanje podataka. Atribut s najvećom informacijskom dobiti se odabire kao onaj koji donosi odluku. Odnosno, sam algoritam ima nekoliko temeljnih slučajeva navedenih ispod.

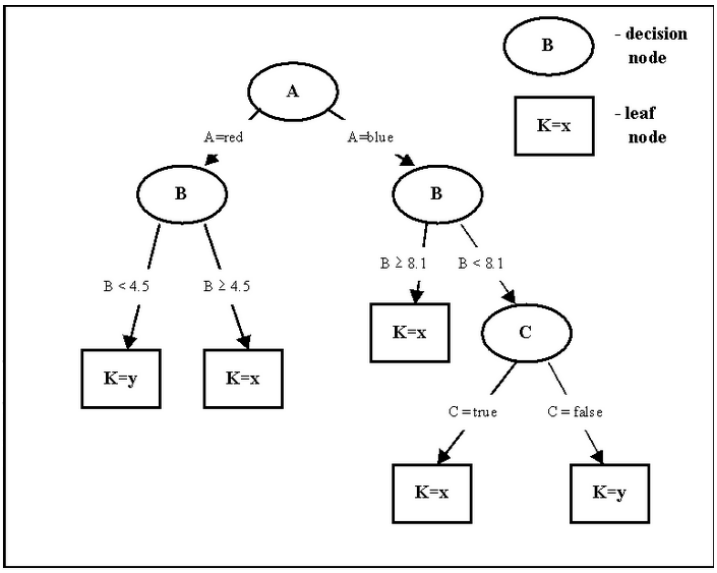
- Svi uzorci u listu pripadaju istoj klasi. Tada se stvara stablo lista za stablo odlučivanja, koje kaže da se odabere ta klasa.
- Niti jedno od svojstava ne omogućava informacijski dobit. U tom slučaju, algoritam kreira viši čvor odlučivanja korištenjem očekivane vrijednosti klase.
- Pojavljuje se slučaj klase koja se dosad pojavljivala. C4.5 kreira viši čvor odlučivanja na stablu korištenjem očekivane vrijednosti.

Još jedan primjer je CART algoritam koji gradi binarno stablo odlučivanja, koje sadrži dvije grane za svaki čvor odlučivanja i nastavlja dijeljenje, sve dok pronalazi nove dijelove koji povećavaju sposobnost podjele podataka u kategorije. Algoritam razmatra kako napraviti svaku podjelu, odlučuje kada je čvor stabla vrh i kako dodijeliti određenu kategoriju svakom završnom čvoru. [11]

Dakle, stablo odlučivanja je klasifikacijski algoritam u kojemu se razlikuju dva tipa čvora povezani granama. Krajnji čvor (eng. *leaf node*) jest čvor kojim završava određena grana stabla. Krajnji čvorovi definiraju klasu kojoj pripadaju primjeri, koji zadovoljavaju uvjete na toj grani stabla. Čvor odluke (eng. *decision node*) jest čvor koji definira određeni uvjet u obliku vrijednosti određenog atributa, iz kojeg izlaze grane koje zadovoljavaju određene vrijednosti tog atributa. Stablo odlučivanja se može koristiti za klasifikaciju primjera, tako da se krene od prvog čvora odlučivanja u korijena stabla pa se kreće po onim granama stabla, koje primjer sa svojim vrijednostima zadovoljavaju sve do krajnjeg čvora, koji klasificira primjer u jednu od postojećih klasa problema. [11]

Osnovni preduvjeti za korištenje tehnike stabla odlučivanja su:

- opis u obliku parova vrijednost – atribut, podaci o primjeru moraju biti opisani u obliku konačnog broja atributa;
- prethodno definiran konačan broj klasa (vrijednost ciljnog atributa), kategorije kojima pripadaju primjeri moraju biti definirane i mora ih biti konačan broj;
- klase moraju biti diskretne, svaki primjer mora pripadati jednoj od postojećih klasa, kojih mora biti znatno manje od broja primjera;
- značaj broj primjera, poželjno je da u skupu primjera za generiranje stabla odlučivanja postojim barem nekoliko stotina primjera. [11]



Slika 8 Primjer stabla odlučivanja (Izvor: Hamilton, Gurak, Findlater i Olive, 2001)

Većina postojećih algoritama odlučivanja jesu varijacije osnovnog algoritma, koji se zove ID3, a razvio ga je John Ross Quinlan. Taj algoritam je prethodnik C4.5 algoritma,a funkcionira tako da pretražuje preko atributa svih primjera u skupu podataka, te pronalazi atribut koji najbolje odvaja određene klase. Ukoliko atribut savršeno razdvaja klase, ID3 algoritam se zaustavlja, dok se inače rekurzivno izvršava na $\boldsymbol{m}$ podskupova, $\boldsymbol{m}$ je broj mogućih vrijednosti atributa, tražeći najbolje atribute za njihovo razdvajanje. Moguće je da generira i stabla koja rade pogrešne klasifikacije na skupu primjera za učenje. Središnji dio algoritma je selekcija atributa za stvaranje čvora odlučivanja, odnosno atributa koji će poslužiti za razdvajanje određene grane stabla. Za selekciju atributa najmješovitije strukture vrijednosti ciljnog atributa, algoritam koristi koncept entropije. Kriterij kvalitete u algoritmu stabla odlučivanja je vezan uz selekciju atributa, koji će

poslužiti kao kriterij za razdvajanje primjera u određenom čvoru odlučivanja stabla. Cilj je odabrati atribut koji je najupotrebljiviji s obzirom na osnovni cilj, a to je klasifikacija primjera. [11]

Dobra kvantitativna mjera vrijednosti atributa jest statistička vrijednost nazvana informacijska dobit (eng. *information gain*), kojom se mjeri koliko dobro atribut razdvaja primjere prema njhovoj klasifikaciji. Ova se mjera koristi kako bi se odabrao najbolji kandidat od atributa, za svaki novi korak prilikom stvaranja stabla odlučivanja. Kako bi se precizno definirala informacijska dobit, definira se mjera koja se naziva entropija i predstavlja mjeru nehomogenosti skupa primjera iz baze podataka, to jest entropija je mjera nereda sistema. Na primjer, ako je zadan skup S i sadrži dvije klase, pozitivne i negativne primjere, entropija takve binarne klasifikacije je definirana sljedećim izrazom:

$\boldsymbol{S} = -\boldsymbol{p_p log_2 p_p}\ \boldsymbol{p_n log_2 p_n}$,

gdje $\boldsymbol{p_p}$ označava postotak pozitivnih primjera u $\boldsymbol{S}$, a $\boldsymbol{p_n}$ postotak negativnih primjera u skupu $\boldsymbol{S}$. Ukoliko ciljni atribut poprima više od dvije vrijednosti, tada je entropija u skupu $\boldsymbol{S}$, u odnosu na klasifikaciju, definirana s:

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

gdje je $\boldsymbol{p_i}$ postotak klase u skupu $\boldsymbol{S}$. Ako ciljni atribut poprima $\boldsymbol{c}$ različitih vrijednosti, maksimalna entropija iznosi $\boldsymbol{log_2 c}$.

