TARTU ÜLIKOOL

Majandusteaduskond

Krister Jaanhold

RAHAPESU TUVASTAMINE MASINÕPPE MEETODITE ABIL TRANSFERWISE LTD NÄITEL

Bakalaureusetöö

Juhendaja: Oliver Lukason

Ettevõttepoolne juhendaja: Taavi Tamkivi

Tartu 2016

Soovitan suunata kaitsmisele ……..

teadur Oliver Lukason

Kaitsmisele lubatud ..... 2016. a

Olen koostanud töö iseseisvalt. Kõik töö koostamisel kasutatud teiste autorite tööd, põhimõttelised seisukohad, kirjandusallikatest ja mujalt pärinevad andmed on viidatud.

……..

Krister Jaanhold

SISUKORD

SISSEJUHATUS

Seoses tehnoloogia arengu ning Interneti leviku eksponentsiaalse kasvuga on kurjategijatel tekkinud aina rohkem uusi võimalusi, kuidas kriminaalsel teel saadud vahendeid konverteerida õiguspärasteks, teisisõnu – tegeleda rahapesuga. Buchanani (2004: 117) kohaselt on rahapesu finantskuritegu, mis võib esineda nii kuritegelikul teel saadud raha päritolu hägustamise kui ka rahvusvahelise terrorismi finantseerimise näol. Peamiselt pärineb kuritegelikul teel saadud raha uimastikaubandusest, relvade müügist, prostitutsiooni vahendamisest või mõnest muust organiseeritud kuriteo vormist, seega on sellel ulatuslikud tagajärjed nii inimeste igapäeva elule, ühiskonna turvalisusele kui ka majanduse stabiilsusele. Probleemi mastaapsusest ajendatuna keskendub käesolev bakalaureusetöö rahapesu ja rahvusvahelise terrorismi finantseerimise tuvastamisele, rakendades selleks erinevaid masinõppe ning andmekaeve meetodeid.

Suurenenud konkurents finantsteenuseid osutavate ettevõtete vahel on loonud ideaalse keskkonna rahapesuks, sest paljud rahapesuga seotud tehingukulud ning riskid on elimineeritud või nende osatähtsust vähendatud. Lisaks hindadega konkureerimisele on saanud aktuaalseks ka võimalikult paindliku ja kasutajasõbraliku teenuse osutamine, millega kaasneb suurenenud anonüümsus – kliente ei tülitata täiendava informatsiooni küsimisega. Sellest tulenevalt on Le-Khac ja Kechadi (2010: 577) toonud välja, et raha pesemine terrorismi finantseerimise eesmärgiga on muutumas aina keerulisemaks ja aktuaalsemaks probleemiks, sest aina kergem on jääda anonüümseks. 9/11 sündmuste Rahvusliku Komisjoni raportis (*The 9/11…* 2004, 254) toodi välja, et rahaülekanded olid üks peamisi viise, kuidas Al-Qaeda terrorirünnakut finantseeris, seega on eriti oluline finantsteenuseid osutavates ettevõtetes implementeerida sobivaid kontrollsüsteeme, et ennetada rahapesu ning ülemaailmset terrorismi finantseerimist.

Rahapesu tuvastamine ettevõtte tasandil on väga keeruline protsess, sest erinevalt teistest finantspettuse vormidest ei kaasne sellega rahalisi kulusid ning sellest tulenevalt ei saa

kunagi täie kindlusega väita, milline klient tegeles rahapesuga. Olgugi, et rahapesu ei too ettevõtte tasandil kaasa otsest rahalist kahju, võivad sellega kaasneda ulatuslikud tagajärjed nii maine languse, litsentsist ja partneritest ilma jäämise kui ka üüratute trahvide näol, mis omakorda destabiliseerivad majandust. Peale selle muudab rahapesu tuvastamise keeruliseks protsessi dünaamilisus – ka rahapesijad õpivad ja võtavad kasutusele uusi meetodeid.

Zdanowics (2004: 53) on toonud välja, et 2001. aasta 11. septembri terrorirünnakud tähistavad uut ajastut terrorismi finantseerimise ja rahapesu tuvastamises. Olulisteks tööriistadeks on saanud nii infotehnoloogia kui ka andmekaeve – levinuimad statistilised meetodid pettuste tuvastamiseks tuginevad tegelike ja eeldatavate andmete võrdlemisele. Zhang ja Zhou (2004: 513) kohaselt on andmekaeve paljutõotav lahendus dünaamiliste ja mittelineaarsete seoste tuvastamisel ning sellest tulenevalt võib see pakkuda lahendust nii andmete rohkusest kui ka rahapesu dünaamilisusest tingitud probleemidele. Bolton ja Hand (2002: 236) on toonud välja, et pettuste tuvastamiseks loodud statistilised mudelid on üldiselt küll efektiivsed, ent ettevõttespetsiifilisusest tingituna ei ole olemas üht ja universaalset mudelit, mis töötaks alati.

Käesoleva bakalaureusetöö eesmärk on luua masinaõppe meetodeid ning rahvusvahelisi rahaülekandeid osutava ettevõtte TransferWise LTD andmeid kasutades mudel, mis suudab tuvastada rahapesukahtlusega kliente. Vastavalt eesmärgile on autor püstitanud järgnevad uurimisülesanded:

- anda ülevaade rahapesu olemusest ja selle aktuaalsusest,
- tuua varasema teaduskirjanduse põhjal välja finantspettusi tuvastada võimaldavad masinõppe meetodid,
- anda varasema kirjanduse põhjal ülevaade rahapesu tuvastada võimaldavatest muutujatest,
- kirjeldada uurimismetoodikat ning lähteandmeid,
- rakendada ümbernäidistamismeetodeid andmete puhastamiseks ja tasakaalustamiseks,
- püstitada rahapesu tuvastada võimaldav mudel.

Lähtudes püstitatud eesmärgist ja uurimisülesannetest on bakalaureusetöö jagatud kaheks osaks. Esimeses, teoreetilises peatükis annab autor esmalt ülevaate rahapesu olemusest ja aktuaalsusest ning toob varasema teaduskirjanduse põhjal välja seda tuvastada võimaldavad muutujad ja statistilised meetodid. Töö empiirilise osa esimeses alapeatükis kirjeldab autor uurimismetoodikat ja lähteandmeid. Empiirilise osa teises alapeatükis rakendab autor erinevaid statistilisi meetodeid andmete eeltöötlemiseks ning viimaks püstitab mudeli, mis võimaldab tuvastada rahapesukahtlusega kliente. Töö on väärtuslik nii antud ettevõttele kui ka ühiskonnale üldiselt – tugevam riskijuhtimine tähendab ettevõttele kiiremat kasvu ning aitab vältida rahvusvahelist terrorismi finantseerimist ja sellega kaasnevaid ohte. Peale selle esitab töö autor raamistiku ja toob välja statistilised meetodid ning nende kombinatsioonid, millele tuginedes on võimalik püstitada rahapesu efektiivselt tuvastada võimaldav mudel.

Tööd iseloomustavad märksõnad – rahapesu, organiseeritud kuritegevus, finantskuriteod, uimastikaubandus, andmekaeve, terrorismi finantseerimine, masinõpe.

1. RAHAPESU OLEMUS JA SEDA TUVASTADA VÕIMALDAVATE STATISTILISTE MEETODITE TUTVUSTUS

1.1 Rahapesu olemus ja aktuaalsus

Käesolevas alapeatükis antakse varasema empiirilise kirjanduse põhjal ülevaade rahapesu aktuaalsusest ning selle olemusest. Rahapesu kui üks finantspettuse vorm on defineeritud erinevate autorite poolt üsna üheselt – tegu on finantskuriteoga, mis võib esineda nii kuritegelikul teel saadud raha päritolu hägustamise kui ka rahvusvahelise terrorismi finantseerimise näol (Buchanan 2004: 117). Finantspettused, sealhulgas rahapesu on probleem, millel on ulatuslikud tagajärjed nii finantsmaailmas kui ka inimeste igapäevaelus – Bhattacharya ja West (2015: 47) on toonud välja, et erinevad pettused võivad vähendada ettevõtete usaldusväärsust, destabiliseerida majandust ning seeläbi mõjutada inimeste elukallidust. Raha pesemine on üks suurimaid takistusi rahvusvahelise finantssüsteemi efektiivse toimimise tagamisel – varimajandus ning korruptsioon finantsturgudel vähendab süsteemi usaldusväärsust ning võib kaasa tuua majanduse destabiliseerumise. Gilmour (2015: 2) on toonud välja, et rahapesu on väga lähedaselt seotud organiseeritud kuritegevusega, mis genereerib suure hulga sularaha, mis tuleb sisestada finantssüsteemi selleks, et vältida võimudele vahele jäämist või varguse ohvriks langemist. Buchanani (2004: 117) kohaselt pärineb kuritegelikul viisil saadud raha peamiselt järgnevatest tegevustest – relvade müük, uimastikaubandus, prostitutsiooni vahendamine, organiseeritud kuritegevus, ametiseisundi kuritarvitamine, insaidertehingud ja väljapressimised, mis mõjutavad nii inimeste turvalisust kui ka varimajanduse aktiivsust.

Vaatamata sellele, et teadlikkus rahapesu ja terrorismi finantseerimise ohust on kasvanud, ei ole antud teemat siiani piisavalt uuritud virtuaalkeskkondade näitel (Liu 2011: 85). Interneti leviku eksponentsiaalne kasv on toonud kaasa ka hulgaliselt uusi võimalusi

pettusteks. Irwin *et al.* (2014: 70) on toonud välja, et virtuaalkeskkonnad on eriti sobilikud terrorismi finantseerimiseks, sest paljud rahapesuga seotud riskid on elimineeritud või nende tähtsust vähendatud. Konkurents rahvusvaheliste maksete osutamisel läbi virtuaalkeskkondade on oluliselt kasvanud ning selle tagajärjel on oluliselt vähenenud ka rahapesuga kaasnevad tehingukulud, mis muudab virtuaalkeskkonnad ideaalseks rahapesuplatvormiks. Buchanan (2004: 116) kohaselt on uute finantsinstrumentide ja kauplemisvõimaluste leviku ning finantsturgude likviidsuse paranemise tulemusena aina kergem luua uusi rahapesu süsteeme. Sarnaselt on Sullivan (2015: 18) on toonud välja, et finantsteenuseid osutavate ettevõtete puhul esineb palju kanaleid, mille kaudu raha on võimalik pesta.

Käesolev bakalaureusetöö on üles ehitatud rahvusvaheliste rahaülekannetega tegeleva Eesti päritolu idufirma TransferWise andmetel. TransferWise on Finantsinspektsiooni (*Financial Conduct Authority*) poolt reguleeritud kui e–raha teenuse osutaja (*e-money provider*). Rahapesu regulatsioonid (*The Money Laundering Regulations,* 2007: 6) kohustavad Ühendkuningriigis nii krediidiasutusi, finantsintiutsioone kui ka audiitoreid implementeerima sobivaid kontroll- ja raporteerimissüsteeme, et ennetada rahapesu ning rahvusvahelist terrorismi finantseerimist. Iga ettevõtte töötaja peab olema lugenud ettevõtte rahapesu tõkestamise juhendit, mis annab ülevaate sellest, kuidas ettevõte käsitleb rahapesust tulenevaid riske. Poliisist ning rahvusvahelistest soovitustest tulenevalt on iga töötaja kohustus raporteerida riskijuhtimise osakonnale, kui täheldatakse mõne kliendi puhul käitumist, mis võib viidata rahapesule. Kahtlase käitumise raporteid (*suspicious activity report*) uurivad põhjalikult ettevõttesiseselt rahapesu tõkestamisega tegelevad agendid, kes vastutavad ka tõenäoliste kurjategijate politseile raporteerimise eest. Kõik põhjendatud kahtlusega rahapesu juhtumid tuleb ette kanda Riiklikule Kuritegevuse Agentuurile (*National Crime Agency*), kes analüüsides kõikide ettevõtete poolt tehtud raporteid otsustab, kas konkreetne juhtum vajab politsei sekkumist ja põhjalikumat uurimist või mitte.

Rahapesu tuvastamise puhul on tegu väga keerulise ülesandega, sest tegu on dünaamilise protsessiga, läbi mille ka petturid õpivad ning mõtlevad välja uusi viise, kuidas oma varade algupära varjata (Gao, Xu 2009: 1495). Rahapesu tuvastamine ettevõttetasandil on oluliselt raskendatud, sest kurjategijaid eelistavad kasutada mitmeid teenuspakkujaid

ja iga finantsteenuseid osutav ettevõte näeb vaid üht osa üldpildist. Näiteks on väga kerge rahapesijatel jätta legitiimse kliendi mulje, kui maksed on jagatud kümnete erinevate teenuspakkujate vahel. Teiseks – erinevalt teistest pettuse liikidest ei kaasne rahapesuga otsest rahalist kulu, seega on võimatu määrata üheselt, kas klient tegeles rahapesuga või mitte. Olgugi, et rahapesu ise ei too ettevõttele rahalist kahju, võivad sellega peale maine languse, litsentsist ja partneritest ilma jäämise kaasneda väga suured trahvid. Näiteks aastal 2011 trahviti HSBC panka 1,9 miljardi dollari väärtuses, kuna regulaatorid leidsid, et HSBC teenuseid kasutasid Mehhiko narkodiilereid (Barrett, Perez 2012). Üüratute trahvide maksmine pankade poolt omakorda destabiliseerib majandust ning seega ei avalda rahapesu ja terrorismi finantseerimine ohtu mitte ainult turvalisusele, vaid ka finantsmaailmale üldisemalt.

Zdanowics (2004: 53) on toonud välja, et pärast 2001. aasta terrorirünnakuid USA-s on eriti aktuaalseks saanud terrorismi rahastamine kui rahapesu üks vorm. 9/11 sündmuste Rahvusliku Komisjoni raportis (*The 9/11*... 2004, 254) toodi välja, et rahaülekanded olid üks peamisi viise, kuidas Al-Qaeda 2001. aasta 11. septembri terrorirünnakuid finantseeris. Rääkides terrorismi finantseerimisest, tuleb täheldada, et mitte kogu raha ei pruugi olla kriminaalsel teel omandatud, vaid võib olla täiesti legaalselt teenitud. Terroriaktide finantseerimise puhul ei ole peamiseks eesmärgiks rahaline tulu, vaid terroriaktide julgustamine, planeerimine või sellele kaasa aitamine. Palmer (2005: 14) on toonud intervjuu põhjal välja, et piirid rahapesu ja terrorismi finantseerimise vahel on küll hägused, ent neid ei tohiks kohelda kui erinevaid sündmusi. Nii terrorismi finantseerimise kui ka rahapesu puhul on ühine eesmärk jaotada raha erinevate kontode vahel nii, et see ei ärataks kellegi tähelepanu. Irwin *et al.* (2011: 85) on toonud välja, et rahapesijatel ja terrorismi finantseerijatel on pisut erinevad eelistused kasutatavate meetodite suhtes – terrorirühmitused eelistavad neid kanaleid, kus on tagatud maksimaalne anonüümsus. Selleks sobivad väga hästi virtuaalkeskkonnad ja -valuutad.

On viidud läbi mitmeid uurimusi eesmärgiga leida efektiivset ja usaldusväärset hinnangut sellele, kui palju raha aastaringselt pestakse. Seniste uuringute kohaselt pestakse ainuüksi Aasias iga aasta ligi 200 miljardit dollarit (Lilley 2003: 50), seevastu globaalselt 100 triljonit, ehk ligikaudu 2% maailma kogutoodangust aastas (*World Bank...* 2003: vii). Rahapesu puhul on tegu kriminaaltegevuse kõrvalproduktiga, seega on Bagella *et al.*

(2013: 207) kohaselt eriti oluline osata mõõta efektiivselt rahapesu mastaapi selleks, et selgitada välja kriminaalsel teel omandatud raha allikas. Mitterahalise mõõdikuna on välja pakutud kahtlase tegevuse või ebatavalise makse (*unusual transaction*) raportite esitamise trendi (Unger, Hertog 2012: 294). Paraku võib raportite arvu suurenemine viidata kolmele erinevale aspektile: rahapesu on hakanud laialdasemalt levima, selle ennetamisele on hakatud rohkem tähelepanu pöörama, st tuvastatakse rohkem kahtlaseid juhtumeid või on ettevõtjad hakanud järelvalvet rohkem kartma ning teevad seetõttu rohkem raporteid.

Tuginedes Buchanani (2004: 117) tööle, võib raha pesemise jagada kolme järgnevasse etappi:

- paigutamine (*placement*),
- kihitamine (*layering*),
- integratsioon (*integration*).

Paigutamine on etapp, mille käigus sisestatakse finantsvarad süsteemi. Varasemas kirjanduses on toodud välja, et tegu on kõige riskantsema etapiga, kus vahele jäämise tõenäosus on kõige suurem. Rahapesijad paigutavad illegaalsel teel saadud varad finantssüsteemi, kasutades selleks erinevaid tehnikaid – peamiselt peetakse silmas sularaha deposiidi tegemist pangakonto(de)le ja/ või muudele finantsinstrumentidele. Irwin et al. (2011: 94) on oma uurimuses toonud välja, et rahapesu ja terrorismi finantseerimise puhul on enimlevinud paigutamismeetoditeks "smurfimine" (*smurfing*) ja struktureerimine (*structuring*). Smurfimine on meetod, mille puhul pesemist vajav raha jagatakse ära mitmete inimeste vahel, kes selle enda nime alt erinevatele kontodele edasi saadavad. Mida rohkemate inimeste vahel must raha ära jaotatakse, seda väiksem tõenäosus on kahtlust äratada. Struktureerimine seevastu on tegevus, mille käigus jaotatakse must raha väiksemateks osadeks nii, et tehakse ühe suure makse asemel mitu väikest makset. Peamiselt kasutavad neid kahte meetodit maksude vältijad, narkodiilerid ja inimkaubitsejad selleks, et äratada võimalikult vähe tähelepanu (Tupman 2010: 152, Irwin *et al*. 2011: 94). Seega on rahapesijatel ja terrorismi finantseerijatel ohutum teha väikseid makseid mitme inimese poolt, kui kanda ühe korraga üle suur summa.

