TARTU ÜLIKOOL
Arvutiteaduse instituut
Andmeteaduse õppekava

Kerdo Puusalu

RAHAPESUKAHTLUSEGA TEHINGUTE TUVASTAMINE JUHENDAMATA VARJATUD MARKOVI AHELATE ABIL

Magistritöö (15 EAP)

Juhendaja: Kaur Lumiste, PhD

Tartu 2022

RAHAPESUKAHTLUSEGA TEHINGUTE TUVASTAMINE JUHENDAMATA VARJATUD MARKOVI AHELATE ABIL

Lühikokkuvõte:

Magistritöö sisaldab endas meetodit rahapesukahtlusega tehingute tuvastamiseks, kasutades varjatud Markovi mudelit ja *DBSCAN* (*Density-based spatial clustering of applications with noise*) juhendamata masinõppealgoritmi. Töö eesmärgiks on asendada praktikas laialdaselt levinud reeglipõhiste rahapesu monitooringusüsteemide aluseks olevad kontroll-laused juhendamata masinõppealgoritmi *DBSCAN* klastritega. Leitud klastreid kasutatakse käesoleva töö raames rahapesu riskiskoorina, mis on varjatud Markovi mudeli vaadeldavaks kihiks.

Võtmesõnad:

Rahapesu, varjatud Markovi mudel (*HMM*), *DBSCAN*, reeglipõhine rahapesu monitooringusüsteem

CERCS: P170 Arvutiteadus, arvanalüüs, süsteemid, kontroll

SUSPECTED MONEY LAUNDERING DETECTION USING UNSUPERVISED HIDDEN MARKOV CHAINS

Abstract:

This master's thesis contains a method for detecting transactions with money laundering suspicion, using a hidden Markov model and DBSCAN (Density-based spatial clustering of applications with noise) unsupervised machine learning algorithm. The aim of the work is to replace widely used rule-based money laundering systems risk score with clusters of the DBSCAN algorithm. The clusters found are used as the money laundering risk score, which will be hidden Markov´s model observable layer.

Keywords:

Money laundering, hidden Markov model (HMM), DBSCAN, rule-based money laundering monitoring system

CERCS: P170 Computer science, numerical analysis, systems, control

Sisukord

Sissejuhatus

Kuritegevus kujutab endas suurt ohtu finantssüsteemi ja majanduskeskkonna stabiilsusele ning kuvandile. Lebid *et al.* [1] toovad välja, et kuritegevusest saadud varaga kerkib esile vajadus seda seaduslikuna näidata, sest vara omanikul eksisteerib soov ilma karistuseta vara realiseerida ja näiteks kinnisvara osta. Sellist vara allika peitmist nimetatakse rahapesuks. USA-s kriminaliseeriti rahapesu aastal 1986, kui jõustus *The Money Laundering Control Act*, mille peamiseks eesmärgiks oli vähendada kuritegelike organisatsioonide narkootikumide müügist saadud tulu ning seeläbi vähendada nende motivatsiooni nimetatud valdkonnas tegutseda [2]. Kui puuduks võimalus kuritegevusest teenitud varasid kasutada, puuduks kurjategijatel ratsionaalne põhjus kuritegevusega tegeleda.

Rahapesu ja selle tõkestamine on aina enam sattunud Eesti üldsuse huviorbiiti. Näiteks sai Swedbank AB perioodi 2015-2019 rahapesu vastu võitlemise reeglite rikkumise eest trahvi neli miljardit Rootsi krooni [3], AS SEB Pank sai perioodi 2017-2019 rahapesu tõkestamise reeglite rikkumise eest trahvi miljon eurot [4] ning lisaks suunas Eesti Finantsinspektsioon AS LHV Panga [5] ja Luminor AS-i [6] parendama rahapesu tõkestamise kontrollisüsteeme. Kõik nimetatud pangad on 30.06.2021 seisuga Eestis bilansimahu [7] poolest neli suurimat panka ja seejuures tuleb tõdeda, et nimetatud uudised pärinevad aastatest 2020-2021.

Kute *et al.* [8] hinnangul ilmestavad määratud trahvid hetkeolukorda, kus olemasolevad rahapesu tuvastussüsteemid pole asjakohased ja efektiivsed võitlemaks rahapesuga, kuigi finantsasutustel, kui kohustatud isikutel, on regulatiivne surve pidevalt oma süsteeme ja meetmeid ajakohastada ja viia vastavusse. Näiteks tutvustas Euroopa Komisjon 2021. aasta juulis uut seadusandlike ettepanekute paketti (*6th Anti Money Laundering Directive - AMLD 6*), et tugevdada ELi rahapesu ja terrorismi rahastamise tõkestamise eeskirju [9]. Rahapesu tuvastussüsteemide arvustuste kõrval on kriitika alla sattunud ka rahapesuregulatsioonid. Näiteks koostasid Tiwari *et al.* [2] ülevaate artiklitest, kus on kirjeldatud regulatsioonide ebaefektiivsust, mittevastavust reaalsele praktikale ja regulaatorite ebaoperatiivsust. Suur hulk tehinguid ning ootus kiirele raha liikumisele, seab kohustatud isikud surve alla, kus rahapesukahtlusega tehingute tuvastamine peab olema tasakaalus täpsuse ja rahapesuvastase võitluse protsessi kiirusega [10].

Rahapesu on „Eesti rahapesu ja terrorismi rahastamise tõkestamise seaduse“ § 4. tuginedes defineeritav kui kuritegelikust tegevusest saadud vara või selle asemel saadud vara muundamine, üleandmine, omandamine, valdamine, kasutamine või tõelise olemuse, päritolu, asukoha, käsutamisviisi, ümberpaigutamise, omandiõiguse või muude õiguste varjamine [11].

Selge ja lihtsa definitsiooni on andnud rahapesule Tiwari *et al.* [2], kes defineerib rahapesu kui protsessi, mille käigus üritatakse kuritegelikul teel saadud rahale omistada legitiimset kuvandit.

Rahapesu ja terrorismi rahastamise tõkestamise seaduses on määratletud kohustatud isikud, kes peavad läbi viima mitmeid tegevusi, et tõkestada rahapesu ja terrorismi rahastamist. Eesti Finantsinspektsiooni (FI) rahapesu ja terrorismi rahastamise tõkestamise järelevalvepoliitikas on välja toodud olulised aspektid, mida FI kohustatud isikutelt järelevalves ootab. FI järelevalve alla kuuluvad näiteks krediidiasutused, finantseerimisasutused, kinnisasjade vahendajad, hasartmängude korraldajad jne [11]. Käesoleva töö raames keskendutakse finantsinstitutsioonidele. Üheks nimetatud tegevuseks järelevalvepoliitikas on hoolsuskohustuse täitmine, mille kohaselt peab kohustatud isik tundma enda klienti ning kliendi poolt teostatavaid tehinguid. Finantsinstitutsioonid on kohustatud pöörama kõrgendatud tähelepanu tegevustele, mis viitavad kuritegevusele, rahapesule või terrorismi rahastamisele, sealhulgas keerukatele, suure väärtusega ja ebatavalistele tehingutele ja tehingumustritele [12]. Käesolevas magistritöös keskendutakse tehingute monitooringule. Tehingute monitooringu eesmärk on tuvastada ebaharilike tunnusmärkidega tehingud, mis võivad viidata rahapesule.

Enamasti kasutavad finantsinstitutsioonid reeglipõhiseid tehingute monitooringusüsteeme rahapesukahtlusega tehingute tuvastamiseks [8, 13]. Reeglitel põhinevad monitooringusüsteemid sisaldavad endas enamasti rangeid eeldefineeritud piirmääradega kontroll-lauseid (nt tehingusumma > 10 tuh. €, viimase 3 päeva tehingupartnerite arv > eelmise kuu tehingupartnerite arv), mis reegli kohandamisel käivitavad edasise kontrolli vajaduse. Zhang *et al.* [14] kirjeldavad kõrgetasemeliselt rahapesukahtlusega tehingute tuvastamise protsessi järgnevalt: esmalt määrab tuvastussüsteem tehingule rahapesukahtluse riski, mis põhineb kindlaks määratud reeglitel; seejärel suunatakse suure rahapesukahtluse tehing finantsinstitutsiooni töötajatele analüüsimiseks, kus otsustatakse, kas tehingust (koos lisainfoga) raporteerida kohalikule järelevalveasutusele või mitte. Mitmed teadusartiklid aga kinnitavad, et reeglitel põhinevad infosüsteemid on ebaefektiivsed - umbes 90% rahapesukahtlaseks märgitud tehingutest on valepositiivsed [13, 15, 16] (isegi kuni 98% valepositiivseid [8]). Seega otsitakse tehingute monitooringus uusi lahendusi.

Varjatud Markovi mudel (ingl *Hidden Markov Model, HMM*), on aastatega kogunud populaarsust ning see on rakendust leidnud mitmetel elualadel. *HMM* põhineb Markovi ahela reeglil, kus iga seisund (varjatud) ahelas genereerib vastavalt tihedusfunktsioonile vaadeldava tunnuse [17]. Käesoleva magistritöö kontekstis soovitakse *HMM*-iga modelleerida vaadeldava

kihi ehk tehingu- ja kliendiprofiili tunnuste järjestuse põhjal varjatud seisundeid ehk rahapesukahtlust või selle puudumist. Töö eesmärk on edasi arendada Kasianova [18] tööd, kus kasutatakse kontroll-lausetel põhinevat rahapesumonitooringu meetodit ja *HMM*-i rahapesukahtlusega tehingute tuvastamiseks. Kontroll-lausetega leitud rahapesu riskiskoor on *HMM*-i vaadeldavaks kihiks ning ennustatavateks (varjatud) seisunditeks on rahapesukahtluse esinemine või selle puudumine. Käesolevas magistritöös asendatakse Kasianova [18] töös kasutatud reeglitel põhinev rahapesu riskiskoor juhendamata masinõppealgoritmiga *DBSCAN* (*Density-based spatial clustering of applications with noise*), mille klastreid kasutatakse *HMM*-i vaadeldava kihina.

Käesolev magistritöö koosneb kirjanduse ülevaatest, metodoloogiast, tulemuste tutvustamisest ja arutelust. Kirjanduse ülevaate peatükis tutvustatakse rahapesu taustinformatsiooni ja antakse ülevaade senistest rahapesu tuvastamise meetoditest. Metodoloogia peatükk sisaldab rahapesutehingute avastamiseks kasutatavate mudelite kirjeldust ning eksperimentide ülesehituse kirjeldust. Tulemuste peatükk annab ülevaate mudelite tulemustest headusmõõdikutel ning kokkuvõttes kajastub lühiülevaade tehtud tööst ja saadud tulemustest.

Töö praktilise osa läbiviimisel kasutati *Pythoni* programmeerimiskeelt (versioon 3.9) [19], *Pandas* [20] *Pythoni* teeki, *NumPy* [21] *Pythoni* teeki, *scikit-learn hmmlearn* [22] *Pythoni* teeki, *scikit-learn DBSCAN Pythoni* teeki [23], *scikit-learn StandardScaler* [24] *Pythoni* teeki ja *scikit-learn metrics* [25] *Pythoni* teeki.

1. Kirjanduse ülevaade

Käesolevas peatükis antakse ülevaade rahapesust ning rahapesuvastases võitluses kasutatavatest tehingumonitooringu meetoditest.

1.1 Sissejuhatus rahapesuvastasesse võitlusesse

Rahapesu on tegevuste hulk, millega soovitakse kuritegelikul teel saadud vara esitleda legaalsena [2, 26] ning seejärel kasutada seda legaalses majandustegevuses või rahastada edaspidist kuritegevust [13]. Eksisteerib väheseid kuritegusid, mille eesmärgiks pole varalise kasu saamine ning seetõttu pole võimalik seostada rahapesu kindla kuritegevuse tüübiga [26]. Pestavate vahendite allikaks võib olla narkootikumide müük, inimkaubandus, terrorismi rahastamine, korruptsioon ja paljud muud kuriteod, kust saadakse varalist kasu. Rahapesu takistab majanduskeskkonna arengut, kahjustab finantssüsteemi mainet ja motiveerib kuritegevuse teket [2][27]. Näiteks on korruptsioon üheks kuritegevusliigiks, millega on võimalik otseselt kahjustada turumajanduse konkurentsi.

Tiwari *et al.* [2] koostasid ülevaate kirjandusest, kus on proovitud hinnata pestud raha mahtu, aga tehtud töid on kritiseeritud ebatäpseks ja nimetatud eksitavaks, seda tulenevalt kuritegevuse mahu hindamise keerukusest ja teadmatusest, kui palju teatud kuritegevusliik tulu genereerib. Walkeri mudeliga hinnati 1995. aastal maailmas pestavaks raha mahuks 2,85 miljardit USD aastas [28], aga 2004. aastal hinnati ainuüksi Austraalias pestavaks raha mahuks 2,8-6,3 miljardit AUD ehk 2,07-4,62 miljardit USD (2004. aasta aritmeetiline keskmine AUD vahetuskurss 0,7362 [29]) [2].

Rahapesu kätkeb endas peamiselt kolme järgnevat protsessi [18, 30, 31]:

1. Paigutamine (ingl *placement*) ehk kuritegelikul teel saadud varade esialgsest allikast eemaldamine ning legaalsesse süsteemi kandmine;
2. Kihistamine (ingl *layering*) ehk legaalsel ja illegaalsel viisil teenitud varade ühildamine ja/või nendega tehingute tegemine;
3. Integratsioon (ingl *integration*) ehk kuritegelike vahendite kasutamine legaalses majanduses.

Rahapesu vastu võideldakse nii globaalsel kui ka kohalikul tasandil. Aastal 1989 loodi rahapesu ja terrorismi rahastamise vastu võitlemiseks rahvusvaheline *Financial Action Task Force* (*FATF*) [32]. *FATF* on poliitikat kujundav organ, mille ülesandeks on panna paika standardid rahapesu ja terrorismi rahastamise vastu võitlemise regulatsioonidele, neid regulatsioone propageerida ning anda ülevaade riikide kaupa seadusandluse hetkeseisust ja

meelsusest. Üle kahesaja jurisdiktsiooni maailmas järgib *FATF* ja *FATF-Style Regional Bodies* (*FSRB*) soovitusi [33]. Paljud riigid on küll kasutusele võtnud ühtse standardi, aga tänase päevani nähakse ühe suure puudusena piiriülest kommunikatsioon, mida kurjategijad kasutavad illegaalsete vahendite kihistamisel [34].

Erinevates riikides on kohalik regulatsioon, mis kohustab finantsinstitutsioone tõkestama rahapesu ja terrorismi rahastamist. Finantsinstitutsioonid on kohustatud koguma kliendi kohta andmeid, et rakendada "tunne oma klienti" põhimõtteid (ingl *know your customer,* KYC). Finantsinstitutsioonidel on kliendi kohta teada n-ö klienti kirjeldavad andmed ehk kliendihaldusprogrammi (ingl *customer relationship management, CRM*) andmed. Näiteks võib kliendihaldusprogramm sisaldada juriidilise isiku vaates andmeid ettevõtte nime, registrikoodi, tegelike kasusaajate, tegevusvaldkonna kohta. Kliendi poolt tehtavate tehingutega kaasnevad andmed nagu näiteks makse sihtriik, tehingu tüüp ja tehingu valuuta ehk tehinguandmed. Nimetatud andmeatribuute võib olla kliendi kohta sadades ning olemasolevatest andmetest on võimalik luua lisanduvaid tunnuseid (ingl *feature engineering*). Rahapesu monitooringusüsteemid kasutavad nimetatud andmeatribuute, et automaatselt tuvastada rahapesukahtlusega tehinguid, mis suunatakse edasisele analüüsile.

1.2 Rahapesuvastase võitluse takistused

Rahapesukahtlusega tehingute tuvastamiseks on välja pakutud erinevaid uusi meetodeid, aga tehtud teadustöödest kerkib esile mitmeid probleeme. Tundis *et al.* [35] koondasid rahapesukahtlusega tehingute tuvastamise teadustööde probleemid kolme järgnevasse kategooriasse:

1. andmestikud pole avalikult kättesaadavad, mistõttu on tulemuste reproduktsioon võimatu;
2. mudeli headuse hindamisel pole kasutatud mõõdikuid või on seda tehtud ebapiisavalt;
3. kasutatud tunnuseid pole formuleeritud.

Samas andmeid on – kuna finantsinstitutsioonid on kohustatud koguma kliendi kohta andmeid, siis on nende käsutuses sadu andmeatribuute, millest on võimalik genereerida lisanduvaid tunnuseid. Näiteks kasutasid Ketenci *et al.* [13] reaalsete tehinguandmete ajalise sageduse analüüsi (ingl *time-frequency analysis*) andmetest lisanduvate tunnuste genereerimiseks. Sageduste arvutamisel kasutati kolmekümne päeva pikkust ajaakent päevase sammuga aritmeetilise keskmise, dispersiooni, asümmeetriakordaja (ingl *skewness*), ekstsessi (ingl *kurtosis*), ajalise tiheduse (ingl *time sparsity*) arvutamisel – kokku kasutati ühteteist erinevat

mõõdikut uute tunnuste arvutamiseks. Alshantti *et al.* [27] kasutasid Norra DNB panga andmestikku, mis peale andmetöötlusmeetodite rakendamist koosnes 522-st andmeatribuudist. Jullum *et al.* [16] kasutasid samuti Norra DNB panga tehinguandmeid. Kasutatavate andmeatribuutide kohta on autorid välja toonud, et andmestik sisaldab taustinformatsiooni ja varasemate tehingute kokkuvõtet tehingu tegijast ja saajast ning kokkuvõtet kahtlasest tehingust ja varasematest kahtlastest tehingutest, millega osapool on seotud olnud. Lõplik andmestik, millele rakendati *XGBoost* masinõppealgoritmi, sisaldas 1100-t tunnust. Nimetatud töödest võib järeldada, et finantsinstitutsioonide käsutuses on palju andmeid, millele tuginedes on võimalik teostada tehingumonitooringut.

Siiski tulenevalt privaatsusnõuetest ei ole finantssektori andmestikud avalikult kättesaadavad ning teadustöödes kasutatakse enamasti tehislikult genereeritud andmestikke [18, 35, 36]. Tehislikult genereeritud andmestike kasutamisel puudub teadmine, kas mudelid ka praktikas häid tulemusi näitavad [27]. Lisanduvalt on Tundis *et al.* [35] viidanud asjaolule, et kuna andmestikud pole kättesaadavad, pole võimalik ka tulemuste reproduktsioon. Suurbritannia finantsinspektsiooni eestvedamisel viidi 2020. aasta juulis ja augustis läbi pilootprojekt, kus osales 120 organisatsiooni ning mille eesmärgiks oli sünteetilise finantsandmestiku loomine [37]. Käesoleva magistritöö koostamise hetkel puudub teadmine, kunas on võimalik andmestikku kasutama hakata.

