Rudarenje podataka u svrhu otkrivanja sumnjivih bankovnih transakcija

Tadić, Anita

Sveučilište u Zagrebu

Ekonomski fakultet

Integrirani preddiplomski i diplomski sveučilišni studij

Poslovna ekonomija – smjer Menadžerska informatika

RUDARENJE PODATAKA U SVRHU OTKRIVANJA SUMNJIVIH BANKOVNIH TRANSAKCIJA

Diplomski rad

Anita Tadić

Zagreb, travanj 2024.

Sveučilište u Zagrebu

Ekonomski fakultet

Integrirani preddiplomski i diplomski sveučilišni studij

Poslovna ekonomija – smjer Menadžerska informatika

RUDARENJE PODATAKA U SVRHU OTKRIVANJA SUMNJIVIH BANKOVNIH TRANSAKCIJA

DATA MINING FOR THE PURPOSE OF DETECTING SUSPICIOUS BANK TRANSACTIONS

Diplomski rad

Student: Anita Tadić

JMBAG studenta: 0067586566

Mentor: prof. dr. sc. Mirjana Pejić Bach

Zagreb, travanj 2024.

IZJAVA O AKADEMSKOJ ČESTITOSTI

Izjavljujem i svojim potpisom potvrđujem da je diplomski rad isključivo rezultat mog vlastitog rada koji se temelji na mojim istraživanjima i oslanja se na objavljenu literaturu, a što pokazuju korištene bilješke i bibliografija.

Izjavljujem da nijedan dio rada nije napisan na nedozvoljen način, odnosno da je prepisan iz necitiranog izvora te da nijedan dio rada ne krši bilo čija autorska prava.

Izjavljujem, također, da nijedan dio rada nije iskorišten za bilo koji drugi rad u bilo kojoj drugoj visokoškolskoj, znanstvenoj ili obrazovnoj ustanovi.

_____

(vlastoručni potpis studenta)

_____

(mjesto i datum)

STATEMENT ON THE ACADEMIC INTEGRITY

I hereby declare and confirm by my signature that the final thesis is the sole result of my own work based on my research and relies on the published literature, as shown in the listed notes and bibliography.

I declare that no part of the thesis has been written in an unauthorized manner, i.e., it is not transcribed from the non-cited work, and that no part of the thesis infringes any of the copyrights.

I also declare that no part of the thesis has been used for any other work in any other higher education, scientific or educational institution.

_____

(personal signature of the student)

_____

(place and date)

SAŽETAK

Financijske institucije se, u modernom digitalnom dobu, susreću sa raznim preprekama u poslovanju, a glavne su transakcije koje vode prijevarama i krađi identiteta. Primjenama metoda za rudarenje podataka pokušava se preventivno djelovati na možebitne negativne posljedice koje sumnjive transakcije mogu uzrokovati. Banke u poslovanje uvode metode rudarenja podataka kako bi efikasno i efektivno djelovale na sigurnost svojih klijenata. Sve veća digitalizacija dovodi i do većih količina podataka, tako da je bankama nužno uvođenje metoda koje će brzo prepoznati potencijalne probleme i pokušati riješiti iste. Sumnjive transakcije se događaju iz dana u dan, a da bi se pokušale spriječiti potrebno je preventivno djelovati.

Ključne riječi: financijske institucije, banke, sumnjive transakcije, baza podataka, otkrivanje znanja

ABSTRACT

In the modern digital age, financial institutions encounter various challenges in their operations, with fraudulent transactions and identity theft being the main concerns. By using data mining methods, banks attempt to proactively address the potential negative consequences that suspicious transactions can cause. Banks are incorporating data mining methods into their operations to enhance the security of their clients effectively and efficiently. The increasing digitalization results in larger data volumes, making it essential for banks to implement methods that could quickly recognize potential issues and attempt to resolve them. Suspicious transactions occur daily, so preventive actions are necessary to prevent them.

Keywords: financial institutions, banks, suspicious transactions, databases, knowledge discovery

SADRŽAJ

1. UVOD

U današnjem digitalnom dobu, financijske institucije suočavaju se sa izazovima koji su vezani za otkrivanje i suzbijanje nezakonitih aktivnosti koje uključuju bankovne transakcije. Porastom kompleksnosti financijskih sustava tradicionalni sustavi postaju ograničeni u otkrivanju sumnjivih aktivnosti. U ovom kontekstu, rudarenje podataka predstavlja moćan alat koji omogućava analizu velikih skupova podataka kako bi se identificirale nepravilnosti transakcija.

1.1. Predmet i cilj rada

Cilj ovoga rada je korištenjem tehnika rudarenja podataka pokušati identificirati odstupanja i nepravilnosti pri određenim bankovnim transakcijama te na temelju toga prepoznati one sumnjive. Isto tako, cilj je istražiti primjenu rudarenja podataka u bankovnim sustavima te prikazati konkretne primjere primjene istoga na stvarnim i dostupnim podacima.

Istražit će se kako rudarenje podataka može pridonijeti poboljšanju procesa otkrivanja sumnjivih bankovnih transakcija, identificirati specifične metode analize podataka koje su najučinkovitije te bi dobiveni rezultati istraživanja mogli poslužiti u svrhu potencijalnih budućih promjena u bankovnim sustavima ili kao smjernice u radu i poslovanju.

1.2. Izvor podataka i metode prikupljanja

Kao izvor podataka koristit će se dostupna statistika i informacije o obavljenim transakcijama. Također, koristit će se i podaci o statistici sumnjivih transakcija, napadima, pokušajima krađe i slično. Analizirat će se podaci o transakcijama iz baza podataka te primijeniti određene metode kako bi se otkrile sumnjive bankovne transakcije. Podaci korišteni u cijelom sadržaju rada nastali su istraživanjem službene literature Ekonomskog fakulteta u Zagrebu, akademskih članaka, knjiga i ostalih relevantnih izvora koji su mogli poslužiti kao baza za nastali rad.

1.3. Metodologija istraživanja

Iako se u svijetu događa puno neovlaštenih i sumnjivih transakcija, ne postoji relevantna baza podataka koja može pouzdano prikazati način i nastanak istih. Ono što se može pronaći su periodične transakcije nastale u nekom dijelu svijeta. Baza koja će se koristiti za obradu podataka, u ovom radu, pronađena je na Web stranici 'Kaggle' koja sadrži baze podataka u raznim formatima, o većini aktualnih tema današnjice. Odabir ove baze podataka opravdan je njenom dostupnošću i obuhvatom relevantnih podataka za istraživanje transakcija putem kreditnih kartica. Rad će razmatrati pristup analizi transakcija kreditnih kartica s naglaskom na rudarenje podataka. Metoda rudarenja podataka koja će se koristiti u ovom radu je metoda stabala odlučivanja, a ista će poslužiti u analizi i interpretaciji pronađenih podataka. Metoda stabala odlučivanja je odabrana zbog sposobnosti iste da jasno interpretira i vizualizira procese donošenja odluka te da istovremeno pruži dobre rezultate u analizi podataka. Metodologija istraživanja temelji se na analizi podataka transakcija kreditnih kartica. Podaci će se preuzeti iz dostupne online baze podataka, a već dolaze pripremljeni u Excelu u .csv formatu, a potrebno ih je prvo pregledati u Excelu radi lakšeg razumijevanja. Nakon toga, budući da su već u traženom formatu, učitat će se u softverski alat Weka. Ovaj alat omogućuje naprednu analizu podataka i primjenu različitih tehnika strojnog učenja kako bi se identificirali uzorci, trendovi i relevantne informacije iz podataka o transakcijama kreditnih kartica. Kao što je već spomenuto, za obradu podataka koristit će se metoda stabala odlučivanja, ali s obzirom da su podaci neujednačeni, pokušat će se i putem drugih metoda navedenog alata, doći do što boljih rezultata.

1.4. Sadržaj i struktura rada

Rad se sastoji od pet glavnih dijelova, odnosno naslova. Izuzev uvoda i zaključka, rad obuhvaća sljedeće dijelove: rudarenje podataka, primjena rudarenja podataka u bankovnom sektoru, rudarenje podataka u svrhu otkrivanja sumnjivih bankovnih transakcija.

Uvod se sastoji od opisa predmeta i cilja rada, izvora podataka, kao i metoda i metodologije istraživanja.

Prvi dio nakon uvoda pruža pregled općenitog opisa rudarenja podataka te definira procese, modele i metode rudarenja podataka.

Nakon općenitog opisa slijedi i konkretan opis primjene rudarenja podataka u bankovnom sektoru, a počevši sa uvodom i trendovima u bankovnom sektoru.. Prikazat će se i područje primjene rudarenja podataka, kao i važnost primjene istoga. Osim toga, prikazat će se i detaljno opisati metode rudarenja podataka u bankovnom sektoru, njihove primjene i važnosti korištenja pri istraživanjima.

Glavni dio rada predstavlja rudarenje podataka u svrhu otkrivanja sumnjivih bankovnih transakcija, a započinje metodologijom istraživanja. Nakon metodologije istraživanja opisat će se podaci koji će se koristiti u samom istraživanju. Rezultati istraživanja temelj su za preporuke za bankovno poslovanje, a također su i zadnji dio ovog rada, prije samog zaključka.

2. RUDARENJE PODATAKA

U modernom digitalnom dobu organizacije se suočavaju sa ogromnim količinama podataka koji svakodnevno kruže. U cilju prikupljanja korisnih podataka i informacija uvelike pomaže rudarenje podataka.

Rudarenje podataka predstavlja proces sortiranja velikih setova podatka koji služe kako bi se identificirali uzorci i uspostavile veze za rješavanje problema, a sve kroz analizu podataka (Šimec, 2020).

Također, predstavlja proučavanje prikupljanja, čišćenja, obrade, analize i stjecanja korisnih spoznaja iz podataka. Rudarenje podataka koristi se kao opći pojam koji opisuje različite aspekte obrade podataka, a u modernom digitalnom dobu gotovo svi automatizirani sustavi generiraju neki oblik podataka ili u svrhu dijagnostike ili u svrhu analize (Aggarwal, 2015).

Rudarenje podataka se ne bavi samo analizom strukturiranih podataka iz postojećih baza podataka, već i nestrukturiranim podacima poput zvuka, teksta ili slika. Koristi napredne statističke, matematičke i računalne tehnike za sami proces rudarenja (Larose i Larose, 2014).

2.1. Uvod u rudarenje podataka

Banke svakodnevno bilježe velike količine podataka koji uključuju informacije o računima, transakcijama, kreditnim obvezama i demografskim podacima. Prikupljeni podaci bilježe se u transakcijske baze podataka koje obavljaju tri osnovne funkcije, a u iste spadaju: vođenje evidencije poslovnih događaja, generiranje poslovnih dokumenata te izvještavanje o stanju poslovnih procesa (Pejić Bach, 2005).

Napretkom tehnologije i digitalizacije života nastaje sve više podataka koje je potrebno obraditi. Skupljaju se sirovi podaci koji mogu biti proizvoljni i nestrukturirani i u formatu koji nije pogodan za automatsku obradu. Takvi podaci mogu potjecati iz različitih izvora, a iste je potrebno obraditi kako bi se pretvorili u standardizirani format, a potom rudarenjem podataka iz takvih podataka mogu se izvući potrebni podaci (Demetis, 2018).

Rudarenje podataka predstavlja ključni aspekt doba podataka u kojem živimo. S obzirom na eksplozivni rast dostupnih podataka, postoji potreba za snažnim alatima koji automatski otkrivaju vrijedne informacije i pretvaraju ih u organizirano znanje. Rudarenje podataka je proces otkrivanja zanimljivih obrazaca i modela u velikim skupovima podataka. Također, omogućava automatsko otkrivanje vrijednih informacija iz ogromnih količina podataka, poput poslovnih transakcija, znanstvenih eksperimenata i medicinskih istraživanja. Ovaj proces pomaže transformirati sirove podatke u organizirano znanje, čime pridonosi donošenju informiranih odluka u poslovnom svijetu, znanstvenim istraživanjima te medicinskim i drugim industrijskim sektorima. Rudarenje podataka ima ključnu ulogu u evoluciji prema dobu informacija, gdje se naglasak stavlja na efikasno iskorištavanje ogromnih podatkovnih resursa radi postizanja dubljeg razumijevanja i optimizacije različitih aspekata društva (Han et al., 2022).

Rudarenjem podataka otkrivaju se odnosi, logika, pravilnosti te općenito strukture među podacima. Ključni aspekt rudarenja podrazumijeva organiziranje i čišćenje podataka kako bi se pristupilo znanju i steklo isto na temelju postojećih podataka u bazama. Unapređenje tehnologije, računala i interneta značajno olakšava organizaciju podataka, ali kako bi oni postali korisni, nužno je njihovo pretvaranje u informacije i znanje (Rovčanin et.al., 2012).

2.2. Proces rudarenja podataka

Proces rudarenja podataka možemo opisati kao aktivnost pronalaženja korisnih informacija i znanja unutar velikih skupova podataka. Ova metodologija poboljšava proces donošenja odluka na strateško-poslovnoj razini, pružajući uvid u "skrivene" podatke putem business intelligence (BI) pristupa (Rovčanin et.al., 2012).

Porast količine podataka dovodi do povećanog interesa, a i potrebe za obradom istih. Time se dolazi do samog procesa obrade podataka koji se sastoji nekoliko koraka, a to su redom: definicija poslovnog problema, priprema podataka, modeliranje, implementacija.

U samom procesu rudarenja podataka moguće je u svakom trenutku sa nekog koraka vratiti se na prethodni jer je proces iterativan. Povratak na prethodni korak je zapravo više pozitivna nego negativna stvar jer je odabir podataka i tehnike, odnosno definiranja problema najbitnije kako bi se proces kvalitetno odradio (Pejić Bach, 2005).

2.2.1. Definiranje poslovnog problema

Prvi korak u procesu rudarenja podataka je definiranje poslovnog problema i izražavanje tog problema u obliku pitanja koja će biti odgovorena na kraju procesa. Analizom područja gdje je rudarenje podataka već uspješno primijenjeno, može se pronaći najbolji pristup definiranju poslovnog problema. Na temelju uspješnih primjera, odabire se područje od najveće važnosti za poduzeće. U ovom koraku također se određuju i odabiru osobe koje će sudjelovati u projektu rudarenja podataka, uključujući stručnjaka za rudarenje podataka, informatičara koji poznaje baze podataka banke, te bankarskog stručnjaka upućenog u potencijalnu primjenu. Važno je da na vrhu bude ključna osoba iz menadžmenta, koja možda neće izravno raditi na projektu, ali će pružiti podršku i pomoći u rješavanju eventualnih prepreka i poteškoća (Pejić Bach, 2005).

Definiranje poslovnog problema ima za cilj precizno odrediti što se želi postići analizom podataka i kako će ta analiza doprinijeti rješavanju poslovnog izazova. Kako je već navedeno postoje ključni aspekti procesa rudarenja podataka, odnosno definiranja poslovnog problema. Kod analize poslovnog konteksta identificira se poslovni kontekst u kojem se rudarenje podataka primjenjuje te se pokušava razumjeti specifičnost industrije, sektora ili područja poslovanja. Nakon analize poslovnog konteksta bitna je identifikacija ciljeva i izazova u kojem se jasno definiraju poslovni ciljevi i izazovi, odnosno utvrđuju problemi koji bi se mogli riješiti analizom podataka. Identifikacijom ciljeva i izazova dolazi se do formulacije poslovnog problema u obliku pitanja koja će se rješavati tijekom procesa rudarenja podataka. Formulacijom poslovnog problema i postavljanjem poslovnih ciljeva dolazi se do analize dostupnih podataka koji se potom proučavaju te se identificiraju relevantni izvori podataka koji će se koristiti u analizi. Potom dolazi do provjere utjecaja rješenja poslovnog problema na ostvarenje, odnosno doprinos ostvarenju strateških ciljeva organizacije. Kako bi se svi ovi aspekti ispunili te osigurala cjelovitost i relevantnost problema, potrebno je, kako je već spomenuto, uključiti ključne sudionike procesa (Marbán et.al., 2009).

