UNIVERZA V LJUBLJANI
FAKULTETA ZA DRUŽBENE VEDE

Špela Valand

Odkrivanje zavarovalniških goljufij s pomočjo podatkovnega rudarjenja

Magistrsko delo

Ljubljana, 2013

UNIVERZA V LJUBLJANI
FAKULTETA ZA DRUŽBENE VEDE

Špela Valand

Mentor: doc. dr. Damjan Škulj

Odkrivanje zavarovalniških goljufij s pomočjo podatkovnega rudarjenja

Magistrsko delo

Ljubljana, 2013

Za Izabelo in Andraža

Zahvala

Za strokovno pomoč in koristne nasvete pri izdelavi magistrskega dela se zahvaljujem mentorju doc. dr. Damjanu Škulju.

Andraž, iskrena hvala za vse spodbudne besede v trenutkih šibkosti in krize ツ

Odkrivanje zavarovalniških goljufij s pomočjo podatkovnega rudarjenja

Zavarovalniške goljufije so z vidika zavarovalništva izjemno velik problem, saj tovrstni gospodarski panogi povzročajo visoke letne izgube. Same zavarovalniške hiše se z reševanjem tega problema ukvarjajo znotraj posebnih oddelkov za odkrivanje goljufij, seveda pa je raziskovanje tega področja med drugim privlačno tudi strokovnjakom, ki se ukvarjajo s podatkovnim rudarjenjem. V procesu podatkovnega rudarjenja operiramo z ogromnimi zbirkami podatkov, kar pri uporabi modelov, ki temeljijo na osnovi algoritmov, omogoča napovedovanje in odkrivanje vzorcev iz podatkov. Tako se ta metoda z vidika odkrivanja goljufij izkaže kot ena izmed najbolj primernih. Tudi sami se v magistrskem delu lotimo raziskovanja tega področja in sicer na osnovi teoretskih izhodišč predstavimo model za odkrivanje zavarovalniških goljufij. Naš model temelji na tehniki analize odkrivanja anomalij, s katero v zbirki podatkov odkrivamo odstopanja skritih vzorcev. Na koncu model tudi testiramo na realnih podatkih, in sicer na primeru zavarovalniških zahtevkov, kjer za njegovo izvedbo uporabimo program SPSS Modeler. Natančnost načrtovanega modela je 87%, kar pomeni, da je med tridesetimi primeri pravilno identificiral šestindvajset primerov. Če pa se omejimo samo na sumljive primere, pa je izmed desetih pravilno identificiral šest le-teh.

Ključne besede: zavarovalniške goljufije, podatkovno rudarjenje, analiza povezav, odkrivanje anomalij, načrtovanje modela

Insurance fraud detection by means of data mining

Insurance fraud represents extremely large problem to insurance companies, causing large annual financial losses. Insurers deal with this problem in special departments for fraud detection, however this research field is also engaging for data mining specialists. Data mining process requires operating with large data sets, which by the use of algorithm based models, enables predicting and detecting patterns in the data. From that point of view, this method seems one of the most adequate in the process of fraud detection. Thus, the purpose of this masters thesis is to build and represent the model for fraud detection. Our model is based on technique of anomaly detection analysis, which enables detection of hidden anomalous patterns in the data set. For the purpose of model evaluation we test our model on the real data of insurance claims, by the use of program SPSS Modeler. The accuracy of the built model is equal to 87%, and it correctly identifies twenty-six out of thirty cases. But considering only the suspicious cases, it correctly identifies six out of ten suspicious cases.

Keywords: insurance fraud, data mining, link analysis, anomaly detection, model development

KAZALO

KAZALO GRAFOV, TABEL IN SLIK

1 UVOD

V zavarovalništvu se velikokrat srečamo z goljufijami, ki so ene izmed razlogov visokih izgub in globalno presegajo sto milijard evrov (Furlan in Bajec 2009), zato predstavljajo zaskrbljujoč in velik problem zavarovalnic. Te se srečujejo z izzivom preiskovanja in odkrivanja tega hitro rastočega pojava, ki ga razumemo kot nezakonito dejanje (Derrig 2002; Whitaker 2009). Veliko pomoč pri reševanju tega perečega problema jim nedvomno predstavljajo lastne baze podatkov, kjer se kot najbolj primerna tehnika obdelave tovrstnih podatkov izkaže podatkovno rudarjenje, saj nam omogoča odkrivanje in analiziranje različnih podatkovnih vzorcev, informacij in trendov. Seveda pa z razvojem in uporabo tehnik podatkovnega rudarjenja za odkrivanje goljufij, le-to postaja izredno priljubljeno in hkrati nujno področje raziskovanja te storitvene dejavnosti (Bolton in Hand 2002; Fawcett in Provost 2002; Thiruvadi in Patel 2011; Gepp in drugi 2012).

Pri podatkovnem rudarjenju operiramo z veliko spremenljivkami in pogoji, kar predstavlja svojevrsten izziv, saj jih moramo pravilno izbrati in uporabiti (Larose 2006). V magistrskem delu se bomo posvetili odkrivanju zavarovalniških goljufij s pomočjo podatkovnega rudarjenja. Večji del magistrskega dela bo namenjen načrtovanju modela, kar vključuje teoretično načrtovanje na osnovi literature. Iz teoretičnih izhodišč bo nato sledila izbira relevantnih spremenljivk in ustrezne metode. Za izhodiščno metodo bomo uporabili analizo odkrivanja anomalij, ki je ena od zelo primernih metod za odkrivanje zavarovalniških goljufij. Obravnavali pa bomo tudi možnost uporabe drugih metod. Dobljeni primer bomo na koncu testirali na realnih podatkih in na podlagi teoretično utemeljenih kriterijev preverili, kakšna je stopnja njegova uspešnosti.

Magistrsko delo se nanaša na področje družboslovne informatike, kjer informacijsko-komunikacijsko tehnologijo v družboslovju uporabimo kot aplikacijo.

2 RAZISKOVALNO VPRAŠANJE

Pri odkrivanju zavarovalniških goljufij veliko vlogo igrajo odškodninski zahtevki, ki jih za izplačilo denarne odškodnine v primeru nezgodnih, škodnih ali drugih zavarovalniških primerih vložijo zavarovanci. Zavarovalnice se zavedajo, da je med njimi lahko nekaj takih, ki so plod goljufije. Ob upoštevanju tega dejstva, nas tako v zavarovalniških zbirkah podatkov zanima prisotnost anomalij, ki nam omogoča odkrivanje goljufivih odškodninskih zahtevkov. Rešitev za to lahko iščemo v različnih tehnikah, med katerimi poznamo tudi analizo odkrivanja anomalij. S pomočjo analize anomalij lahko poleg odkrivanja zavarovalniških goljufij, odkrivamo tudi goljufije s kreditnimi karticami, mobilnimi telefoni, trgovanja z notranjimi informacijami, ipd. Tehnika je primerna tudi za odkrivanje raznovrstnih napak, vdorov v računalniške sisteme, vojaškemu nadzorovanju sovražnikovih dejavnosti ter drugih aktivnosti (Chandola in drugi 2009, 14–16).

Tudi sami se bomo v magistrskem delu osredotočili na odkrivanje zavarovalniških goljufij z analizo odkrivanja anomalij. Skozi delo pa bomo tako poskušali odgovoriti na naslednji raziskovalni vprašanji:

- Kako učinkovito načrtovati model za odkrivanje zavarovalniških goljufij?
    - Kakšna je vloga izbire posameznih spremenljivk pri načrtovanju modela?
    - Kako je pri načrtovanju pomembna izbira statističnega modela?
- Kakšna je stopnja uspešnosti odkrivanja zavarovalniških goljufij s tehniko podatkovnega rudarjenja s poudarkom na analizi odkrivanja anomalij?

3 ZAVAROVALNIŠKE GOLJUFIJE

Zavarovalniški strokovnjaki razlikujejo med težjimi in lažjimi oblikami goljufij. V prvem primeru gre za goljufije s področja gospodarskega kriminala, v drugem pa za zlorabe oz. neetično vedenje posameznika (Tennyson 2008). Na splošno zavarovalnice zavarovalniške goljufije definirajo kot jasno in zavestno protipravno ravnanje z zavajanjem zavarovalnice, katerega namen je finančno okoriščanje v lastno ali tujo korist (Derrig in drugi 2006).

Razlikujemo tudi med naslednjimi skupinami goljufij, kjer gre pri prijavi škode za (Kopše 2004, 130):

- odstopanja od dejanskega škodnega stanja, kjer govorimo o povečevanju oz. »napihovanju« škode;
- zavarovalni dogodek, ki se v resnici ni primeril ali izmišljevanje zavarovalnega dogodka v celoti, in
- namerno povzročanje oz. uprizarjanje zavarovalnega dogodka.

Zavarovalniške goljufije se pojavljajo na različni področjih in v različnih oblikah, v nadaljevanju tako sledi pregled in opis le-teh, kasneje pa bomo predstavili še teoretične vidike zavarovalniških goljufij.

3.1 PODROČJA IN OBLIKE ZAVAROVALNIŠKIH GOLJUFIJ

Z zavarovalniškimi goljufijami se pogosto srečamo na področju avtomobilskih, premoženjskih in nezgodnih zavarovanj, seveda pa so storilci goljufij lahko tudi agenti ali posredniki sami. V nadaljevanju si bomo znotraj posameznih področij ogledali nekatere oblike zavarovalniških goljufij.

3.1.1 ZAVAROVALNIŠKE GOLJUFIJE AVTOMOBILSKIH ZAVAROVANJ

Zavarovalniške goljufije na področju avtomobilskih zavarovanj se zagotovo uvrščajo med ene izmed najbolj razširjenih oblik tovrstnih dejanj. Poznamo več oblik goljufij na

tem področju, katerim je skupno prikazovanje drugačnih okoliščin od dejanskih, pri čemer zavarovanec spravi zavarovalnico v zmotno prepričanje. Dokazovanje goljufij na tem področju je kompleksen in dolgotrajen proces, tako da pogosto zavarovalnici ne preostane drugega, kot da s sumljivo stranko ne sklepa več (novih) zavarovalnih poslov (Lamberger 2004, 117–118).

Sledijo opisi nekaterih oblik zavarovalniških goljufij na področju avtomobilskih zavarovanj, ki so povzete po Whitakerju (2009, 4–5; 2013, 14–16).

Opustitev

Lastnik opusti vozilo, v upanju, da bo le-to ukradeno ali bodo ukradeni njegovi deli, odpeljan na odlagališče in uničen. Namen tovrstnega dejanja je izplačilo denarja na podlagi zavarovalne police ali v izogib poravnave posojila, kjer gre navadno za drago vozilo z nizkim plačilom akontacije.

Naknadna prijava

Gre za obliko goljufije, kjer je v prometni nesreči udeležena trenutno nezavarovana oseba. Ta kasneje pridobi zavarovanje in po preteku določenega obdobja prijavi nesrečo, kot npr. nesrečo s pobegom ali nepooblaščeno uporabo vozila.

Popravilo vozila

Pri tej obliki goljufije gre predvsem za zaračunavanje novih ali originalnih nadomestnih delov za nadaljnjo prodajo, medtem ko so bili v vozilo v resnici nameščeni že uporabljeni ali neoriginalni nadomestni deli. Pri tej obliki goljufije, se srečujemo tudi z nepotrebnimi avtoličarskimi popravili.

Tihotapljenje vozil

Ta oblika goljufije vključuje nakup vozila z najvišjo stopnjo financiranja, pri kateri se pridobi ponarejeno potrdilo o legalni lastninski pravici. Oseba za zavarovanje ne plača veliko, saj je vozilo zavarovano po najnižji odbitni franšizi. Vozilo, ki je nato poslano v tuje pristanišče, se prijavi kot ukradeno in se na novi lokaciji proda na črnem trgu. Na koncu je za »ukradeno« vozilo zavarovancu izplačana tudi zavarovalna premija.

Fantomska vozila

V tem primeru se izda potrdilo o legalni lastninski pravici na vozilu, za katerega ni dokazano, da je kadarkoli obstajalo ali pa je bilo ocenjeno kot totalna škoda, pri čemer potrdilo o tem ni bilo izdano. Pri menjavi identifikacijske številke vozila se le-to nadomesti s številko ukradenega vozila iste znamke in letnice izdelave.

Uprizorjene prometne nesreče

Gre za prometne nesreče, v katerih je vnaprej določeno na kakšen način bo vozilo udeleženo v nesreči. Storilci v nesrečo vključijo nedolžne udeležence v prometu ali pa je le-ta plod organiziranega dejanja skupine ljudi. Največkrat storilci za uprizorjene prometne nesreče uporabijo eno in isto vozilo.

Napihnjena škoda

To je oblika goljufije, kjer so v ocenjene stroške potrebnega popravila vključeni tudi predvideni stroški škode, ki je na vozilu nastala že pred primerom nezgode. Namen te oblike goljufije je tudi pokritje stroškov odbitne franšize, da zavarovanec ne bi utrpel izgube.

3.1.2 ZAVAROVALNIŠKE GOLJUFIJE PREMOŽENJSKIH ZAVAROVANJ

Pri zavarovalniških goljufijah premoženjskih zavarovanj gre predvsem za vložitev odškodninskih zahtevkov za premoženje, kjer so vrednosti izgube precenjene. Na odškodninskem zahtevku je tako poleg dejansko izgubljenega premoženja prikazano tudi premoženje, ki (Whitaker 2009, 5-6; Whitaker 2013, 12–13):

- ni nikoli obstajalo,
- ni bilo nikoli v lasti zavarovanca ali
- je bilo prodano že pred nastankom nezgodnega primera.

S to obliko goljufij se v največji meri srečamo v primeru požarov ali požigov zaradi dobička, kraj, vandalizma, poplav, ipd. Poseben primer so t.i. »papirnate ladje«, kjer je se vloži odškodninski zahtevek za potopljeno vodno plovilo. Dostikrat je namreč težko

dokazati, da je vodno plovilo dejansko obstajalo, saj ga lahko registriramo samo s kupoprodajno pogodbo (Whitaker 2013, 12–13).

3.1.3 ZAVAROVALNIŠKE GOLJUFIJE NEZGODNIH IN ŽIVLJENJSKIH ZAVAROVANJ

V primerih zavarovalniških goljufij na področju **nezgodnih zavarovanj** gre za uveljavljanje prekomernih dnevnih odškodninskih zneskov, ki so posledica samopoškodb ali zlorabe zdravniških spričeval o poškodbi (tudi bolniških dopustov). Med indice, ki so lahko pokazatelj zavarovalniških goljufij na tem področju, lahko štejemo naslednje (Lamberger 2004, 114–115):

- zdravniški pregled vseh oškodovancev pri istem zdravniku;
- poškodbe, ki naj bi jih oškodovanci utrpeli, so zelo subjektivne narave in težko dokazljive, kot so npr. poškodbe vratnih vretenc;
- nesorazmernost obsega telesnih poškodb (pretiravanje) s poškodbami na vozilu, v katerem se je oškodovanec ponesrečil;
- vedno je predpisano enako in dolgotrajno zdravljenje za različne poškodbe;
- ne glede na vrsto poškodbe in nesreče se izdajo vedno ista zdravniška potrdila in izvidi;
- zdravniška spričevala rednih kontrol so izdana na dela proste dneve (nedelja, prazniki ali drugi dela prosti dnevi).

Med najmanj razširjene zavarovalniške goljufije štejemo tiste, ki nastanejo pri uveljavljanju odškodnine iz naslova **življenjskih zavarovanj**. Zavarovalnica se ob dogodku zavarovalnega primera (doživetje ali smrt zavarovanca) obveže, izplačati odškodnino zavarovancu ali pooblaščeni osebi (Petrović 2004, 73–74). Med oblike zavarovalniških goljufij na področju življenjskih zavarovanj štejemo lažno smrt in umor iz koristoljubja, ki ga izvršijo svojci oz. upravičenci do odškodnine (Lamberger 2004, 113).

3.1.4 ZAVAROVALNIŠKE GOLJUFIJE AGENTOV ALI POSREDNIKOV

Gre za področje goljufij, za katerega je značilno nepošteno in nekorektno opravljanje dela zavarovalniških agentov oz. posrednikov, v katero je zavarovanec vpleten brez njegove vednosti. Na tem področju poznamo naslednje oblike goljufij (Whitaker 2013, 11):

- **zamenjava**: agent obstoječo polico zamenja z novo, ki jo brez vedenja zavarovanca lahko prenese na drugo zavarovalnico ali pa jim proda vrsto zavarovanja, ki je ne potrebujejo ali celo ne želijo;
- **podtikanje**: v zavarovalno polico agent vključi dodatno kritje brez vedenja ali privolitve zavarovanca;
- **prekomerno trgovanje**: agent pri podaljšanju zavarovanja zavarovanca prepriča v nakup dodatnega zavarovanja brez dodatnih stroškov, kar največkrat seveda ne drži, saj je strošek nove police precej višji kot strošek stare;
- **goljufija s premijo**: agent si prilasti plačano premijo, ki je pri zavarovalnici ne prijavi, in nato sestavi lažno polico za zavarovanca, če le-ta ne prijavi škode oz. nezgode, za prevaro nihče ne izve.

3.2 TEORETIČNI VIDIKI ZAVAROVALNIŠKIH GOLJUFIJ

Izvor potrošniške nepoštenosti Tennysonova (2008, 1181) vidi v kompleksnem medsebojnem delovanju interesov in okoliščin, ki jo usmerjajo morala, priložnost, družbene norme in institucionalni kontekst. Kompleksnost tega delovanja je še posebej očitna na področju zavarovalniških goljufij, ki so lahko rezultat dolgotrajnega zavestnega **načrtovanja** ali trenutnega navdiha (**priložnostna goljufija**). O načrtovani zavarovalniški goljufiji govorimo takrat, ko do spontanega škodnega oz. nezgodnega primera ne pride, če pa do tega pride nenamerno, ampak so okoliščine pri prijavi prirejene, govorimo o priložnostni goljufiji (Tennyson 2008, 1183).