Enamasti kasutavad finantsinstitutsioonid reeglipõhiseid monitooringusüsteeme rahapesukahtlusega tehingute tuvastamiseks [8, 13]. Kontroll-lausetel põhinev tehingute monitooringu süsteem sisaldab endas määratletud reeglite kogumit, mis põhineb valdkonnaekspertide teadmistel [27]. Reeglipõhised monitooringusüsteemid märgivad reegli rakendumise korral tehingu rahapesukahtlusega tehinguks, mis suunatakse edasi rahapesu tuvastamise spetsialistidele analüüsimiseks ning lisanduva informatsiooni hankimiseks [15, 27]. Kui analüüsitav tehing on ka manuaalse kontrolli tulemusel rahapesukahtlusega, edastatakse info ja kahtlus järelevalveasutusele kahtlase tegevuse raportina (ingl *Suspicious Activity Report*, *SAR*) [8, 15]. Käesolevas töös kasutatakse *SAR* tehinguid ja rahapesukahtlusega tehinguid sünonüümidena. Eestis tuleb rahapesukahtlusega isikutest teavitada Rahapesu Andmebürood (RAB).

Mitmed teadusartiklid viitavad probleemile, kus reeglitel põhinevad rahapesu monitooringusüsteemid on ebaefektiivsed [15, 16]] ning 90% rahapesukahtlusega tehingutest märgitakse vale-positiivseteks [13, 27]. Suur hulk vale-positiivseid tulemusi tekitab liigset müra ning erineva kogemuste pagasiga finantsasutuse töötajad peavad üle vaatama suurel

hulgal tehinguid [16], mis põhjustab finantsasutusele aja- ja finantskulu [13]. Ketenci *et al.* [13] väidavad, et suurel hulgal vale-positiivseid tehinguid üle kontrollinud töötajad ei suuda tuvastada reaalselt kahtlaseid tehinguid. Lisanduvalt toovad Kute *et al.* [8] välja, et ausate klientide raporteerimine kohalikule järelevalveasutusele võib päädida ebavajaliku kriminaaluurimise alustamisega ning kulutada järelevalveasutuse ressursse, mida saaks suunata reaalsete rahapesukahtlusega tehingute uurimisele. Seega kätkevad nii vale-positiivsed kui ka vale-negatiivsed tulemused rahapesukahtlusega tehingute tuvastamises riske.

Ketenci *et al.* [13] hinnangul on üheks reeglipõhise rahapesumonitooringu süsteemi puuduseks asjaolu, et reeglid on rahapesijatele teada, sest rahapesuvaldkonda reguleerivad seadused ja soovitused on avalikult kättesaadav informatsioon. Paljud rahapesumonitooringu reeglid on koostatud regulatsioonidest lähtuvalt, et täita kohaliku regulaatori poolt seatud kohustusi. Eestis on näiteks välja antud FI soovituslik juhend „Rahapesu ja terrorismi rahastamise tõkestamise meetmed krediidi- ja finantseerimisasutustes“, kus on välja toodud aspektid, mis peaksid kliendi rahapesu riskitaset mõjutama [38]. Reeglid võivad olla kurjategijatele teada ka tulenevalt finantsasutuse poolt kasutatavast infosüsteemist või võivad reeglid olla jõudnud rahapesijate valdusesse korrumpeerunud finantsasutuse töötaja abil. Teades lävendeid, millal reegel rakendub, on võimalik teostada tehinguid ning jääda monitooringureeglitele märkamatuks [13].

Rahapesijad leiavad pidevalt uusi võimalusi kuidas kuritegelikul teel saadud raha pesta. Reeglipõhiste süsteemide nõrkuseks on asjaolu, et uute reeglite defineerimine on aeglane ning enamasti tagajärgedega tegelev protsess [1, 13]. Jullum *et al.* [16] hinnangul on uusi reegleid võimalik rahapesujuhtumite analüüsiga leida, aga siiski jäävad reeglid liiga lihtsustatuks. Näiteks kasutas Kasianova [18] varjatud Markovi mudeli vaadeldava kihi leidmiseks reeglit, mis lisas rahapesu komposiitskoorile peale kella 21-t või enne kelle 7-t tehtud tehingu eest 15 lisapunkti (mis võis moodustada 20% koguskoorist, ptk 4.2). Nimetatud reeglit on võimalik lihtsal viisil vältida ja saavutada madalam skoor. Teisalt võidakse omistada ausatele klientidele, kellele on omane õhtune või varahommikune tehingute tegemine (nt välisturgudel kauplemine) ebavajalikult kõrge skoor. Tehingumonitooringu koosseisu kuuluv suur hulk kontroll-lauseid võib põhjustada palju vale-positiivseid tulemusi (üks tegevus võib põhjustada mitme reegli rakendumise), aga liialt lihtsustatud lähenemine võib päädida rahapesutehingute mitte avastamisega, mis omakorda võib kaasa tuua mainekahju, trahvid ning võimaldab kurjategijatel saadud vahendeid kasutada.

Chen *et al.* [39] toovad välja, et reeglipõhistes süsteemides pole võimalik määrata õigeid reegli rakendumise piirmäärasid (ingl *threshold*) ning nimetatud põhimõttel toimivad süsteemid ei suuda oma jäikuse tõttu toime tulla suures koguses struktureeritud, semi-struktureeritud või struktureerimata andmetega (nt kvalitatiivsete andmetega). Seetõttu peavad eelnevalt nimetatud autorid masinõpet sobilikumaks lahenduseks, sest masinõppealgoritmid suudavad teha andmetel põhinevaid üldistusi ning määravad sobilikud piirmäärad maksimeerides või minimeerides määratletud funktsiooni rahapesu tuvastamiseks.

Rahapesujuhtumite uurimine on aeganõudev protsess ning finantsinstitutsioon on harva teadlik, kas rahapesukahtlus osutus tõeseks või mitte, sest lõpliku otsuse teeb kohus [16]. Alshantti *et al.* [27] viitavad asjaolule, et hinnanguliselt moodustavad tuvastatud rahapesujuhtumid mitte rohkem kui 10% kogu pestavast rahast. Seetõttu võivad juhendatud masinõppemeetodid anda valesid tulemusi, sest õppimine toimub suure tõenäosusega ebakorrektselt märgendatud andmestikul. Lisanduvalt toovad Shokry *et al.* [10] ühe puudusena välja probleemi, et juhendatud masinõppemeetodid suudavad tuvastada rahapesukahtlusega tehinguid, mis kajastuvad vaid treeningandmestikus ehk mudel ei pruugi leida uusi mustreid. Tulenevalt asjaolust, et rahapesujuhtumid on muutlikud ning rahapesijad leiavad uusi mooduseid kuritegeliku raha pesemiseks, on juhendamata masinõppemeetodid sobilikumad [8, 36]. Siiski hindavad Gupta *et al.* [15], et masinõpet ja statistilisi mudeleid kasutavate süsteemide kõrval jäävad siiski püsima reeglitel põhinevad monitooringusüsteemid, sest need on laialdaselt levinud ning nende asendamine uute süsteemidega on aeganõudev ja kulukas tegevus.

Kuna paljude masinõppemeetodite tulemusi on keeruline või võimatu põhjendada, võib see luua segadust rahapesu tuvastamise spetsialistile, kes peab teostama edasist analüüsi, et lõplikult välja selgitada, kas tegemist on rahapesukahtlase tehinguga või mitte. Reeglitel põhineva rahapesu monitooringusüsteemi eeldefineeritud kontroll-laused võimaldavad koheselt suunata rahapesuspetsialisti tähelepanu tegevusele, mis reegli rakendumise põhjustas. Nimetatud probleemile on tähelepanu juhtinud Kute *et al.* [8], kes annavad ülevaate kasutatud masinõppemeetoditest rahapesus ning nende tõlgendatavusest (*Explainable Artificial Intelligence, XAI*). Autorid analüüsivad kokku 43 artiklit, mis on jaotatud järgnevatesse kategooriatesse: juhendatud masinõppemeetodid (65% analüüsitud artiklitest), juhendamata masinõppemeetodid (28% analüüsitud artiklitest), pool-juhendatud masinõppemeetodid (3% analüüsitud artiklitest), stiimulõpe (ingl *reinforcement learning*) (2% analüüsitud artiklitest) ja koosmõjus juhendatud ning juhendamata masinõpe (2% analüüsitud artiklitest). Autorite

hinnangul on koondatud artiklitest 51% meetoditest tõlgendamatud, 7% artiklitest jääb tõlgendatus ebaselgeks ning 42% on tõlgendatavad artiklid. Nimetatud töö autorid peavad tõlgendatavuse all silmas võimet mudeli poolt tehtud otsuseid selgitada või põhjendada. Nimetatud puudus masinõppemeetodites võib osutuda suureks takistuseks nende rakendamisel praktikas, sest rahapesukahtlust või selle puudumist võib olla vajalik põhjendada järelevalveasutusele või audiitorile.

1.3 Rahapesukahtlusega tehingute tuvastamise meetodid

Rahapesu tuvastamise meetodid võib jaotada kolme peamisesse gruppi:

- juhendatud masinõppemeetodid,
- juhendamata masinõppemeetodid ja
- muud statistilised meetodid.

Juhendamata masinõppemeetodid püüavad tuvastada andmetest mustreid ilma teadmiseta, kas tehing on rahapesukahtlusega või mitte. Juhendatud masinõppemeetodite kasutamisel antakse algoritmile ette märgendid, mille põhjal püüab algoritm leida treeningandmestikust mustreid, mis eristavad erinevate märgenditega vaatlusi üksteisest. [16] Muude statistiliste meetodite all on silmas peetud kombinatsioone erinevatest meetoditest ja meetodeid, mis ei sobi juhendatud või juhendamata masinõppemeetodite alla (nt visualiseerimistehnikad, geograafiline analüüs).

Rahapesukahtlusega tehingute tuvastamiseks on välja pakutud mitmeid meetodeid ning koostatud on mitmeid ülevaatlike artikleid kasutatavatest meetoditest [8, 39, 40, 41, 42, 43]. Salehi *et al.* [40] andsid ülevaate 25-st enda hinnangul olulisemast rahapesu tuvastamise artiklist perioodil 2005-2017. Autorid jagavad kasutatud meetodid seitsmesse erinevasse kategooriatesse:

- klasterdamine,
- AML (ingl *anti-money laundering)* tüpoloogiatel põhinevad meetodid,
- tehisnärvivõrgud,
- tugivektor-klassifitseerijad,
- otsustuspuud,
- sotsiaalvõrgustike analüüs ja
- muud meetodid.

Muude meetodite all peavad autorid silmas meetodeid, mis ei sobinud eelnevalt nimetatud kategooriatesse või on kombinatsioon erinevatest meetoditest. Kute *et al.* ülevaates [8] on

rahapesu tuvastamise meetodite all kasutatud AML tüpoloogiaid, seoste analüüsi, käitumise modelleerimist, riskide hindamist, anomaaliate tuvastamist ja geograafilist analüüsi.

Erinevaid ülevaatlike artikleid [8, 39, 40, 41, 42, 43] rahapesu avastamisest on palju, mistõttu käesolevas töös sellele suurt rõhku ei asetata ning tuuakse välja vaid mõned näited tehtud töödest. Välja toodud tööde eesmärk on anda ülevaade erinevate kasutatud meetodite tulemuste suurusjärkudest headusmõõdikutel. Siinkohal tuleb mainida, et tulemusi headusmõõdikutel pole otseselt võimalik omavahel võrrelda tulenevalt peatükis 1.2 mainitud põhjustest (nt erinevad andmestikud, erinev andmestike kompositsioon).

Ketenci *et al.* [13] rakendasid reaalsetele tehinguandmetele ajalise sageduse analüüsi (ingl *time-frequency analysis*) lisanduvate tunnuste leidmiseks ning seejärel kasutasid juhumetsa (ingl *Random Forest*) masinõppealgoritmi rahapesukahtlusega tehingute tuvastamiseks. Parim tulemus, milleks on 91,49% *ROC*-kõvera joonealusest pindalast (ingl *Area under the ROC Curve, AUC*), saavutatakse kasutades nii klienti kirjeldavaid andmeid, tehinguandmeid kui ka tehinguandmete ajalise sageduse andmeid. Nimetatud töö autorid tõdevad, et ajalise sageduse andmed ja tehinguandmed on suuresti korreleeruvad ning mudeli headusmõõdikute tulemusi suuresti ei paranda. Kõige enam aitavad kaasa rahapesukahtlusega tehingute tuvastamisele klienti kirjeldavad andmed.

Klienti kirjeldavad andmed (nt nimi, elukoht, isikukood) on suuresti muutumatud, aga kliendi tehinguandmed on seevastu pidevas muutuses. Klienti kirjeldavad andmed on andmestike liitmisel duplikaatidena (ühe kliendi erinevate tehinguandmete kõrval samad klienti kirjeldavad andmed), aga omavad tehinguandmete kõrval n-ö kliendiprofiili iseloomu. Kui isiku tehinguandmete tunnused pole omased tema kliendiprofiilile, võib olla see masinõppealgoritmile signaaliks, et tegemist *SAR* tehinguga. Ketenci *et al.* [13] töö põhjal võib järeldada, et tehinguandmete juures olevad klienti kirjeldava andmed aitavad kõige enam kaasa rahapesukahtlusega tehingute tuvastamisele.

Alshantti *et al.* [27] kasutavad Norra DNB panga andmeid, kus finantsinstitutsiooni töötajate poolt on andmetele määratud rahapesukahtluse või selle puudumise märgend. Treeningandmestik sisaldab 9128-t pangakontot, millest 912 on *SAR* märgisega kontod ning testandmestik sisaldab 2282-t kontot, millest 228 on *SAR* märgisega. Ansambli meetodil, mis sisaldab L2-regularisatsiooni (ingl *L2 Regularisation*), Gini ebapuhtuse (ingl *Gini Impurity*), *ANOVA* F-skoori, Fisheri skoori ja kiire korrelatsioonipõhise filtri (ingl *Fast Correlation Based Filter*) algoritme, sooritatakse olulisemate andmeatribuutide valik. Klassifitseerijana

kasutatakse tehisnärvivõrkude perekonda kuuluvat juhendamata masinõppealgoritmi iseorganiseeruv kaart (ingl *Self-Organising Map*) ning treenitakse kaks erinevat strateegiat, milleks on turvaline (keskmise ja kõrge riskiga kontod klassifitseeritakse kahtlaseks) ja kiire (kõrge riskiga tehingud klassifitseeritakse kahtlaseks). Mudeli õigsuse (ingl *accuracy*), täpsuse (ingl *precision*), saagise (ingl *recall*), F1-skoori ja *AUC* tulemused on toodud kahel erineval strateegial – turvaline ja kiire.

Tulemused strateegiaga turvaline:

- õigsus 0,84,
- täpsus 0,37,
- saagis 0,86,
- F1-skoor 0,52 ja
- *AUC* 0,85.

Tulemused strateegiaga kiire:

- õigsus 0,91,
- täpsus 0,55,
- saagis 0,65,
- F1-skoor 0,59 ja
- *AUC* 0,79

AUC 0,79.Iseorganiseeruv kaart on juhendamata masinõppemeetod, mis on heaks lahenduseks olukorras, kus andmestiku märgendid ei pruugi olla korrektsed (vt ka ptk 1.2). Mudeli tulemused on head võttes arvesse, et tegemist on reaalse finantsinstitutsiooni andmestikuga.

Jullum *et al.* [16] kasutasid samuti Norra DNB panga tehinguandmeid. Rakendades andmestikule andmetöötlus- ja andmekaevemeetodeid, sisaldab lõplik andmestik kokku 1100 tunnust, millele rakendatakse juhendatud masinõppealgoritmi *XGBoost*. Autorid raporteerivad mudeli headusmõõdikud kahel erineval põhimõttel koostatud andmestikul. Esimene andmestik sisaldab ainult rahapesu monitooringusüsteemi poolt tuvastatud rahapesukahtlusega tehinguid ning teine andmestik sisaldab lisanduvalt eelmisele andmestikule ka ausaid (edaspidi ka kui tavapäraseid) tehinguid. *XGBoost* algoritmi *AUC*-skoor on esimesel andmestikul 0,82 ja teisel andmestikul 0,91. Võrreldes Alshantti *et al.* [27] tulemusi sama finantsinstitutsiooni andmestikul, on Jullum *et al.* [16] tulemused paremad, aga siinkohal tuleb välja tuua, et *XGBoost* on juhendatud masinõppemeetod, mille puudustest on võimalik lugeda peatükist 1.2.

Zhang *et al.* [14] kasutasid reaalseid tehinguandmeid, mille USA finantsinstitutsiooni poolt kasutatav rahapesu tuvastamise süsteem oli märkinud rahapesukahtlusega tehinguteks ning soovisid neist tehingutest tuvastada järelevalveasutusele raporteeritud tehinguid. Andmestik koosneb 6079 rahapesukahtlusega tehingust, millest 34 on *SAR* tehingud. Mudeli treenimiseks ja ennustamiseks kasutatakse kümmet sõltumatut muutujat, mida pole andmete

konfidentsiaalsuse tõttu avaldatud. Autorid kasutavad enda töös järgnevat kuute erinevat masinõppealgoritmi: *single-hidden layer feedforward network*, Bayesi logistilist regressiooni (ingl *Bayesian logistic regression*), suurima tõepära logistilist regressiooni (ingl *maximum likelihood logistic regression*), juhumetsa (ingl *random forest*), otsustuspuud (ingl *decision tree*) ja tugivektorklassifitseerijat polünoomse tuumaga (ingl *support vector machine with the polynomial kernel function*). Autorid raporteerivad, et nimetatud esimese nelja algoritmi keskmine *ROC*-kõvera joonealuse pindala skoor (*AUC*) on üle 0,74 ja kahe viimase oma alla 0,6.

Tundis *et al.* [35] kasutasid rahvusvahelisi mobiilimakseid pakkuva ettevõtte andmete põhjal genereeritud andmeid rahapesukahtlusega tehingute tuvastamiseks. Andmestik sisaldab endas 164 502 tehingut (treeningandmestikus 98 701 tehingut ja testandmestikus 65 801 tehingut). Autorid pole kirjeldanud kasutatavaid andmeatribuute ega samme, kuidas andmeid töödeldi. Autorid kasutavad juhumetsa, otsustuspuu, tugivektorklassifitseerija, lineaarnse regressiooni ja Bayesi klassifikaatori masinõppealgoritme rahapesukahtlusega tehingute tuvastamiseks. Mudelite headusmõõdikute tulemused on toodud tabelis 1.3.

Tabel. 1.3. Tundis *et al.* [35] töös kasutatud masinõppemudelite tulemused.