Kihitamine on etapp, kus jaotatakse raha korduvalt erinevate kontode vahel eesmärgiga tekitada võimalikult palju kihte, et varade algupära tuvastamine oleks võimalikult

keeruline. Rahapesijatel on tendents jaotada kriminaalsel teel saadud raha erinevate riikide vahel – eelistatakse vähemreguleeritud kontrollmehhanismidega riike (Buchanan, 2004: 116). Simser (2013: 43) on toonud välja, et raha paigutamine leebema kontrolliga riikide finantssüsteemi on levinud meetod eelkõige uimastitega kaubitsejate puhul. Kihitamise faasis on levinuimaks tehnikaks riiulfirmade kasutamine, seevastu narkokaubitsejad eelistavad pangatšekkide ning -vekslite kasutamist, lisaks ka rahaülekannete pakkujaid (Irwin *et al.* 2011: 95). Tulenevalt sellest, et virtuaalkeskkondades nagu TransferWise ei saa kasutada deposiidi tegemiseks veksleid ega tšekke, on lähemat tähelepanu pööratud vaid ülekannete käitumuslikkusele, sh smurfimisele ja struktureerimisele. Sullivan (2015: 19–20) tõi välja, et rahaülekanded on üheks enimlevinud kihitamise meetodiks – raha jaotamine erinevate valuutade ja riikide vahel muudab selle algupära määramise eriti raskeks. Antud aspektist lähtudes on rahvusvahelistele maksetele spetsialiseerunud finantsasutused eriti haavatavad. Rahvusvaheliste soovituste kohaselt (*The FATF...* 2012: 14) peavad finantsinstitutsioonid rakendama hoolsusmeetmeid ja küsima teatud lävendini jõudes klientidelt lisainfot ning sellest tulenevalt võib maksete struktureerimine ja/ või smurfimine olla ka üheks kihitamismeetodiks. Buchanani (2004: 118) kohaselt on nii kihitamise kui ka integratsiooni faasis levinud meetodiks fassaadettevõtete (*front companies*) kasutamine. He (2010: 24) kohaselt on fassaadfirmade puhul tegu kurjategijate poolt loodud ettevõtetega, millel on nii õiguspärane sissetulek kui ka tegevusvaldkond, ent mille eesmärk pole kasumi teenimine vaid raha pesemine. Peamiselt on tegu sularahaintensiivsete ettevõtetega, mille puhul on raha algupära tuvastamine raske. Fassaadettevõtete puhul võib lisaks olla tegu nii eksporditavate kui ka imporditavate kaupade valehindamisega, mille puhul eristatakse peamiselt kahte strateegiat – topeltarvete koostamine ning eksporditavate toodete ala- või ülehindamine (Buchanan 2004: 119). Antud tüüpi ettevõtted on seega rahapesu seisukohast kõrge riskiga, sest raha päritolu võib olla väga raske tuvastada, kuna ettevõtte nii õiguspäraselt kui ka kriminaalsel teel saadud varad on segatud. Buchanan (2004: 118) kohaselt on üheks levinud rahapesu meetodiks valearvete koostamine (*misinvoicing*) rahvusvahelises kaubanduses. Peale fassaadettevõtete eristatakse veel ka riiulifirmasid – He (2010: 24) kohaselt on tegu ettevõtetega, millel ei ole organisatoorset struktuuri ega aktiivset äritegevust. Zeldin (1998: 297) on toonud välja, et rahapesijad kasutavad valearvete

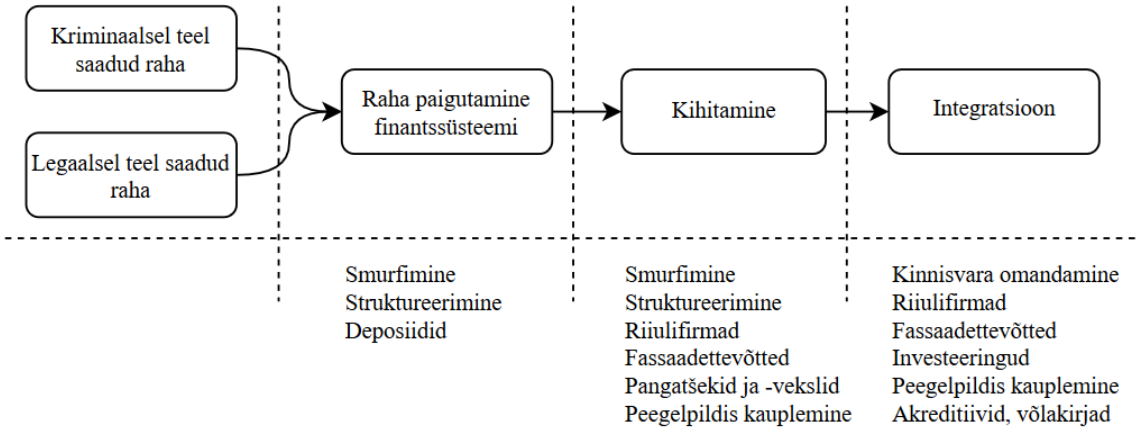
koostamiseks riiulifirmasid selleks, et luua arveid toodete ja teenuste eest, mida tegelikult ei eksisteeri. Näiteks tuvastas USA toll, et 1990. aastal importisid uimastikartelli kuuluvad ettevõtted Boliiviast 129% seal riigis üldse toodetud kullast (*Ibid.*). Buchanan (2004: 119) on toonud välja, et nii kihitamise kui ka integratsiooni faasis on üheks levinud meetodiks peegelpildis kauplemine (*mirror-image trading*), mille puhul tehakse väärtpaberiturul tehinguid kontode vahel, mis mõlemad kuuluvad ühele ja samale inimesele – näiteks ühe kontoga müüakse väärtpabereid ning teise kontoga ostetakse sama arv väärtpabereid tagasi.

Integratsioon on ebaseadusliku tulu konverteerimine legitiimseks tuluks läbi normaalse finantstegevuse. Levinuimaks integratsioonimeetodiks Irwin *et al.* (2011: 96) kohaselt on kinnisvara omandamine. Lisaks sellele on toodud välja investeeringud kapitaliturgu, kaubandusettevõtete loomine, valuutadevahelised ülekanded ja intensiivsete rahavoogudega ettevõtete loomine. Erinevat tüüpi ettevõtte kasutamine nii kihitamise kui ka integratsiooni faasis on tingitud eelkõige sellest, et ebastabiilne maksemuster äriprofiilidel äratab oluliselt vähem kahtlust, sest ettevõtte majandustegevus võib olla sesoonne. Investeeringute all peetakse muuseas silmas ka luksuskaupade soetamist. Buchanani (2004: 117) kohaselt kasutatakse pestud raha süsteemi sisestamiseks nii akreditiive, võlakirju kui ka muid finantsinstrumente. Lisaks on sama autor (2004: 120) toonud välja, et tihti ei pruugi must raha peavoolu finantssüsteemi üldse jõudagi, vaid sisestatakse põrandaalusesse pangandussüsteemi – raha vahetatakse vahetuspunktides nii-öelda piletite vastu, mis on võimalik lunastada vastavalt mõnes teises punktis sama rahasumma vastu. Tulenevalt käesoleva töö uurimiseesmärgist, jääb põrandaalune pangandus põhjalikuma vaatluse alt välja.

Lisaks on He (2010:16) toonud ühe rahapesu meetodina välja smugeldamise (*smuggling*), mille eesmärk on raha salaja üle piiri toimetada ning seejärel finantssüsteemi sisestada ja kasutada seda kinnisvarasse või ettevõtetesse investeerimisel. Sarnaselt on Naheen (2015: 439) juhtinud tähelepanu rahapesule läbi investeeringute, kus raha paigutatakse välismaa pangakontodele ning seejärel saadetakse läbi erinevate investeerimisfirmade tagasi kodumaale. Sarnaselt on Zeldin (1998: 298) toonud välja, et avamere pankade kasutamise peamine eesmärk on maksudest eemale hiilimine ja uimastikaubanduse ning pettuste

varjamine. Sellest tulenevalt on rahvusvahelisi rahaülekandeid pakkuvad ettevõtted eriti haavatavad.

Kokkuvõtvalt on rahapesu ja terrorismi finantseerimise kolm peamist etappi koos levinuimate meetoditega esitatud järgneval joonisel (vt joonis 1).



Joonis 1. Rahapesu ja terrorismi finantseerimise põhiskeem ning levinuimad meetodid (autori koostatud).

Seega on rahapesu ja ülemaailmse terrorismi finantseerimise puhul tegemist probleemiga, millel on ulatuslikud tagajärjed nii inimeste igapäevaelus, ühiskonna turvalisuses kui ka majanduse stabiilsuses. Tulenevalt sellest, et rahapesu puhul on tegemist dünaamilise protsessiga, läbi mille ka kurjategijad õpivad, on võimalike meetodite nimekiri loendamatu, ent üldiselt jagatavad kolme etappi – paigutamine, kihitamine ning integratsioon. Rahapesu kui globaalse probleemi puhul on seega väga oluline implementeerida sobivaid hoolsusmeetmeid rahapesu efektiivseks tuvastamiseks, eriti virtuaalkeskkondades.

1.2 Ülevaade pettusi tuvastada võimaldavatest masinõppe meetoditest

Käesolevas alapeatükis antakse varasema teaduskirjanduse põhjal ülevaade erinevatest pettusi tuvastada võimaldavatest masinõppe meetoditest. Selleks, et statistilised meetodid oleksid efektiivsed, peavad nad olema dünaamilised ning rakendatud piisavalt suurel andmestikul. Pettuste tuvastamise puhul peetakse kõige aktuaalsemaks probleemiks tasakaalustamata andmestikke, mille puhul on ühes klassis oluliselt vähem vaatlusi (vähemusklass), kui mõnes teises klassis (Napierała, Stefanowski 2015: 9468). Rahapesu

tuvastamisel kuuluvad vähemusklassi rahapesukahtlusega kliendid ning enamusklassi heatahtlikud kliendid. Garcia ja He (2009: 1264) on toonud välja, et peamiste masinõppe algoritmide puhul on enamusklassi klassifikaatorid ligi 100 % täpsusega, seevastu vähemusklassi puhul vaid 0–10 % täpsusega. Sarnaselt on Chen *et al.* (2004: 1) toonud välja, et peamised klassifikatsioonialgoritmid ei tööta seetõttu, et nende eesmärk on minimeerida üleüldist veaprotsenti, mitte pöörata konkreetselt tähelepanu vähemusklassile. Näiteks kui mudel klassifitseerib 100 petturlikust kliendist õigesti vaid ühe, kuid 10 000 heatahtlikust kliendist õigesti 9900, siis on üleüldine veaprotsent väga väike, seevastu vähemusklassi puhul aga väga suur.

Chen *et al.* (2004: 1) on toonud välja, et tasakaalustamata andmetega töötamisel on levinuimateks meetoditeks kulutundlik õppimine (*cost sensitive learning*) ja näidistamine (*sampling*). Kulutundliku õppimise puhul määratakse igale vähemusklassis tehtud klassifikatsiooniveale kõrge maksumus (*cost*) ning mudeli eesmärk on minimeerida kogukulu. Sheng *et al.* (2014: 151) tõid välja, et kulupõhine õpe võib kergelt viia mudeli ületreenimiseni, mille puhul klassifikatsioon toimib hästi ainult nende andmete peal, millega mudel treeniti. Seega suudaks mudel väga täpselt kirjeldada olemasolevaid rahapesu juhtumeid, aga ei suudaks tuvastada peaaegu ühtegi uut. Tulenevalt sellest ning antud töö mahulisest piirangust võtab autor vaatluse alla vaid erinevad ümbernäidistamise meetodid. Ümbernäidistamise meetodid võib nende eesmärgi alusel jagada kaheks:

- ülenäidistamine – vähemusklassi tekitatakse uusi vaatlusi, tasakaalustades sellega andmeid;
- alanäidistamine – andmetest võetakse välja need vaatlused, mis kattuvad (*overlap*), ent kuuluvad erinevatesse klassidesse.

Sarnaselt kulutundliku õppega võib ka ülenäidistamine viia ületreenimiseni, kuna korratakse vähemusklassi vaatluseid. Vanitha ja Niraimathi (2013: 3643) on oma töös toonud välja, et esineb ka keerulisemaid meetodeid, näiteks sünteetiline vähemusklassi ülenäidistamine, mille puhul on ületreenimise oht väiksem. Sarnaselt tõid Batista *et al.* (2004:24) välja, et antud meetod aitab vältida ületreenimise probleemi ja laiendab otsustuspiire, viies vähemusklassi vaatlusi lähemale enamusklassi ruumile.

SMOTE (*synthetic minority oversampling technique*) ehk sünteetiline vähemusklassi ülenäidistamine on meetod, mille puhul lisatakse vähemusklassi uusi sünteetilisi vaatlusi interpoleerides lähimaid vaatlusi juhuslikkuse alusel. Chawla *et al.* (2002: 328) poolt esmakordselt välja pakutud meetodi puhul luuakse tunnusteruumis uusi sünteetilisi vaatlusi järgnevalt: esmalt leitakse konkreetsele vähemusklassi vaatlusele lähim sarnane vaatlus kasutades k–lähima naabri algoritmi. Seejärel leitakse lähima vaatluse ja originaalvaatluse vahe ja korrutatakse see läbi arvuga vahemikus 0–1 ning liidetakse algsele vaatlusele, mille põhjal luuakse uus sünteetiline vähemusklassi vaatlus.

Selleks, et mudelil oleks kergem eristada kahte klassi, pakuti 1997. aastal välja alanäidistamise meetod nimega Tomeki sidemed (*Tomek links*), mille puhul multidimensionaalses ruumis eemaldatakse kahe erineva klassi tunnuste piiril olevaid enamusklassi vaatluseid (Gu *et al.* 2008: 1021). Juhul kui kaks vaatlust moodustavad Tomeki sideme, peab neist ühe puhul olema tegu müra või erindiga. Selliseid harva esinevaid väärtusi ei ole mudeli treenimisel hea kasutada, sest need võivad tulemusi moonutada.

Tomeki sidemete olemasolu saab kontrollida järgnevalt (Kotsiantis *et al.* 2006: 3):

- kahe erineva klassi vaatluse a ja b puhul leitakse nende Eukleidese kaugus (*Euclidean distance*) $\delta(a,b)$,
- vaatlused a ja b moodustavad Tomeki sideme siis, kui ei ole ühtegi vaatlust c, mille puhul $\delta(a,c) < \delta(a,b)$ või $\delta(b,c) < \delta(a,c)$.

Seega võib öelda, et Tomeki side on neil vaatlustel, kus a ja b on üksteise lähimad naabrid ning a ja b on erineva klassi vaatlused.

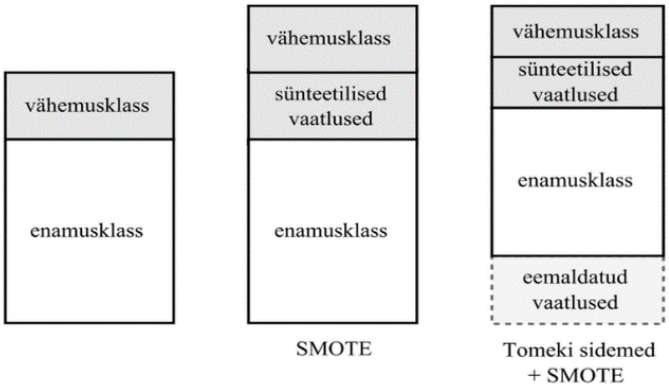
Estabrooks *et al.* (2004: 33) soovitasid kombineerida erinevaid ümbernäidistamise meetodeid selleks, et viia vähemusklassi andmete jaotus optimaalsele tasemele. Samuti tõid Batista *et al.* (2004: 24) välja, et kasutades ainult ülenäidistamise meetodeid, ei suudeta elimineerida kõiki klassifikatsioonivigu põhjustavaid faktoreid. Selleks, et andmeid korrastada ja vältida ületreenimist, tõid autorid (*Ibid.*) välja Tomeki sidemete ja SMOTE kombinatsiooni, mille puhul tuleks rakendada Tomeki sidemeid mitte ainult enamusklassis, vaid kogu andmestikul. Antud meetodite kombinatsiooni kasutatakse ka käesolevas töös andmete tasakaalustamiseks.

Mudeli soorituse hindamisel on üheks levinuimaks meetodiks konfusioonimaatriks, kuhu kuuluvad kõik vaatlused mudeli poolt vastu võetud otsuse (ennustus) ja tegelikkuse lõikes. Eristatakse nii valenegatiivseid kui ka valepositiivseid otsuseid – esimese puhul tehakse statistiliselt I liiki viga, mille kohaselt mudel ennustab, et klient ei ole rahapesukahtlusega, ent tegelikult on. Valepositiivsete otsuste puhul tehakse II liiki viga, mille puhul mudel ütleb, et tegu on rahapesu kahtlusega kliendiga, ent tegelikult on tegu heatahtliku kliendiga. Batista *et al.* (2003: 8) töös tagas SMOTE ja Tomeki sidemete kombineerimine võrreldes teiste meetoditega kõige väiksema valenegatiivsete otsuste, ent suurima valepositiivsete otsuste osakaalu. Mudeli rakendamisel tuleb lõppotsus võtta vastu siiski inimesel, seega on antud töö raames eriti oluline minimeerida valenegatiivsete otsuste osakaalu, fikseerides valepositiivsed otsused mingil vastuvõetaval tasemel, et ükski pahatahtlik klient ei jääks tähelepanuta. Kokkuvõtvalt on konfusioonimaatriksi põhikuju on esitatud järgmisel joonisel (vt joonis 2).

		Tegelikkus: Jah	Tegelikkus: Ei
Ennustus	Jah	Õigepositiivne otsus	Valepositiivne otsus
	Ei	Valenegatiivne otsus	Õigenegatiivne otsus

Joonis 2. Konfusioonimaatriks (autori koostatud, He & Ma 2013: 61 põhjal).

Tuginedes varasemale empiirilisele kirjandusele ja oskusteabele toob autor järgneval joonisel (vt joonis 3) välja meetodid ja nende kombinatsiooni ning tööpõhimõtte, mida käesoleva töö raames andmete tasakaalustamiseks kasutatakse.



Joonis 3. Üle- ja alanäidistusmeetodite võrdlus (autori koostatud).

Varasemas kirjanduses on pettuste tuvastamisel viidatud nii juhendamisega (*supervised*) kui ka juhendamiseta (*unsupervised*) õppele. Juhendamiseta õpet kasutatakse siis, kui puudub informatsioon varasemate pettuste kohta. Peamiselt otsitakse erinevate meetodite abil sarnaste tunnustega vaatlusi, mis kõik koondatakse ühte klastrisse. Zhang ja Zhou (2004: 514) on defineerinud, et klasterdamise põhimõte on maksimeerida klastrisisest ja minimeerida klastrite vahelist sarnasust. Mudeli valik sõltub eelkõige olemasolevate andmete tüübist ja eripärast. Tulenevalt suurest andmestikust ei võeta antud töös kasutusse klasterdamismeetodeid ja juhendamiseta õpet – kõikide klastrite läbivaatamine suurendab agentide igapäevast tööd. Peale selle on olemas informatsioon ka varasemate pettuste kohta ning juhendamiseta õppe valimisel jääks väga oluline osa väärtuslikku informatsiooni kasutamata.

Finantspettuste tuvastamisel on enim levinud juhendamisega õppe mudelid – meetodid, mis kuuluvad statistilise õppimise valdkonda. Statistilisel õppimisel on kaks peamist eesmärki: püstitada mudel, mis suudab kirjeldavate tunnuste põhjal ennustada sõltuva muutuja väärtust või anda ülevaade, kuidas kirjeldavad tunnused mõjutavad sõltuvat muutujat (Altmann *et al.* 2010: 1340). Juhendamisega õpe eeldab, et mudelil on teada lõppotsus (*label*) iga konkreetse vaatluse kohta, mille pealt õppida. Tsang *et al.* (2016: 3030) on toonud välja, et juhendamisega õpe on meetod, mis suudab hõlpsalt luua klassifikaatoreid, võimaldamaks tuvastada erinevat tüüpi pettuseid. Antud meetodi puhul kasutatakse andmestikku, kuhu on koondatud kõik varasemad pettused ja neid iseloomustavad tunnused ning selle andmestiku vastu hakatakse võrdlema igapäevaseid tehinguid ja kliente. Tsang *et al.* (2016: 3028) kasutasid oma töös pettuste tuvastamiseks edukalt juhendamisega õpet koos sünteetiliste andmetega. Tulenevalt sellest, et antud töö raames kasutatakse samuti osaliselt sünteetilisi andmeid, on juhendamisega õpe asjakohane. Juhendamisega õppe puhul esineb nii klassifikatsiooni– kui ka regressioonimudeleid.

Regressioon on statistiline meetod, mida kasutatakse uurimaks seost ühe või mitme selgitava muutuja ja sõltuva muutuja vahel (Han *et al.* 2012: 19). Antud meetod täidab statistilise õppimise mõlemat eesmärki – võimaldab ennustada sõltuva tunnuse väärtuseid ning interpreteerida otsust mõjutavaid tegureid. Ngai *et al.* (2011: 562) kohaselt on regressioon väga levinud meetod kindlustus- ja ettevõttepettuste puhul. Regressiooni- ja

klassifikatsioonimeetodite erinevus seisneb sõltuvas muutujas – kui sõltuv muutuja on pidev, on tegu regressiooniga ning kui diskreetne, on tegu klassifikatsiooniga. Antud uurimisprobleemi puhul on tegu binaarse klassifikatsiooniga, mille eesmärk on eristada häid ja halbu kliente, seega ei ole regressioonimeetodid asjakohased.

