TARTU ÜLIKOOL
Loodus- ja täppisteaduste valdkond
Arvutiteaduse instituut
Informaatika õppekava

Joosep Tavits

Andmete üldistamisel põhineva anonüümimise kvaliteedi hindamine

Bakalaureusetöö (9 EAP)

Juhendaja(d): Sulev Reisberg, PhD

Tartu 2022

Andmete üldistamisel põhineva anonüümimise kvaliteedi hindamine

Lühikokkuvõte:
Isikuandmete väljastamisel kolmandatele osapooltele tuleb tagada andmetes sisalduvate isikute kaitse. Levinud meetodiks on sel juhul kasutada andmete anonüümimist, mille tulemusel ei ole üksikisikud andmestikus enam tuvastatavad. Lisaks viimastel aastakümnetel avaldatud erinevatele anonüümimismeetoditele on publitseeritud ka hulgaliselt erinevaid anonüümimise kvaliteedi hindamise mõõdikuid. Seetõttu ei ole nende mõõdikute seast valikute tegemine ilma valdkonna ekspertteadmisteta triviaalne. Selles bakalaureusetöös käsitletakse andmete anonüümimist ja selle kvaliteedi mõõtmist ühe tüüpilise tihedalt andmeväljastustega kokku puutuva organisatsiooni vaatest. Töö raames tuvastatakse kasutatavate anonüümimismeetodite ja lõppkasutaja kirjelduse põhjal sobiv alamhulk anonüümimise kvaliteedi mõõdikud. Seejärel realiseeritakse need mõõdikud olemasoleva anonüümimistarkvara väljundi analüüsimise komponendina, et kirjeldada anonüümimise tulemusel alles jäänud riske ja andmestikus toimunud muudatusi.

Võtmesõnad:
anonüümimine, privaatsus, mõõdikud, automatiseerimine

CERCS: P175 Informaatika, süsteemiteooria

Analyzing the quality of generalization based data anonymization

Abstract:
When data holders release personal data, it is required that the privacy of data subjects has to be preserved. A widespread method for counteracting the accompanying extensive responsibilities is data anonymization, which largely removes the connection between data records and individuals. In addition to several different anonymization algorithms created throughout the past two decades, a variety of metrics for estimating the quality of the performed anonymization have been published. Therefore choosing the best metrics for specific types of anonymization without extensive knowledge of the field is not a trivial problem. This bachelor's thesis researches data anonymization and ways to measure its quality based on a typical organization that often releases personal data. The focus of this thesis is to identify an optimal subset of anonymization quality metrics based on the description of the end user and the anonymization methods used. Finally the subset of metrics is implemented as a software component, which is then integrated into existing anonymization software as an output validation component in order to describe the remaining risks and changes in the dataset post-anonymization.

Keywords:
anonymization, privacy, metrics, automation

CERCS: P175 Informatics, systems theory

Sisukord

1 Sissejuhatus

Tänapäeva ühiskond on oma olemuselt andmetel põhinev (i. k *data-driven*). See tähendab, et osutatavate teenuste ja otsuste tegemisel tuginetakse põhiliselt andmetele. Andmed on aga tihti otseselt seotud füüsiliste isikutega ning võivad olla tundlikud, mis tähendab, et nende avalikustamine võib kahjustada vastavate isikute mainet vms. Sellest tulenevalt on isikuandmete käitlemine tugevalt regulatsioonidega piiritletud. Selleks, et vähendada füüsiliste isikute seoseid andmetega, on välja pakutud mitmeid lahendusi. Üheks neist on anonüümimine. Seejuures anonüümimine ei ole täiuslik protseduur. See tähendab, et ka anonüümimise kvaliteeti on vaja kuidagi mõõta. Anonüümimise kvaliteedi hindamiseks on loodud väga palju erinevaid mõõdikuid, mis kõik on ühes või teises kontekstis paremad kui teised.

Selles bakalaureusetöös uuritaksegi erinevaid anonüümimise kvaliteedi hindamise mõõdikuid, mis laias laastus jagunevad kaheks - informatsioonikao mõõdikud ja privaatsuse tagamist kontrollivaid mõõdikuid. Neist valitakse välja sobivaimad, võttes arvesse Health Sense projekti raames paika pandud piiranguid. Seejärel implementeeritakse need eraldiseisva tarkvarakomponendina, mis lõpuks integreeritakse Health Sense projekti raames loodava andmete hägustaja lähtekoodi.

Töö põhieesmärk on valida välja mõõdikute komplekt üldistamisel põhineva anonüümimise kvaliteedi hindamiseks, säilitades seejuures tavakasutajale lihtsasti arusaadavat väljundit. Alameesmärk on valitud mõõdikute realiseerimine eraldi tarkvarakomponendina Health Sense projektis. Taustapeatükis kirjeldatakse projekti motivatsiooni, lähteprobleeme, andmekaitse valdkonna hoiakuid anonüümimise osas ja laialdaselt levinud anonüümimismeetodeid ning neile vastavaid mõõdikuid. Metoodika peatükis kirjeldatakse lühidalt tulemuste realiseerimise plaane ja põhjendatakse kasutatavate meetodite valikuid. Tulemuste ja arutelu peatükis arutletakse tulemuste üle ning antakse lühiülevaade loodava tarkvarakomponendi lõplikust väljundist ja integreerimisest.

2 Taust

Selles peatükis tutvustatakse põhilisi käsitletavaid probleeme, analüüsitavaid andmeid ja projekti, mille raames töö tulemus vajalik on. Põhiline eesmärk on andmete väljastamise protseduuri automatiseerimise vajaduse tutvustamine. Teisesed eesmärgid on levinud privaatsuse tagamise meetodite formaalne kirjeldamine ja loodava lahenduse ning selle realiseerimisega seotud organisatsioonide tutvustamine.

2.1 Andmete väljastamine ja seotud probleemid

Isikuandmed on tänapäeval meditsiini-, reklaami-, jm. tööstusvaldkondade kaasajastamisel ning statistika välja andmisel väga kõrge väärtusega. Seejuures on isikuandmed delikaatsed andmesubjektide mõistes. Laialdaselt levinud viisid delikaatsuse kaotamiseks ning väärtuse säilitamiseks on otseselt isikuga seotud tunnuste pseudonüümimine[1] või nende täielik eemaldamine andmestikust. Sel viisil on aga teatud eelteadmistega ründajal endiselt võimalik üsna lihtsasti tuvastada konkreetse isikuga seotud tundlikku informatsiooni, mis tekitab kahju nii andmesubjektist isikule kui ka andmeid väljastavale institutsioonile. Seetõttu tuleb privaatsuse kaitseks rakendada keerulisemaid meetmeid, mis tahes-tahtmata teevad laialdase andmeväljastuse protseduurid suuremates organisatsioonides keerukaks ja aeganõudvaks. Üheks selliseks organisatsiooniks on näiteks Tervise ja Heaolu Infosüsteemide Keskus (edaspidi TEHIK). TEHIK on info- ja kommunikatsioonitehnoloogia kompetentsikeskus tervise-, sotsiaal- ja töövaldkonnas. TEHIKu peamisteks tegevusteks on on avaliku sektori olulisemate IT-projektide elluviimine ja riigiastutustele vajalike arendus- ja haldustööde nõustamine [TEH22]. Andmete väljastamise protseduuride ressursimahukuse paremaks mõistmiseks vaadeldakse järgnevas alapeatükis andmete väljastamise protsessi TEHIKu näitel.

2.1.1 Andmete väljastamise protsess TEHIKus

Selles alapeatükis käsitletav informatsioon tugineb TEHIKu poolt e-kirja teel jagatud dokumendile [Mä22], mis kirjeldab andmete väljastamise protsessi 2022. aasta veebruari seisuga. TEHIKule esitatavad päringud pärivad andmeid põhiliselt Tervise infosüsteemist (TIS). Sarnaselt tehakse päringuid ka retseptikeskusest ja haigekassa andmekogust, mida haldab Haigekassa. Päringuid on laias laastus kahte tüüpi:

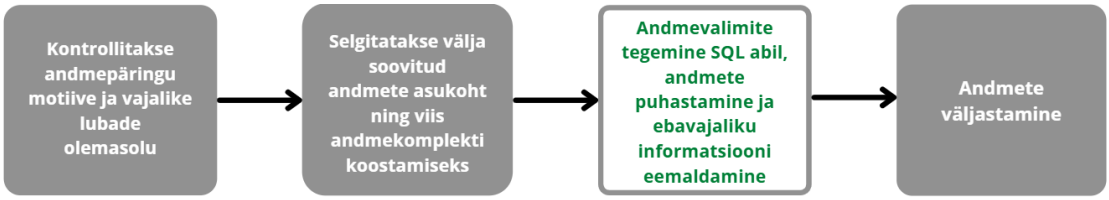
- Üksikpäringud
- Andmepäringud statistika, uuringute jms teostamiseks

[1] https://akit.cyber.ee/term/914-pseudonuumimine

Üksikpäringud: päritakse tavaliselt korraga 1-2 inimese andmed. Võib olla nii isik, kes pärib oma andmeid kui ka mõni organisatsioon või volitatud isik kes pärib andmeid kolmanda(te) isiku(te) kohta. Selliseid päringuid teevad näiteks: politsei/uurimisasutus, Terviseamet, KOV ja muud sarnased organisatsioonid. Kuna aga andmed üksikpäringute raames on alati isikustatud kujul, siis on üksikpäringud väljaspool selle bakalaureusetöö skoopi.