Mjera efektivnosti atributa u klasificiranju primjera je dakle 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:

$(\boldsymbol{S}, \boldsymbol{A}) = \boldsymbol{S} - \sum_{v\ \varepsilon\ \boldsymbol{Values(A)}} \frac{|s_v|}{|s|} (\boldsymbol{s_v})$,

gdje je $\boldsymbol{Values}(\boldsymbol{A})$ skup svih mogućih vrijednosti atributa $\boldsymbol{A}$. Prvi član u jednadžbi za informacijsku dobit je entropija originalnog skupa $\boldsymbol{S}$, dok je drugi član očekivana vrijednost entropije, nakon što je skup $\boldsymbol{S}$ razdvojen korištenjem atributa $\boldsymbol{A}$. Informacijska dobit $(\boldsymbol{S}, \boldsymbol{A})$, je dakle očekivana redukcija entropije uzrokovana poznavanjem vrijednosti atributa $\boldsymbol{A}$, odnosno to je informacija o vrijednosti ciljnog atributa, uz poznate vrijednosti atributa $\boldsymbol{A}$. 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 je 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. [11]

Dobar izbor atributa je zapravo izbor onog atributa koji će dati najmanje stablo. Informacijska dobit se povećava sa srednjom čistoćom podskupova, koju proizvodi neki atribut. Strategija je izabrati atribut koji rezultira s najvećim informacijskom dobiti. S danom vjerojatnošću

distribucije, informacija koja je potrebna kako bi se predvidio neki događaj je entropija distribucije. Informacija se mjeri u bitovima, a informacijska dobit je konkretno razlika informacija prije razdvajanja i informacija poslije razdvajanja. [11]

Proces odabira novog atributa i razdvajanje primjera se ponavlja za svaki čvor odlučivanja, a koriste se samo oni primjeri koji pripadaju tom čvoru. Pri tome su svi atributi korišteni prije tog čvora u istoj grani stabla, isključeni iz daljnjeg odabira, što znači da se mogu pojaviti samo jednom na određenoj grani stabla. Ovaj se proces nastavlja sve dok na određenom čvoru nije zadovoljen jedan od dva kriterija:

1. svi atributi su već bili korišteni u toj grani stabla, ili
2. svi primjeri koji pripadaju tom čvoru imaju istu klasu, dakle radi se o krajnjem čvoru grane (entropije primjera je jednaka 0). [11]

Jedna od poteškoća stabla odlučivanja jer *over-fitting*, kada se smatra da generirano stablo pretjerano dobro aproksimira odnose u podacima. Rješenje je zaustaviti proces rasta stabla, prije nego se postigne savršena klasifikacija primjera iz skupa podataka za učenje ili se najprije generira stablo koje savršeno klasificira primjere, a zatim se to stablo reducira „skraćivanjem“ određenih grana, prema prethodno definiranom kriteriju. Drugi pristup se u praksi prikazao pouzadnijim jer je teško unaprijed definirati željenu kompleksnot stabla. Također je pitanje kako odrediti optimalnu kompleksnost, odnosno veličinu stabla za konkretni problem? Moguća su sljedeća rješenja:

1. korištenje posebnog skupa primjera (validacijski skup), koji je različit od onog korištenog za generiranje stabla, kako bi se ocijenila uspješnost reduciranja stabla;
2. korištenje posebnog statističkog testa na čvorovima koji su kandidati za „skraćivanje“, kojima se pokazuje hoće li se izbacivanjem tog čvora postići poboljšanje;
3. korištenje eksplicitne mjere kompleksnosti kodiranja primjera stablom odlučivanja, koja zaustavlja rast stabla kada je taj kriterij zadovoljen. [11]

Najčešće korišten je prvi naveden pristup, gdje se primjeri dijele u dva skupa: skup za učenje koji se koristi za generiranje stabla i skup za provjeru, koji se koristi za provjeru učinkovitosti metode reduciranja stabla. [11]

Zaključno, prednosti stabla odlučivanja su :

- jednostavna priprema podataka,
- generiranje razumljivih i jednostavnih modela,
- relativno skromni zahtjevi računalnih resursa (vrijeme i memorija),
- sposobnost korištenja numeričkih i kategorijskih tipova atributa,
- jasno se odražava važnost pojedinih atributa za konkretni klasifikacijski, odnosno predikcijski problem,
- modeli se mogu vrednovati statističkim tehnikama, robustni su i brzi jer se brzo izvršavaju s velikom količinom podataka u kratkom vremenu. [11]

Nedostaci su pak sljedeći:

- stabla odlučivanja su manje prikladna za probleme kod kojih se traži predikcija kontinuiranih vrijednosti ciljnog atributa,
- skloni su pogreškama pri radu s više klasa i relativno malim brojem primjera za učenje,
- ponekad mogu biti računalno zahtjevni, na primjer ako je potrebna redukcija stabla, jer se tada mora generirati velik broj stabala kako bi se odabralo najbolje stablo za klasifikaciju određenog problema,
- nisu dobro rješenje za klasifikacijske probleme kod kojih su regije određenih klasa omeđene nelinearnim krivuljama. [11]

5. Opis istraživanja

Nakon što sam opisala područje rudarenja podataka i povezala ga sa samom temom rada, proučila slična istraživanja i korištene tehnike, objasnila sam kako detaljno funkcionira metoda stablo odlučivanja. Naime, uzimajući u obzir prednosti i nedostatke obrađenih metoda rudarenja podataka, i u dogovoru s mentoricom, za potrebe istraživanja ovog rada, uzeta je tehnika stabla odlučivanja za ispitivanje podataka. To je tehnika koja je primjenjiva za potrebe klasificiranja, predviđanja, procjenjivanja, klasteriranja, opisivanja i vizualizaciju podataka. Dakle, uz široku primjenjivost, prednost ove metoda je još što je je jednostavna i razumljiva, a rezultati ispitivanja se lako interpretiraju. Ispitivanje će se provoditi na skupu podataka numeričkog i kategorijskog tipa, koji pripada njemačkoj banci.

U praksi podaci koji se koriste za istraživanje, nikada nisu idealni pa je potrebno uložiti vrijeme i energiju u njihovu pripremu. U selekciji se odabiru podaci koji su najkorisniji za proces ispitivanja, na primjer, samo neke varijable ili samo neki primjeri, ili se uzima samo reprezentativan uzorak. Selekcija može značiti izbor podskupova ili atributa. U predobradi se podaci dovode u prikladniji oblik, „čiste“ se netočni podaci, ispravljaju se neispravni podaci i nadomještavaju se podaci koji nedostaju. Glavni cilj predobrade podataka je dobiti podatke visoke kvalitete metodama čišćenje podataka, integracije podataka i redukcije podataka. Procesom transformacije, podaci se mogu transformirati u prikladniji oblik, na primjer iz više tablica se može napraviti jedna tablica jer ih samo tako mogu obraditi algoritmi koji se koriste. Podatke je moguće transformirati normalizacijom podataka, pretvaranjem atributa iz jednog tipa u drugi i slično. Nadalje, obrada cjelokupnog skupa podataka može biti preskupa ili predugotrajna tako da je glavna tehnika koja se koristi za izbor podataka uzorkovanje. Ako je uzorak reprezentativan, rezultati će biti jednako kvalitetni kao da se radilo o cjelokupnom skupu podataka. Dakle, istraživanje podataka uključuje opisivanje i transformaciju podataka, a ispitivanje kvalitete podataka identificira i rješava probleme kvalitete, kao što su nedostajuće vrijednosti, vrijednosti koje strše i čišćenje podataka. Sve su to važne metode za pripremu podataka, koja je ključni korak za uspješno ispitivanje, tako da sam tom dijelu posvetila veliku pažnju.

Baza podataka kojom raspolažem sadrži 1 000 instanci i 20 vrsta atributa (7 numeričkih, 13 kategorijskih).