2.2.2. Priprema podataka

Drugi korak procesa rudarenja podataka je priprema podataka. Cilj ovoga koraka je osigurati da odabrani podaci budu prikladni za analitičke postupke. Ključni aspekti drugog koraka, odnosno pripreme podataka su, kao prvo, sakupljanje podataka iz različitih izvora, a koji uključuju baze

podataka, datoteke, vanjske izvore podataka i druge primjerene resurse. Sakupljene podatke iz različitih izvora je potrebno agregirati i obraditi kako bi se stvorila sveobuhvatna slika. Nakon sakupljanja slijedi čišćenje podataka, a podrazumijeva detekciju nedostajućih podataka, rješavanje dupliciranih i nekonzistentnih podataka te eliminaciju ili korekciju anomalija i pogrešaka u podacima. Da bi se stvorila baza podataka potrebno je očišćene podatke integrirati, a potom i uspostaviti ključeve povezivanja između različitih skupova podataka. Transformacija, odnosno pretvaranje podataka u odgovarajući format je sljedeći aspekt ovog koraka. Kada su podaci pretvoreni u odgovarajući format slijedi selekcija značajki koje će biti korištene u analizi, a nakon toga razvoj podatkovnog modela koji podrazumijeva projektiranje strukture podatkovnog modela te definiranje organizacije i pohrane podataka. Sljedeći segment je provjera kvalitete podataka kako bi se osigurala točnost, dosljednost i pouzdanost. Nakon provjere potrebno je segmentirati podatke na relevantne segmente ili grupe koje olakšavaju analizu istih, a iza toga i osigurati učinkovitu analizu upravljanjem velikim volumenom podataka. Korak pripreme podataka je ključan jer utječe na kvalitetu rezultata analize, a što je priprema podataka temeljitija i pažljivija stvara se pouzdaniji i relevantniji model te olakšava interpretaciju rezultata (Panian, 2007).

2.2.3. Modeliranje

Modeliranje je treći korak procesa rudarenja podataka, a uključuje izbor odgovarajućih modela i algoritama kako bi se analizirali podaci i izgradili prediktivni modeli. Prvi aspekt trećeg koraka procesa rudarenja je izbor modela, a pod time se smatra identifikacija vrste problema koji se rješava te odabir odgovarajućeg modela ovisno o ciljevima analize. Nakon izbora modela dolazi do podjele podataka te izbora značajki koje će se koristiti u modeliranju, tako što se razmatra važnost tih značajki za poboljšanje performansi modela. Za odabrane stavke je potrebna i optimizacija kako bi se postigla što bolja prilagodba podacima. Nakon toga slijedi izgradnja modela primjenom odabranog algoritma te evaluacija performansi korištenjem određenih mjera. Prije interpretacije rezultata potrebno je provjeriti valjanost modela, a i optimizirati sami model kako bi se poboljšale performanse ili prilagodile promjene u podacima. Model je potrebno kontinuirano pratiti i optimizirati tijekom vremena (Pejić Bach, 2005).

Nakon svih provjera dolazi se i do zadnjeg aspekta ovog koraka procesa rudarenja podataka, a to je implementacija, odnosno integracija modela u stvarno poslovno okruženje. Ovaj korak u procesu

rudarenja podataka je ključan kako za ostvarivanje ciljeva analize tako i za osiguranje pouzdanih i korisnih rezultata.

2.2.4. Implementacija

Četvrti korak procesa odnosi se na implementaciju rezultata koji su dobiveni modeliranjem, a uključuje primjenu i integraciju modela u stvarno poslovno okruženje, kao i interpretaciju rezultata i njihovo korištenje. U fazi implementacije ključnu ulogu igra interpretacija rezultata, a za koje je bitno da budu u jednostavnom obliku kako bi iščitavanje istih bilo jednostavno. Model je potrebno integrirati u postojeći sustav poduzeća te osigurati učinkovito funkcioniranje i komuniciranje sa drugim aplikacijama i sustavima poduzeća. Implementacija modela kao zadnji korak procesa rudarenja podataka, značajan je korak jer određuje stvarnu vrijednost koju model pruža organizaciji, odnosno poduzeću. Održavanje i pravilna implementacija predstavljaju glavnu ulogu u dugoročnom uspjehu u primjeni rudarenja podataka u poslovnom kontekstu (Panian, 2007).

2.3. Modeli i metode rudarenja podataka

Rudarenje podataka ima dva ključna cilja koji se mogu grupirati u dvije osnovne kategorije, a to su predviđanje i deskripcija.

Kod predviđanja naglasak je na predviđanju budućih vrijednosti varijabli, tako što se koriste postojeći podaci kao osnova. U drugoj kategoriji, odnosno deskripciji, fokus je na identifikaciji uzoraka koji se nalaze unutar podataka, a s ciljem objašnjenja ponašanja cijelog sustava. Deskriptivna analiza omogućuje dublje razumijevanje podataka i interpretaciju ključnih faktora koji utječu na ishode (Bhambri, 2011).

Unatoč raznolikosti metoda, mogu se odrediti tri najvažnije, odnosno tri osnovne metode, a to su klasifikacija, klasteriranje te asocijacija.

Klasifikacija je usmjerena na razvrstavanje podataka u određene kategorije koje su unaprijed definirane. Kao primjer može se navesti razvrstavanje osoba koje traže kredit u kategorije niskog, srednjeg ili visokog rizika. Kako bi se postigla navedena klasifikacija algoritam rudarenja podataka analizira podatke o prethodnim korisnicima kredita i identificira karakteristike koje su tipične za korisnike koji nisu redovito vraćali kredit, a na temelju kojih banka može klasificirati tražitelje

kredita u određene kategorije i tako prilagoditi zahtjeve za osiguranjem povrata sredstava prema razini rizika koju svaka kategorija predstavlja (Scott et al., 1998).

Klasifikacija je ključna metoda u području rudarenja podataka i ima široku primjenu u mnogim područjima, kao što su financije, medicina, marketing i ostala područja u kojima se analiziraju podaci kako bi se donosile odluke. Svrha klasifikacije je predviđanje pripadnosti podataka određenoj klasi na temelju njegovih karakteristika. Kako bi se podaci svrstali određene su klase i atributi, klase predstavljaju kategorije u koje se podaci trebaju svrstati, a atributi su zapravo karakteristike podataka koji se koriste za donošenje odluka. Da bi se klasifikacijski model izgradio potrebno je testirati model na neovisnom testnom skupu kako bi se procijenila njegova točnost i učinkovitost. Kako bi se postupak klasifikacije uspješno proveo potrebno je na početku prikupiti podatke, potom ih pripremiti, odnosno očistiti i pripremiti u oblik prilagođen za analizu podataka. Potom izabrati atribute te izgraditi model koji se na kraju evaluira, a može se i optimizirati prilagodbom parametara ili promjenom algoritma za postizanje boljih performansi (Singh i Agray, 2017).

Binarna klasifikacija je osnovni tip klasifikacijskog problema, a ciljni atribut ima samo dvije moguće vrijednosti. Te vrijednosti mogu biti, npr. visoka i niska stopa. Postoje sa druge strane i višestruki ciljevi koji imaju više vrijednosti, npr. niska, srednja, visoka stopa. Dok je model u procesu izgradnje, klasifikacijski algoritmi traže veze između vrijednosti prediktora i vrijednosti cilja. Različiti algoritmi klasifikacije koriste i različite tehnike za otkrivanje veza, a one su sežete u modelu koji se može primijeniti na neki drugi skup podataka sa nepoznatim zadacima klasa. Podaci za klasifikacijski projekt većinom su podijeljeni u dva skupa od kojih se jedan koristi za izgradnju modela, a drugi za testiranje, odnosno ispitivanje modela. Ova podjela pomaže u procjeni učinkovitosti modela i njegovoj sposobnosti klasifikacije (Singh i Agray, 2017).

Algoritmi koji se koriste u metodi klasifikacije su stabla odlučivanja, statističke metode, kao što su Naive Bayes klasifikator i regresijski modeli, zatim algoritmi za učenje pravila i njihove kombinacije, vještačke neuronske mreže, mehanizmi podrške vektora, J48, OneR i ostali (Panian i Spremić, 2007).

Druga metoda je klasteriranje, a uključuje svrstavanje objekata u kategorije. U ovoj metodi kategorije nisu unaprijed definirane pa je taj zadatak malo izazovniji. Kao primjer za ovu metodu Žapčević i Butala (2015) navode klasifikaciju kupaca u određene skupine, a zatim prilagodbu

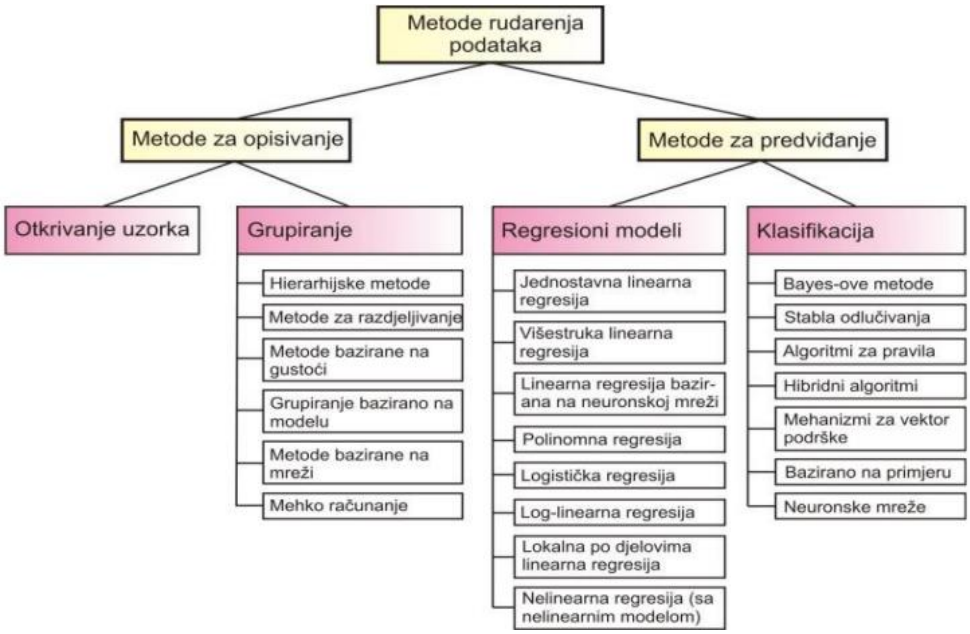
različitih marketinških strategija za svaku skupinu. Kupci se razlikuju po kriterijima koji uključuju ukuse, uvjerenja, stilove kupovine i profitabilnost iste. Tako da je nužno prilagoditi tretiranje kupaca prema njihovim specifičnim karakteristikama kako bi se postigao optimalan rezultat.

Klasteriranje se bavi grupiranjem objekata ili instanci u skupine koje nazivamo klasteri. Objekti unutar jednog klastera su međusobno slični, dok se razlikuju od objekata iz drugog klastera, odnosno slični objekti se grupiraju zajedno. Mjera sličnosti je ključna za određivanje klase objekata. Kao što je već navedeno, nema unaprijed definiranih klasa nego se algoritmom po sličnosti objekti raspoređuju. Nakon pripreme podataka i izbora atributa potrebno je izabrati algoritam za raspored po klasama, a neki od njih su K-means klasteriranje, EM algoritam, hijerarhijsko klasteriranje. Metoda klasteriranja je ključna tehnika rudarenja podataka i koristi se u područjima strojnog učenja, analize slika, prepoznavanja uzoraka te pronalaženja informacija i slično. Već spomenuti algoritmi klasteriranja koriste se u svrhu ove analize, a bitno se razlikuju u načinu definiranja klastera i učinkovitosti pronalaženja sličnih objekata. Kod formiranja klastera dijele se skupovi podataka gdje pripadnost grupi određuje sličnost značajki te se unutar zadane populacije korištenjem algoritama pronalaze te sličnosti. Najpoznatiji od spomenutih algoritama klasteriranja je K-means koji kroz iterativni proces koristeći funkcije za procjenu distance i centroida stvara klastere. Nakon ovoga, najčešće korišten je hijerarhijski algoritam (Han et al., 2022).

Treća metoda je asocijacija, a bavi se istraživanjem pitanja koje se stvari događaju istovremeno. Na primjeru potrošačke košarice može se proučiti koji se proizvodi često zajedno kupuju pa se analizom podataka mogu otkriti neke očigledne veze u kupnji dva proizvoda zajedno. Asocijacijom je moguće identificirati i neke skrivene povezanosti, koje nisu toliko očigledne. Asocijacija se temelji na istraživanju simultanih događanja ili pojava unutar skupa podataka, a kao što je već navedeno, osnovni cilj je otkriti povezanosti između različitih atributa. Pravila asocijacije se često predstavljaju u obliku uvjeta 'ako' i rezultata 'onda'. Takav tip rudarenja podataka omogućuje identifikaciju korisnih veza između različitih varijabli te često otkriva i one skrivene i nejasne. Za procjenu sličnosti često se koriste mjere Euklidske ili Manhattan udaljenosti. Primjene asocijativnog rudarenja su brojne i nalaze se u raznim sektorima, a ponajviše u trgovini, medicini, marketingu i sličnim istraživanjima. Asocijativno rudarenje obuhvaća identifikaciju često

ponavljajućih uzoraka u velikim skupovima podataka te otkrivanje asocijacija pomaže u stvaranju strategija i pravila poslovanja temeljenih na stvarnim podacima (Han et al., 2022).

Slika 1 Metode rudarenja podataka



Izvor: Žapčević, S., Butala, P. (n.d.) OTKRIVANJE ZNANJA I METODE RUDARENJA PODATAKA U PROIZVODNIM SISTEMIMA

Prema Žapčeviću i Butali (n.d.) (slika 1) metode rudarenja podataka dijele se na metode za opisivanje i metode za predviđanje. U metode za opisivanje spada otkrivanje uzoraka i grupiranje, dok u metode za predviđanje ulaze regresijski modeli i klasifikacija. Kod metoda za opisivanje grupiranje obuhvaća metode kao što su hijerarhijske metode, metode za razdjeljivanje, metoda bazirane na gustoći, grupiranje bazirano na modelu, metode bazirane na mreži te meko računanje. U regresijske modele metoda za predviđanje spadaju jednostavna linearna regresija, višestruka linearna regresija, linearna regresija bazirana na neuronskoj mreži, polinomna regresija, logistička regresija, log – linearna regresija, lokalna po dijelovima linearna regresija te nelinearna regresija sa nelinearnim modelom. Na drugoj strani, klasifikacija obuhvaća Bayes – ove modele, stabla odlučivanja, algoritme za pravila, hibridne algoritme, mehanizme za vektor podrške, klasifikacije bazirane na primjeru te neuronske mreže.

3. PRIMJENE RUDARENJA PODATAKA U BANKOVNOM SEKTORU

U ovom poglavlju istražit će se područje otkrivanja znanja iz baza podataka, a koje je neizostavan alat za obradu i analizu velikih količina podataka i pretvorbu istih u smislene podatke i korisne informacije.

Na početku će se razmotriti sama situacija u financijskoj industriji, odnosno u bankovnom sektoru kako bi se spoznale promjene koje ista doživljava u dobu digitalizacije. Nakon spoznaje problema, sljedeći dio je uvod u otkrivanje znanja iz baza podataka u bankarstvu. U ovom dijelu pokušat će se objasniti što zapravo predstavljaju baze podataka bankama, otkud dolaze i zašto su važne u procesu rudarenja podacima. Uvod o otkrivanju znanja iz baza podataka prate područja primjene istih u bankarstvu. Kroz opis i navođenje područja primjene objasnit će se svrha rudarenja podataka, a nakon toga i važnost primjene otkrivanja znanja iz baza podataka. Sve u cilju opisa pozitivnih strana za poslovanje pri implementaciji i korištenju. Opisat će se i modeli i metode primjene kako bi se pojam istih približio i detaljnije pojasnio prije samog procesa obrade podataka.

3.1. Trendovi u bankovnom sektoru

Financijska industrija diljem svijeta doživljava značajne promjene, a najveća od njih je zamjena tradicionalnog osobnog pristupa i kontakta s klijentima onim elektroničkim. Ovakva zamjena ima za cilj smanjenje vremena i troškova obrade zahtjeva te poboljšanje performansi sustava. Upotrebom interneta, računalizacijom financijskih sustava i automatizacijom softvera mijenjaju se osnovni koncepti poslovanja. U bankovnom sektoru, od prošlog stoljeća, nastaju velike promjene sa potpunim prelaskom na online transakcije, centralizirane baze podataka i bankomate. Ove promjene tehnički olakšavaju korištenje usluga banke i klijenti ih rado prihvaćaju. Rudarenje podataka u bankarstvu omogućuje izvlačenje vrijednih informacija iz velikih operativnih baza podataka kako bi se unaprijedile performanse i učinkovitost. Rudarenjem podataka postiže se i donošenje boljih odluka, povećanje vrijednosti korisnika te komunikacija i zadovoljstvo klijenata. Razvojem mobilnog bankarstva i e-bankinga eksponencijalno rastu i informacije u stvarnom vremenu. Takav kontinuirani razvoj i rast dostupnosti velikih podataka vode većoj potrebi za rudarenjem podataka i čine ga jednim od najvažnijih alata u sektoru bankarstva (Stanišić, 2007).