Andmepäringud statistika, uuringute jms teostamiseks: Küsitakse paljude inimeste andmeid korraga, üldjuhul terviseandmeid või tervishoiuteenuse osutamist käsitlevaid andmeid. Vajadusest tingituna on tegemist enamasti kohati suurte andmevalimite ja -mahtudega. Andmed võivad olla nii isikustatud kui ka isikustamata. Selliste päringute korral on võimalik TIS terviseandmeid ühildada ka muudest registritest (rahvastikuregister, hariduse infosüsteem jms) tulevate andmetega. Kliendid on tavaliselt: teadlaste grupid, uurimustööde tegijad (nt teaduslikud uurimustööd, bakalaureuse- ja magistritööd), ravimifirmad, tervishoidu korraldavad ametiasutused jt.

Eelnevalt mainitud põhjustel on üksikpäringud väljaspool töö skoopi, seega nendele kohaldatavaid erisusi andmete väljastamise protsessis ei käsitleta. Suuremate andmepäringute korral on protsess 2022. aasta veebruari seisuga alljärgnev:



Joonis 1. Andmete väljastamise protsess TEHIKus 2022. [Mä22]

Selle protseduuri kõige aeganõudvamaks osaks on reeglina esimene kast (vt joonis 1), kuna on vaja välja selgitada lähteülesande selgus, lahendatavus ja vajalike lubade/õiguslike aluste olemasolu (nt Sotsiaalministeeriumi luba). Andmete asukoha välja selgitamine (vt joonis 1 2. samm) ja füüsiline andmevalimite konstrueerimine (vt joonis 1 3. samm) on samuti aeganõudvad. Hetkeseisuga on mõistlikult automatiseeritav siiski ainult andmevalimite konstrueerimise samm, kuna pseudonüümimine ja anonüümimine kuuluvad selle sammu alla. Selle bakalaureusetöö tulemused on kasutatavad eelkõige andmevalimite konstrueerimise sammus. Järgnevalt sõnastatakse eelnevalt kirjeldatud protseduuri põhilised probleemid.

2.1.2 Protsessi probleemid

Igasugune andmeväljastus reaalsete isikute delikaatseid andmeid sisaldavate andmestike kontekstis on tänapäeval tugevalt reguleeritud. Euroopa raamistikus on õiguslikuks aluseks Isikuandmete Kaitse Üldmäärus (GDPR [PotEU16]). GDPR paneb väga rangelt paika füüsiliste isikute isikuandmete töötlemise piirid. Sellest tulenevalt on andmete väljastamisel kolmandatele isikutele ja organisatsioonidele vaja läbida aeganõudev kooskõlastuste ahel - taotleda ja saada eetikakomitee luba, konsultatsioon valimi osas, valimi kooskõlastamine andmebaaside vastutavate töötlejatega jms. Esimene suurem probleem ongi eelmainitud protsessi ressursimahukus.

Laialdase lahenduse raames on märkimisväärse tähtsusega asjaolu, et vastavalt GDPR põhjendusele nr 26[2] ei reguleerita sellega anonüümitud andmete käitlemist. GDPRis ei ole aga konkreetselt defineeritud, mida tähendavad anonüümitud andmed. Seda käsitletakse lähemalt alapeatükis 2.3. Teiseks põhiliseks probleemiks on see, et hetkel puudub tööriist, millega saab andmeid tõhusalt ja automaatselt anonüümida, säilitades mõistliku kasutatavuse ja pakkudes seejuures põhjalikku tulemuste analüüsi. Hetkeseisuga tehakse seda n-ö käsitsi.

Anonüümimine on üks paljudest privaatsust säilitavatest lahendustest andmete väljastamisel. Selle ja sarnaste lahenduste automatiseerimiseks on vaja nende alustalasid süvitsi mõista. Järgmises alapeatükis tutvustataksegi privaatsust säilitava andmete väljastamise kui laialdasema lahenduse olemust.

2.2 Privaatsust säilitav andmete väljastamine

Andmete väljastamiseks statistiliste andmebaaside mõistes on kolm põhilist formaati: mikroandmed, agregeeritud andmed (sagedustabelid) ja päritavad andmebaasid [DFSSC16]. Selle bakalaureusetöö raames käsitletakse ainult mikroandmeid, agregeeritud andmed ja päritavad andmebaasid on väljaspool skoopi. Mikroandmed on üldjuhul ühe konkreetse füüsilise või juriidilise isiku kohta käivad muutmata kujul andmed, mida väljastatakse hulkadena. Hulkadena tähendab siinkohal ennikute hulka, mis on reeglina viis andmete talletamiseks andmebaasides. Hulk ennikutest (mikroandmetest) on siis vaste andmebaasi päringule.

Privaatsuse tagamise perspektiivist saab mikroandmete hulka kirjeldada formaalsemalt nii:

$$D(ID, QID, TT, MTT), \tag{1}$$

kus D on mikroandmete hulk, mille elemendid jaotuvad alljärgnevalt:

[2] https://gdpr-info.eu/recitals/no-26/

- ID on otseselt identifitseerivate tunnuste[3] hulk, näiteks isikukood, isikut tõendava dokumendi number vms. Need tunnused üldjuhul eemaldatakse enne andmete väljastamist.

- QID on kvaasi-identifikaatorite[4] hulk. Kvaasi-identifikaatorid on tunnused, mis on iseseisvalt andmesubjekti identifitseerimisel kasutud, ent kombineerituna teiste kvaasi-identifikaatoritega võib olla piisav mõne andmesubjekti isiku paljastamiseks teiste sarnaste kirjete välistamise teel. Nende üldistamine on privaatsuse perspektiivist ülimalt oluline, kuna näiteks [Swe00] näitab, et kuni 87% USA elanikkonnast on avaandmetes osaliselt või täielikult identifitseeritavad kvaasi-identifikaatorite kombinatsiooni järgi ning [Gol06] suuresti kinnitab seda, kuid leiab, et identifitseeritavate andmesubjektide osakaal on vähenenud 63%-ni. Sellesse hulka kuuluvatel tunnustel rakendatakse anonüümimismeetodeid. Täpsem definitsioon antakse alapeatükis 2.3.1.

- TT on tundlike tunnuste[5] hulk. Siia kuuluvad reeglina tunnused, mis ei ole üldjuhul avalikud ning võivad avalikuks saamisel andmesubjekti kahjustada. Näiteks on sellisteks tunnusteks palganumber, seksuaalne orientatsioon, krooniline haigus vms. Sellesse hulka kuuluvaid tunnuseid võetakse anonüümimisel arvesse (vt 2.3.1)

- MTT on mittetundlike tunnuste hulk, sellesse hulka kuuluvad kõik tunnused, mis ei kuulu eelkirjeldatud kolme hulka. Mittetundlikke tunnuseid ignoreeritakse anonüümimisel ja tulemuste analüüsi teostamisel.

Mikroandmete hulk D jaotub neljakohaliseks ennikuks, mille elemendid on hulgad, mis moodustavad lahkneva komplekti, s.t neli elementi ID, QID, TT ja MTT on hulgad mille paarikaupa ühisosad on tühjad [BCMFY10]. See tähendab, et ükski tunnus mikroandmete hulgas ei saa korraga kuuluda rohkem kui ühte enniku elementides olevatest hulkadest. Seega näiteks ei saa üks tunnus olla korraga nii kvaasi-identifikaator kui ka tundlik tunnus.

Eelnevalt kirjeldatu põhjal on selge, et mikroandmete struktuur omab päris põhjalikku formaalset definitsiooni. Sama ei saa aga öelda andmete privaatsuse kohta üldisemalt. Esimene katse privaatsust formaliseerida tehti juba 1977. aastal Daleniuse poolt [Dal77]. Dalenius väitis, et ligipääs väljastatud andmetele ei tohi ühelegi võimalikule ründajale anda lisateadmisi mõne konkreetse indiviidi konfidentsiaalse informatsiooni osas. See sisuliselt tähendab, et vaatleva indiviidi uskumused ja teadmised ei tohi erineda enne ja pärast väljastatud andmete nägemist. See definitsioon aga osutus liiga karmiks, et

[3] https://akit.cyber.ee/term/245-identifikaator
[4] https://akit.cyber.ee/term/3535-kvaasi-identifikaator
[5] https://akit.cyber.ee/term/3551-sensitive-attribute

seda praktikas kasutada. Hetkeseisuga on lähimaks õigusaktiks formaalsetele nõuetele privaatsuses USA-s kasutusel olev HIPAA[6] (Health Insurance and Portability and Accountability Act), mis püüab täpsemalt defineerida tunnused, mis tuleks eemaldada või mida tuleks muuta, et tagada privaatsus. Sellest annavad hea ülevaate allikad [DFSSC16] ja [BCMFY10].

Privaatsust säilitav andmete väljastamine on vajalik, kuna isikuandmed on peaaegu alati delikaatsed, aga samas digiühiskonna arendamiseks, reklaamimiseks, teadustööks ja paljuks muuks väga väärtuslikud. Seejuures kergekäeliselt isikuandmeid käideldes võivad tekkida olukorrad, kus mõni konkreetne isik on kergesti seostatav tema mainet kahjustavate andmetega vms. Üldjuhul tagatakse andmete väljastamisel siiski see, et andmetest ei ole otseselt võimalik sääraseid seoseid välja lugeda. Küll aga olukorrad, kus läbi erinevate ründevektorite sääraste seoste paljastamine on väga lihtne, ei ole eriti haruldased. Järgnevalt tutvustataksegi klassikalisi ründemudeleid ja nende taga olevaid võimalikke motiive.