1. atribut (kategorijski): iznos na postojećem tekućem računu

   A11: < 0 DM
   A12: >= 0 ... < 200 DM
   A13: >= 200 DM / primanje plaće barem godinu dana
   A14: bez tekućeg računa

2. atribut (numerički): trajanje u mjesecima

3. atribut (kategorijski): povijest kreditiranja

   A30: nije bilo kredita / svi krediti su vraćeni propisno
   A31: svi krediti ovoj banci su vraćeni propisno
   A32: postojeći krediti su otplaćivani propisno
   A33: kašnjenje s plaćanjima u prošlosti
   A34: račun je kritično / ostali postojeći krediti (ne u ovoj banci)

4. atribut (kategorijski): namjena

   A40: novi automobil
   A41: korišten automobil
   A42: namještaj/oprema
   A43: radio/televizija
   A44: kućanski aparati
   A45: popravci
   A46: edukacija
   A47: odmor – postoji li?!
   A48: treniranje
   A49: posao
   A410: ostalo

5. atribut (numerički): iznos kredita

6. atribut (kategorijski): štedni račun / obveznice

   A61: < 100 DM
   A62: <= 100 ... < 500 DM
   A63: <= 500 ... < 100 DM
   A64: >= 1 000 DM
   A65: nepoznato / nepostojeći štedni račun

7. atribut (kategorijski): trenutno zaposlenje od

   A71: nezaposlen(a)
   A72: < 1 godine
   A73: >= 1 ... < 4 godine
   A74: >= 4 ... < 7 godina
   A75: >= 7 godina

8. atribut (numerički): stopa rate u postotku s obzirom na raspoloživi dohodak

9. atribut (kategorijski): spol i bračni status

   A91: muškarac, razveden / razdvojen
   A92: žena, razvedena / razdvojena / udana
   A93: muškarac, slobodan
   A94: muškarac, oženjen / udovac
   A95: žena, slobodna

10. atribut (kategorijski): ostali dužnici / jamci

    A101: nitko
    A102: supotpisnik
    A103: jamac

11. atribut (numerički): prijavljeno prebivalište od

12. atribut (kategorijski): vlasništvo

    A121: nekretnina
    A122: ako nije A121 – stambena štedionica / životno osiguranje
    A123: ako nije A121 / A122 – automobil ili drugo, što nije u atributu broj 6
    A124: nepoznato / nema vlasništva

13. atribut (numerički): godine

14. atribut (kategorijski): ostali obročni planovi

    A141: banka
    A142: trgovine
    A143: ništa

15. atribut (kategorijski): stanovanje

    A151: najam
    A152: vlastito
    A153: besplatno

16. atribut (numerički): broj postojećih kredita u ovoj banci

17. atribut (kategorijski): posao

    A171: nezaposlen / nekvalificiran – nerezident
    A172: nekvalificiran – rezident
    A173: kvalificiran zasposlenik / službeno
    A174: menadžment / samozaposlen / visokokvalificiran zaposlenik / službenik

18. atribut (numerički): broj odgovornih osoba za uzdržavanje

19. atribut (kategorijski): telefon

    A191: nema

A192: da, registriran na ime klijenta

20. atribut (kategorijski): inozemni radnik
    A201: da
    A202: ne

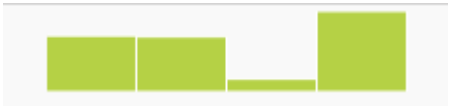
21. atribut (numerički): klasifikacija klijenta
    1 = dobar klijent
    2 = loš klijent

Podaci ove baze podataka su bili dostupni u tekstualnom formatu, te sam ih prilagodila i pretvorila u Excel format. Prilikom prebacivanja podataka u prikladan format za daljni rad, došlo je do poremećaja u vrijednostima nekih redaka, što sam ručno pregledavala i ispravljala. Prilikom ispravljanja i proučavanja podataka, uočila sam kako nema stršećih ili nedostajućih vrijednosti. Sve atribute smatram prikladnim i korisnim za provedbu ispitivanja, tako da je priprema podataka završila.

Kako bih što preciznije opisala skup podataka s kojim ću raditi, odnosno atribute koji će se koristiti za istraživanje, pripremljenu bazu podataka sam povezala s alatom BigML pa ću prikazati slike koje odgovaraju svakom atributu te ih opisati. BigML je platforma namijenjena istraživanju podataka i pronalaženju korisnih informacija u podacima. Nastala je iz želje da se strojno učenje približi što većoj populaciji ljudi, te da se kroz učenje i istraživanje uživa. Opisivanje podataka ću započeti tabularni prikazom za kategorijske atribute, koji će sadržavati karakteristike svakog od atributa, izražene standardnim statističkim mjerama. Nakon toga slijede slike grafova za svaku vrstu atributa, gdje histogram prikazuje bilo koji numerički atribut, a stupičasti graf bilo koji kategorijski atribut.

Tablica 1 Opis numeričkih atributa

	Minimum	Aritmetička sredina	Medijan	Maksimum	Standardna devijacija	Koeficijent spljoštenosti	Koeficijent asimetrije
Trajanje u mjesecima	4.00	20.90	18.00	72.00	12.06	0.91	1.09
Iznos kredita	250.000	3 254.41	2 309.00	18 424.00	2 808.03	4.35	1.96
Stopa rate u postotku s obzirom na raspoloživi dohodak	1.00	2.97	3.00	4.00	1.12	-1.21	-0.53
Prijavljeno prebivalište od	1.00	2.85	3.00	5.00	1.10	-1.37	-0.27
Godine	19.00	35.55	33.00	75.00	11.38	0.59	1.02
Broj postojećih kredita u ovoj banci	1.00	1.41	1.00	4.00	0.58	1.58	1.27
Broj odgovornih osoba za uzdržavanja	1.00	1.16	1.00	2.00	0.36	1.64	1.91
Klasifikacija klijenta	1.00	1.30	1.00	2.00	0.46	-1.24	0.87



Slika 9 Iznos na postojećem tekućem računu, kategorijski atribut – multimodalna distribucija

Slika 10 Trajanje u mjesecima, numerički atribut – uniformna distribucija iskrivljena udesno

Slika 9 Povijest kreditiranja, kategorijski atribut – multimodalna distribucija

Slika 10 Namjena, kategorijski atribut – multimodalna distribucija

Slika 13 Iznos kredita, numerički atribut – uniformna distribucija iskrivljena udesno

Slika 14 Štedni račun/obveznice, kategorijski atribut – multimodalna distribucija

Slika 11 Trenutno zaposlenje od, kategorijski atribut – uniformna distribucija

Slika 12 Stopa rate u postotku s obzirom na raspoloživi dohodak, numerički atribut – uniformna distribucija

Slika 13 Spol i bračni status, kategorijski atribut – unimodalna distribucija

Slika 14 Ostali dužnici/jamci, kategorijski atribut – uniformna distribucija iskrivljena udesno

Slika 15 Prijavljeno prebivalište od, numerički atribut – uniformna distribucija

Slika 16 Vlasništvo, kategorijski atribut – uniformna distribucija

Slika 17 Godine, numerički atribut – uniformna distribucija iskrivljena udesno

Slika 18 Ostali obročni planovi, kategorijski atribut – multimodalna distribucija

Slika 19 Stanovanje, kategorijski atribut – multimodalna distribucija

Slika 20 Broj postojećih kredita u ovoj banci, numerički atribut – eksponencijalna distribucija

Slika 21 : Posao, kategorijski atribut – multimodalna distribucija

Slika 22 Broj odgovornih osoba za uzdržavanje, numerički atribut – uniformna distribucija iskrivljena udesno