Tehnike rudarenja podataka u bankarstvu najviše se koriste za sigurnost i otkrivanje prijevara praćenjem i analizom sekundarnih podataka, kao što su transakcijski zapisi kako bi se poboljšala sigurnost u bankarstvu i prepoznala neuobičajena ponašanja koji ukazuju na prijevaru, pranje novca i slično. Također, rudarenje podataka u bankarstvu koristi se za upravljanje rizicima i investicijsko bankarstvo, tako što se analiziraju interni podaci o kreditnim karticama te se tako bankama omogućuje procjena kreditnog rizika i odobravanje kredita (Bhambri, 2011).

Bankarska i financijska industrija je izrazito konkurentska te je podložna političkim i ekonomskim uvjetima u domaćim zemljama i zemljama svijeta. Kako bi se banke nosile sa mnogim rizicima, ključna strategija je poboljšanje učinkovitosti putem smanjenja troškova i povećanja prihoda. S porastom transakcija klijenata, pohranjeni podaci u računalnim sustavima postaju sve obimniji. Podaci, za banke, predstavljaju sredstvo organizacije jer sadrže važna saznanja i korisne i zanimljive uzorke. U konzervativnim pristupima, donošenje odluka se često obavlja ručno, a korisnici, također mogu koristiti različite alate za analizu podataka pri donošenju ključnih odluka. Ručna analiza može biti neprecizna jer se velike količine podataka mogu analizirati u ograničenom opsegu (Preethi i Vijayalakshmi, 2017).

Kako se u bankarstvu događaju promjene u načinu obavljanja poslova, primjenom rudarenja podataka i implementacijom elektroničkog bankarstva prikupljanje transakcijskih podataka postaje lakše. Analiza velike količine podataka nadilazi granice ljudskih mogućnosti, a primjena rudarenja podataka može pomoći pri rješavanju poslovnih problema pronalaženjem obrazaca, veza i korelacija skrivenih u poslovnih informacijama koje su pohranjene u bazama podataka. Otkrivanjem znanja iz baza podataka banke mogu uvidjeti, odnosno, sa povećanom preciznošću, predvidjeti kako će kupci reagirati na promjene kamatnih stopa te koji će kupci potencijalno prihvatiti ponude novih proizvoda, a koji će imati veći rizik od neplaćanja kredita. Banke sve više prepoznaju važnost informacija koje posjeduju o svojim klijentima. Održavanje odnosa sa klijentima i učinkovito upravljanje istim je ključno za banke i poslovanje. Da bi to postigle, banke moraju uložiti svoje resurse kako bi što bolje razumjele svoje postojeće i potencijalne klijente (Srivastava, 2021).

Prema Preethi i Vijayalakshimi (2017) rudarenje podataka banke implementiraju u CRM sustav (engl. Customer Relationship Management)., koji predstavlja ključnu poslovnu strategiju banaka i omogućuje stvaranje vrijednosti brenda, kao i razumijevanje te identifikaciju potreba klijenata

pružanjem relevantnih informacija. Ova tehnologija pruža alate za segmentiranje klijenata i pružanje pravovremenih usluga istovremeno omogućujući izgradnju snažnih odnosa sa profitabilnim klijentima. Kroz korištenje rudarenja podataka moguće je identificirati potencijalne klijente i prilagoditi im usluge. Isto tako CRM kroz rudarenje podataka omogućuje bankama da prilagode svoje usluge prema potrebama i ponašanju klijenata radi poboljšanja korisničkog iskustva i lojalnosti.

3.2. Uvod u rudarenje podataka u bankovnom sektoru

Prema Kantardžiću (2011), rudarenje podataka, proizlazi iz potrebe za razumijevanjem kompleksnih skupova podataka obogaćenih informacijama.

Baze ili skladišta podataka mogu se opisati na više načina, a predstavljaju kolekciju podataka strukturiranih prema fizičkom, ali i posredno prema logičkom i konceptualnom modelu informacijskog sustava. Sa druge strane, skladište podataka je baza koja sadrži pažljivo pripremljene podatke koji su nužni za analizu i stvaranje znanja ključnog za donošenje poslovnih odluka. Da bi baza podataka bila kvalitetna mora posjedovati određena obilježja, uključujući jednostavan pristup potrebnim informacijama, dosljednu prezentaciju korporativnih podataka, prilagodljivost i otpornost na promjene te sigurnost za čuvanje korporativnih podataka i funkcionalnost kao temelj za poboljšanje donošenja poslovnih odluka i prihvaćenost od strane korisnika (Ćurko i Španić Kezan, 2016).

Baze podataka se mogu stvarati iz različitih izvora, podijeljenih na unutarnje (npr. plaće zaposlenika, upravljanje kvalitetom, prodaja) i vanjske podatke (npr. podaci o konkurenciji, cijenama na tržištu, demografski podaci). Znanje, kao nematerijalan resurs koji uključuje intuiciju, iskustvo, vještine i učenje, može se povezati s bazama podataka, što dovodi do pojma otkrivanja znanja iz baza podataka. Različite definicije opisuju otkrivanje znanja iz baza podataka, uključujući traženje vrijednih informacija u velikim količinama podataka, netrivijalan postupak pronalaženja novih, valjanih i razumljivih podataka te istraživanje i analizu velikih količina podataka s ciljem otkrivanja smislenih pravilnosti. Za otkrivanje znanja iz baza podataka koriste se različiti alati, uključujući open source alate poput Rapid Minera, KNIME-a, Weke, R-a i SPSS Modelera (Provost i Fawcett, 2013).

3.3. Područja primjene rudarenja podataka u bankovnom sektoru

Kako Pejić Bach (2005) navodi, područja primjene rudarenja podacima u bankovnom sektoru su: rizik, prodaja dodatnih proizvoda postojećim klijentima, zadržavanje postojećih klijenata, segmentacija, životna vrijednost klijenta, odaziv, aktivacija, racionalizacija poslovanja.

Prvo područje je rizik, a model rizika specifičan je za bankovni sektor i osiguravajuća društva, igrajući ključnu ulogu u odobravanju kredita i upravljanju osiguranim rizicima. Bankama je bitna preciznost u procjeni vjerojatnosti vraćanja kredita, što dovodi do korištenja modela rizika. Ovi modeli su važni kako za kredite sa osiguranjem, kao što su jamci, hipoteke i založna prava, tako i za neosigurane kredite poput revolving kreditnih kartica ili prekoračenja na tekućem računu. U slučaju ozljede klijenta, osiguravajuća društva suočavaju se s rizikom da će klijenti iskoristiti osiguranje. Vodeće banke i osiguravajuća društva u Hrvatskoj u svoje poslovanje integriraju modele previđanja rizika (Pejić Bach, 2005).

Model rizika u bankarstvu je ključni instrument za procjenu i upravljanje potencijalnim gubicima ili nepoželjnim događajima koji mogu utjecati na poslovanje ili financijsku stabilnost banaka i institucija. Ima zadaću identificirati moguće rizike za koje se pretpostavlja da negativno utječu na financijsku dobit, a u prvom planu su to rizici kredita, ulaganja i operativnih procesa. Tipovi rizika koji su obuhvaćeni u ovom području su kreditni, operativni i tržišni rizik. Kreditni rizik odnosi se na mogućnost gubitka koji banka može pretrpjeti zbog nemogućnosti klijenta u ispunjavanju svojih obveza prema banci. Operativni rizik obuhvaća rizik od gubitka koji proizlazi iz internih procesa, sustava, ljudskih faktora, tehnologije ili nekih vanjskih događaja i utjecaja. Tržišni rizik se odnosi na gubitak koji proizlazi iz promjene tržišnih uvjeta, a u koje spadaju kamatne stope, tečaj, dionice i slično. Ove vrste rizika su međusobno povezane i često se mogu javiti istovremeno, a banke primjenjuju razne analitičke metode kako bi njima upravljale na pravi način (Mocanu, 2016).

Prodaja dodatnih proizvoda postojećim klijentima jedno je od područja primjene rudarenja podataka u bankarstvu. Cilj je povećanje prihoda, poboljšanje odnosa sa klijentima te ukupne profitabilnosti. Otkrivanje znanja iz baza podataka igra glavnu ulogu u optimizaciji procesa prodaje. Upotrebom rudarenja podataka na početku omogućuje bankama analizu ponašanja klijenata na temelju njihovih transakcija, uplata, isplata i sličnih aktivnosti. Identifikacija uzoraka ponašanja pomaže pri razumijevanju potreba klijenata i predviđanju njihovih budućih aktivnosti.

Nakon analize ponašanja slijedi segmentacija klijenata. Na temelju preferencija, životnih situacija, demografskih čimbenika i drugi podataka koji su relevantni za segmentaciju, otkrivanjem znanja segmentiraju se klijenti, a takav pristup omogućuje bankama prilagodbu ponuda dodatnih proizvoda za specifične potrebe pojedine skupine klijenata (Ostapchenya, 2021).

Rudarenjem podataka pomaže se i pri stvaranju personaliziranih preporuka za dodatne proizvode kao što su krediti, štednje, investicije. Kao što je navedeno, prilagođene preporuke temelje se na analizi povijesnih podataka o transakcijama klijenata i njihovim preferencijama te na usporedbi sa ponašanjem klijenata sa sličnim aktivnostima. Na ovaj način se održavaju dugoročni odnosi sa klijentima, a kako navodi Pejić Bach (2005) klijenti često prelaze konkurenciji zbog pogodnosti koje im se nude.

Sljedeća primjena rudarenja podataka u bankarstvu je na području zadržavanja postojećih klijenta. Prema Pejić Bach (2005) odlazak klijenata konkurenciji problem je mnogih djelatnosti. Zasićenje tržišta vodi tome da mogućnost rasta jednog poduzeća postoji preotimanjem klijenata od konkurencije. Globalna praksa pokazuje da određeni segment klijenata uspješno iskorištava niske kamatne stope kod više kartičnih kompanija. Primjenom rudarenja podataka razvijaju se modeli s ciljem predviđanja vjerojatnosti da će klijent, nakon što se kamatne stope povise na uobičajenu razinu, preći konkurenciji ili će prilagoditi svoje potrošačke navike. Ovakvi modeli omogućuju kartičnim kompanijama unaprijed prepoznati rizike od gubitka klijenata ili smanjenja potrošnje te prilagoditi svoje strategije zadržavanja klijenata i prilagodbe kamatnih stopa kako bi održale konkurentske pozicije na tržištu. Zadržavanje postojećih klijenata je nužno za dugoročni uspjeh banaka.

Segmentacija klijenata u bankarstvu omogućuje bankama dublje razumijevanje ponašanja klijenata te prilagodbu poslovnih strategija prema specifičnim potrebama pojedinih skupina. Banke kreiraju modele ponašanja koji omogućuju identifikaciju uzoraka u transakcijama, preferencijama ili interakcijama klijenata sa bankom. (Rovčanin et al., 2012).

Životna vrijednost klijenta kao sljedeće područje, prema Pejić Bach (2005), predstavlja očekivanu vrijednost zarade od pojedinog klijenta kroz neko razdoblje. Navodi primjer interesa banke za studentima sa kojima može stvoriti dobre odnose. Zarada od istih će u početku biti mala, ali nakon diplomiranja oni će postati profitabilni klijenti zbog novih životnih potreba.

Rudarenje podataka je efikasan alat za ekstrakciju znanja iz postojećih podataka, a u bankarstvu igra ključnu ulogu u analizi i obrazi transakcijskih podataka i profila klijenata. Primjenom tehnika rudarenja podataka, korisnici mogu donositi učinkovite odluke. U bankarstvu se ističu dva osnovna područja primjene, a to su upravljanje odnosima s klijentima i otkrivanje prijevara (Preethi i Vijayalakshmi, 2017).

Prema Preethi i Vijayalakshmi (2017) otkrivanje znanja iz baza podataka se koristi u sustavu upravljanja odnosima s klijentima (eng. Customer Relationship Management, CRM).. Ovaj sustav predstavlja strategiju banaka koja pomaže u stvaranju vrijednosti, identifikaciji i razumijevanju korisnika i njihovih potreba , a kroz osiguravanje relevantnih i pravovremenih informacija koje mogu dodati vrijednost klijentima. CRM sustavi pružaju alate koji omogućuju segmentaciju klijenata i pružanje pravovremenih informacija i usluga djelujući na dinamičke informacije o klijentima. To omogućuje praćenje i izgradnju snažnih i stabilnih odnosa sa profitabilnim klijentima, kao i identifikaciju specifičnih proizvoda i usluga koje klijentima mogu biti od koristi. Otkrivanje znanja iz baza podataka može se koristiti i za stvaranje profila klijenata kako bi se klijenti grupirali u određenu skupinu, a ta skupina se potom obrađivala sukladno njihovim profilima.

Otkrivanje prijevara značajno je područje primjene u bankarstvu. Prioritet mnogih poduzeća je otkrivanje prijevara i aktivnosti istih, a zabrinutost za otkrivanjem svakim danom raste. Primjenom otkrivanja znanja iz baza podataka identificiraju se i prijavljuju brojne prijevare. Kako Preethi i Vijayalakshmi (2017) navode, financijske institucije razvile su dva pristupa za otkrivanje obrazaca prijevara. Prvi pristup uključuje pristupanje skladištu podataka treće strane i upotrebu programa rudarenja podataka za identifikaciju obrazaca prijevare. Nakon toga, banka može usporediti te obrasce sa vlastitom bazom podataka kako bi pronašla znakove internih problema. Drugi pristup predstavlja identifikaciju prijevarnih obrazaca isključivo na internim informacijama same banke. Većina banaka koristi hibridni pristup kombinirajući ova dva pristupa.

3.4. Važnost primjene rudarenja podataka u bankovnom sektoru

Primjena rudarenja podataka u bankovnom sektoru donosi niz značajnih prednosti koje poboljšavaju učinkovitost poslovanja, marketinške strategije te omogućuju brže i informiranije donošenje odluka. Rudarenje podataka je zapravo ključan alat za ostvarivanje konkurentske prednosti i optimizaciju različitih aspekata poslovanja.

Jedna od ključnih prednosti rudarenja podataka u bankovnom sektoru leži u sposobnosti kreiranja modela aktivacije klijenata, što omogućuje predviđanje vjerojatnosti da će novi klijent postati profitabilan. Kroz ovakve modele, banke mogu prilagoditi svoje marketinške strategije kako bi privukle i zadržale klijente sa visokim potencijalom (Srivastava, 2021).

Rudarenje podataka također pruža sredstva za poboljšanje odnosa sa postojećim klijentima, omogućujući personalizaciju ponuda dodatnih proizvoda i usluga. Analizom ponašanja klijenta, banke mogu identificirati segmente klijenata koji su skloni koristiti usluge jednokratno te im pružiti dodatne pogodnosti kako bi ih potaknule na aktiviranje i korištenje novih usluga. Ovaj pristup, osim što povećava profitabilnost poslovanja, također, jača i odnose sa klijentima (Singh i Agray, 2017).

U procesu odobravanja kredita rudarenje podataka ima ključnu ulogu u predviđanju rizika. Modeli rudarenja podataka analiziraju različite faktore, uključujući prethodno ponašanje klijenata, informacije o transakcijama te socioekonomske podatke. Ova analiza pomaže bankama u donošenju brzih i preciznih odluka o odobravanju kredita, smanjujući rizik od neuspjeha u otplati (Voican, 2020).

Otkrivanje znanja iz baza podataka omogućuje i selektivno usmjeravanja marketinških kampanja prema segmentima klijenata sa visokom vjerojatnošću odaziva, čime se smanjuju uzaludni napori i troškovi marketinških inicijativa. Analiza transakcijskih podataka, uz integraciju informacija o plaćanju računa, socioekonomskom statusu i podacima o poslovanju, pruža sveobuhvatan uvid u profil klijenta (Bhambri, 2011).

Primjena rudarenja podataka u bankovnom sektoru pridonosi povećanju konkurentske prednosti, smanjenju troškova, optimizaciji marketinških strategija te jačanju odnosa sa klijentima. Integracija različitih izvora podataka omogućuje bankama donošenje valjanih odluka, čime se

postiže dugoročna profitabilnost i održivost poslovanja u dinamičnom okruženju suvremenog tržišta.