2.2.1 Ründemudelid

Enne kui on selge, mida ründemudelitega üldse ette näha proovitakse, on vaja mõista võimalikke desanonüümimise motiive. Need motiivid on reeglina samad ka umbisikustatud, krüpteeritud jm. viisidel isikuandmeid varjavate andmestike puhul. Sellest tulenevalt käsitletakse neid järgnevalt ühtsena andmekaitse mõiste all. Seni praktikas esinenud motiivid on näiteks [Gar15]:

- Andmekaitse kvaliteedi testimine
- Kuulsuse või professionaalse tunnustuse saamine andmekaitse nõrkuse näitamise eest
- Andmeid väljastanud organisatsiooni häbistamine/haavamine
- Otsese kasu saamine andmete seostamisest konkreetse füüsilise isikuga
- Konkreetse füüsilise isiku häbistamine läbi tema tundliku informatsiooni paljastamise

Ülal kirjeldatud motiivid jaotuvad abstraktsematesse ründemudelitesse. Neid on mitmeid, ent selle bakalaureusetöö raames on on olulised vaid kolm:

1. Prokuröri ründemudel, i.k *Prosecutor re-identification scenario* hindab ühe konkreetse indiviidi tuvastamise riski eeldusel, et ründaja on teadlik vastava indiviidi olemasolust vastavas andmestikus [ED08].

[6] https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/administrative/combined/hipaa-simplification-201303.pdf

2. Ajakirjaniku ründemudel, i. k *Journalist re-identification scenario* hindab ühe suvalise indiviidi tuvastamise riski. Eesmärgiks on reeglina näidata, et kedagi on endiselt võimalik tuvastada ning seeläbi kahjustada andmeid väljastanud organisatsiooni mainet [ED08].

3. Turundaja ründemudel, i. k *Marketer re-identification scenario* hindab protsentuaalselt tuvastatavate andmesubjektide kogust andmestikus. Selles stsenaariumis on reeglina eesmärgiks tuvastada võimalikult suur hulk identiteete ja selle pealt otsest kasu saada [DE10].

Kõikide eelmainitud ründemudelite praktikas rakendamiseks on vaja teostada mingit tüüpi andmete seostamine[7] i. k *linkage*, mis tähendab siinkohal loogiliste sidemete loomist eri objektide vahel. Seostamist hõlmavaid ründeid käsitleme edaspidi üldisemalt seostamisrünnetena. Järgnev on refereeritud privaatsust säilitava andmete väljastamise kirjanduse ülevaatest [BCMFY10]. Seostamisründeid on kolme tüüpi: kirjete seostamine, tunnuste seostamine ja tabelite seostamine. Kõigi kolmes seostamisründe eelduseks on, et ründaja teab ohvri kvaasi-identifikaatoreid. Kirjete ja tunnuste seostamisel eeldatakse veel, et ründaja teab, et ohvri kirje eksisteerib vaadeldavas andmestikus. Tabelite seostamise eesmärk ongi teada saada, kas ohvri kirje eksisteerib vaadeldavas andmestikus või mitte. Öeldakse, et andmestik on privaatsust säilitav, kui sellele rakendatud andmekaitse meetodid suudavad efektiivselt kaitsta eelmainitud seostamisrünnete vastu.

Siinkohal tuleb tähele panna, et tabelite seostamine on väljaspool selle bakalaureusetöö skoopi, kuna töö raames loodav tarkvara ei ole mõeldud töötamiseks mitme andmetabeliga. Lisaks on olemas ka taustteaberünded, kuid ka neid selle bakalaureusetöö raames ei käsitleta. Järgnevalt tutvustatakse lähemalt selle töö raames olulisi seostamisründeid lähemalt.

2.2.2 Seostamisründed

Kirjete seostamine i. k *record linkage*, tähendab, et ründaja kasutab ära taustateadmisi ohvri kvaasi-identifikaatorite kohta ja seostab seeläbi ohvrit vaid väikese hulga mikroandmetega või halvimal juhul unikaalsete mikroandmetega, tuvastades ohvri 100%-lise kindlusega. Tugevalt lihtsustatud näitena oletame, et joonis 2 (tabel A) on mingi haigla poolt avalikustatud andmestik ja et joonis 3 (tabel B) on Tartumaa elanike andmebaas. Eraldiseisvana ei ole tabeli B põhjal võimalik diagnoosiga kokku panna konkreetset isikut. Küll aga kui on olemas ligipääs nii tabelitele A kui B, siis nende ühisosa leidmisel üle QID $< Sugu, Vanus, Postiindeks >$ saadakse tulemuseks tabel C (vt joonis 4), kus on inimeste nimed koos haiglast saadud diagnoosidega. Näiteks on võimalik pärast ühisosa leidmist kindlalt väita, et Aivil on HIV.

[7] `https://akit.cyber.ee/term/2397-linkage-1`

			Tabel A
Sugu	**Vanus**	**Postiindeks**	**Diagnoos**
N	40	50602	Leukeemia
N	42	50709	HIV
N	42	50709	HIV
M	47	51006	Gripp
N	41	50101	Tuulerõuged
M	45	51522	HIV
M	49	51011	Angiin

Joonis 2. Haigla avaandmed.

			Tabel B
Nimi	**Sugu**	**Vanus**	**Postiindeks**
Meeli	N	40	50602
Aivi	N	42	50709
Pille	N	42	50709
Salme	N	41	50602
Ervin	M	48	51011
Peeter	M	47	51006
Leida	N	41	50101
Jaan	M	45	51522
Margus	M	49	51011
Jessica	N	43	50702

Joonis 3. Tartumaa elanike andmebaas.

Inner join				Tabel C
Nimi	**Sugu**	**Vanus**	**Postiindeks**	**Diagnoos**
Meeli	N	40	50602	Leukeemia
Aivi	N	42	50709	HIV
Pille	N	42	50709	HIV
Peeter	M	47	51006	Gripp
Leida	N	41	50101	Tuulerõuged
Jaan	M	45	51522	HIV
Margus	M	49	51011	Angiin

Joonis 4. Tabelite A ja B ühisosa.

Tunnuste seostamine i. k *attribute linkage*, juhtub enamasti siis, kui puhtalt kirjete seostamise abil ei õnnestu ühtegi andmesubjekti identifitseerida. Sellisel juhul võib ründaja suuta väikesest mikroandmete hulgast tundlike tunnuste põhjal tuletada, et sinna hulka kuuluvad andmesubjektid on suure tõenäosusega seotud mingi konkreetse tundliku tunnusega. Seda juhtu on hästi kirjeldanud [MKGV07]. Siinkohal saab samuti tuua lihtsustatud kujul näite. Kuna tabelis A (vt joonis 2) on kaks kirjet mille QID $< Sugu, Vanus, Postiindeks >$ on samad ja diagnoos on ka sama, siis vaadates tabelit B (vt joonis 3) on näha, et Tartumaal on ainult kaks isikut kellel QID $< Sugu, Vanus, Postiindeks >$ on <N, 42, 50709> ja haigla tabeli järgi on neil mõlemal HIV, siis saab sellest tuletada, et näiteks Pillel on HIV.

Tunnuste seostamise näidet aitavad täpsemalt mõista k-anonüümsuse ja l-hajutuse põhimõtted. Selleks tutvustatakse järgmises alapeatükis täpsemalt anonüümimise mõistet ja põhilisi praktikas rakendatavaid privaatsusmudeleid.

2.3 Anonüümimine

Andmekaitse ja infoturbe leksikoni [AS11] järgi on anonüümimine[8] pööramatu protseduur, mille tulemuseks on isikutuvastusteabe kõrvaldamine või muutmine nii, et selle

[8] https://akit.cyber.ee/term/630-anonuumimine

subjekti ei saa tuvastada. Anonüümimist ei tohi segamini ajada umbisikustamisega[9], tuntud ka kui pseudonüümimine, mis on pööratav protseduur andmesubjektide identiteedi varjamiseks statistilistes andmestiks ning hõlmab otseste ja kaudsete isikuandmete peitmist või maskeerimist. Erinevalt anonüümimisest säilitab umbisikustamine isikuandmed ja võib olla pööratav. Oma olemuselt on anonüümitud andmed ka umbisikustatud andmed, aga mitte vastupidi. See tähendab, et anonüümimine on privaatsuse mõistes turvalisem kui umbisikustamine. Selles bakalaureusetöös vaadeldakse erinevaid mõõdikuid tuvastamaks anonüümimise kvaliteeti, s.t eeldatakse, et väljundandmed on anonüümitud, mitte umbisikustatud. Selle eelduse kohaselt on igasugune anonüümimise protseduuri pöördkonstrueerimine[10] tulemuste analüüsi ajal väljaspool bakalaureusetöö skoopi.

Alljärgnev definitsioon on refereeritud elektroonse kogumiku JMIR Publications artiklist [ZWB+21]. Anonüümimise olemuse mõistmiseks on see defineeritud ka matemaatiliselt alljärgnevalt: Olgu meil anonüümimisfunktsioon A ja algsel kujul andmestik X. Anonüümitud andmestik on

$$X' = A(X) \tag{2}$$

kui ei leidu sellist funktsiooni R, mis taastab anonüümitud andmetest X' algsed andmed X, s.t ei leidu funktsiooni R nii, et

$$R(X') = R(A(X)) = X \tag{3}$$

Kui selline funktsioon leidub, siis ei ole tegemist anonüümitud andmetega, vaid hoopis umbisikustatud ehk pseudonüümitud andmetega. Anonüümimise peamine eesmärk on ründemudelitest tulenevate riskide minimeerimine. Selleks on loodud suur hulk erinevaid privaatsusmudeleid, mille rahuldamise korral saab andmestiku avalikustamisest tulenevaid riske formaalsemalt kirjeldada. Sellest lähemalt järgmises alapeatükis.