Poleg finančnega okoriščanja, povode za zavarovalniške goljufije iščemo tudi v posameznikovem občutku upravičenosti do imetja odškodnine, obupa, zamere, lahko pa

do teh pride nenamerno, in sicer kot posledica napačnega razumevanja pogodbenih pogojev. Gledano različnih zornih kotov tako zavarovalniške goljufije obravnavamo kot **ekonomsko-pogodbeni**, **moralno-psihološki**, **moralno-sociološki** problem in problem na področju **gospodarskega kriminala** (Tennyson 2008, 1181). V nadaljevanju bomo predstavili naštete teoretične vidike zavarovalniških goljufij.

3.2.1 EKONOMSKO-POGODBENI VIDIK

Govorimo o vidiku zavarovalniških goljufij, ki temelji na ekonomski teoriji moralnega tveganja, kjer do goljufije pride znotraj okvira pogodbenega razmerja med zavarovalnico in zavarovancem, saj jo razumemo kot ekonomski odziv na pogodbo. Zavarovanec se na podlagi pričakovanega dobička iz odškodnine odloči, ali bo vložil odškodninski zahtevek. Če ta ugotovi, da bo v primeru izplačila odškodnine finančno pridobil, se na koncu odloči za vložitev goljufivega zahtevka. Moralno tveganje zavarovalnicam predstavlja bistven problem pri njihovem poslovanju, ki tako stremijo k zmanjšanju teženj zavarovancev po tovrstnem ravnanju (Baker v Tennyson 2008, 1188–1189).

Zato so zelo pomemben odgovor na goljufije preiskovanja odškodninskih zahtevkov, saj le-ta za zavarovalnice predstavljajo veliko korist, če je njihova posledica odkritje goljufivega odškodninskega zahtevka ali zavrnitev le-tega. Vendar pa taka preiskovanja v nekaterih primerih ne pridejo v poštev, saj so njihovi stroški previsoki ali pa je težko odkriti resnično ozadje nastale škode oz. nezgode. V takih primerih zavarovalnice lahko uporabijo učinke odvračanja oz. strategije, ki omejujejo izplačila zahtevkov. Neposredna korist takega ukrepa je seveda zmanjšanje stroškov izplačil odškodninskih zahtevkov, na drugi strani pa se posredna korist kaže v zmanjševanju vlaganj naklepnih goljufivih zahtevkov. Ker pa se škodni oz. nezgodni primeri med seboj razlikujejo, morajo zavarovalnice v najprej predvideti, kakšno je razmerje med pričakovanim znižanjem vlaganja goljufivih zahtevkov in pričakovanimi stroški reševanja sumljivih zahtevkov (Crocker in Tennyson 2002, 470; Tennyson 2008, 1189). V primeru, da bi bili dokazi o spornosti zahtevkov s pravnega vidika prešibki oz. sodni stroški previsoki, pa se morajo zavarovalnice soočiti s pritiskom izplačil spornih zahtevkov (Abraham v Tennyson 2008, 1189).

3.2.2 MORALNO-PSIHOLOŠKI VIDIK

Psihološki vidik zavarovalniških goljufij poudarja, da lahko obseg ponotranjenih in aktivnih družbenih norm vpliva na to, ali se bo ustaljeni posameznikovi vedenjski vzorci odražali tudi takrat, ko ga bomo opazovali. Tako so znotraj tega vidika posameznikovi notranji mehanizmi nagrajevanja pomemben dejavnik poštenega vedenja. Ko je pri posamezniku poštenost le šibko razvita, obstaja velika verjetnost, da notranji mehanizem nagrajevanja posameznika ne bo odvrnil od goljufije. V primeru, da so tovrstni mehanizmi nedejavni, bo imel posameznik težave pri oceni svojega vedenja v odnosu do sprejetih družbenih norm, kar vodi k višji stopnji nepoštenega vedenjskega vzorca (Tennyson 2008, 1193–1194).

Mazar, Amir in Ariely (2008, 5–6) so v eksperimentih na področju goljufij domnevali, da se posameznik, ko mu je dana priložnost za denarno okoriščanje z goljufijo, odloči za tak korak, in da je obseg goljufanja odvisen od verjetnosti razkrinkanja oz. od pričakovane kazni za goljufivo dejanje. Raziskovalci so prišli do zaključkov, da so nekateri posamezniki, ko se jim je ponudila priložnost, res goljufali, vendar pa je bil obseg goljufij precej nizek, pri čemer je le nekaj posameznikov goljufalo do maksimalno možne stopnje. Dokazali so tudi, da se pri kontekstualni manipulaciji obseg goljufij zmanjša – torej, ko se posameznika opomni na moralne standarde (Mazar in drugi 2008, 34). Tako Tennysonova (2008, 1194) na podlagi tega primera trdi, da ima večina ljudi ponotranjene družbene norme o poštenosti, in da le-te lahko vplivajo na posameznikovo vedenje, s čemer so dokazali, da so notranji mehanizmi nagrajevanja pomemben dejavnik v procesu odločanja.

3.2.3 MORALNO-SOCIOLOŠKI VIDIK

Nagnjenost h goljufijam v okviru sociološke dimenzije je odvisna od zavarovančevega odnosa do vlaganja goljufivih zahtevkov. Na intenziteto tega odnosa vplivajo konstantni denarni stroški, ki nastanejo pri vzdrževanju zavarovanja, sama korist od goljufije, možno družbeno stigmatiziranje in psihične posledice na posameznika zaradi vpletanja v goljufije. Sam odnos do neprimernega vedenja se bo tako v določeni meri odražal v

sprejetih družbenih normah. Teorija razvoja družbenih norm med drugim poudarja tudi vpliv primerljivih skupin (peer gorup) ali omrežij ljudi, kar pomeni naslednje – če so znotraj teh zavarovalniške goljufije sprejete, potem obstaja večja verjetnost, da jih bo sprejel tudi posameznik, ki pripada tej skupini ali omrežju. Torej, višja kot je toleranca do goljufij oz. je sprejeta percepcija, da so le-te nekaj vsakdanjega, bolj bodo v družbi sprejete in manjše bodo družbene posledice na posameznika, ki bo v goljufiji sodeloval (Tennyson 2008, 1192).

Poznamo tudi teorije, ki predpostavljajo, da lahko posameznikova percepcija o zavarovalnicah, torej pozitivna ali negativna, vpliva tudi na posameznikovo poštenost poslovanja z zavarovalnicami (Axelrod v Tennyson 2008, 1193). To pomeni, da bodo zavarovanci, ki imajo o zavarovalnicah negativno percepcijo ali so imeli z njimi negativne izkušnje, v večji meri sprejemali zavarovalniške goljufije. Na drugi strani pa Cialdini (Tennyson 2008, 1193) govori o ravno obratnih modelih, ki se navezujejo na percepcijo o poštenosti poslovanja zavarovalnic. Torej, če posameznik meni, da zavarovalnice z zavarovanci ravnajo nepošteno, da veliko preveč služijo ali so zavarovalne premije nepravično določene, bo tudi bolj nagnjen h sprejemanju zavarovalniških goljufij. Tako na podlagi omenjene teorije, Tennysonova (2008, 1193) predlaga, da bi se bilo potrebno reševanja problema zavarovalniških goljufij in zmanjšanju sprejemanja le-teh, lotiti na sistematičen način in izboljšati podobo zavarovalniške industrije ter okrepiti odnos med zavarovanci in zavarovalnicami.

3.2.4 KRIMINALNI VIDIK

Ko zavarovalniško goljufijo obravnavamo kot kriminalno dejanje, se moralno tveganje omeji v okvir potencialnih kazenskih sankcij, ki lahko vključujejo denarne kazni, zaporno kazen in izgubo lastnega ugleda. Po tradicionalni teoriji kriminalitete, naj bi posameznikova odločitev o vložitvi goljufivega zahtevka, temeljila na podlagi ocene morebitnega dobička v primeru uspešne zoper višino kazni v primeru neuspešne prijave škode oz. nezgode in stopnjo verjetnosti odkritja goljufije. Zavarovanec bo najverjetneje vložil goljufiv zahtevek, če bo pričakovan dobiček v primeru uspešne prijave odtehtal predvideno kazen v nasprotnem primeru. Seveda se odločitev o takem dejanju med

ljudmi razlikuje, in odvisna od posameznikove sposobnosti zaznavanja verjetnosti uspeha goljufije, njihove stopnje naklonjenosti k tveganju, obrestne mere in občutljivost na izgubo ugleda (Tennyson 2008, 1190–1191).

Tennysonova (2008, 1191) obenem trdi, da tudi, če je verjetnost odkritja goljufije razmeroma majhna, vendar so predvidene kazni dovolj visoke, obstaja verjetnost, da do kaznivega dejanja kljub vsemu ne bo prišlo. V primeru, da kazni ne morejo biti neomejene, saj so te lahko določene z zakonom, so zavarovalnice prisiljene v iskanje rešitev za odvračanje od goljufij, ampak kot nam je že znano so te metode lahko neuspešne. Kljub temu pa Nagin in Pogarsky (v Tennyson 2008, 1192) menita, da so dobro izdelane metode odvračanja od goljufij, lahko učinkovitejše od visokih kazni.

4 PODATKOVNO RUDARJENJE

V tem poglavju se bomo osredotočili na podatkovno rudarjenje, in sicer bomo najprej predstavili nekaj osnovnih pojmov, ki se navezujejo na podatkovno rudarjenje, kasneje sledi njegova definicija, predstavljene pa bodo tudi njegove korake. V nadaljevanju se bomo osredotočili na opis različnih tehnik podatkovnega rudarjenja, za zaključek pa bomo ovrednotili njegov pomen pri odkrivanju zavarovalniških goljufij.

4.1 PREDSTAVITEV OSNOVNIH POJMOV

Za začetek so podane opredelitve nekaterih ključnih pojmov, ki služijo lažjemu razumevanju spodaj predstavljene teorije podatkovnega rudarjenja.

Vzorec je lahko kakršnakoli kombinacija vrednosti, ki nosijo pomen znotraj obravnavanega konteksta ali domene. Vzorci temeljijo na individualni interpretaciji podatkov, okolja, okoliščin in kvaliteti zbranih podatkov (Westpahl 2009, 68).

Model je skupek pravil, sestavljen iz podatkov, statistik in vzorcev, ki se uporabijo za generiranje novih podatkov, s katerimi lahko kasneje napovedujemo in pojasnjujemo relacije med podatki (Rajaraman in Ullman 2012).

Zbirka podatkov je nabor meritev, ki smo jih črpali iz določenega okolja ali procesa. Kot najbolj osnoven primer zbirke podatkov lahko razumemo tisti skupek elementov (posamezniki, entitete, predmeti, zapisi, ipd), kjer ima vsak od elementov isti nabor meritev (spremenljivke, lastnosti, ipd) (Hand in drugi 2001, 2). Ko govorimo o **transakcijskih zbirkah podatkov**, pa imamo v mislih tiste, ki obsegajo le del celotne zbirke podatkov.

Algoritem je nabor korakov, operacij ali postopkov, s pomočjo katerih bomo prišli do končnega rezultata (Nisbet in drugi 2009, 791).

Asociacijska pravila so osnova za iskanje razmerij in povezav med atributi oz. lastnostmi v zbirki podatkov. Asociacijsko pravilo zavzema naslednjo opisno obliko, »če se zgodi predhodni dogodek, potem sledi posledica« (Larose 2005, 180).

4.2 OPREDELITEV PODATKOVNEGA RUDARJENJA

Poznamo več definicij podatkovnega rudarjenja, njegov osnovni namen pa je, s pomočjo konstruiranja statističnega modela (Rajaraman in Ullman 2012, 1), iskanje skritih vzorcev v velikih zbirkah podatkov z uporabo algoritmov strojnega učenja (Nisbet in drugi 2009). Operiranje z velikimi zbirkami podatkov je v procesu podatkovnega rudarjenja tako nujno, saj bi v nasprotnem primeru težko govorili o klasični eksplorativni (poizvedovalni) analizi podatkov, ki jo izvajajo statistiki (Hand in drugi 2001, 2).

Pri podatkovnem rudarjenju gre za predstavitev velike količine in, na prvi pogled, nepovezanih podatkov na drugačen način, s katerim dosežemo njihovo večjo koristnost. Iščemo torej relacije in strukture znotraj zbirke podatkov, ki morajo seveda predstavljati novost. Kljub temu pa le iskanje razmerij in struktur ni dovolj, saj morajo biti le-te tudi razumljive. V procesu podatkovnega rudarjenja tako operiramo z vzorci in modeli, ki vključujejo linearne enačbe, pravila, razvrščanja v skupine, grafične prikaze, drevesne strukture in ponavljajoče se vzorce v časovnih vrstah (Hand in drugi 2001, 1–2). Seveda pa se znotraj podatkovnega rudarjenja srečujemo z različnimi izzivi, kot so uporaba podatkov v transakcijskih zbirkah podatkov, redukcija, transformacija podatkov, čiščenje, raztresenost in redkost podatkov (Nisbet in drugi 2009, 25).

Proces iskanja razmerij in struktur znotraj zbirke podatkov, torej iskanja točnih, ustreznih in uporabnih načinov predstavitve nekaterih vidikov podatkov, vključuje številne korake in sicer (Hand in drugi 2001, 3):

- določanje narave in strukture načina predstavitve;
- odločanje o oceniti in primerjavi, kako dobro različni načini predstavitev ustrezajo podatkom;
- izbira algoritma za optimizacijo uspešnosti napovedne funkcije in
- odločanje o načelih upravljanja s podatki za učinkovito izvajanje algoritmov.

Navadno gre pri podatkovnem rudarjenju za sekundarno analizo podatkov, saj so v večini primerov le-ti zbrani s povsem drugim namenom. To pomeni, da v procesu podatkovnega rudarjenja, metodologija zbiranja podatkov ne igra nobene oz. minimalno vlogo, zaradi česar se podatkovno rudarjenje razlikuje od ostalih statističnih metod

raziskovanja, saj gre pri ostalih za usmerjeno in namensko zbiranje podatkov (Hand in drugi 2001, 1–2).

4.3 KORAKI PODATKOVNEGA RUDARJENJA

Med glavne korake podatkovnega rudarjenja uvrščamo naslednja splošna opravila (Hand in drugi 2001, 11–15; Hand v Nisbet in drugi 2009, 23–24):

1. **Eksplorativna (poizvedovalna) analiza podatkov:** Cilj omenjene analize je raziskovanje novega in na začetku še neznanega področja, brez jasno razdelanih idej o tem, kaj naj bi v podatkih iskali. Gre torej za nekakšno osnovo kasnejšega podrobnejšega raziskovanja tega področja. Navadno se za to vrsto raziskovanja uporablja vizualne in interaktivne tehnike, s katerimi dobimo prvi vpogled v zbirko podatkov, ki nam omogočajo prepoznati morebitne vzorce in trende.

2. **Opisno (deskriptivno) modeliranje:** Na tej stopnji gre za vpogled v zbirko podatkov na višji ravni, ki lahko vključuje naslednje:
   - Določitev celotne verjetnostne porazdelitve podatkov (*ocena gostote*);
   - Modele, ki opisujejo relacije med spremenljivkami (*modeliranje odvisnosti*);
   - Delitev podatkov v skupine, bodisi z *metodo razvrščanja v skupine* ali *segmentacijo*. Pri slednji gre za iskanje homogenih skupin, povezane s spremenljivko, vključeno v modeliranje (npr. segment strank *zvesti*). Na drugi strani, pa je cilj metode razvrščanja v skupine, iskanje naravnih skupin.

3. **Napovedno modeliranje (klasifikacija in regresija):** Cilj je načrtovanje modela, kjer vrednost ene spremenljivke lahko napovemo iz vrednosti drugih spremenljivk. Klasifikacijo in logistično regresijo se uporablja v primeru opisnih oz. nominalnih spremenljivk (npr. spol), medtem ko so za

linearno regresijo rezervirane številske oz. kvantitativne spremenljivke (npr. starost).

4. **Odkrivanje vzorcev in pravil:** V ta korak je lahko vključeno vse, kar je povezano z iskanjem kombinacij, ki se v transakcijski zbirki podatkov pogosto pojavljajo skupaj. S pomočjo tovrstnih analiz lahko generiramo asociacijska pravila.

5. **Iskanje po vsebini:** V tem koraku na podlagi že znanega vzorca, v novi zbirki podatkov iščemo podobne vzorce.

4.4 TEHNIKE PODATKOVNEGA RUDARJENJA

Opisali bomo tehnike podatkovnega rudarjenja, ki se najpogosteje uporabljajo na področju odkrivanja zavarovalniških goljufij, in sicer odločitvena drevesa, nevronske mreže, Bayesova klasifikacija, regresija, razvrščanje v skupine in analiza povezav.