Rudarenje podataka i poslovna inteligencija imaju važnu primjenu u nekoliko ključnih kategorija koje spominju Stancu i Mocanu (2016). Na prvom mjestu nalazi se upravljanje portfeljem, bankarske institucije koriste podatke o klijentima iz svojih velikih baza podataka kako bi upravljale portfeljima i pružile relevantne informacije za donošenje odluka o investicijama i upravljanju rizicima. Banke koriste rudarenje podataka u važnu svrhu upravljanja odnosa sa klijentima, a koristi se za razumijevanje potreba klijenata i poboljšanje njihovog iskustva. Podaci se koriste u svim fazama odnosa sa klijentima, a uključuju i privlačenje novih klijenata, povećanje vrijednosti postojećih klijenata te kontrolu i održavanje tih odnosa.

Rudarenje podataka i poslovna inteligencija igraju važnu ulogu u kategorijama upravljanja rizikom i kreditnog rizika, gdje se fokus stvara na analizi rizika kao ključnoj komponenti u procesu kreditiranja (Stancu i Mocanu, 2016).

3.5. Prikaz algoritama rudarenja podataka

Rudarenje podataka sa sobom nosi mnogobrojne algoritama koji mogu poslužiti pri istraživanjima i obradi podataka. U ovom dijelu opisat će se neke od najčešćih metoda koje se koriste za rudarenje podataka. Među takve ubrajamo stabla odlučivanja, neuronske mreže, algoritam k srednjih vrijednosti te asocijativna pravila.

3.5.1. Stabla odlučivanja

Stabla odlučivanja ističu se kao izuzetno efikasna metoda koja se može primijeniti na različite područja poput klasifikacije, predviđanja, procjene vrijednosti, klasteriranja, opisivanja i vizualizacije podataka. U usporedbi sa drugim metodama, stabla odlučivanja se ističu svojom jednostavnošću i lakoćom razumijevanja te je to zbog toga jedna od popularnijih metoda i opcija. Jednostavnost stabla odlučivanja dolazi do izražaja u činjenici da pravila koja generira mogu biti jasno formulirana na čitljivom jeziku, koji je pristupačan svakome. Ova svojstva omogućuju direktnu primjenu u radu sa bazama podataka, gdje se pravila stabla odlučivanja mogu koristiti za izdvajanje određenih primjera iz baze podataka (Maimon i Rokach, 2014).

Osim toga, stablo odlučivanja pokazuje i svestranost i primjenjivost u istraživačkim radovima, omogućujući prepoznavanje veza između velikog broja ulaznih varijabli i tražene vrijednosti. Kao što je navedeno, stabla odlučivanja koriste se za klasifikaciju podataka, a imaju sposobnost naučiti kompleksne odnose između različitih atributa i kategorizirati podatke u klase ili kategorije. Osim klasifikacije, stabla odlučivanja koriste se i za predviđanja numeričkih vrijednosti (Magee, 1964).

Moguće je stvoriti regresijske modele koji procjenjuju vrijednost na temelju ulaznih varijabli. Ova funkcionalnost ima veliku primjenu u financijskom sektoru, gdje stabla odlučivanja pomažu u predviđanju budućih financijskih performansi temeljem različitih faktora.

Još jedan ključan aspekt stabala odlučivanja je i procjena važnosti različitih atributa u donošenju odluka. Ovaj aspekt je od velike koristi jer omogućuje identifikaciju ključnih faktora ili varijabli koje značajno utječu na određeni ishod, a istovremeno pružajući dublje razumijevanje podataka. Stabla odlučivanja, osim funkcionalnosti donošenja odluka nude i transparentnost samog procesa. Opsežni opisi pri donošenju odluka omogućuju korisnicima bolje razumijevanje logike koja stoji iza rezultata modela. Nadalje, vizualizacija stabala odlučivanja je jednostavna, a to olakšava interpretaciju rezultata i čini ovu metodu pristupačnu svima, uključujući i one koji nisu stručnjaci za područje analize podataka. Sve navedene značajke čine stabla odlučivanja snažnim alatom za otkrivanje znanja iz baza podataka te pridonose razumijevanju, analizi i donošenju odluka koje se temelje na podacima u raznolikim industrijskim i istraživačkim kontekstima (Maimon i Rokach, 2014).

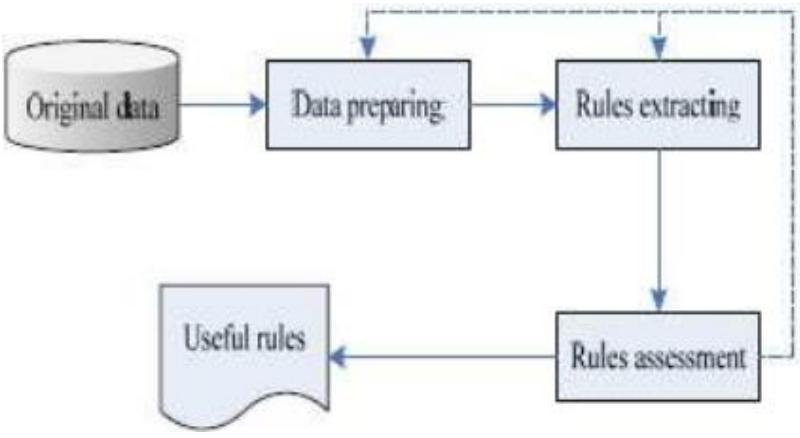
Metoda stabala odlučivanja imaju veliku ulogu u otkrivanju znanja iz baza podataka u bankovnom sektoru jer pružaju intuitivne i lako interpretirane modele koji pomažu donositeljima odluka u raznim aspektima poslovanja. Stabla odlučivanja koriste se u analizi kreditne sposobnosti klijenata. Mogu doprinijeti i pružiti jasne podatke za odlučivanje o odobravanju kredita, a koji su temeljeni na raznim faktorima poput prihoda, prethodnih kredita, transakcija i ostalih financijskih područja i parametara. Metoda stabala odlučivanja posebno je korisna i za praćenje transakcija koje odstupaju od uobičajenih obrazaca te tako, stabla odlučivanja, pridonose identifikaciji nepravilnosti u bankovnim transakcijama. Stabla odlučivanja pomažu bankama i kod segmentacije klijenata prema različitim karakteristikama poput ponašanja, preferencija, kreditnih sposobnosti i ostalih parametara. Osim segmentacije, metoda služi i za predviđanje izgleda zaduživanja klijenta. Analizom različitih varijabli, poput radne stabilnosti ili prijašnjih zaduživanja, ova metoda

pridonosi i smanjenju rizika za banku. Stabla odlučivanja, kroz spomenute primjene, unaprjeđuju analizu podataka u bankovnom sektoru te čine procese donošenja odluka učinkovitima, transparentnima i prilagodljivima okruženju financijskog sektora (Apté i Weiss, 1997).

3.5.2. Neuronske mreže

Neuronske mreže su sljedeća metoda koja se koristi u otkrivanju znanja iz baza podataka. Često se koriste za analizu kompleksnih skupova podataka i otkrivanje skrivenih uzoraka. Također, koriste se u analizi rizika i prognoziranju, a proces rudarenja podataka započinje učenjem mreže na skupu podataka s poznatim vrijednostima koje se žele prognozirati. Metoda neuronskih mreža je inspirirana strukturom ljudskog mozga te simulira način na koji neuroni međusobno komuniciraju kako bi odradili informacije. Primjena ove metode značajno raste, a korištenje iste je započelo kao alat za prepoznavanje uzoraka, poput rukopisa ili geometrijskih likova, temeljeći se na principu podudarnosti. Neuronske mreže se sastoje od slojeva neurona, a najčešće se dijele na tri vrste slojeva, a to su ulazni sloj, skriveni slojevi te izlazni sloj. Ulazni sloj prima početne podatke, skriveni slojevi ih obrađuju, a izlazni sloj daje konačni rezultat. Veze između neurona imaju različite težine koje se prilagođavaju tijekom procesa učenja, a ne postoji univerzalni model neuronskih mreža koji bi odgovarao svim područjima. Za svaku domenu, odnosno dio koji se proučava, potrebno je razviti specifičan model sa određenim prednostima i mogućnostima (Banov et al., 2022).

Slika 2 Proces rudarenja podataka baziran na neuronskim mrežama



Izvor: Gaur, P. (2012), *Neural networks in data mining. International Journal of Electronics and Computer Science Engineering*

Na slici 2 prikazan je proces rudarenja podataka koji je baziran na metodi neuronskih mreža, a obuhvaća pripremu podataka, ekstrakciju pravila te procjenu istih i time se dolazi do korisnih podataka.

Temeljni princip učenja neuronskih mreža leži u vezama između eksperimentalnih uzoraka. Razlikujemo autoasocijativno učenje, gdje se uzorci pridružuju sami sebi te heteroasocijativno učenje, gdje se različiti tipovi uzoraka pridružuju jedni drugima. Kako bi interpretacija rezultata bila što uspješnija, neuronske mreže se često kombiniraju sa drugim metodama. Iste se najčešće primjenjuju za segmentaciju, predikciju, klasifikaciju, regresiju i analizu skupa podataka. Prednosti neuronskih mreža kao metode za otkrivanje znanja iz baza podataka leži u sposobnosti obrade kompleksnih odnosa u podacima, a izazovi uključuju potrebu za velikim skupovima podataka i računalnim resursima te složenost interpretacije rezultata. Zbog složenosti interpretacije, kao što je već navedeno, neuronske mreže kombiniraju se sa drugim metodama za uspješno provođenje interpretacije (Cios et al., 2012).

Neuronske mreže igraju ključnu ulogu u bankovnom sektoru kroz razne primjene koje doprinose optimizaciji poslovnih procesa, poboljšanju sigurnosti, kao i pružanju boljeg korisničkog iskustva. Ova metoda koristi se za detekciju prijevara u financijskim transakcijama. Analizom uzoraka potrošačkih navika i transakcija korisnika, mogu se identificirati neobične ili sumnjive aktivnosti pridonoseći smanjenju rizika prijevare. Osim detekcije prijevara, metoda neuronskih mreža koristi se i da procjenu kreditne sposobnosti klijenata, kao i personaliziranih ponuda i preporuka. Neuronske mreže mogu pratiti tržišne promjene i analizirati financijske pokazatelje kako bi identificirale anomalije koje ukazuju na potencijalne rizike. Neuronskim mrežama mogu se predvidjeti i ekonomski trendovi, analizom velikog skupa podataka (Zekić et al., 2009).

3.5.3. Algoritam K srednjih vrijednosti

Za razliku od prethodno opisanih metoda, algoritam K srednjih vrijednosti je metoda klasteriranja, a ne metoda klasifikacije. Algoritam K srednjih vrijednosti, eng. K-means, ima za cilj podijeliti skup podataka u K klastera, a pri čemu svaki klaster ima slične primjene, a atributi iz različitih klastera su međusobno različiti. Ova metoda, kao što je navedeno pripada metodama klasteriranja. Metode klasteriranja nalaze se u grupi neusmjerenih metoda koje imaju cilj otkrivanje globalne strukture podataka. (Mirošević, 2016).

Metode klasteriranja dijele primjere u skup klastera, odnosno podskupova, a isti moraju zadovoljavati dva osnovna kriterija. Prvi kriterij je taj da svaki klaster mora predstavljati homogeni skup, tj. atributi koji pripadaju istom klasteru moraju biti međusobno slični. Drugi kriterij koji se mora zadovoljiti je taj da se svaki klaster mora razlikovati od ostalih klastera. Odnosno, atributi koji se nalaze u određenom klasteru moraju se razlikovati od atributa iz ostalih klastera (Kodinariya i Makwana, 2013).

$$W(S, C) = \sum_{k=1}^{K} \sum_{i \in S_k} \|y_i - c_k\|^2 \qquad (1)$$

U formuli (1) S predstavlja podjelu K-klastera skupa entiteta koji su predstavljeni vektorima y_i (i € I), a sastoje se od nepraznih i nepreklapajućih S_k klastera, svakih sa težištima c_k (k=1, 2, …, K).

Algoritam se sastoji od par koraka, od kojih je prvi postavljanje k točaka u prostor predstavljen objektima, koji se koriste pri klasteriranju. Ove točke predstavljaju početne točke težišnih grupa. Svaki objekt se dodjeljuje grupi sa najbližim težištem. Slijedi ponovni izračun položaja težišta, a potom se ovi koraci ponavljaju dok se ne dođe do stadija gdje se težišta više ne pomjeraju (Kodinariya i Makwana, 2013).

Metoda K-means ima dosta važnu ulogu u bankovnom sektoru. Analiza podataka putem ove metode omogućuje bankama bolje razumijevanje klijenata, identifikaciji obrazaca ponašanja te prilagodbi svojih usluga i marketinških aktivnosti. Određivanje broja K klastera omogućuje bankama razdvajanje klijenata u grupe sličnih profila. Klasteri se mogu formirati na temelju različitih kriterija, kao što su, transakcije, demografski podaci, preferencije i slično. Identifikacijom sličnosti u obrascima potrošnje, štednje i investiranja pomoću ove metode omogućuje bankama prilagodbu svojih proizvoda i usluga prema potrebama svakog klastera. K-means metoda može pomoći i pri prevenciji prijevara. Analizom transakcijskih uzoraka za otkrivanje neobičnih obrazaca ponašanja koji mogu ukazivati na prijevaru stvaraju se klasteri prema vrstama sumnjivih transakcija te tako banke mogu doći do brzog rezultata i brze i učinkovite reakcije. Algoritam K srednjih vrijednosti je potrebno redovito održavati i ažurirati kako bi se održavale promjene u ponašanju klijenata i dinamici tržišta. Ova metoda služi kao vrijedan alat bankama za optimizaciju analize podataka i segmentaciju klijenata. Integracija K-means metode omogućuje bolje

razumijevanje potreba klijenata, povećanje učinkovitosti marketinških strategija te doprinosi očuvanju sigurnosti i integriteta bankarskih podataka. Održavanje modela ključno je za prilagodbu brzim promjenama u bankovnom okruženju (Kodinariya i Makwana, 2013).

3.5.4. Asocijativna pravila

Asocijativna pravila su još jedna od metoda za rudarenje podataka, a koriste se za otkrivanje skrivenih veza i zavisnosti među podacima. Ova metoda je poznata i kao analiza potrošačke košarice. Metoda asocijativnih pravila, u kontekstu bankarstva, ključan je alat za otkrivanje skrivenih veza između transakcija, ponašanja klijenta i ostalih relevantnih podataka koji mogu poslužiti za obradu podataka. Asocijativna pravila omogućuju bankama bolje razumijevanje potreba klijenata, prilagodbu ponuda i povećanje sigurnosti financijskih transakcija. Korištenjem asocijativnih pravila u analizi podataka o transakcijama mogu se otkriti povezanosti između različitih financijskih aktivnosti klijenata. Identifikacijom uzoraka potrošnje i ponašanja klijenata, asocijativnim pravilima moguće je prilagoditi personaliziranu ponudu bankarskih proizvoda. Prednosti metode asocijativnih pravila je to što je jednostavna za korištenje i daje jasne rezultate, namijenjena je problemima koji nisu klasifikacijskog tipa, a može se koristiti i kod primjera sa varijabilnim brojem atributa (Miyan, 2017).

$$Rule: X \rightarrow Y \begin{cases} \nearrow \; Support = \dfrac{frq(X,Y)}{N} & (2) \\ \rightarrow \; Confidence = \dfrac{frq(X,Y)}{frq(X)} & (3) \\ \searrow \; Lift = \dfrac{Support}{Supp\,(X) \times Supp(Y)} & (4) \end{cases}$$

Formule (2), (3), (4) prikazuju, zapravo, način na koji asocijativna pravila funkcioniraju. Ista pomažu otkrivanju odnosa između nepovezanih podataka u bazama podataka. Kriteriji određivanja asocijativnih pravila sa slike 4 su podrška, povjerenje i ohrabrenje. Ovi kriteriji identificiraju odnose i pravila koja su generirana analizom podataka prema ako/tada obrascima (Kumbhare i Chobe, 2014).