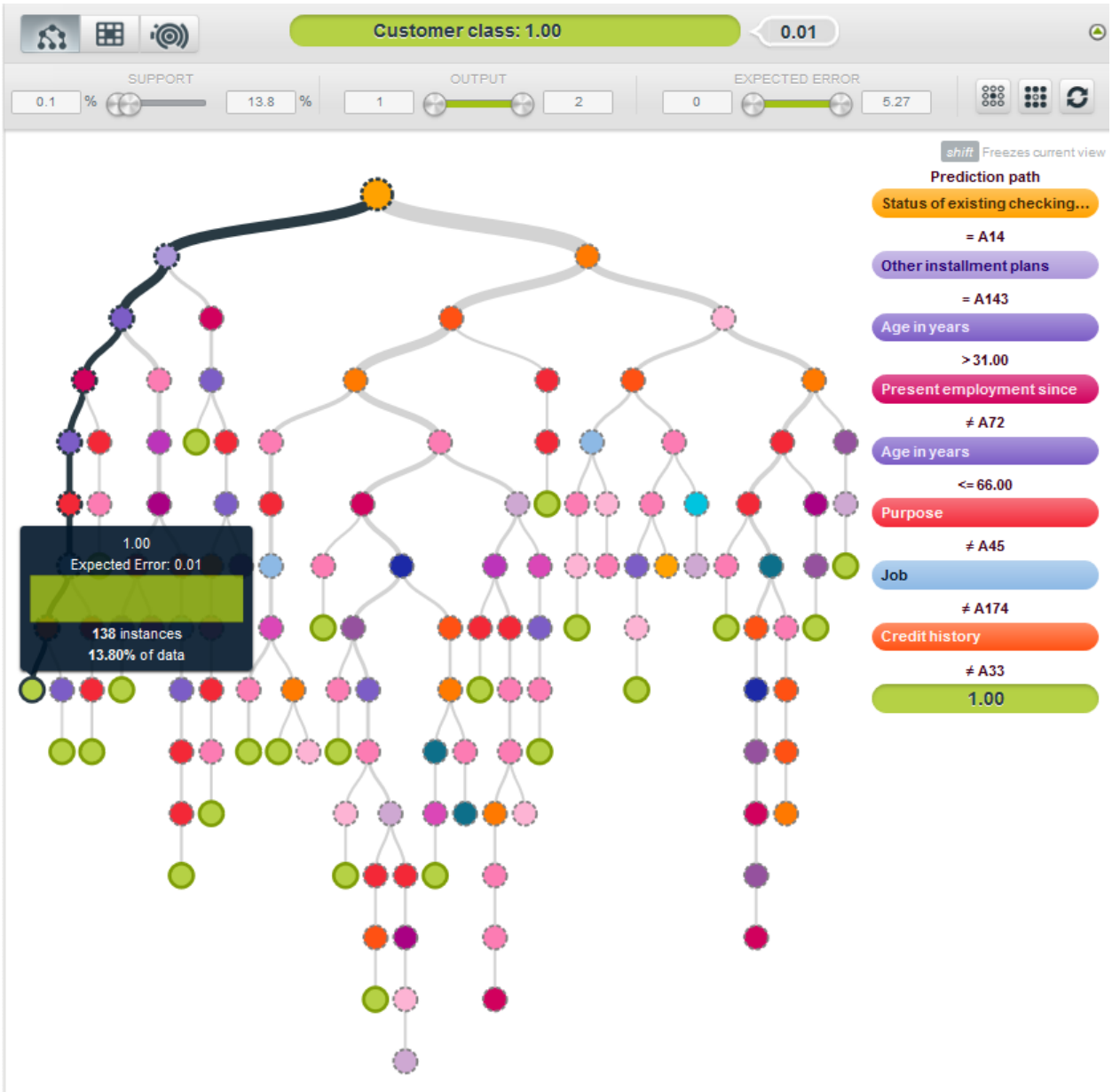
Slika 23 Telefon, kategorijski atribut – uniformna distribucija iskrivljena udesno

Slika 24 Inozemni radnik, kategorijski atribut – uniformna distribucija iskrivljena udesno

Slika 25 Klasifikacija klijenta, numerički atribut – uniformna distribucija iskrivljena

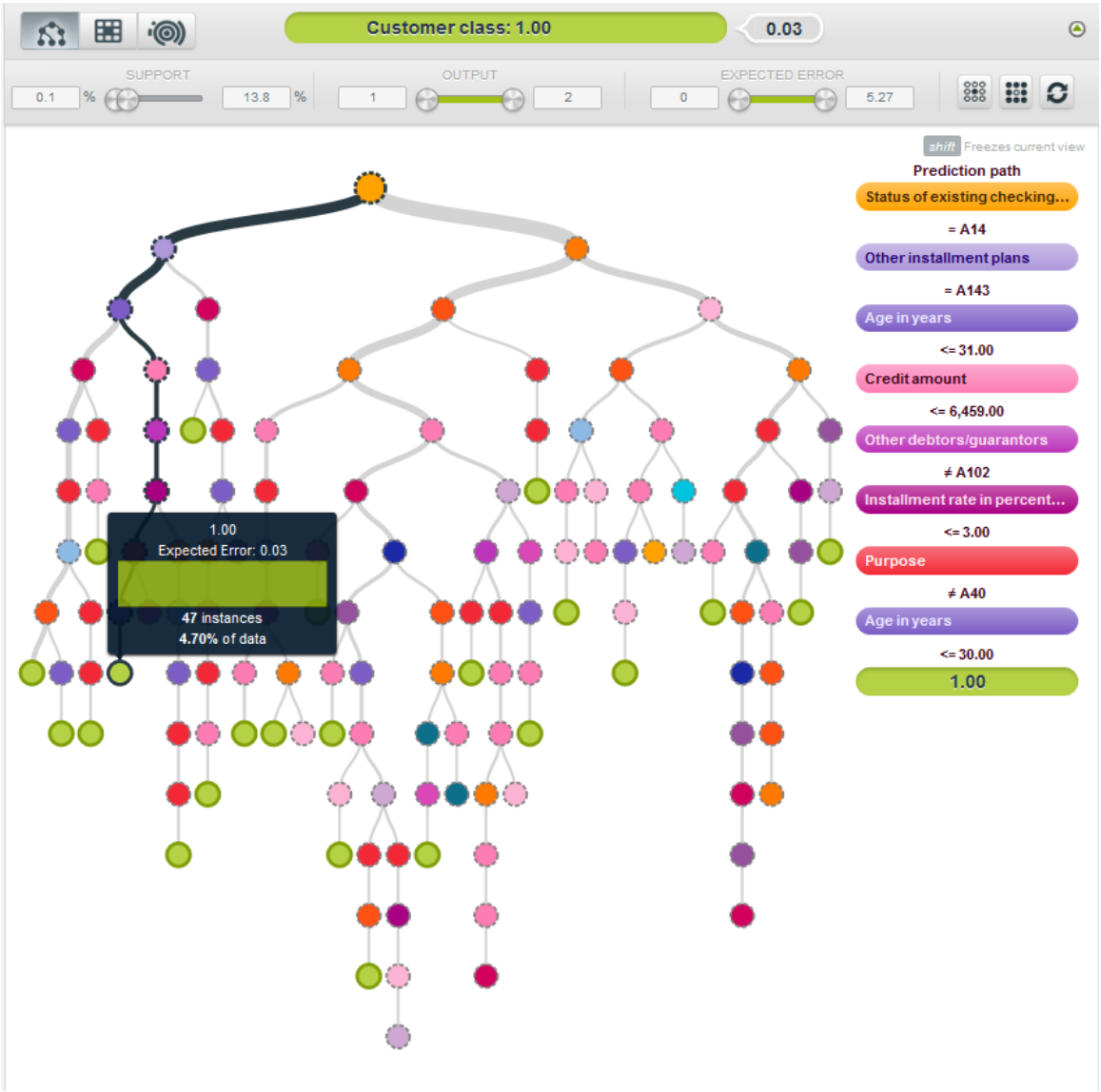
6. Rezultati istraživanja

Dakle, za provedbu samog istraživanja i analiziranje podataka, koristit ću metodu stablo odlučivanja, koje ću izraditi u alatu BigML na temelju unesene baze podataka i skupa atributa, koje sam prethodno pripremila i opisala. Prikazat ću slike samog stabla, koje ću opisati s skladu s izvedenim poslovnim pravilima.