Med drugim lahko tehnike podatkovnega rudarjenja glede na način klasifikacije delimo v dve skupini, in sicer na nadzorovane in nenadzorovane tehnike modeliranja. Operacije nadzorovanih tehnik temeljijo na relacijah podatkov, ki imajo razrede in lastnosti entitet določene vnaprej. Na drugi strani pa govorimo o nenadzorovanih tehnikah takrat, ko vhodne informacije o razredih še niso znane (Nisbet in drugi 2009, 235). Tehnike opisane v nadaljevanju se glede na načine klasifikacije uvrščajo na naslednji način (Phua in drugi 2005, 5–8):

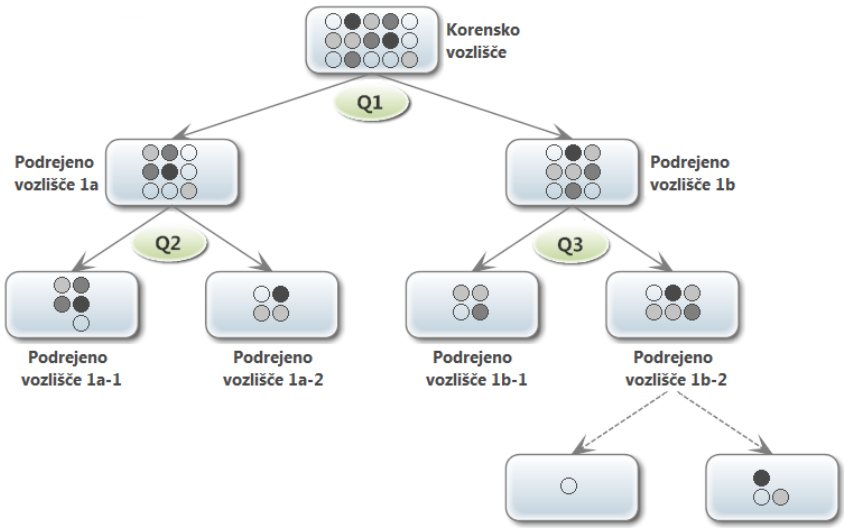
- nadzorovane tehnike: odločitvena drevesa, nevronske mreže, Bayesova klasifikacija in regresija;
- nenadzorovane tehnike: razvrščanje v skupine, analiza odkrivanja anomalij in analiza povezav.

4.4.1 ODLOČITVENA DREVESA

Odločitvena drevesa so močna in priljubljena tehnika podatkovnega rudarjenja, namenjena klasificiranju skupin podatkov in napovedovanju vzorcev. Odločitveno drevo (glej Sliko 4.1) predstavlja hierarhično skupino razmerij, organizirano v drevesno strukturo (Nisbet in drugi 2009, 241; Sherly 2012, 2).

Začetek drevesa predstavlja spremenljivka, imenovana korensko vozlišče, in se nadaljuje v najmanj dve veji, ki predstavljajo ločena razreda (če gre za kategorično korensko vozlišče) oz. specifične nize (če gre za zvezno korensko vozlišče). Veje dobimo po predhodno zastavljenem vprašanju, odgovore nanj pa razumemo kot liste drevesa. Vsak razcep vozlišča se torej nanaša na starševsko vozlišče (ang. *parent node*), ki je razdeljeno v podrejena vozlišča (ang. *child nodes*). Ta proces, ki ga imenujemo rekurzivna delitev (ang. *recursive partitioning*), se ponavlja, dokler niso izpolnjeni vsi pogoji (Nisbet in drugi 2009, 241).

Slika 4.1: Preprosto binarno odločitveno drevo



Vir: Nisbet in drugi (2009, 241).

4.4.2 NEVRONSKE MREŽE

Nevronske mreže so uporabne za različne namene, pomembne pa so predvsem na področju opisnega in napovednega podatkovnega rudarjenja. Sprva so bile na področju

strojnega učenja zasnovane kot poskus imitiranja nevrofiziologije človeških možganov s pomočjo kombinacije enostavnih računalniških elementov, imenovani umetni nevroni, v med seboj visoko povezan sistem (Giudici in Figini 2009, 76).

Umetni nevroni so osnovni elementi nevronskih mrež, saj sestavljajo njihovo arhitekturo oz. operacijske procese. Ti nevroni so organizirani v sloje treh vrst in sicer: vhodni, skriti in izhodni sloj. **Vhodni sloj** sprejema informacije le iz zunanjega okolja, vsak nevron vhodnega sloja pa se navadno ujema s pojasnjevalno spremenljivko. Njegova vloga ni izvajanje izračunov, temveč posredovanje informacij na naslednjo stopnjo. Ta stopnja predstavlja **skriti sloj**, ki ni direktno povezan z zunanjim okoljem. Ta sloj je namenjen izključno analizam, katerih naloga je prenos relacij med vhodnimi in izhodnimi spremenljivkami. **Izhodni sloj** poda končni rezultat, ki ga pošlje omrežje iz sistema. Vsi nevroni izhodnega sloja ustrezajo odzivni spremenljivki. V nevronski mreži navadno obstajata vsaj dve odzivni spremenljivki (Giudici in Figini 2009, 79; Nisbet in drugi 2009, 251).

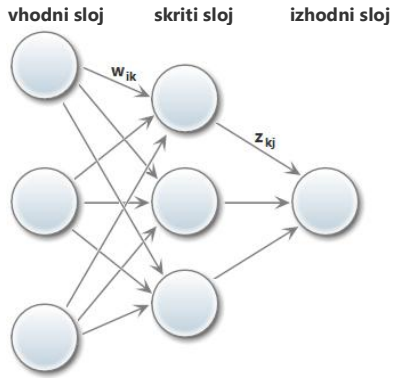
Arhitekture različnih nevronskih mrež se med seboj razlikujejo glede na stopnjo diferenciacije vhodnega in izhodnega sloja, število slojev, smer toka izračunov in vrsto povezav (Giudici in Figini 2009, 79).

Razlikujemo med enoslojnimi in večslojnimi perceptroni. Enoslojni perceptron predstavlja najenostavnejši primer arhitekture nevronske mreže, kjer se vhodni sloji ujemajo s številom izhodnih, med katerimi ni diferenciacije. Tovrstne mreže imajo $\boldsymbol{n}$ vhodnih enot ($\boldsymbol{x_1}$, …, $\boldsymbol{x_n}$), povezanih v sloj $\boldsymbol{p}$ izhodnih enot ($\boldsymbol{y_1}$, …, $\boldsymbol{y_p}$). Vsaka povezava ima svojo utež, ki jo lahko predstavimo z naslednjo matriko (Giudici in Figini 2009, 79–80):

$$\begin{bmatrix} \boldsymbol{w_{1}}1 & \cdots & \boldsymbol{w_{1j}} & \cdots & \boldsymbol{w}_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ w_{i1} & \cdots & w_{ij} & \cdots & w_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ w_{n1} & \cdots & w_{nj} & \cdots & w_{np} \end{bmatrix}, \text{ kjer je } i = 1, \dots, n \text{ in } j = 1, \dots, p.$$

w_{ij} predstavlja utež povezave med i-tim nevronom vhodnega sloja in j-tim nevronom izhodnega sloja (Giudici in Figini 2009, 80).

Slika 4.2: Arhitektura dvoslojne nevronske mreže



Vir: Nisbet in drugi (2009, 252).

Nevronsko mrežo z več kot enim slojem nevronov, ki vsebuje vsaj en skriti sloj, pa imenujemo večslojni perceptron. Dvoslojna nevronska mreža (glej Sliko 4.2) ima tako en skriti sloj, v katerem obstaja n nevronov v vhodnem, h v skritem in p v izhodnem sloju. Uteži w_{ik} ($i = 1, \dots, n$; $k = 1, \dots, h$) povezujejo vozlišča vhodnega sloja z vozlišči skritega sloja, medtem ko uteži z_{kj} ($k = 1, \dots, h$; $j = 1, \dots, p$) povezujejo vozlišča skritega sloja z vozlišči izhodnega sloja. Nevroni skritega sloja prejemajo informacije iz vhodnega sloja, utežene z utežmi w_{ik}, katerih izid je $h_k = f(x, w_k)$, kjer f predstavlja aktivacijsko funkcijo enot v skritem sloju. Na koncu nevroni izhodnega sloja prejmejo informacije iz skritega sloja, utežene z utežmi z_{kj}, s končnim izidom mreže $y_j = g(h, z_j)$. Končni izid nevrona j v izhodnem sloju tako predstavlja naslednja enačba (Giudici in Figini 2009, 80):

$$y_j = g\left(\sum_k h_k z_{kj}\right) = g\left(\sum_k z_{kj} f\left(\sum_i x_i w_{ik}\right)\right).$$

4.4.3 BAYESOVA KLASIFIKACIJA

Bayesova klasifikacija je metoda podatkovnega rudarjenja, ki temelji na podlagi Bayesovega izreka. S pomočjo te klasifikacije lahko napovemo verjetnost pripadnosti posameznega elementa določeni skupini. Razlikujemo med naivnim Bayesovim klasifikatorjem in Bayesovimi verjetnostnimi mrežami (Sherly 2012, 4).

Naivni Bayesov klasifikator predpostavlja, da je vpliv vrednosti atributov na dani razred neodvisen od vrednosti drugih atributov, zaradi česar je metoda tudi poimenovana kot naivna. Kljub njegovi premočni predpostavki, pa deluje precej dobro na različnih zbirkah podatkov, saj na njem delujejo tako številske (kvantitativne) kot nominalne spremenljivke. Njegova prednost je tudi preprostost in hitrost delovanja, saj bistveno poenostavi proces klasifikacije, glede na to, da izračuna razred pogojne gostote posamično za vsako spremenljivko, kar pomeni, da reducira večrazsežnostne naloge v več enorazsežnostne (Nisbet in drugi 2009, 256; Sherly 2012, 4).

Na drugi strani pa **Bayesove verjetnostne mreže** upoštevajo tudi odvisnost med spremenljivkami. Verjetnostna mreža je definirana z dvema komponentama, in sicer z neposrednim acikličnim diagramom in nizom pogojnih verjetnosti. Posamezno vozlišče diagrama predstavlja slučajno spremenljivko, njihove relacije pa so sledeče. Če je vozlišče Y usmerjeno v vozlišče Z, Y predstavlja neposrednega predhodnika (»starš«), Z pa je njegov naslednik (»potomec«) (Sherly 2012, 4).

4.4.4 REGRESIJA

Regresija, kot tehnika podatkovnega rudarjenja, predstavlja močno in elegantno metodo ocenjevanja vrednosti spremenljivk (Larose 2006, 33), njeni koraki pa so naslednji (Škulj 2006, 4):

- postavitev regresijskega modela oz. teoretičnih predpostavk o odvisnosti med spremenljivkami:
    - izbira odvisne spremenljivke in tistih, ki nanjo pomembno vplivajo;
    - določitev vrste odvisnosti med izbranimi spremenljivkami, kar je predpogoj za naslednji korak;
- testiranje na vzorcu oz. ocena parametrov modela
    - ta temelji na določenih predpostavkah, ki jih je potrebno preveriti;
- dvojna vloga regresijskega modela:
    - pojasnjevalna: parametri nam lahko pojasnijo, kako posamezna neodvisna spremenljivka vpliva na odvisno;

- napovedovalna: regresijski model nam iz vrednosti neodvisne lahko napove vrednost odvisne spremenljivke.

Poznamo več tipov regresije, v nadaljevanju bomo opisali linearno in logistično.

Linearna regresija velja za glavno metodo v statistiki, s katero poskušamo izraziti odvisno spremenljivko (y) kot linearno kombinacijo ene ali več neodvisnih spremenljivk (x_1, …, x_k), s predhodno določenimi utežmi oz. regresijski parametri (w_0, w_1, …, w_k) in napako (e). Oblika regresijske enačbe je tako sledeča (Witten in drugi 2011, 124):

- za bivariatno linearno regresijo: $y = w_0 + w_1 x + e$, kjer w_0 predstavlja sečišče osi y ($x = 0$) in w_1 nagib regresijske premice v razsevnem grafikonu;
- za multivariatno linearno regresijo: $y = w_0 + \ w_1 x_1 + \ldots + w_k x_k + e$, kjer uteži oz. regresijske parametre imenujemo tudi **regresijski koeficienti –** w_0 je konstantni člen, w_1, …, w_k pa parcialni regresijski koeficienti.

Glavne naloge linearne regresije pa so naslednje (Nisbet in drugi 2009, 264):

- ugotoviti, če med dvema ali več spremenljivkami obstaja odvisnost;
- opisati za kakšno odvisnost gre, če le-ta obstaja;
- oceniti oz. izmeriti natančnost te odvisnosti;
- ovrednotiti relativne doprinose posamezne spremenljivke.

Linearna regresija je tako namenjena analizi številskih (kvantitativnih) spremenljivk, vendar se v statistiki dostikrat srečamo tudi z odvisnimi nominalnimi spremenljivkami. Za tovrstne odvisne spremenljivke uporaba te metode ni primerna, ampak je za te primere rezervirana logistična regresija (Larose 2006, 155).

Logistična regresija je ena od metod nelinearne regresije, namenjena klasifikaciji. Gre za metodo zelo podobno linearni regresiji, saj opisuje relacijo med nominalno spremenljivko (odvisna spremenljivka) in enim ali več prediktorji (neodvisne spremenljivke). Kot pove že sama definicija, gre za operiranje z binarnimi spremenljivkami, ki zavzemajo samo dve vrednosti, in sicer 0 in 1 (Larose 2006, 155; Nisbet in drugi 2009, 272).

4.4.5 RAZVRŠČANJE V SKUPINE

Razvrščanje v skupine ali grozdenje (ang. *clustering*) je metoda razvrščanja zapisov oz. enot v skupine, v katerih je medsebojna podobnost med zapisi maksimizirana, na drugi strani pa je podobnost zapisov izven skupine minimizirana. Razvrščanje v skupine je metoda, katere namen ni klasificiranje, ocenjevanje ali napovedovanje vrednosti ciljne spremenljivke, temveč segmentiranje celotne zbirke podatkov v relativno homogene skupine (Larose 2005, 16). Ločimo med hierarhičnimi in nehierarhičnimi metodami razvrščanja v skupine.

V primeru **hierarhičnega razvrščanja v skupine**, delitev le-teh grafično prikažemo z drevesno strukturo, ki jo imenujemo hierarhično drevo ali dendrogram. Enote predstavljajo liste drevesa, točke združitve pa sestavljene skupine. Nivo združenja ali višina točke je sorazmerna meri različnosti med skupinama (Ferligoj 1989, 68).

Med ene najpogostejših metod hierarhičnega razvrščanja v skupine uvrščamo (Ferligoj 1989, 76–77):

- minimalno metodo (enojna povezanost), kjer je razdalja med dvema skupinama minimalna;
    - združita se tisti skupini, med katerima obstaja največja povezanost oz. najmanjša različnost;
    - primerna je pri razkrivanju dolgih verižnih struktur in neuporabna pri neizrazito ločenih skupinah;
- maksimalno metodo (polna povezanost), ki je osredotočena na razkrivanje znotraj kohezivnih skupin;
    - primerna je za razkrivanje okroglih in prekrivajočih skupin;
- Wardovo metodo, ki upošteva največjo notranjo kohezivnost in najdaljšo zunanjo razdaljo
    - gre za najpogosteje uporabljeno metodo v družboslovju
    - najprimernejša je za eliptično strukturirane podatke.

Glavna razlika med hierarhičnimi in nehierarhičnimi metodami je v tem, da potrebno pri slednjih število skupin določiti vnaprej. **Nehierarhično razvrščanje v skupine**

imenujemo tudi metode razvrščanja v skupine z lokalno optimizacijo, saj pregledajo le del množice razvrstitev. Pri teh metodah različnosti med enotami prikažemo na dva načina, in sicer tako, da podamo matriko različnosti med enotami ali pa različnost med enotami računamo sproti (Ferligoj 1989, 87–88).

V okviru nehierarhičnega razvrščanja v skupine poznamo metodo voditeljev in metodo prestavljanj. Metoda voditeljev je iteracijska metoda, kjer moramo število skupin, v katere bodo enote razvrščene, določiti že v začetku. Postopek iteracije se prične z vnaprej podano množico predstavnikov posameznih skupin, ki jih imenujemo voditelji. Metoda nato voditeljem priredi najbližje enote in poišče centroide novonastale skupine, ki predstavljajo nove voditelje, v nadaljevanju voditeljem zopet priredi najbližje enote, itd. Ko se nova in predhodna množica voditeljev med seboj več ne razlikujeta, je postopek iteracije končan (Ferligoj 1989, 93).

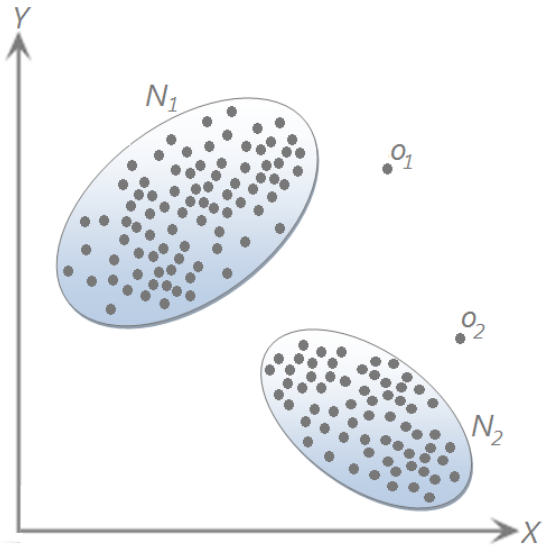
Metoda prestavljanj temelji na že izračunanih različnostih med enotami in se jo uporablja za razvrščanje v skupine, kjer je v zbirki podatkov nekaj sto enot. Postopek se začne z določitvijo dopustne razvrstitve. Če med tekočo razvrstitvijo in sosednjimi razvrstitvami obstaja nova razvrstitev, za katero velja, da je manjša ali enaka od predhodne, se postopek pomakne v novo razvrstitev. Postopek se ponavlja, dokler se lahko (Ferligoj 1989, 88–89).