2.3.1 Põhilised privaatsusmudelid

Privaatsusmudelid üldisemas raamistikus on justkui reeglid, mis täpsustavad tingimusi, millele anonüümitud andmestik peab vastama, et desanonüümimise[11] risk oleks kontrollitud. Tavaliselt sõltuvad privaatsusmudelid ühest või mitmest parameetrist, mis täpsustavad maksimaalse lubatud desanonüümimise riski [SCDF15]. Enamik olemasolevaid laialdaselt kasutatavaid privaatsusmudelid on välja arendatud töötama ühe staatilise andmestikuga. Kuna selle bakalaureusetöö raames käsitletakse anonüümimist ainult ühe eraldiseisva tabeli mõistes, mis sobib eelneva tingimusega, siis keskendutakse tulemuste analüüsis eelkõige üldtuntud privaatsusmudelitele. Erinevatest privaatsusmudelitest, nende loomismotivatsioonist ja kasutusjuhtudest annab hea ülevaate [BCMFY10]. Järgnevalt

[9] https://akit.cyber.ee/term/2848-umbisikustamine
[10] https://akit.cyber.ee/term/913-reverse-engineering
[11] https://akit.cyber.ee/term/4177-desanonuumimine-nuumimine

kirjeldatakse täpsemalt selle bakalaureusetöö raames olulisi privaatsusmudeleid. Privaatsusmudelite paremaks mõistmiseks on vaja süvitsi mõista kvaasi-identifikaatorite (edaspidi QID) definitsiooni ja olemust. QID on formaalselt defineeritud järgnevalt [SS98]:

Definitsioon (QID) Olgu $T(A_1, ..., A_n)$ tabel. T mõistes on QID-ks hulk tunnuseid $\{A_i, ..., A_j\} \subseteq \{A_1, ..., A_n\}$ mille väljastamine peab olema kontrollitud.

Kontrollitud tähendab siin kontekstis, et QID kuuluvaid tunnuseid tuleb muuta nii, et iga võimalik QID kombinatsioon vastaks vähemalt k-le kirjele tabelis T, kus k on mingi naturaalarv. See aga eeldab, et kõik tunnused, mis on kombinatsioonidena avalikud väljaspool tabelit T on defineeritud T-ga seotud QID-s. See aga omakorda eeldab, et kõik sellised tunnused on andmete hoidja poolt korrektselt tuvastatud. See eeldus ei pruugi alati täies mahus kehtida, ning lohakuse tõttu QID tuvastamisel ei ole vastavus lubatud privaatsusmudelitega enam garanteeritud [OMSG$^+$21]. QID tuvastamine on endiselt üsna lahtine probleem, siiski on selle jaoks pakutud välja mitmeid üldjuhul hästi toimivaid algoritme. Üks uudsemaid on näiteks [OMSG$^+$21]. Nõue, et QID-d vastaksid vähemalt k-le kirjele tabelis T ongi k-anonüümsuse privaatsusmudeli aluseks. K-anonüümsus defineeriti esimest korda formaalselt L. Sweeney ja P. Samarati töös [SS98] järgnevalt:

Definitsioon (K-anonüümsus) Olgu $T(A_1, ..., A_n)$ tabel ja Q_T T-ga seotud QID. T rahuldab k-anonüümsuse privaatsusmudelit siis ja ainult siis kui iga $QID \in Q_T$ iga väärtuste jada $T[QID]$-s ilmub $T[QID]$-s vähemalt k korda.

Kui eelnev definitsioon on rahuldatud, siis iga ennik $T[QID]$-s esineb vähemalt k korda. Sellest tulenevalt on igasuguse kirjete linkimise ründe tulemuslikkuse tõenäosus maksimaalselt $1/k$. Sellest on omakorda tuletatav, et mida suurem on valitud k väärtus, seda väiksema kindlusega saab ründaja väita, et ta on tuvastanud ohvri kirje tabelis T. Joonisel 5 olev tabel A on joonisel 2 oleva tabeli 3-anonüümne vorm ning joonisel 6 on joonisel 3 oleva tabeli 4-anonüümne vorm. QID-d on samad: <Sugu, Vanus, Postiindeks>. Kirjete linkimise ründe neutraliseerimist läbi k-anonüümsuse rakendamise illustreerib hästi joonisel 7 olev tabel, mis on tabelite A ja B anonüümsete vormide ühisosa. Joonisel 7 on selgelt näha, et tulemus säilitab 3-anonüümsuse. Siin on märkimisväärse tähtsusega ka asjaolu, et anonüümitud tabelitest eemaldatakse kõik otseselt identifitseerivad atribuudid, milleks on siinkohal isikute nimed. Eemaldamise sümboliseerimiseks on eemaldatavad otsesed identifikaatorid värvitud oranžiks ning alles jäetud vaid näite selguse tagamiseks. Tulemuseks on veidi üldisemat informatsiooni andvad tabelid, võrreldes originaalkujul tabelitega, ent see-eest on kirjete seostamise rünne täielikult neutraliseeritud. Nagu eelnevalt mainitud, on k-anonüümsuse rakendamisel kriitilise tähtsusega QID tuvastamine. Näites eeldatakse, et see on õnnestunud.

Sellisel viisil anonüümitud tabelid A ja B ning nende ühisosa C ei võimalda siiski meil mõista allesjäänud ohtu, mis seisneb tunnuste seostamises. Ohu kirjeldamiseks sobib

K = 3; L = 3			
3-Anonüümne			
3-Hajutus	Avalik		Anon A
Sugu	**Vanus**	**Postiindeks**	**Diagnoos**
N	[40-45)	50XXX	Leukeemia
N	[40-45)	50XXX	HIV
N	[40-45)	50XXX	HIV
M	[45-50)	51XXX	Gripp
N	[40-45)	50XXX	Tuulerõuged
M	[45-50)	51XXX	HIV
M	[45-50)	51XXX	Angiin

Joonis 5. 3-anonüümne tabel A.

	k = 4		
	4-Anonüümne		
Eemaldatud	Avalik		Anon B
Nimi	**Sugu**	**Vanus**	**Postiindeks**
Meeli	N	[40-45)	50XXX
Aivi	N	[40-45)	50XXX
Pille	N	[40-45)	50XXX
Salme	N	[40-45)	50XXX
Ervin	M	[45-50)	51XXX
Peeter	M	[45-50)	51XXX
Leida	N	[40-45)	50XXX
Jaan	M	[45-50)	51XXX
Margus	M	[45-50)	51XXX
Jessica	N	[40-45)	50XXX

Joonis 6. 4-anonüümne tabel B.

Inner join	K = 3; L = 3			
Eemaldatud	Avalik			Anon C
Nimi	**Sugu**	**Vanus**	**Postiindeks**	**Diagnoos**
Meeli	N	[40-45)	50XXX	Leukeemia
Aivi	N	[40-45)	50XXX	HIV
Pille	N	[40-45)	50XXX	HIV
Peeter	M	[45-50)	51XXX	Gripp
Leida	N	[40-45)	50XXX	Tuulerõuged
Jaan	M	[45-50)	51XXX	HIV
Margus	M	[45-50)	51XXX	Angiin

Joonis 7. Anonüümitud tabelite A ja B ühisosa.

järgnev stsenaarium. Oletame, et Aivi naaber Toomas teab, et Aivi käis nädal aega tagasi haiglas. Sellest tulenevalt on Toomas kindel, et Aivi kohta on joonisel 8 kuvatud tabelis kirje. Samuti teab Toomas, et Aivi on naine ja ta on 42-aastane. Nende teadmistega 2-anonüümset haigla avaandmete tabelit vaadates saab Toomas kindlasti öelda, et Aivil on HIV, kuna QID <N, [40-45), 50XXX> kohta on tabelis A ainult 2 kirjet ja neil mõlemal on diagnoositud HIV.

Inner join	K = 2; L = 1			
Eemaldatud	Avalik			Anon C
Nimi	**Sugu**	**Vanus**	**Postiindeks**	**Diagnoos**
Aivi	N	[40-45)	50XXX	HIV
Pille	N	[40-45)	50XXX	HIV
Peeter	M	[45-50)	51XXX	Gripp
Jaan	M	[45-50)	51XXX	HIV
Margus	M	[45-50)	51XXX	Angiin

Joonis 8. Tunnuste seostamise võimalus 2-anonüümses haigla avaandmete tabelis.

Sellised stsenaariumid ei ole üldse haruldased ja tähendavad seega tugevat privaatsusriski andmestikus esindatud isikutele. Tunnuste seostamise rünnete vastu töötati sellistest stsenaariumitest tulenevalt välja l-hajutuse privaatsusmudel. Esmase formaalse definitsiooni

l-hajutuse privaatsusmudelile andis [MKGV07]. Enne l-hajutuse definitsiooni selgitamist on vaja defineerida $Q*$-plokk.

Definitsioon ($Q*$-plokk) $Q*$-plokk on ennikute hulk tabelis T mille **mitte** tundlike tunnuste hulkadesse kuuluvate tunnuste väärtused on üldistatavad $Q*$-ks. Siinkohal on mitte tundlike tunnuste all mõeldud kõiki tunnuseid a, kus $a \notin TT$.

Definitsioon (L-hajutus) Vaatleme tabelisse T kuuluvat $Q*$-plokki. Öeldakse, et $Q*$-plokk on l-hajutatud kui see sisaldab vähemalt l "hästi esindatud" väärtust tundlike tunnuste hulgas TT. Tabel T on l-hajutatud kui iga $Q*$-plokk $\in T$ on l-hajutatud.