Masinõppealgoritm	**Õigsus**	**Saagis**	**Täpsus**	**F1-skoor**
Juhumets	95.44%	97.22%	94.59%	95.89%
Otsustuspuu	91.57%	94.67%	91.03%	92.81%
Tugivektorklassifitseerija	89,14%	93%	87%	89.9%
Lineaarne regressioon	86,45%	87%	82%	84.4%
Bayes'i klassifikaator	78,54%	81%	74%	77.3%

Allikas: Tundis *et al.* [35], autori kohandatud

Varjatud Markovi mudelit on kasutatud mitmetel elualadel. Käesolevale tööle sarnases valdkonnas on *HMM*-i kasutatud krediitkaardipettuste [44, 45], pangakaardi varguste [46], mobiilimaksete pettuste [47], rahapesukahtlusega tehingute tuvastamisel [18, 48].

Kasianova [18] kasutas *HMM*-i rahapesukahtlusega tehingute tuvastamiseks nii tehislikul andmestikul kui ka Baltikumi finantsinstitutsiooni andmestikul. Varjatud Markovi ahela mudelis on defineeritud kaks varjatud seisundit („rahapesukahtlusega tehing“ ja „aus tehing“) ning kaks vaadeldavat väärtust („kõrge risk“ ja „madal risk“). Varjatud Markovi mudeli vaadeldavaks kihiks on komposiitskoor, mis on leitud reeglitel põhineva rahapesumonitooringu süsteemi põhimõttel. Kasianova [18] leiab manuaalselt parimad reeglid tehislikul andmestikul. Peatükis 1.2 on välja toodud mitmeid puuduseid seoses reeglitel

põhinevate rahapesu monitooringusüsteemidega. Kasianova [18] toob ka enda töö kokkuvõttes välja, et komposiitskoor ei ole dünaamiline, põhineb hetkelisel üldisel arusaamal rahapesukahtlusega tehingute omadustest ning reegleid tuleks pidevalt uuendada ja täiendada. Baltikumi finantsinstitutsiooni andmestikul saadakse täpsuseks 67%, saagiseks 100% ja F1-skooriks 0.81.

Nkemnole *et al.* [46] kasutasid pangakaardi pettuse tuvastamiseks varjatud Markovi ahela mudelit Poissoni jaotusega, üldistatud Poissoni jaotusega ja normaaljaotusega. Kasutatavaid andmeid ja andmete eeltöötlust töö autorid ei kirjelda. *HMM*-i vaadeldava kihi leidmine toimub reeglitel põhineval meetodil, kus võrreldakse kaardiomaniku varasemat käitumisharjumust uue tehtava tehinguga. Tehtud töö tulemusena osutub mudeli headusmõõdikute vaates parimaks normaaljaotusega *HMM*, mille saagis on 85,6% ja vale-negatiivsete osakaal 14,4%.

Danaa *et al.* [47] kasutasid *DBSCAN* algoritmi ja K-keskmiste algoritmi *HMM*-i vaadeldava kihi leidmiseks. Nimetatud töö autorid kasutavad klassifitseerimiseks tehisliku mobiilimaksete andmestikku, mis sisaldab üle kuue miljoni tehingu ja millest 8 213 tehingut on märgitud pettuseks. Esmalt kasutavad autorid andmestiku tasakaalustamiseks SMOTE algoritmi ja seejärel arvutatakse K-keskmiste algoritmiga tehingu kaugus klastri tsentroidist. Klastrite leidmiseks ning anomaaliate tuvastamiseks rakendatakse varasemalt arvutatud kaugustele *DBSCAN* algoritmi, mille tulem on varjatud Markovi mudeli vaadeldavaks kihiks. Kahe kuni viie varjatud seisundiga *HMM*-i tulemused täpsusel, saagisel ja F1-skooril jäävad vastavalt vahemikesse 0,8-1, 0,8-1 ja 0,9-1.

2. Metodoloogia

Kirjanduse ülevaate peatükis tuvastati kaks peamist probleemi. Üheks peamiseks probleemiks on reeglitel põhinevad monitooringusüsteemid, mis pole piisavalt paindlikud, märgivad suurel hulgal tehinguid rahapesukahtlusega tehinguteks, on avalik informatsioon (vt ka ptk 1.2).

Teise probleemina tuvastati, et juhendatud masinõppemeetodid ei ole sobilikud rahapesukahtlusega tehingute tuvastamiseks, sest suure tõenäosusega võivad andmestiku märgendid olla ebakorrektsed, juhendatud masinõppemeetodid ei pruugi olla võimelised tuvastama uusi rahapesumustreid (vt ka ptk 1.2).

Käesolev töö on Kasianova [18] töö edasiarendus. Eesmärgiks on asendada eelnevalt nimetatud töös kasutatud reeglitel põhinev rahapesu riskiskoor juhendamata masinõppealgoritmi *DBSCAN*-i klastritega.

Teadaolevalt on tehingute andmestik tugevalt kallutatud (ingl *imbalanced*) ehk eksisteerib väga väike osakaal rahapesukahtlusega kliente (nt 0,1%) ja suur osakaal kliente, kelle tehingud on tavapärased. Samas puudub teadmine, mitu erinevat tehingu- või kliendiprofiili andmestikus eksisteerib. *DBSCAN* võimaldab andmestikust leida teadmata arvu klastreid - käesoleval juhul on klastriteks n-ö tehingu- ja kliendiprofiilid. Tehingu- ja kliendiprofiilid, mis erinevad suurest hulgast (ausatest) klientidest, määratakse müraks.

DBSCAN on oma olemuselt juhendamata masinõppemeetod, aga parimate hüperparameetrite leidmiseks kasutatakse andmestiku rahapesukahtluse märgendeid. *DBSCAN*-iga leitud klastrid (edaspidi ka kui rahapesu riskiskoor) on varjatud Markovi mudeli vaadeldavaks kihiks. Seejärel ennustatakse *HMM*-iga, kas tegemist on rahapesukahtlusega tehinguga või mitte.

Järgnevad peatükid annavad ülevaate *DBSCAN*-i algoritmist, varjatud Markovi mudelist ning headusmõõdikutest.

2.1 *DBSCAN* masinõppealgoritm

Järgnev alampeatükk põhineb algoritmi *Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise* (*DBSCAN*) autorite samanimelisel artiklil [49].

Klasterdamine on juhendamata masinõppemeetod, mille võib omakorda tinglikult jagada peamiselt kahte kategooriasse: jaotusalgoritmid ja hierarhilised algoritmid. Jaotusalgoritmid konstrueerivad n objektist koosnevast andmestikust k klastrit. Klastrite arv k on nende algoritmide sisendparameeter, mistõttu on vaja omada valdkonnaekspertiisi ehk teada, mitu

klastrit andmestikus eksisteerib või proovida erinevaid klastrite arve ja valida optimaalne. Käesoleva töö kontekstis puudub teadmine, mitu erinevat klastrit tehingu- või kliendiprofiile eksisteerib.

Hierarhilised algoritmid loovad hierarhilise jagunemise H. Hierarhilist jagunemist esindab dendrogramm, mis jagab H iteratiivselt väiksemateks alamhulkadeks, kuni iga alamhulk koosneb ainult ühest objektist. Sellise hierarhia iga sõlm (ingl *node*) esindab jagunemise H klastrit. Dendrogrammi saab luua lehtedest juureni (koonduv lähenemine) või juurest lehtedeni (jagav lähenemine). Hierarhilised algoritmid ei vaja võrreldes jaotusalgoritmidega sisendiks klastrite arvu k-d. Siiski tuleb määratleda lõpetamise tingimus, mis näitab, millal ühinemis- või jagunemisprotsess tuleks lõpetada.

DBSCAN on hierarhiline algoritm, mis põhineb ideel, et punktide tihedus klastris on suurem kui väljaspool klastrit. Klastri iga punkti raadiuses peab olema minimaalsete punktide arv, st tihedus naabruses peab ületama määratletud lävendit. Naabruskonna kuju määrab kahe punkti p ja q kaugusfunktsiooni valik, mida tähistatakse järgnevalt: $dist(p, q)$. Käesolevas töös on kaugusfunktsiooniks valitud Manhattani kaugus. Manhattani kaugus kahe punkti $x = (x_1, x_2, \dots, x_m)$ja $y = (y_1, y_2, \dots, y_m)$ m-mõõtmelises ruumis on kauguste summa absoluutvahe igas mõõtmes ehk $d(x, y) = \sum_{i=1}^{m} |x_i - y_i|$ [50].

Definitsioon 1: Punkti p epsilon (edaspidi ε) naabrus andmestikus D, mida tähistab $N_\varepsilon(p)$, on defineeritud järgmiselt:

$$N_\varepsilon(p) = q \in D | \text{dist}(p, q) \leq \varepsilon.$$

Kuna klastris eksisteerib punkte, mis on raadiuse piiril olevad punktid (ingl *border point*) ja nende punktide ε-naabruse tihedus on madalam kui tuumpunktide (ingl *core point*) tihedus, tuleks määratleda minimaalsete punktide arvu suhteliselt väikeseks, et kaasata kõik punktid, mis kuuluvad samasse klastrisse. Seetõttu nõuame, et klastri C iga punkti p jaoks oleks klastris C punkt q, viisil, kus punkt p asub punkti q ε-naabruses ja $N_\varepsilon(q)$ sisaldab vähemalt minimaalse arvu punkte (edaspidi $MinPts$). Seda määratlust kirjeldatakse allpool.

Definitsioon 2: Otseselt tiheduse-ulatuses (ingl *directly density-reachable*): Punkt p on tihedusega vahetult saavutatav punktist q tulenevalt parameetritest ε, $MinPts$ kui

1) $p \in N_\varepsilon(q)$
2) $|N_\varepsilon(q)| \geq MinPts$ (põhipunktitingimus)

Tuumpunktide otsene tihedus-ulatus on sümmeetriline, aga juhtudel, kus üheks punktiks on tuumpunkt ja teiseks punktis on piiripunkt, ei ole tihedus-ulatus üldiselt sümmeetriline.

Definitsioon 3: Tihedus-ulatuses (ingl *density-reachable*): Punkt p on tihedus-ulatuses punktist q tulenevalt parameetritest ε ja $MinPts$, kui on olemas punktide ahel $p_1, p_2, \dots, p_n,\ p_1 = q, p_n = p$, kus p_{i+1} on punktist p_i otseselt tihedus-ulatuses.

Sama klastri C kaks piiripunkti ei pruugi olla tihedus-ulatuses, sest tuumpunkti tingimus ei pruugi mõlema piiripunkti puhul kehtida. Samas peab olema klastris C tuumpunkt, millest C piiripunktid on tihedus-ulatuses.

Definitsioon 4: Tihedus-ühenduvus (ingl *density connected*): Punkt p on tihedus-ühenduv punktiga q tulenevalt parameetritest ε ja $MinPts$, kui on punkt o, mis on mõlema punkti p ja q tihedus-ulatuses punktist o tulenevalt parameetritest ε ja $MinPts$.

Intuitiivselt määratletakse klastrina tihedus-ühenduvate punktide kogumit, mis on maksimaalne tulenevalt tihedus-ulatusest. Müra määratletakse punktide kogumina andmestikus D, mis ei kuulu ühtegi klastrisse.

Definitsioon 5: Klaster: Olgu D punktide andmestik. Klaster C tulenevalt parameetritest ε ja $MinPts$ on D mittetühi alamhulk mis vastab järgmistele tingimustele:

1) $\forall\, p, q$: kui $p \in C$ ja q on tihedus-ulatuses punktist p tulenevalt parameetritest ε ja $MinPts$, siis $q \in C$.

2) $\forall\, p, q \in C: p$ on tihedus-ühendatud punktiga q tulenevalt parameetritest ε ja $MinPts$.

Definitsioon 6: Müra: Olgu $C_1, \dots, C_k$ klastrid andmestikus D tulenevalt parameetritest ε ja $MinPts, i = 1, \dots, k$. Määratleme müra andmestikus D punktide kogumina, mis ei kuulu ühtegi C_i klastrisse, st. $müra = \{p \in D \mid \forall\, i: p \notin C_i\}$.

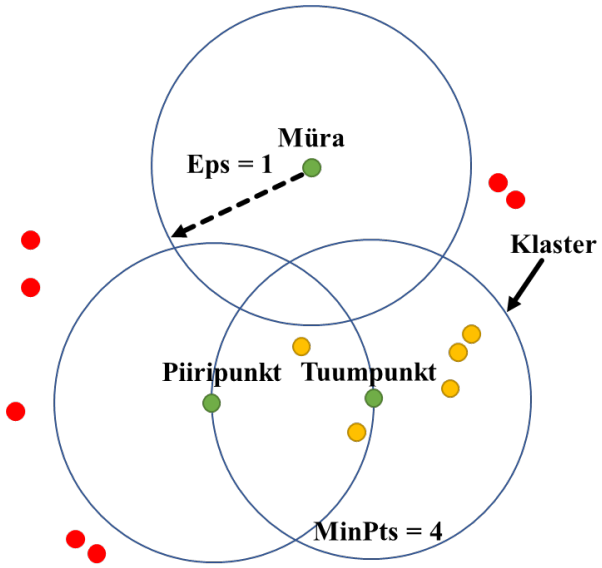
Nüüd saame esitada järgmise olulise tulemuse.

Lemma 1: Olgu p punkt andmestikus D ja $|N_\varepsilon(p)| \geq MinPts$. Siis hulk $O = \{o \mid o \in D$ ja o on tihedus-ulatuses punktist p tulenevalt parameetritest ε ja $MinPts\}$ on klaster tulenevalt parameetritest ε ja $MinPts$.

Iga punkt klastris C on tihedus-ulatuses mis tahes C põhipunktist ja seetõttu sisaldab klaster C punkte, mille tihedus-ulatus on saavutatav mistahes C põhipunktist.

Lemma 2: Olgu C klaster tulenevalt parameetritest ε ja $MinPts$ ja p on mis tahes punkt klastris C tingimusel $|N_\varepsilon(p)| \geq MinPts$. Siis on C võrdne hulgaga $O = \{o|o$ on tihedus − ulatuses punktist p tulenevalt parameetritest eps ja $MinPts$.

DBSCAN valib klastri leidmiseks suvalise punkti p ja otsib punktile p kõik tihedus-ulatuses olevad punktid tulenevalt parameetritest ε ja $MinPts$ (joonis 2.1).



Joonis 2.1. *DBSCAN* algoritmi tööpõhimõte. Allikas: Armstrong *et al.* [51], autori kohandatud

Kui p on põhipunkt, annab see protseduur tulenevalt parameetritest ε ja $MinPts$ klastri vastavalt Lemmale 2. Kui punkt p on piiripunkt ja ühtegi punkti ei ole punkti p tihedus-ulatuses, valitakse järgmine punkt. Kõik punktid, mis ei kuule mitte ühtegi klastrisse määratletakse müraks.

2.2 Varjatud Markovi mudel

Varjatud Markovi mudel on aastatega kogunud populaarsust ning see on rakendust leidnud mitmetel elualadel. Näiteks on kasutatud *HMM*-i kõnetuvastuses, näoilmete tuvastamises, geenide ennustamises, žestide tuvastuses, muusikalises kompositsioonis ja bioinformaatikas. Varjatud Markovi mudelil on mitmeid alamjaotusi *First-order HMM*, *Second-Order HMM*, *Higher-Order HMM*, *Hidden-Semi Markov Model*, *Factorial HMM*, *Layered HMM*, *Autoregressive HMM*, *Non-Stationary HMM*, *Hierarchal HMM* jne. [52] Käesolevas töös kasutatakse klassikalist esimest järku varjatud Markovi mudelit.

HMM-i peamisteks komponentideks on algjaotus, üleminekutõenäosuste maatriks emiteerimistõenäosuste maatriks, vaadeldavate tunnuste järjend ja varjatud seisund. *HMM* hakkab mööda vaadeldavaid tunnuseid liikuma ning arvama, tuginedes seisundite algjaotusele, seisundite üleminekutõenäosustele ja emiteerimistõenäosustele, mis varjatud seisund võis põhjustas vaadeldava tunnuse. Klassikaline esimest järku varjatud Markovi mudel järgib Markovi ahela reeglit, kus seisund on sõltuv ainult talle vahetult eelnevast seisundist.

Rahapesu tuvastamisel on kohustatud isikule teada nii klienti kirjeldavad andmed kui ka tehinguandmed ehk vaadeldavad tunnused, millele tuginedes on võimalik tuvastada tehingu varjatud seisundit ehk kas tegemist on rahapesukahtlusega tehinguga või mitte. Seetõttu sobib varjatud Markovi mudel oma ülesehituselt rahapesukahtlusega tehingute tuvastamiseks. Varjatud Markovi mudeli eeliseks on asjaolu, et tegemist on statistilise mudeliga, mis ei vaja märgendeid mudeli treenimiseks.

Järgnev osa peatükist põhineb Lawrence R. Rabineri 1989. aastal avaldatud artiklil "*A tutorial on hidden Markov models and selected applications in speech recognition*" [17].

HMM on statistiline mudel, mille abil on võimalik modelleerida juhuslikku protsessi. *HMM* põhineb Markovi ahela reeglil, kus iga seisund on sõltuv vahetult eelnevast seisundist (varjatud) ahelas ning genereerib vastavalt tihedusfunktsioonile vaadeldava tunnuse. *HMM*-iga modelleeritakse vaadeldava järjestuse põhjal varjatud seisundeid.

Vaatleme juhuslike suuruste jada Q, mis võib igal ajal olla N erinevas seisundis $S_1, S_2, \dots, S_N$. Teatud ajaintervalli järel toimub süsteemis muutus ja liigutakse edasi järgmisesse seisundisse (võimalik liikuda ka samasse seisundisse tagasi). Tähistame seisundimuutuse ajahetke $t = 1, 2, \dots$ ja seisundit ajahetkel t kui q_t ning järjestikuste seisundite jada $Q = q_1, q_2, \dots, q_t$.

Esimest järku diskreetse Markovi ahela reegli järgi on tegelik seisund q_t ajahetkel t tulenev ainult eelnevast seisundist q_{t-1}, ehk

$$P(q_t = S_j | q_{t-1} = S_i, q_{t-2} = S_k \dots, q_1 = S_j) = P(q_t = S_j | q_{t-1} = S_i).$$

Protsess algab algjaotusest (ingl *initial state probabilities*) $\pi = \pi_1, \pi_2, \dots, \pi_N$, kus π_i on tõenäosus, et Markovi ahel algab seisundiga S_i, seega $\sum_{i=1}^{n} \pi_i = 1$.