3.5.5. J48 algoritam

Algoritam J48 se koristi za klasifikaciju i predikciju u području strojnog učenja. Temelji se na popularnom klasifikacijskom algoritmu C4.5, koji generira stabla odlučivanja koristeći teoriju informacija. Zapravo je proširenje ranijeg ID3 algoritma, u Weki poznatog kao J48. Implementacija C4.5 algoritma u J48 ima brojne dodatne značajke, uključujući i nedostajuće vrijednosti, podrezivanje stabala odlučivanja, kao i raspon vrijednosti kontinuiranih atributa i slično (Ibrahim et. al., 2016)

$$GainRatio(S, A) = \frac{Gain\ (S, A)}{SplitInformation\ (S, A)} \quad (5)$$

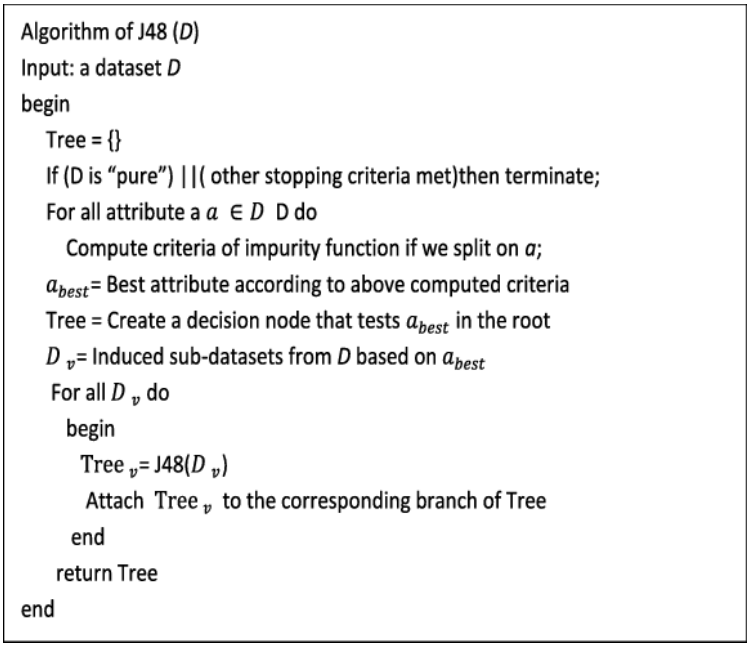
Jednadžba (5) koja se koristi za izračunavanje omjera, odnosno koeficijenta dobivanja informacija i podjele informacija, a čija je jednadžba (6) prikazana ispod.

$$SplitInformation\ (S, A) = -\sum_{j=1}^{|k|} \frac{S_i}{S} \log_2 \frac{S_i}{S} \quad (6)$$

S_i do S_k su podskupovi instanci, a prikupljeni su podjelom informacija po „k“ vrijednostima atributa.

U softverskom alatu Weka, koji će se, u ovom radu koristiti za obradu baze podataka, J48 predstavlja implementaciju otvorenog tipa C4.5. algoritma u Javi. J48 omogućuje klasifikaciju putem stabala odlučivanja ili pravila generiranih iz njih. Algoritam J48 gradi stabla odlučivanja na temelju skupa podataka na isti način kao i spomenuti ID3 algoritam, a koristeći koncept informacijske entropije. Trening set podaci su skup S={s1,s2, …} već klasificiranih uzoraka. Svaki takav uzorak se sastoji o vektora (x1, i, x2, i, … xp, i), a tu xj predstavlja vrijednost atributa (Ibrahim et. al., 2016).

Slika 3 Pseudo kod algoritma J48



Izvor: Ibrahim et. al. (2016)

Na slici 3 prikazan je princip djelovanja J48 klasifikatora. Na početku se stvara stablo odlučivanja koje se temelji na vrijednosti atributa u skupu trening podataka, odnosno podataka za obuku. Nakon toga je potrebno izabrati najbolji atribut koji može razlikovati različite instance, a to je obično atribut sa najvećim priljevom informacija. Potom se, za svaku vrijednost odabranog atributa, stvara tzv. dječji korijen. Zatim se instance raspoređuju po čvorovima i postupak se tako ponavlja koristeći trening instance, koje su povezane sa svakim dječjim čvorom, da bi se odabrao najbolji atribut za testiranje. Grana se završava ciljanom vrijednošću koja je dobivena na kraju, a čini ju kombinacija atributa.

Ključni koraci uključuju odabir najboljeg atributa za podjelu skupa podataka i primjenu algoritma na podskupovima podataka koji su stvoreni podjelom prema odabranom atributu. Najvažniji dio pseudokoda je odluka o tome koji atribut odabrati za podjelu skupa podataka. Obično se to radi koristeći neku mjeru informativnosti, poput entropije ili Ginijevog koeficijenta. Cilj je odabrati atribut koji će rezultirati što čišćim podjelama skupa podataka. Osim toga, proces se rekurzivno poziva na podskupovima podataka koji su generirani podjelom prema odabranom atributu, sve dok se ne ispuni neki od uvjeta zaustavljanja (Ibrahim et.al, 2016).

4. RUDARENJE PODATAKA U SVRHU OTKRIVANJA SUMNJIVIH BANKOVNIH TRANSAKCIJA

Korištenje kreditnih kartica pri obavljanju transakcija, u današnjem svijetu, sve je učestalije. Upravo ta učestalost korištenja dovodi i do sumnjivih transakcija, odnosno do prijevara, koje u financijskom sektoru predstavljaju značajnu brigu.

Ručna analiza transakcija prijevare nije izvediva zbog ogromne količine podataka i složenosti istih. Iz milijuna transakcija provedenih kreditnim karticama, koje se događaju, doslovno, u trenutku, nije moguće ručno identificirati prijevare, stoga postoji potreba za automatiziranim sustavima detekcije prijevara. Rudarenje podataka pruža automatiziran i brz način otkrivanja prijevara među milijunima transakcija bez potrebe ljudske intervencije.

U ovom poglavlju obradit će se podaci pronađeni na web stranicama koje sadržavaju baze podataka o određenim područjima. Na početku će se opisati metodologija istraživanja podataka, izvori istih te metode prikupljanja i obrade. Nakon metodologije istraživanja opisat će se pronađeni podaci, a potom će se prikazati rezultati istraživanja. Rezultati istraživanja pomoći će i kod preporuka za bankovno poslovanje koje će ujedno biti i zadnji dio ovoga rada prije zaključka istoga.

4.1. Metodologija istraživanja

Metodologija istraživanja temelji se na analizi podataka transakcija kreditnih kartica. Podaci su preuzeti sa dostupne online baze podataka. Isti su već bili spremni u Excel-u u .csv formatu. Radi lakšeg razumijevanja podaci su prvo pregledani u Excelu, a nakon toga, s obzirom da su već bili u traženom formatu, učitani u softverski alat Weka.

Weka je softverski alat za strojno učenje, a napisan je u programskom jeziku Java. Ovaj alat je razvijen na Sveučilištu Waikato na Novom Zelandu. Omogućuje identifikaciju skrivenih informacija iz baza podataka, sadrži jednostavne opcije i vizualno sučelje. Koristi alate za vizualizaciju i algoritme za rješavanje problema u rudarenju podataka (Kulkarni i Kulkarni, 2016).

Prema Ilic i suradnicima (2016) Weka je prenosivi softver napisan u programskom jeziku Java, što ga čini kompatibilnim sa gotovo svim modernim računalnim platformama. Weka omogućuje izvršavanje standardnih zadataka rudarenja podataka, a kao što su obrada podataka, klasteriranje,

klasifikacija, asocijacijska vizualizacija i odabir značajki. Korisnici pristupaju Weka grafičkom okruženju putem Weka GUI izbornika, koji nudi četiri glavne funkcionalnosti: 'Explorer', 'Experimenter', 'Knowledge Flow' i 'Simple CLI', a u verziji koja će se koristiti za obradu podataka, u ovom radu, postoji i funkcionalnost 'Workbench'.

Tehnike i algoritmi rudarenja podataka smješteni su u 'Explorer' sučelju, koje omogućuje učitavanje podataka, pregled atributa i primjenu različitih tehnika na podacima. Kroz panel za klasifikatore, korisnici mogu konfigurirati i izvršiti klasifikaciju nad podacima te vizualizirati greške klasifikacije. Panel za klasteriranje omogućuje, slično kao i prethodni, izvršavanje klastera na podacima, uz mogućnost vizualizacije istih. Dodatni paneli pružaju mogućnost za asocijaciju podataka i vizualizaciju, uključujući rudarenje asocijacija i odabir značajki. Vizualizacijski panel omogućuje prikaz podataka z jednoj i dvije dimenzije, a koristi boje za prikaz diskretnih atributa i veličinu točaka za prikaz kontinuiranih atributa (Ilic et al., 2016).

Za analizu podataka odabrana je metoda stabala odlučivanja. Kao što je već ranije spomenuto, stabla odlučivanja su jedna od najjednostavnijih i najpopularnijih oblika strojnog učenja, koji koristi induktivno učenje za generiranje modela.

Zbog jednostavnosti i veće točnosti analize, neki od atributa iz pronađene baze podataka su uklonjeni.

4.2. Opis podataka

U tablici 1 prikazani su atributi, njihovi opisi, format i modaliteti te najmanja i najveća vrijednost za numeričke atribute. Baza podataka sadrži 15 atributa, od kojih su 8 numerički i 7 nominalni. U tablici 1 su prikazani modaliteti samo nominalnih atributa jer numerički atributi imaju velik broj modaliteta. Baza podataka sadrži 19963 instance, što možemo vidjeti na slici 4.

Svaki atribut redom ima svoj naziv i opis. Nakon naziv i opisa u tablici 1 su prikazani formati atributa, numerički ili nominalni. Osim ovih stavki i modaliteta nominalnih atributa, prikazana je najmanja i najveća vrijednost numeričkih atributa, kao posljednja stavka tablice 1.

Tablica 1 Prikaz atributa sa formatima, modalitetima te najmanjom i najvećom vrijednosti

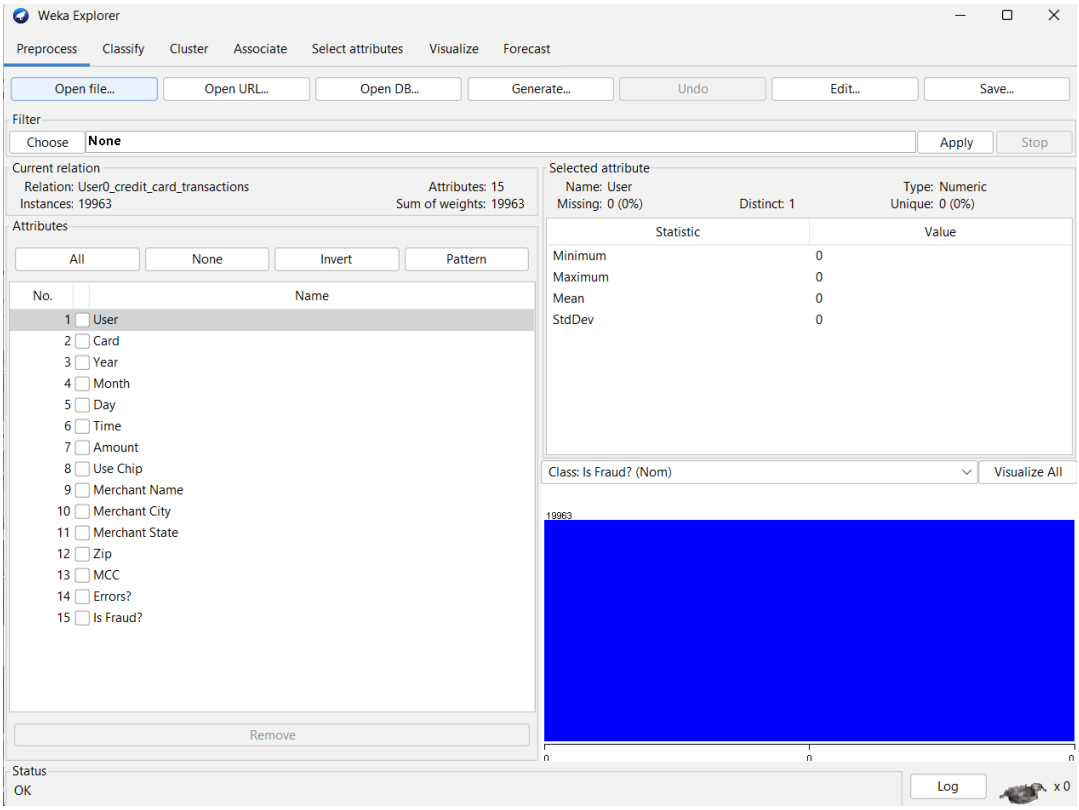
Naziv atributa	Opis atributa	Format atributa	Modaliteti atributa (nominalnih)	Najmanja i najveća vrijednost
User	**Korisnik**	**Numerički**		**0, 0**
Card	**Kartica**	**Numerički**		**0, 4**
Year	**Godina transakcije**	**Numerički**		**2002, 2020**
Month	**Mjesec transakcije**	**Numerički**		**1, 12**
Day	**Dan transakcije**	**Numerički**		**1, 31**
Time	**Vrijeme transakcije**	**Nominalni**	**Popis vremena nastanka transakcija**	
Amount	**Iznos transakcije**	**Nominalni**	**Popis iznosa transakcija**	
Use Chip	**Način transakcije**	**Nominalni**	**Swipe Trancastion, Online Transaction, Chip Transaction**	
Merchant Name	**Kod trgovca**	**Nominalni**	**Popis kodova trgovaca**	
Merchant City	**Grad trgovca**	**Nominalni**	**Popis od 295 gradova**	
Merchant State	**Država trgovca**	**Nominalni**	**CA, NE, IL, MO, Switzerland, IA, TX, Estonia, NJ, NV, Japan, AZ, UT, FL, MI, Mexico, WY, OH, Dominican Republic, NM, China, SC, AK, PA, VA, Portugal, HI, CT, MA, MN, CO, Italy, GA, Philippines, Jamaica, AR, Canada, OR, WI**	
Zip	**Zip kod**	**Nominalni**	**Kodovi za 99504 naselja**	
MCC	**Kodovi kategorije trgovca (engl. Merchant Category Codes)**	**Nominalni**	**Kodovi za 9042 kategorija trgovaca**	

Errors?	**Pogreška u transakciji**	**Nominalni**	**Technical Glitch, Insufficient Balance, Bad PIN, Bad PIN;Insufficient Ballance, Bad Expiration, Bad PIN;Technical Glitch, Bad card Number, Bad CVV**	
Is Fraud?	**Je li transakcija prijevara?**	**Nominalni**	**No, Yes**	

Izvor: rad autorice

Na slici 4 prikazani su podaci učitani u Preprocess panel softverskog alata Weka.

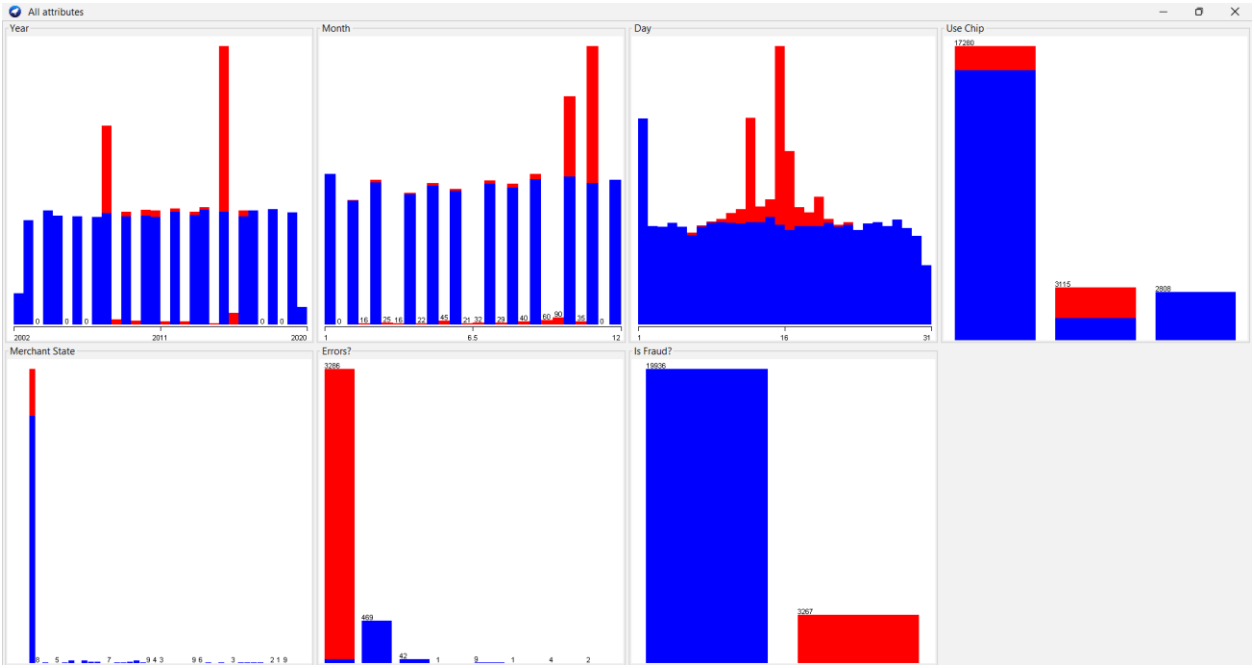
Slika 4 Prikaz podataka učitanih u Preprocess panel



Izvor: rad autorice, Weka

Kao što je navedeno, prvi korak ovoga procesa je priprema podataka u .csv formatu, pregled u Excelu te učitavanje istih u Preprocess panel u Weka-i.