Ngai *et al.* (2011: 562) tõid välja, et peamised klassifikatsioonimeetodid on närvivõrgud (*neural network*), Bayes'i võrgud, otsustuspuud ja tugivektor masinad (*support vector machines*). Finantsvaldkonnas on levinuimate meetodite seast toodud esile närvivõrgud ja reeglipõhised meetodid (Zhang, Zhou 2004: 514). Järgnevalt annab autor lühidalt ülevaate erinevate meetodite sobivusest.

Närvivõrkude puhul on Mar ja Naing (2008: 154) kohaselt väga raske määrata optimaalset arhitektuuri, sest ei ole ilmne, mitu sisend (*input*) või peidetud (*hidden*) sõlme (*node*) mudelis kasutatakse. Tuginedes Zhang ja Zhou (2004: 515) tööle on närvivõrkude poolt leitud tulemusi väga raske tõlgendada. Juhul, kui mudel teeb klassifitseerimisel vigu, on parandusi sisse viia väga raske, sest pole teada, mis antud otsust mõjutas. See omakorda aeglustab optimaalse mudeli püstitamist. Teiseks peab rahapesu tuvastada võimaldav mudel olema interpreteeritav selleks, et nii rahapesu tuvastamisega tegelevatel agentidel kui ka audiitoritel oleks selge arusaam, mille alusel ettevõtte masinõppe mudel kahtlaseid vaatlusi leiab. Seega ei sobi antud meetod käesoleva uurimisprobleemi lahendamiseks ja jääb käesoleva töö raames vaatluse alt välja.

Otsustuspuude (*decision trees*) puhul on vaatlused esitatud tunnusvektorina (*feature vector*), mis koosnevad tunnuse ja väärtuse paarist. Antud meetodi puhul moodustatakse klassifikaatoreid „kui" (kindlatel tingimustel) ja „siis" (sellest tulenev) reeglite põhjal. Otsustuspuud töötavad ülevalt–alla meetodi kohaselt: igal tasemel otsitakse tunnuse jaoks teda kirjeldavaid omadusi, mis eristaks klasse kõige täpsemalt ning seejärel töödeldakse tekkinud jaotusi rekursiivselt (Zhang, Zhou 2004: 515). Otsustuspuud jagavad keerulise ülesande mitmeks väikseks alamülesandeks ja nende eesmärk on luua võimalikult väike aga maksimaalselt täpne puu (Shen *et al.* 2007: 2). Otsustuspuudel on mitmeid eeliseid – nad on paindlikud ning neid on kerge implementeerida klassifikatsioonireegliteks, sest tunnusvektori põhjal on leitav selline tunnuste kombinatsioon, mis viitab rahapesukahtlusele. Levinuimad otsustuspuu meetodid on ID3 ja C4.5, mis on edasi arendatud kategooria õppimisest (*category learning*) ning suudavad käsitleda pidevaid

andmeid (Shen *et al.* 2007: 2). Tuginedes otsustuspuude tööpõhimõttele võib ühe miinusena tuua välja mudeli staatilisuse – puu püstitatakse kogu andmestikule, võttes arvesse kõiki tunnuseid ning seega ei pruugi olla skaleeruv dünaamilistel reaalandmetel. Teine oluline probleem otsustuspuude puhul on ületreenimine, mille ennetamiseks tuleb kasutada pügamist (*pruning*), mis omakorda raskendab kvaliteetse mudeli püstitamist.

Juhumets (*random forest*) on 2001. aastal Breimani (2001: 29) poolt välja pakutud meetod, kus püstitatakse mitu otsustuspuud juhuslikult valitud tunnuste põhjal. Tavalistes otsustuspuudes jagatakse iga sõlm (*node*) kasutades parimat jaotust kõikide tunnuste ja andmete lõikes. Juhumetsa puhul treenitakse iga otsustuspuu juhuslikult valitud andmete alamhulga ja tunnuste põhjal ning seejärel klassifitseeritakse vaatluseid klassidesse selle alusel, kui suur osa otsustuspuudest sama näitavad. Bhattacharyya *et al.* (2011: 605) tõid välja, et vaid ühest puust koosnevad otsustuspuu mudelid võivad olla väga ebastabiilsed ja liiga tundlikud andmete eripärast tingituna. Sarnaselt tõid Khoshgoftaar *et al.* (2007: 310) tõid välja, et mida rohkem otsustuspuid püstitatakse, seda enam väheneb oht mudelit üle treenida. Seega sobib see eriti hästi kombineerituna erinevate ala– ning ülenäidistamismeetoditega. Breimani (2001) kohaselt töötavad juhumetsad efektiivselt suurte andmestikega, suudavad kasutada tuhandeid tunnuseid, annavad tunnustele tõenäosushinnangu ning nende sooritus on tihti parem kui tugivektor masinatel ja närvivõrkudel. Lopez-Rojas ja Axelssoni töös (2012: 7) tagas juhumetsa kasutamine erinevalt teistest reeglitepõhistest meetoditest kõige parema tulemuse konkreetselt rahapesukahtlusega juhtumite klassifitseerimisel.

Tuginedes varasemale empiirilisele kirjandusele ja oskusteabele on autor toonud välja järgnevad meetodite kombinatsioonid, mida käesolevas töös rakendatakse:

- juhumets,
- SMOTE + juhumets,
- SMOTE + Tomeki sidemed + juhumets.

Seega võetakse antud bakalaureusetöös vaatluse alla nii ala- kui ka ülenäidistamismeetodeid ja nende kombinatsioon juhendamisega õppega selleks, et õppida ekstremaalselt tasakaalust väljas andmete põhjal.

1.3 Ülevaade rahapesu tuvastada võimaldavatest muutujatest

Lähtudes vaid alapeatükis 1.1 välja toodud rahapesu meetodite põhimõttest, ei ole võimalik rahapesu efektiivselt tuvastada – ka nii-öelda heatahtlikud kliendid kasutavad sarnaseid meetodeid, mida kasutavad rahapesijad – näiteks investeerivad kinnisvarasse või luksuskaupadesse. Põhjendamatu oleks peatada manuaalseks ülevaatamiseks kõik välismaale tehtud investeeringud vaid seetõttu, et sarnast meetodit võivad ka rahapesijad kasutada – see tekitaks väga palju valealarmi ning selle tagajärjel kannataks oluliselt klientide rahulolu. Sellest tulenevalt on käesoleva alapeatüki eesmärk tuua varasema kirjanduse põhjal välja täiendavaid tunnuseid, mis võivad viidata rahapesule. Kusjuures tasub nentida, et rahapesu tuvastada võimaldavate muutujate kohta on varasem kirjandus piiratud, sest ka rahapesijad võivad sellest õppida.

Rahapesule viitavad tunnused võib jagada profiili-, makse- ning saajapõhisteks. Tuginedes Irwin *et al.* (2011: 94) tööle, on struktureerimine ja smurfimine ühed enimlevinud rahapesu ja terrorismi finantseerimise tehnikad ning tulenevalt sellest kasutab autor ühe maksemustrit kirjeldava tunnusena sama suurte maksete arvu kliendipõhiselt. Struktureerimise, smurfimise ning maksete võrgustike tuvastamiseks on Drezewski *et al.* (2015: 18) toonud välja sotsiaalvõrgustiku analüüsi (*social network analysis*), mis võimaldab tuvastada keerulisi struktuure ja tuvastada võrgustikesse kuuluvaid inimesi ja nende vahelisi seoseid. Samad autorid (2015: 31) kasutasid efektiivselt võrgustike loomist rahapesu tuvastamiseks, ent soovitasid seda kombineerida masinõppega maksimaalse efektiivsuse tagamiseks. Käesoleva töö raames moodustatakse võrgustikke maksete põhjal ja nende kirjeldamiseks kasutatakse võrgustiku suurust ja võrgustikust läbi käinud rahasummat. Lisaks sellele arvutatakse välja, kui palju raha on saadetud igale võrgustikku kuuluvale inimesele ning kui suure summa sellest on inimene edasi saatnud. See aitab tuvastada võimalikku struktureerimist juhul, kui mitu inimest saadavad läbi kolmandate isikute kõik samale lõppkontole.

Ülemaailmselt on rahapesu tõkestamises väga suur roll 1989. aastal loodud valitsustevahelisel Rahapesu Toimkonnal (*Financial Action Task Force* ehk FATF*),* mis sätestab rahvusvahelised rahapesu tõkestamise standardid, mida riigid peavad jälgima, et mitte sattuda musta nimekirja (*blacklist*), millega kaasnevad tõsised negatiivsed majanduslikud tagajärjed (Unger, Hertog 2012: 287). Tulenevalt FATF soovitustest

(FATF... 2012: 14) peavad finantsinstitutsioonid rakendama hoolsusmeetmeid ja küsima teatud lävendini jõudes klientidelt lisainfot nende maksepõhjuse ja identiteedi kohta. Maksemustri järsk muutumine võib tähendada, et rahapesijad arvavad, et on lävendi ületanud ning rohkem lisainfot neilt ei küsita ja seetõttu võib rahapesu protsessi kiirendada suuremate summade jaotamisega. Sellest tulenevalt on Le Khac ja Kechadi (2010: 578) toonud välja, et kahtlaste juhtumite tuvastamiseks on oluline jälgida kliente, kelle maksemuster on ebajärjepidev. Samad autorid (*Ibid.*) pakkusid välja kaks järgnevat tõeväärtusega reeglit:

- kas käesoleva perioodi (nädal, kuu, kvartal) maksete summa ületab varasemate perioodide keskmist 2 standardhälbe võrra või on käesoleva perioodi maksete summa 200% keskmisest,
- kas väljaminevate maksete summa on 90–110% sissetulekust või kas kogu sissetulek on rohkem kui 50% eelmisest kontojäägist.

Tulenevalt sellest, et rahapesu on dünaamiline protsess, mille konkreetsed meetodid muutuvad ajas, leiab käesoleva töö autor, et jäikade piiride kehtestamine ei ole asjakohane ja tuleks kasutada hoopis pidevat tunnust, mis kirjeldab käesoleva perioodi maksete summa erinevust keskmisest. Paljude virtuaalkeskkonnas rahvusvahelisi rahaülekandeid osutavate ettevõtete puhul ei eksisteeri nii-öelda e-rahakoti teenust, sest seadustest tulenevalt ei ole finantsteenuseid osutavatel ettevõtetel (*Money Service Business*) õigust kliendile ettevõtte juures kontot avada. Le Khac ja Kechadi (2010: 578) poolt pakutud teise reegli rakendamiseks mudelis tuleks seega jälgida hoopis seda, kui palju raha on saadetud läbi ettevõtte konkreetsele saajale (kontole) ning kui suur osa maksetest on finantseeritud selle sama konto poolt.

Ühe võimaliku rahapesu vahendina on välja toodud ettemakstud (*prepaid*) kaardid (Irwin *et al.* 2014: 58). Visa ja MasterCardi kinkekaardid annavad võimaluse jääda anonüümseks ning seega tuleb olulist tähelepanu pöörata ka sellele, mis vahendiga tehakse finantsteenuseid osutavate ettevõtete juures deposiidid. Näiteks 2015. aasta 13. novembri terrorirünnakute puhul Pariisis olid ettemakstud kaardid üheks viisiks, kuidas kurjategijad oma ettevalmistusi finantseerisid (Mathers, 2016). Antud juhul sobib mudelis kasutatavaks tunnuseks tõeväärtus (*boolean*) – kas raha tuli ettemakstud kaardilt või mitte.

FATF poolt kehtestatud soovitustes on toodud välja, et finantsinstitutsioonid peaksid täiendavaid hoolsusmeetmeid (*enhanced due diligence*) rakendama kõrge rahapesuriskiga riikide puhul (FATF... 2012: 19). Sellest tulenevalt on käesoleva töö raames autor jaganud riigid nende riski alusel kolme kategooriasse – madala, keskmise ja kõrge rahapesuriskiga riigid. Rahapesu tuvastada võimaldavas mudelis saab antud tunnust kasutada nii kliendi profiilis oleva aadress, sisselogimise aadressi kui ka saaja riigi (*recipient country*) lõikes.

Irwin *et al.* (2014: 58) kohaselt on üheks rahapesu indikaatoriks ebajärjekindel käitumine äriprofiilidel. Suurem osa andmekaeve meetoditest jätavad müra filtreerimiseks välja erindid, ent Han *et al*. (2012: 20–21) tõid välja, et pettuste tuvastamise puhul võib olla oluline rakendada erindite tuvastamist ja jälgida just neid nähtusi, mis erinevad keskmisest. Tulenevalt sellest on oluline virtuaalkeskkondades profileerida ärikliente sektoripõhiselt ja pöörata tähelepanu neile, kelle maksesagedus või summad erinevad antud sektori keskmisest oluliselt. Wang ja Yang (2007: 285) kohaselt kohaselt on oluline rahapesu riski hindamisel eristada ettevõtteid täiendavalt nende suuruse ning tootmisharu põhiselt. Samade autorite (*Ibid.*) riskihinnang, mis põhineb otsustuspuu meetodil, on esitatud järgnevalt:

- börsifirmad ja suurettevõtted – madal risk,
- keskmise suurusega ettevõtted – keskmine risk,
- väikeettevõtted – kõrge risk.

Seega on oluline finantsteenuseid osutavatel ettevõtetel koguda täiendavat informatsiooni äriklientide kohta ja jälgida maksemustri seost ettevõtte suurusega. Autoripoolse edasiarendusena riiulifirmade teooriast tuleks täiendavat tähelepanu pöörata erinevaid teenuseid osutavatele ettevõtetele, sest raske on määrata, kas teenus tegelikult ka eksisteeris.

Moustafa *et al.* (2015: 315) on oma töös toonud välja magavate kontode mõiste, mille kohaselt terrorismi finantseerijate üheks tehnikaks võib olla ebaaktiivsete pangakontode kasutamine lühikese aja vältel. Seega tasuks tähelepanu pöörata neile kontodele, mis on olnud pikalt ebaaktiivsed, ent siis teinud lühikese aja jooksul mitu makset. Autoripoolse

edasiarendusena tasuks vaadata ka neid kontosid, mis on tehtud samas ajavahemikus, olnud pikka aega ebaaktiivsed ja siis teinud mõne makse ühele ja samale pangakontole.

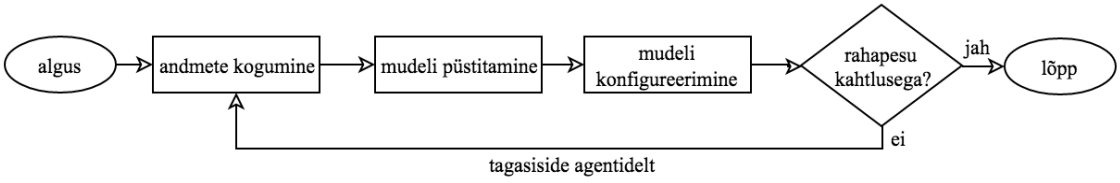
Tuginedes FATF soovitustele (2012: 17) peavad ettevõtted enne uute toodete või teenuste käiku laskmist hindama nendega kaasnevaid riske. Tulenevalt uute toodete ja teenustega kaasnevatest riskidest leiab autor, et pideva tunnusena sobiks mudelisse kaasata ka nende vanust kirjeldav tunnus. Kusjuures toodete all peetakse eelkõige silmas uusi valuutasid ning teenuste all näiteks raha küsimise võimalust (*request money*), seega tasub hinnata, kui kaua mingi konkreetse valuuta makseid on toetatud ning kui kaua mingi teenus turul on olnud.

2. RAHAPESU TUVASTADA VÕIMALDAVA MASINÕPPE MUDELI LOOMINE

2.1 Uurimismetoodika ning lähteandmete kirjeldus

Käesolevas bakalaureusetöös kasutatakse TransferWise LTD andmeid mis on eeltöödeldud sõltuvalt uurimiseesmärgist. Mudeli püstitamisel kasutatakse treeningandmetena ettevõttesiseseid rahapesukahtluse raporteid ning rakendatakse erinevaid andmekaeve meetodeid. Andmete kogumisel piiranguid ei ole – kasutatakse MySQL päringuid, mille läbi on autoril ligipääs ettevõtte olemasolevatele andmetele. Tulenevalt sellest, et tegu on konfidentsiaalsete isikuandmetega, ei ole nende sisu ning arvutusvalemeid käesoleva töö raames võimalik avaldada. Mudeli treenimisel kasutatavate sõltuvate tunnuste puhul piirdub autor vaid tunnuse tüübi ja üldise kirjelduse andmisega. Mudeli treenimisel ja andmete töötlemisel kasutatakse statistilise arvutamise vabavara R.

Zhang ja Zhou (2004: 515) on toonud välja, et andmekaeve puhul on erinevad meetodid väga tundlikud muutuste suhtes andmetes ning sellest tulenevalt toimub käesoleva töö raames mudeli treenimine kolmes tsüklis, kus iga tsükliga kogutakse juurde täiendavaid andmeid rahapesu tõkestamiseks ning treenitakse mudel minimeerides viga ennustuse ja tegelikkuse vahel. Raamistik rahapesu tuvastada võimaldava masinõppe mudeli loomiseks on esitatud joonisel (vt joonis 3). Kusjuures iteratsioonide arvul piiranguid ei ole – mudelit tuleks treenida nii mitme tsükli vältel, kuni on saavutatud soovitud tulemus.



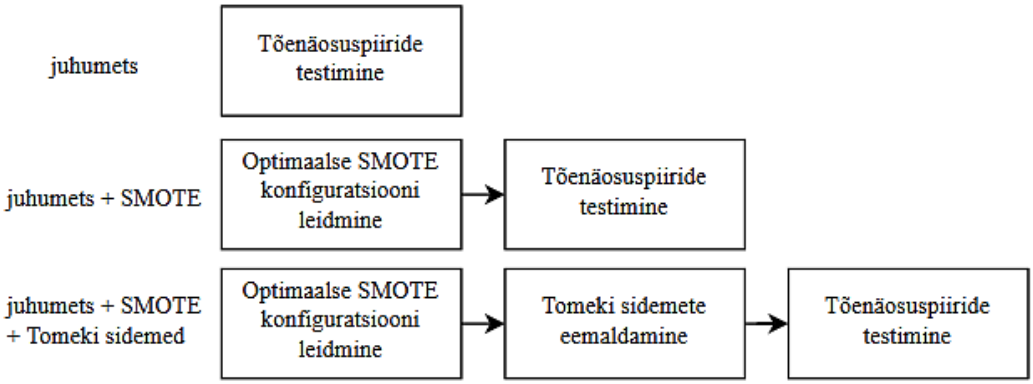
Joonis 3. Parima mudeli püstitamise metoodika (autori koostatud).

Andmete eeltöötlus ja kogumine on oluline etapp andmekaeve puhul, millesse kuuluvad peamiselt neli tegevust: andmete puhastamine, integratsioon, transformatsioon ning vähendamine (Catal *et al.* 2011: 4867). Andmete eeltöötluse all peab käesoleva töö autor eelkõige silmas andmete puhastamist empiirilise vaatluse põhjal – modelleeritavate juhtumite manuaalne ülevaatamine ja nende sobivuse hindamine. Juhtumite sobivuse hindamiseks kasutab autor rahapesu tõkestamisega tegelevate agentide eksperthinnangut sellele, kas ja kui kahtlase juhtumiga tegu oli. Andmete kogumine on protsess, mille käigus lisatakse agentide tagasiside põhjal andmestikku täiendavaid tunnuseid, mille põhjal kahtlus ümber lükati. Mudeli treenimine on esitatud tsüklitena selleks, et korrigeerida ning eeltöödelda andmeid vastavalt tagasisidele ja mudeli sooritusele. Rakendatakse nii andmete puhastust kui ka vähendamist ning tsüklite tulemusena üritatakse luua selline andmestik, kuhu kuuluvad vaid väga hästi rahapesu kirjeldavad juhtumid ning kõik väline müra on jäetud vaatluse alt välja.