Protsessi ühest seisundist teise liikumist iseloomustavad üleminekutõenäosused a_{ij}, mis esitletakse üleminekutõenäosuste maatriksina (ingl *state transition matrix*) A.

$$A = \{a_{ij}\} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1N} \\ a_{21} & a_{22} & \dots & a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N1} & a_{N2} & \dots & a_{NN} \end{pmatrix}$$

kus $a_{ij} = P(q_{t+1} = S_j | q_t = S_i), 1 \leq i, j \leq N$ ning üleminekukoefitsendid järgivad tingimust $a_{ij} \geq 0$ ning $\sum_{j=1}^{N} a_{ij} = 1, \ \forall i$.

Varjatud Markovi mudelis on seisundid varjatud, seetõttu tähistame M võimalikku vaadeldavate tunnuste hulga $V = \{v_1, v_2 \dots v_M\}$, mida varjatud tunnus S põhjustab. *HMM* hakkab liikuma mööda varjatud jada $Q = q_1, q_2, \dots, q_t$ ja emiteerib seejärel t vaadeldavat tunnust $O = o_1, o_2, \dots, o_t$. Emiteerimistõenäosuseks nimetatakse tõenäosusjaotust, et varjatud tunnus ajahetkel t, $q_t = S_i$ emiteerib vaadeldava tunnuse $O_t = v_k$, tähistatakse emissioonimaatriksiga (ingl *state emission matrix*) B.

$$B = \{b_i(v_k)\} = \begin{pmatrix} b_1(v_1) & b_1(v_2) & \dots & b_1(v_M) \\ b_2(v_1) & b_2(v_2) & \dots & b_2(v_M) \\ \vdots & \vdots & \ddots & \vdots \\ b_N(v_1) & b_N(v_2) & \dots & b_N(v_M) \end{pmatrix}$$

kus $b_i(v_k) = P(O_t = v_k | q_t = S_i)$,

Varjatud Markovi mudeli kasutamiseks reaalelulistes rakendustes peab lahendama järgmised kolm olukorda:

1. Tõenäosus: kui on teada *HMM*-i parameetrite kolmik $\lambda = (\pi, A, B)$ ja vaatluste jada O, siis kuidas efektiivselt leida tõenäosused $P(O|\lambda)$.

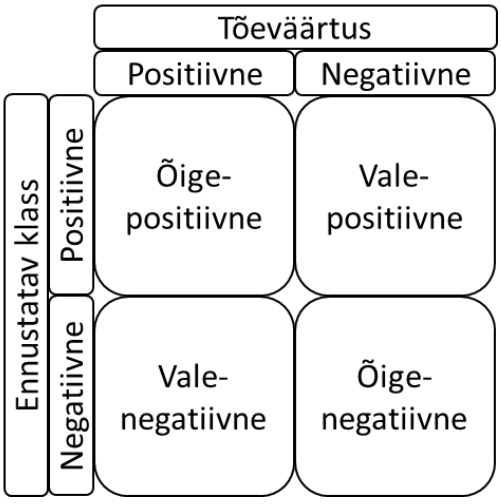
2. Lahtikodeerimine: kui on teada *HMM*-i parameetrite kolmik $\lambda = (\pi, A, B)$ ja vaatluste jada O, siis kuidas leida parim peidetud olekute jada Q.
3. Õppimine: leida vaatlusjärjestuse O ja võimalike seisundite S põhjal parameetrid π, A ja B. Kuidas leida parameetrid $\lambda = (\pi, A, B)$ nii, et $P(O|\lambda)$ oleks maksimaalne.

Käesolevas töös kasutatakse lahtikodeerimisel Viterbi algoritmi ja õppimisel Baum-Welch algoritmi.

2.3 Tulemusmõõdikud

Mudeli headusmõõdikuteks on valitud täpsus (ingl *precision*), saagis (ingl *recall*), F1-skoor, F2-skoor, *ROC*-kõver (ingl *Receiver Operating Characteristic curve*) ja *AUC* (ingl *Area Under the ROC Curve*). Kuna käesoleva töö eesmärgiks on ennustada binaarset diskreetset tunnust („rahapesukahtlusega tehing“, „aus tehing“) tugevalt tasakaalustamata andmestikul, on nimetatud headusmõõdikud sobilikud andmaks ülevaadet mudeli headusest. Lisaks on nimetatud headusmõõdikud klassifitseerimise valdkonnas laialdaselt levinud [14, 16, 27, 35, 47] mis võimaldab erinevaid töid omavahel võrrelda või vähemalt anda indikatsiooni mudeli tulemustest, sest kasutatavad andmed on erinevad.

Klassifitseerimismaatriksiga on võimalik illustreerida headusmõõdikute arvutamiseks vajalike väärtusi (Joonis 2.3) ja selle komponentideks on õige-positiivne, vale-negatiivne, vale-positiivne ja õige-negatiivne klassifitseerimine.



Joonis 2.3 Klassifitseerimismaatriks.

Klassifitseerimismaatriksi komponentide definitsioonid on järgnevad:

1. Õige positiivne (ÕP) - loendatud arv juhte, kus ennustatud väärtus on positiivne ja tegelik väärtus on positiivne.

2. Õige negatiivne (ÕN) - loendatud arv juhte, kus ennustatud väärtus on negatiivne ja tegelik väärtus on negatiivne.
3. Vale-negatiivne (VN) - loendatud arv juhte, kus ennustatud väärtus on negatiivne ja tegelik väärtus on positiivne. Tuntud ka kui II liiki viga.
4. Vale-positiivne (VP) - loendatud arv juhte, kus ennustatud väärtus on positiivne ja tegelik väärtus on negatiivne. Tuntud ka kui I liiki viga.

Täpsus näitab, et kui suur osakaalu ennustatud positiivsetest väärtustes oleme ennustatud korrektselt kõigist ennustatud positiivsetest väärtustest ehk kui suure osakaalu õigesti ennustatud rahapesukahtlusega tehingutest oleme ennustanud kõigist rahapesukahtluseks märgitud tehingutest.

$$\text{Täpsus} = \frac{\text{ÕP}}{\text{ÕP} + \text{VP}}$$

Saagis näitab, et kui suure osakaalu positiivsetest väärtustes on mudel ennustatud korrektselt ehk kui suure osa rahapesukahtlusega tehingutest oleme suutnud tuvastada kõigist rahapesukahtlusega tehingutest.

$$\text{Saagis} = \frac{\text{ÕP}}{\text{ÕP} + \text{VN}}$$

F1-skoor on harmooniline keskmine täpsusest ja saagisest.

$$\text{F1-skoor} = 2 * \frac{\text{Täpsus} \times \text{Saagis}}{\text{Täpsus} + \text{Saagis}}$$

Tulenevalt rahapesukahtlusega tehingute tuvastamise protsessist on finantsinstitutsioonid sunnitud valima, kas määrata tehing rahapesukahtlusega tehinguks või mitte. Määrates tehingu rahapesukahtlusega tehinguks, riskitakse ausale kliendile rahapesukahtluse omistamisega ja ebavajaliku uurimise alustamisega, mis võib omakorda kaasa tuua kliendisuhete halvenemise. Teisalt jättes tehingu rahapesukahtlaseks märkimata, riskitakse võimalusega, et lastakse kurjategijatel edasi tegutseda, millega võivad kaasneda trahvid FI-lt, mainekahju ja finantskahju. Näiteks omasid rahapesuskandaalid tugevat mõju Swedbank AB A-seeria aktsiahinnale (vt lisa I). 18. veebruaril 2019. aastal avaldati raport, kus viidati Swedbanki osalusele Danske panga rahapesuskandaalis ning 21. veebruariks oli aktsia kaotanud 20% oma väärtusest [54]. Teine suur aktsiahinnalangus on seotud 23. märtsil 2019. aastal avaldatud raportiga, kus selgub, et Swedbankis teostati viie aastase perioodi jooksul 37 miljardi euro väärtuses kõrge rahapesukahtlusega tehinguid (28. märtsiks oli aktsia kaotanud 17% oma

väärtusest) [55]. Seetõttu on nii moraalsetel kui ka finantsilistel põhjustel olulisem minimeerida II liiki vea toimumist ehk vale-negatiivseid tulemusi. Seetõttu arvutatakse headusmõõdikuna ka F2-skoor. F2-skoor on harmooniline keskmine täpsusest ja saagisest, aga annab saagisele suurema kaalu.

$$\text{F2-skoor} = \frac{(1+2^2) * \tilde{\text{O}}\text{P}}{(1+2^2) \times \tilde{\text{O}}\text{P} + 2^2 \times \text{VN} + \text{VP}}$$

ROC-kõver on graafiline viis illustreerimiseks mudeli klasside ennustamist erinevate tõenäosuste tasemetel. *ROC*-kõvera komponentideks on õige-positiivsete määr (saagis) ja vale-positiivsete määr.

$$\text{Vale-positiivsete määr} = \frac{\text{VP}}{\text{VP} + \tilde{\text{O}}\text{N}}$$

AUC on *ROC*-kõvera joonealune pindala, mida kasutatakse *ROC*-kõvera üldistamiseks ühe numbriga.

3. Empiiriline uurimus

Käesolev peatükk jaguneb peamiselt kaheks osaks. Esimeses osas antakse ülevaade kasutatud andmestikust ning andmestiku eeltöötluse töövoost ning teises osas tutvustatakse saadud tulemusi.

3.1 Andmed

Käesoleva töö empiiriline osa põhineb tehislikult genereeritud andmestikul. Kasutusel olev andmestik koosneb tehinguandmetest ja *CRM* andmetest.

Tehinguandmestiku andmeatribuudid ja nende tähendused on järgnevad:

1. *id* - unikaalne tehingunumber
2. *user_id* - kliendi identifitseerimisnumber
3. *type* - sissetuleva või väljamineva tehingu indikaator
4. *date_created* - tehingu toimumise aeg
5. *sender_account* - tehingu saatja kontonumber
6. *receiver_account* - tehingu saaja kontonumber
7. *from_cur* - tehingu saatmise valuuta
8. *to_cur* - tehingu saamise valuuta
9. *amount_in_eur* - tehingu väärtus eurodes
10. *transaction_type* - siseriikliku makse, välisriikliku makse, sularahatehingu või finantsasutusesisese makse indikaator,
11. *meta_sar_id* – rahapesukahtluse või selle puudumise indikaato

Klienti kirjeldavast andmestikust kasutatakse järgnevaid andmeatribuute:

1. *id* - kliendi identifitseerimisnumber
2. *date_created* – kliendilepingu sõlmimise kuupäev,
3. *dob* - kliendi sünnikuupäev;
4. *country_of_residence* - kliendi residentsusriik
5. *country* - kliendi elukohariik;
6. *sic_code* - juriidilise isiku tegevusvaldkonna kood
7. *customer_type* - juriidilise või füüsilise isiku indikaator

Andmestike eeltöötluse töövoog on järgnev:

1. Tehinguandmestikule ja *CRM* andmestikule leitakse lisanduvad tunnused (vt ka Lisa II).
2. Tehinguandmestik ühendatakse üheks andmestikuks *CRM* andmestikuga.

3. Andmestiku kategoorilistele tunnustele rakendatakse ühega kodeerimist (ingl *one-hot encoding*).
4. Teostatakse andmestiku numbriliste väärtuste standardiseerimine (ingl *standard scaling*) ehk samale skaalale viimini läbi keskväärtuse lahutamise ja standardhälbega jagamise.
5. Andmestiku jagamine treenimis- ja testandmestikuks põhimõttel, et andmestikud sisaldaksid võrdse arvu kliente, kellel on esinenud vähemalt üks rahapesukahtlusega tehing (tabel 3.1). Kuna rahapesu riskiskoori leidmiseks kasutatakse tihedusfunktsioonil põhinevat *DBSCAN* algoritmi, on oluline hoida andmestike ridade arv samaväärne tulenevalt *DBSCAN* tööpõhimõttest. Seetõttu on tehingute arv treening- ja testandmestikus ligilähedane (edaspidi ka kui 50%/50% kriteerium). *DBSCAN* masinõppealgoritmi rakendatakse treening- ja testandmestikule eraldi.

Tabel 3.1. Treening- ja testandmestiku kokkuvõte.

Andmestik	**Tehingute arv**	***SAR* kliente**	***SAR* tehinguid**	***SAR* tehingute osakaal**
Treeningandmestik	67071	36	135	0,2%
Testandmestik	65717	37	132	0,2%

Kokkuvõttes toimus klientide andmestikesse jagamine juhuslikult ning täita tuli varasemalt nimetatud rahapesuklientide võrdse määra ja andmestike ridade arvu 50%/50% tingimus. Tabelist 3.1 nähtub, et *SAR* tehingute osakaal treening- ja testandmestiku on juhuslikult sama suur. Töö autori eesmärk oli võimalikult vähe seada reegleid andmestike koostamisele.

3.2 Reeglitel põhinev rahapesu riskiskoor koos varjatud Markovi mudeliga

Baasmudelina kasutatakse Kasianova [18] töös välja töötatud metodoloogiat, mis koosneb reeglitel põhineval rahapesu riskiskoori leidmise meetodil ning varjatud Markovi mudelil. Kasianova [18] kasutas reeglite väljatöötamiseks ja optimeerimiseks sama andmebaasi tehislike andmeid, mida kasutakse ka käesolevas töös. Tema poolt väljatöötatud reeglid varjatud Markovi mudeli vaadeldava kihi (rahapesu komposiitskoori) leidmiseks on toodud tabelis 3.2.

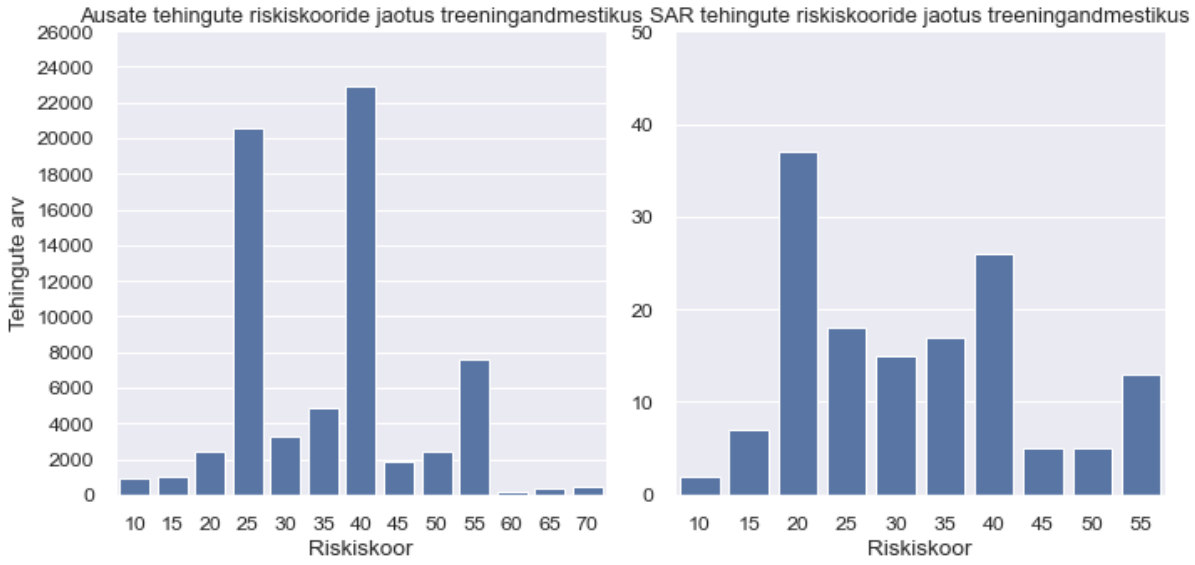
Tabel 3.2. Rahapesu riskiskoori arvutamise reeglid

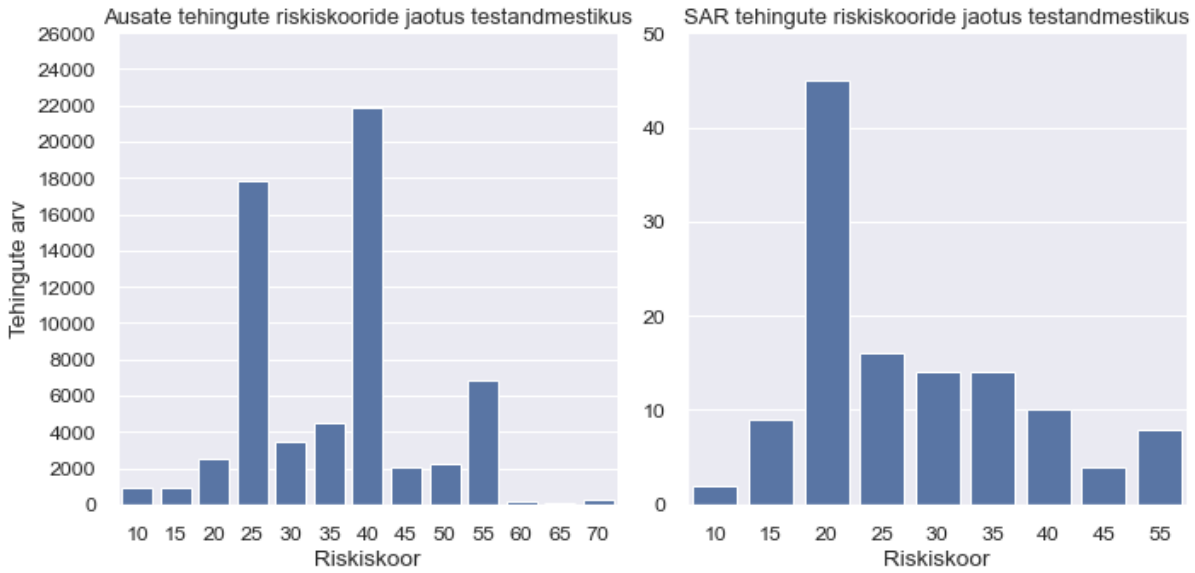
Reegel	**Riskiskoori tõus**
Tehingu toimumise aeg enne kella 7:00	15
Tehingu toimumise aeg peale kella 21:00	15
Tehingu osapoole riik on kõrge rahapesuriskiga riik	10
Tehinguvaluuta ei ole EUR või USD	10

Viimase 3 päeva sissetulevate maksete summa vahemikus [5 000; 10 000)	5
Viimase 3 päeva sissetulevate maksete summa vahemikus [10 000; 20 000)	10
Viimase 3 päeva sissetulevate maksete summa suurem kui 20 000	15
Viimase 3 päeva väljaminevate maksete summa vahemikus [5 000; 10 000)	5
Viimase 3 päeva väljaminevate maksete summa vahemikus [10 000; 20 000)	10
Viimase 3 päeva väljaminevate maksete summa suurem kui 20 000	15
Viimase 7 päeva maksete arv vahemikus [3; 5)	5
Viimase 7 päeva maksete arv vahemikus [5; 10)	10
Viimase 7 päeva maksete arv suurem kui 10	15
Viimase 3 päeva väljuvate maksete osakaal sissetulevatest maksetest >=90%	15

Autor: Kasianova [18], autori kohandatud

Joonise 3.2.1 vasakus tulbas on kujutatud ausate tehingute riskiskooride jaotused treening- ja testandmestiku lõikes. Treening- ja testandmestiku tehingud on peamiselt saanud riskiskooriks 25, 40 ja 55 punkti. Nimetatud riskiskooriga klastritesse kuulub 74,17% treeningandmestiku ja 73,10% testandmestiku kõigist ausatest tehingutest.