Slika 5 Prikaz odnosa pojedinih atributa sa klasnim atributom



Izvor: rad autorice, Weka

Na slici 5 prikazan je odnos pojedinog atributa sa klasnim atributom 'Is Fraud'. Zbog jednostavnije i pouzdanije analize neki atributi su uklonjeni.

Prvotno učitani podaci u Preprocess panel softverskog alata Weka su zbog problema klasne neravnoteže podvrgnuti filtriranju. Jedan od primijenjenih filtera je promjena numeričkih atributa u nominalne, a potom korištenje filtera SMOTE u svrhu rješavanja klasne neravnoteže i približavanja rezultata istraživanja stvarnom svijetu.

4.2.1. Problem klasne neravnoteže

Neuravnoteženim skupom podataka smatraju se podaci kod kojih jedna od dvije klase ima vrlo malo uzoraka u usporedbi sa uzorcima iz druge klase. Što bi značilo da prva klasa ima više uzoraka od druge klase, a ona klasa koja je u manjku je, zapravo, ona koja nas zanima. Algoritam strojnog učenja, u slučaju uravnoteženih podataka, uvijek daje povjerljive i dobre rezultate. Ali, kada su u pitanju neuravnoteženi podaci, tada dolazi do problema pri obradi podataka. Predikcija je, u ovom slučaju, naklonjena klasi sa većim brojem uzoraka (Verma, 2019).

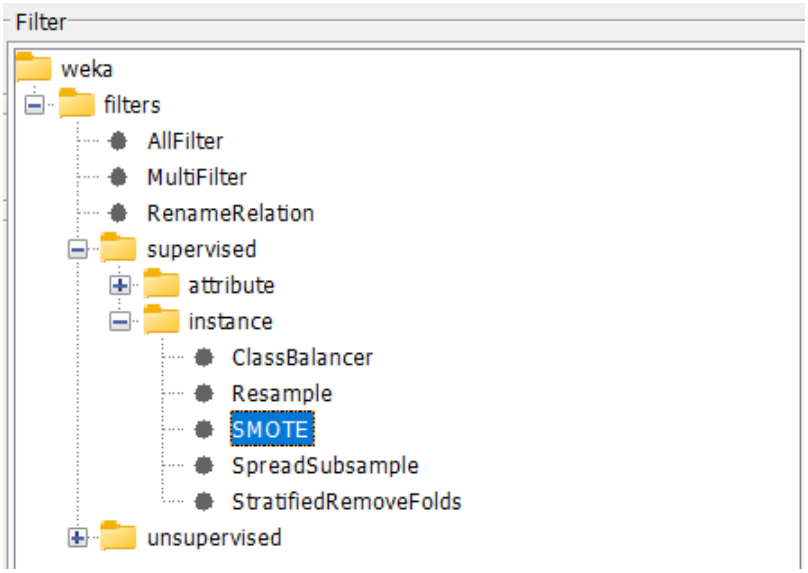
Da bi se problem klasne neravnoteže riješio, primjenjuju se brojni algoritmi i filteri kako bi se dobili jasniji rezultati istraživanja. Jedno od rješenja je korištenje algoritma SMOTE (engl. Synthetic Minority Oversampling Technique). SMOTE je metoda stvaranja sintetičkih primjera manjinske klase. Skup podataka mijenja se dodavanjem sintetički generiranih primjeraka manjinske klase, a što rezultira uravnoteženijom distribucijom klasa. Dodani primjeri su nazvani sintetičkim iz razloga što su stvoreni iz postojećih primjera manjinske klase. Kako bi se ti sintetički primjeri stvorili, SMOTE prvo nasumično odabire primjer, npr. 'X' manjinske klase i dalje nastavlja sa pronalaskom najbližih 'k' susjeda u toj klasi. 'X' se tada povezuje sa jednim od najbližih susjeda 'k', npr. 'Y' te se povezivanjem 'X' i 'Y' stvara novi sintetički primjerak 'Z', a predstavlja kombinaciju, spomenutih 'X' i 'Y' primjeraka (Verma, 2019).

SMOTE je zapravo tehnika balansiranja klasa, a koja se koristi u strojnom učenju, posebno za nebalansirane skupove podataka. Generira sintetičke primjere manjinske klase kako bi se povećala njezina zastupljenost u skupu podataka. U Weki je SMOTE implementiran kao filter koji se može primijeniti na skup podataka prije obrade samog modela, odnosno primjene odabranog algoritma. Omogućuje podešavanje različitih parametara, uključujući sintetički broj primjera koji će biti generirani za manjinsku klasu te broj najbližih susjeda koji se koriste za generiranje novih primjera. Prije primjene filtera, isti se može konfigurirati određivanjem spomenutih parametara. Primjena ovog filtera može poboljšati performanse modela klasifikacije nad neuravnoteženim skupovima podataka jer pomaže modelima lakšu obradu manjinske klase. Filter SMOTE može dovesti do povećanja veličine skupa podataka, budući da generira dodatne sintetičke primjere. Osim toga, rezultati mogu varirati ovisno o odabranim parametrima (Verma, 2019).

4.3. Rezultati istraživanja primjenom tehnike SMOTE

Kako bi istraživanje nad odabranim podacima rezultiralo pouzdanijim uvidima, isti će se, na početku, filtrirati algoritmom SMOTE. Na slici 6 prikazan je odabir spomenutog filtera, a isti se nalazi pod primjerima koji su svrstani u one koji se nadziru pri obradi.

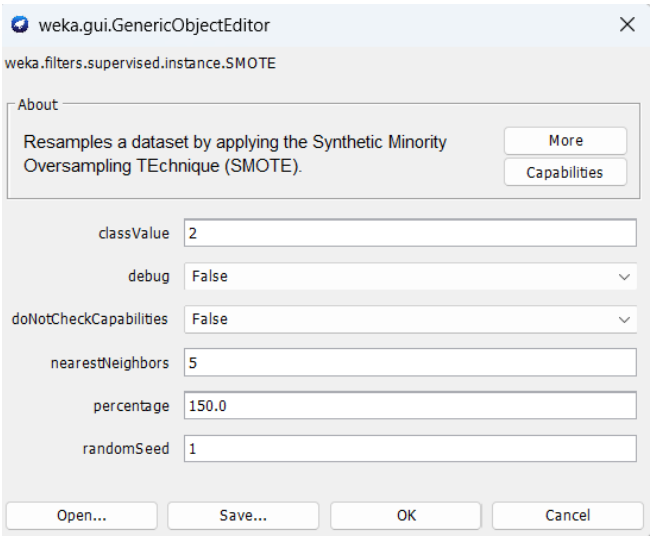
Slika 6 Prikaz odabranog filtera SMOTE



Izvor: rad autorice, weka

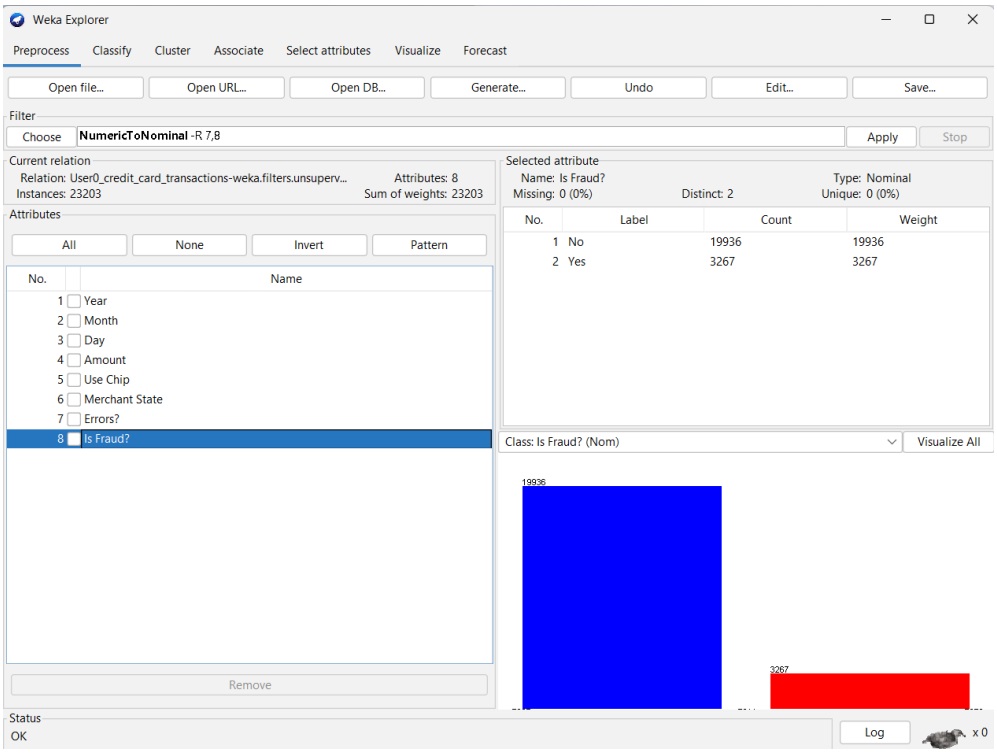
Na slici 7 prikazane su postavke filtera SMOTE nad klasnim atributom, a postotak je povećan na 150% kako bi se povećao i broj sumnjivih transakcija, odnosno instanci modaliteta 'Yes'.

Slika 7 Prikaz postavki filtera SMOTE, rad autorice



Kod prvotno učitanih podataka broj sumnjivih transakcija, odnosno instanci modaliteta 'Yes' iznosio je svega 27, a ovim povećanjem i primjenom filtera SMOTE, isti iznosi 3267, što se može vidjeti na slici 8.

Slika 8 Prikaz odnosa modaliteta klasnog atributa



Izvor: rad autorice, Weka

Na slici 9 prikazani su korišteni atributi. Zbog lakše vizualizacije stabla odlučivanja, neki atributi su uklonjeni te se, u ovom slučaju, obrađuje njih 7. Za obradu podataka, nakon filtriranja, korišten je, već spomenuti, algoritam J48 sa određenim postavkama parametara (batchsize = 1000, confidenceFactor = 0,9, minNumObj = 700).

Slika 9 Prikaz korištenih atributa i postavki algoritma

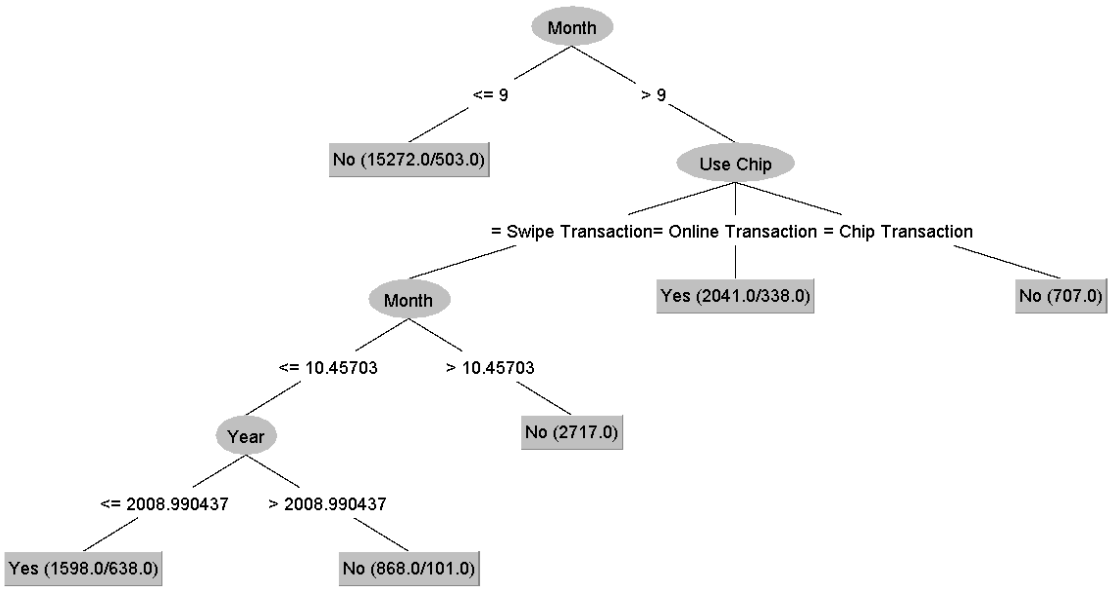
```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.9 -M 700
Relation:     User0_credit_card_transactions-weka.filters.unsupervised.attribute.Remove-R1-weka.filte
Instances:    23203
Attributes:   7
              Year
              Month
              Day
              Use Chip
              Merchant State
              Errors?
              Is Fraud?
Test mode:    10-fold cross-validation
```

Izvor: rad autorice, Weka

Stablo odlučivanja koje se dobije obradom ovog seta podataka, a pod filterom SMOTE prikazano je na slici 10. U ovom slučaju, broj instanci je porastao na 23203. Broj listova koje ovo stablo odlučivanja sadrži je 6, a veličina samog stabla je ukupno 10.

Slika 10 Vizualni prikaz stabla odlučivanja



Izvor: rad autorice, Weka

Slika 11 Prikaz stabla odlučivanja

```
J48 pruned tree
------------------

Month <= 9: No (15272.0/503.0)
Month > 9
|   Use Chip = Swipe Transaction
|   |   Month <= 10.45703
|   |   |   Year <= 2008.990437: Yes (1598.0/638.0)
|   |   |   Year > 2008.990437: No (868.0/101.0)
|   |   Month > 10.45703: No (2717.0)
|   Use Chip = Online Transaction: Yes (2041.0/338.0)
|   Use Chip = Chip Transaction: No (707.0)

Number of Leaves  :     6

Size of the tree :      10
```

Izvor: rad autorice, Weka

Na slikama 10 i 11 može se vidjeti prikaz stabla odlučivanja. Atribut 'Month', kao korijen stabla, poprima modalitete <=9 i >9. Prema vizualnom prikazu, transakcije koje su se dogodile do 9. mjeseca u godini klasificirane su kao pouzdane, a u kasnijim mjesecima, nastale su nepouzdane transakcije. Najviše nepouzdanih ili sumnjivih transakcija dogodilo se online putem. Slijede transakcije provlačenjem kartice, a kojih se najviše dogodilo od 2002. do 2008. godine, do 10. mjeseca.

Slika 12 Prikaz konfuzijske matrice

```
=== Confusion Matrix ===

     a     b   <-- classified as
 18999   937 |     a = No
   651  2616 |     b = Yes
```

Izvor: rad autorice, Weka

Konfuzijska matrica (slika 12) prikazuje da je 18999 točno klasificirana instanca na poziciji 'aa' te 2616 na poziciji 'bb'. Elementi na poziciji 'aa' su one transakcije koje su pozitivne, odnosno, klasificirane kao ispravne, dok su na poziciji 'bb' transakcije koje su klasificirane kao one koje vode prijevari.

Elementi na pozicijama 'ab' i 'ba' su netočno klasificirani pa tako imamo 651 transakciju koja je klasificirane kao ispravne, a zapravo su prijevara te 937 transakcija koje su klasificirane kao negativne, a zapravo su ispravne.

Ovim modelom je 21615 instanci točno klasificirano, a 1588 njih netočno.

4.4. Primjena filtera Resample za rješavanje problema klasne neravnoteže

Set podataka koji će se koristiti za prikaz primjene filtera 'Resample' pri rješavanju problema klasne neravnoteže preuzet je sa web stranice Kaggle, a sadrži podatke o transakcijama izvršenim putem kreditnih kartica, u rujnu 2013. godine, na području Europe. Ovaj skup podataka predstavlja transakcije koje su se dogodile tijekom dva dana, a pri čemu je, od ukupno 284807 transakcija, njih 492 klasificirano kao prijevara. Iz ove informacije odmah se da zaključiti da je riječ o neuravnoteženom skupu podataka te, da bi rezultati istraživanja bili korisniji, potrebno je

primijeniti određene filtere ta rješavanje navedenog problema. Spomenuti skup sadrži samo numeričke ulazne varijable V1, V2, …, V28, a koji predstavljaju podatke poput imena korisnika, mjesta provedbe transakcije, načina provođenja iste i slično. Zbog pitanja povjerljivosti, originalne značajke nisu prikazane.