Mudeli püstitamise etapis treenitakse mudel erinevatel treeningandmetel, mille tarbeks rakendatakse selles faasis nii üle- kui ka alanäidistamismeetodeid. Kasutades sünteetilist vähemusklassi ülenäidistamismeetodit leitakse esmalt selline vähemusklassi vaatluste genereerimise koefitisient, andmestiku suurus ja tasakaal, mis tagab kõrgeima F1–skoori ning seejärel treenitakse neil andmetel mudel. Kombineerides nii üle- kui ka alanäidistamismeetodeid luuakse esmalt uusi, sünteetilisi vaatlusi ning seejärel eemaldatakse andmestikust Tomeki sidemed ja treenitakse selle põhjal mudel. Seega püstitatakse igas tsüklis kolm mudelit, mis erinevad treeningandmete statistilise töötlemise meetodite poolest.

Mudeli konfigureerimine tähendab mudelispetsiifiliste konfiguratsioonide katsetamist eesmärgiga leida sellised parameetrite väärtused, mis tagavad mudelile parima soorituse. Antud töö raames on mudelispetsiifiliste konfiguratsioonide all peetud silmas eelkõige tõenäosuspiiri (*cutoff*). Tõenäosuspiiriks nimetatakse sellist tõenäosusväärtust, millest alates mudel klassifitseerib vaatluse kahtlaseks (standardselt 0,5). On selge, et mida madalamale me tõenäosuspiiri seame, seda rohkem rahapesu kahtlusega kliente leitakse üles, ent seda rohkem tekitatakse ka müra valepositiivsete vaatluste näol. Seega lähtutakse stabiilsuse analüüsil F1–skoorist – juhul, kui mudel üle- või alahindab vaatlusi, kaasneb sellega tulemusmõõdiku mitmekordne muutus tõenäosuspiiride nihutamisel.

Seega kasutatakse tõenäosuspiiride testimist eelkõige mudeli stabiilsuse analüüsiks, mitte soorituse parandamiseks. Kokkuvõtvalt on mudeli püstitamise ja konfigureerimise faasis rakendatavad tegevused iga meetodi puhul esitatud järgnevalt joonisel (vt joonis 4).



Joonis 4. Iga meetodi puhul mudeli püstitamisel ja konfigureerimisel rakendatavad tegevused (autori koostatud).

Tagasiside etapis ei katsetata mudeli sooritust enam testandmetel vaid reaalsetel andmetel, ehk klientidel, keda rahapesu tõkestamisega tegelevad agendid veel ei ole üle vaadanud ning kelle kohta puudub otsus, kas tegu on kahtlase kliendiga või mitte. Antud etapis vaadatakse need kliendid spetsiaalselt agentide poolt üle, eesmärgiga hinnata, kas mudel leiab üles ka päriselus kahtlased kliendid või mitte. Juhul, kui mudeli jaoks kahtlased kliendid agentide jaoks ei ole kahtlased, kogutakse täiendavat infot, mille põhjal nad kahtluse ümber lükkasid ning lisatakse need andmete kogumise etapis tunnuste näol mudelisse.

Nagu alapunktis 1.1 välja toodud, on ettevõtte tasandil väga keeruline määrata, milline klient tegeles rahapesuga ning milline mitte. Sellest tulenevalt on mudeli treenimiseks kasutatud andmetesse koondatud need rahapesu kahtluse raportid ja neid kirjeldavad tunnused, mis vajasid agentide tähelepanu ja lähemat uurimist ning lisainfo küsimist. Enamikul juhtudel kulmineerus täiendav uurimine ka politseile välise raporti tegemisega. Mudeli püstitamisel on oluline treeningandmetesse koondada ka nii–öelda heatahtlike kliente, kelle puhul igasugune rahapesu kahtlus puudub selleks, et mudel oleks ka reaalandmetel võimeline eristama rahapesu kahtlusega kliente. Heatahtlikke kliente võib eristada teistest pika ja stabiilse makseajaloo ning olemasoleva informatsiooni rohkuse põhjal – mida rohkem on kliendi kohta andmeid, seda suurem on tõenäosus, et klient ei tegele rahapesuga.

Mudeli treenimise esimeses tsüklis on analüüsi kaasatud 18 tunnust, mis võib jagada kolme järgnevasse kategooriasse: profiili, makseid ning saajaid kirjeldavad tunnused. Profiili kirjeldavatest tunnustest on vaatluse alla võetud profiili tüüpi (äri- või erakonto) kirjeldav binaarne tunnus, kliendi vanus pideva muutujana ning kaks profiili aktiivse kasutamise algust kirjeldavat tunnust. Saajaid kirjeldavate tunnuste puhul on analüüsi lisatud nii saajate asukohta kui ka saajate arvu iseloomustavad karakteristikud. Kõige enam on makseid kirjeldavaid tunnuseid, mis peegeldavad nii maksete arvu (2 tunnust), summat (7 tunnust) kui ka kasutatud valuutasid (2 tunnust). Mudeli treenimisel esimeses tsüklis kasutatud tunnused koos selgitustega on kokkuvõtvalt esitatud järgnevas tabelis (vt tabel 1).

Tabel 1. Esimeses tsüklis kasutatud tunnused koos selgitustega.

Muutuja	Selgitus	Tunnuse grupp
x_0	Identifitseerimiskood, mida ei kasutata mudeli treenimisel	–
x_1	Binaarne profiili tüüpi kirjeldav tunnus – kas tegu on äri- või eraprofiiliga	Profiili kirjeldav
$x_{2\ldots3}$	Maksete arvu kirjeldavad tunnused	Makseid kirjeldav
$x_{4\ldots5}$	Saajate arvu kirjeldavad tunnused	Saajaid kirjeldav
$x_{6\ldots12}$	Maksete rahalist väärtust kirjeldavad tunnused	Makseid kirjeldav
$x_{13\ldots14}$	Kliendi poolt kasutatud valuutasid kirjeldavad tunnused	Makseid kirjeldav
x_{15}	Kliendi vanus	Profiili kirjeldav
$x_{16\ldots17}$	Profiili aktiivse kasutamise algust kirjeldavad tunnused	Profiili kirjeldav
x_{18}	Saajate asukohta kirjeldav tunnus	Saajaid kirjeldav

Allikas: autori koostatud.

Teises tsüklis lisati analüüsi kõige enam makseid kirjeldavaid tunnuseid – täiendati kasutatud valuutasid iseloomustavaid karakteristikuid kahe uue muutuja näol ning lisati saatja ja saaja omavahelist seost kirjeldavad tunnused, mis olid jagatud seitsmesse kategooriasse. Peale selle lisati makse tüüpi ehk finantseerimisviisi kirjeldav kategooriline tunnus, millel oli neli erinevat taset (*level*), kliendi poolt raha ülekandmise kiirust iseloomustav pidev tunnus, makse selgitusi kirjeldavad tunnused, mis olid jagatud 25 erinevasse kategooriasse ning maksete loomise platvormi ja sagedust kirjeldavad tunnused. Profiili iseloomustavatest karakteristikutest jagati kliendid 6 erinevasse vanuserühma ja lisati profiili loomise sessiooni ning liitumiskanalit kirjeldavad tunnused, mis olid jagatud 14 erinevasse kategooriasse. Kusjuures kategoorilised tunnused on

käesoleva töö raames esitatud läbi fiktiivsete muutujate. Kokkuvõtvalt on teises tsüklis analüüsi lisatud tunnused esitatud järgnevas tabelis (vt tabel 2).

Tabel 2. Teises tsüklis analüüsi lisatud tunnused koos selgitustega.

Muutuja	Selgitus	Tunnuse grupp
x_{19}	Kliendi vanuserühm	Profiili kirjeldav
$x_{20…22}$	Kliendi liitumiskanalit kirjeldavad tunnused	Profiili kirjeldav
x_{23}	Profiili loomise sessiooni kirjeldav tunnus	Profiili kirjeldav
$x_{24…25}$	Kliendi poolt kasutatud valuutasid täiendavalt kirjeldavad tunnused	Makseid kirjeldav
x_{26}	Makse saajate seost saatjaga kirjeldavad tunnused jagatuna seitsmesse kategooriasse	Makseid kirjeldav
x_{27}	Makse tüüpi kirjeldavad tunnused jagatuna nelja kategooriasse	Makseid kirjeldav
$x_{28…31}$	Kliendi poolt raha ülekandmise kiirust kirjeldavad tunnused	Makseid kirjeldav
x_{32}	Makse selgitusi kirjeldavad tunnused jagatuna 25 kategooriasse	Makseid kirjeldav
x_{33}	Makse loomise platvormi kirjeldavad tunnused jagatuna nelja kategooriasse	Makseid kirjeldav
x_{34}	Maksete sagedust kirjeldav arvuline tunnus	Makseid kirjeldav

Allikas: autori koostatud.

Viimases tsüklis lisati üks maksete sagedust kirjeldav tunnus, mis oli jagatud nelja kategooriasse ning kliendi asukohta kirjeldav tunnus, mis oli jagatud üheksasse kategooriasse. Lisaks sellele esitati varasemates tsüklites defineeritud maksete käitumuslikkust kirjeldavad tunnused osakaaludena. Kolmandas tsüklis analüüsi lisatud muutujad koos selgitustega on esitatud järgnevas tabelis (vt tabel 3).

Tabel 3. Viimases tsüklis analüüsi lisatud tunnused koos selgitustega.

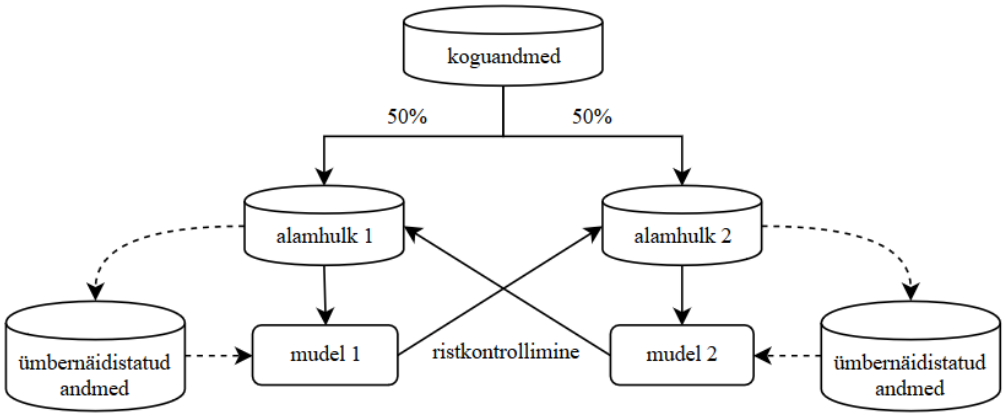
Muutuja	Selgitus	Tunnuse grupp
x_{35}	Maksete sagedust kirjeldavad tunnused jagatuna nelja kategooriasse	Makseid kirjeldav
x_{36}	Kliendi asukohta kirjeldavad tunnused jagatuna üheksasse kategooriasse	Profiili kirjeldav

Allikas: autori koostatud.

Parima mudeli püstitamine koosneb nii empiirilisest kui ka matemaatiliselt mõõdetavast tagasisidest. Masinõppe mudelite soorituse matemaatilisel hindamisel ei tohi kasutada samu andmeid, mille põhjal mudel treeniti – hinnates mudeli sooritust treeningandmetel, on tulemus alati väga hea, sest mudel on püstitatud samadele andmetele. Tulenevalt sellest tuleb andmed jagada kaheks: treening- ning testandmed. Mudeli soorituse

testimisel võrreldakse mudeli poolt klassifitseeritud vaatlusi tegelikkusega. Kuhn ja Johnson (2013: 77–78) kohaselt ei ole sobilik hinnata mudeli sooritust vaid ühel testandmete kogumil, sest andmete jagamisel kaheks võivad treening ja testandmed olla liiga heterogeensed ning sellest tingituna võivad tulemused olla kallutatud. Dietterich (1998: 1894) kohaselt on tüüpiline meetod klassifikatsioonialgoritmide soorituse hindamiseks k–kordne ristkontrollimine (*k-fold crossvalidation*), mille puhul jagatakse kogu andmestik ligikaudu võrdsetesse, ent mitte kattuvatesse komplektidesse, kusjuures komplektide arvu tähistab konstant k. Ristkontrollimise puhul treenitakse mudel alati k–1 komplektil ning tulemust hinnatakse komplekti põhjal, mis jäi treeningust välja. Nii tehakse läbi kõik võimalikud kombinatsioonid ning lõpuks võetakse tulemusmõõdikute keskmine. Paraku ei pruugi antud lähenemine olla efektiivne ekstremaalselt tasakaalust väljas andmetel – jagades koguandmestiku näiteks viieks komplektiks võib juhtuda, et igas osas ei ole piisavalt vähemusklassi vaatlusi, millelt õppida või mille põhjal mudelit testida. Dietterich (1998: 1905) soovitas kasutada 5x2 ristkontrollimist, mille kohaselt treenitakse mudel viie iteratsiooni vältel 2–kordse ristkontrollimisega. Sama meetodit kasutatakse ka käesolevas töös – andmed jagatakse viiel korral juhuslikkuse alusel kaheks ning seejärel treenitakse mõlemal andmete alamhulgal mudelid, mille tulemust testitakse teise alamhulga peal ja lõpuks leitakse kümne tulemusmõõdiku keskmine. Mudeli soorituse korrektseks hindamiseks rakendatakse igas tsüklis ümbernäidistamismeetodeid vaid treeningandmetel ning testandmed jäetakse puutumata.

Antud töös kasutatava ristkontrollimise põhimõte on kokkuvõtvalt esitatud joonisel (vt joonis 5), kusjuures sellist tsüklit läbitakse viiel korral, kus igal korral jagatakse andmed alamhulkadesse erinevalt.

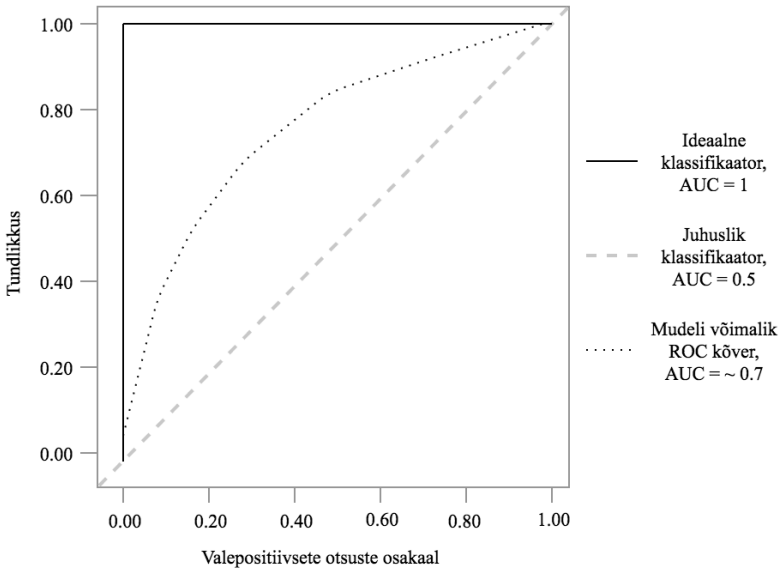


Joonis 5. Mudeli soorituse hindamine ristkontrollimise abil (autori koostatud).

Mudeli matemaatilist täpsust hindab töö autor konfusioonimaatriksi põhjal kolmel järgneval viisil: õigepositiivsete otsuste osakaal, õigesti tehtud otsuste osakaal ja F1-skoor. Õigepositiivsete otsuste osakaal ehk tundlikkus (*true positive rate, recall, sensitivity)* näitab vähemusklassi klassifikaatori täpsust – mitu protsenti kõikidest rahapesu kahtlusega klientidest mudel testandmetest üles leidis. Pettuste tuvastamisel on klassifitseerimisvigade tähtsused erinevad – kõige riskantsem ja kulukam on ennustada, et mõni rahapesu kahtlusega klient on heatahtlik ning jätta ta agentide tähelepanuta. Paraku ei saa mudelit hinnata vaid tundlikkuse järgi, sest isegi kui mudel suudab klassifitseerida täpselt vähemusklassi, võib ta tekitada agentidele palju lisatööd valepositiivsete otsuste näol. Valepositiivsete otsuste mõõtmiseks kasutab autor õigesti tehtud otsuste osakaalu ehk täpsust (*positive predictive value, precision),* mis näitab, kui suur osa mudeli jaoks kahtlastest klientidest olid ka päriselus kahtlased. Optimaalse mudeli puhul on täpsus 1 ehk mudeli jaoks kahtlastest klientidest olid ka päriselus kõik kahtlased. Gu *et al.* (2008: 1023) on toonud välja, et õigesti tehtud otsuste osakaal ega õigepositiivsete otsuste osakaal üksi ei ole sobivad mõõdikud mudeli täpsuse hindamiseks ja soovitatav on kasutada F1-skoori. Antud tulemusnäidiku puhul on tegemist õigesti tehtud otsuste osakaalu ja õigepositiivsete otsuste osakaalu harmoonilise keskmisega, mis võimaldab võrrelda erinevate mudelite sooritust võttes arvesse nii nende tundlikkust kui ka täpsust. F1–skoor on alati vahemikus [0;1] ning mida lähemal on see ühele, seda täpsema mudeliga on tegu.

Lisaks sellele visandatakse iga mudeli hindamiseks ROC (*receiver operating characteristic*) kõver ja leitakse selle alune pindala. Shatnawi *et al.* (2010: 7) on toonud välja, et ROC kõverad on andmekaeve puhul üheks laialdaselt kasutust leidnud tulemusnäidikuks. ROC kõvera puhul on x–teljel kujutatud valepositiivsete otsuste osakaal ning y–teljel tundlikkus. Mudeli hindamisel arvutatakse kõvera alla jääva osa pindala, ehk AUC (*area under curve*), mis on alati vahemikus [0;1] ning mida lähemal on see väärtus ühele, seda parem on mudeli ennustustäpsus. Fawcett (2003: 3) on toonud välja, et ROC analüüs on väga efektiivne meetod tasakaalust väljas andmetel klassifikatsiooni testimiseks, sest kujutab nii-öelda suhtelist kompromissi tulude (õiged otsused) ja kulude (valepositiivsed otsused) vahel.

Kokkuvõtvalt on ROC kõvera ning AUC abil mudeli täpsuse hindamise põhimõte esitatud järgneval joonisel (vt joonis 6).



Joonis 6. ROC kõvera ning AUC abil mudeli täpsuse hindamine (autori koostatud Carter *et al.* 2016: 1642 põhjal)

Tundlikkuse kasv toob alati kaasa ka valepositiivsete otsuste osakaalu kasvu ning sellest tulenevalt võib öelda, et ideaalse mudeli puhul, mis ei tee klassifitseerimisel ühtegi viga, on ROC kõvera alune pindala 1 ning juhuslikkuse alusel klassifitseerival mudelil 0,5.

Eespool mainitud tulemusmõõdikuid kasutatakse iga tsükli lõikes parima mudeli leidmiseks ning viimaks kolme tsükli parima mudeli leidmiseks. Kokkuvõtvalt on mudeli täpsuse hindamiseks kasutatavad meetrikad koos selgitustega esitatud järgnevas tabelis (vt tabel 4).

Tabel 4. Mudeli täpsuse hindamiseks kasutatavad meetrikad.