Joonis 3.2.1. Ausate ja rahapesukahtlusega tehingute riskiskooride jaotus andmestike lõikes.

Joonise 3.2.1 paremas tulbas on kujutatud *SAR* tehingute riskiskooride jaotused andmestike lõikes. Treening- ja testandmestiku *SAR* tehingute jaotuses näeme, et võrreldes ausate tehingute riskiskooride jaotusega on esile kerkinud riskiskoor 20. Väike osakaal *SAR* tehingutest on saanud andmestikus madalaima riskiskoori (10 punkti), aga mitte ükski rahapesukahtlusega tehingutest pole saanud riskiskooriks üle 55 punkti (maksimaalne riskiskoor andmestikus 70 punkti). Treening- ja testandmestiku *SAR* tehingute riskiskooride jaotused jäävad esimesse graafiku poolde, kus rahapesukahtluse riskiskoor on madalam.

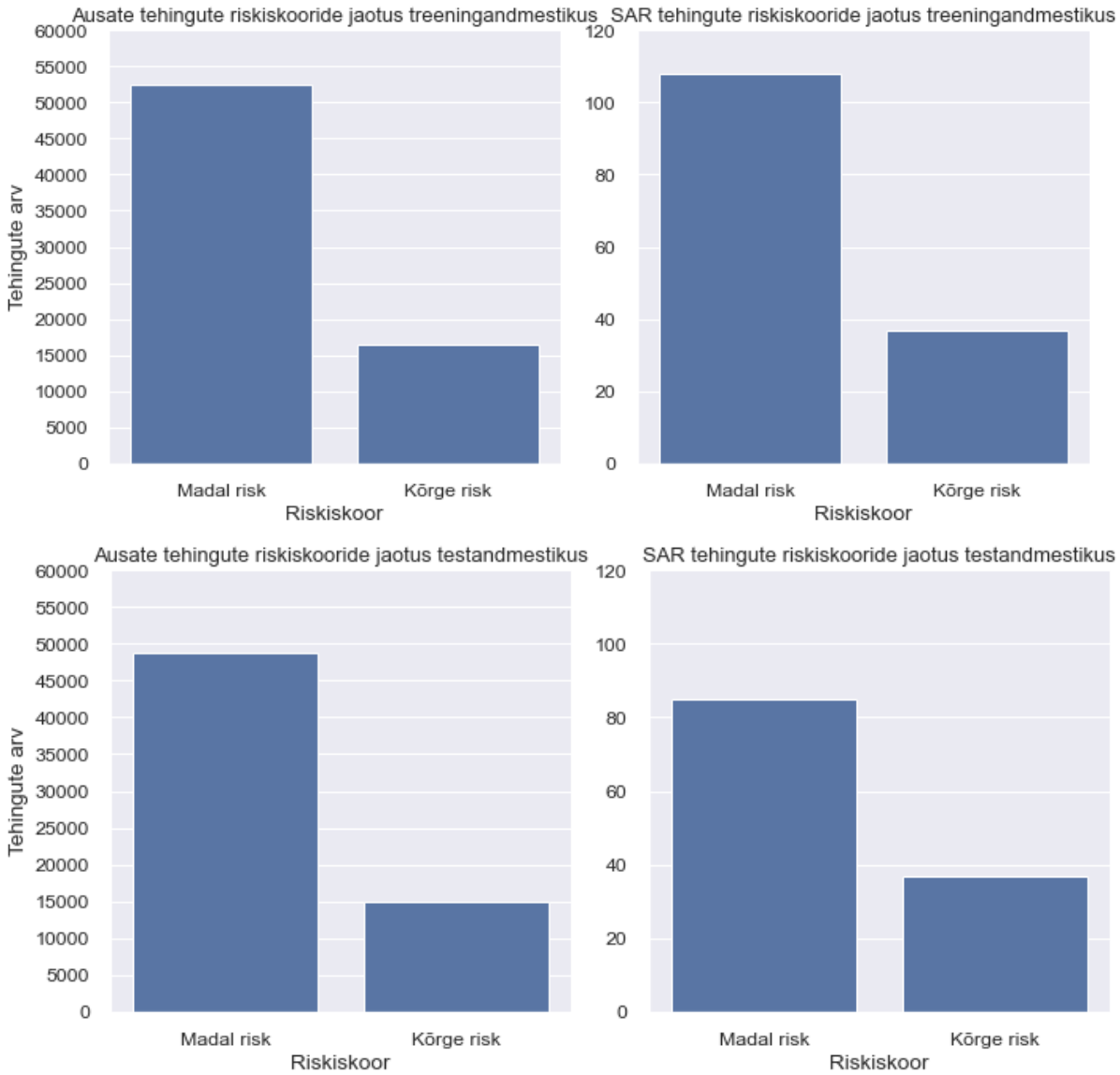
Mõlemas andmestikus on suurim arv *SAR* tehinguid saanud riskiskooriks 20 punkti, mis on pigem madal, võttes arvesse, et ausate tehingute maksimaalne riskiskoor on 70. Riskiskoori määramisel reeglitel põhineval meetodil oleks olnud loogiline eeldada, et andmestike jaotused *SAR* tehingute vaates on kaldu kõrgemate riskiskooride suunas, mis hetkel ei pea paika. Mistõttu võime järeldada, et reeglitel põhineva rahapesu riskiskoori määramine on suure tõenäosusega ebaõnnestunud.

Varasemalt leitud rahapesu riskiskooridele rakendab Kasianova [18] reegleid, et saada varjatud Markovi mudeli vaadeldava kihi väärtusteks „madal risk“ ja „kõrge risk“. Rakendatavad reeglid on järgnevad:

1. kui riskiskoor ≤ kliendi maksimaalsest riskiskoorist, millest on lahutatud 5 ühikut, on tehingu riskiskooriks „madal risk“
2. kui riskiskoor > kliendi maksimaalsest riskiskoorist, millest on lahutatud 5 ühikut, on tehingu riskiskooriks „kõrge risk“

3. kui kliendi maksimaalne riskiskoor on 0, on kliendi kõik tehingud riskiskooriga „madal risk“.

Joonisel 3.2.2 on kujutatud riskiskooride jaotus peale reeglite rakendamist. Joonise 3.2.2 vasakus tulbas on ausate tehingute riskiskooride jaotus treening- ja testandmestiku lõikes. Ausate tehingute jaotuses on suur osakaal madala riskiga tehinguid, aga ka kõrge riskiga tehingute osakaal ausates tehingutes on üsna kõrge (treeningandmestikus 23,81% ja testandmestikus 23,56% kõigist ausatest tehingutest).



Joonis 3.2.2. Ausate ja rahapesukahtlusega tehingute riskiskooride jaotus andmestike lõikes.

Joonise 3.2.2 paremas tulbas on kujutatud *SAR* tehingute riskiskooride jaotus treening- ja testandmestikus. Treeningandmestikus on saanud 74,48% *SAR* tehingutest madala riskiskoor ning testandmestikus on samaks näitajaks 69,67%. Teadaolevalt kasutas metodoloogia autor

reeglite optimeerimiseks sama andmebaasi, aga autori tehingud pärinesid teisest ajaperioodis. Seetõttu loodud reeglid sama andmebaasi erineva perioodi andmetel häid tulemusi ei anna, mis on kooskõlas peatükis 1.2 välja toodud reeglitel põhineva monitooringusüsteemi puudustega.

Magistritöö praktilises osas kasutati varjatud Markovi mudeli valmis funktsioone *scikit-learn hmmlearn* teegist. Varjatud Markovi mudeli parameetrid pärinevad Kasianova [18] tööst. Varjatud Markovi mudeli varjatud kihi võimalikeks väärtusteks määratakse "aus tehing" ja "rahapesukahtlusega tehing". Reeglitel põhineva riskiskoori meetodil leitud *HMM*-i vaadeldava kihi väärtusteks on "madal risk" ja "kõrge risk". Varjatud Markovi mudeli parameetrid on järgnevad:

- Varjatud kihi võimalikud väärtused:

$$S = (S_1 = \text{“aus tehing”};\ S_2 = \text{“rahapesukahtlusega tehing”})$$

- Vaadeldava kihi võimalikud väärtused:

$$(\text{“madal risk”};\ \text{“kõrge risk”})$$

- Algjaotus:

$$\pi = (\pi_1 = 0{,}9;\ \pi_2 = 0{,}1)$$

- Üleminekumaatriks:

$$A = \begin{pmatrix} a_{11} = 0{,}8;\ a_{12} = 0{,}2 \\ a_{21} = 0{,}1;\ a_{22} = 0{,}9 \end{pmatrix}$$

- Emiteerimismaatriks:

$$B = \begin{pmatrix} b_1(\text{“madal risk”}) = 0{,}90; b_1(\text{“kõrge risk”}) = 0{,}10 \\ b_2(\text{“madal risk”}) = 0{,}01; b_2(\text{“kõrge risk”}) = 0{,}99 \end{pmatrix}$$

HMM mudeli parameetrid treeniti Baum-Welch algoritmiga ja kõige tõenäolisemat järjestust ennustati Viterbi algoritmiga.

3.3 *DBSCAN* masinõppealgoritm koos varjatud Markovi mudeliga

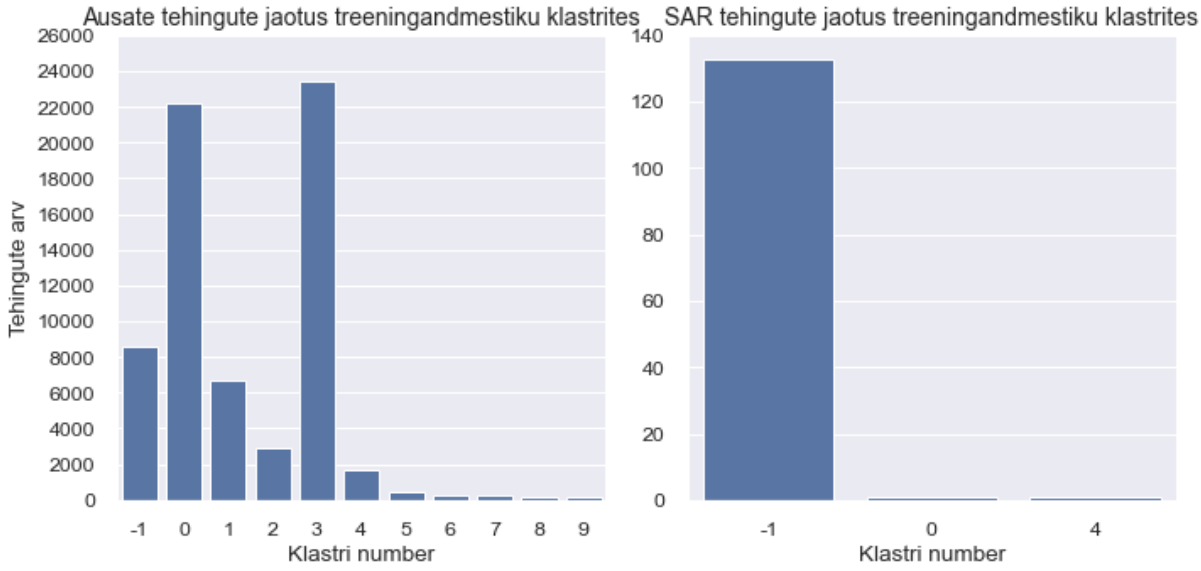
DBSCAN algoritmi kasutuseesmärgiks on leida rahapesu riskiskoor, mis oleks varjatud Markovi mudeli vaadeldavaks kihiks. Magistritöö praktilises osas on kasutatud *scikit-learn DBSCAN* algoritmi valmis funktsioone. *DBSCAN* algoritmi kaks peamist sisendparameetrit on *eps* (maksimaalne distants kahe punkti vahel, teooria osas mainitud ε) ja *min_samples* (minimaalne vaatluste arv, et moodustada klaster, teooria osas mainitud $MinPts$). Tuginedes teadmisele, et rahapesukahtlusega tehingute tuvastamisel on tegemist tugevalt

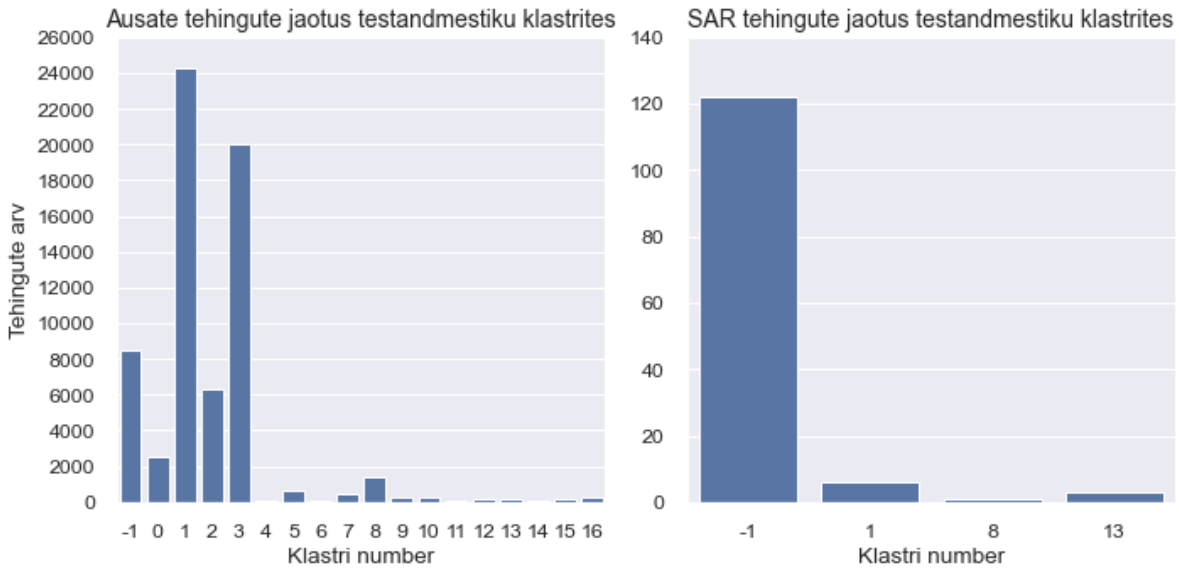
tasakaalustamata andmestikuga, soovime leida võimalikult väikse *eps* parameetri ja võimalikult suure *min_samples* parameetri, mille tulemusena määratakse rahapesukahtlusega tehingud ühte klastrisse (*DBSCAN* algoritmi teoreetilises vaates müraks).

Parimate hüperparameetrite leidmiseks kasutati võrguotsimise (ingl *grid search*) meetodit. Võrguotsimise headusmõõdik koosnes kahest järgnevast kriteeriumist:

1. Müraks märgitud *SAR* tehingute osakaal kõigist *SAR* tehingutest treeningandmestiku peab olema üle 80%.
2. Seejärel valiti parameetrid kõrgeima täpsuse väärtuse põhjal.

Võrguotsimise meetodil tuvastati, et parimaks *eps* väärtuseks on 4 ja *min_samples* väärtuseks 90. Leitud parimate parameetrite korral on treeningandmestikus müraks märgitud *SAR* tehingute osakaal kõigist *SAR* tehingutest 98,52% ja täpsuseks 1,52%. Joonise 3.3.1 vasakus tulbas on kujutatud ausate tehingumärgendiga *DBSCAN* algoritmi klastrite suurused treening- ja testandmestiku lõikes. Treeningandmestikus on kaks suurimat klastrit numbriga 0 (33,12% kõigist tehingutest) ja 3 (34,97% kõigist tehingutest). Testandmestiku kaks suurimat klastrit on numbritega 1 ja 3, kuhu kuuluvad vastavalt 36,95% ja 30,47% tehingutest andmestikus.





Joonis 3.3.1. *DBSCAN* algoritmi klastrite suurused ausate ja rahapesutehingute lõikes.

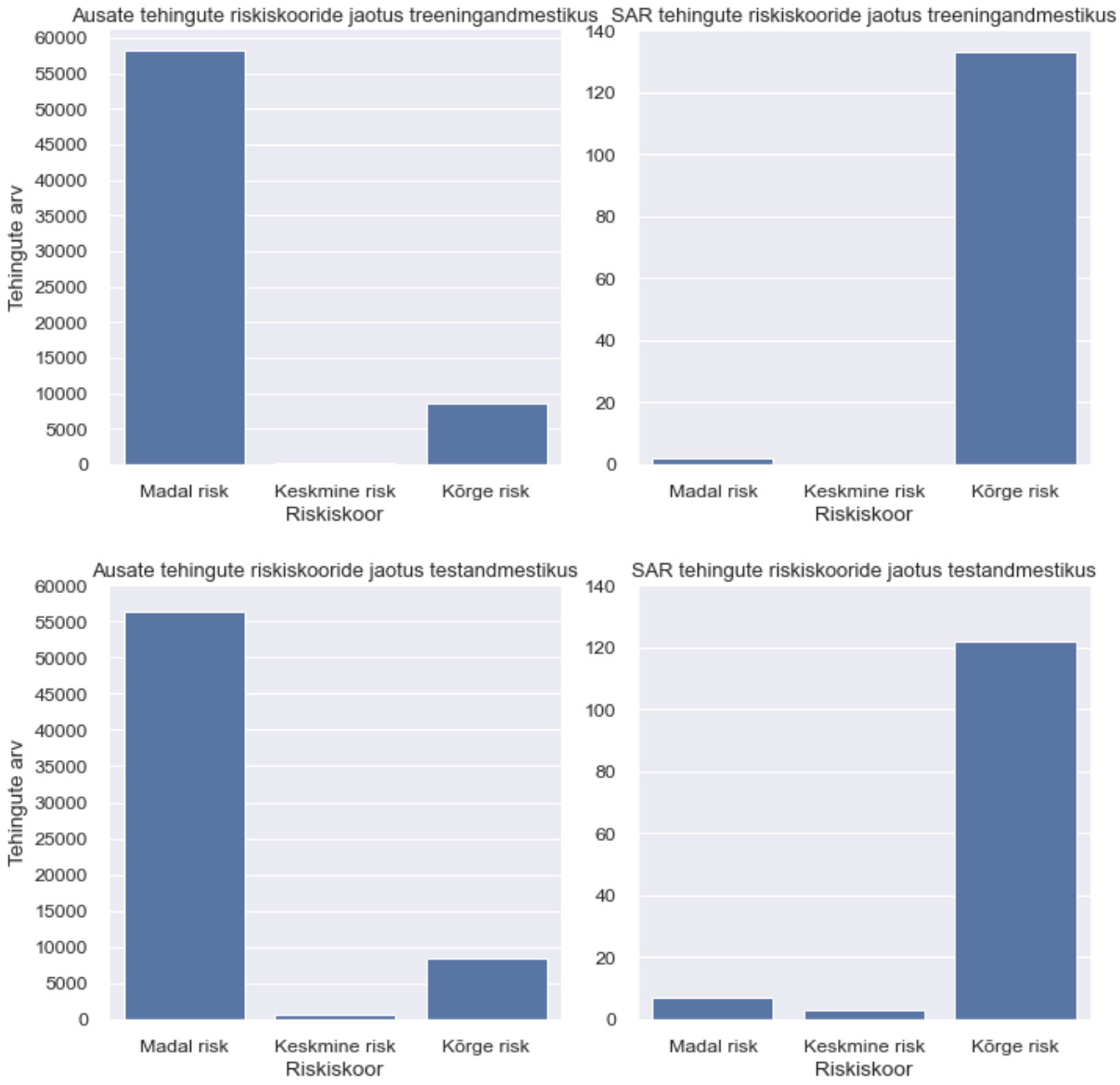
Joonise 3.3.1 paremas tulbas on kujutatud *SAR* tehingute *DBSCAN* algoritmi klastrite suurused treening- ja testandmestiku lõikes. Jooniselt on näha, et kõigis andmestikes on suurimaks klastriks -1 ehk müraks klassifitseeritud tehingud. Tulenevalt võrguotsimise esimesest kriteeriumist, on loogiline, et treeningandmestikus on ülekaalukalt suurimaks klastriks müraks klassifitseeritud tehingud. Positiivne on näha, et ka testandmestiku *SAR* tehingute suurimaks klastriks on müraks klassifitseeritud tehingud. Teisal jääb jooniselt 3.3.1 silma ka asjaolu, et treening- ja testandmestikus klasterdati suur hulk (vastavalt 12,87% ja 12,97% kõigist ausatest tehingutest) ausatest tehingutest samuti müraks.