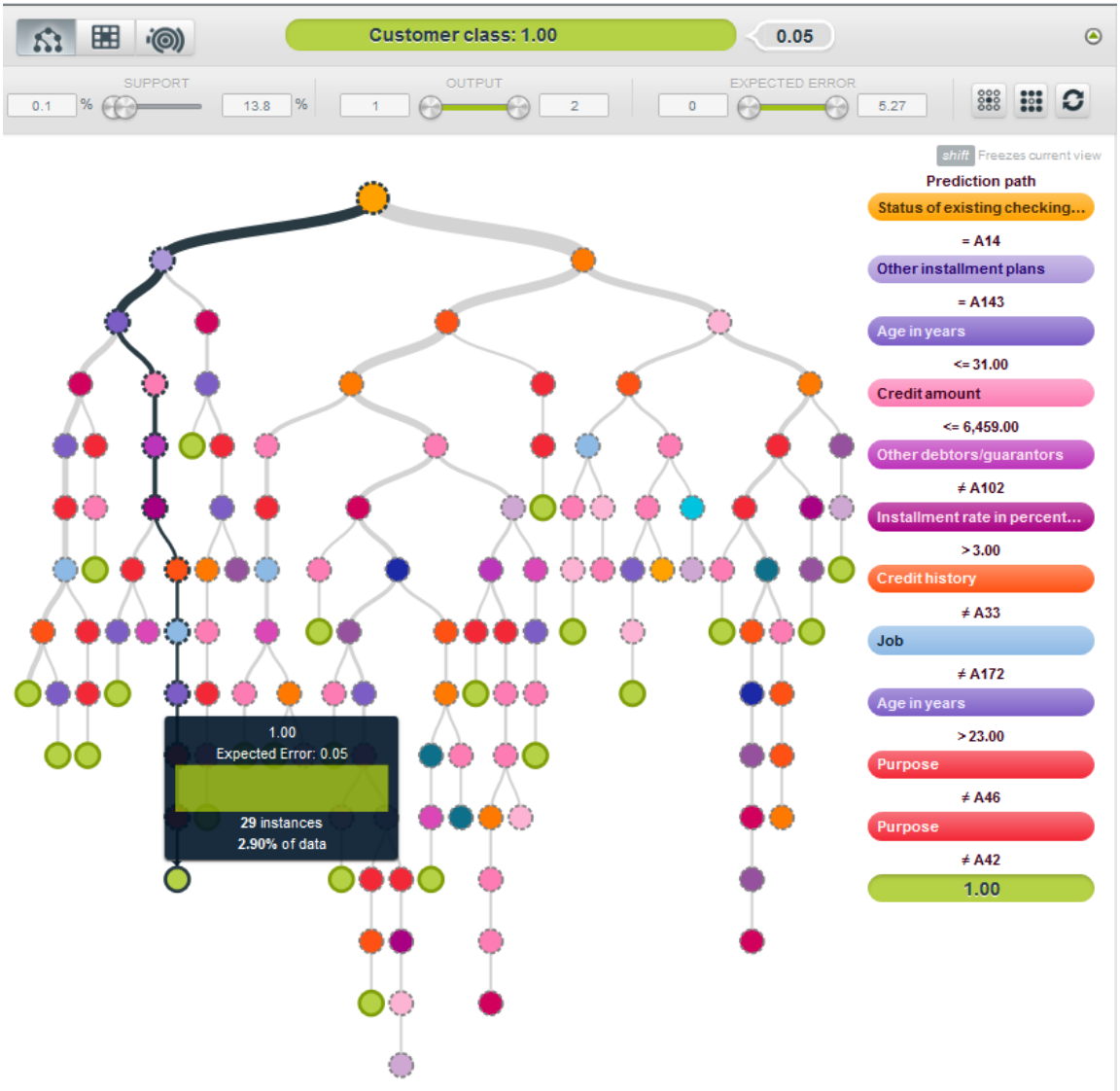
Slika 26 Stablo odlučivanja – dobar klijent, prvi primjer

Prethodna slika predstavlja poslovna pravila za dobrog klijenta, gdje je iznos predviđajuće pogreške samo 0.01. Izabrana grana predviđa, na temelju 138 instanci iz skupa podataka, kako će dobar klijent biti onaj koji nema otvoren tekući račun niti bilo kakve obročne planove. Takav klijent ima više od 31 godinu i jednako ili manje od 66 godina, dok što se trenutnog zaposlenja tiče on je nezaposlen, radi između 1 – 4 godine, 4 – 7 godina ili više od 7 godina. Namjena kredita ovakvog klijenta nije za nikakve popravke, a što se posla tiče on je dakle nezaspolen, nekvalificiran nerezident, nekvalificiran rezident ili pak kvalificiran zaposlenik. Klijent koji odgovara ovom opisu nema nikakvih kašnjenja prilikom isplaćivanja kredita u prošlosti.



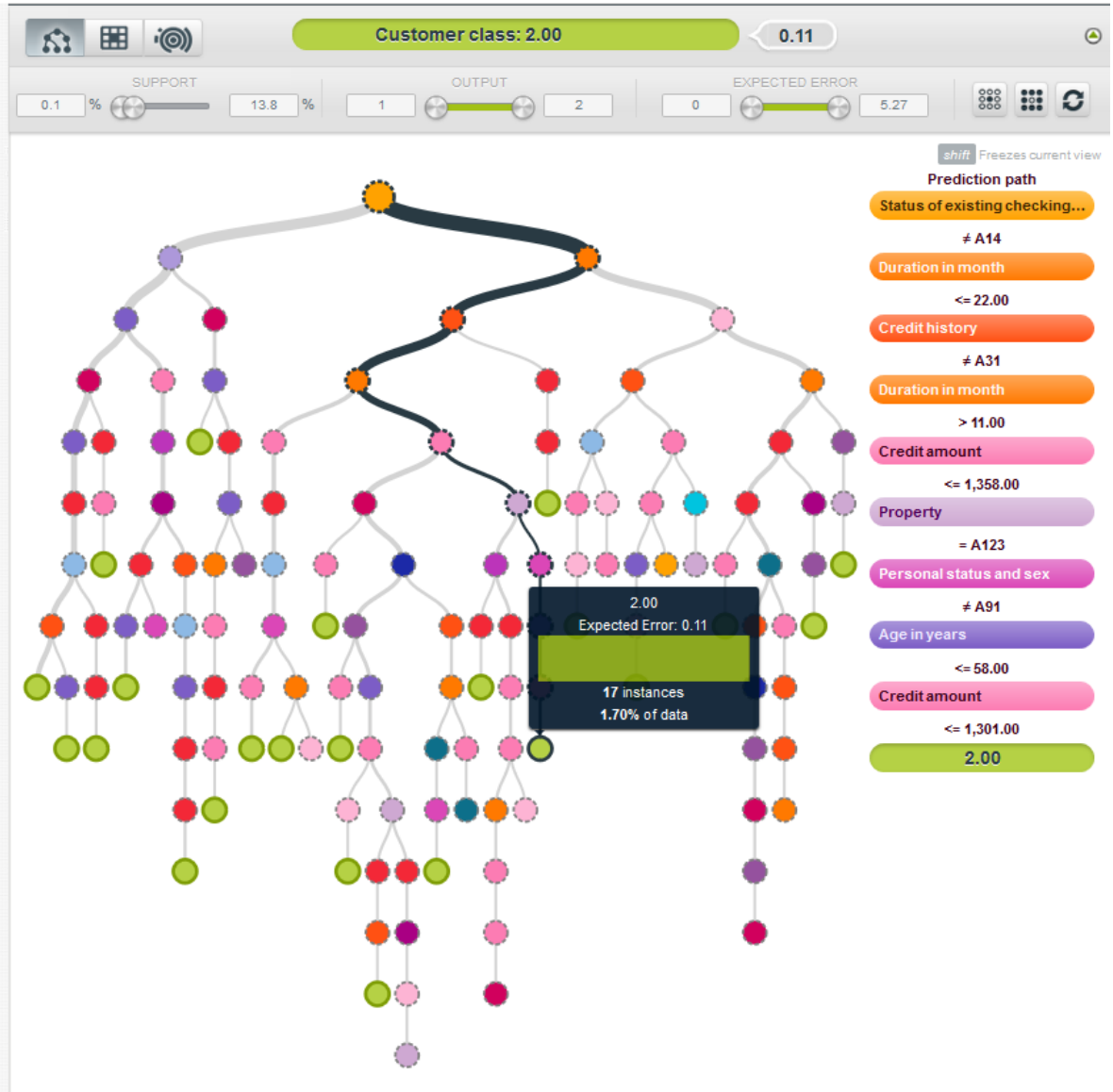
Slika 27 Stablo odlučivanja – dobar klijent, drugi primjer