Tulemusmõõdik	**Selgitus**
Õigepositiivsete otsuste osakaal, tundlikkus	$\frac{\text{õigepositiivsed otsused}}{\text{õigepositiivsed} + \text{valenegatiivsed otsused}}$
Õigesti tehtud otsuste osakaal, täpsus	$\frac{\text{õigepositiivsed otsused}}{\text{õigepositiivsed} + \text{valepositiivsed otsused}}$
F1–skoor	$2\frac{\text{õigesti tehtud otsuste osakaal} * \text{õigepositiivsete otsuste osakaal}}{\text{õigesti tehtud otsuste osakaal} + \text{õigepositiivsete otsuste osakaal}}$
AUC	ROC kõvera alune pindala

Allikas: autori koostatud.

2.2 Andmete statistiline eeltöötlus ja masinõppe mudeli püstitamine

Käesolevas alapeatükis treenib autor masinõppe mudelit kolmes tsüklis eespool mainitud põhimõtte kohaselt, kus igas tsüklis katsetatakse nii juhumetsa kui ka üle– ja alanäidistamise meetodeid. Kombineerides nii empiirilist kui ka matemaatilist tagasisidet toob autor välja sellise meetodite kombinatsiooni koos vastavate konfiguratsioonidega, mis on kõige sobivamad rahapesu kahtlusega klientide leidmiseks. Lisaks toob autor tagasiside põhjal välja ka selle, milliseid tulemusmõõdikuid optimaalse mudeli püstitamisel tuleks kasutada.

Esimeses tsüklis kasutati mudeli treenimisel 170 varasemat rahapesu kahtlusega ning 14008 rahapesu kahtluseta juhtumit. Antud tsükli puhul moodustas vähemusklass ~1% kogu andmetest, seega andmed olid ekstremaalselt tasakaalust väljas. Olgugi, et varasemas empiirilises kirjanduses soovitatakse püstitada mudel enam–vähem tasakaalus andmetel, ei ole see antud uurimisprobleemi puhul põhjendatud – treenides mudeli vaid 170 rahapesu kahtlusega ja 170 rahapesu kahtluseta kliendi põhjal, ei ole mudel skaleeruv ning ei suuda kogu kliendibaasil täpselt eristada ei enamus- ega ka vähemusklassi. Mudeli sooritust iga iteratsiooni vältel hinnati testandmetel, kuhu kuulusid vastavalt 172 rahapesukahtlusega ning 14006 heatahtlikku klienti. Kusjuures nagu eespool mainitud, siis jagati vaatlused treening ning testandmetesse igas iteratsioonis suvaliselt, fikseerides vaid selle, kui palju vähemusklassi vaatlusi mõlemas andmete alamhulgas peab olema.

Kõige parema F1–skoori tagas SMOTE kombinatsioon 1200/700, mille puhul genereeriti iga vähemusklassi vaatluse kohta 12 uut, sünteetilist vaatlust. Iga genereeritud sünteetilise vähemusklassi vaatluse kohta kaasati treeningandmetesse 7 enamusklassi vaatlust. Seega koosnes lõplik andmestik SMOTE rakendamisel 2210 rahapesu kahtlusega ning 14280 rahapesu kahtluseta vaatlusest. Tomeki sideme olemasolu tuvastati SMOTE poolt genereeritud andmetes 460 vaatlusel, seega jäi pärast üle- ja alanäidistamist andmestikku 2210 rahapesu kahtlusega ning 13820 rahapesu kahtluseta vaatlust. Ülevaade mudeli treenimiseks ning testimiseks kasutatud andmete struktuurist on kokkuvõtvalt koondatud järgnevasse tabelisse (vt tabel 5).

Tabel 5. Mudeli treenimiseks ja testimiseks kasutatud andmete struktuur erinevate meetodite puhul esimeses tsüklis.

Meetod	Treening		Test	
	Kahtlusega	Kahtluseta	Kahtlusega	Kahtluseta
Juhumets	170	14008	172	14006
SMOTE + juhumets	2210	14280	172	14006
SMOTE + Tomeki sidemed + juhumets	2210	13820	172	14006

Allikas: autori koostatud.

Järgnevalt annab autor ülevaate esimeses tsüklis viie iteratsiooni vältel treenitud mudeli F1–skooridest (vt tabel 6). Kusjuures tulenevalt eespool kirjeldatud 5x2 ristkontrollimise (vt joonis 5) põhimõttest on siin ja ka edaspidi iga iteratsiooni tulemusmõõdik esitatud mudeli 1 ja mudeli 2 tulemuse keskmisena. Tabelist ilmneb, et sünteetiliste vaatluste genereerimine parandas mudeli keskmist sooritust F1–skoori lõikes üle kolme korra. Nii üle- kui ka alanäidistamismeetodite kombineerimine käesolevas tsüklis mudeli tulemusi täiendavalt ei parandanud – keskmine sooritus on sama, mis vaid SMOTE rakendamisel. Tomeki sidemete eemaldamine parandas tulemust küll võrreldes ainult juhumetsa rakendamisega, ent tagas sama tulemuse, mis juhumetsa ja SMOTE kombineerimine. Tomeki sidemete eemaldamisel oli mudeli sooritus ebastabiilsem, sest viie iteratsiooni vältel on parima ning halvima tulemusmõõdiku vahe (0,07) ligi kaks korda sama palju, kui SMOTE ja juhumetsa kombineerimisel (0,04). Vaatamata sellele tagas iteratsioonide lõikes parima F1–skoori (0,14) SMOTE ja Tomeki sidemete kombineerimine, mis on ligi kolm korda parem, kui parima juhumetsa sooritus. Nagu tabelist ilmneb, siis juhumetsa rakendamisel viiendal iteratsioonil ei olnud võimalik F1–skoori välja arvutada, sest tundlikkus või täpsus olid 0. F1–skoori lõikes tagas antud tsüklis parima ning stabiilsema soorituse SMOTE ja juhumetsa kombineerimine, ent vaatamata sellele ei ole mudelit võimalik rahapesu tuvastamiseks kasutada, sest tundlikkuse ja täpsuse harmooniline keskmine on vaid veidi suurem, kui 0.

Tabel 6. Mudeli F1–skoor viie iteratsiooni vältel esimeses tsüklis.

Meetod	Iteratsiooni number					Keskmine	Max–min
	1	2	3	4	5		
Juhumets	0,03	**0,05**	0,01	0,02	NA	0,03	0,04
SMOTE + juhumets	**0,13**	0,10	0,09	0,08	0,10	0,10	0,04
SMOTE + Tomeki sidemed + juhumets	0,13	0,09	0,10	0,07	**0,14**	0,10	0,07

Allikas: autori koostatud.

Selleks, et näha mis mõjutas F1–skoori ning mudelit parandada, annab autor ülevaate mudeli tundlikus- ning täpsusmõõdikutest (vt tabel 7). Tabelist ilmneb, et rakendades vaid juhumetsa, on mudeli tundlikkus nullilähedane, ent täpsus on kahe iteratsiooni vältel 1, ehk mudeli jaoks kahtlastest vaatlustest olid kõik ka tegelikkuses kahtlased. Mudeli halvast tundlikkusest tingituna ei ole vaid juhumetsa rakendamine sobiv rahapesu tuvastamiseks, sest mudel leidis üles vaid 1% kõikidest kahtlastest vaatlustest. Suure erinevusega täpsuse ja tundlikkuse vahel kaasneb väga palju valepositiivseid otsuseid ehk statistiliselt II liiki vigu. Nii üle- kui ka alanäidistamismeetodi kombineerimine parandas mudeli sooritust tundlikkuse lõikes seitsmekordselt. Näiliselt vähendab üle– ning alanäidistamismeetodite rakendamine küll mudeli täpsus kuid see ei ole antud olukorras tähendusrikas, sest mudeli tundlikkus paranes seitse korda täpsuse vaid kolmekordse vähenemise arvelt. Parima mudeli puhul peaks tundlikkus ning täpsus olema mõlemad võimalikult suured ning nende erinevus võimalikult väike – antud juhul tagas parima soorituse juhumetsa kombineerimine sünteetilise vähemusklassi ülenäidistamisega. Ümbernäidistamise tulemusena paranesid kõige rohkem mudeli täpsus ning F1–skoor, seega tagas käesolevas tsüklis parima stabiilsuse ning soorituse SMOTE ja juhumetsa kombineerimine.

Tabel 7. Mudeli tundlikkus ning täpsus viie iteratsiooni vältel esimeses tsüklis.

Mõõdik	**Meetod**	**Iteratsiooni number**					**Keskmine**	**Max–min**
		1	2	3	4	5		
Täpsus	Juhumets	**1**	0,67	0,5	1	0	0,63	1
	SMOTE + juhumets	0,23	0,2	**0,37**	0,16	0,26	0,24	0,21
	SMOTE + Tomeki sidemed + juhumets	0,21	0,17	**0,32**	0,13	0,32	0,23	0,2
Tundlikkus	Juhumets	0,01	**0,03**	0,01	0,01	0,00	0,01	0,03
	SMOTE + juhumets	**0,09**	0,07	0,05	0,06	0,07	0,07	0,04
	SMOTE + Tomeki sidemed + juhumets	**0,09**	0,06	0,06	0,05	0,09	0,07	0,05

Allikas: autori koostatud.

Selleks, et saada paremat ülevaadet tulemusmõõdikuid mõjutanud teguritest, esitab autor kolme erineva mudeli keskmisest sooritusest lähtuvalt konfusioonimaatriksi (vt tabel 8). Vaatamata sellele, et ümbernäidistamismeetodite rakendamisel mudeli sooritus oluliselt paranes, ei ole ükski mudel skaleeruv kogu kliendibaasil II liigi vigade rohkuse tõttu – iga õige otsusega kaasneb keskmiselt 3–4 valepositiivset otsust. Nagu eespool mainitud

leidis kõige rohkem rahapesukahtlusega kliente üles juhumetsa kombineerimine ümbernäidistamismeetoditega. Rahapesu tuvastamisel on oluline minimeerida I liiki vigu, et ükski rahapesu kahtlusega klient ei jääks märkamata. Käesolevas tsüklis vähendas SMOTE rakendamine küll I liiki vigu, kuid suurendas II liiki vigade arvu. Olgugi, et üle- ja alanäidistamismeetodi kombineerimise puhul oli I liiki vigu kõige vähem, kaasnes sellega võrdlemisi palju II liiki vigu, seega tundub antud tsükli näitel kõige optimaalsem vaid juhumetsa kombineerimine vaid SMOTE–ga. Antud juhul võib tuua välja kaks peamist põhjust, miks mudeli sooritus ei ole rahuldav – treenimisel kasutatud tunnuseid on liiga vähe või treeningul kasutatud vähemusklassi vaatluseid ei ole piisavalt palju.

Tabel 8. Kolme erineva mudeli konfusioonimaatriks esimeses tsüklis.

		Tegelikkus		
		Kahtlusega	Kahtluseta	
Ennustus	Kahtlusega	37	1	juhumets
	Kahtluseta	135	14005	
	Kahtlusega	44	31	SMOTE + juhumets
	Kahtluseta	128	13975	
	Kahtlusega	45	36	SMOTE + Tomeki sidemed + juhumets
	Kahtluseta	127	13970	

Allikas: autori koostatud.

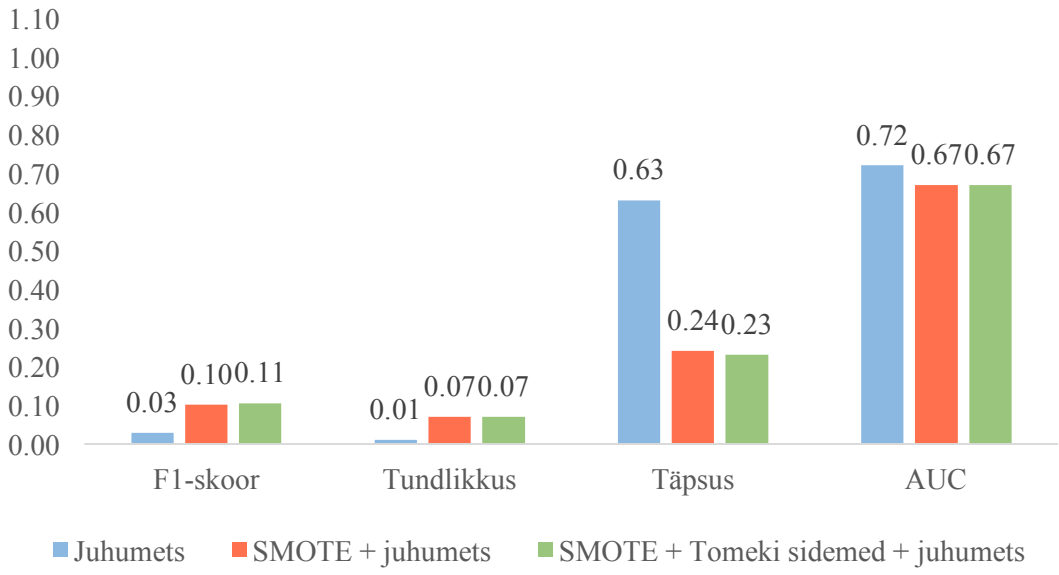
Järgnevalt on autor koondanud tabelisse ROC kõvera alla jääva pindala igal iteratsioonil (vt tabel 9). Olgugi, et F1–skoor ning täpsus paranesid ümbernäidistamismeetodite rakendamisel oluliselt, kinnitab AUC näitaja väärtus asjaolu, et antud mudelid reaalsetel andmetel kasutatavad ei ole – väärtus 0,5 kirjeldab otsuseid, mis on tehtud täielikult juhuslikkuse alusel, st praegused mudelid on veidi täpsemad juhuslikkusest ja teevad võrdlemisi palju I ja II liiki vigu. Erinevalt täpsusest on mudeli soorituste erinevus AUC näitaja lõikes viie iteratsiooni vältel võrdlemisi väike. Nii iga iteratsiooni kui ka keskmise näitaja lõikes tagas kõige parema soorituse juhumets. Mõlemad ümbernäidistamismeetodid tagasid sama hea tulemuse nii iga iteratsiooni lõikes (0,69) kui ka keskmiselt (0,67).

Tabel 9. Mudeli ROC kõvera alune pindala viie iteratsiooni vältel esimeses tsüklis.

Meetod	Iteratsiooni number					Keskmine	Max–min
	1	2	3	4	5		
Juhumets	0,72	0,72	**0,73**	0,73	0,71	0,72	0,02
SMOTE + juhumets	0,66	0,67	0,67	**0,69**	0,67	0,67	0,02
SMOTE + Tomeki sidemed + juhumets	0,66	0,67	0,67	**0,69**	0,67	0,67	0,03

Allikas: autori koostatud.

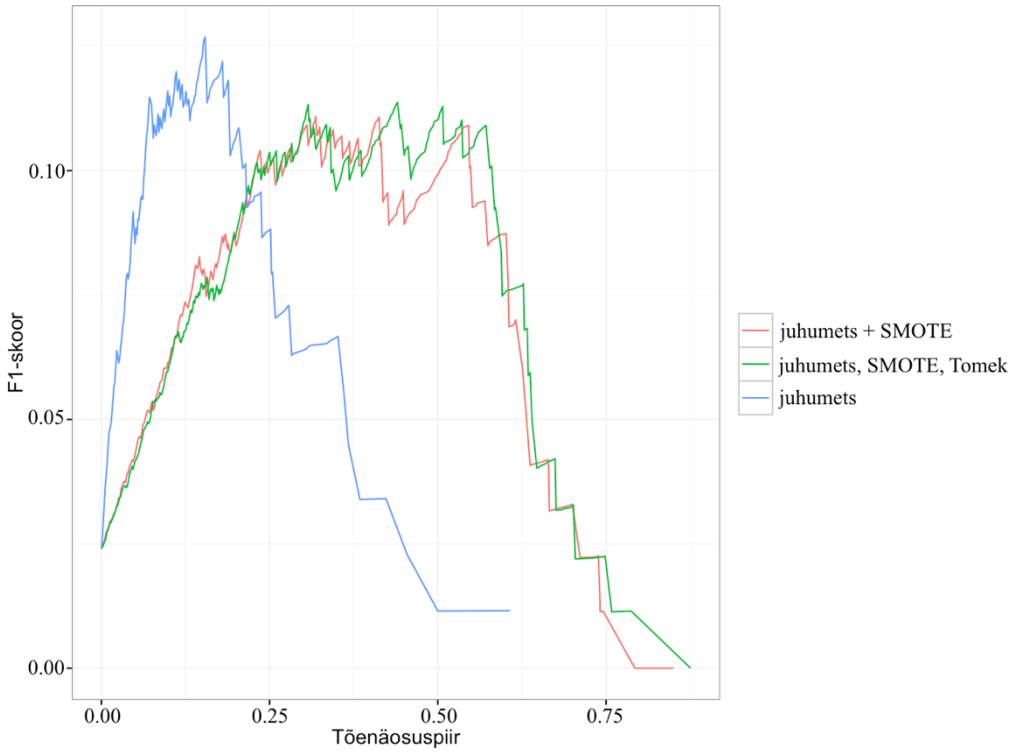
Kokkuvõtvalt on mudeli keskmised tulemusmõõdikud esitatud järgneval joonisel (vt joonis 7) ning nagu ilmneb, siis ümbernäidistamismeetodite rakendamine parandas mudeli sooritust kõige rohkem F1–skoori ning tundlikkuse lõikes. Vaatamata suurele täpsusele ning AUC näitajale, mille väärtus on üle 0,5, võib öelda, et eespool esitatud nelja tulemusmõõdiku lõikes on parim sooritus SMOTE ja juhumetsa kombineerimisel.



Joonis 7. Esimeses tsüklis treenitud mudelite keskmine sooritus (autori koostatud).

Lisaks iga iteratsiooni tulemusmõõdikute hindamisele skitseerib autor mudeli stabiilsuse täiendavaks hindamiseks iga mudeli F1–skoori seose tõenäosuspiiriga (vt joonis 8). Jooniselt ilmneb, et iga meetodi puhul on tegu väga ebastabiilse mudeliga – sooritus F1–skoori lõikes muutub hüppeliselt tõenäosuspiiride nihutamisel. Juhumetsa puhul tagas kõige kõrgema F1–skoori tõenäosuspiir 0,13, seega võib öelda, et mudel alahindab oluliselt vaatlusi, sest ei suuda kunagi öelda piisava kindlusega ($p > 0,5$), kas klient

tegeles rahapesuga. Mudelite ebastabiilsus on tingitud tõenäoliselt ekstremaalselt tasakaalust väljas andmetest ja/ või kirjeldavate tunnuste vähesusest.



Joonis 8. Mudelite F1–skoori seos tõenäosuspiiriga (autori koostatud).

Teises tsüklis vaadati kriitilise pilguga üle mudeli treenimiseks kasutatud rahapesu kahtlusega juhtumid ning tagasisidest lähtuvalt jäeti alles vaid 251 vaatlust, millest mudeli treenimisse kaasati igas iteratsioonis 123. Sarnaselt esimesele tsüklile moodustas vähemusklass ~1% kogu andmetest. Nagu eespool kirjeldatud, lisati tagasisidest tulenevalt mudelisse ka 16 uut tunnust. Mudeli sooritust iga iteratsiooni vältel hinnati testandmetel, kuhu kuulusid 128 rahapesu kahtlusega ning 13990 heatahtlikku klienti. Sarnaselt esimesele tsüklile jagati vaatlused treening- ning testandmetesse igas iteratsioonis suvaliselt. Antud tsüklis tagas kõrgeima F1–skoori SMOTE kombinatsioon 1200/700 ning lõplik andmestik SMOTE rakendamisel koosnes 1599 rahapesu kahtlusega ning 10332 rahapesu kahtluseta vaatlusest. Tomeki sideme olemasolu tuvastati SMOTE poolt genereeritud andmetes 284 vaatlusel, seega jäi pärast üle- ja alanäidistamist andmestikku 2210 rahapesu kahtlusega ning 10048 rahapesu kahtluseta vaatlust. Ülevaade mudeli treenimiseks ning testimiseks kasutatud andmete struktuurist on kokkuvõtvalt esitatud järgnevas tabelis (vt tabel 10).