4.4.6 ANALIZA ODKRIVANJA ANOMALIJ

Analiza odkrivanja anomalij je tehnika, ki temelji na osnovi razvrščanja enot v skupine, torej jo uvrščamo med tehnike nenadzorovanega podatkovnega rudarjenja. Bistvo analize odkrivanje anomalij je iskanje tistih vzorcev v zbirki podatkov, ki glede na ostale enote odstopajo od pričakovanega vedenja. Gre torej za odkrivanje neskladij, napak, posebnosti oz. izjem (Chandola in drugi 2009, 1).

Chandola in drugi (2009, 2) anomalije definirajo kot vzorce, ki med ostalimi podatki ne ustrezajo normalnemu vedenju. Grafično jih lahko predstavimo v razsevnem grafikonu (glej Sliko 4.3), kjer N_1 in N_2 predstavljata skupini enot normalnega vedenja, saj je večina teh razvrščenih prav tam, $o_1$in o_2 pa anomaliji, ki ležita daleč stran od ostalih enot.

Slika 4.3: Preprost razsevni grafikon anomalij



Vir: Chandola in drugi (2009, 2).

Za izvedbo analize odkrivanja anomalij potrebujemo zbirko vhodni podatkov, ki lahko predstavljajo entiteto, vzorec, dogodek, ipd, ki so opisani z različnimi atributi. V našem primeru, ko se z analizo odkrivanja anomalij ukvarjamo v okviru statistike, atribute tako predstavljajo nominalne in zvezne spremenljivke (Chandola in drugi 2009, 6).

Chandola in drugi (2009, 7) predstavijo tudi različne vrste anomalij, to so:

- **Točkovne anomalije** predstavljajo najbolj osnovno vrsto anomalij. Pod to besedno zvezo razumemo tisto vrsto anomalije, kjer lahko posamezno enoto glede na ostale primere obravnavamo kot anomalijo. Na primer, oseba glede na ostale, dvigne enormno vsoto denarja.
- **Anomalije konteksta** so tiste, kjer je enota anomalija le v posebnih okoliščinah, kot sicer. Na primer, 10 °C v zimskem času predstavlja normalno temperaturo, medtem ko bi v poletnem predstavljala anomalijo.
- O **anomalijah skupine** pa govorimo takrat, ko je glede na ostale enote anomalija manjša skupina enot. Na primer, med večjo skupino oseb, so tudi tri živali.

Zelo pomemben vidik analize odkrivanja anomalij je, na kakšen način bodo rezultati anomalij predstavljeni, in sicer imamo dve možnosti. Predstavimo jih z *ocenami*, kjer je posamezni enoti dodeljena vrednost, ki pove v kolikšni meri enoto lahko obravnavamo kot anomalijo. Drugi način pa je *označevanje*, in sicer tako, da posamezno enoto označimo kot anomalijo ali normalno vedenje (Chandola in drugi 2009, 10).

4.4.7 ANALIZA POVEZAV

Analizo povezav je tehnika nenadzorovanega podatkovnega rudarjenja, katere cilj je iskanje razmerij med enotami (vozlišči) in skritih povezav med vzorci, za katere se zdi, da so med seboj nepovezani. Gre torej za povezovanje skupin in aktivnost v odnosu do določenega vedenja. Vozlišča, ki so večvrstna, pa lahko predstavljajo ljudi, dogodke ali sredstva (Phua in drugi 2005, 8; Zacharias in drugi 2008, 233; Nisbet in drugi 2009, 351).

Zlasti na področju kazenskega pregona, so na začetku analitiki z besedno zvezo *analiza povezav* opisovali pristope, ki so jim omogočili prikaze in iskanje vzrokov za povezave med vozlišči. Danes pa analizo povezav razumemo kot novo področje, ki se je ustalilo znotraj računalništva in statistike (Zacharias in drugi 2008, 231–233).

Temelji na teoriji grafov in analizi vzorcev, poznamo pa jo tudi iz področja diskretne matematike. Proces izvajanja analize povezav se imenuje odkrivanje povezav (ang. *link discovery*), samo tehniko pa se uporablja, ko predhodni vzorci vedenja niso poznani. V začetku torej načrtujemo izhodiščni model, ki predstavlja normalno vedenje, nato pa poskušamo odkriti vzorce, ki nakazujejo večja odstopanja od predhodno določene norme (Bolton in Hand 2002, 237; Nisbet in drugi 2009, 351).

Za razliko od standardnih tehnik, analiza povezav zahteva popolnoma drugačen pristop odkrivanja znanja, tako v metodah, kot v pristopih ocenjevanja rezultatov algoritma. Odstopanja se kažejo v naslednjih lastnostih obravnavane tehnike (Schroeder in drugi 2007, 842):

- Gre za operiranje s heterogenimi podatki, saj so v proces iskanja povezav in vzorcev iz podatkov vključeni ljudje, organizacije, enote, aktivnosti in dogodki. Vsaka od teh entitet ima lasten nabor atributov, med katerimi lahko obstaja mnogo relacij.
- Za razliko od standardnih tehnik podatkovnega rudarjenja, kjer spremenljivke predstavljajo vozlišča, povezave pa statistične relacije med spremenljivkami, so pri analizi povezav entitete obravnavane kot vozlišča, povezave pa kot relacije med njimi.

Dostikrat je za samo iskanje povezav potrebna predpriprava ogromne količine podatkov (Goldberg in Wong v Zacharias in drugi 2008, 231). Kljub temu pa ne glede na zahtevnost procesa, kombinacija napredne tehnike za obdelavo podatkov in strojnega učenja omogoči hitro predelovanje in transformacijo zbirke podatkov, kot tudi pridobivanje vzorcev. Osnova tega procesa je preučevanje poti, ki nam pomaga identificirati anomalije teh poti in vzorce, ki jih lahko sklepamo iz njih. Te vzorce lahko kasneje uporabimo za njihovo pojasnjevanje ali napovedovanje nadaljnjih povezav (Zacharias in drugi 2008, 231–232).

Znotraj analize povezav obstajajo trije temeljni koncepti (Zacharias in drugi 2008, 232–233):

- *podobnost ali razdalja*: Razdalja se nanaša na pojasnjevanje povezljivosti vozišč pod predpostavko, da se bodo tista, ki so si podobna ali blizu, med seboj tudi povezala na podoben način.
- *interakcija*: Skupine in enote, ki navadno predstavljajo vozlišča, so v medsebojni interakciji.
- *povezovalna funkcija*: Na osnovi vrste podobnosti povezav se le-te pretvorijo v njihovo prisotnost, odsotnost ali vlogo uteževanja. Preko povezovalne funkcije pa se kažejo tudi ključne razlike med algoritmi.

4.5 POMEN PODATKOVNEGA RUDARJENJA NA PODROČJU ODKRIVANJA ZAVAROVALNIŠKIH GOLJUFIJ

Uporaba podatkovnega rudarjenja je, z vidika uspešnosti, učinkovit način odkrivanja potencialnih goljufivih zavarovalniških primerov (Phua in drugi 2005, 1), saj velja za napredno analitično orodje, ki je lahko zavarovalnicam v veliko pomoč pri ključnih poslovnih odločitvah in preprečevanju goljufij (Sharma in Panigrahi 2012, 38).

Vendar pa navkljub temu, da podatkovno rudarjenje predstavlja primerno in učinkovito metodo v procesu odkrivanja zavarovalniških goljufij, obstajata dve glavni kritiki njegove uporabe na tem področju (Phua in drugi 2005, 1):

- redkost javno dostopnih podatkov za izvedbo analiz oz. testiranj modelov in
- pomanjkanje informacij o preverjenih oz. dobro raziskanih metodah in tehnikah podatkovnega rudarjenja na področju odkrivanja zavarovalniških goljufij.

Nisbeth in drugi (2009, 350) podajo zelo dobro pojasnilo, zakaj so ti podatki javno dostopni v tako majhni meri in v čem je razlog za pomanjkanje informacij o metodah in tehnikah. Zavarovalnice se seveda otepajo razkritju njihovih načinov in tehnik odkrivanja ter odvračanja od goljufij, saj bi na drugi strani potemtakem potencialni goljufi lahko našli svoje načine izogibanja tem preprečevalnim ukrepom.

Z vidika odkrivanja zavarovalniških goljufij, se analiza povezav navezuje na identifikacijo, analizo in vizualizacijo relacij ter povezav med sumljivimi entitetami (osebe, organizacije, vozila, lokacije, neetično vedenje, ipd). S pomočjo analize povezav, lahko z določitvijo asociacijskih pravil, ki povezujejo entitete, kot so osumljenec in žrtev ali oškodovanec v primeru neetičnega vedenja, pridemo do informacij kot so motiv tovrstnega dejanja in le-to tudi razkrijemo (Schroeder in drugi 2007, 842).

5 METODOLOGIJA

V prvem koraku se bomo lotili pregleda literature s področja odkrivanja zavarovalniških goljufij s statističnimi metodami oz. tehnikami podatkovnega rudarjenja, kar bo predstavljalo sekundarno analizo strokovne literature. Pregled obstoječe teorije in praktičnih primerov bo osnova za naslednji korak, t.j. določitev kriterijev pri izbiri modela, kjer bo ob primerjavi različnih modelov, podana tudi utemeljitev izbire metode analize povezav. Temu bo sledila priprava podatkov, na katerih bo izvedena implementacija in testiranje uspešnosti modela z računalniškim orodjem SPSS Modeler. Na koncu pa bodo predstavljeni še kriteriji učinkovitosti modela in kritična evalvacija končnega rezultata z utemeljitvijo obnašanja modela.

5.1 PREGLED MODELOV ZA ODKRIVANJE ZAVAROVALNIŠKIH GOLJUFIJ

Za začetek si bomo na podlagi strokovne literature ogledali primere modelov, ki so jih avtorji uporabili pri odkrivanju zavarovalniških goljufij.

5.1.1 MODEL PRI UPORABI BAYESOVE KLASIFIKACIJE IN ODLOČITVENIH DREVES

Bhowmik (2011) je za odkrivanje zavarovalniških goljufij na področju avtomobilskih zavarovanj za modeliranje uporabil Bayesovo klasifikacijo in odločitvena drevesa.

Naivna Bayesova klasifikacija

1. Oblikovanje Bayesovih omrežij: V prvem koraku je Bhowmik (2011, 157) za potrebe odkrivanja goljufij oblikoval dve Bayesovi omrežji. Pri prvem gre za postavitev modela vedenja pod predpostavko, da je oseba oz. v tem primeru voznik goljufiv, drugi model pa predpostavlja, da je njegovo vedenje legalno. Postavitev omrežja, ki predstavlja goljufije, temelji na uporabi strokovnega znanja, medtem ko drugo omrežje temelji na skupini legalnih voznikov. Do verjetnosti merjenega E pridemo, ko so v obe omrežji dodani pogoji. Pojasnjeno z drugimi besedami, to pomeni, da na podlagi dodanih pogojev lahko dobimo ocene o tem, do katere mere opazovano voznikovo vedenje velja za tipično

goljufivo ali legalno. Te ocene so poimenovane kot $P(E|\ rezultat = legalno)$ in $P(E|\ rezultat = goljufija)$.

Ker imamo le dva možna izida, torej *legalno* ali *goljufivo* dejanje, velja naslednje $P(rezultat = goljufija)$ in $P(rezultat = legalno) = 1 - P(rezultat = goljufija)$. Z uporabo Bayesove formule lahko izračunamo verjetnost goljufije, pogojno na meritev *E*:

$$P(rezultat = goljufija \mid E) = \frac{P(rezultat{=}goljufija)\ P(E|rezultat{=}goljufija)}{P(E)} \quad (5.1)$$

kjer je $P(rezultat = goljufija)$ apriorna verjetnost, $P(E|\ rezultat = goljufija)$ pogojna verjetnost, $P(E)$ pa verjetnost dogodka.

2. Uporaba: V drugem koraku Bhowmik (2011, 157–158) predstavi Bayesov algoritem, katerega namen je napovedovanje goljufivih dejanj. Za izračun verjetnosti, da bo izhodni atribut *goljufija* dejanje, uporabimo osnovni enačbo (5.1).

Apriorna verjetnost $P(rezultat = goljufija)$ je v tem primeru izpeljana po naslednjem principu $P(goljufija) = d_i \,/\, d$, kjer d predstavlja celotno populacijo primerov dejanj, d_i pa število goljufivih.

Pogojna verjetnost $P(E|\ rezultat = goljufija)$ temelji na poenostavljeni predpostavki za naivni Bayesov klasifikator, ki predpostavi, da med atributi ni povezanosti, in sledi iz pravila za produkt neodvisnih dogodkov:

$P(E|\ rezultat = goljufija)$ = $\prod_{k=1}^{n} P\ (x_k \mid rezultat = goljufija)$, , kjer x_k predstavlja vrednosti atributov za $(A_1,\ A_2, \ldots, A_n)$.

Verjetnost dogodka $P(E)$ izračunamo po formuli za popolno verjetnost, in sicer:

$P(E) = P(rezultat = goljufija)\ P(E|\ rezultat = goljufija) + P(rezultat = legalno)\ P(E|\ rezultat = legalno)$.

Tako lahko na osnovi obstoječe zbirke podatkov ocenimo verjetnosti $P(x_k \mid rezultat = goljufija)$ za primere goljufivih dejanj, in sicer po naslednjem principu:

$P(x_k \mid rezultat = goljufija)$ = d_{ik} / d_i, kjer d_i predstavlja število izhodiščnih zapisov za primere *goljufij*, d_{ik} pa število izhodiščnih zapisov za razrede *goljufij*, katerih atribute predstavlja x_k.

Podobna operacija je izvedena tudi na primerih legalnih dejanj, kjer njihovo verjetnost izračunamo kot $P(x_k \mid rezultat = legalno)$ = d_{ik} / d_i, kjer je d_i število izhodiščnih zapisov za primere *legalnih* dejanj in d_{ik} število izhodiščnih zapisov za razrede *legalnih* dejanj, katerih atribute predstavlja x_k.

Iz teh verjetnosti je sedaj po predhodnih formulah mogoče izračunati potencialne verjetnosti za goljufiva oz. legalna dejanja.

Odločitvena drevesa

1. Osnovni algoritem: Odločitveno drevo se začne z enim vozliščem m, ki predstavlja celotno zbirko podatkov ali prostor R_m z N_m opazovanimi enotami ali primeri, in je definirano kot:

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k),$$

kjer je k del razreda primerov vozlišča m. Primere vozlišča m razporedimo v razred $k_m = arg\ max_k\ \hat{p}_{mk}$, ki predstavlja glavni razred vozlišča m (Hastie in drugi 2009, 309).

V prvem koraku Bhowmik (2011, 158) opiše osnovni algoritem za odločitvena drevesa, kjer sta možna dva izida – *goljufija* in *legalno* dejanje. Če so primeri istega tipa, torej *goljufija*, vozlišče postane list in je označen kot *goljufija*.

V nasprotnem primeru za meritev stopnje ločevanja primerov, algoritem lahko uporabi eno od spodnjih metod:

- križno entropijo, ki jo izračunamo po naslednji formuli:

$$ent = -\sum_{k=1}^{K} \hat{p}_{mk}\ log\ \hat{p}_{mk}$$

- Ginijev indeks:

$$GI = \sum_{k \neq k'} \hat{p}_{mk}\ \hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk}\ (1 - \hat{p}_{mk})$$

- ali klasifikacijsko napako, ki atribute klasificirajo v svoje razrede:

$$KN = \frac{1}{N_m} \sum_{i \in R_m} I(y_i \neq k(m)) = 1 - \hat{p}_{mk(m)}.$$

V našem primeru sta primernejši metodi križna entropija in Ginijev indeks, saj sta bolj odvedljivi, in tako primernejši za numerične spremenljivke. Obenem pa sta tudi bolj občutljivi na spremembe verjetnosti izidov (Hastie in drugi 2009, 309). Seveda pa bi bile v model lahko vključene tudi druge metode odločitvenih dreves (dvojiški, χ^2, ipd).

2. Algoritem: V naslednjem sklopu korakov Bhowmik (2011, 159) prikaže algoritem na podlagi entropije, katerega naloga je klasificiranje danega primera, in sicer kot:

$$E(goljufija, legalno) = - (goljufiviPrimeri \ / \ Primeri) \ log_2 \ (goljufiviPrimeri \ / \ Primeri) - (legalniPrimeri \ / \ Primeri) \ log_2 \ (legalniPrimeri \ / \ Primeri)$$

ter entropijo glede na atribut:

$$E(A) = \{(goljufiviAtributi \ / Primeri) + (legalniAtributi \ / Primeri)\} * \{E(goljufiviAtributi, legalniAtributi)\}$$

Vrednost (oz. prispevek k informacijam) atributa je izračunana kot:

$$pridobitev(atr) = (informacija \ pred \ razdelitvijo) - (informacija \ po \ razdelitvi).$$

Pričakovana redukcija entropije je:

$$pridobitev(atr) = entropija \ nadrejene \ tabele - E(atr).$$

Algoritem računa, koliko informacij je pridobil vsak atribut. Tisti, ki doseže najvišjo stopnjo, je izbran za testni atribut.

Vsakič, ko je znana nova vrednost testnega atributa, je ustvarjena tudi nova veja odločitvenega drevesa. Particije drevesa se oblikujejo v procesu iteracij, ki deluje na podlagi algoritma. Takoj, ko atribut zazna vozlišče, le-ta ni več upoštevan v nobenem naslednjem vozlišču.

Iterativni proces se konča, ko je zadoščen en od pogojev:

- vsi primeri danega vozlišča pripadajo istemu razredu;
- atributov, na katerem bi se vzorci nadalje delili, ni več na voljo
- predstavniki oz. kandidati za nadaljevanje oblikovanja drevesa oz. vej, niso več na voljo.

5.1.2 MODELI PRI UPORABI REGRESIJE

Belhadji in drugi (2000) v svojem delu predstavijo model za odkrivanje zavarovalniških goljufij na področju avtomobilskih zavarovanj s pomočjo regresije. Model opisujemo v nadaljevanju.