Slika iznad predstavlja poslovna pravila za dobrog klijenta, gdje je iznos predviđajuće pogreške 0.03. Izabrana grana predviđa, na temelju 47 instanci iz skupa podataka, kako će dobar klijent biti onaj koji nema otvoren tekući račun niti bilo kakve obročne planove. Takav klijent ima jednako ili manje od 30 godinu, dok je iznos kredita jednak ili manji od 6 459, 00 DM. Nadalje, takav klijent nema dužnika ili ima jamca, a stopa rate u ovisnosti postojećeg dohotka iznosi manje ili jednako 3.00 %. Namjena kredita ovakvog klijenta nije za kupovinu novog automobila.



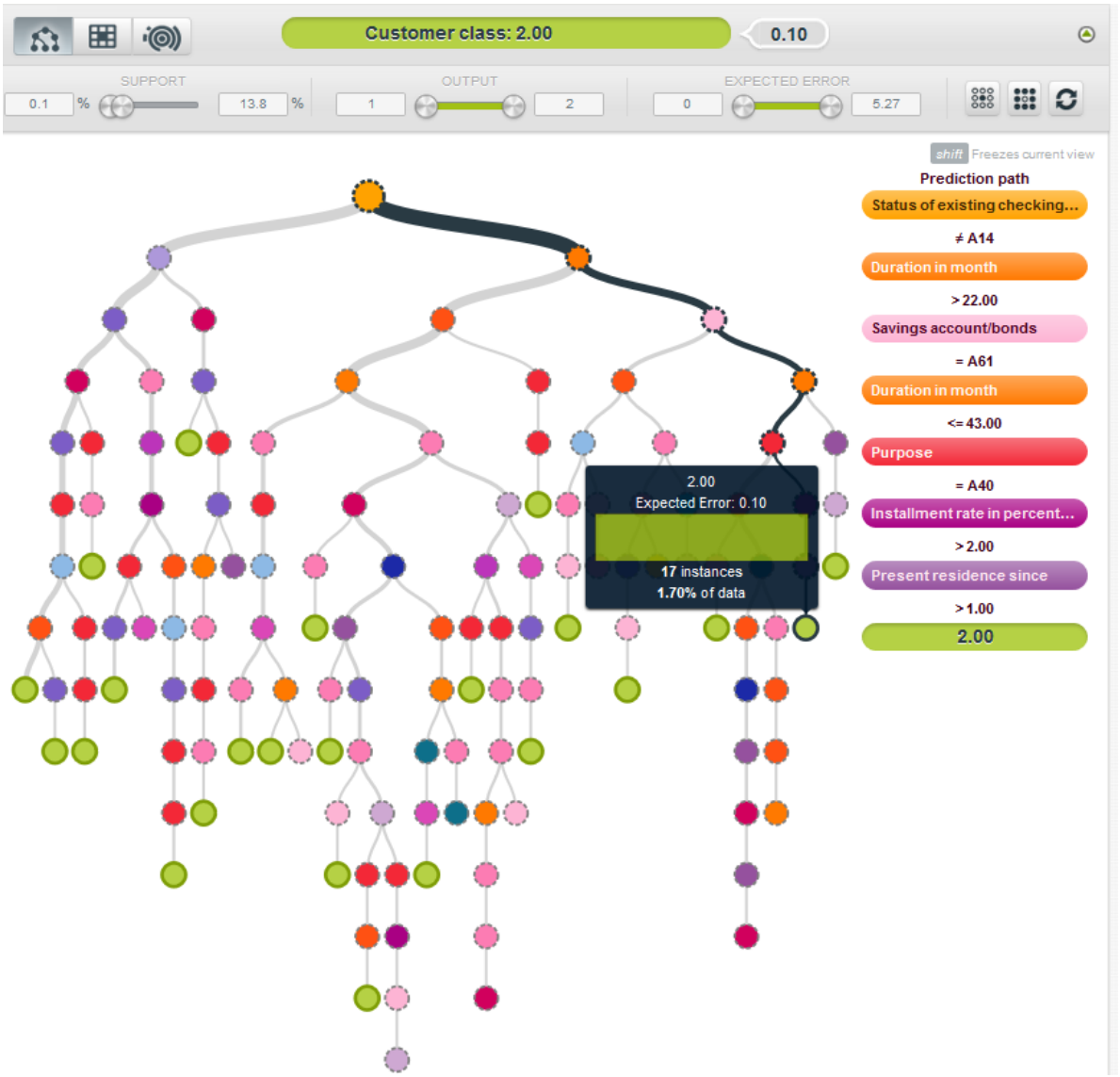
Slika 28 Stablo odlučivanja – dobar klijent, treći primjer

Prethodna slika predstavlja poslovna pravila također za dobrog klijenta, gdje je iznos predviđajuće pogreške 0.05. Izabrana grana predviđa, na temelju 29 instanci iz skupa podataka, kako će dobar klijent biti onaj koji nema otvoren tekući račun niti bilo kakve obročne planove. Takav klijent ima jednako ili manje od 31 godinu, a više od 23 godine, dok je iznos kredita jednak ili manji od 6 459, 00 DM. Nadalje, takav klijent nema dužnika ili ima jamca, a stopa rate u ovisnosti postojećeg dohotka iznosi više od 3.00 %. Klijent koji odgovara ovom opisu nema nikakvih kašnjenja prilikom isplaćivanja kredita u prošlosti, a što se posla tiče nije nekvalificiran niti rezident. Namjena kredita nije za educiranje niti namještaj ili opremu.