Käesoleva töö konteksti arvesse võttes on loogiline anda tehingutele rahapesu riskiskooriks märgend „madal risk“, „keskmine risk“ ja „kõrge risk“. *DBSCAN* algoritm leiab tuginedes *min_samples* ja *eps* parameetritele teadmata arvu klastreid. Seetõttu rakendati rahapesu riskiskoori leidmiseks klastritele järgnevaid reegleid:

1. Kui klastri numbriks on -1 ehk tehing ei kuulu mitte ühtegi klastrisse, siis tehingu märgendiks on “kõrge risk”.
2. Kui klastri märgend ei ole -1 ja klastris ei ole kaks korda rohkem tehinguid, kui on parameetri *min_samples* väärtus, on tehingu rahapesu riskiskooriks “keskmine risk”.
3. Kui klastri märgend ei ole -1 ja klastris on kaks korda rohkem tehinguid, kui on parameeter *min_samples*, on tehingu rahapesu riskiskooriks “madal risk”.

Reeglite rakendamisel saadi vaadeldava kihi võimalikeks väärtusteks „madal risk“, „keskmine risk“ ja „kõrge risk“. Joonise 3.3.2 vasakus tulbas on kujutatud tehingute rahapesu

riskiskooride jaotust treening- ja testandmestiku lõikes, kus märgendiks on rahapesukahtluse puudumine (peale reeglite rakendamist). Treening- ja testandmestikus on riskiskooride jaotus samas suurusjärgus, aga testandmestikus on keskmise riskiga tehingute osakaal vähesel määral suurem. Seda tulenevalt asjaolust, et treeningandmestikust leiti suurem arv väikse arvu tehingutega klastreid (vt ka joonis 3.3.1).



Joonis 3.3.2. Riskiskooride jaotus ausate ja rahapesukahtluse tehingute lõikes.

Joonise 3.3.2 paremast tulbast nähtub, et treening- ja testandmestikus on suur osakaal *SAR* märgendiga tehingutest saanud riskiskooriks „kõrge risk“. Samuti on treening- ja testandmestiku ausate tehingute hulgas suur osakaal kõrge riskiga tehinguid (vastavalt 14,78% ja 14,90% kõigist ausatest tehingutest).

Praktikas võiks eeldada, et andmestikus on suur osakaal madala riskiga tehinguid, seejärel keskmise riskiga tehinguid ja väga väike osakaal kõrge riskiga tehinguid. Käesoleva eksperimendi ülesehituse juures näeme, et väga väiksele osakaalule tehingutest määratakse riskiskoor „keskmine risk“. Nimetatud põhjus võib tuleneda asjaolust, et *DBSCAN* valitud parameetrid pole parimad. Kuna tegemist on tihedusfunktsioonil põhineva algoritmiga on oluline vaatluste maht andmestikus. Võrguotsingumeetodi kasutamine parimate parameetrite leidmiseks on ajamahukas ja ressursimahukas tegevus. *DBSCAN* algoritm nõuab sisendparameetrina kaugusmaatriksit, mis on dimensioonides $n \times n$ (ehk andmestiku ridade arv × ridade arv). Lisanduvalt on suurel andmestikul algoritmi komputatsioonide arv suur ning seetõttu ka ajamahukas.

Tähelepanu tuleb ka juhtida asjaolule, et tegemist on tehislikult genereeritud andmestikuga, kus tehingu- ja kliendiprofiilide jaotus ei pruugi sarnaneda reaalsele finantsinstitutsiooni andmestikule ning andmestiku märgendid ei pruugi olla korrektselt määratud. Tabelis 3.3 on välja toodud näide ühe kliendi tehingutest. *DBSCAN* algoritm määrab ülemineku tavapärastelt väikese suurustega tehingutelt suure väärtusega tehingule kõrge riskiklassi. Ei ole ebatavaline, et isik, kes teeb madala väärtusega tehinguid, teeb vahel mõne suure väärtusega tehingu, aga nimetatud asjaolu võiks kahtlust äratada (märgendiveerus „*SAR* kahtlus“ kõik „Ei“ väärtused).

Tabel 3.3 Kliendi 17297 tehingute osaline väljavõte

Kliendi ID	**Tehingu kuupäev**	**Tehingu väärtus, EUR**	***DBSCAN* riskiklass**	***SAR* kahtlus**
17297	2020-03-12 07:55:40	5	Madal risk	Ei
17297	2020-03-12 14:05:24	5	Madal risk	Ei
17297	2020-03-13 09:52:29	10	Madal risk	Ei
17297	2020-03-13 11:24:51	5	Madal risk	Ei
17297	2020-03-13 13:21:33	1	Madal risk	Ei
17297	2020-03-14 06:14:44	11	Madal risk	Ei
17297	2020-03-14 08:29:30	6	Madal risk	Ei
17297	2020-03-14 14:03:50	3	Madal risk	Ei
17297	2020-03-15 18:14:16	9	Madal risk	Ei
17297	2020-03-16 14:15:13	73	Madal risk	Ei
17297	2022-04-25 06:43:44	900	Kõrge risk	Ei

17297	2022-04-26 20:43:45	900	Keskmine risk	Ei
17297	2022-04-27 06:43:45	900	Kõrge risk	Ei
17297	2022-04-27 19:43:45	900	Kõrge risk	Ei
17297	2022-04-28 06:43:44	900	Kõrge risk	Ei
17297	2022-04-28 19:43:45	900	Kõrge risk	Ei
17297	2022-04-29 07:43:45	900	Kõrge risk	Ei
17297	2022-04-29 20:43:45	900	Kõrge risk	Ei

Magistritöö praktilises osas kasutati varjatud Markovi mudeli funktsioone *scikit-learn hmmlearn* teegist. Varjatud Markovi mudeli varjatud kihi väärtusteks määratakse "aus tehing" ja "rahapesukahtlusega tehing". *DBSCAN* algoritmiga leitud *HMM*-i vaadeldava kihi väärtusteks on "madal risk", "keskmine risk" ja "kõrge risk". Algjaotuse ja ülemineku- ning emiteerimismaaktriksi väärtused on võetud võimalikult sarnased Kasianova [18] töös välja toodud väärtustega. Varjatud Markovi mudeli parameetrid on järgnevad:

- Varjatud kihi võimalikud väärtused:

$$S = (S_1 = \text{“aus tehing”}; S_2 = \text{“rahapesukahtlusega tehing”})$$

- Vaadeldava kihi võimalikud väärtused:

$$(\text{“madal risk”};\ \text{“keskmine risk”};\ \text{“kõrge risk”})$$

- Algjaotus:

$$\pi = (\pi_1 = 0.9; \pi_2 = 0.1)$$

- Üleminekumaatriks:

$$A = \begin{pmatrix} a_{11} = 0{,}8;\ a_{12} = 0{,}2 \\ a_{21} = 0{,}1;\ a_{22} = 0{,}9 \end{pmatrix}$$

- Emiteerimismaatriks:

$$B = \begin{pmatrix} b_1(\text{“madal risk”}) = 0{,}90; & b_1(\text{“keskmine risk”}) = 0{,}05; & b_1(\text{“kõrge risk”}) = 0{,}05 \\ b_2(\text{“madal risk”}) = 0{,}01; & b_2(\text{“keskmine risk”}) = 0{,}01; & b_2(\text{“kõrge risk”}) = 0{,}98 \end{pmatrix}$$

HMM mudeli parameetrid treeniti Baum-Welch algoritmiga ja kõige tõenäolisemat järjestust ennustati Viterbi algoritmiga.

3.4 Tulemused

Käesolevas peatükis antakse ülevaade mudelist, mis koosnes *DBSCAN*-ist ja varjatud Markovi mudelist („*DBSCAN+HMM*“) ja mudelist, mis koosnes reeglitel põhinevast rahapesu riskiskoorist ja varjatud Markovi mudelist („Reeglid+*HMM*“). Lisanduvalt loodi mudelid „reeglid“, „*DBSCAN* turvaline“ ja „*DBSCAN* kiire“ (mudelite loomise loogikast tuleb juttu lõigu alumises osas), et hinnata, kas *HMM*-i kasutamine õigustab mudeli keerukust. Varjatud Markovi mudelite parameetrid peale treenimist Baum-Welch algoritmiga on toodud tabelis 3.4.1.

Tabel 3.4.1. Varjatud Markovi mudelite parameetrid peale Baum-Welch algoritmiga treenimist.

Parameeter	***DBSCAN+HMM***	**Reeglid+*HMM***
Algjaotus	$(\pi_1 = 0{,}872; \pi_2 = 0{,}128)$	$(\pi_1 = 0{,}903; \pi_2 = 0{,}097)$
Ülemineku-maatriks	$\begin{pmatrix} a_{11} = 0{,}993;\ a_{12} = 0{,}007 \\ a_{21} = 0{,}036;\ a_{22} = 0{,}964 \end{pmatrix}$	$\begin{pmatrix} a_{11} = 0{,}966;\ a_{12} = 0{,}034 \\ a_{21} = 0{,}084;\ a_{22} = 0{,}916 \end{pmatrix}$
Emiteerimis-maatriks	$\begin{pmatrix} b_1(\text{"}MR\text{"}) = 0{,}997; & b_2(\text{"}MR\text{"}) = 0{,}113 \\ b_1(\text{"}KeR\text{"}) = 0{,}001; & b_2(\text{"}KeR\text{"}) = 0{,}017 \\ b_1(\text{"}K\tilde{o}R\text{"}) = 0{,}002; & b_2(\text{"}K\tilde{o}R\text{"}) = 0{,}870 \end{pmatrix}$	$\begin{pmatrix} b_1(\text{“}MR\text{”}) = 0{,}986; b_1(\text{“}K\tilde{o}R\text{”}) = 0{,}014 \\ b_2(\text{“}MR\text{”}) = 0{,}068; b_2(\text{“}K\tilde{o}R\text{”}) = 0{,}932 \end{pmatrix}$

Legend: MR – „madal risk“; KeR – „Keskmine risk“; KõR – „kõrge risk“

Tabelist 3.4.1 nähtub, et algjaotuses ja üleminekumaatriksites on väärtused mõlemal mudelil üsna sarnased. Mudelite emiteerimismaatriksi parameetrid on erinevad, sest „*DBSCAN+HMM*“ mudelis oli kolm võimalikku vaadeldavat väärtust ja mudelis „Reeglid+*HMM*“ kaks võimalikku vaadeldavat väärtust. Näiteks kui võtta mudeli „Reeglid+*HMM*“ emiteerimismaatriksist kohalt $b_2(\text{"kõrge risk"})$, on emiteerimistõenäosus 0.932, aga kui võtta mudeli „*DBSCAN+HMM“* emiteerimismaatriksist kohalt $b_2(\text{"keskmine risk"})$ ja $b_2(\text{"kõrge risk"})$ ning tõenäosused omavahel liita, oleks tõenäosuseks 0,887, mis on üsna sarnane mudeli „Reeglid+*HMM*“ $b_2(\text{"}k\tilde{o}rge\ risk\text{"})$ väärtusega. Algolekumaatriksi, üleminekumaatriksi ja emiteerimismaatriks parameetrid on sarnased, sest Baum-Welch algoritm ei garanteeri globaalset maksimumi.

Kui *HMM*-i vaadeldavaks tunnuseks on „kõrge risk“, eeldab varjatud Markovi mudel, et ollakse suure tõenäosusega seisundis „rahapesukahtlusega tehing“ ning riskiskoori „madal risk“ korral eeldatakse, et ollakse seisundis „aus tehing“. Kui mudelis „*DBSCAN+HMM“* on vaadeldavaks tunnuseks „keskmine risk“, eeldab varjatud Markovi mudel, et vaadeldav tunnus on põhjustatud suure tõenäosusega seisundist „rahapesukahtlusega tehing“.

Mudeli headusmõõdikute arvutamiseks teostati 10 eksperimenti, et vähendada headusmõõdikute tulemuste juhuslikkust. Igal eksperimendil jagati andmestik juhuslikult treening- ja testandmestikuks (täites rahapesuklientide võrdse määra ja andmestike ridade arvu sama suurusjärgu kriteeriumi), rakendati kontroll-lauseid või *DBSCAN* algoritmi (vastavalt mudelile) ning teostati *HMM*-i treenimine ning rahapesukahtlusega tehingute ennustamine. Erinevate mudelite headusmõõdikute tulemuste aritmeetiline keskmine ja standardhälve on toodud tabelis 3.4.2.

Mõlema mudeli täpsus on kehv, mistõttu on ka tulemused F1-skooril ja F2-skooril kehvad. Üheltpoolt on tegemist tugevalt kallutatud andmestikega – treening- ja testandmestikus on rahapesukahtlusega tehingute osakaal 0,2% ning piisab väiksest arvust valepositiivsetest, et tulemusmõõdik täpsus näitaks kehva tulemust. Teisalt märkisid mudelid „*DBSCAN*+*HMM*“ ja „reeglid+*HMM*“ keskmiselt testandmestikus 16,46% ja 31,92% kõigist tehingutest rahapesukahtlusega tehinguteks, mis on väga suur osakaal.

Tuleb tõdeda, et tegemist on tehislikult genereeritud andmestikuga, mistõttu ei pruugi andmestik peegeldada reaaleluliste rahapesukahtlusega tehingute olemust ning märgendid andmestikus ei pruugi olla korrektsed. Mõlemad mudelid tuvastasid varjatud Markovi mudeli vaadeldava kihi leidmisel väga suure osakaalu kõrge riskiga tehinguid. Näiteks märkis reeglitel põhinev meetod testandmestikus 23,57% kõigist tehingutest kõrge riskiga tehinguteks ning *DBSCAN*-iga leitud meetodil märgiti 13,13% kõigist tehingutest kõrge riskiga kategooriasse, mis on väga suur osakaal kõrge riskiga tehinguid andmestikus.

Markovi mudelile on oluline n-ö signaal vaadeldavast kihist, mille põhjal eeldatakse, milline seisund võis vaadeldava tunnuse põhjustada. Peatükis 3.2 nägime, et reeglitel põhineva meetodiga leitud *HMM*-i vaadeldav kiht ei suuda piisavalt hästi tuvastada n-ö rahapesukahtlusega tehingute signaali. *DBSCAN*-i parimad parameetrid on leitud kasutades käesolevat andmestikku, aga Kasianova [18] reeglid on leitud küll sama andmebaasi andmete pealt, aga mitte kasutades samu andmeid. Siinkohal tuleb ära märkida, et Kasianova [18] tulemused reaalsel Baltikumi finantsinstitutsiooni andmestikul olid järgnevad: täpsus 67%, saagis 100% ja F1-skoor 81%, mis on tunduvalt parem tulemus, kui seda on käesoleval andmestikul. Käesoleval andmestikul on mudeli „reeglid+*HMM*“ tulemused kõigil mõõdikutel kehvad.