Joonisel 9 on üks 5-anonüümne $Q*$-plokk, mis on 4-hajutatud. Plokk on $Q*$-plokk, kuna ploki mitte tundlike tunnuste hulkadesse kuuluvad tunnused on kõik ühtlaselt üldistatud QID-ks $< Sugu : N, Vanus : [40-50), Postiindeks : 50XXX >$ ning vastav plokk on 4-hajutatud, kuna plokis oleva TT hulka kuuluva tunnuse $Diagnoos$ väärtuste hulgas on 4 unikaalset väärtust. K-anonüümne tabel koosneb sarnastest plokkidest, aga neid nimetatakse üldiselt ekvivalentsiklassideks. Seda sellepärast, et k-anonüümsuse kehtimiseks peavad kõik konkreetses plokis olevad QID hulka kuuluvad väärtused olema ekvivalentsed. Siit on tuletatav, et kuna l-hajutuse definitsioon on üles ehitatud $Q*$-plokkidele (edaspidi ekvivalentsiklass) ja ka k-anonüümsuse definitsioon on üles ehitatud QID mõistes ekvivalentsiklassidele ning l-hajutus nõuab lisaks mitmekesisust TT hulgas olevate tunnuste väärtuste seas, siis l-hajutus on k-anonüümsusest rangem nõue, ning seetõttu kehtib alati $k \leqslant l$. See tähendab, et stsenaariumit, kus anonüümitud andmestikus on rahuldatud näiteks $k \leqslant 5$ ja $l = 6$, ei saa tekkida. See on põhiliselt oluline mõistmaks, miks selliste olukordade testimist peatükis 4.3 ei käsitleta.

K = 5; L = 4				
Eemaldatud	Avalik			Anon D
Nimi	**Sugu**	**Vanus**	**Postiindeks**	**Diagnoos**
Aivi	N	[40-50)	50XXX	Köha
Pille	N	[40-50)	50XXX	Gripp
Peeter	N	[40-50)	50XXX	Gripp
Jaan	N	[40-50)	50XXX	Seljavalu
Margus	N	[40-50)	50XXX	Angiin

Joonis 9. Näide ühest 5-anonüümsest ekvivalentsiklassist, mis sümboliseerib terviklikku tabelit ja on 4-hajutatud.

K-anonüümsuse ja l-hajutuse kombinatsioonist ei piisa siiski olukordades, kus k-anonüümsus ei ole tagatud indiviidi tasemel, vaid kirje tasemel. Kirje tasemel tagatud k-anonüümsus vastab nõuetele seni, kuni anonüümitud tabelis $T*$ ei ole ühtegi ekvivalentsiklassi üle QID $< Amet, Postiindeks >$ kus üks konkreetne indiviid on esindatud rohkem kui üks kord. Probleemi visualiseerimiseks vaadeldakse joonisel 10 olukorda, kus 5-anonüümne

ja 5-hajutatud ekvivalentsiklass sisaldab tegelikult oodatavast viiest unikaalsest indiviidist ainult kahte unikaalset indiviidi. Seda kirjeldab tunnus ID, mis sümboliseerib indiviidi unikaalset identifikaatorit andmebaasis.

ID	Amet	Postiindeks	Haigus
377891	Kunstnik	5070*	Palavik
368821	Kunstnik	5070*	Köha
377891	Kunstnik	5070*	Kasvaja
377891	Kunstnik	5070*	HIV
377891	Kunstnik	5070*	Düsleksia

Joonis 10. Olukord kus k-anonüümsuse privaatsusmudel ei taga lubatud nõudeid.

Selle probleemi lahendamiseks pakkusid Wang ja Fung [WF06] välja (X, Y)-anonüümsuse privaatsusmudeli kus X ja Y on tunnuste lahknevad komplektid. Definitsioon on alljärgnev:

Definitsioon ((X, Y)-anonüümsus) Olgu x väärtus X-s. x-i anonüümsus Y suhtes, sümboliseeritud kui $a_Y(x)$, on x-ga koos esinevate unikaalsete väärtuste arv Y-s. Kui Y on võti tabelis T, siis $a_Y(x)$ on võrdne x-i sisaldavate kirjete arvuga. Olgu $A_Y(X) = min\{a_Y(x) \mid x \in X\}$. Tabel T rahuldab (X, Y)-anonüümsust suvalise täisarvu k korral kui $A_Y(X) \geq k$.

Näiteks olgu $X = \{Amet,\ Postiindeks\}$ ja $Y = \{ID\}$. Sellisel juhul on (X, Y)-anonüümsus tagatud joonisel 10 oleval tabelil tagatud $k = 2$ väärtuse jaoks. Seda sellepärast, et iga QID $< Amet:\ Artist,\ Postiindeks:\ 5070* >$ vastavuses oleva ID tunnuses olevate kirjete hulgas on ainult kaks unikaalset ID-d. Selleks, et kehtiks ka (X, Y)-anonüümsus väärtusele 5, peaks igale QID $< Amet:\ Artist,\ Postiindeks:\ 5070* >$ kirjele vastama unikaalne ID. Sellisel juhul oleks k-anonüümsus tagatud nii kirje kui ka indiviidi tasemel. Sisuliselt saab (X, Y)-anonüümsust abstraheerida indiviidi tasemel k-anonüümsuseks. Olukorda, kus k-anonüümsus on tagatud nii indiviidi kui kirje tasemel, illustreerib joonis 11.

ID	Amet	Postiindeks	Haigus
338274	Kunstnik	5070*	Palavik
368821	Kunstnik	5070*	Köha
532423	Kunstnik	5070*	Kasvaja
212341	Kunstnik	5070*	HIV
324235	Kunstnik	5070*	Düsleksia

Joonis 11. Olukord kus k-anonüümsus on tagatud nii indiviidi kui kirje tasemel.

Selle bakalaureusetöö ja laialdasema Health Sense projekti (lähemalt alapeatükis 2.6) raames rakendatakse anonüümimisprotseduuridel hetkeseisuga k-anonüümsuse ja l-hajutuse nõudeid. Siiski otsustas projekti töögrupp, et analüüsikomponendis on mõistlik kontrollida ka (X, Y)-anonüümsuse kehtivust ning tuvastada rikkuvad ekvivalentsiklassid. Seda käsitletakse lähemalt sektsioonis 3.2. Arvesse tuleb siiski võtta asjaolusid, et ülalkirjeldatud privaatsusmudelite formaalsetest definitsioonidest ja põhilistest desanonüümimise ründevektorite tõhususe vähendamisest hoolimata, on anonüümitud andmete väljastamine alati riskantne, kui ei väljastata täielikult pimendatud andmeid, aga sellisel juhul kaob andmete väljastamise mõte üldse. Pimendatud andmed tähendavad siinkohal andmeid, mille sisu ei ole enam informatiivne. Teisisõnu andmed, mille QID-sse kuuluvad väärtused on üldistatud neile vastavate taksonoomia puude juurtippudeni (vt alapeatükk 2.4). Lisaks on anonüümimisel veel mitmeid pisemaid probleeme, aga nendest räägitakse lähemalt juba järgmises alapeatükis.

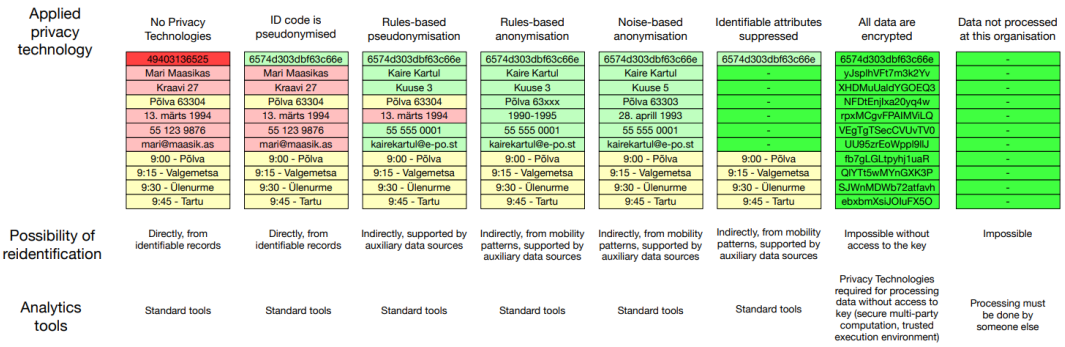
2.3.2 Anonüümimise probleemid

Anonüümitud andmetele kehtivaid konkreetseid nõudmisi ei ole siiski õnnestunud täielikult formaalselt piiritleda, kuna hoolimata vastavusest tingimustele 2 ja 3 selle peatüki sissejuhatuses antud definitsioonis on välja töötatud väga palju erinevaid anonüümimisfunktsioone. Kuna neil on sisulised erinevused, siis on raske panna paika üldiseid nõudmisi, millele nad kõik vastama peaksid. Anonüümimisfunktsioonide arvukus tuleneb sellest, et neid on välja töötatud alates 1998. aastast, mil [SS98] tutvustas esmakordselt k-anonüümsust (lähemalt alapeatükis 2.3.1). Loeme seda alguspunktiks, kuna kõik selle töö raames tutvustatavad privaatsusmudelid on üles ehitatud k-anonüümsuse täiustustena. Sellest tulenevalt on k-anonüümsus ka *de facto* standard maailmas isikuandmete kaitsmisel. 24-aastasest uurimistööst tulenevalt on variantide hulk läinud väga laialdaseks ning enamik uudsemaid privaatsusmudeleid on üles ehitatud mõne vanema privaatsusmudeli pisemate puuduste likvideerimiseks, mis teeb uudsed privaatsusmudelid väga olukorraspetsiifiliseks ning õigusruumides ei ole võimalik üheselt defineerida konkreetseid piire kõikvõimalikke stsenaariume ja privaatsusmudeleid arvesse võttes.