Nad klasnim atributom, na početku, je primijenjen filter 'NumericToNominal', a koji služi da se određenom atributu promijeni tip, točnije, da se iz numeričkog pretvori u nominalni atribut.

Prvotno učitani set podataka (slika 13) sadrži 284315 pozitivno klasificiranih transakcija te 492 negativno klasificirane transakcije.

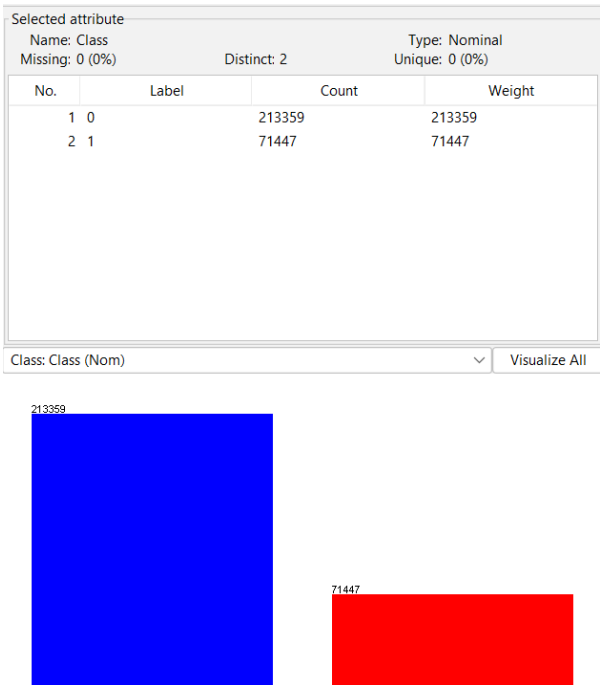
Slika 13 Prikaz modaliteta klasnog atributa

Izvor: rad autorice, Weka

Kao što je spomenuto, set podataka je neuravnotežen pa će se nad istim primijeniti filter 'Resample', kako bi se problem neuravnoteženosti donekle riješio.

Na slici 14 prikazani su modaliteti klasnog atributa, a sada broj pozitivno klasificiranih transakcija iznosi 213359, a broj negativnih transakcija je porastao na 71447.

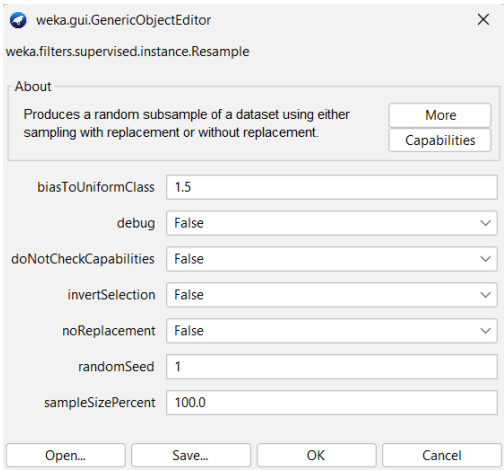
Slika 14 Prikaz modaliteta klasnog atributa nakon primjene filtera 'Resample'



Izvor: rad autorice, Weka

Filter 'Resample' u softverskom alatu Weka služi sa nasumično preuzimanje uzoraka manjinske klase, a proizvodi slučajni poduzorak skupa podataka koristeći ili uzrokovanje sa zamjenom ili bez zamjene, ovisno o postavkama parametara. Ovim filterom postiže se uravnoteženija distribucija klasa i izbjegava se neuravnoteženost iste pri daljnjoj obradi. Zamjensko uzrokovanje koristi se kada se uzorci skupa podataka uzimaju sa mogućnošću ponavljanja, a to znači da jedna instanca može biti odabrana više puta u uzorku. Uzrokovanje bez zamjene osigurava da svaka instanca bude odabrana samo u jednom uzorku, što može biti korisno u nekim situacijama. Parametar 'biasToUniformClass' (slika 15) omogućuje korisniku kontrolu nad distribucijom klasa u generiranom skupu podataka. Koristeći filter 'Resample', mogu se učinkovito rješavati problemi sa neuravnoteženim klasama ili prilagoditi veličina skupa podataka prema potrebama za analizu podataka ili treniranje modela strojnog učenja (Verma, 2019).

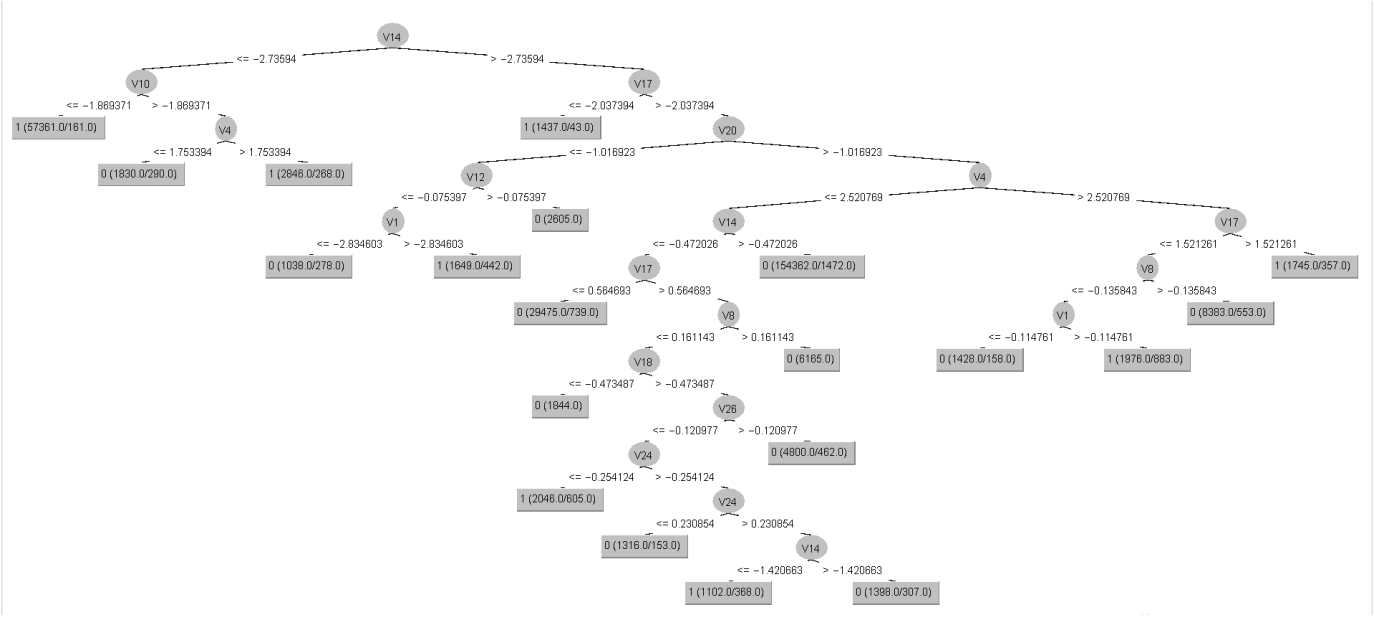
Slika 15 Prikaz postavki parametara filtera 'Resample'



Izvor: rad autorice, Weka

Nakon korištenja filtera 'Resample', slijedi, kao i u prethodnom slučaju, obrada podataka algoritmom J48. U ovom slučaju, za analizu je korišten ukupno 31 atribut, odnosno svi atributi koji su i prvotno učitani u 'Preprocess panel'. Stablo odlučivanja (slika 16) koje se dobije ovom analizom sadrži 20 lista i ukupne je veličine 39.

Slika 16 Vizualni prikaz stabla odlučivanja



Izvor: rad autorice, Weka

Konfuzijska matrica (slika 17) nastala ovim modelom prikazuje da je 210360 transakcija klasificirano kao pozitivno, a 66666 njih je klasificirano kao negativne, odnosno prijevarne transakcije.

Slika 17 Prikaz konfuzijske matrice

```
=== Confusion Matrix ===

      a      b   <-- classified as
 210360   2999 |      a = 0
   4781  66666 |      b = 1
```

Izvor: rad autorice, Weka

Ukupan broj točno klasificiranih transakcija je 277026, a netočno klasificiranih 7780.

Slika 18 Prikaz stabla odlučivanja

```
V14 <= ?2.73594
|   V10 <= ?1.869371: 1 (57361.0/161.0)
|   V10 > ?1.869371
|   |   V4 <= 1.753394: 0 (1830.0/290.0)
|   |   V4 > 1.753394: 1 (2846.0/268.0)
V14 > ?2.73594
|   V17 <= ?2.037394: 1 (1437.0/43.0)
|   V17 > ?2.037394
|   |   V20 <= ?1.016923
|   |   |   V12 <= ?0.075397
|   |   |   |   V1 <= ?2.834603: 0 (1038.0/278.0)
|   |   |   |   V1 > ?2.834603: 1 (1649.0/442.0)
|   |   |   V12 > ?0.075397: 0 (2605.0)
|   |   V20 > ?1.016923
|   |   |   V4 <= 2.520769
|   |   |   |   V14 <= ?0.472026
|   |   |   |   |   V17 <= 0.564693: 0 (29475.0/739.0)
|   |   |   |   |   V17 > 0.564693
|   |   |   |   |   |   V8 <= 0.161143
|   |   |   |   |   |   |   V18 <= ?0.473487: 0 (1844.0)
|   |   |   |   |   |   |   V18 > ?0.473487
|   |   |   |   |   |   |   |   V26 <= ?0.120977
|   |   |   |   |   |   |   |   |   V24 <= ?0.254124: 1 (2046.0/605.0)
|   |   |   |   |   |   |   |   |   V24 > ?0.254124
|   |   |   |   |   |   |   |   |   |   V24 <= 0.230854: 0 (1316.0/153.0)
|   |   |   |   |   |   |   |   |   |   V24 > 0.230854
|   |   |   |   |   |   |   |   |   |   |   V14 <= ?1.420663: 1 (1102.0/368.0)
|   |   |   |   |   |   |   |   |   |   |   V14 > ?1.420663: 0 (1398.0/307.0)
|   |   |   |   |   |   |   |   V26 > ?0.120977: 0 (4800.0/462.0)
|   |   |   |   |   |   V8 > 0.161143: 0 (6165.0)
|   |   |   |   V14 > ?0.472026: 0 (154362.0/1472.0)
|   |   |   V4 > 2.520769
|   |   |   |   V17 <= 1.521261
|   |   |   |   |   V8 <= ?0.135843
|   |   |   |   |   |   V1 <= ?0.114761: 0 (1428.0/158.0)
|   |   |   |   |   |   V1 > ?0.114761: 1 (1976.0/883.0)
|   |   |   |   |   V8 > ?0.135843: 0 (8383.0/553.0)
|   |   |   |   V17 > 1.521261: 1 (1745.0/357.0)
```

Izvor: rad autorice, Weka

Slika 18 prikazuje stablo odlučivanja, a može se vidjeti da je korijen stabla odlučivanja atribut V14, a taj atribut je zapravo na pola ukupnog broja „V“ atributa. Grana se na listove sa atributima V10

i V17, a koji se potom granaju na varijable sa većom ili manjom vrijednošću. Najveći broj transakcija koje su klasificirane kao negativne i sumnjive nalazi se na listu V10.

Atributi V1, V2, …, V28 nose negativne vrijednosti pa je teško odrediti što zapravo znače i koje prave vrijednosti nose. Ovom analizom pokušao se riješiti problem klasne neravnoteže kroz primjenu filtera 'Resample'.

4.5. Izazovi izrade studije slučaja u svrhu otkrivanja sumnjivih bankovnih transakcija

U svijetu bankovnih transakcija, otkrivanje onih sumnjivih predstavlja ključni element u borbi protiv financijskog kriminala i pokušaja krađe identiteta ili određenih iznosa novca. Kreiranje studije slučaja kako bi se identificirale potencijalne nezakonite aktivnosti zahtijeva preciznu analizu podataka, ali se i suočava sa brojnim izazovima, a posebno u pronalaženju povjerljivih baza podataka.

Jedan od glavnih izazova pri izradi studije slučaja, u primjeru ovog rada, je pronalazak povjerljive baze podataka sa uravnoteženim podacima, koja bi pružila relevantne informacije potrebne za analizu sumnjivih bankovnih transakcija. Iako su brojne financijske institucije spremne surađivati u borbi protiv ovakvih nezakonitih događanja, osiguravanje pristupa njihovim podacima složen je i dugotrajan proces. Strogi propisi o zaštiti podataka i osjetljivost informacija često ograničavaju pristup podacima koji bi mogli biti od velikog značaja pri istraživanju. Pristup relevantnim podacima predstavljao je izazov u više faktora. Kao što je već spomenuto, ovakvi podaci su često zaštićeni sigurnosnim protokolima i regulativama, što je ograničavalo mogućnost pristupa potrebnim podacima. Osim toga, čak i kada je pristup bio omogućen, raspoložive baze podataka nisu uvijek bile u formatu koji odgovara potrebama istraživanja. Različiti formati, nedostatak standardizacije i specifičnih informacija otežali su proces samog istraživanja.

Osim sa pristupom podacima, prilikom izrade studije slučaja, došlo je i do suočavanja sa izazovima uravnoteženosti, odnosno neuravnoteženosti podataka. Neuravnoteženost podataka unutar dostupnih baza podataka znači da postoji značajna razlika između broja regularnih transakcija i transakcija koje su potencijalno sumnjive ili vode prijevari. Poznato je da se, u današnjem svijetu, događa sve više sumnjivih transakcija te je nemoguće da stvarna baza podataka sadrži toliku neuravnoteženost i većinu pozitivno klasificiranih transakcija. Ovakva neuravnoteženost podataka,

zapravo, narušava učinkovitost algoritama za analizu podataka jer algoritmi većinom bolje funkcioniraju na uravnoteženom skupu podataka.

Kako bi se ovaj problem neuravnoteženosti podataka riješio, bar jednim dijelom, u procesu izrade studije slučaja korišteni su relevantni filteri i postavke parametara u svrhu približavanja rezultata dostupne baze podataka stvarnom svijetu.

Interpretacija rezultata studije slučaja je također jedan od izazova, a iz razloga što ju treba približiti i prilagoditi svima, a posebno publici koja možda nema znanje iz ovoga područja.

Iako su izazovi u pronalaženju relevantnih podataka predstavljali prepreku, primjenom, u radu već spomenutih filtera i tehnika, kroz ovo istraživanje uspjelo se identificirati obrasce i anomalije koje bi mogle ukazivati na potencijalni financijski kriminal i voditi do sumnjivih bankovnih transakcija.