V študiji je sodelovalo 18 kanadskih zavarovalniških podjetij, kar predstavlja 70% celotnega zavarovalniškega trga v Kanadi. Vsaka zavarovalnica je prejela vprašalnike o karakteristikah zaključenih zavarovalniških primerov. Na koncu je zbirka podatkov, na kateri je bilo izvedeno modeliranje, obsegala informacije o 2509 zaključenih zavarovalniških primerih, med katerimi je obstajalo 116 sumljivih primerov, kar predstavlja 4,62% vseh primerov in 18 potrjenih goljufij oz. 0,72% vseh primerov (Belhadji in drugi 2000, 518–519).

1. Indikatorji: Avtorji se najprej osredotočijo na določitev niza indikatorjev, ki so občutljivi na goljufive in sumljive primere. Izbor indikatorjev na koncu tudi utemeljijo.

(I.) *Izbor indikatorjev:* Izbira indikatorjev, ki so v procesu odkrivanja goljufij bolj pomembni od ostalih, je temeljila na pregledu obstoječe literature. Avtorji študije so tako izbirali med 50. indikatorji, ki so znotraj vprašalnikov predstavljala posamezna vprašanja.

(I.) *Kriteriji za omejitev števila izbranih indikatorjev, ki bodo vključeni v model:* Za vsak indikator je bila izračunana pogojna verjetnost, ki bi lahko nakazovala goljufijo. V spodnji tabeli (glej Tabelo 5.1) je prikazan primer tabele, ki je osnova za izračun pogojnih verjetnosti.

Tabela 5.1: Primer tabele za izračun pogojne verjetnosti goljufije

Primer	Indikator	Vsota vseh zapisov	Vsota sumljivi in dokazani	Razmerje (%)
1	ni_zapisnika	363	34	9,37
2	min_škoda	74	6	8,11
3,4	nekons	57	18	31,58
5	kraja	32	12	37,50
⋮	⋮	⋮	⋮	⋮

Vir: Belhadji in drugi (2000, 521).

Za vse ocene indikatorjev iz tabele so bili izračunani standardni odkloni in sicer kot: $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, kjer $\hat{p}$ predstavlja vrednost ocene indikatorjev in n frekvenco posameznega indikatorja. Enačba za izračun intervala zaupanja:

$\left[\hat{p} - Z_{a/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}};\ p + Z_{a/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$; kjer $Z_{a/2}$ predstavlja $1 - a/2$-ti kvantil normalne porazdelitve. Standardni odklon za oceno tretjega indikatorja, je tako: $\sqrt{\frac{(0{,}316)(0{,}684)}{57}} = 0{,}062.$

95% interval zaupanja za oceno obravnavanega indikatorja pa naslednji:

$[0{,}316 - 1{,}96(0{,}062);\ 0{,}316 + 1{,}96(0{,}062)]$ ali $[19{,}5\%;\ \ 43{,}7\%)]$.

Na koncu so bili izločeni vsi indikatorji, ki so imeli število zapisov manjše ali enako 15 (tretji stolpec v Tabeli 5.1)

(II.) *Regresijski model in rezultati:* V študiji je uporabljen probit model, ki ga izrazimo na naslednji način:

$y_i^* = b^{'}x_i + u_i,$

kjer je y_i^* odvisna spremenljivka, x_i predstavlja indikatorje, $b^{'}$ pa regresijske parametre. Binarna odvisna spremenljivka y_i je definirana kot:

$y = 1$, če $y_i^* > 0$, drugače $y = 0$.

Na obravnavanem primeru to pomeni naslednje: $y = 1$, če je zapis ocenjen kot sumljiv ali goljufiv, drugače $y = 0$. Iz tega sledi:

$$\mathrm{P}(y_i^* = 1) = P(u_i > - b^{'} x_i)$$

$$= - \mathrm{F}(- b^{'} x_i);$$ kjer je F porazdelitvena funkcija u. Funkcija verjetja je tako naslednja: $L = \prod_{y_i=o} \mathrm{F}(- b^{'} x_i) \prod_{y_i=1} [1 - \mathrm{F}(- b^{'} x_i)]$, ki jo uvedemo za izračun goljufivih primerov.

u_i se v probit modelu porazdeljuje normalno po $N(0, \acute{o}^2)$. V tem primeru:

$$\mathrm{F}(- b' x_i) = \int_{-\infty}^{-b' \frac{x_i}{\acute{o}}} \frac{1}{(2đ)^{1/2}} \exp\left(-\frac{t^2}{2}\right) dt.$$

Avtorji so dobljene rezultate probit regresije uporabili kot osnovo za izbor indikatorjev, ki so jih kasneje uporabili za izračun verjetnosti goljufije. Uporabljeni pa so bili le statistično značilni indikatorji, in sicer 23 indikatorjev od 50.

2. Model, kjer so pomembne le predpostavke likvidatorjev: V tem koraku Belhadji in drugi (2000, 525) najprej med seboj primerjajo indikatorje modela, z dobljenimi rezultati iz vprašalnikov, ki so jih izpolnile zavarovalnice. Kasneje predstavijo primere odločitev o tem, ali je ponovno oz. nadaljnje preiskovanje še potrebno.

(I.) *Primerjava rezultatov:* Napovedi modela avtorji primerjajo z rezultati iz vprašalnikov, ki predstavljajo predpostavke likvidatorjev, te veljajo tudi za končne, kar pomeni, da v vzorcu goljufivi ali sumljivi primeri ne obstajajo (Belhadji in drugi 2000, 525). Kot predstavljeno v uvodu, vemo, da so likvidatorji izmed 2509 primerov našli 134 sumljivih in goljufivih, kar skupno predstavlja 5,34%.

V naslednjem koraku so v študiji predpostavili mejno vrednost za 10%, kar pomeni, da model ustvari tolikšen vzorec, ki ustreza deležu primerov, ki nosijo vrednosti spremenljivke znotraj te meje, to pa predstavlja 296 primerov. Med temi primeri je tako 93 goljufivih in/ali sumljivih, s stopnjo pravilne napovedi 31,42% (ali $93/296$). Število primerov, ki jih je model prepoznal kot legalne, ustreza 2213 (ali $2509 - 296$), med katerimi je s strani likvidatorjev kot legalnih označenih 2172 primerov, kar predstavlja

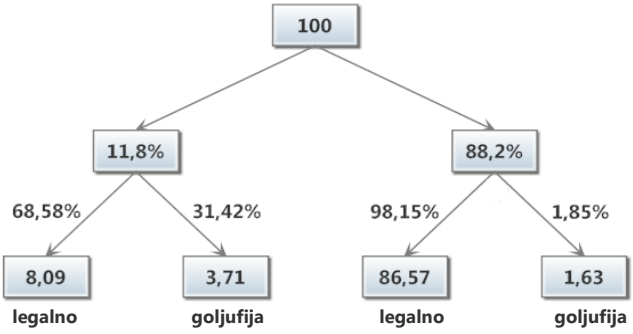
98,15% stopnjo natančnosti (ali $2172/2213$). Stopnja natančnosti je tako osnova za določitev razmerja, ki določa, kako velik mora biti vzorec – podjetje bo tako v pregled izbralo določen odstotek (X) primerov, na osnovi naslednje enačbe (Belhadji in drugi 2000, 526):

á . $X\%$ + $(1 - \hat{a})(1 - X) = 5{,}34\%$, kjer je:

- á stopnja natančnosti v % za goljufive primere, ki jih je zaznal model – torej, če je mejna vrednost za 10%, je stopnja natančnosti enaka 31,42%.
- â pa predstavlja stopnjo natančnosti v % za legalne primere, ki jih je generiral model. V primeru, da je mejna vrednost za 10%, pomeni, da je stopnja natančnosti enaka 98,15%.
- á . $X\%$ izraža odstotek goljufij (kot jih je predvidel model sam), ki obstaja med izbranimi primeri (kjer je verjetnost presega mejne vrednosti)
- $(1 - \hat{a})(1 - X)$ izraža odstotek goljufij, ki so jih zaznali preiskovalci, vendar je manjši od izbrane mejne vrednosti (neizbrani primeri za pregled). Torej bi ob mejni vrednosti za 10%, bilo za ponovni pregled primernih 11,8% primerov.

(II.) *Odločitev o ponovnem preiskovanju:* V nadaljevanju predstavijo model za odločitev o ponovnem preiskovanju. Glede na zgodnje podatke, bi na primeru 100 zavarovalniških zahtevkov dobili rezultate prikazane na spodnji sliki (glej Sliko 5.1):

Slika 5.1: Rezultati pri mejni vrednosti za 10%



Vir: Belhadji in drugi (2000, 527).

Kjer je odstotek goljufije enak 5,34% ($3{,}71\% + 1{,}63\%$), odstotek natančnosti za primere goljufij 31,42% in odstotek odkritih primerov 69,4% ($3{,}71\%$ / $5{,}34\%$). Rezultati za ostale mejne vrednosti pa so predstavili v tabelo (glej Tabelo 5.2).

Tabela 5.2: Odstotki natančnosti in odkritih primerov za različne mejne vrednosti

Verjetnost	Odstotek vzorca	Odstotek natančnosti	Odstotek odkritih primerov
10%	11,8	31,42	69,4
15%	8,61	38,43	61,94
20%	6,7	42,86	53,73
25%	4,9	50,41	46,27
30%	14,11	53,4	41,04
35%	13,07	61,04	35,07
40%	2,75	62,32	32,09
45%	1,99	62	23,13
50%	1,79	60	20,15
55%	1,63	58,54	17,91
60%	1,16	58,62	12,69
65%	0,8	55	8,21
70%	0,6	53,33	5,97
75%	0,48	58,33	5,22
80%	0,44	54,55	4,48
85%	0,32	62,5	3,73
90%	0,2	80	2,99

Vir: Belhadji in drugi (2000, 529).

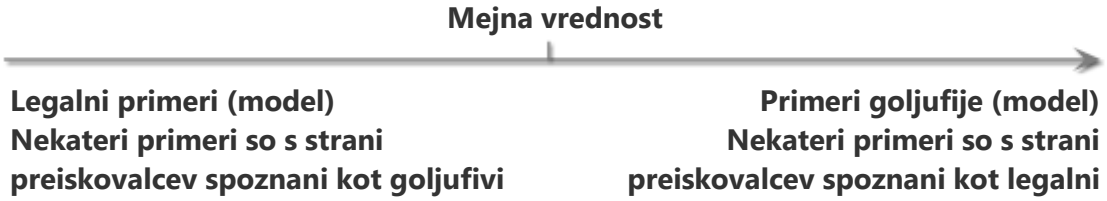
Belhadji in drugi (2000, 529) pridejo do ugotovitve, če se pri odločitvi o ponovnem preiskovanju upošteva samo odstotek natančnosti, bi bilo takih primerov zelo malo. Kar seveda podjetjem predstavlja prednost te metode, saj imajo, s ponovnim preiskovanjem manjšega števila primerov, nižje stroške. Ugotovili so tudi, da med zaznavanjem goljufije in natančnostjo obstaja povezanost – višja kot je mejna vrednost, boljša je natančnost in manjša stopnja zaznavanja.

(III.) *Primeri, ko model napoveduje pravilno:* Obstajajo tri situacije, ko model napove pravilni izid (Belhadji in drugi 2000, 531):

- Model in preiskovalci zaznajo enak delež goljufivih oz. legalnih primerov. V okviru modela so to tisti primeri, kjer je verjetnost goljufije višja od mejne vrednosti.
- Situacija, ko določen delež goljufivih primerov, ki jih zazna model, preiskovalci pa ne, se zopet zgodi v primeru, ko je verjetnost goljufije višja od mejne vrednosti.
- Z zadnjo situacijo pa se srečamo v primeru, ko verjetnost goljufije pade pod mejno vrednost, torej, ko preiskovalci zaznajo določen delež goljufij, model pa ne.

Zgoraj opisane situacije so avtorji prikazali s spodaj prikazano sliko (glej Sliko 5.2):

Slika 5.2: Prikaz primerov, ko model napoveduje pravilno



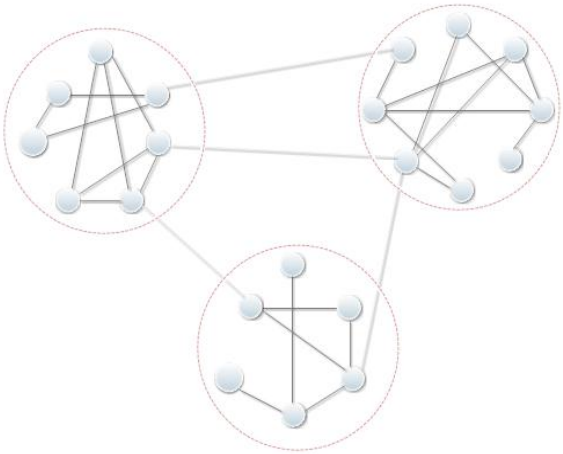
Vir: Belhadji in drugi (2000, 531).

5.1.3 MODELI PRI UPORABI ANALIZE SOCIALNIH MREŽ

Šubelj in drugi (2008) predstavijo model za odkrivanje goljufij na področju avtomobilskih zavarovanj, na osnovi analize socialnih mrež.

V okviru te tehnike operiramo z omrežji, katerih sestavni deli so med seboj povezani nizi točk, ki jih imenujemo vozlišča. Večina omrežij je nehomogoenih, kar se ne nanaša na nediferencirana vozlišča, temveč na različne skupine, ki jih lahko najdemo v omrežju. Med vozlišči znotraj skupin obstaja veliko povezav, medtem ko jih med samimi skupinami obstaja le nekaj (Newman 2004, 312). Primer preprostega omrežja si lahko ogledamo na zgornji sliki (glej Sliko 5.3).

Slika 5.3: Primer preprostega omrežja



Vir: Newman (2004, 312).

Model temelji na podatkih iz policijskih zapisnikov avtomobilskih nesreč, ki vsebujejo podatke o udeležencih, vozilih, dogodku (nesreča), najverjetneje tudi policistu, ki je sestavil zapisnik, in pričah, ki v modelu niso upoštevane, saj se v zapisnikih pojavljajo le redko. Ostali navedeni podatki v modelu predstavljajo posamezno entiteto, ki nosijo svoje atribute, npr. entiteta *udeleženec* ima atribute *ime*, *spol* in *starost* (Šubelj in drugi 2008, 24–25).

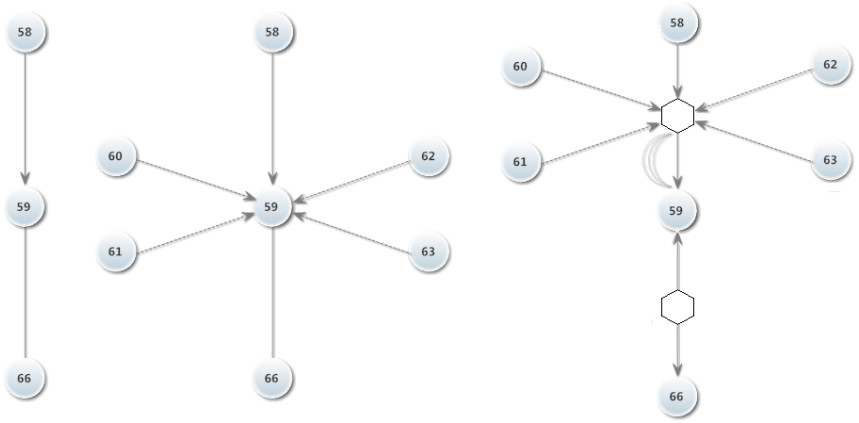
Zbirko podatkov so oblikovali na podlagi 40 policijskih zapisnikov, iz katerih so pridobili informacije o 71 udeležencih (47 voznikov), 48 policistih, 68 vozil in 35 lokacijah nesreč. Sumljivo pa je že dejstvo, da je bilo med 40 nesrečami, kar 47 voznikov (Šubelj in drugi 2008, 40).

V nadaljevanju so predstavljeni koraki, ki so jih avtorji uporabili pri načrtovanju obravnavanega modela.

(I.) *Mreže:* V prvem koraku so avtorji osnovali tri vrste mrež, in sicer mreže voznikov, sopotnikov in nesreč. Mreže temeljijo na povezavah med entitetami in atributi, kjer so entitete vozlišča mrež, povezave pa predstavljajo relacije med njimi. Najprej so osnovali *mrežo voznikov*, in sicer tako, da so med seboj povezali tiste, ki so bili udeleženi v isti nesreči, kjer smer povezave kaže iz krivega v oškodovanega voznika. Neusmerjene povezave v mreži pomenijo, da povzročitelj nesreče ni bil znan. *Mrežo sopotnikov* so dobili tako, da so v drugi fazi mreži voznikov dodali tudi sopotnike, ki so z vozniki povezani tako, da je povezava usmerjena iz sopotnika v voznika. Tretja mreža, ki predstavlja *mrežo nesreč*, ostaja podobna mreži sopotnikov. Razlika je v tem, da so oba voznika oz. sopotnika in voznika preko vozlišča povezali v pripadajočo nesrečo, voznika in nesrečo pa povezujejo še dodatne neusmerjene povezave, ki ustrezajo številu udeleženih sopotnikov v nesreči (Šubelj in drugi 2008, 26–27).

Primere vrst mrež so predstavili tudi slikovno (glej Sliko 5.4). Prva mreža predstavlja mrežo voznikov, druga mrežo sopotnikov, tretja pa mrežo nesreč. Udeleženci nesreče so prikazani z okroglimi vozlišči, nesreče s šest-kotniki, povezave s polno črto pripadajo voznikom, sopotniki s črtkano črto, pikčaste črte med voznikom in nesrečo pa predstavljajo število sopotnikov.