Täiusliku anonüümimise saavutamise võimatuse tõttu on akadeemiline debatt sellel teemal endiselt käimas ja täielik üksmeel puudub. Üldpildis on kritiseerivateks osapoolteks formalistid ja pooldavateks osapoolteks pragmatistid. Kõige põhjalikumalt arvesse võetud kriitilise ülevaate annab Paul Ohmi artikkel [Ohm09], kus ta analüüsib põhjalikult anonüümimise puudusi, võttes aluseks mitmed eelnevad teadustööd, kus eeldatavalt anonüümsetest andmetest tuvastatakse andmesubjekte vähese vaevaga. Lõpuks jõuab ta järeldusele, et anonüümimise garantiisid on tugevalt üle hinnatud ja anonüümimist ei tohiks suurandmete ajastul tõlgendada kui lahendust kõigile privaatsusprobleemidele. Alternatiiviks pakub ta riskipõhise lähenemise, kus iga andmeväljastuse puhul tuleks riske hinnata olukorraspetsiifiliselt ning olukorrad, kus potentsiaalne risk ületab võimalikku

kasu, peaksid olema tugevalt reguleeritud.

Vastuseks Ohmile arutlevad Schwartz ja Solove [SS13] isikutuvastusteabe[12] mõiste ümber sõnastamist paindlikuma lähenemise suunas. Nad leiavad, et desanonüümimise riski tuleks hinnata skaalal, kus ühes otsas on täielikult isikustatud andmed ja teises otsas on andmed, mida on täielikult võimatu mõne andmesubjektiga seostada. Jooniselt 12 on näha, et selline skaala ka eksisteerib. Laialdasemat kasutust leiabki just nn hall ala, kus säilitatakse võimalikult suur hulk informatsiooni, aga tagatakse suure tõenäosusega ka anonüümsus. Selle bakalaureusetöö raames käsitletakse eelkõige reeglipõhist anonüümimist, i. k *rules-based anonymization*, mis jääb täpselt skaala keskele 4. kasti.



Joonis 12. Skaala visualiseerimaks desanonüümimise riski astmeid, mille keskel on risk mõõdukas - andmesubjekt on tuvastatav aga mitte tuvastatud ning tuvastamiseks vajalikud sammud ei ole triviaalsed. Allikas: [Bog22]

Lisaks kritiseerib Ohmi soovitust anonüümimisest loobuda ka Bambauer artiklis [Bam11], öeldes, et Ohm ülehindab anonüümimisest tulenevaid riske spetsiifiliste näidete põhjal ja alahindab avaandmete väärtust. Kuna mõlema osapoole väidetel on alust ja anonüümimise laialdasem kasutuselevõtt on akadeemilise debati viigiseisu taha ootama jäänud, pakuvad Rubenstein ja Hartzog [SH16] välja protsessipõhise lähenemise desanonüümimise ja tundlike tunnuste paljastamise riski minimeerimiseks, leppides seejuures asjaoluga, et täiuslik anonüümimine on võimatu. See aga eeldab spetsiifilisi teadmisi ja kontekstitundlikkust andmeid väljastavatelt institutsioonidelt, mis on laialdases kasutuses tugev eeldus, ent siiski võimaldab laialdasemat avaandmete väljastamist kui Ohmi pakutud jäik kasutegurite ja riski võrdlemise protseduur.

[12] https://akit.cyber.ee/term/728

Täpsema ülevaate akadeemilise debati sisust annab [Col19], kus on koostatud põhjalik kokkuvõte kaasaegsetest probleemidest anonüümimises. Tähtsaim väide eelmainitud ülevaates on, et praktiliselt kasulik andmestik ei saa mitte kuidagi olla täielikult riskivaba, igasugune isikuandmete käitlemine toob kaasa mingi riskiteguri.

Selle alapeatüki eesmärgiks oli tutvustada kaasaegseid probleeme ja vaateid anonüümimise raamistikus. Akadeemilise debati hetkeseisuga kõige populaarsem seisukoht on, et täiuslik anonüümimine on võimatu ning mõistlik praktiline eesmärk oleks püsida joonisel 12 tutvustatud skaala keskel, mis eeldab kvaasi-identifikaatorite üldistamist või nendesse müra lisamist, võttes seejuures arvesse andmeväljastuse konteksti. Selle bakalaureusetöö raames vaadeldakse ainult kvaasi-identifikaatorite põhist üldistamist, teisisõnu anonüümimine läbi kvaasi-identifikaatorite hierarhilise üldistamise. Kõik teised skaalal kirjeldatud meetodid jäävad skoobist välja.

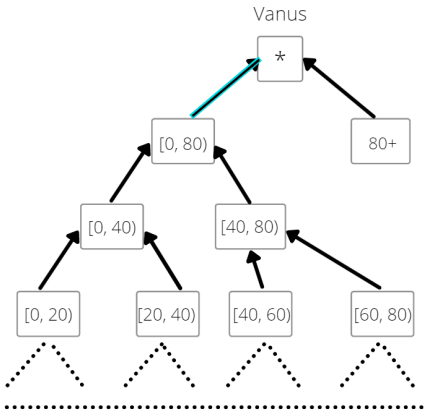
2.4 Informatsioonikadu

Siiamaani on keskendutud põhiliselt anonüümimise olemusele, põhimõtetele ja formaliseerimise viisidele. Ilmselt aga on üsna selge see, et privaatsusmudelite nõuete täitmiseks on vaja andmete informatiivsust olulisel määral vähendada.

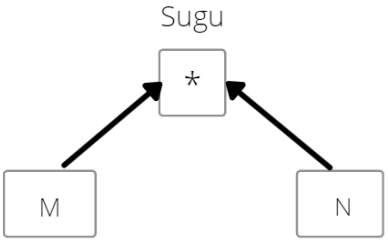
Vaadates näiteks jooniseid 2, 3 ja 4 näeme selgelt, et tagatud ei ole ükski eelnevalt tutvustatud privaatsusmudelite nõuetest. Ilmselge on see, et anonüümimise tagajärjel tuleb loobuda otsestest identifikaatoritest, nagu isikute nimed ja neile unikaalsed koodid nagu näiteks Eesti isikukood vms. See tähendab, et kui näiteks kliinilistes uuringutes tuvastatakse anonüümitud andmestikust andmeanalüüsi põhjal mõni huvitav indiviid, siis isikustatud andmetega on võimalik selle isiku haigusloo ja muude täpsemate andmete uurimine. Anonüümitud andmete puhul ei tohiks teoreetiliselt olla võimalik ühte kirjet andmetabelis seostada konkreetse indiviidiga, nii et anonüümitud andmetega saab praktikas teostada ainult üldisemaid uuringuid. Siiski, uuringute üldistamine vähesel määral, säilitades andmesubjektide privaatsust on tunduvalt väärtuslikum kui uuringute üldse mitte tegemine. Küll aga identifitseerivate tunnuste kaotamine ei ole kogu informatsioonikadu. Informatsioonikadu võib edaspidi käsitleda kui abstraktset mõistet mis kirjeldab mingit informatsiooni sisaldavate väärtuste üldistamist ja osalist või täielikku maskeerimist.

Siinkohal tuleks meenutada $Q*$-ploki või siis ekvivalentsiklassi definitsiooni, mille kohaselt mitte tundlike tunnuste hulk peab olema üldistatav. Üldistamise tulemuseks on alati informatsioonikadu, kuna üldistamine taksonoomia puude või teisisõnu hierarhiate abil muudab täpseid väärtusi vahemikeks ja üldistab näiteks linnaosad maakondadeks jms. Võrreldes siinkohal joonistes 5, 6 ja 7 olevaid tabeleid nende isikustatud versioonidega joonistel 2, 3 ja 4 ja vaadeldes taksonoomia puid 14, 13 ja 15, on näha kui

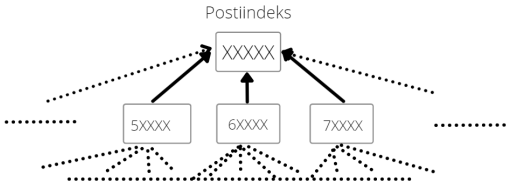
palju informatsiooni on kaotatud liikudes anonüümsuse tagamiseks taksonoomia puudes ülespoole.



Joonis 13. Vanuse taksonoomia puu.



Joonis 14. Soo taksonoomia puu.



Joonis 15. Postiindeksi taksonoomia puu.

Täpsem võrdlusmoment tekib kui vaadata kõrvuti näiteks joonistel 2 ja 5 olevaid tabeleid. On näha, et sugu ei ole kuidagi üldistatud, vaid 3-anonüümsuse ja 3-hajutuseni on jõutud teisi tunnuseid üldistades. Näiteks vanused on üldistatud ühe sammu võrra 5 aastastesse vahemikesse ja postiindeksid on üldistatud esimesele kahele numbrile, viitamaks täpsema piirkonna asemel laiemale piirkonnale. Postiindeksite mõistes on kasutatud ühe üldistussammuna ühe numbri peitmist, s.t igale postiindeksile on rakendatud kolm üldistussammu, s.t postiindeksi taksonoomia puus on liigutud ülespoole kolm taset. Sellise lähenemise kasutamisel on intiutiivne, et mida tugevamad privaatsusnõuded esitada, seda rohkem üldistussamme on vaja teha, et nõudmisi rahuldada. Kõrgeim privaatsus on tagatud taksonoomia puude juurtippudesse jõudmisel, mis on samaväärne tunnuste andmestikust eemaldamisega (vt joonis 12 paremalt kolmandat kasti). QID tunnuste eemaldamise tulemusel jõutakse peaaegu samale isikukaitsetasemele andmete krüpteerimisega. Krüpteeritud andmed ei ole ilma krüpteerimisvõtmeta enam informatiivsed, sama on täiesti olematute andmetega, mis tagavad kõrgeima privaatsuse. Kui aga nüüd läheneda andmeväljastuse kliendi perspektiivist, kes tahab nii kõrget informatiivsust kui

võimalik, siis on ilmselge, et krüpteeritud ja tühjad andmed on väärtusetud. Täielikult eemaldatud QID väärtustega andmestik ei anna samuti informatsiooni näiteks selle kohta, kus või millises vanuseklassis inimestega midagi juhtub või mis ametikohtadel inimesed külastavad kõige tihedamini silmaarsti. Täieliku pimendamise teel jääb alles vaid tühine informatsioon haiguste nimedega ja visiitide kogust kirjeldava arvuga. Selle info ainus väärtus on kinnitus, et keegi on selle haigusega nakatunud või keegi silmaarsti külastanud x korda. Keerulisematele uurimis- ja analüüsiküsimustele vastamine ei ole enam võimalik.