Tabelisse 3.4.2 on loodud kolm lisanduvat mudelit („*DBSCAN* turvaline“, „*DBSCAN* kiire“ ja „Reeglid“) illustreerimaks asjaolu, et vaadeldav kiht mängib olulist rolli varjatud Markovi

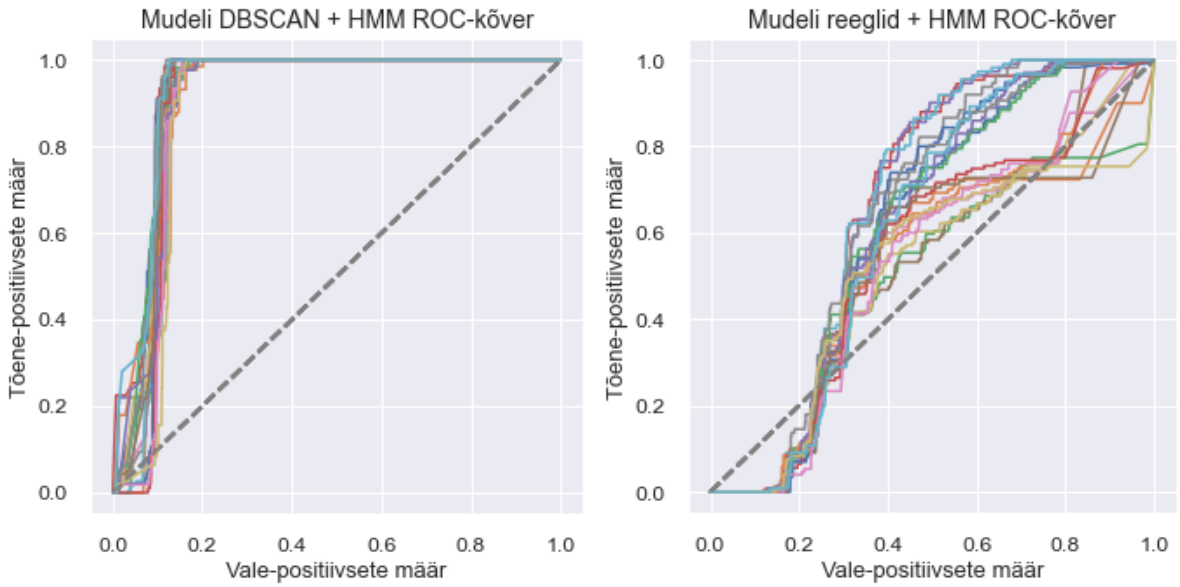
mudelis. Mudel „*DBSCAN* turvaline“ on loodud loogikal, kus riskiskoor „madal risk“ ennustab märgendit „aus tehing“ ja riskiskoorid „keskmine risk“ ja „kõrge risk“ ennustavad märgendit „rahapesukahtlusega tehing“. Mudel „*DBSCAN* kiire“ on loodud samal loogikal, aga riskiskoor „keskmine risk“ ennustab märgendit „aus tehing“. Mudel „Reeglid“ on loodud loogikal, kus riskiskoor „madal risk“ ennustab märgendit „aus tehing“ ja riskiskoor „kõrge risk“ ennustab märgendit „rahapesukahtlusega tehing“. Tabelist 3.4.2 nähtub, et mudelid, mis sisaldasid *DBSCAN* masinõppealgoritmi, on saavutanud sarnaseid tulemusi headusmõõdikutel. Sama kehtib reegleid sisaldanud mudelite kohta.

Tabel 3.4.2. Mudelite headusmõõdikute keskmised tulemused protsentides (st. hälve protsentides)

	Andmestik	Saagis	Täpsus	F1-skoor	F2-skoor	*AUC*
***DBSCAN* turvaline**	Treeningandmestik	96,47 (1,52)	1,56 (0,31)	3,08 (0,60)	7,33, (1,36)	
	Testandmestik	96,11 (1,52)	1,37 (0,30)	2,70 (0,58)	6,47, (1,33)	
***DBSCAN* kiire**	Treeningandmestik	96,14 (1,46)	1,34 (0,31)	3,22 (0,59)	7,66 (1,35)	
	Testandmestik	95,93 (1,73)	1,45 (0,29)	2,86 (0,56)	6,83 (1,27)	
DBSCAN* +*HMM	Treeningandmestik	98,21 (1.26)	1,28 (0.26)	2,53 (0,51)	6,08 (1,19)	90,54 (1,36)
	Testandmestik	99,05 (1,15)	1,49 (0,28)	2,93 (0,55)	7,01 (1,25)	91,23 (1,44)
Reeglid	Treeningandmestik	27,31 (3,62)	0,24 (0,02)	0,47 (0,05)	1,14 (0,12)	
	Testandmestik	28,90 (3,84)	0,23 (0,02)	0,46 (0,05)	1,13 (0,12)	
Reeglid+ *HMM*	Treeningandmestik	21,03 (3,99)	0,18 (0,03)	0,35 (0,07)	0,86 (0,16)	57,23 (6,04)
	Testandmestik	20,78 (4,37)	0,17 (0,03)	0,33 (0,07)	0,80 (0,16)	59,08 (6,03)

Tabelist 3.4.2 nähtub, et mudelid „*DBSCAN* turvaline“ ja „*DBSCAN* kiire“ näitavad peaaegu kõigil headusmõõdikutel (v.a saagis) vähesel määral paremaid tulemusi võrreldes mudeliga „*DBSCAN+HMM*“. Mudel „*DBSCAN+HMM*“ suudab tuvastada kõigil andmestikel kõige enam rahapesukahtlusega tehinguid võrreldes teiste mudelitega (keskmine eksperimendi saagis treening- ja testandmestikul vastavalt 98,21%, 99,05%). Kunas *HMM* suurt lisandväärtust rahapesukahtlusega tehingutele ei anna, ei pruugi mudeli „*DBSCAN+HMM*“ keerukus olla õigustatud. Teisalt võib esimest järku *HMM* osutuda liialt lihtsustatuks tulenevalt Markovi ahela reeglist ning vajalik oleks modelleerida pikemaajalist sõltuvust ahelas.

Joonisel 3.4 on toodud mudelite „reeglid+*HMM*“ ja „*DBSCAN+HMM*“ *ROC*-kõverad kümnel eksperimendil (ühel joonisel nii treening- kui ka testandmestiku *ROC-kõver*). Mudeli „reeglid+*HMM*“ *ROC*-kõver on suuresti varieeruv ja kohati alla 45-kraadi joont. Mudeli „*DBSCAN+HMM*“ *ROC*-kõver on stabiilsem. Nimetatud asjaolule viitab ka tabelis 3.4.2 toodud testandmestiku *AUC*-skoori standardhälve, mille kohaselt on mudeli „*DBSCAN+HMM*“ standardhälve üle nelja korra väiksem võrreldes mudeliga „reeglid+*HMM*“.



Joonis 3.4. Mudelite „reeglid+*HMM*“ ja „*DBSCAN+HMM*“ ROC-kõver.

Mudeli „*DBSCAN+HMM*“ *ROC*-kõvera jooniselt näeme, et 20% vale-positiivsete määraga saavutatakse kõigil andmestikel 100%-ne õigete-positiivsete määr.

4. Kokkuvõte

Rahapesu tuvastamiseks tehingumonitooringus on välja pakutud mitmeid erinevaid meetodeid. Käesoleva töö eesmärgiks oli edasi arendada Kasianova [18] rahapesukahtlusega tehingute tuvastamise mudelit, mis põhines laialdaselt levinud reeglipõhisel rahapesumonitooringu meetodil ning varjatud Markovi mudelil. Kirjanduse ülevaate peatükis selgus, et reeglipõhisel meetodil on mitmeid puuduseid, aga ka positiivseid aspekte. Positiivse aspektina on välja toodud, et reeglitel põhineva rahapesumonitooringu kontroll-lauseid on lihtne mõista, mis võimaldab rahapesukahtlust kontrollival spetsialistil mõista, mis põhjustas reegli rakendumise. Puuduste all on välja toodud näiteks asjaolu, et reeglite väljamõtlemine on tagajärgedega tegelev protsess, keeruline on määrata reeglite rakendumise piirmäärasid, loodud kontroll-laused põhinevad avalikult kättesaadaval informatsioonil jne. Seetõttu otsustati asendada reeglitel põhinev rahapesu riskiskoor *DBSCAN* masinõppealgoritmiga.

Teadaolevalt on tehingumonitooringu andmestikud tugevalt kallutatud, sest rahapesu juhtumid on üldiselt väga harvad. Finantsinstitutsioonile on teada klienti kirjeldavad andmed, mille põhjal on võimalik luua n-ö kliendiprofiil ja kliendi tehinguandmed ehk andmed, mis iseloomustavad kliendi poolt tehtavaid tehinguid. Peamine loogika *DBSCAN*-iga rahapesu riskiskoori leidmisel seisneb selles, et sarnase kliendiprofiiliga isikud peaksid tegema sarnaseid tehinguprofiiliga tehinguid. Kui kliendi profiil ei lähe kokku kliendi tehinguprofiiliga, on tegemist nö. anomaaliaga. *DBSCAN* algoritmiga on võimalik leida teadmatu arv klastreid tuginedes naabruses olevate punktide tihedusele ja minimaalsele klastri punktide arvule. Kui tehing ei kuulu mitte ühtegi klastrisse, on tegemist anomaaliaga ehk kliendi tehingu- ja kliendiprofiil erineb teistest klientidest. Nimetatud anomaaliad saavad rahapesu riskiskooriks „kõrge risk“. Klastrid, mis ei ole märgitud müraks, aga sisaldavad vähesel määral tehinguid, määratakse riskiskooriks „keskmine risk“. Ülejäänud, klastritesse kuuluvad tehingud saavad riskiskooriks „madal risk“. *DBSCAN*-iga leitud rahapesu riskiskoor on varjatud Markovi mudeli vaadeldavaks kihiks.

Kasianova [18] poolt välja töötatud reeglid rahapesu riskiskoori leidmiseks põhinevad sama andmebaasi andmetel, aga mitte samadel tehingutel. Kirjanduse ülevaate peatükis ja ka Kasianova [18] enda töös välja toodud puudused seoses reeglitel põhineva komposiitskooriga (ei ole dünaamiline, põhineb hetkelisel üldisel arusaamal rahapesukahtlusega tehingute omadustest ning reegleid tuleks pidevalt uuendada ja täiendada) said kinnitust ka käesolevas töös. Kontroll-lausete põhimõttel leitud rahapesu riskiskoor ei suutnud käesoleval andmestikul tuvastada rahapesukahtlusega tehinguid, mistõttu on ka varjatud Markovi mudeli tulemused

headusmõõdikutel kehvad. Eksperimenti läbi viies sai kinnitust asjaolu, et tunnuste positsioon vaadeldavas kihis on varjatud Markovi mudelis olulise tähtsusega. *DBSCAN*-iga leitud varjatud Markovi mudeli vaadeldav kiht suutis paremini iseloomustada rahapesukahtlusega tehinguid ning seetõttu oli mudeli "*DBSCAN+HMM*" tulemused ka paremad. "*DBSCAN+HMM*" suudab edukalt tuvastada rahapesukahtlusega tehinguid (saagis kõrge), aga seda suure hulga vale-positiivsete arvelt (madal täpsus). Kasianova [18] poolt loodud mudel "Reeglid+*HMM*" saavutas Baltikumi finantsinstitutsiooni andmestikul saagiseks 100% ja F1-skooriks 0.81. Tulevikus on soov testida mudelit "*DBSCAN+HMM*" ka reaalelulisel andmestikul, sest käesoleva töö andmestiku rahapesukahtlusega tehingud ei pruugi peegeldada reaalelulisi rahapesukahtlusega tehinguid.

Töö esmaseks edasiarenduse ideeks oleks leida parem headusmõõdik *DBSCAN*-i parimate hüperparameetrite leidmiseks. Hetkel kasutati *DBSCAN*-i hüperparameetrite leidmiseks andmestiku märgendeid, aga kirjanduse ülevaate peatükis on viidatud probleemile, et tehingumonitooringu andmestiku märgendid võivad suure tõenäosusega olla ebakorrektsed. Lisaks on valikukriteeriumites tingimus, mis nõuab, et vähemalt 80% rahapesukahtlusega tehingutest peab olema määratud müraks ehk saama rahapesu riskiskooriks „kõrge risk". Kaheksakümne protsendi lävend ei ole reaalelulises riskipõhises tehingumonitooringus aktsepteeritav. Seetõttu tuleks kasutada *DBSCAN*-i parimate hüperparameetrite leidmisel headusmõõdikuid, mis ei põhine andmestiku märgenditel.

Üheks lähenemisideeks oleks kombineerida rahapesumonitooringu reegleid *DBSCAN*-iga. *DBSCAN* võimaldaks vähendada reeglitel põhineva rahapesu monitooringusüsteemi riske, aga samas säiliks valdkonnaekspertide teadmine rahapesu olemusest. Näiteks oleks võimalik rakendada *DBSCAN*-i andmestikule, mis on koostatud reeglitel põhineva rahapesu riskiskoori alusel. Ühte andmestikku koondatakse kokku sarnaste rahapesu riskiskooriga isikud, mis võimaldaks vähendada *DBSCAN*-i naabruskonna parameetrit. Lisanduvalt võimaldaks nimetatud lähenemine maandada riski, kus isik satub juhuslikult endale mitteomasesse kliendi- ja tehinguprofiiliga klastrisse. Antud täiendused aga ei mahtunud antud lõputöö raamesse.

5. Viidatud kirjandus

[1] O. Lebid and O. Veits, “Search for statistically approved criteria for identifying money laundering risk,” Banks and Bank Systems, vol. 15, no. 4, pp. 150–163, Detsember 2020, doi: 10.21511/bbs.15(4).2020.13 (Jaanuar 07, 2022).

[2] M. Tiwari, A. Gepp, and K. Kumar, “A review of money laundering literature: the state of research in key areas,” Pacific Accounting Review, vol. 32, no. 2, pp. 271–303, Märts 2020, doi: 10.1108/par-06-2019-0065 (Jaanuar 07, 2022).

[3] Finantsinspektsioon, “Swedbank saab trahvi ja ettekirjutuse rahapesu vastu võitlemise reeglite rikkumise eest,” Finantsinspektsioon, Märts 19, 2020. https://www.fi.ee/et/uudised/swedbank-saab-trahvi-ja-ettekirjutuse-rahapesu-vastu-voitlemise-reeglite-rikkumise-eest (Jaanuar 07, 2022).

[4] Finantsinspektsioon, “AS SEB Pank sai rahapesu tõkestamise reeglite rikkumise eest trahvi,” Finantsinspektsioon, Juuni 25, 2020. https://www.fi.ee/et/uudised/seb-pank-sai-rahapesu-tokestamise-reeglite-rikkumise-eest-trahvi (Jaanuar 07, 2022).

[5] Finantsinspektsioon, “Finantsinspektsioon suunas LHV Panka parendama rahapesu tõkestamise kontrolli süsteeme,” Finantsinspektsioon, Märts 09, 2021. https://www.fi.ee/et/uudised/finantsinspektsioon-suunas-lhv-panka-parendama-rahapesu-tokestamise-kontrolli-susteeme (Jaanuar 07, 2022).

[6] Finantsinspektsioon, “Balti riikide järelevalveasutused suunavad Luminori parendama rahapesu tõkestamise süsteeme,” Finantsinspektsioon, August 24, 2021. https://www.fi.ee/et/uudised/balti-riikide-jarelevalveasutused-suunavad-luminori-parendama-rahapesu-tokestamise-susteeme (Jaanuar 07, 2022).

[7] Finantsinspektsioon, “Krediidiasutuse bilanss,” Finantsinspektsioon. https://www.fi.ee/koond/bilanss_kred.php (Jaanuar 07, 2022).

[8] D. V. Kute, B. Pradhan, N. Shukla, and A. Alamri, “Deep Learning and Explainable Artificial Intelligence Techniques Applied for Detecting Money Laundering–A Critical Review,” IEEE Access, vol. 9, pp. 82300–82317, 2021, doi: 10.1109/access.2021.3086230 (Jaanuar 17, 2022)

[9] Euroopa Komisjon, “Beating financial crime: Commission overhauls anti-money laundering and countering the financing of terrorism rules,” European Commission - European Commission, Detsember 20, 2021. https://ec.europa.eu/commission/presscorner/detail/en/ip_21_3690 (Jaanuar 07, 2022).

[10] A. Shokry, M. A. Rizka, and N. M. Labib, “Counter Terrorism Finance by Detecting Money Laundering Hidden Networks Using Unsupervised Machine Learning Algorithm,” Juuli 2020. Saadaval: http://dx.doi.org/10.33965/ict_csc_wbc_2020_202008l012 (Jaanuar 17, 2022).

[11] Riigi Teataja, “Rahapesu ja terrorismi rahastamise tõkestamise seadus–Riigi Teataja,” Riigi Teataja, Juuni 02, 2021. https://www.riigiteataja.ee/akt/102062021009?leiaKehtiv (Jaanuar 07, 2022)

[12] Finantsinspektsioon, “Finantsinspektsiooni järelevalvepoliitika rahapesu ja terrorismi rahastamise tõkestamisel,” Finantsinspektsioon, November 2018 (Jaanuar 17, 2022)

[13] U. G. Ketenci, T. Kurt, S. Önal, C. Erbil, H. Ş. İlhan, and S. Aktürkoğlu, “A Time-Frequency Based Suspicious Activity Detection for Anti-Money Laundering,” IEEE Access, vol. 9, pp. 59957–59967, Aprill 2021, doi: 10.1109/ACCESS.2021.3072114. (Veebruar 13, 2022).

[14] Y. Zhang and P. Trubey, “Machine Learning and Sampling Scheme: An Empirical Study of Money Laundering Detection,” Computational Economics, vol. 54, no. 3, pp. 1043–1063, Oktoober 2018, doi: 10.1007/s10614-018-9864-z. (Veebruar 13, 2022).

[15] A. Gupta, D. N. Dwivedi, and A. Jain, “Threshold fine-tuning of money laundering scenarios through multi-dimensional optimization techniques,” Journal of Money Laundering Control, vol. 25, no. 1, pp. 72–78, Märts 2021, doi: 10.1108/jmlc-12-2020-0138. (Märts 16, 2022).

[16] M. Jullum, A. Løland, R. B. Huseby, G. Ånonsen, and J. Lorentzen, “Detecting money laundering transactions with machine learning,” Journal of Money Laundering Control, vol. 23, no. 1, pp. 173–186, Jaanuar 2020, doi: 10.1108/jmlc-07-2019-0055. (Mai 5, 2022).

[17] L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” Proceedings of the IEEE, vol. 77, no. 2, pp. 257–286, 1989, doi: 10.1109/5.18626. (Juuni 20, 2022).

[18] K. Kasianova, “Detecting Money Laundering Using Hidden Markov Model,” Tartu Ülikool, pp. 1–47, 2020. (Märts 16, 2022).

[19] “Welcome to Python.org,” Python.org. https://www.python.org/ (August 08, 2022).

[20] “pandas,” Python Data Analysis Library. https://pandas.pydata.org/

[21] “NumPy.” https://numpy.org/ (August 08, 2022).

[22] “hmmlearn — hmmlearn 0.2.7.post13+g6d3900d documentation.” https://hmmlearn.readthedocs.io/en/latest/ (August 08, 2022).

[23] “sklearn.cluster.DBSCAN,” scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html (August 08, 2022).

[24] “sklearn.preprocessing.StandardScaler,” scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html (August 08, 2022).

[25] “3.3. Metrics and scoring: quantifying the quality of predictions,” scikit-learn. https://scikit-learn.org/stable/modules/model_evaluation.html (August 08, 2022).

[26] Al-Rashidi, “‘Indirect Method of Proof’ and the Kuwaiti Anti-Money Laundering Law: A Lesson from the UK,” Criminal Law Forum, vol. 32, no. 3, pp. 405–433, Märts 2021, doi: 10.1007/s10609-021-09415-3. (Märts 25, 2022).

[27] A. Alshantti and A. Rasheed, “Self-Organising Map Based Framework for Investigating Accounts Suspected of Money Laundering,” Frontiers in Artificial Intelligence, vol. 0, Jaanuar 2021, doi: 10.3389/frai.2021.761925. (Detsember 20, 2021).