Slika 5.4: Vrste mrež



Vir: Šubelj in drugi (2008, 28).

Za odkrivanje goljufij zadostujeta le mreža sopotnikov in mreža nesreč, ki zagotavljata, da med njunimi komponentami obstaja bijektivna relacija. Ta pride do izraza v naslednjem koraku, saj model uporabi obe vrsti mrež ločeno, kljub vsemu pa upravlja z istim naborom podatkov (Šubelj in drugi 2008, 28–29).

Kjer je potrebno, model v prvem koraku komponente mreže deli v manjše, s čemer so modelu zagotovili olajšanje dela v nadaljnjih fazah in tudi zaradi tega, da bodo skupine posameznikov vključenih v goljufije, sestavljene iz manjšega števila oseb. Delitev poteka tako, da je na vsakem koraku odstranjena povezava z največjo vmesnostjo, tako bo komponenta najverjetneje razpadla na dve manjši (Šubelj in drugi 2008, 29).

(II.) *Iskanje sumljivih mrež:* Druga faza modela, ki uporablja samo mrežo sopotnikov, je namenjena identifikaciji sumljivih komponent znotraj mrež, ki so jih avtorji oblikovali v prvi fazi. Sumljive komponente so s stališča vozlišč in povezav od nesumljivih praviloma večje, med drugim pa je zelo majhno tudi razmerje med številom nesreč in udeležencev. Model torej med komponentami išče tiste, ki po opisanih lastnostih značilno odstopajo od običajnih vzorcev ali pa presega neke mejne vrednosti, torej (Šubelj in drugi 2008, 29):

$$S_i\ (K) = \begin{cases} 1 \text{ komponenta } K \text{ glede na } i\text{-to lastnost izstopa} \\ 0 \text{ sicer} \end{cases}$$

, kjer K predstavlja komponento mreže, S_i pa preslikavo, ki ustreza i-ti lastnosti.

Za sumljive K komponente znotraj mreže velja:

$S\ (K) = \sum_{i=1}^{h} S_i\ \ (K) \geq \frac{h}{2}$, kjer h predstavlja število opazovanih lastnosti. Model zavrže tiste komponente, kjer je $S\ (K) < \frac{h}{2}$. Goljufive skupine, pridobljene na koncu te faze, predstavljajo tudi vhodne podatke za zadnjo fazo modela (Šubelj in drugi 2008, 30).

(III.) *Odkrivanje pomembnih entitet:* V tej fazi podrobno raziščemo sumljive komponente, ki jih je model za tovrstne identificiral v drugi fazi. To pomeni, da model za vsako entiteto sumljive komponente izračuna stopnjo sumljivosti, ki omogoča identifikacijo pomembnih entitet, torej udeležencev in nesreč. Skupino organiziranih goljufov tako najverjetneje predstavljajo tisti udeleženci, ki so v komponenti najbolj sumljivi. Stopnja sumljivosti deluje na podlagi iterativne metode. V tem postopku je pomembno, da je vsaka entiteta dobro opisana s svojimi lastnostmi in lastnostmi entitet, s katerimi tvori povezave. Model na tej stopnji podatke črpa iz mreže nesreč (Šubelj in drugi 2008, 33).

V nadaljevanju Šubelj in drugi (2008, 34) predstavijo metodo za določanje stopnje sumljivosti entitet, ki poteka v treh korakih, kjer $(u_1, \ldots, u_s)$ predstavlja udeležence, $(a_1, \ldots, a_t)$ nesreče komponente K, $V_K(\cdot)$ bijektivno preslikavo, ki vsaki entiteti pripiše ustrezno vozlišče, $s^i(\cdot)$ pa sumljivost določene entitete na i-ti iteraciji:

1. stopnja sumljivosti entitet je enaka $\forall i : s^0\ (u_i) = \frac{1}{s}$ ter $\forall i : s^0\ (a_i) = \frac{1}{s}$,
2. dokler velja $\sum_{i=1}^{s}(s^k\ (u_i) - s^{k-1}(u_i))^2 > e^2$, ponovi:

$$\forall i : s^{k+1}\ (a_i) = f_{ent}(a_i) \sum_{e=\{v,V_K(a_i\}\in E(K),\ \ x=V_K^{-1}(v)} f_e(e,a_i)\ s^k(x)$$

$\forall i : s^{k+1}\ (u_i) = ăs^k(u_i) + (1 - ă)\ f_{ent}(u_i) \sum_{e=\{v,V_K(u_i\}\in E(K),\ \ a_j=V_K^{-1}(v)} f_e(e,u_i)\ s^{k+1}(a_j)$, kjer je $f_{ent}(x)$ faktor lastnosti entitete x , $f_e(e,x)$ pa predstavlja utež, odvisno od tipa povezave e.

$$\forall i : s^{k+1}\ (u_i) = \frac{s^{k+1}(u_i)}{\sum_{j=1}^{s} s^{k+1}(u_j)}\ .$$

Kar pomeni, da v tej fazi model najprej na vsaki iteraciji oceni sumljivost nesreč, ki je izražena kot utežena linearna kombinacija sumljivosti njenih sosedov (prva enačba druge točke). Ocena sumljivosti nesreče je tako osnova za oceno sumljivosti udeležencev (druga enačba druge točke), in predstavlja kombinacijo stare in nove vrednosti. Model na koncu iteracije sumljivost udeležencev še normalizira (tretja enačba druge točke) (Šubelj in drugi 2008, 34).

3. Zadnji korak je osnova za medsebojno primerjavo sumljivosti entitet različnih komponent. To je zagotovljeno tako, da model sumljivost udeležencev še enkrat normalizira, ampak tokrat glede na sumljivost komponente $CS(K)$, ki predstavlja vsoto sumljivosti vseh nesreč komponente K (Šubelj in drugi 2008, 35):

$$\forall i : s^{k+1}(u_i) = s^{k+1}(u_i)\ CS(K).$$

(IV.) *Uporaba na dejanskih podatkih:* V prvi fazi modela, torej predstavitev z mrežami, se podatki v mreži nesreč razdelijo na štiri povezane komponente. Druga faza modela dve komponenti opusti (14 udeležencev, med katerimi so trije goljufi), preostali dve pa identificira kot sumljivi. Model z metodo za določanje stopnje sumljivosti entitet, le-to oceni za udeležence sumljivih komponent, ostalim pa pripiše vrednost 0. Na podlagi izračunane sumljivosti udeleženca ($\bar{x} = 0{,}30$, s standardnim odklonom $s = 0{,}33$), so avtorji nato kot sumljive izpostavili vse udeležence, katerih sumljivost je bila nadpovprečna. Natančnost modela je enaka 83,10%, med vsemi udeleženci (71) nesreč, pa je kot goljufe prepoznal 24 udeležencev (Šubelj in drugi 2008, 41).

6 PREDSTAVITEV MODELA

V tem poglavju se lotimo načrtovanja in predstavitvi našega modela za odkrivanje zavarovalniških goljufij, ki temelji na analizi odkrivanja anomalij. Na začetku predstavimo program, v katerem bo izvedeno testiranje načrtovanega modela, nato opišemo sam način, s katerim program upravlja s podatki in uporaba katerih kriterijev je potrebna za izvedbo analize, kasneje pa se posvetimo natančnemu opisu modela.

6.1 DOLOČITEV KRITERIJEV MODELA

Kot je bilo že omenjeno, bo za testiranje modela uporabljen program SPSS Modeler, ki ga je razvilo podjetje IBM. Gre za bogato zbirko orodij za podatkovno rudarjenje, pri katerem ni potrebno znanje programiranja, saj program nudi vrsto vgrajenih modulov, ki delujejo na različnih tehnikah podatkovnega rudarjenja. Med moduli najdemo tudi take, ki omogočajo odkrivanje ali napovedovanje goljufij.

Med omenjenimi moduli poznamo orodje za **analizo entitet** (ang. *entity analytics*), katere metodologija temelji na razreševanju nekonsistentnosti med prvotnimi podatki, in tako izboljša njihovo povezanost in usklajenost. Ko s programom poskušamo odkrivati goljufije, algoritmi analize entitet potencialno goljufive primere prepoznajo tako, da jih povežejo z že znanimi goljufivimi vzorci, ali jih prepoznajo na podlagi njim podobnih predhodnikov (IBM 2012a, 1–3).

Na drugi strani pa modul za **odkrivanje anomalij** (ang. *anomaly detection*) omogoča prepoznavanje izstopajočih ali nenavadnih primerov v podatkih, in sicer tako, da na podlagi obravnavanih podatkih skladišči informacije o vzorcih vedenja, ki se zdijo normalni. Torej, za razliko od ostalih metod, ta modul omogoča prepoznavanje izstopajočih ali nenavadnih primerov, tudi če ti ne ustrezajo nobenemu znanemu vzorcu. Ker pa se vzorci, ki nakazujejo goljufijo vedenje, lahko konstantno spreminjajo, je obravnavani modul pri odkrivanju goljufij zelo uporaben. Naj izpostavimo tudi to, da analiza odkrivanja anomalij temelji na metodi razvrščanja v skupine (IBM 2012b, 77).

Skozi analizo odkrivanja anomalij operiramo z indeksom anomalije, s pomočjo katerega izvedeni model določi stopnjo, ki nam pove v kolikšni meri primer odstopa od skupine vzorcev normalnega vedenja. Večji kot je indeks, večje je odstopanje, kar pomeni, da moramo primer obravnavati kot sumljivega. Navadno so kot sumljivi primeri obravnavani tisti, ki imajo vrednost indeksa anomalije večjo od 1,5 (IBM 2012b, 78).

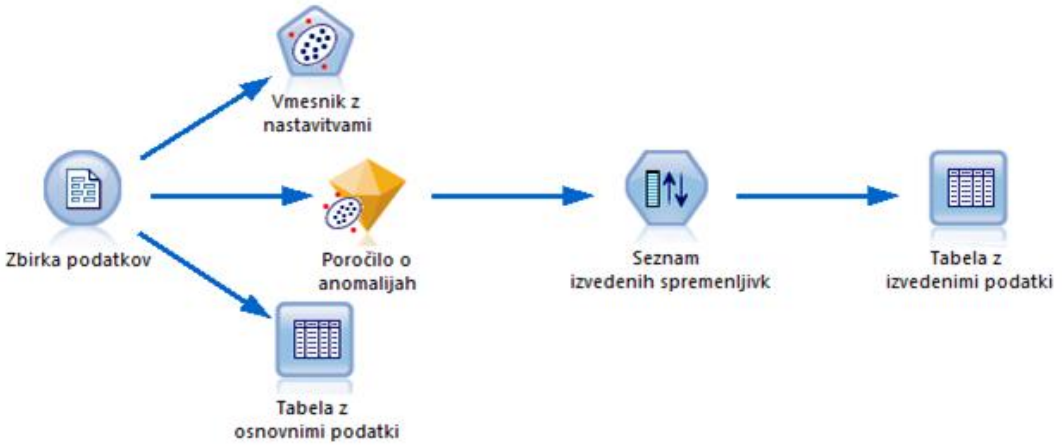
Za izvedbo analize najprej določimo vrednost odstopanja, ki pove pri kateri naj program primer označi kot anomalijo. In sicer imamo tri možnosti, določimo lahko minimalno vrednost indeksa odstopanja, odstotek najbolj odstopajočih primerov ali število najbolj odstopajočih primerov v zbirki podatkov. V naslednjem koraku lahko določimo tudi število najbolj odstopajočih spremenljivk, s pomočjo katerih lahko pojasnimo, zakaj je določen primer anomalija. S koeficientom prilagoditve pri izračunu razdalj določimo vrednost, ki uravnovesi utež zveznim in nominalnim spremenljivkam. Večjo vrednost kot določimo, večji je vpliv zvezne spremenljivk. Ne nazadnje lahko določimo tudi, ali naj program upošteva manjkajoče vrednosti spremenljivk.

Primere, ki smo jih skozi analizo odkrivanja anomalij prepoznali kot sumljive, moramo v nadaljevanju obravnavati natančneje in ugotoviti ali so primeri tudi v resnici goljufija dejanja ali ne. Za nadaljnje raziskovanje strokovnjaki najpogosteje uporabijo metodo nevronskih mrež (He in drugi v Chandola in drugi 2009, 15; IBM 2012b, 78).

6.2 OPIS MODELA

V nadaljevanju opisujemo algoritem, ki ga SPSS Modeler uporablja skozi proces odkrivanja anomalij v zbirki podatkov. Na začetku pa za lažje razumevanje celotnega procesa analize, model prikazujemo tudi orisno (glej Sliko 6.1).

Slika 6.1: Proces iskanja anomalij



Vir: SPSS (2000, 218).

1. Modeliranje:

(I.) Program na začetku odstrani tiste zvezne spremenljivke, ki imajo izjemno visoke vrednosti in manjkajoče spremenljivke, razen v primeru, ko smo v nastavitvah določili drugače. Program preostale spremenljivke uporabi za izgradnjo modela odkrivanja anomalij (IBM 2012c, 4).

(II.) Na podlagi hierarhičnega razvrščanja program vsakemu primeru, v našem te predstavljajo zavarovalniški zahtevki, dodeli ustrezno skupino, ki temelji na podobnostih vhodnih spremenljivk. Gre torej za identifikacijo razvrstitvenih skupin (IBM 2012c, 4–5).

(III.a.) Program za vsako vhodno zvezno spremenljivko znotraj skupin izračuna povprečje vseh spremenljivk ali t.i. glavno povprečje (ang. *grand mean*) in standardni odklon vseh spremenljivk ali t.i. glavni standardni odklon (ang. *grand standard deviation*). V okviru nominalnih spremenljivk pa določi, katera vrednost nosi največjo težo oz. odstotek (IBM 2012c, 4–5).

(III.b.) Če smo v nastavitvah določili upoštevanje manjkajočih spremenljivk, program vsem manjkajočim zveznim spremenljivkam nato dodeli glavno povprečje. Na drugi strani pa v primeru manjkajočih nominalnih spremenljivk, vsem dodeli kategorijo manjkajočih vrednosti, ki so upoštevane kot veljavne. Program za obe vrsti spremenljivk,

izračuna novo zvezno spremenljivko, ki predstavlja odstotek manjkajočih spremenljivk (IBM 2012c, 4–5).

2. Faza ocenjevanja:

Predelane vhodne spremenljivke iz prvega koraka so tako pripravljene za izgradnjo modela, ki ga skupaj s predhodno izračunanimi statistikami program uporabi v tem koraku. Če so spremenljivke zvezne, sta torej upoštevana glavno povprečje in standardni odklon, ter omenjeni statistiki znotraj posameznih skupin, opisani nekoliko kasneje. V primeru nominalnih spremenljivk pa frekvenca posamezne kategorije (IBM 2012c, 5).

(I.) Vsakemu zavarovalniškemu zahtevku model dodeli ustrezno, torej najbližjo skupino, vsaki skupini identifikacijsko številko, vsaki spremenljivki pa nato še indeks odklona. Indeks odklona spremenljivke je definiran kot razdalja spremenljivke od skupine, ki ustreza povprečju skupine. Kot omenjeno pa program izračuna tudi standardni odklon skupine posameznega zavarovalniškega zahtevka, ki predstavlja vsoto indeksov odklona vseh spremenljivk (IBM 2012c, 5).

(II.) Kasneje program izračuna nova dva indeksa, ki sta z vidika razumljivosti lažja za interpretacijo, to sta indeks anomalije in mera prispevka posamezne spremenljivke. **Indeks anomalije** je izpeljan iz indeksa standardnega odklona skupine, in sicer predstavlja kvocient indeksa standardnega odklona skupine posameznega zavarovalniškega zahtevka in povprečje vseh, ki pripadajo isti skupini. **Mera prispevka** posamezne spremenljivke pa predstavlja kombinacijo obeh indeksov (indeks odklona spremenljivke in standardnega odklona skupine posameznega zavarovalniškega zahtevka) (IBM 2012c, 5–6).

3. Faza sklepanja:

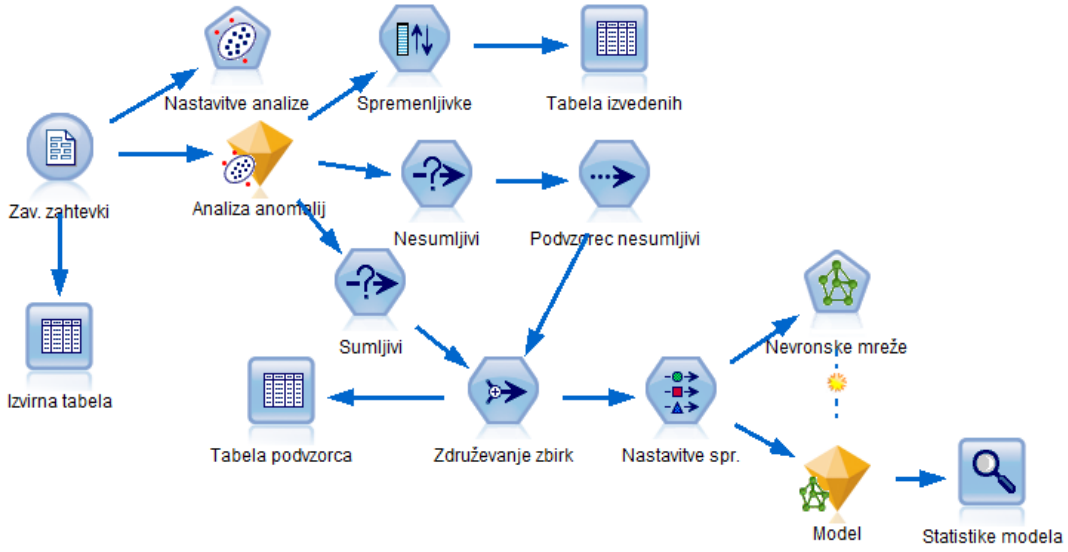
(I.) V tej fazi nastopi razvrščanje nenavadnih oz. sumljivih zavarovalniških zahtevkov. Zavarovalniški zahtevki so razvrščeni glede na indeks anomalije od najbolj do najmanj sumljivega. Prikazanih je le toliko zavarovalniških zahtevkov, kolikor smo jih določili v nastavitvah, torej določeno število ali delež vseh zahtevkov (IBM 2012c, 7).