Sellest tulenevalt on eesmärk andmeid anonüümida täpselt nii palju, et säilib maksimaalne kasutatavus. Kõige laialdasemat kasutust leiab praktikas 5-anonüümsus. Üle selle väärtuste kasutamine andmeid talletavate organisastioonide poolt on üsnagi ebatavaline ning kõrgemat kui 15-anonüümsust ei kasutata peaaegu üldse [EED08]. Oluline tähelepanek siinkohal on see, et anonüümsusele saab väiksema informatsioonikaoga esitada kõrgemaid nõudmisi sõltuvalt ka andmete kogusest. Kui nõuda 10-realiselt tabelilt 5-anonüümsust, kus kõik read on unikaalsed, siis tuleb tõenäoliselt eesmärgi saavutamiseks kaotada suurem osa kasutatavat informatsiooni. Samas kui nõuda 50000-realiselt sarnaste väärtustega tabelilt 10-anonüümsust, siis protsentuaalne infokao suurus on suure tõenäosusega kordades väiksem. Sellised juhtumid on aga väga olukorraspetsiifilised ning neid on raske üldistada, sellest tulenevalt on keeruline, kui mitte võimatu, panna paika fikseeritud standardit. Sellest tulenevalt on andmete anonüümimisel tihti vaja erinevate parameetritega katsetada, võttes seejuures arvesse andmete väljastamise vajaduse konteksti, andmestiku suurust ja selles esinevat mitmekesisust. Alla 5-anonüümsete andmete avalikustamine on aga üldiselt ebatavaline [EED08].

See kõik tähendab, et informatsioonikao ja anonüümsuse vahelist tasakaalu oleks vaja kuidagi mõttekalt mõõta nii, et mõõtmistulemuste põhjal oleks võimalik otsustada, kas saavutatud anonüümsus on piisav ning kas andmed on piisavalt informatiivsed, et nende väljastamine oleks endiselt mõttekas. Selleks on välja pakutud väga palju erinevaid informatsioonikao mõõdikuid. Paljud neist põhinevad erinevate karistuspunktide määramisel igasuguste üldistussammude suhtes. Informatsioonikao mõõdikuid tutvustatakse lühidalt näiteks [BCMFY10]-s.

Selle alapeatüki eesmärk oli lühidalt tutvustada anonüümimise ja informatsioonikao suhet ning täpsemalt kirjeldada anonüümimise kui andmekaitse lahenduse teoreetilisi piire. Olles tutvunud anonüümimise ja levinud privaatsusmudelite definitsioonide ja hüpoteetiliste näidisolukordadega ning informatsioonikao ja tagatud privaatsuse vaheliste kompromissidega i. k *trade-off*, saab vastu võtta optimaalsemaid otsuseid, valides mõõdikuid, mida riskianalüüsi ja informatsiooni kao hindamise protseduurides rakendada. Esmalt oleks aga vaja ülevaadet mõõdikute tüüpidest abstraktsemalt. Järgnevas alapeatükis tutvustataksegi mõõdikute tüüpe ja nende olemusi.

2.5 Mõõdikute tüübid

Anonüümumisalgoritmide kasulikkuse parandamiseks ja anonüümitud andmete kvaliteedi analüüsimiseks on alates k-anonüümsuse mõiste tutvustamisest loodud väga suur hulk erinevaid mõõdikuid. Need mõõdikud on ka väga varieeruvad. Probleemiks on, et teadmata nende tausta, on väga raske otsustada milliseid neist praktilistes tööriistades rakendada, kuna enamik spetsiifilisemaid mõõdikuid põhinevad põhjalikul teadustööl ja tulevad koos veenvate argumentidega. Siiski on mõõdikud bakalaureusetöö autori hinnangul võimalik jaotatada viite abstraktsemasse klassi:

- Statistilised mõõdikud
- Üldisemad informatsioonikao mõõdikud
- Privaatsusmudelite tagamist kontrollivad mõõdikud
- Tuntud rünnete riski hindavad mõõdikud
- Hierarhiapõhised informatsioonikao mõõdikud

Statistilised mõõdikud võivad olla mistahes andmeanalüüsis kasutatavad näitajad, nagu näiteks mood, mediaan, miinimum, maksimum jms. Üldisemad informatsioonikao mõõdikud on mõõdikud, mis kasutavad kao arvutamiseks ainult sisend- ja väljundandmestikku, võrreldes nendes tekkinud erinevusi ja ka näiteks neis esinevate ekvivalentsiklasside kirjeldavaid väärtusi. Privaatsusmudelite tagamist kontrollivad mõõdikud on selle töö raames üles ehitatud nende eelmainitud definitsioonidele, kontrollimaks, kas väljundandmestik rahuldab nõutud tingimusi.Tuntud rünnete riski hindavad mõõdikud põhinevad privaatsusmudelite tagamist kontrollivate mõõdikutega sarnastel ideedel, ent nende abil arvutatakse ka näiteks keskmist ja vähimat riski, kui otseselt privaatsusmudelite tagamist kontrollivad mõõdikud arvutavad ainult suurimat riski ning võrdlevad seda suurima lubatud riskiga. Hierarhiapõhised informatsioonikao mõõdikud on väga lai valdkond. Nende hulka kuuluvad näiteks: moonutuste mõõdik, i.k *distortion* [LwWFP06], kaotatud lehtede mõõdik, i.k *lost leaves metric* [MMD20] jt.

Selle bakalaureusetöö tulemuse raames rakendatakse kõiki abstraktsemaid mõõdikute klasse, välja arvatud hierarhiapõhiste informatsioonikao mõõdikute klass. Sellest aga lähemalt juba järgmises alapeatükis, kus tutvustatakse laiemat projekti, mille raames see bakalaureusetöö tehti.

2.6 Health Sense projekt

Health Sense[13] on andmete teisendamise keskkonna loomise projekt. Projekti raames luuakse keskkond, mille abil saab töödelda andmeid nii, et pärast töötlust võib neid

[13]Health Sense `https://www.tehik.ee/projektid`

käsitleda avaandmetena. See tähendab, et isikud andmestikus ei ole tuvastatavad ja seeläbi ei ole andmete välja andmiseks vaja läbida pikka kooskõlastuste ahelat.

Projekti laialdasemaks eesmärgiks on parandada andmete kättesaadavust, et soodustada uudseid lähenemisi ja lahendusi nii tervishoiu juhtimises kui ka erasektoris, sidudes need teadus- ja arendustegevustes tekkivate teadmistega. Konkreetsem eesmärk on suurendada andmete välja andmise kiirust ja vähendada sellega seonduvat tööjõukulu. Projekti kirjelduse kohaselt keskendutakse põhiliselt just terviseandmetele, kuna terviseandmete kasutamise vastu on maailmas kõige suurem huvi ja just nende andmete väljastamine on üldjuhul kõige aeganõudvam. Siiski plaanitakse projekti raames realiseeritav keskkond luua selliselt, et selle abil on võimalik töödelda ka mistahes teiste valdkondade andmeid.

Projekt jaguneb viieks alamtööks, mis on vastavalt:

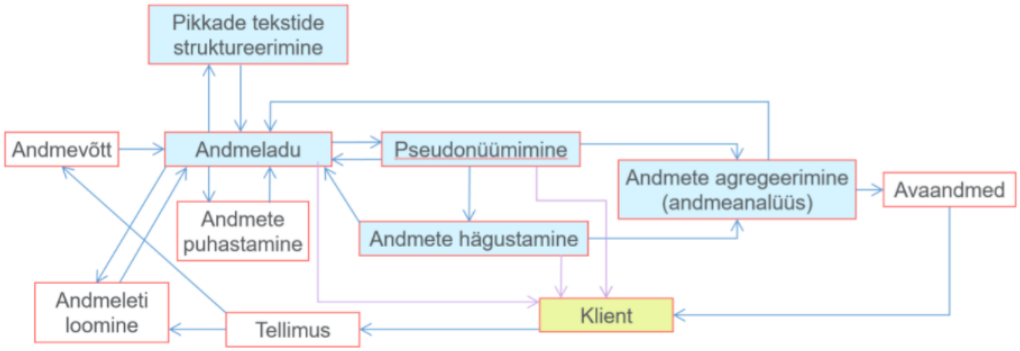
- Pseudonüümija
- Andmete hägustaja
- Pikkade tekstide struktureerija
- Andmelao andmemudelite loomine
- Analüüsikeskkonna loomine

Selles bakalaureusetöös keskendutakse andmete hägustaja alamtöö raames realiseeritava tarkvara väljundile analüüsikomponendi loomisele. Andmete hägustaja alamtöö eesmärk on luua tarkvara, mis võimaldab konkreetseid andmeid üldistada hierarhilise üldistamise abil. See tähendab, et sisuliselt on tegemist ühe laialdaselt levinud anonüümimise meetodi implementeerimisega praktilise kasutuse eesmärgil. Olulised skoobikitsendused andmete hägustajale, mis rakenduvad ka selle bakalaureusetöö raames loodavale analüüsikomponendile, on järgnevad:

- Sisendiks on **üks staatiline** tabel. See tähendab, et ei vaadelda anonüümimismeetodeid ja mõõdikuid, mis on mõttekad jooksvalt kasvavate või muutuvate andmestike raamistikus.
- Sisend **ei ole relatsiooniline**. Sisend võib küll olla väljavõte mõnest relatsioonilisest andmebaasist, ent sellisel juhul ei võeta andmete hägustaja töö käigus arvesse teisi sellega seoseid omavaid tabeleid.
- Lõppkasutaja võib olla **kes iganes**. See tähendab, et kasutajalt ei eeldata valdkonna ekspertteadmisi. Andmete hägustaja väljund peaks andma piisavat ja seejuures hoomatavat tagasisidet selle kohta, mis anonüümimise tulemusel paranes ka ekspertteadmisteta kasutajale.