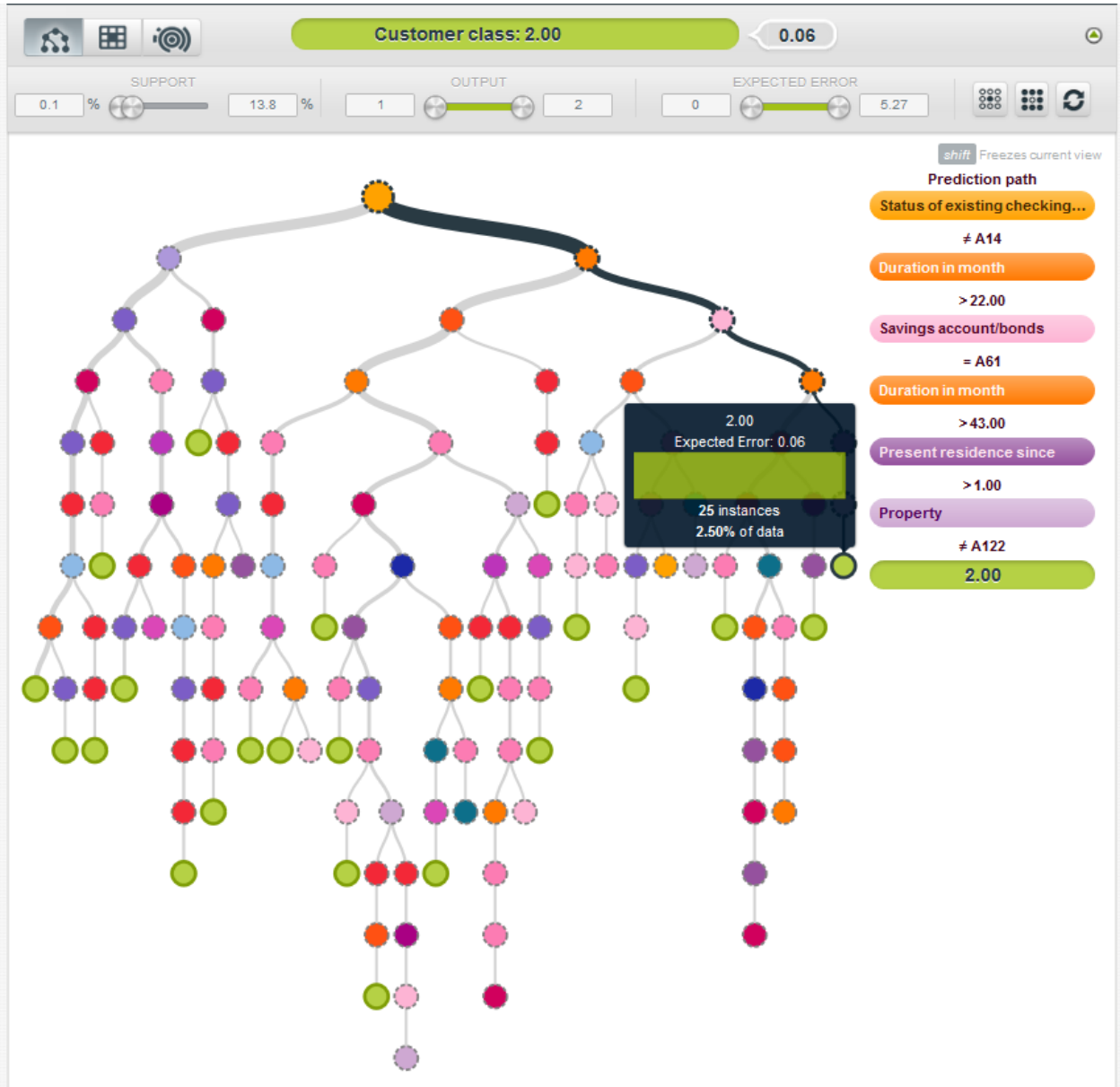
Slika 29 Stablo odlučivanja – loš klijent, prvi primjer

Slika iznad je primjer lošeg klijenta, gdje mogućnost predviđajuće pogreške iznosi 0.11, a grana predviđa na temelju 17 instanci iz skupa podataka. Izabrana grana predviđa kako će loš klijent biti onaj koji ima otvoren tekući račun s iznosom od 0 – 200 DM ili prima plaću barem godinu dana. Otvoren račun u banci ima više od 11 te jednako ili manje od 22 mjeseca. Što se povijesti kreditiranja tiče, ovakav klijent je isplaćivao dugovanja banci koja je vlasnik ovih podataka na vrijeme, dok je iznos kredita jednak ili manji od 1 301, 00 DM. Nadalje, takav klijent u vlasništvu ima automobil ili nešto drugo, što nije nekretnina, stambena štedionica ili životno osiguranje. Ne radi se o muškarcu koji je razveden ili razdvojen, dok se radi o osobi koja ima jednako ili manje od 58 godina.



Slika 30 Stablo odlučivanja – loš klijent, drugi primjer

Prethodna slike je još jedan primjer mogućeg lošeg klijenta, iznos predviđajuće pogreške je 0.10 i grana predviđa na temelju 17 instanci primjera iz skupa podataka. Tako izabrana grana predviđa kako će loš klijent biti onaj koji ima otvoren tekući račun s iznosom od 0 – 200 DM ili prima plaću barem godinu dana. Otvoren račun u banci ima više od 22 te jednako ili manje od 43 mjeseca. Štedni račun ili obveznice ovakvog klijenta iznose manje od 100 DM, dok je namjena kredita za novi automobil. Iznos stope rate u postotku, s obzirom na dohodak je više od 2 %, dok takva osoba ima prijavljeno prebivalište više od 1 godine.



Slika 31 Stablo odlučivanja – loš klijent, treći primjer

Slika iznad je treći primjer mogućeg lošeg klijenta, iznos predviđajuće pogreške je 0.06 i grana predviđa na temelju 25 instanci primjera iz skupa podataka. Tako izabrana grana predviđa kako će loš klijent biti onaj koji ima otvoren tekući račun s iznosom od 0 – 200 DM ili prima plaću barem godinu dana. Otvoren račun u banci ima više od 43 mjeseca. Štedni račun ili obveznice ovakvog klijenta iznose manje od 100 DM i takva osoba ima prijavljeno prebivalište više od 1 godine. Što se vlasništva tiče, takva osoba posjeduje nekretninu, automobil ili nešto drugo, što nije stambena štedionica ili životno osiguranje, no muguće je da takav klijent ne posjeduje ništa u svom vlasništu.

Zaključak je kako sve dobre klijente karakterizira nepostojanje otvorenog tekućeg računa. Dobrim klijentima je još zajedničko kako nemaju nikakvih obročnih planova. Dobri klijenti će imati vjerojatno oko 30 godina i tražit će iznos kredita od otprilike 6 490, 00 DM. Nemaju dužnike ili jamce te je stopa rate izražena postotkom, u ovisnosti od raspoloživog dohotka oko 3.00 %. Namjena kredita vjerojatno neće biti za bilo kakve popravke, kupovinu novog automobila, namještaj ili opremu niti za edukaciju. Dobrim klijentima bi još bilo zajedničko kako nemaju kašnjenja prilikom isplaćivanja svojih obveza. Lošim klijentima je zajedničko kako imaju otvoren tekući račun na kojemu je iznos između 0 – 200 DM ili primaju plaću barem godinu dana. Također, imat će prijavljeno prebivalište duže od godinu dana, a njihov iznos na štednom računu ili obveznicama iznosi manje od 100 DM. Obilježje koje bi još kvalificiralo loše klijente, bilo bi da vjerojatno posjeduju automobil.

7. Zaključak

Izrada završnog rada je ispunila moja očekivanja i ambicije. Pobliže sam se upoznala sa sektorom koji me zainteresirao, uz primjenu tehnika, koje sam također htjela bolje svladati. Smatram kako bi motivacija za izradu završnog rada svakako trebala biti intrizična, te je jako važno pronaći odgovarajućeg mentora koji će podržati ideju pojedinca i usmjeravati ga. Zadovoljna sam jer sam uspjela odabrati temu koju sam htjela, i jer sam kontinuirano imala podršku mentorice.