(II.) V drugem koraku te faze zavarovalniške zahtevke razvrstimo po indeksu odklona spremenljivke in podamo razloge, zakaj so zavarovalniški zahtevki obravnavani kot sumljivi (IBM 2012c, 7).

4. Evalvacija modela:

Na koncu za pojasnjevanje modela uporabimo metodo nevronskih mrež, ki nam bo v pomoč za evalvacijo modela. Tako bomo lahko prišli do zaključkov in ugotovitev, katere zavarovalniške zahtevke smo z analizo odkrivanja anomalij pravilno identificirali. Celoten oris modela si lahko ogledamo na spodnji sliki (glej Sliko 6.2).

Slika 6.2: Grafični prikaz modela



7 ŠTUDIJA PRIMERA

Zbirka podatkov je del testnih zbirk podatkov programa Clementine, ki je predhodnik programa SPSS Modeler. V magistrskem delu so obravnavani zahtevki premoženjskih zavarovalniških podatkov na področju kmetijstva, s katerimi imajo imetniki polic zavarovane svoje pridelke in zaloge. S pomočjo podatkovnega rudarjenja poskušamo na tem primeru ugotoviti, kolikšen delež zavarovalniških zahtevkov je sumljivih.

Zbirka podatkov obsega 300 zavarovalniških zahtevkov, zavedamo pa se, da je takšno število primerov za prave analize v procesu podatkovnega rudarjenja malo. Vsak zavarovalniški zahtevek ima svojo identifikacijsko številko, ime, poleg tega pa vsebuje še informacije o regiji, velikosti njive, obsegu padavin, prihodku kmetije, vrsti glavnega pridelka, vrsti zavarovalniškega zahtevka ter vrednosti vloženega zahtevka. Na spodnji sliki (glej Sliko 7.1) prikazujemo del obravnavane zbirke podatkov.

Slika 7.1: Del zbirke podatkov

	id	name	region	farmsize	rainfall	landquality	farmincome	maincrop	claimtype	claimvalue
30	id630	name630	midlands	680	81	4	221391.000	potatoes	arable_dev	68370.900
31	id631	name631	midlands	960	52	9	408440.000	maize	decommission_land	176464.0...
32	id632	name632	midlands	1260	82	3	319386.000	wheat	decommission_land	63715.600
33	id633	name633	southeast	1820	90	9	1621150.0...	maize	decommission_land	574478.0...
34	id634	name634	north	820	101	6	516064.000	rapeseed	decommission_land	187782.0...
35	id635	name618	southeast	520	107	7	392856.000	maize	arable_dev	116072.0...
36	id636	name636	southeast	1160	21	8	185939.000	potatoes	arable_dev	52599.800
37	id637	name637	midlands	940	106	6	622450.000	maize	arable_dev	170310.0...
38	id638	name638	midlands	1480	64	6	586185.000	wheat	arable_dev	158310.0...
39	id639	name639	southwest	1240	44	7	346747.000	rapeseed	decommission_land	134176.0...
40	id640	name640	southeast	960	60	4	215908.000	wheat	arable_dev	69410.800

V nadaljevanju predstavljamo rezultate analize, in sicer skozi posamezne faze modela, ki smo jih predstavili v prejšnjem poglavju. Zadnja faza, evalvacija modela, je prikazana v svojem poglavju (glej osmo poglavje).

1. Modeliranje:

(I.) Program izpusti prvi korak te faze, saj v zbirki podatkov ni nobenega zavarovalniškega zahtevka ali spremenljivke, ki bi imela izjemno visoke ali manjkajoče vrednosti.

(II.) V fazi razvrščanja v skupine je program zavarovalniške zahtevke razdelil v dve skupini (glej Prilogo A). Prva skupina vsebuje 202 zavarovalniška zahtevka, druga pa 98 zavarovalniških zahtevkov (glej Prilogo B).

(III.a.) Kot vemo v tej fazi program za vsako zvezno spremenljivko posamezne skupine izračuna glavno povprečje in glavni standardni odklon, vsaki vrednosti nominalne spremenljivke pa je pripisana frekvenca oz. odstotek od celote.

Tabela 7.1: Statistike modeliranja zveznih spremenljivk

	Prva skupina		**Druga skupina**	
Ime spremenljivke	Glavno povprečje	Glavni std. odklon	Glavno povprečje	Glavni std. odklon
VrednostZahtevka	86185,011	42808,141	271474,49	102289,578
PrihodekFarme	291679,895	142112,438	865520,51	292275,943
VelikostFarme	1087,624	439,787	1462,449	360,678
KakovostZemlje	5,634	1,984	7,122	1,664
Padavine	56,594	27,178	85,949	16,338

Ločeno prikazujemo statistike zveznih (glej Tabelo 7.1) in nominalnih spremenljivk (glej Tabelo 7.2) glede na skupini. Pri slednjih so najpomembnejše označene z odebeljenimi številkami. V prilogi (glej Prilogo B) prikazujemo tudi poročilo iz programa.

Tabela 7.2: Statistike modeliranja nominalnih spremenljivk

		Prva skupina	**Druga skupina**
Ime spremenljivke	Razred spremenljivke	Odstotek	Odstotek
VrstaZahtevka	uničenje	**56,93%**	**63,27%**
	površine	43,07%	36,73%
GlavniPridelek	pšenica	**43,56%**	31,63%
	koruza	28,22%	**34,69%**
	repa	17,33%	14,29%
	krompir	10,89%	19,39%
Regija	notranjost	37,13%	18,37%
	sever	11,39%	16,33%
	jugozahod	13,37%	16,33%
	jugovzhod	**38,12%**	**48,98%**

(III.b.) Podobno kot v prvem koraku te faze, program tudi tega koraka ne izvede, saj v zbirki podatkov ni manjkajočih vrednosti spremenljivk.

2. Faza ocenjevanja:

(I.) Gre za preliminarni korak ocenjevanja anomalij. V ozadju programa se na podlagi skupin, v katere se zavarovalniški primeri uvrščajo in rezultatov iz prve faze tretjega koraka, dodelijo prve ocene o izstopajočih zavarovalniških zahtevkih.

(II.) Ocene iz prejšnjega koraka model uporabi za izračun **indeksa anomalije** in **mere prispevka** posamezne spremenljivke za vsak zavarovalniški zahtevek. V spodnji tabeli (glej Tabelo 7.3) si lahko ogledamo osnovne statistike indeksa anomalije. Vsakemu zavarovalniškemu zahtevku so pripisane še najvišje mere prispevka, ki nam povedo, katere tri spremenljivke najbolj vplivajo na sumljivost posameznega zavarovalniškega zahtevka.

Tabela 7.3: Statistike indeksa anomalije

N	**300**
Povprečje	1,000
Minimum	0,652
Maksimum	1,770
Razpon	1,118
Varianca	0,031
Standardni odklon	0,175
Standardna napaka povprečja	0,010

Za primer si poglejmo zavarovalniški zahtevek »*id633*« iz spodnje slike (glej Sliko 7.2). Indeks anomalije je tem primeru enak 1,6, kar glede na zgornjo tabelo (glej Tabelo 7.3) predstavlja nadpovprečno vrednost indeksa, vendar ne najvišje. Nanj v največji meri vpliva vrednost vloženega zahtevka (*claimvalue* ≅ 0,36), prihodek kmetije (*farmincome* ≅ 0,28) in vrsta glavnega pridelka (*maincrop* ≅ 0,09).

3. Faza sklepanja:

Program kreira novo zbirko podatkov izvedenih spremenljivk, kjer najdemo podatke o indeksu anomalije, skupini v katero je zavarovalniški zahtevek uvrščen, mere prispevka

posamezne spremenljivke, vsakemu pa tudi pripiše vrednost, ki nam pove, ali je zahtevek sumljiv ali ne. Na spodnji sliki (glej Sliko 7.2) tako vidimo, da je zavarovalniški zahtevek »*id633*« označen kot anomalija.

Slika 7.2: Del zbirke podatkov izvedenih spremenljivk, s poudarjenim sumljivim zavarovalniškim zahtevkom

	id	$O-Anomaly	$O-AnomalyIndex	$O-PeerGroup	$O-Field-1	$O-FieldImpact-1	$O-Field-2	$O-FieldImpact-2	$O-Field-3	$O-FieldImpact-3
30	id630	F	1.027	1	maincrop	0.318	region	0.143	farmsize	0.129
31	id631	F	1.054	1	landquality	0.249	maincrop	0.178	claimvalue	0.160
32	id632	F	0.806	1	landquality	0.238	region	0.182	rainfall	0.166
33	id633	T	1.600	2	claimvalue	0.358	farminco...	0.275	maincrop	0.093
34	id634	F	1.227	2	maincrop	0.220	region	0.206	farmsize	0.189
35	id635	F	1.109	1	rainfall	0.267	maincrop	0.169	farmsize	0.166
36	id636	F	1.103	1	maincrop	0.297	rainfall	0.170	landquality	0.154
37	id637	F	1.112	2	region	0.212	farmsize	0.155	maincrop	0.133
38	id638	F	0.980	1	farminco...	0.183	region	0.150	farmsize	0.131
39	id639	F	1.005	1	region	0.296	maincrop	0.258	landquality	0.109
40	id640	F	0.721	1	region	0.198	claimtype	0.173	maincrop	0.171

(I.) Glede na razvrščanje po sumljivosti, je model skupno zaznal 10 sumljivih zavarovalniških zahtevkov, zbirko podatkov le-teh si lahko ogledamo v prilogi (glej Prilogo C). Glede na skupine, je v prvi med 202 zavarovalniškimi zahtevki kot sumljive zaznal 3, v drugi pa 7 med 89 (glej Prilogo D). V spodnji tabeli (glej tabelo 7.4) opazimo, da ima druga skupina višje povprečje indeksa anomalije, med drugim pa zajema najvišjo in najnižjo vrednost. Pri primerjavi sumljivih in nesumljivih zavarovalniških zahtevkov pa opazimo značilno višje vrednosti statistik pri sumljivih, razen v primeru razpona vrednosti.

Tabela 7.4: Statistike indeksa anomalije glede na skupine in sumljivost zavarovalniških zahtevkov

	Prva skupina		**Druga skupina**		**Skupaj**	
	Sumljivi	Nesumljivi	Sumljivi	Nesumljivi	Sumljivi	Nesumljivi
N	3	199	7	91	10	290
Povprečje	1,462	0,993	1,533	0,959	1,512	0,982
Minimum	1,358	0,652	1,350	0,705	1,350	0,652
Maksimum	1,641	1,335	1,770	1,323	1,770	1,335
Razpon	0,283	0,683	0,420	0,618	0,420	0,683
Varianca	0,024	0,021	0,024	0,022	0,023	0,022
Standardni odklon	0,156	0,146	0,156	0,149	0,151	0,147
Std. napaka povprečja	0,090	0,010	0,059	0,016	0,048	0,009

(II.) V prvi skupini (glej Prilogo D) lahko razloge za sumljivost zavarovalniških zahtevkov v največji meri iščemo v spremenljivkah regija (*region*), vrsta glavnega pridelka (*maincrop*) in obseg padavin (*rainfall*), saj imajo najvišji povprečni indeks (≅ 0,2) in se pojavljajo kot odstopanja na vsaj dveh (od treh) zavarovalniških zahtevkih.

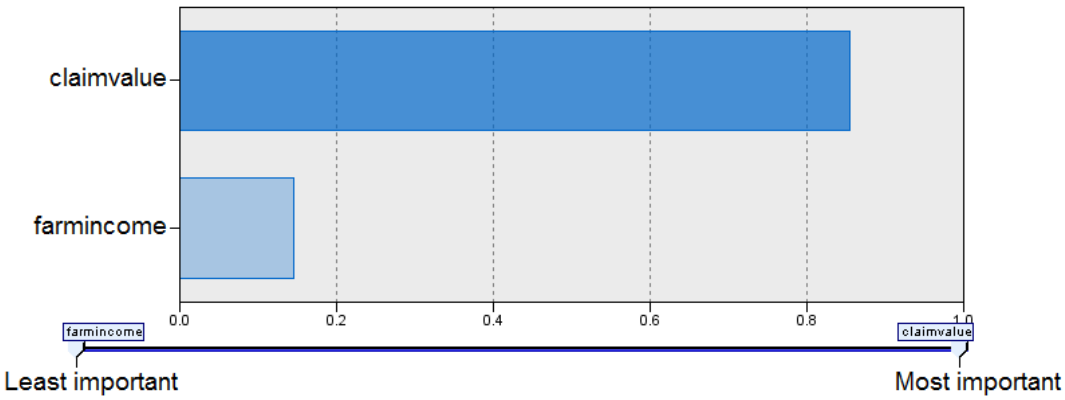
Medtem pa v drugi skupini (glej Prilogo D) največji razlog za sumljivost zavarovalniških zahtevkov nosijo spremenljivke vrednost vloženega zahtevka (*claimvalue*), prihodek kmetije (*farmincome*) in vrsta glavnega pridelka (*maincrop*). Povprečni indeks prve spremenljivke znaša 0,27 in se kot odstopanje pojavi pri šestih (od sedmih) zavarovalniških zahtevkih. Druga spremenljivka nosi približno enak povprečni indeks (≅0,27), in predstavlja odstopanje pri petih zavarovalniških zahtevkih, medtem ko tretja predstavlja odstopanja pri šestih zavarovalniških zahtevkih, vendar ima glede na ostali spremenljivki povprečni indeks precej manjši (≅ 0,14).

8 EVALVACIJA MODELA

V zadnjem koraku modela izvedemo podrobnejše raziskovanje ob uporabi metode nevronskih mrež. Z analizo odkrivanja anomalij se je kot sumljivih izkazalo 10 zavarovalniških primerov. Ker pa mora biti za uspešno pojasnjevanje modela razmerje med sumljivimi in legalnimi zavarovalniškimi zahtevki vsaj $30\% : 70\%$, smo v prvem koraku segment legalnih primerno zreducirali. Na koncu dobimo zbirko podatkov, ki obsega 30 zavarovalniških zahtevkov, kar sicer predstavlja razmerje $33{,}3\% : 66\%$ sumljivih in nesumljivih zahtevkov.

Na združeni zbirki podatkov sumljivih in pod-vzorca nesumljivih zavarovalniških zahtevkov zaženemo model nevronskih mrež, kjer odvisno spremenljivko predstavlja spremenljivka *anomalija* oz. tista, ki določa ali je zahtevek sumljiv ali nesumljiv. Na drugi strani pa smo na podlagi statistik izvedene analize odkrivanja anomalij kot neodvisni določili *claimvalue* (vrednost vloženega zahtevka) in *farmincome* (prihodek kmetije). Kot vidimo na spodnjem grafu (glej Graf 8.1), ima največji vpliv na sumljivost zavarovalniškega zahtevka njegova vrednost.

Graf 8.1: Pomembnost neodvisnih spremenljiv



Bolj kot to pa nas seveda zanima, kakšna je sama stopnja uspešnosti modela. V tabeli (glej Tabelo 8.1), vidimo, da je model pravilno odkril 6 sumljivih zavarovalniških zahtevkov, ostale 4 pa je z analizo odkrivanja anomalij identificiral napačno. Medtem, ko

je na drugi strani vse nesumljive identificiral pravilno. Model je tako pravilno identificiral 87% vseh zavarovalniških zahtevkov (glej Prilogo G).

Tabela 8.1: Matrika sovpadanja

	Legalen	Nelegalen
Nesumljiv	20	0
Sumljiv	4	6

9 ZAKLJUČEK

Na področju odkrivanja goljufij se soočamo z različnimi izzivi, saj sodi v zelo specifično in kompleksno področje raziskovanja. Nepoznavanje tega področja in uporaba manj ali celo neustreznih metod lahko predstavlja dolgotrajen ali neuspešen proces njihovega odkrivanja. V tem procesu se osredotočamo na iskanje novih znanj in odkrivanje vzorcev, ki se nam zdijo nenavadni, predvsem pa moramo znotraj tega praviloma operirati z veliko količino podatkov. Ker pa se pri podatkovnem rudarjenju srečujemo s sorodnim procesom, se na področju odkrivanja goljufij izkaže kot zelo učinkovita metoda.

S preučevanjem tega področja se je ukvarjalo že veliko raziskovalcev in posameznikov različnih strok. Ti so v svojih raziskavah in delih obravnavali odkrivanje goljufij z uporabo različnih tehnik podatkovnega rudarjenja, med katerimi se pojavijo Bayesova klasifikacija, odločitvena drevesa, regresija, nevronske mreže, analiza socialnih mrež, analiza odkrivanja anomalij, in druge. Nekateri pa so se raziskovanja lotili celo z uporabo kombinacije dveh ali več tehnik.

Ker je to področje, kljub svoji kompleksnosti, vseeno privlačno in zanimivo, smo se tudi sami v magistrskem delu osredotočili na odkrivanje goljufij, in sicer na področju zavarovalništva. Najprej smo ob pregledu strokovne literature predstavili nekaj modelov za odkrivanje zavarovalniških goljufij, ki temeljijo na različnih tehnikah podatkovnega rudarjenja. Na naslednji stopnji smo se načrtovanju modela posvetili tudi sami, v katerega smo v prvi fazi aplicirali tehniko analize odkrivanja anomalij, v drugi pa še tehniko nevronskih mrež. Na koncu je sledilo testiranje načrtovanega modela na realnih podatkih, ki predstavljajo del zbirke podatkov z informacijami o zavarovalniških zahtevkih.