5. ZAKLJUČAK

Banke se svakodnevno suočavaju sa problemima prevare i prevencije istih. Nastoje suzbiti i otkriti nezakonite aktivnosti koje uključuju bankovne transakcije. Kako raste kompleksnost financijskih sustava, tako tradicionalni sustavi postaju sve manje podobni za otkrivanje sumnjivih transakcija. Kako bi se posao otkrivanja takvih transakcija proveo učinkovitije, banke posežu za jednostavnijim načinom obrade podataka te u svoje sustave implementiraju rudarenje podataka. U ovom radu predstavljen je proces rudarenja podataka, kako općenito, tako i u bankovnom sektoru, kao glavnoj temi rada. Rudarenjem podataka banke dolaze do jednostavnijeg načina obrade podataka koje svakodnevno bilježe, a koji sadrže informacije o računima, transakcijama, kreditima i demografskim podacima klijenata. Zaključeno je da povećana digitalizacija vodi i većem broju podataka koji svakodnevno kolaju, a samim time i većoj potrebi za obradom tih podataka. Da bi banke otkrile korijen problema, potrebno je na početku definirati poslovni problem i izraziti ga u obliku pitanja koje će na kraju procesa dobiti odgovor. Nakon definicije problema potrebno je identificirati ciljeve i pripremiti podatke za obradu. Podaci su prikupljanjem sirovi te ih treba agregirati i obraditi kako bi bili prikladni sustavu za analizu. Modeliranjem se biraju odgovarajući modeli i algoritmi analize za izradu prediktivnih modela. Rezultati analize bitni su za razumijevanje problema, a implementacija istih u stvarno poslovno okruženje osigurava učinkovito funkcioniranje i povezivanje sa drugim aplikacijama i sustavima poduzeća. Zaključuje se da je bitno povezati rudarenje podataka, odnosno rezultate istoga sa drugim sustavima poduzeća kako bi se brzo i efikasno dobila cjelokupna slika o nastalom problemu. Primjena rudarenja podataka bankovnom sektoru donosi veliki niz prednosti koje su značajne pri poboljšanju učinkovitosti poslovanja i donošenja odluka. Rudarenje podataka omogućuje bankama poboljšanje odnosa sa postojećim klijentima i izgradnju povjerenja na temelju prediktivnih modela koje banka stvori iz postojećih baza podataka. Osim dobrog odnosa klijenta prema bankama, bitan je i odnos banke prema klijentu, a rudarenjem podataka baze koja sadrži podatke o već nastalim transakcijama i ostalim bitnim informacijama, baka može spoznati rizik davanja kredita. To je još jedan vid zaštite od neželjenih događaja i posljedica prijevare. Metoda otkrivanja znanja iz baza podataka korištena u ovom radu, stabla odlučivanja, daje nam jednostavne rezultate velikog problema današnjih banaka. Ključan aspekt stabala odlučivanja je procjena važnosti različitih atributa pri donošenju odluka. Ovaj aspekt igra veliku ulogu jer omogućuje identifikaciju ključnih faktora koji utječu na

određeni ishod i istovremeno pruža uvid i dublje razumijevanje podataka. Stabla odlučivanja nude transparentnost samog procesa te je moguće uvidjeti sve događaje iz baze podataka. Online transakcije klasificirane su kao najnepouzdanija i najsumnjivija vrsta transakcija zbog lakoće presretanja sustava i upadanja u podatke korisnika. Iz vizualnog prikaza stabla odlučivanja, koje je napravljeno u istraživanju ovoga rada, jasno se vidi koje transakcije glase za sumnjive, a najviše njih je bilo provedeno online. Stoga se zaključuje da su online transakcije, među ostalim tipovima, jedne od najnesigurnijih te je potrebno provesti dodatne mjere zaštite kako bi se stvorila prevencija pronevjere podataka i novčanih iznosa pri plaćanjima. Može se primijetiti kako se većina transakcija, iz prvog seta podataka, koje su klasificirane kao sumnjive, provelo online. Kako bi se ovakve neželjene situacije izbjegle može se poraditi na mjerama sigurnosti koje dovode do smanjenja sumnjivih transakcija i prijevara pri plaćanju. Prva razina zaštite pri online plaćanju je, zapravo, korištenje stabilne i sigurne Wi-Fi mreže. Često su javne mreže lako dostupne svima i ne zahtijevaju nikakve provjere za korištenje istih, a što dosta puta može dovesti do krađe podataka, što nekih nebitnih do onih najpovjerljivijih, kao što su bankovni podaci. Na taj način podacima korisnika se pristupa lako i velikom brzinom te korisnik, bez da je svjestan situacije, može ostati bez privatnosti. U zadnje vrijeme se često koristi dvofaktorska autentifikacija, čak i u primjerima društvenih mreža, kao što je Instagram i ostale Meta aplikacije. Dvofaktorska autentifikacija osigurava identifikaciju korisnika putem lozinke i npr. SMS koda, tako da korisnik mora znati, ne samo lozinku, nego i jedinstveni kod koji ima ograničeno vrijeme korištenja i vidljiv je samo na uređaju korisnika. Ova metoda zaštite primjerena je i za bankovno poslovanje i online plaćanja jer osigurava dodatnu zaštitu i smanjuje postotak mogućnosti za neželjenim gubitkom podataka. Kako bi online transakcije bile korisnije, preporuka je i uvesti autentifikaciju putem tokena, biometrijskih podataka ili jednokratnih lozinki. Također, sve online transakcije trebale bi biti kriptirane kako bi se spriječilo presretanje i neovlašten pristup osjetljivim podacima. Ključnu ulogu u zaštiti transakcija igra i edukacija korisnika. Banka korisnike treba educirati o sigurnosnim praksama poput prepoznavanja phishing napada, sigurnog rukovanja lozinkama i prepoznavanja sumnjivih aktivnosti na svojim računima. Korištenje sigurnosnih platformi za plaćanje te redovito praćenje bankovnih transakcija, također su važni za održavanje sigurnosti tijekom online transakcija. Što se tiče i drugih načina plaćanja, a ne samo online transakcija, banke bi trebale osigurati i odrediti jasnu politiku identifikacije korisnika, a koja zahtijeva provjeru identiteta svih korisnika koji obavljaju transakcije, posebno one koje su povezane sa velikim iznosima i sumnjivim plaćanjima. Rudarenje

podataka u svrhu otkrivanja sumnjivih bankovnih transakcija može značajno poboljšati sigurnost i spriječiti prijevare, a u kontekstu otkrivanja prijevara banke bi trebale točno odrediti što žele postići korištenjem rudarenja podataka. Potrebno je skupljati relevantne podatke koji mogu poslužiti za analizu transakcija, a oni bi trebali uključivati podatke o transakcijama, korisnicima i demografske podatke. Automatizacijom procesa rudarenja podataka banke bi omogućile brzu i učinkovitu analizu velikih količina podataka, kao i identifikaciju sumnjivih transakcija u stvarnom vremenu. Banke bi trebale integrirati rezultate rudarenja podataka sa drugim sigurnosnim alatima i sustavima kako bi se detekcija i prevencija prijevara dodatno poboljšala, a također je potrebno redovito i kontinuirano optimizirati model rudarenja podataka kako bi se održala učinkovitost analize.

LITERATURA

Knjige i znanstveni članci:

1. Aggarwal, C. C. (2015) *Data mining: The textbook* (Vol. 1). New York: Springer.
2. Apté, C., Weiss, S. (1997) *Data mining with decision trees and decision rules. Future generation computer systems, 13*(2-3), 197-210
3. Banov, R., Valent, A., Anušić, J. (2022) *Neuronske mreže za početnike. Poučak: Časopis za metodiku i nastavu matematike, 23*(90), 24-34
4. Bhambri, V. (2011) *Application of Data Mining in Banking Sector. International Journal of Computer Science and Technology, 2*. Preuzeto s http://www.ijcst.com/vol22/1/vivek.pdf
5. Cios, K. J., Pedrycz, W., Swiniarski, R. W. (2012) *Data mining methods for knowledge discovery* (Vol. 458). Springer Science & Business Media
6. Ćurko, K., Španić Kezan, M. (2016) *Skladištenje podataka: Put do znanja i poslovne inteligencije.* Zagreb, Sveučilište u Zagrebu, Ekonomski fakultet
7. Demetis, D. S. (2018) Fighting money laundering with technology: A case study of Bank X in the UK. *Decision Support Systems, 105*, 96-107
8. Gaur, P. (2012) *Neural networks in data mining. International Journal of Electronics and Computer Science Engineering (IJECSE, ISSN: 2277-1956), 1*(03), 1449-1453
9. Han, J., Pei, J., Tong, H. (2022) *Data mining: Concepts and techniques*. Morgan Kaufmann
10. Ibrahim, H., Yasin, W., Udzir, N. I., Hamid, N. A. W. A. (2016) *INTELLIGENT COOPERATIVE WEB CACHING POLICIES FOR MEDIA OBJECTS BASED ON J48 DECISION TREE AND NAIVE BAYES SUPERVISED MACHINE LEARNING ALGORITHMS IN STRUCTURED PEER-TO-PEER SYSTEMS. Journal of Information and Communication Technology, 15*(2), 85-116
11. Ilic, M., Spalevic, P., Veinovic, M., Alatresh, W. S. (2016) *Students' success prediction using Weka tool. Infoteh-Jahorina, 15*, 684-688

12. Kantardzic, M. (2011) *Data mining: Concepts, models, methods, and algorithms*. John Wiley & Sons

13. Kodinariya, T. M., Makwana, P. R. (2013) Review on determining number of Cluster in K-Means Clustering. *International Journal, 1*(6), 90-95

14. Kulkarni, E. G., Kulkarni, R. B. (2016) Weka powerful tool in data mining. *International Journal of Computer Applications, 975*, 8887

15. Kumbhare, T. A., Chobe, S. V. (2014) An overview of association rule mining algorithms. *International Journal of Computer Science and Information Technologies, 5*(1), 927-930

16. Larose, D. T., Larose, C. D. (2014) *Discovering knowledge in data: an introduction to data mining* (Vol. 4) John Wiley & Sons

17. Magee, J. F. (1964) Decision trees for decision making (pp. 35-48) Brighton, MA, USA: Harvard Business Review

18. Maimon, O. Z., Rokach, L. (2014) *Data mining with decision trees: theory and applications* (Vol. 81). World Scientific

19. Marbán, Ó., Mariscal, G., Segovia, J. (2009) A data mining & knowledge discovery process model. In *Data mining and knowledge discovery in real life applications*. IntechOpen.

20. Mirošević, I. (2016) Algoritam k-sredina. *KoG, 20*(20), 91-98

21. Miyan, M. (2017) Applications of Data Mining in Banking Sector. *International Journal of Advanced Research in Computer Science, 8*(1)

22. Mirza, S., Mittal, S., Zaman, M. (2018) Decision support predictive model for prognosis of diabetes using SMOTE and decision tree. *International Journal of Applied Engineering Research, 13*(11), 9277-9282

23. Mocanu, M. (2016) Data mining in the banking system. *Cogito, 8*(1), 78

24. N. Vlahović (ur.). (2013)., Temeljne vještine poslovne informatike, 1. izdanje, Mikrorad, Zagreb

25. Ostapchenya, D. (2021) The Role of Big Data in Banking: How do Modern Banks Use Big Data. Preuzeto s https://www.finextra.com/blogposting/20446/the-role-of-big-data-in-banking-how-do-modern-banks-use-big-data

26. Panian, Ž., Spremić, M. (2007) *Korporativno upravljanje i revizija informacijskih sustava*. Zagreb, Zgombić i partneri

27. Panian, Ž. (2007) *Poslovna inteligencija: Studije slučajeva iz hrvatske prakse*. Narodne novine

28. Pejić Bach, M. (2005) Rudarenje podataka u bankarstvu. *Zbornik ekonomskog fakulteta u Zagrebu, 3*(1), 181-193

29. Preethi, M., Vijayalakshmi, M. (2017) Data Mining In Banking Sector. *International Journal of Advanced Networking and Applications, 8*(5), 1-4.

30. Provost, F., Fawcett, T. (2013) *Data Science for Business: What you need to know about data mining and data-analytic thinking.* O'Reilly Media, Inc.

31. Rovčanin, A., Mataradžija, A., Mataradžija, A. (2012) Upravljanje znanjem kroz primjenu alata poslovne inteligencije

32. Scott, I., Svinterikou, S., Tjortjis, C., Keane, A. (1998) Experiences of using Data Mining in a Banking Application. Preuzeto s https://www.ihu.edu.gr/tjortjis/Experiences%20of%20using%20Data%20Mining%20in%20a%20Banking%20Application.pdf

33. Singh, A., Agray, S. K. (2017). Role of Data Mining: A Survey and its Implications. *International Journal of Advanced Research in Computer Science*, *8*(5).

34. Spremić, M. (2017) *Sigurnost i revizija informacijskih sustava u okruženju digitalne ekonomije*. Zagreb, Sveučilište u Zagrebu, Ekonomski Fakultet

35. Srivastava, S. (2021) Process mining techniques for detecting fraud in banks: A study. *Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12*(12), 3358-3375

36. Stancu, A. M. R., Mocanu, M. (2016) Expanding K-means algorithm for absolute data. *Knowledge Horizons. Economics*, *8*(2), 163

37. Stanišić, M. (2007) Uloga interne revizije u otkrivanju i sprečavanju prevara u bankama. *Bankarstvo*, *36*(1-2), 20-33

38. Šimec, A., Lozić, D. (2020) Rudarenje podataka. U A. Karmela & S. Stojakovic-Celustka (Ur.), Nove tehnologije u primjeni (str. 57-65). Zagreb: Zagrebačka škola ekonomije i managementa

39. Varga Mladen, S. (ur.). (2016) *Informacijski sustavi u poslovanju*. Zagreb, Sveučilište u Zagrebu, Ekonomski fakultet Zagreb

40. Verma, A. (2019) Evaluation of classification algorithms with solutions to class imbalance problem on bank marketing dataset using WEKA. *International Research Journal of Engineering and Technology*, *5*(13), 54-60

41. Voican, O. (2020) Using data mining methods to solve classification problems in financial-banking institutions. *Economic Computation and Economic Cybernetics Studies and Research*, *54*(1), 159-176

42. Zekić-Sušac, M., Frajman-Jakšić, A., Drvenkar, N. (2009) Neuronske mreže i stabla odlučivanja za predviđanje uspješnosti studiranja. *Ekonomski vjesnik: Review of Contemporary Entrepreneurship, Business, and Economic Issues*, *22*(2), 314-327.

43. Žapčević, S., Butala, P. (n.d.) OTKRIVANJE ZNANJA I METODE RUDARENJA PODATAKA U PROIZVODNIM SISTEMIMA

Web stranice:

1. mehmoodabbas616 (2023, September 20) Creditcardfrauddetection, Kaggle. https://www.kaggle.com/code/mehmoodabbas616/creditcardfrauddetection/input
2. Fraud Detection. (2022, September 12) Kaggle. https://www.kaggle.com/datasets/whenamancodes/fraud-detection
3. Sayad, S. (n.d.), Assotiation Rules, preuzeto s: https://saedsayad.com/association_rules.htm

POPIS SLIKA I TABLICA

ŽIVOTOPIS STUDENTA

Anita Tadić

Državljanstvo: hrvatsko **Datum rođenja:** 01/10/1999 **Spol:** Žensko **Telefonski broj:** (+385) 921776788

E-adresa: tadicanita99@gmail.com

Kućna: Zlatarska ulica, 10000 Zagreb (Hrvatska)

O MENI

Apsolventica na studiju Poslovne ekonomije, smjer Menadžerska informatika. Bivša članica karate reprezentacije Hrvatske.

RADNO ISKUSTVO

Konobarica
Vertigo bar [07/2018 – 09/2018]

Mjesto: Župa Dubrovačka
Zemlja: Hrvatska

Konobarica
Pepper's Bar [09/2019 – 04/2022]

Mjesto: Zagreb
Zemlja: Hrvatska

Promocija
Carlsberg Croatia d.o.o. [01/05/2021 – 31/05/2021]

Mjesto: Zagreb
Zemlja: Hrvatska

Pomoćnica u trgovini
Terranova [10/2021 – 04/2022]

Mjesto: Zagreb
Zemlja: Hrvatska

Administrativna radnica
Agencija za strukovno obrazovanje i obrazovanje odraslih [01/04/2022 – 30/06/2022]

Mjesto: Zagreb
Zemlja: Hrvatska

Konobarica
Ugostiteljski obrt Bonaca [01/07/2022 – 31/08/2022]

Mjesto: Supetar, Brač
Zemlja: Hrvatska

Asistentica u marketingu
Apcom d.o.o. [11/2022 – Trenutačno]

Mjesto: Zagreb
Zemlja: Hrvatska

Online

Asistentica u prodaji
Optika Stepinac [01/05/2023 – 15/07/2023]

Mjesto: Zagreb
Zemlja: Hrvatska

Pomoćni tajnički poslovi
Karate klub Hercegovina-Zagreb [01/09/2023 – Trenutačno]

Mjesto: Zagreb
Zemlja: Hrvatska

OBRAZOVANJE I OSPOSOBLJAVANJE

Opća gimnazija
Srednja škola Prozor [09/2014 – 06/2018]

Adresa: 88440 Prozor-Rama (Bosna i Hercegovina)

Studentica
Ekonomski fakultet Zagreb [10/2018 – Trenutačno]

Adresa: 10000 Zagreb (Hrvatska)

JEZIČNE VJEŠTINE

Materinski jezik/jezici: **hrvatski**

Drugi jezici: **engleski** | **njemački**

DIGITALNE VJEŠTINE

MS Office (MS Word, MS Powerpoint, MS Excel, MS) / Osnovno poznavanje programskih jezika Python i C# / Osnove poznavanja rada u Oracle SQL Devepoleru, Bizagi Modelaru te Doctusu

VOZAČKA DOZVOLA

Vozačka dozvola: B

POČASTI I NAGRADE

Dekanova nagrada
Ekonomski fakultet Sveučilišta u Zagrebu [02/12/2019]

Dekanova nagrada za izniman doprinos u području sveučilišnog i fakultetskog sporta