Samo kreditiranje je nešto što mene osobno tek najvjerojatnije čeka. Svjesna sam kako je kreditiranje zapravo uvijek financijski gubitak, jer se radi o zaduživanju. Nažalost, mogućnosti pojedinaca su različite i za veliku većinu ljudi to je jedan od osnovnih načina financiranja, kako bi ostvarili nešto što im je potrebno za privatne ili poslovne potrebe. Odgovornost je svakog pojedinca da realno procijeni što je to što mu doista treba i kakve su mu mogućnosti, što se zaduživanja, odnosno otplate duga tiče. S druge strane, svjesna sam kako je loše kreditiranje dovelo do mnogih financijskih potresa i kriza, koje nisu utjecale samo na ljude koji su se zadužili, već i na mnogo širu populaciju. Zato smatram kako je doista važno ulagati u politiku kreditiranja, za svaku banku i sličnu organizaciju. Naravno, njima je važan profit, ali se nadam kako gledaju i šire od toga, jer posljedice loše kreditne politike doista mogu biti strašne.

Pripremanjem izabranog skupa podataka, opisivanjem svakog od atributa te izradom stabla odlučivanja, upoznala sam se s nekim poslovnim pravilima koja razlikuju dobre i loše potencijalne kandidate za kreditiranje. Vjerojatno je najizraženija razlika između njih ta da dobri kandidati nemaju otvoreni tekući račun, dok loši kandidati imaju. Svakako, doista je puno faktora koji utječu na mogućnost isplaćivanja obveza pojedinaca, pa je važno dobro se upoznati s osobnim potrebama i mogućnostima, kako se ne bi doveli u financijske probleme, dok je za banke i slične organizacije važno razlikovati dobre od loših potencijalnih kanditata za kreditiranje, odnosno važno je utjecati i poboljšavati politiku kreditiranja, koliko god je to moguće.

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] 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 15.08.2018].

[3] Ž. Garača i M. Jadrić, *Različiti aspekti informacijskog društva*. Split: Ekonomski fakultet u Splitu. 2011

[4] "Rudarenje podataka" (09.09.2013.) [Na internetu]. Dostupno: **https://hr.wikipedia.org/wiki/Rudarenje_podataka** [pristupano 16.08.2018].

[5] G. Santini i S. Bebek: *Vodič za razumijevanje osobnih financija.* Zagreb: Rifin d.o.o. 2005

[6] P. S. Rose i S. C. Hudgins: *Upravljanje bankama i financijske usluge*. Zagreb: Mate d.o.o. 2015

[7] D. Jakovčević: *Upravljanje kreditnim rizikom u suvremenom bankarstvu.* Zagreb: Teb poslovno savjetovanje d.o.o. 2000

[8] Che-hui Lien i I-Cheng Yeh, "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients“, *Expert Systems with Applications,* 2007. [Na internetu]. Dostupno: ScienceDirect, https://www.sciencedirect.com.[pristupano 16.08.2018].

[9] Nor Huselina Mohamed Husain, Seng Huat Ong i Bee Wah Yap, "Using data mining to improve assessment of credit worthiness via credit scoring models", *Expert Systems with Applications,* 2011. [Na internetu]. Dostupno: ScienceDirect, https://www.sciencedirect.com. [pristupano 16.08.2018].

[10] D. Oreški, G. Oreški i S. Oreški, "Hybrid system with generic algorithm and artificial neural networks and its application to retail credit risk assessment", *Expert Systems with Applications,* svi. 2015. [Na internetu]. Dostupno: ScienceDirect, https://www.sciencedirect.com. [pristupano 16.08.2018].

[11] "Stablo odlučivanja" (14.09.2018.) [Na internetu]. Dostupno: http://dms.irb.hr/tutorial/hr_tut_dtrees.php [pristupano 14.09.2018].

Popis slika

Slika 1 Grafikon i krivulje (Izvor: I-Cheng, Che-hui, 2009) 11
Slika 2 Grafikon neuronske mreže (Izvor: I-Cheng, Che-hui, 2009) 11
Slika 3 Dijagram raspršivanja neuronske mreže (Izvor: I-Cheng, Che-hui, 2009) 13
Slika 4 Dijagram toka (Izvor: Wah Yap, Huat Ong, Huain 2011) 15
Slika 5 Model stabla odlučivanja (Izvor: Wah Yap, Huat Ong, Huain 2011) 16
Slika 6 Dijagram rada za GA-NN tehniku (Izvor: Oreški, Oreški, Oreški 2012) 20
Slika 7 Dijagram rada za procjenu kreditnog rizika (Izvor: Oreški, Oreški, Oreški 2012) 20
Slika 8 Primjer stabla odlučivanja (Izvor: Hamilton, Gurak, Findlater i Olive, 2001) 23
Slika 9 Trajanje u mjesecima, numerički atribut – uniformna distribucija iskrivljena udesno 322
Slika 10 Iznos na postojećem tekućem računu, kategorijski 32
Slika 11 Povijest kreditiranja, kategorijski atribut – multimodalna distribucija 32
Slika 12 Namjena, kategorijski atribut – multimodalna distribucija 32
Slika 13 Štedni račun/obveznice, kategorijski atribut – multimodalna distribucija 32
Slika 14 Iznos kredita, numerički atribut – uniformna distribucija iskrivljena udesno 32
Slika 15 Trenutno zaposlenje od, kategorijski atribut – uniformna distribucija 32
Slika 16 Stopa rate u postotku s obzirom na raspoloživi dohodak, numerički atribut – uniformna distribucija 33
Slika 17 Spol i bračni status, kategorijski atribut – unimodalna distribucija 33
Slika 18 Ostali dužnici/jamci, kategorijski atribut – uniformna distribucija iskrivljena udesno 33
Slika 19 Prijavljeno prebivalište od, numerički atribut – uniformna distribucija 333
Slika 20 Vlasništvo, kategorijski atribut – uniformna distribucija 333
Slika 21 Godine, numerički atribut – uniformna distribucija iskrivljena udesno 33
Slika 22 Ostali obročni planovi, kategorijski atribut – multimodalna distribucija 33
Slika 23 Stanovanje, kategorijski atribut – multimodalna distribucija 33
Slika 24 Broj postojećih kredita u ovoj banci, numerički atribut – eksponencijalna distribucija .. 34
Slika 25 : Posao, kategorijski atribut – multimodalna distribucija 34
Slika 26 Broj odgovornih osoba za uzdržavanje, numerički atribut – uniformna distribucija iskrivljena udesno 34
Slika 27 Telefon, kategorijski atribut – uniformna distribucija iskrivljena udesno 34
Slika 28 Inozemni radnik, kategorijski atribut – uniformna distribucija iskrivljena udesno 34
Slika 29 Klasifikacija klijenta, numerički atribut – uniformna distribucija iskrivljena udesno 34
Slika 30 Stablo odlučivanja – dobar klijent, prvi primjer 35
Slika 31 Stablo odlučivanja – dobar klijent, drugi primjer 36
Slika 32 Stablo odlučivanja – dobar klijent, treći primjer 37
Slika 33 Stablo odlučivanja – loš klijent, prvi primjer 38
Slika 34 Stablo odlučivanja – loš klijent, drugi primjer 39
Slika 35 Stablo odlučivanja – loš klijent, treći primjer 40

Popis tablica

Tablica 1 Opis numeričkih atributa 31