Prva faza modela nam je omogočila identificiranje sumljivih zavarovalniških zahtevkov in relevantnih spremenljivk, katerih vloga je do izraza prišla v naslednji fazi. Z implementacijo odvisne in neodvisnih spremenljivk v drugo fazo modela, smo tako lahko preverili stopnjo uspešnosti modela. Tudi izbira tehnik obeh faz modela je v našem primeru odigrala pomembno vlogo. Prva nam je tako omogočila natančnejši vpogled v

same podatke, torej jo lahko razumemo kot raziskovalno vlogo modela. Druga faza pa pojasnjevalno vlogo, saj smo z njo lahko model ovrednotili. Kot že vemo, pa je predpogoj za učinkovito načrtovanje modela v prvi vrsti dobro poznavanje področja problematike.

Sama stopnja uspešnosti modela je načeloma visoka, saj je pravilno identificiral 87% zavarovalniških zahtevkov. Če pa se osredotočimo le na sumljive zahtevke, jih je izmed desetih pravilno uvrstil šest. Kljub temu se moramo zavedati, da je zbirka podatkov obsegala relativno majhno število zavarovalniških zahtevkov, zato bi bilo za prikaz realnejše slike uspešnosti modela smiselno model testirati še večjem vzorcu.

Tukaj nastopi nov problem, kar se je seveda v našem primeru izkazalo kot pomanjkljivost magistrskega dela, saj do zavarovalniških zbirk podatkov, ki bi nam jih posredovale zavarovalnice, ne moramo priti zlahka. Ob razmisleku ugotovimo, da je to ravnanje smiselno in logično, saj gre v končni fazi le za poslovne podatke. Posledično pa se tako srečamo tudi s primanjkljajem prosto dostopnih podatkov.

10 LITERATURA

Belhadji, El Bachir, Georges Dionne in Faouzi Tarkhani. 2000. A Model for the Detection of Insurance Fraud. *The Geneva Papers on Risk and Insurance* 25. Dostopno prek: https://www.genevaassociation.org/media/236369/ga2000_gp25%284%29_belhadji,dionnetarkhani.pdf (12. september 2013).

Bhowmik, Rekha. 2011. Detecting Auto Insurance Fraud by Data Mining Techniques. Journal of Emerging Trends in Computing and Information Sciences 2 (4). Dostopno prek: http://www.cisjournal.org/archive/vol2no4/vol2no4_1.pdf (12. september 2013).

Bolton, Richard J. in David J. Hand. 2002. Statistical Fraud Detection: A Review. *Statistical Science* 17 (3). Dostopno prek: http://projecteuclid.org/DPubS/Repository/1.0/Disseminate?view=body&id=pdf_1&handle=euclid.ss/1042727940 (13. februar 2013).

Chandola, Varun, Arindam Banerjee in Vipin Kumar. 2009. Anomaly Detection: A Survey. *ACM Computing Surveys* 41 (3). Dostopno prek: http://www-users.cs.umn.edu/~banerjee/papers/09/anomaly.pdf (20. september 2013).

Crocker, Keith J in Sharon Tennyson. 2002. Insurance Fraud and Optimal Claims Settlement Strategies. *Journal of Law & Economics* 45 (2). Dostopno prek: http://tennyson.human.cornell.edu/research/Insurance%20Fraud%20and%20Optimal%20Claims%20Settlement%20Strategies.pdf (3. september 2013).

Derrig, Richard A. 2002. Insurance Fraud. *The Journal of Risk and Insurance* 69 (3). Dostopno prek: http://www.derrig.com/research/InsuranceFraud.pdf (12. februar 2013).

Derrig, Richard A. in Daniel J. Johnston in Elizabeth A. Sprinkel. 2006. Auto insurance fraud: Measurements and efforts to combat it. *Risk Management and Insurance Review* 9 (2): 109–130.

Fawcett, Tom in Foster Provost. 2002. Fraud Detection. V *Handbook of Data Mining and Knowledge Discovery*, ur. Jan Zytkow in Willi Klösgen, 726–731. Oxford University: Press.

Ferligoj, Anuška. 1989. Razvrščanje v skupine: Teorija in uporaba v družboslovju. *Metodološki zvezki* 4. Ljubljana: Fakulteta za sociologijo, politične vede in novinarstvo, Raziskovalni inštitut.

Furlan, Štefan in Marko Bajec. 2009. Celovit pristop k obvladovanju zavarovalniških goljufij. *Uporabna informatika* 17 (2): 72–78.

Gepp, Adrian, J. Holton Wilson, Kuldeep Kumar in Sukanto Bhattacharya. 2012. A Comparative Analysis of Decision Trees Vis-à-vis Other Computational Data Mining Techniques in Automotive Insurance Fraud Detection. *Journal of Data Science* 10: 537–561.

Giudici, Paolo in Silvia Figini. 2009. *Applied data mining for business and industry: Second Edition*. West Sussex, UK: John Wiley & Sons, Ltd.

Hand, David, Heikki Mannila in Padhraic Smyth. 2001. *Principles of Data Mining*. Cambridge, MA: MIT Press.

Hastie, Trevor, Robert Tibshirani in Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction: Second Edition*. New York, NY: Springer.

IBM. 2012a. *IBM SPSS Modeler Entity Analytics 15 User Giude*. New York: IBM Corporation.

--- 2012b. *IBM SPSS Modeler 15 Modeling Nodes*. New York: IBM Corporation.

Kopše, Vilijem. 2004. Zavarovalniške goljufije, povezane s tatvinami motornih vozil. V *Goljufije v zavarovalništvu*, ur. Anton Dvoršek in Liljana Selinšek, 127–146. Ljubljana: Fakulteta za policijsko varnostne vede; Maribor: Pravna fakulteta.

Lamberger, Igor. 2004. Zavarovalniške goljufije v Sloveniji; preiskovanje s policijskega zornega kota. V *Goljufije v zavarovalništvu*, ur. Anton Dvoršek in Liljana Selinšek, 107–126. Ljubljana: Fakulteta za policijsko varnostne vede; Maribor: Pravna fakulteta.

Larose, Daniel T. 2005. Discovering knowledge in data: An Introduction to Data Mining. New Jersey: John Wiley & Sons, Inc.

--- 2006. *Data mining methods and models*. New Jersey: John Wiley & Sons, Inc.

Mazar, Nina, On Amir in Dan Ariely. 2008. The Dishonesty of Honest People: A Theory of Self-Concept Maintenance. *Journal of Marketing Research* 45 (6). Dostopno prek: http://econ.ucsd.edu/~jandreon/Seminar/OnAmir.pdf (3. september 2013).

Newman, M. E. J. 2004. Detecting community structure in networks. *The European Physical Journal B* 38 (2). Dostopno prek: http://www-personal.umich.edu/~mejn/papers/epjb.pdf (16 september 2013).

Nisbet, Robert, John Elder in Gary Miner. 2009. *Handbook of Statistical Analysis and Data Mining Applications*. Burlington, MA: Elsevier.

Petrović, Tomislav. 2004. Prevare u osiguranju života. V *Goljufije v zavarovalništvu*, ur. Anton Dvoršek in Liljana Selinšek, 61–75. Ljubljana: Fakulteta za policijsko varnostne vede; Maribor: Pravna fakulteta.

Phua, Clifton, Vincent Lee in Kate Smith in Ross Gayler. 2005. A Comprehensive Survey of Data Mining-based Fraud Detection Research. *Artificial Intelligence review,* 9. februar. Dostopno prek: http://arxiv.org/ftp/arxiv/papers/1009/1009.6119.pdf (12. februar 2013).

Rajaraman, Anand in Jeffrey David Ullman. 2012. *Mining of Massive Datasets*. Cambridge: Cambridge University Press.

Schroeder, Jennifer, Jennifer Xu, Hsinchun Chen in Michael Chau. 2007. Automated Criminal Link Analysis Based on Domain Knowledge. *Journal of the american society for information science and technology* 58(6). Dostopno prek: http://ai.arizona.edu/intranet/papers/JASIST_v58_n6.pdf (9. september 2013).

Sharma, Anuj in Prabin Kumar Panigrahi. 2012. A Review of Financial Accounting Fraud Detection based on Data Mining Techniques. *International Journal of Computer Applications* 39 (1). Dostopno prek: http://research.ijcaonline.org/volume39/number1/pxc3877016.pdf (9. september 2013).

Sherly, K. K. 2012. A comparative assessment of supervised data mining techniques for fraud prevention. *TIST International Journal for Science, Technology & Research* 1. Dostopno prek: http://www.tochjournal.in/journals/FraudDetection_newformat_.pdf (9. september 2013).

Škulj, Damjan. 2006. *Statistika 2: Prosojnice s predavanj*. Ljubljana: Fakulteta za družbene vede, 19. maj.

SPSS. 2000. *Clementine 10.0 Desktop User's Guide*. Chicago, IL: SPSS Inc.

Šubelj, Lovro, Marko Bajec in Matjaž Kukar. 2008. *Odkrivanje goljufij na osnovi analize socialnih mrež, diplomsko delo*. Ljubljana: Fakulteta za računalništvo in informatiko, Fakulteta za matematiko in fiziko.

Tennyson, Sharon. 2008. Moral, social and Economic Dimensions of Insurance Claims Fraud. *Social Research* 75 (4): 1181–1204.

Thiruvadi, Sheela in Sandip C. Patel. 2011. Survey of Data-mining Techniques used in Fraud Detection and Prevention. *Information Technology Journal* 10 (4): 710–716.

Westpahl, Christopher. 2009. *Data mining for intelligence, fraud & criminal detection: advanced analytics & information sharing technologies*. Boca Raton, FL: CRC Press.

Whitaker, James E. 2009. *Insurance Fraud Handbook*. Austin, TX: ACFE.

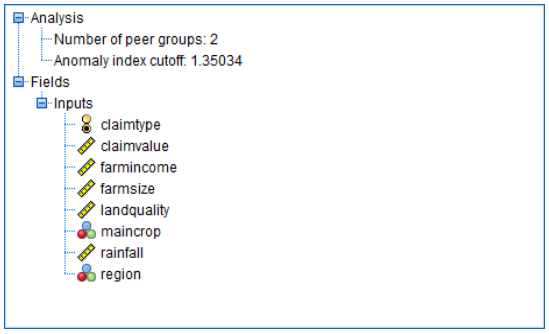
--- 2013. *Zavarovalniške goljufije – trenutno stanje in izzivi*. Ljubljana, Grand Hotel Union, 14. maj 2013.

Witten, Ian H., Eibe Frank in Mark A. Hall. 2011. *Data Mining: Practical Machine Learning Tools and Techniques: Third Edition*. Burlington, MA: Elsevier Inc.

Zacharias, Greg L., Jean MacMillan in Susan B. Van Hemel. 2008. *Behavioral Modeling and Simulation: From Individuals to Societies*. Washington, DC: The National Academies Press.

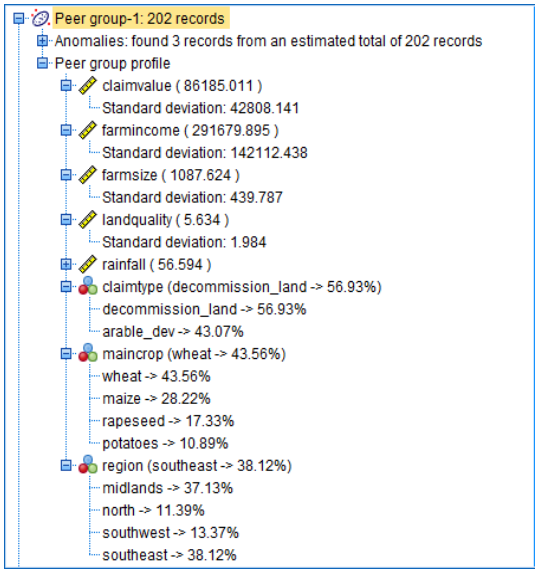
PRILOGE

Priloga A: Osnovne informacije modela

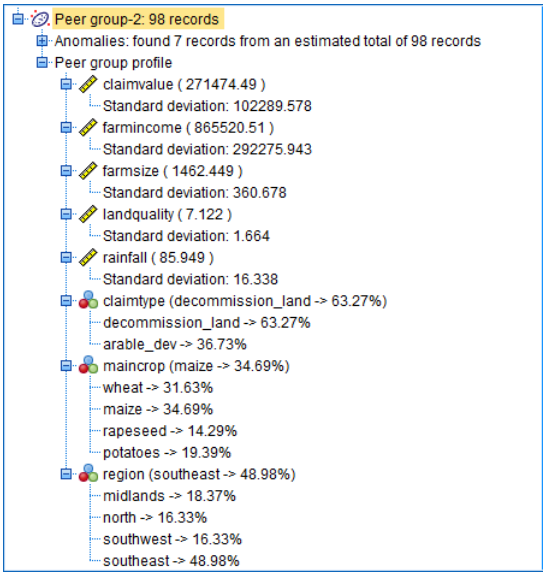


Priloga B: Statistike modeliranja zveznih in nominalnih spremenljivk za obe skupini

Prva skupina



Druga skupina



Priloga C: Prikaz zbirke podatkov sumljivih zavarovalniških primerov

	id	$O-Anomaly	$O-AnomalyIndex	$O-PeerGroup	$O-Field-1	$O-FieldImpact-1	$O-Field-2	$O-FieldImpact-2	$O-Field-3	$O-FieldImpact-3
1	id601	F	0.825	1	landquality	0.206	region	0.178	rainfall	0.169
2	id602	F	1.285	2	rainfall	0.287	region	0.196	maincrop	0.115
3	id603	F	0.917	1	maincrop	0.283	farmsize	0.209	region	0.160
4	id604	F	1.323	2	landquality	0.303	region	0.191	maincrop	0.173
5	id605	F	1.244	1	region	0.258	farminco...	0.177	farmsize	0.161
6	id606	F	0.917	1	maincrop	0.204	farmsize	0.172	region	0.156
7	id607	F	0.980	1	farmsize	0.259	maincrop	0.191	rainfall	0.147
8	id608	F	1.080	2	claimvalue	0.299	maincrop	0.137	claimtype	0.130
9	id609	F	0.971	2	region	0.260	maincrop	0.166	landquality	0.155
10	id610	F	0.784	1	farmsize	0.199	region	0.182	rainfall	0.160

Priloga D: Statistika skupin zbirk podatkov

Prva skupina

Peer group-1: 202 records
Anomalies: found 3 records from an estimated total of 202 records

Contribution	Count	Average index
region	2	0.2
landquality	2	0.176
farmsize	1	0.155
maincrop	2	0.197
rainfall	2	0.197

Residual of the unreported reasons: 43.48%
Peer group profile
claimvalue (86185.011)
farmincome (291679.895)
farmsize (1087.624)
landquality (5.634)
rainfall (56.594)
claimtype (decommission_land -> 56.93%)
maincrop (wheat -> 43.56%)
region (southeast -> 38.12%)
Peer group-2: 98 records
Anomalies: found 7 records from an estimated total of 98 records
Peer group profile

Druga skupina

Peer group-1: 202 records
Peer group-2: 98 records
Anomalies: found 7 records from an estimated total of 98 records

Contribution	Count	Average index
region	3	0.163
claimvalue	6	0.274
maincrop	6	0.139
rainfall	1	0.322
farmincome	5	0.257

Residual of the unreported reasons: 34.67%
Peer group profile
claimvalue (271474.49)
farmincome (865520.51)
farmsize (1462.449)
landquality (7.122)
rainfall (85.949)
claimtype (decommission_land -> 63.27%)
maincrop (maize -> 34.69%)
region (southeast -> 48.98%)

Priloga E: Statistike indeksa anomalije sumljivih zahtevkov

Skupaj

$O-AnomalyIndex
Statistics

Count	10
Mean	1.512
Min	1.350
Max	1.770
Range	0.420
Variance	0.023
Standard Deviation	0.151
Standard Error of Mean	0.048

Prva skupina

$O-AnomalyIndex
Statistics

Count	3
Mean	1.462
Min	1.358
Max	1.641
Range	0.283
Variance	0.024
Standard Deviation	0.156
Standard Error of Mean	0.090

Druga skupina

$O-AnomalyIndex
Statistics

Count	7
Mean	1.533
Min	1.350
Max	1.770
Range	0.420
Variance	0.024
Standard Deviation	0.156
Standard Error of Mean	0.059

Priloga F: Statistike indeksa anomalije nesumljivih zahtevkov

Skupaj

$O-AnomalyIndex
Statistics

Count	290
Mean	0.982
Min	0.652
Max	1.335
Range	0.683
Variance	0.022
Standard Deviation	0.147
Standard Error of Mean	0.009

Prva skupina

$O-AnomalyIndex
Statistics

Count	199
Mean	0.993
Min	0.652
Max	1.335
Range	0.683
Variance	0.021
Standard Deviation	0.146
Standard Error of Mean	0.010

Druga skupina

$O-AnomalyIndex
Statistics

Count	91
Mean	0.959
Min	0.705
Max	1.323
Range	0.618
Variance	0.022
Standard Deviation	0.149
Standard Error of Mean	0.016

Priloga G: Uspešnost modela

Results for output field $O-Anomaly
Comparing $N-$O-Anomaly with $O-Anomaly

Correct	26	86.67%
Wrong	4	13.33%
Total	30	

Coincidence Matrix for $N-$O-Anomaly (rows show actuals)

	F	T
F	20	0
T	4	6

Performance Evaluation

F	0.223
T	1.099