[28] J. Walker and B. Unger, “Measuring Global Money Laundering: ‘The Walker Gravity Model,’” Review of Law & Economics, vol. 5, no. 2, Jaanuar 2009, doi: 10.2202/1555-5879.1418. (Detsember 20, 2021).

[29] International Monetary Fund, “Exchange Rate Report Wizard,” Imf.org. https://www.imf.org/external/np/fin/ert/GUI/Pages/Report.aspx?CU=%27EUR%27,%27USD%27,%27AUD%27&EX=REP&P=DateRange&Fr=632085120000000000&To=632400480000000000&CF=Compressed&CUF=Period&DS=Ascending&DT=Blank (Veebruar 13, 2022). (Detsember 20, 2021).

[30] L. Butgereit, “Anti Money Laundering: Rule-Based Methods to Identify Funnel Accounts,” IEEE Xplore, Märts 10, 2021. https://ieeexplore.ieee.org/document/9394990

[31] Ühinenud Rahvaste Organisatsioon, “Money Laundering,” United Nations : Office on Drugs and Crime. https://www.unodc.org/unodc/en/money-laundering/overview.html (Jaanuar 07, 2022). (Veebruar 15, 2022).

[32] J. Han, Y. Huang, S. Liu, and K. Towey, “Artificial intelligence for anti-money laundering: a review and extension,” Digital Finance, vol. 2, no. 3–4, pp. 211–239, Juuni 2020, doi: 10.1007/s42521-020-00023-1.

[33] Financial Action Task Force, “Countries,” Financial Action Task Force (FATF). https://www.fatf-gafi.org/countries/ (Jaanuar 15, 2022).

[34] E. L. Paula, M. Ladeira, R. N. Carvalho, and T. Marzagao, “Deep Learning Anomaly Detection as Support Fraud Investigation in Brazilian Exports and Anti-Money

Laundering," Detsember 2016. (Jaanuar 16, 2022). Saadaval: http://dx.doi.org/10.1109/icmla.2016.0172

[35] A. Tundis, S. Nemalikanti, and M. Mühlhäuser, "Fighting organized crime by automatically detecting money laundering-related financial transactions," The 16th International Conference on Availability, pp. 1–10, August 2021, (Jaanuar 08, 2022). doi: https://doi.org/10.1145/3465481.3469196.

[36] Y. Feng et al., "Anti-money Laundering (AML) Research: A System for Identification and Multi-classification," in Web Information Systems and Applications, Cham: Springer International Publishing, 2019, pp. 169–175. (Jaanuar 08, 2022). Saadaval: http://dx.doi.org/10.1007/978-3-030-30952-7_19

[37] Financial Conduct Authority, "Digital sandbox pilot: FCA DataSprint," FCA, August 27, 2020. https://www.fca.org.uk/firms/innovation/digital-sandbox-pilot-datasprint (Märts 17, 2022).

[38] Finantsinspektsioon, "Rahapesu ja terrorismi rahastamise tõkestamise meetmed krediidi- ja finantseerimisasutustes," Juuli 2013. (November 02, 2021)

[39] Z. Chen, L. D. Van Khoa, E. N. Teoh, A. Nazir, E. K. Karuppiah, and K. S. Lam, "Machine learning techniques for anti-money laundering (AML) solutions in suspicious transaction detection: a review," Knowledge and Information Systems, vol. 57, no. 2, pp. 245–285, Veebruar 2018, doi: 10.1007/s10115-017-1144-z. (Juuli 2, 2022)

[40] A. Salehi, M. Ghazanfari, and M. Fathian, "Data Mining Techniques for Anti Money Laundering," International Journal of Applied Engineering Research, vol. 12, no. 20, pp. 10084–10094, 2017. (November 02, 2021)

[41] A. A. S. Alsuwailem and A. K. J. Saudagar, "Anti-money laundering systems: a systematic literature review," Journal of Money Laundering Control, vol. 23, no. 4, pp. 833–848, May 2020, doi: 10.1108/jmlc-02-2020-0018. (Juuli 2, 2022)

[42] A. Sudjianto, S. Nair, M. Yuan, A. Zhang, D. Kern, and F. Cela-Díaz, "Statistical Methods for Fighting Financial Crimes," Technometrics, vol. 52, no. 1, pp. 5–19, Veebruar 2010, doi: 10.1198/tech.2010.07032. (Detsember 20, 2021)

[43] Leite, Albuquerque, and Pinheiro, "Application of Technological Solutions in the Fight against Money Laundering—A Systematic Literature Review," Applied Sciences, vol. 9, no. 22, p. 4800, November 2019, doi: 10.3390/app9224800. (Veebruar 06, 2022)

[44] A. Srivastava, A. Kundu, S. Sural, and A. K. Majumdar, "Credit Card Fraud Detection Using Hidden Markov Model," IEEE Transactions on Dependable and Secure Computing, vol. 5, no. 1, pp. 37–48, Jaanuar 2008, doi: 10.1109/tdsc.2007.70228. (Veebruar 06, 2022)

[45] Divya. Iyer, A. Mohanpurkar, S. Janardhan, D. Rathod, and A. Sardeshmukh, “Credit card fraud detection using Hidden Markov Model,” Detsember 2011. (Veebruar 06, 2022). Saadaval: http://dx.doi.org/10.1109/wict.2011.6141395

[46] E. B. Nkemnole and A. A. Akinsete, “Hidden Markov Model using transaction patterns for ATM card fraud detection,” Theoretical and Applied Economics, vol. 0(4(629), pp. 51–70, 2021, doi: http://store.ectap.ro/articole/1566.pdf. (Veebruar 04, 2022)

[47] A. A. A. Danaa, M. I. Daabo, and A. Abdul-Barik, “Detecting Electronic Banking Fraud on Highly Imbalanced Data using Hidden Markov Models,” Earthline Journal of Mathematical Sciences, vol. 7, no. 2, pp. 315–332, September 2021, doi: 10.34198/ejms.7221.315332. (Veebruar 03, 2022)

[48] I. Aghahasanli, “Detecting money laundering in transaction monitoring using hidden Markov model,” DSpace, Jaanuar 01, 2021. http://hdl.handle.net/10062/72862 (Märts 15, 2022).

[49] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,” KDD, 1996. (Juuli 26, 2022)

[50] J. Fürnkranz et al., “Manhattan Distance,” in Encyclopedia of Machine Learning, Boston, MA: Springer US, 2011, pp. 639–639. (Juuli 26, 2022). Saadaval: http://dx.doi.org/10.1007/978-0-387-30164-8_506

[51] T. Armstrong, A. M. Brown, P. M. Chadwick, and S. J. Nolan, “The detection ofFermiAGN above 100 GeV using clustering analysis,” Monthly Notices of the Royal Astronomical Society, vol. 452, no. 3, pp. 3159–3166, Juuli 2015, doi: 10.1093/mnras/stv1398. (August 05, 2022).

[52] B. Mor, S. Garhwal, and A. Kumar, “A Systematic Review of Hidden Markov Models and Their Applications,” Archives of Computational Methods in Engineering, vol. 28, no. 3, pp. 1429–1448, May 2020, doi: 10.1007/s11831-020-09422-4. (Juuli 26, 2022)

[53] Yahoo Finance, “Swedbank AB A-seeria ajalooline aktsiahind,” Yahoo Finance. https://finance.yahoo.com/quote/SWED-A.ST/history?period1=1546300800&period2=1564531200&interval=1d&filter=history&frequency=1d&includeAdjustedClose=true (August 05, 2022).

[54] Reuters, “Swedbank slips 9 percent on lingering money laundering worries,” CNBC, Veebruar 21, 2019. Saadaval: https://www.cnbc.com/2019/02/21/swedbank-shares-slip-on-lingering-money-laundering-worries.html (August 05, 2022).

[55] R. Milne, “Swedbank failings on €37bn of transactions revealed in report,” Financial Times, Märts 23, 2020. (August 05, 2022). Saadaval: https://www.ft.com/content/86d4736a-6cd2-11ea-89df-41bea055720b (August 05, 2022).

[56] NAICS Association, “High Risk/Cash Intensive NAICS Industries*,” Naics.com. https://www.naics.com/wp-content/uploads/2014/10/NAICS-ASSOCIATION-High-Risk-and-Cash-Intensive-NAICS-Codes-List.pdf (Märts 16, 2022).

Lisad

I. Swedbank AB rahapesuskandaali mõju aktsiahinnale

Joonis I. Andmed: Yahoo Finance [53], autori koostatud

II. Lisanduvate tunnuste leidmine

Lisanduvalt eksisteerivatele tunnustele arvutati tehinguandmestikku igale tehingule järgnevad tunnused:

1. *sender_country* - saatja kontonumbri riik eraldatud kontonumbri algusest.
2. *receiver_country* - saaja kontonumbri riik eraldatud tähed kontonumbri alguses.
3. *hour* – tehingu teostamise kellaaeg tundides.
4. *transaction_time_high_risk* - binaarne tunnus, mis näitab, kas tehing on teostatud enne kella 7-t või peale kella 21-t.
5. *time_difference* - kliendi kahe tehingu vahele jääv aeg minutites. Kliendi esimesed tehingud said imputeeritud andmestiku keskmise tehingute vahele jääva ajaga.
6. *7d_incoming_amount* - kliendi seitsme päeva sissetulevate maksete libisev summa eurodes.
7. *7d_outgoing_amount* - kliendi seitsme päeva väljaminevate maksete libisev summa eurodes.
8. *30d_incoming_amount* - kliendi kolmekümne päeva sissetulevate maksete libisev summa eurodes.
9. *30d_outgoing_amount* - kliendi kolmekümne päeva väljaminevate maksete libisev summa eurodes.
10. *7d_incoming_count* - kliendi seitsme päeva libisev sissetulevate maksete arv.
11. *7d_outgoing_count* - kliendi seitsme päeva libisev väljaminevate maksete arv.
12. *30d_incoming_count* - kliendi kolmekümne päeva libisev sissetulevate maksete arv.

13. *30d_outgoing_count* - kliendi kolmekümne päeva libisev väljaminevate maksete arv.
14. *7d_incoming_accounts* - kliendi seitsme päeva libisev sissetulevate unikaalsete osapoolte arv.
15. *7d_outgoing_accounts* - kliendi seitsme päeva libisev väljaminevate unikaalsete osapoolte arv.
16. *30d_incoming_accounts* - kliendi kolmekümne päeva libisev sissetulevate unikaalsete osapoolte arv.
17. *30d_outgoing_accounts* - kliendi kolmekümne päeva libisev väljaminevate unikaalsete osapoolte arv.
18. *7d_incoming_std* - kliendi seitsme päeva sissetulevate maksete libisev standardhälve.
19. *7d_outgoing_std* - kliendi seitsme päeva väljaminevate maksete libisev standardhälve.
20. *30d_incoming_std* - kliendi kolmekümne päeva sissetulevate maksete libisev standardhälve.
21. *30d_outgoing_std* - kliendi kolmekümne päeva väljaminevate maksete libisev standardhälve.
22. *7d_incoming_std* - kliendi seitsme päeva sissetulevate maksete libisev keskmine.
23. *7d_outgoing_std* - kliendi seitsme päeva väljaminevate maksete libisev keskmine.
24. *30d_incoming_std* - kliendi kolmekümne päeva sissetulevate maksete libisev keskmine.
25. *30d_outgoing_std* - kliendi kolmekümne päeva väljaminevate maksete libisev keskmine.
26. *cash_deposit* - kliendi sularahadeposiitide arv vaatlusperioodis kokku.
27. *cash_withdrawal* - kliendi sularahaväljavõtete arv vaatlusperioodis kokku.
28. *domestic_transfer* - kliendi siseriiklike maksete arv vaatlusperioodis kokku.
29. *internal_transfer* - kliendi pangasiseste maksete arv vaatlusperioodis kokku.
30. international_transfer - kliendi rahvusvaheliste maksete arv vaatlusperioodis kokku.

Lisanduvalt eksisteerivatele tunnustele arvutati/genereeriti *CRM* andmestikust järgnevad andmeatribuudid:

1. *birthyear* - füüsilisest isikust kliendi sünniaasta.
2. *customer_age* - füüsilisest isikust kliendi vanus aastates.
3. *company_age* - juriidilisest isikust kliendi vanus aastates.
4. *customer_country_high_risk* - binaarne tunnus, mis näitab, kas kliendi kodakondsus on AML kõrge riskiga riikide nimekirjas.
5. *country_of_residence_high_risk* - binaarne tunnus, mis näitab, kas kliendi residentsusriik on AML kõrge riskiga riikide nimekirjas.
6. *industry_high_risk* - juriidilise isiku tegevusvaldkond on AML kõrge riskida. tegevusvaldkondade hulgas.

III. Kõrgendatud rahapesuriskiga riikide nimekiri

Lisanduvate tunnuste genereerimiseks kasutatud abitabel.

Tabel III. Kõrgendatud rahapesuriskiga riikide nimekiri

Tunnuskood	Riigi nimi
AF	Afghanistan
AI	Anguilla
AG	Antigua and Barbuda
AW	Aruba

PT-20	Azores
BS	Bahamas
BH	Bahrain
BB	Barbados
BY	Belarus
BZ	Belize
BM	Bermuda
BA	Bosnia and Herzegovina
BN	Brunei Darussalam
BF	Burkina Faso
KH	Cambodia
KY	Cayman Islands
CF	Central African Republic
CG	Congo
CK	Cook Islands
CI	Cote d'Ivoire
CU	Cuba
CW	Curacao
KP	Democratic People's Republic of Korea (DPRK)
DJ	Djibouti Republic
DM	Dominica
DO	Dominican Republic
EC	Ecuador
ER	Eritrea
ET	Ethiopia
GH	Ghana
GI	Gibraltar
WG	Grenada
GT	Guatemala
GG	Guernsey
DW	Guinea Bissau
GY	Guyana
HK	Hong Kong
IR	Iran
IQ	Iraq
IM	Isle of Man
JA	Jamaica
JE	Jersey
KE	Kenya
LA	Lao People's Democratic Republic
LB	Lebanon
LR	Liberia
LY	Libya
MO	Macao
PT-30	Madeira
MV	Maldives

MH	Marshall Islands
MU	Mauritius
MS	Montserrat
MZ	Mozambique
MM	Myanmar (Burma)
NA	Namibia
NR	Nauru
NC	New Caledonia
NU	Niue
PK	Pakistan
PW	Palau
PS	Palestine State of
PA	Panama
RU	Russian Federation
SH	Saint Helena, Ascension and Tristan da Cunha
KN	Saint Kitts and Nevis
PM	Saint Pierre and Miquelon
VC	Saint Vincent and the Grenadines
WS	Samoa
RS	Serbia
SC	Seychelles
SX	Sint Maarten
SO	Somalia
SS	South Sudan
LK	Sri Lanka
SD	Sudan
SZ	Swaziland (Eswatini)
SY	Syria
PF	Tahiti (French Polynesia)
TL	Timor-Leste
TO	Tonga
TT	Trinidad and Tobago
TN	Tunisia
TC	Turks and Caicos Islands
UG	Uganda
UY	Uruguay
VU	Vanuatu
VE	Venezuela
VG	Virgin Islands, British
VI	Virgin Islands, U.S.
YE	Yemen

Allikas: Kasianova [18], autori kohandatud

IV. Kõrgendatud rahapesuriskiga tegevusvaldkondade nimekiri

Lisanduvate tunnuste arvutamiseks kasutatud abitabel.

Tabel IV. Kõrgendatud rahapesuriskiga tegevusvaldkondade nimekiri

Tegevusvaldkond	Risk	Esimese taseme kood	Teise taseme kood	Kolmanda taseme kood
Auto Dealers	high_risk	44111	4411	441
Auto Dealers	high_risk	44112	4411	441
Recreational Vehicles	high_risk	44121	4412	441
Motorcycle	high_risk	44122	4412	441
Boat Dealer	high_risk	44122	4412	441
Aircraft Dealer	high_risk	44122	4412	441
Automotive Parts	high_risk	44131	4413	441
Automotive Repair	high_risk	81111	8111	811
Automotive Repair	high_risk	81112	8111	811
Automotive Repair	high_risk	81122	8112	811
Casinos	high_risk	71321	7132	713
Travel agency	high_risk	56151	5615	561
Check Cashing	money_services	52239	5223	522
Currency Exchange	money_services	52313	5231	523
Electronic Fund Transfer	money_services	52232	5223	522
Money Transmitter	money_services	52239	5223	522
Money Transmitter	money_services	52229	5222	522
Money Order Sales	money_services	52239	5223	522
Pawn Shop	non_bank	52229	5222	522
Jewelry Store	non_bank	42394	4239	423
Jewelry Store	non_bank	44831	4483	448
Jewelry/gem	non_bank	44831	4483	448
Consumer Loans	non_bank	52229	5222	522
Convenience Store	cash_intensive	44512	4451	445
Convenience W/Gas	cash_intensive	44711	4471	447
Restaurant	cash_intensive	72211	7221	722
Restaurant	cash_intensive	72221	7222	722
Liquor Store	cash_intensive	44531	4453	445
Tobacco Distributors	cash_intensive	42494	4249	424
Vending Machine	cash_intensive	45421	4542	454
Parking garage	cash_intensive	81293	8129	812
Retail	cash_intensive	44211	4421	442
Retail	cash_intensive	44421	4442	444
Retail	cash_intensive	45111	4511	451
Retail	cash_intensive	45211	4521	452
Retail	cash_intensive	44221	4422	442
Retail	cash_intensive	44422	4442	444
Retail	cash_intensive	44229	4422	442
Retail	cash_intensive	44611	4461	446

Retail	cash_intensive	45112	4511	451
Retail	cash_intensive	45299	4529	452
Retail	cash_intensive	44311	4431	443
Retail	cash_intensive	44812	4481	448
Retail	cash_intensive	45311	4531	453
Retail	cash_intensive	44815	4481	448
Retail	cash_intensive	45113	4511	451
Retail	cash_intensive	45321	4532	453
Retail	cash_intensive	44312	4431	443
Retail	cash_intensive	44819	4481	448
Retail	cash_intensive	45114	4511	451
Retail	cash_intensive	45322	4532	453
Retail	cash_intensive	44411	4441	444
Retail	cash_intensive	44821	4482	448
Retail	cash_intensive	45121	4512	451
Retail	cash_intensive	45331	4533	453
Retail	cash_intensive	44412	4441	444
Retail	cash_intensive	45122	4512	451
Retail	cash_intensive	45399	4539	453
Private ATMs	ATMS	45421	4542	454
Non-governmental charity	charity	81321	8132	813

Allikas: NAICS [56], autori kohandatud

I. Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Kerdo Puusalu,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose

Rahapesukahtlusega tehingute tuvastamine juhendamata varjatud markovi ahelate abil,

mille juhendaja on Kaur Lumiste,

reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Kerdo Puusalu

08.08.2022