Tabel 10. Mudeli treenimiseks ja testimiseks kasutatud andmete struktuur erinevate meetodite puhul teises tsüklis.

Meetod	Treening		Test	
	Kahtlusega	Kahtluseta	Kahtlusega	Kahtluseta
Juhumets	123	13996	128	13990
SMOTE + juhumets	1599	10332	128	13990
SMOTE + Tomeki sidemed + juhumets	1599	10048	128	13990

Allikas: autori koostatud.

Teises tsüklis treenitud mudelite F1–skoor viie iteratsiooni vältel on esitatud tabelis (vt tabel 11). Erinevalt esimesest tsüklist on tulemused iga iteratsiooni vältel oluliselt stabiilsemad – keskmise soorituse erinevus parimast ja halvimast näitajast on võrdlemisi väike. Sarnaselt esimesele tsüklile tagasid kõige parema soorituse nii SMOTE ja juhumetsa kui ka SMOTE, Tomeki sidemete ja juhumetsa rakendamine. Ümbernäidistamismeetodite rakendamisel paranes F1–skoor ligi 0,2 korda, kusjuures esimeses tsüklis, kus oli vähem tunnuseid, oli sama mõõdiku paranemine kolmekordne. Võrreldes esimese tsükliga parandas täiendavate tunnuste lisamine ning andmete puhastamine empiirilise vaatluse teel mudeli sooritust mitmekümnekordselt. Nagu ka esimeses tsüklis, on erinevate ümbernäidistamismeetodite rakendamisel treenitud mudelite sooritus samaväärne, ent viie iteratsiooni vältel kõige stabiilsem SMOTE ja juhumetsa kombineerimisel. SMOTE ja juhumetsa kombineerimine tagas keskmiselt 0,06 võrra parema soorituse kui lihtsalt juhumetsa rakendamine.

Tabel 11. Mudeli F1–skoor viie iteratsiooni vältel teises tsüklis.

Meetod	Iteratsiooni number					Keskmine	Max–min
	1	2	3	4	5		
Juhumets	0,28	**0,28**	0,24	0,23	0,25	0,26	0,06
SMOTE + juhumets	0,31	**0,35**	0,30	0,32	0,30	0,32	0,05
SMOTE + Tomeki sidemed + juhumets	0,34	**0,35**	0,31	0,32	0,27	0,32	0,08

Allikas: autori koostatud.

Järgnevalt annab autor ülevaate teises tsüklis viie iteratsiooni vältel treenitud mudeli täpsus- ning tulemusmõõdikutest (vt tabel 12). Täpsuse ja tundlikkuse lõikes ilmneb juhumetsa rakendamisel sama tendents, mis esimeses tsüklis – tundlikkus on väga kõrge kuid täpsus madal. Võrreldes esimese tsükliga paranesid nii tundlikkus kui ka täpsus mitmekordselt. Ühtlasi ilmneb, et täiendavate tunnuste lisamine ning andmete puhastamine parandasid ümbernäidistamise meetoditest oluliselt rohkem juhumetsa

tundlikkust (15-kordselt). Nagu F1–skoori puhul juba ilmnes, tagas kõige parema sooritusе SMOTE ja juhumetsa rakendamine, sest seos tundlikkuse ja täpsuse vahel on kõige optimaalsem. Seevastu hinnates mudeli stabiilsust täpsuse ja tundlikkuse lõikes ilmneb, et erinevalt teistest meetoditest on kõige parem sooritus vaid juhumetsa rakendamisel – parima täpsuse ning tundlikkuse erinevus halvimast on kõige väiksem. Rakendades vaid juhumetsa, leidis mudel üles 15% kõikidest kahtlastest vaatlustest, ent ülejäänud 12% mudeli jaoks kahtlastest klientidest olid tegelikult heatahtlikud, SMOTE rakendamisel leiti üles vastavalt 20% ja ülejäänud 24% mudeli jaoks kahtlastest klientidest olid heatahtlikud. Rakendades nii SMOTE–d kui ka Tomeki sidemete eemaldamist leidis mudel üles 21% kahtlastest klientidest kuid ülejäänud 28% mudeli jaoks kahtlastest klientidest olid heatahtlikud.

Tabel 12. Mudeli tundlikkus ning täpsus viie iteratsiooni vältel teises tsüklis.

Mõõdik	**Meetod**	**Iteratsiooni number**					**Keskmine**	**Max–min**
		1	2	3	4	5		
Täpsus	Juhumets	0,84	**0,91**	0,9	0,9	0,9	0,88	0,07
	SMOTE + juhumets	**0,86**	0,8	0,8	0,8	0,6	0,76	0,28
	SMOTE + Tomeki sidemed + juhumets	**0,87**	0,74	0,78	0,7	0,5	0,72	0,38
Tundlikkus	Juhumets	0,17	**0,17**	0,14	0,13	0,15	0,15	0,04
	SMOTE + juhumets	0,19	**0,22**	0,19	0,2	0,2	0,20	0,04
	SMOTE + Tomeki sidemed + juhumets	0,21	**0,23**	0,19	0,2	0,19	0,21	0,04

Allikas: autori koostatud.

Selleks, et paremini mõista, mis mõjutasid täpsust ning tundlikkust esitab autor järgnevalt kolme erineva mudeli keskmisest sooritusest lähtuvalt konfusioonimaatriksi (vt tabel 13). Nagu tabelist ilmneb, tehti keskmiselt kõige vähem esimest liigi vigu erinevate ümbernäidistamismeetodite rakendamisel (102). Antud tulemusmõõdiku lõikes on kõige optimaalsem mudel, mis teeb kõige vähem I liiki vigu ja tekitab võrdlemisi vähe II liiki vigu. Esimest ja teist liiki vigu ei ole kerge võrrelda, sest nende kaalud on väga erinevad. Pannes arvud konteksti võib öelda, et rakendades vaid juhumetsa, jäänuks 7 rahapesijat ja/ või terrorismi finantseerijat tuvastamata, seega leiab autor, et kõige sobivam on kasutada SMOTE ja juhumetsa kombinatsiooni, sest tekitab kõige vähem esimest liiki vigu. Nagu ka teoreetilises osas välja toodud, suudab mudel võrdlemisi täpselt klassifitseerida enamusklassi vaatluseid, ent mitte vähemusklassi.

Tabel 13. Kolme erineva mudeli konfusioonimaatriks teises tsüklis.

		Tegelikkus		
		Kahtlusega	Kahtluseta	
Ennustus	Kahtlusega	19	3	juhumets
	Kahtluseta	109	13987	
	Kahtlusega	26	9	SMOTE + juhumets
	Kahtluseta	102	13981	
	Kahtlusega	26	12	SMOTE + Tomeki sidemed + juhumets
	Kahtluseta	102	13978	

Allikas: autori koostatud.

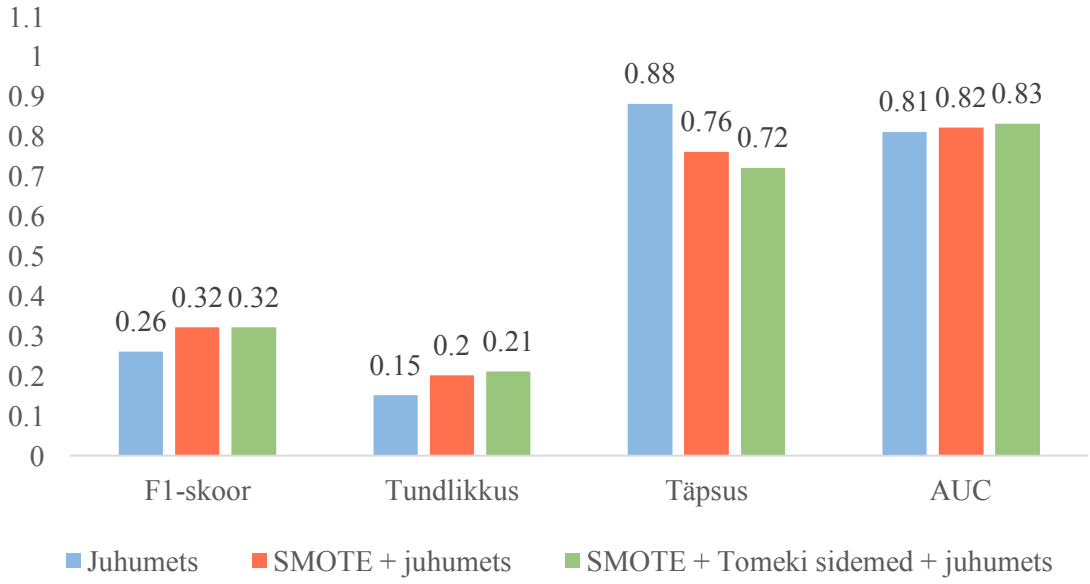
Järgnevalt toob autor välja ROC kõvera aluse pindala viie iteratsiooni vältel teises tsüklis (vt tabel 14). Nagu tabelist ilmneb on AUC näitaja kolme erineva mudeli puhul ligilähedaselt sarnane ning viie iteratsiooni vältel üsna stabiilne. Erinevalt esimesest tsüklist tagas parima soorituse AUC lõikes nii üle- kui ka alanäidistamismeetodite kombineerimine. Juhumetsa sooritus antud tulemusmõõdiku lõikes oli kõige halvem ning ka ebastabiilsem – parima ja halvima soorituse vahe iga iteratsiooni lõikes oli 0,06. Võrreldes esimese tsükliga paranes kõige rohkem SMOTE ja Tomeki sidemete kombineerimisel treenitud mudeli sooritus (0,16 võrra).

Tabel 14. Mudeli ROC kõvera alune pindala viie iteratsiooni vältel teises tsüklis.

Meetod	**Iteratsiooni number**					**Keskmine**	**Max–min**
	1	2	3	4	5		
Juhumets	0,78	0,82	0,81	0,81	**0,84**	0,81	0,06
SMOTE + juhumets	0,82	0,82	0,81	0,82	**0,85**	0,82	0,04
SMOTE + Tomeki sidemed + juhumets	0,82	0,83	0,82	0,82	**0,85**	0,83	0,03

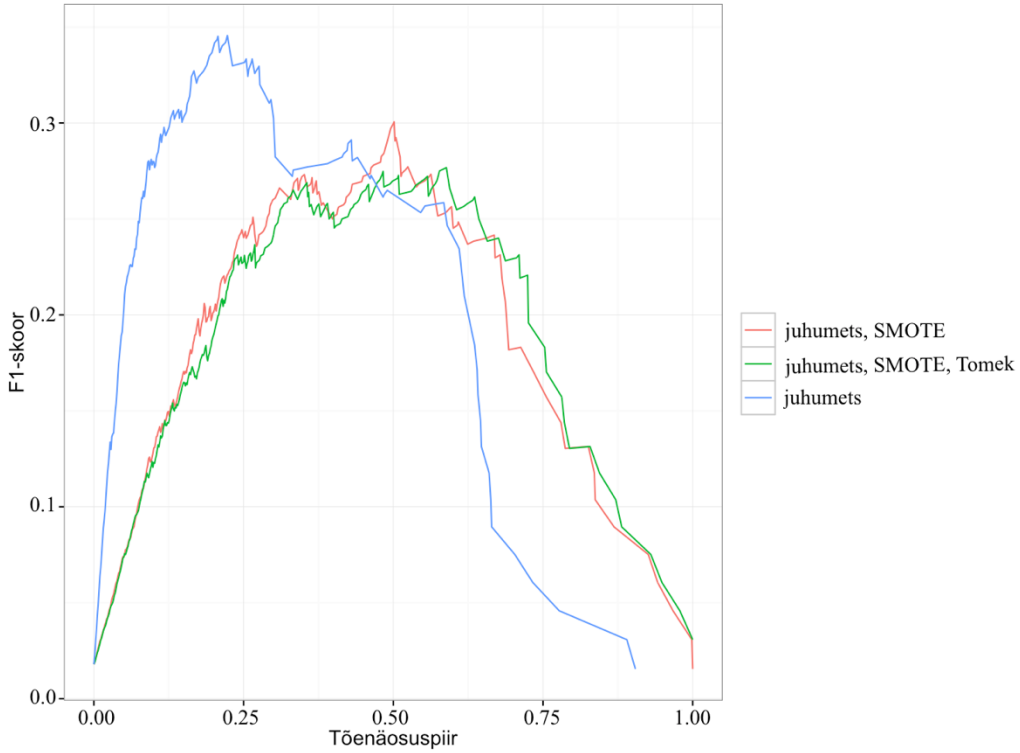
Allikas: autori koostatud.

Kokkuvõtvalt on iga mudeli keskmised tulemusmõõdikud esitatud järgneval joonisel (vt joonis 9). Nagu eespool argumenteeritud ning ka jooniselt ilmneb, siis teises tsüklis tagas kõige parema soorituse SMOTE ning juhumetsa kombineerimine. Ligilähedaselt sarnase soorituse tagas SMOTE ja Tomeki sidemete kombineerimine kuid kuna AUC ning F1–skoorid on mudelitel võrdsed, lähtub autor mudelite võrdlemisel nende täpsusest. Rakendades vaid SMOTE–d on mudeli täpsus 0,04 võrra suurem 0,01 täpsuse arvelt.



Joonis 9. Teises tsüklis treenitud mudelite keskmine sooritus (autori koostatud).

Lisaks iga iteratsiooni tulemusmõõdikute hindamisele skitseerib autor mudeli stabiilsuse täiendavaks hindamiseks iga mudeli F1–skoori seose tõenäosuspiiriga (vt joonis 10). Sarnaselt esimesele tsüklile alahindab vaid juhumetsa rakendamine vaatlusi, ent üldiselt on stabiilsus oluliselt parem, kui esimeses tsüklis, sest ei esine nii järske kõikumisi.



Joonis 10. Mudelite F1–skoori seos tõenäosuspiiriga (autori koostatud).

Kolmas tsükkel toimus ajaliselt oluliselt hiljem kui esimesed kaks tsüklit, seega oli vahepeal kogunenud nii mudeli rakendamisest kui ka igapäevasest monitoorimisest juurde rahapesu kahtlusega vaatluseid, mida analüüsi lisada. Treeningandmetesse oli koondatud iga viie iteratsiooni vältel 239 rahapesu kahtlusega ning 17693 rahapesu kahtluseta vaatlust ning mudeli treenimisel saadud tulemusi kontrolliti testvalimil, kuhu kuulusid vastavalt 248 rahapesu kahtlusega ning 17683 rahapesu kahtluseta vaatlust. Sarnaselt varasemate tsüklitega tagas kõige kõrgema F1–skoori SMOTE konfiguratsioon 1200/700. Nagu eespool mainitud, lisati käesolevas tsüklis 2 uut tunnust ning esitati maksete käitumuslikkust kirjeldavad tunnused osakaaludena. Mudeli treenimiseks ning testimiseks kasutatud andmete struktuur on esitatud järgnevas tabelis (vt tabel 15).

Tabel 15. Mudeli treenimiseks ja testimiseks kasutatud andmete struktuur erinevate meetodite puhul kolmandas tsüklis.

Meetod	Treening		Test	
	Kahtlusega	Kahtluseta	Kahtlusega	Kahtluseta
Juhumets	239	17693	248	17683
SMOTE + juhumets	3107	20076	248	17683
SMOTE + Tomeki sidemed + juhumets	3107	19926	248	17683

Allikas: autori koostatud.

Järgnevalt annab autor ülevaate kolmandas tsüklis treenitud mudelite F1–skoorist viie iteratsiooni vältel (vt tabel 16). Nagu tabelist ilmneb, tagas viimases tsüklis kõige kõrgema F–1 skoori SMOTE ja Tomeki sidemete kombinatsioon, mis oli ka iga iteratsiooni vältel parima sooritusega. Vähemusklassi suurendamise ning tunnuste osakaaludena esitamise tulemusena paranes tulemusmõõdik üle kahe korra. Võrreldes esimese tsükliga on kõige rohkem F1–skoor paranenud juhumetsal, ligi 25 korda ning võrreldes teise tsükliga ligi kolm korda. Ümbernäidistatud andmetele püstitatud mudeli F1–skoor paranes võrreldes eelmise tsükliga ~2.3 korda ning võrreldes esimese tsükliga ligikaudu 8 korda. Seega on kolme tsükliga parandatud mudeli sooritust oluliselt ning lähtudes F1–skoorist võiks sobida rahapesu efektiivseks tuvastamiseks.

Tabel 16. Mudeli F1–skoor viie iteratsiooni vältel kolmandas tsüklis.

Meetod	Iteratsiooni number					Keskmine	Max–min
	1	2	3	4	5		
Juhumets	0,71	0,76	0,77	**0,78**	0,70	0,74	0,08
SMOTE + juhumets	0,71	0,76	0,75	**0,77**	0,77	0,75	0,06
SMOTE + Tomeki sidemed + juhumets	0,72	0,76	0,77	**0,80**	0,77	0,77	0,08

Allikas: autori koostatud.

Mudeli sooritusest parema ülevaate saamiseks esitab autor järgnevas tabelis mudeli treenimise kolmanda tsükli tulemus- ja täpsusmõõdikud (vt tabel 17). Nagu tabelist ilmneb, siis peaaegu iga iteratsiooni vältel olid täpsused üsna sarnased. Keskmiselt leidis mudel vaid juhumetsa rakendamisel üles 60% kõikidest kahtlastest vaatlustest ning 98% mudeli jaoks kahtlastest vaatlustest olid ka päriselus kahtlased. Sellest tulenevalt tehti võrdlemisi vähe II liiki vigu. Ümbernäidistamismeetodite rakendamine tekitas küll rohkem II liiki vigu, ent vähendas I liiki vigu – mudel leidis üles vastavalt 63% ja 65% kahtlastest klientidest, ent 6% mudeli jaoks kahtlastest klientidest olid tegelikult heatahtlikud. Kolmandas tsüklis tagas nende kahe tulemusmõõdiku lõikes seega kõige parema tulemuse SMOTE ja Tomeki sidemete kombineerimine.

Tabel 17. Mudeli tundlikkus ning täpsus viie iteratsiooni vältel kolmandas tsüklis.

Mõõdik	Meetod	Iteratsiooni number					Keskmine	Max–min
		1	2	3	4	5		
Täpsus	Juhumets	0,99	**1,00**	0,99	0,96	0,98	0,98	0,04
	SMOTE + juhumets	0,92	0,95	0,94	**0,96**	0,91	0,94	0,05
	SMOTE + Tomeki sidemed + juhumets	0,91	**0,96**	0,94	0,95	0,92	0,94	0,05
Tundlikkus	Juhumets	0,56	0,61	0,63	**0,66**	0,54	0,60	0,11
	SMOTE + juhumets	0,58	0,64	0,63	0,64	**0,66**	0,63	0,08
	SMOTE + Tomeki sidemed + juhumets	0,60	0,63	0,66	**0,69**	0,67	0,65	0,09

Allikas: autori koostatud.

Selleks, et täpsus ning tulemusmõõdikuid paremini tõlgendada annab autor ülevaate I ja II liiki vigade arvu kohta viie iteratsiooni vältel mudeli treenimise kolmandas tsüklis (vt tabel 18). Nagu eespool mainitud leidis testandmetest kõige rohkem kahtlaseid vaatlusi üles SMOTE ning Tomeki sidemete kombineerimisel saadud mudel. Lähtudes sellest, et mudel leiab üles ligi 65% kahtlastest klientidest ning tekitab seejuures väga vähe müra II liiki vigade näol, on kolmandas tsüklis jõutud mudelini, mida saab efektiivselt kasutada rahapesu tõkestamiseks. Nagu eespool mainitud, tegelevad rahapesu tõkestamisega igapäevaselt vastava väljaõppe saanud agendid, seega on masinõppe mudel pigem tugisüsteem agentidele mitte miski, millele ainsana tuginetakse. Kusjuures tasub nentida, et mudeli puhul, mille täpsus on võrdlemisi suur, ei pruugi kõik II liiki vead olla tegelikkuses valepositiivsed otsused – ka testandmetest on võimalik tuvastada uusi rahapesu kahtlusega vaatlusi.

Tabel 18. Kolme erineva mudeli konfusioonimaatriks kolmandas tsüklis.

		Tegelikkus		
		Kahtlusega	Kahtluseta	
Ennustus	Kahtlusega	151	3	juhumets
	Kahtluseta	97	17680	
	Kahtlusega	158	10	SMOTE + juhumets
	Kahtluseta	90	17673	
	Kahtlusega	162	11	SMOTE + Tomeki sidemed + juhumets
	Kahtluseta	86	17672	

Allikas: autori koostatud.

Selleks, et mudeleid võrrelda annab autor ülevaate ka ROC kõvera alla jääva osa pindalast (vt tabel 19). Nagu tabelist ilmneb, on see iga meetodi puhul pindala 0,98 ehk tegu on peaaegu ideaalse klassifikaatoriga. Iga mudeli puhul oli iteratsioonide lõikes parimaks soorituseks 0,99 ning halvima ning parima soorituse vahe maksimaalselt 0,03. ROC kõvera alla jääva osa pindala kohaselt on kõik mudelid võrdselt head, ent tuginedes eespool toodud argumentatsioonile leiab autor, et kolmanda tsükli parim mudel saavutatakse SMOTE ja Tomeki sidemete rakendamisel. Võrreldes eelmise tsükliga paranes ROC näitaja juhumetsa puhul 0,18 võrra, SMOTE ja juhumetsa puhul 0,17 võrra ning SMOTE ja Tomeki kombineerimisel juhumetsa treenimisel 0,16 võrra.

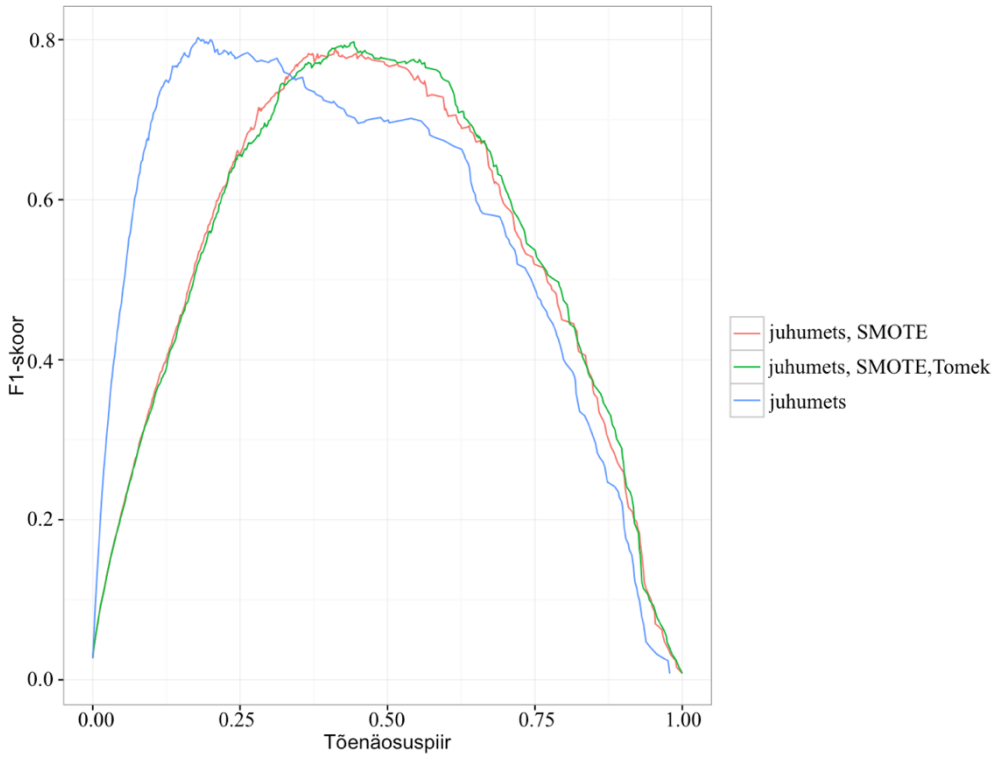
Tabel 19. Mudeli ROC kõvera alune pindala viie iteratsiooni vältel kolmandas tsüklis.

Meetod	**Iteratsiooni number**					**Keskmine**	**Max–min**
	1	2	3	4	5		
Juhumets	0,98	0,98	0,98	**0,99**	0,97	0,98	0,01
SMOTE + juhumets	0,98	0,97	0,98	**0,99**	0,97	0,98	0,02
SMOTE + Tomeki sidemed + juhumets	0,98	0,96	0,98	**0,99**	0,97	0,98	0,03

Allikas: autori koostatud.

Lisaks iga iteratsiooni tulemusmõõdikute hindamisele skitseerib autor mudeli stabiilsuse täiendavaks hindamiseks iga mudeli F1–skoori seose tõenäosuspiiriga (vt joonis 10). Nagu tabelist ilmneb, alahindab ka viimases tsüklis juhumets vaatlusi – suurim F1–skoor saavutatakse tõenäosuspiiri 0,2 juures. Nii üle- kui ka alanäidistamismeetodite rakendamise puhul võib öelda, et on saavutatud stabiilne ning skaleeruv mudel – F1–skoori ja tõenäosuspiiri seos sarnaneb normaaljaotusele, st kõige kõrgema F1–skoori tagab tõenäosuspiir ~0,5. Lisaks viitab joonte võrdlemisi sujuv kulgemine sellele, et

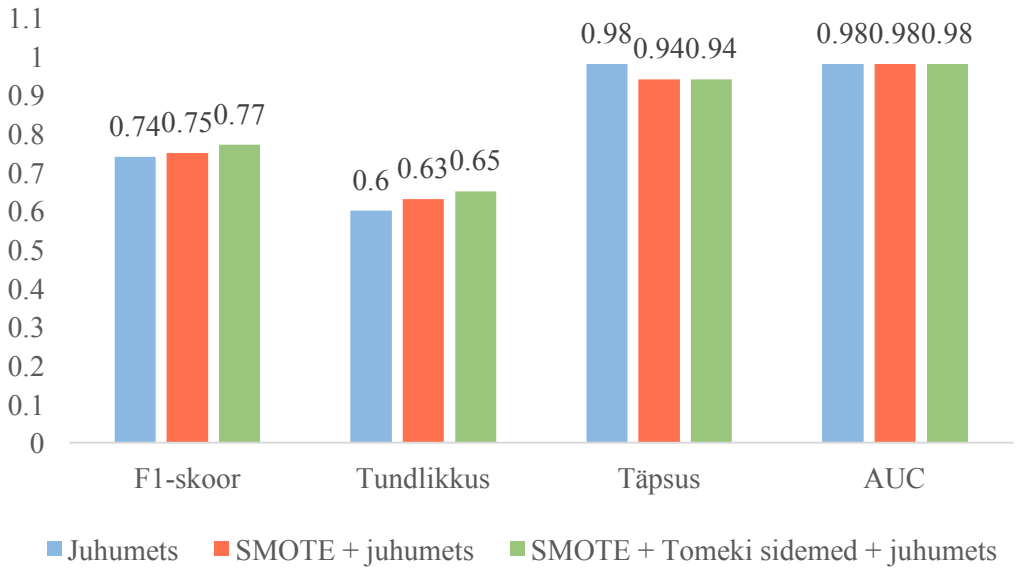
mudelid suudavad piisava kindlusega eristada rahapesu kahtlusega kliente headest klientidest ja ei tee põhjendamatuid vigu.



Joonis 10. Mudelite F1–skoori seos tõenäosuspiiriga (autori koostatud).

Kokkuvõtvalt on iga iteratsiooni keskmised tulemusmõõdikud esitatud joonisel (vt joonis 11). Nagu eespoolt argumenteeritud osutus kolmandas tsüklis parimaks mudeliks ülening alanäidistamismeetodite kombineerimine, mis tagas kõige suurema F1–skoori ning tundlikkuse. AUC näitaja oli iga mudeli puhul võrdne ning täpsuse lõikes osutus parimaks mudeliks juhumets. Nagu eespool ilmnes, alahindab juhumets vaatlusi, mis on otseses seoses täpsusnäitajaga ning sellest tulenevalt ei ole juhumetsa parim sooritus täpsuse lõikes asjakohane. Võrreldes teise tsükliga paranes kõige rohkem F1–skoor juhumetsa rakendamisel (0,48 võrra) kuid ligikaudu sama hea paranemine oli ka SMOTE ja juhumetsa rakendamisel (0,43) ning SMOTE, Tomeki sidemete ning juhumetsa kombineerimisel (0,45). Võrreldes teise tsükliga paranes tundlikkus iga mudeli lõikes ligi kolm korda. Olgugi, et juhumets tagas kõige parema täpsuse, oli tulemusmõõdiku paranemine võrreldes teise tsükliga kõige suurem SMOTE, Tomeki sidemete ja juhumetsa rakendamisel, 0,22 võrra. Tuginedes joonisel esitatud tulemusmõõdikutele leiab autor, et SMOTE ja Tomeki sidemete kombineerimine juhumetsaga on käesoleva

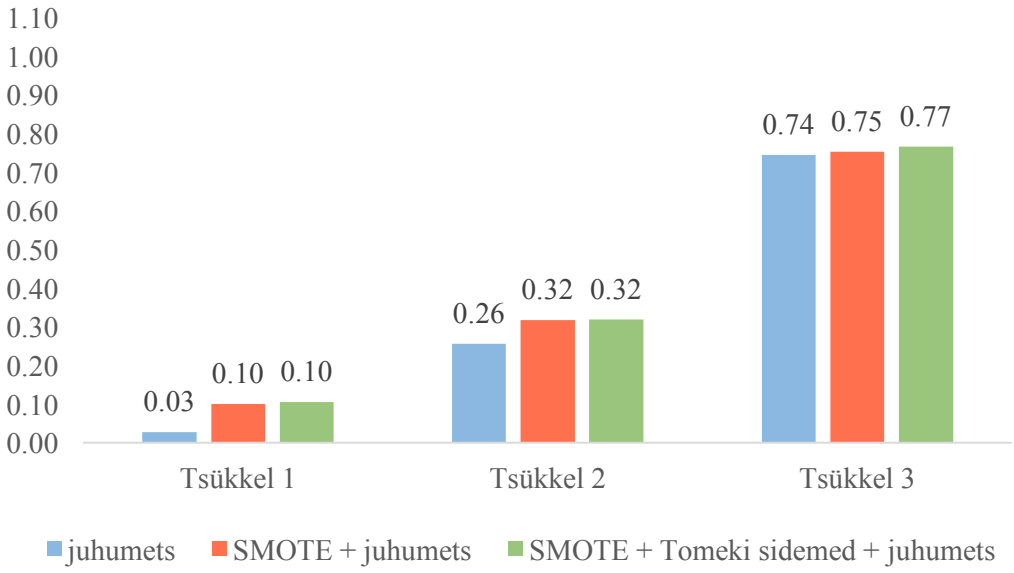
uurimisprobleemi puhul asjakohane ning võimaldab efektiivselt aidata kaasa rahapesu tuvastamisele.



Joonis 11. Kolmandas tsüklis treenitud mudelite keskmine sooritus (autori koostatud).

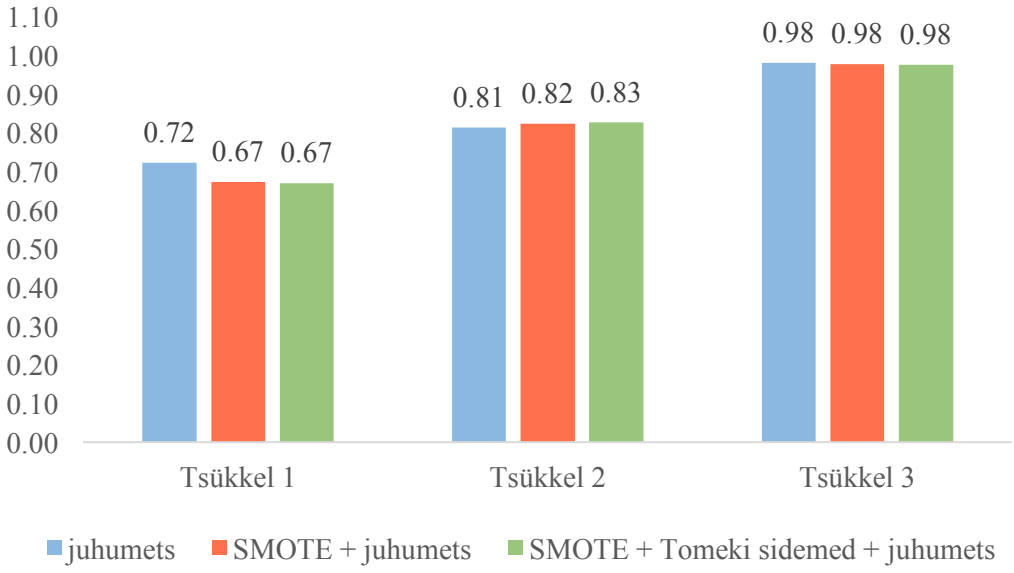
Järgnevalt annab autor kokkuvõtvalt ülevaate mudelite F1–skoorist kolme tsükli vältel (vt joonis 12). Nagu näha paranes kolme tsükli jooksul F1–skoor ligi 7 korda ümbernäidistamismeetodite rakendamisel ning ligi 25 korda vaid juhumetsa rakendamisel. Kolme tsükli jooksul paranes kõige rohkem SMOTE ning Tomeki sidemete kombineerimisel treenitud mudel, mis saavutas lõpuks ka kõige kõrgema F1–skoori (0,77). Kõige suurem paranemine antud tulemusmõõdiku lõikes oli teises tsüklis, kus paranemine oli peamiselt tingitud kahest asjaolust – andmete empiiriline eeltöötlus ning tunnuste arvu kahekordistamine. Andmete empiirilisel eeltöötlemisel vähendati küll kahtlaste juhtumite arvu, ent täiendavate tunnuste lisamine kompenseeris selle ning mudeli sooritus paranes oluliselt. Kõige parema sooritusega mudel siis, kui teises tsüklis püstitatud mudeli ning igapäevase agentidepoolse monitoorimise tulemusena leiti juurde rahapesu kahtlusega kliente, mille põhjal mudel treenida. Peale selle võib öelda, et mudeli sooritust parandas ka maksete käitumuslike tunnuste esitamine osakaaludena. Lähtudes F1–skoorist on mudeli treenimine tsüklites põhjendatud, sest sooritus paranes oluliselt ning viimases tsüklis jõuti mudelini, mille abil on võimalik efektiivselt rahapesu tuvastada. Vaatamata sellele, et mudeli F1–skoor oli vaid juhumetsa rakendamisel ligilähedalselt sarnane teiste mudeli omadele, tasuks ekstremaalselt tasakaalust väljas

andmetelt õppimisel eelistada ümbernäidistatud meetodite rakendamist, sest tagavad mudelile suurema stabiilsuse ning lõpuks ka parema soorituse.



Joonis 12. Mudelite F1–skoor kolme tsükli vältel (autori koostatud).

Lisaks F1–skoorile võrdleb autor kolme tsükli jooksul treenitud mudeleid ka AUC näitaja lõikes (vt joonis 13). Viimases tsüklis oli igal mudelil AUC võrdne ning teises tsüklis ligilähedaselt võrdne. Võrreldes esimese tsükliga paranes kõige rohkem erinevate ümbernäidistamismeetodite rakendamisel treenitud mudeli AUC näitaja.



Joonis 13. Mudelite AUC näitaja kolme tsükli vältel (autori koostatud).

Tuginedes nii F1–skoorile kui ka AUC näitajale võib tuua välja, et rahapesu tuvastada võimaldava mudeli treenimine on efektiivne tsüklites – kui esimeses tsüklis treenitud mudelit ei olnud võimalik kasutada, siis viimases tsüklis jõuti mudelini, mida on võimalik efektiivselt kasutada rahapesu tuvastamiseks. Tsüklite käigus ilmnes, et püstitades mudelit ekstremaalselt tasakaalust väljas andmetele tuleks kasutada erinevaid ümbernäidistamismeetodeid, mis parandasid mudeli sooritust oluliselt iga tulemusmõõdiku lõikes. Kahe tsükli vältel tagas kõige parema soorituse juhumetsa kombineerimine SMOTE–ga ning viimases tsüklis, kus tunnuseid ning vähemusklassi vaatlusi oli kõige rohkem, tagas parima soorituse juhumetsa, SMOTE ning Tomeki sidemete kombineerimine. Seega võib öelda, et nii SMOTE kui ka Tomeki sidemete rakendamine on efektiivsed ning tagavad mudelile parima soorituse ka siis, kui andmetes esineb teatavat müra, nagu näiteks esimeses tsüklis. Peale üldise soorituse paranemise tagab ümbernäidistamismeetodite rakendamine ka stabiilsema soorituse, st iga iteratsiooni parima ja halvima tulemusmõõdiku erinevus on võrdlemisi väike ning tõenäosuspiiride liigutamisel muutub F1–skoor oluliselt sujuvamalt, kui vaid juhumetsa rakendamisel.

KOKKUVÕTE

Rahapesul on ulatuslikud tagajärjed nii majanduse stabiilsusele, ühiskonna turvalisusele kui ka inimeste heaolule üldiselt, sest peamiselt pärineb kuritegelikul teel omandatud raha uimastikaubandusest, relvade müügist, prostitutsiooni vahendamisest või organiseeritud kuritegevusest. Rahapesuga on väga lähedaselt seotud ka terrorismi finantseerimine – mõlema protsessi eesmärk on varjata varade algupära, kasutades selleks samu meetodeid. Tehnoloogia areng ning Interneti levik on loonud üha uusi võimalusi, kuidas kuritegelikul viisil saadud vahendeid õiguspärasteks konverteerida ning paljud rahapesuga seotud riskid on elimineeritud või nende osatähtsust vähendatud. Virtuaalkeskkonnas finantsteenusi osutavaid ettevõtteid kasutades on aina kergem jääda anonüümseks ning sellest tulenevalt on oluline ettevõttetasandil implementeerida sobivaid kontrollmehhanisme, et takistada rahapesu ja terrorismi finantseerimist ning vältida riigivõimude poolt tehtavaid üüratuid trahve, mis majandust destabiliseerivad.

Käesoleva bakalaureusetöö raames püstitati masinõppe meetodeid ning rahvusvahelisi rahaülekandeid osutava ettevõtte TransferWise LTD andmeid kasutades mudel, mis suudab tuvastada rahapesu kahtlusega kliente. Selleks anti teoreetilises osas ülevaade rahapesu olemusest, aktuaalsusest ning kurjategijate poolt kasutatavatest meetoditest. Analüüsides erinevaid masinõppe meetodeid valiti argumenteeritult välja kolm mudelit, mida käesoleva töö raames rakendati. Varasemale teaduskirjandusele tuginedes toodi teoreetilises osas ka välja rahapesu tuvastada võimaldavaid tunnuseid.

Töö empiirilises osas esitas autor raamistiku rahapesu tuvastada võimaldava mudeli püstitamiseks ning rakendas seda kolme tsükli vältel. Autori poolt välja pakutud raamistik osutus käesoleva töö raames igati efektiivseks – mudelite sooritus paranes kolme tsükli vältel mitmekordselt. Mudeli soorituse paranemine oli eelkõige tingitud sellest, et treenides mudelit tsüklites sai võimalikuks koguda nii empiirilist kui ka matemaatilist tagasisidet, millest lähtuvalt mudelit parandati. Empiiriline tagasiside tuli kasuks nii

andmete puhastamisel kui ka mudelisse täiendavate tunnuste lisamisel. Kolmandas tsüklis ilmnes, et peale tagasiside kogumise võimaldab mudeli treenimine tsüklites koguda ka täiendavalt rahapesu kahtlusega vaatlusi – varasemates tsüklites püstitatud mudelid olid leidnud uusi vaatlusi, mida analüüsi kaasata.

Mudeli püstitamisel kasutas autor nii makseid, saajaid kui ka profiile kirjeldavaid tunnuseid ning soorituse hindamiseks rakendas 5x2 ristkontrollimist, mis oli antud uurimisülesande puhul igati asjakohane, sest võimaldas lisaks hinnata ka mudeli stabiilsust. Iteratsioonide vältel võimalikult vähe muutuva sooritusega mudel on oluline selleks, et see oleks rakendatav ka reaalsetel andmetel, mis võivad samuti olla heterogeensed. Antud töö puhul selgus, et üldjuhul parandab erinevate ümbernäidistamismeetodite rakendamine mudeli stabiilsust – näiteks esimeses tsüklis oli vaid juhumetsa rakendamisel täpsus viie iteratsiooni vältel minimaalselt 0 kuid maksimaalselt 1, seevastu kombineerides samas tsüklis juhumetsa SMOTE ja Tomeki sidemetega oli parima ja halvima sooritusе erinevus viis korda väiksem. Seega tuleks rahapesu tuvastada võimaldava mudeli sooritusе hindamiseks rakendada 5x2 ristkontrollimist.

Rahapesu efektiivselt tuvastada võimaldava mudeli püstitamiseks kasutas autor nii tundlikkust, täpsust, F1–skoori kui ka AUC näitajat. Rahapesu tuvastamisel on väga oluline minimeerida I liiki vigu fikseerides samal ajal II liiki vigade arvu mingile vastuvõetavale tasemele. Antud tulemusmõõdikute lõikes tagas kahe tsükli vältel kõige parema sooritusе juhumetsa ning SMOTE kombineerimine ning viimases tsüklis SMOTE, Tomeki sidemete ning juhumetsa kombineerimine. Lisaks sellele hindas autor ka mudeli erinevate konfiguratsioonide mõju sooritusele ehk F1–skoori ning tõenäosuspiiri seost. Ilmnes, et iga tsükli vältel alahindas juhumets vaatlusi – mida madalamale tõenäosuspiir seati, seda paremaks muutus F1–skoor. Teisisõnu – mudel ei olnud kunagi võimeline ütlema täie kindlusega, kas klient on rahapesu kahtlusega või mitte. Sellest lähtudes võib öelda, et õppides ekstremaalselt tasakaalus väljas andmetelt on oluline rakendada erinevaid statistilisi meetodeid andmete eeltöötlemiseks.

Bakalaureusetöö algne eesmärk sai täidetud – autori poolt püstitatud parim mudel leidis testandmetel rakendades üles 65% rahapesu kahtlusega klientidest, tekitades vaid 6% II liiki vigu. Vaatamata sellele, et mudel ei suutnud tuvastada ~35% rahapesu kahtlusega

klientidest tuleb nentida, et masinõppe mudeli rakendamine on kõigest tugisüsteem rahapesuga igapäevaselt tegelevatele agentidele, mitte ei asenda nende tööd ning sellest tulenevalt on püstitatud mudel igati efektiivne.

Mahulisest piirangust tingituna ei olnud käesoleva töö raames võimalik lähemat tähelepanu pöörata empiirilisele tagasisidele ja selle olulisusele ning sellest tulenevalt on üheks töö edasiarendamise võimaluseks viia läbi kvalitatiivne uuring selle kohta, mille alusel rahapesu tõkestamisega tegelevad agendid hindavad klientide puhul nende rahapesu kahtlust. Lisaks näeb autor töö edasiarendamise võimalusena sama raamistiku katsetamist erinevate masinõppe ja ümbernäidistamise meetoditega.

VIIDATUD ALLIKAD

1. Altmann, A., Tolosi, L, Sander, O., Lengauer T. Permutation importance: a corrected feature importance measure. – Bioinformatics, Oxford Journals, 2010, Vol. 26, No. 10, pp. 1340–1347. DOI: 10.1093/bioinformatics/btq134

2. Bagella, M., Busato, F., Argentiero, A. Using dynamic macroeconomics for estimating money laundering: a simulation for the EU, Italy and the United States. – Research Handbook on Money Laundering, 2013, pp. 207–223.

3. Barret, D., Perez, E. HSBC to pay record U.S. penalty. – The Wall Street Journal (U.S. Edition), 2011, URL: http://www.wsj.com/articles/SB10001424127887324478304578171650887467568

4. Batista, G. E. A. P. A., Bazzan, C. L. A., Monard, C. M. Balancing Training Data for Automated Annotation of Keywords: a Case Study. 2003, 9 p. URL: http://www.icmc.usp.br/~gbatista/files/wob2003.pdf

5. Batista, G. E. A. P. A., Prati, C. R., Monard, C. M. A Study of Behavior of Several Methods for Balancing Machine Learning Training Data. – ACM SIGKDD Explorations – Special issue on learning from imbalanced datasets, 2004, Vol. 6, No. 1, pp. 20-29, URL: http://sci2s.ugr.es/keel/dataset/includes/catImbFiles/2004-Batista-SIGKDD.pdf

6. Bolton, R.J., Hand, D. J. Statistical Fraud Detection: A Review. – Statistical Science. 2002, Vol. 17, No. 3, pp. 235–249, URL: http://projecteuclid.org-/download/pdf_1/euclid.ss/1042727940

7. Breiman, L. Random Forests. – Machine Learning, 2001, Vol 45, pp. 5–32, URL: https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf

8. Buchanan, B. Money laundering—a global obstacle. – Research in International Business and Finance, 2004, Vol. 8, pp. 115–127, DOI: 10.1016/j.ribaf.2004.02.001

9. Catal, C., Oral, A., Balkan, K. Class noise detection based on software metrics and ROC curves. – Information Sciences, 2011, Vol. 181, pp. 4867–4877, DOI: 10.1016/j.ins.2011.06.017

10. Chen, C., Liaw, A., Breiman, L. Using Random Forest to Learn Imbalanced Data. – Journal of Machine Learning Research, Department of Statistics, University of California, Berkeley, 2004, No. 666, pp. 1–12, URL: http://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf

11. Dietterich, T. G. Approximate statistical tests for comparing supervised classification learning algorithms. – Neural Computation, 1998, Vol 10, No. 7, pp. 1895–1923, URL: http://web.cs.iastate.edu/~honavar/dietterich98approximate.pdf

12. Drezewski, R., Sepielak, J., Filipkowski, W. The application of social network analysis algorithms in a system supporting money laundering detection. – Information Sciences, 2015, Vol 295, pp. 18–32, DOI: 10.1016/j.ins.2014.10.015

13. FATF – International Standards on Combating Money Laundering and the Financing of Terrorism & Proliferation, 2012, 130 p. URL: http://www.fatf-gafi.org/-media/fatf/documents/recommendations/pdfs/FATF_Recommendations.pdf

14. Fawcett, T. ROC Graphs: Notes and Practical Considerations for Data Mining Researchers, 2003, Technical Report, HP Laboratories, pp. 1–28, URL: http://www.hpl.hp.com/techreports/2003/HPL-2003-4.pdf

15. Gao, S., Xu, D. Conceptual modeling and developments of an intelligent agent-assisted decision support system for anti-money laundering. – Expert Systems with Applications, 2009, Vol. 36, pp. 1493–1504, URL: https://www.semanticscholar.org/paper/-Conceptual-modeling-and-development-of-an-Gao-Xu/d08ae505c53b31b5eff5cb7ab42a2d92d70fa3c5/pdf

16. Gilmour, N. Understanding the practices behind money laundering – A rational choice interpretation. – International Journal of Law, Crime and Justice, 2015, Vol. 44, pp. 1–13, URL: http://modirprozhe.ir/wp-content/uploads/edd/2016/04/Understanding-the-practices-behind-money-laundering.pdf

17. Han, J., Kamber, M., Pei, J. Data Mining Concepts and Techniques: Third Edition, 2012, 740 p.

18. He, P. A typological study on money laundering. – Journal of Money Laundering Control, 2010, Vol. 13, No. 1, pp. 15–32, DOI: http://dx.doi.org-/10.1108/13685201011010182

19. He, H., Garcia, E. A. Learning from Imbalanced Data. – Institute for Electrical and Electronics Engineers (IEEE) Transactions on Knowledge and Data Engineering,

2009, Vol. 21, No. 9, pp. 1263–1284, URL: http://www.cs.utah.edu-/~piyush/teaching/ImbalancedLearning.pdf

20. Irwin, S. M. A., Slay, J., Choo, R. K. K. Money laundering and terrorism financing in virtual environments: a feasibility study. – Journal of Money Laundering Control, 2014, Vol. 17, No. 1, pp. 50–75, DOI: http://dx.doi.org/10.1108/JMLC-06-2013-0019

21. Irwin, M. S. A., Choo, R. K., Liu, L. An analysis of money laundering and terrorism financing typologies. – Journal of Money Laundering Control, 2011, Vol 15, No. 1, pp. 85–111, DOI: http://dx.doi.org/10.1108/13685201211194745

22. Kuhn, M., Johnson K. Applied Predictive Modeling. – Springer Science + Business Media, New York, 574 p.

23. Le-Khac, A. N., Kechadi, M. T. Application of Data Mining for Anti-Money Laundering Detection: A Case Study. – International Conference on Data Mining Workshops, 2010, pp. 577–584, DOI: 10.1109/ICDMW.2010.66

24. Lopez-Rojas, A, E., Axelsson, S. Multi Agent Based Simulation (MABS) of Financial Transactions for Anti Money Laundering (AML). – Nordic Conference on Secure IT Systems, Blekinge Institute of Technology, 2012, pp. 1–8, URL: https://www.researchgate.net/profile/Edgar_Alonso_Lopez-Rojas/publication/232806257_Multi_Agent_Based_Simulation_(MABS)_of_Financial_Transactions_for_Anti_Money_Laundering_(AML)/links/0fcfd5097b30c54354000000.pdf

25. Mar, W. K., Naing, T. T. Optimum Neural Network Architecture for Precipitation Prediction of Myanmar. – International Journal of Environmental, Chemical, Ecological, Geological and Geophysical Engineering, 2008, Vol. 2, No. 12, pp. 154–158, URL: http://www.waset.org/publications/13861

26. Mathers, C. Terrorists Used Prepaid Cards to Finance Preparations For Paris Attacks. – Pulse on LinkedIn, 2016, URL: https://www.linkedin.com/pulse/terrorists-used-prepaid-cards-finance-preparations-paris-mathers

27. Moustafa, H. T., El-Megied, A. Z. M., Sobh, S. T., Shafea, M. K. Anti money laundering using a two-phase system. – Journal of Money Laundering Control, 2015, Vol 18, No. 3, pp. 304–329, DOI: http://dx.doi.org/10.1108/JMLC-05-2014-0015

28. Naheen, A. M. Money Laundering using investment companies. – Journal of Money Laundering Control, 2015, Vol. 18, No. 4, pp 438–446, DOI: http://dx.doi.org/10.1108/JMLC-10-2014-0031

29. Ngai, E. W. T., Yong, H., Wong, Y. H., Sun, X. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. – Decision Support Systems, 2011, Vol. 50, pp 559–569, DOI: 10.1016/j.dss.2010.08.006

30. Palmer, C. A picture of terrorist financing. – AML Newsletter, 2005 December, pp 14–16, URL: http://www.afma.com.au/afmawr/_assets/main/lib90059/2005-_december_aml%20magazine.pdf

31. Shatnawi, R., Li, W., Swain, J., Newman, T. Finding software metrics threshold values using ROC curves. – Journal of Software Maintenance and Evolution: Research and Practice, 2010, Vol. 22, pp 1–16, DOI: 10.1002/smr.404

32. Shen, A., Tong, R., Deng, Y. Application of Classification Models on Credit Card Fraud Detection. – International Conference on Service Systems and Service Management, 2007, pp. 1–4, DOI: 10.1109/ICSSSM.2007.4280163

33. Sheng, V. S., Gu, B., Fang, W., Wu, J. Cost-sensitive learning for defect escalation. – Knowledge-Based Systems, 2014, Vol. 66, pp. 146–155, DOI: 10.1016/j.knosys.2014.04.033

34. Simser, J. Money laundering: emerging threats and trends. – Journal of Money Laundering Control, 2013, Vol. 16, No. 1, pp. 41–54, DOI: http://dx.doi.org/10.1108/13685201311286841

35. Sullivan, K. Methods of Money Laundering. – Anti-Money Laundering in a Nutshell, 2015, pp. 15–35, DOI: 10.1007/978-1-4302-6161-2_2

36. The 9/11 commission report. – National Commission on Terrorist Attacks upon the United State, New York, NY, 2004, 585 p.

37. Tsang, S., Koh, S. Y., Dobbie, G., Alam, S. Detecting online auction shilling frauds using supervised learning. – Expert Systems with Applications, 2014, Vol. 41, pp. 3027–3040, DOI: 10.1016/j.eswa.2013.10.033

38. The Money Laundering Regulations, Financial Services, 2007, 50p. URL : http://www.legislation.gov.uk/uksi/2007/2157/pdfs/uksi_20072157_en.pdf

39. Unger, B., Hertog, J. Water always finds its way: Identifying new forms of money laundering. – Crime, Law and Social Change, 2012, pp. 287–304, DOI: 10.1007/s10611-011-9352-z

40. Vanitha, A., Niraimathi, S. Conversion of Imbalanced Data Into A Stream Using SMOTE Algorithm. – International Journal of Advanced Research in Computer and Communication Engineering, 2013, Vol. 2, No. 9, pp. 3640–3644, URL: http://ijarcce.com/upload/2013/september/61-h-vanitha_anandh_-conversion_of_imbalanced_data_into_a_stream.pdf

41. Wang, S. U., Yang, J. G. A Money Laundering Risk Evaluation Method Based on Decision Tree. – Machine Learning and Cybernetics International Conference, Hong Kong, 2007, pp. 283–286, DOI: 10.1109/ICMLC.2007.4370155

42. West, J., Bhattacharya, M. Intelligent financial fraud detection: A comprehensive review. – Computers & Security, 2015, Vol. 57, pp. 47–66, DOI: 10.1016/j.cose.2015.09.005

43. World Bank and International Monetary Fund. Anti Money Laundering and Combating the Financing of Terrorism. – World Bank and IMF Global Dialogue Series, 2003, 38 p, URL: http://siteresources.worldbank.org/INTAML/265197-1135187891284/20766146/GPD_Booklet_sar02_102402.pdf

44. Zdanowics, S. J. Detecting Money Laundering and Terrorist Financing via Data Mining. – Communications of the ACM, 2004, Vol. 47, No. 5, pp. 53–55, URL: http://www2.econ.uu.nl/users/unger/papers/Zdanowicz%202.pdf

45. Zeldin, M. Money Laundering. – Journal of Money Laundering Control, 1998, Vol. 1, No. 4, pp. 295–302, DOI: http://dx.doi.org/10.1108/eb027152

46. Zhang, D., Zhou, L. Discovering Golden Nuggets: Data Mining in Financial Application. – IEEE Transaction on Systems, Man, and Cybernetics – Part C: Applications and Reviews, 2004, Vol. 34, No. 4, pp. 513–522, DOI: 10.1109/TSMCC.2004.829279

47. Zhu, X. Semi-Supervised Learning Literature Survey. – Computer Sciences TR 1530 University of Wisconsin – Madison, 2008, 60 p, URL: http://pages.cs.wisc.edu/~jerryzhu/pub/ssl_survey.pdf

SUMMARY

DETECTING MONEY LAUNDERING USING MACHINE LEARNING METHODS BASED ON A CASE STUDY OF TRANSFERWISE LTD

Krister Jaanhold

Due to the development of technology and the exponential growth of Internet distribution, there have arised more and more opportunities for criminals to convert their illicit funds to legitimate. According to Buchanan (2004: 117) money laundering is a financial crime which may be in form of disguising the origin of criminally acquired funds or financing international terrorism. Illicit funds mainly originate from drug trafficking, the sale of weapons, mediation of prostitution or any other form of organised crime, hence it affects people's everyday life, public safety and even the economic stability. Driven by the scale of the problem, the focus of this Bachelor's thesis was on detecting money laundering and international terrorism financing using machine learning and data mining methods.

The increased competition between financial institutions has established a perfect environment for money laundering – many transaction costs and risks have been eliminated or their intensity has been reduced and it's becoming increasingly easier to stay anonymous. The 9/11 commission report pointed out, that wire transfers were one of the main methods for Al-Qaeda to finance the terrorist attacks, therefore it is important to implement suitable control mechanisms to prevent money laundering and international terrorism financing. Detecting money laundering on the corporate level is a complex task, because unlike other forms of fraud, it does not result in any financial loss and hence it's really hard to distinguish, whether someone was involved in money laundering schemes or not. Even though money laundering does not result in financial loss for the company, it may have extensive effects on the result of reputational decline, losing the licence or partners and paying enormous penalties, which in turn may destabilise the economy.

According to Zdanowics (2004: 53) the terrorist attacks on 11. September 2001 mark a new era of detecting money laundering and international terrorism financing – data mining and computer science have become essential tools. Furthermore, Zhang and Zhou (2004: 513) have pointed out, that data mining is a promising solution for detecting non-linear and dynamic connections and therefore may offer a solution to problems that arise from the abundance of data and dynamism of money laundering.

The purpose of this Bachelor's thesis was to develop a model using machine learning methods, that is able to detect customers with a suspicion of money laundering. For creating the model, data from a company which offers international transfers, TransferWise LTD, was used. In the theoretical part the autor gave an overview of the nature and topicality of money laundering and the characteristics that refer to it. In the second subchapter an overview of machine learning methods suitable for detecting money laundering was provided. In the empirical part, the author gave an overview of research methodology and described the initial data. The author proposed a framework for establishing the model, which was used throughout three cycles, during which three different combinations of machine learning models were trained. The proposed framework turned out to be efficient – the performance of different models improved significantly. Training the model in cycles provided an opportunity to gather both empirical and mathematical feedback, which were used to elaborate the model. Empirical feedback from anti-money laundering agents was used both for cleaning the data and gathering additional features to use in the model. For evaluating the models, the author used 5x2 crossvalidation recommended by Dietterich (1998: 1905) and it was confirmed to be essential, when training a model on extremely unbalanced data – in some cases the performance on each of the subsets differed significantly from the average.

Applying different resampling methods to balance the data improved the model's performance significantly regarding F1–score and AUC metric. For detecting money laundering, it is really important to minimise type I error, keeping type II errors at an acceptable level. The best performance during the first two cycles was achieved combining SMOTE and random forest. In the third cycle applying SMOTE, removing Tomek links and training the random forest ensured the best results. To establish a model on extremely unbalanced data, it is really important to evaluate the relationship between

F1–score and cutoff – it appeared, that through each cycle, using only random forest, the model was underestimating. In other words – the model was never confident enough for classifying minority class observations.

The initial purpose of this Bachelor's thesis was accomplished – the final model was able to detect 65% of customers with money laundering suspicion, causing only 6% of false positives. Even though the model couldn't detect ~35% of customers with money laundering suspicion, it has to be stated, that machine learning model is applied as a support system for anti money laundering agents and doesn't substitute their work, thus the model in general is beneficial.

As a next step the author sees conducting a qualitative research on the empirical feedback provided by anti money laundering agents – how they evaluate risk and what are the main features based on what they override suspicion. Besides that, the proposed framework could be used for applying and evaluating other machine learning and data mining methods for detecting money laundering.

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Krister jaanhold,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „Rahapesu tuvastamine masinõppe meetodite abil TransferWise LTD näitel“

mille juhendaja on Oliver Lukason,

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, **24.05.2016**