Projekti rahastajaks on Norra Kuningriigi poolt asutatud meetme Norwegian Financial Mechanism poolt käivitatud toetusprogramm NORWAY GRANTS PROGRAMME "GREEN ICT". Partneriteks on Sotsiaalministeerium, Tervise ja Heaolu Infosüsteemide Keskus (TEHIK), Tartu Ülikool, Tallinna Tehnikaülikool, Helse West IKT (Norra) ja Western Norway University of Applied Sciences.

Järgnevalt tuleme tagasi töö alguses sõnastatud TEHIKu andmeväljastuse protsessi ja selles olevate probleemide juurde (vt alapeatükk 2.1.1) ning sõnastame andmete hägustaja loomisega kaasnevad lahendused püstitatud probleemidele. Andmeväljastuste parema korraldamise suurimaks eesmärgiks TEHIKu perspektiivist (saab üldistada ka muudele organisatsioonidele, kellel pole andmete väljastamine automatiseeritud) on vähendada andmete väljastamise protsessile kuluvaid ressursse läbi automatiseerimise, säilitades vastavuse andmekaitse regulatsioonidele. Health Sense raames loodava vahendite komplekti ülesandeks ongi soovitud andmete võimalikult kiire ja vähese tööjõukuluga komplekteerimine ning andmesubjektide ja muude seotud organisatsioonidega vajalike kooskõlastuste minimeerimine.



Joonis 16. Visioon andmete väljastamise protseduuri komponentidest pärast Health Sense'i. Allikas: [Mä22].

Joonisel 16 on kirjeldatud andmete väljastamise protseduuri oodatavaid tarkvarakomponente ja nendevahelisi seoseid pärast Health Sense projekti lõppu. Valgega märgitud tarkvarakomponendid on juba olemas, kollane sümboliseerib klienti (andmepäringu esitaja) ja sinisega märgitud tarkvarakomponendid plaanitakse realiseerida Health Sense raames. Health Sense projekti töögrupi ülesandeks on andmete hägustamise tarkvarakomponendi loomine. Selle bakalureusetöö raames valitakse välja mõõdikute komplekt üldistamisel põhineva anonüümimise kvaliteedi hindamiseks. Seejärel luuakse andmete hägustajale riskianalüüsi ja informatsioonikao mõõtmist teostavad alammoodulid ning komplekteeritakse need suuremasse analüüsikomponenti. Andmete hägustamise all on siin silmas peetud anonüümimist läbi hierarhilise üldistamise.

3 Metoodika

Selles peatükis kirjeldatakse lähemalt püstitatud probleemi lahendamiseks kasutatavaid meetodeid, tööriistade ja raamistike valikuid ning skoobikitsendusi. Lisaks põhjendatakse ka erinevate mõõdikute kasutuselevõtu osas tehtud otsuseid.

3.1 Küsimuste identifitseerimine

Selleks, et otsustada, milliseid mõõdikuid Health Sense'i raames projekti raames loodavas tarkvaras kasutada, on kõigepealt vaja tuvastada peamised kasutusjuhud ja küsimused, millele väljundi tulemused peaksid vastama. Kasutusjuhtude kohta pole eriti palju täpsemat infot kui TEHIKut ja andmeväljastuse protsessi probleeme üldiselt kirjeldav dokument, mille kohaselt oleks tellitud andmeid kokku panevatel töötajatel ressursside kokku hoidmiseks vaja automatiseeritud anonüümimistööriista. Siiski on Healt Sense'i projekti kirjelduse kohaselt laiem eesmärk, et tarkvara oleks üldkasutatav anonüümimiseks ka muudes organisatsioonides ning ideaalis võiks olla isegi vabalt alla laetav kõigile huvilistele.

Nendest eesmärkidest järeldub, et kasutaja võib olla kes iganes, s. t kasutajalt ei saa eeldada valdkonna põhjalikku tundmist. See tähendab, et küsimused, millele kasutaja analüüsikomponendilt vastuseid saada tahab, on tõenäoliselt üsna üldised ning peaksid olema koondatud fundamentaalsete nõuete ümber nagu näiteks privaatsusmudelite tagatuse kinnitus ja protsentuaalne muudetud väärtuste osakaal. Lisaks on tähtis ka ära märkida, et Health Sense'i töögrupi poolt loodav tööriist ei ole mõeldud anonüümimise analüüsimiseks ega ka kõikvõimalike tõhususe- ja kvaliteediprobleemide lahendamiseks, vaid mõistliku kvaliteediga automatiseeritud andmete anonüümimiseks. Analüüsi teostamiseks on sobivamad teistsugused tööriistad, mis proovivad endas implementeerida kõikvõimalikke anonüümimise valdkonnas tutvustatud lähenemisi, nagu näiteks ARX[14]. Kuna aga ARXis on proovitud implementeerida kõikvõimalikke lähenemisi, siis selle abil mõistlike tulemuste saamine on aega- ja põhjalikke teadmisi nõudev tegevus. Eelnevalt mainiti, et kasutajalt ei saa antud juhul põhjalikke teadmisi eeldada ja eesmärk on aja kokkuhoid, mitte erinevate anonüümimisviiside rakendamise võrdlusmomendi loomine vms. Sellest tulenevalt ei ole ka ARX ning sarnased tarkvarad üldjuhul sobivad.

Võttes arvesse eelnevalt välja toodud asjaolusid, formuleerisin viis põhilist küsimust, millele analüüsikomponent peaks tulemustega vastama. Nende küsimuste formuleerimine on üks selle bakalaureusetöö põhilisi tulemusi. Lähemalt on neid käsitletud peatükis 4. Järgnevalt tutvustatakse metoodikat kasutatavate mõõdikute valimisel.

[14] ARX anonüümija `https://arx.deidentifier.org/`

3.2 Mõõdikute valik

Mõõdikute valikul tulevad sisse Health Sense'i töögrupi ja TEHIKu poolt varasemates faasides vastu võetud otsused. Selle bakalaureusetöö raames ei ole oluline, milliseid parameetreid või millist tarkvara kasutatakse anonüümimisel. Küll aga on olulisteks skoobikitsendusteks otsused, et anonüümimisel kasutatakse ainult hierarhilist üldistamist.

See tähendab, et pole mõtet käsitleda näiteks ainult arvulistele väärtustele lisatud nn "müra"mõju andmestikule, kuna kõiki tunnuseid käsitletakse kategoorilistena ning et arvulised väärtused üldistatakse vahemikesse või peidetakse teatud osi neist, s.t müra lisamist ei toimugi. Selle eelduse kohaselt on koheselt välistatud mõõdikud, mis mõõdavad numbriliste väärtuste otsesest muutmisest tulenevaid erinevusi. Mõned numbrilisi väärtusi otseselt muutvad anonüümimisoperatsioonid on näiteks [JDF01]:

- Mikroagregeerimine
- Andmete moonutamine läbi tõenäosusliku jaotuse
- Müra lisamine

Nende strateegiate mittekasutamine tähendab, et sisend- ja väljundandmestikel ei ole tähenduslikku erinevust näiteks algse ja anonüümitud andmestiku mõne pidevaid väärtusi sisaldava tunnuse võrdlemisel leitud keskmisel ruutveal, keskmisel absoluutveal jms. väärtustel, seega ei ole nende võrdlemine informatiivne.

Veel üheks tugevaks kitsenduseks on kokkulepe, et Health Sense'i raames tegeletakse **ühe staatilise** tabeliga. See tähendab, et dünaamilised tabelid, kuhu kirjeid jooksvalt juurde lisatakse ja tabelite kogumid relatsiooniliste andmebaaside mõistes on väljaspool selle bakalaureusetöö ja Health Sense projekti skoopi. Eelmainitud põhjustel ei vaadelda ka näiteks diferentsiaalprivaatsust[15], multirelatsioonilist k-anonüümsust[16] jm. sarnaseid anonüümimismeetodeid. Töögrupi sisese otsusena võeti algselt kasutusele k-anonüümsus ja l-hajutus. Kaaluti ka t-lähedust[17], kuid otsustati mitte kasutusele võtta, kuna see on l-hajutuse rangem edasiarendus ning nõuab seetõttu kõrgemat privaatsust, kuid tekitab ka suuremat informatsioonikadu. Arutelu tulemusel otsustati, et liigse informatsioonikao tekkimise vältimiseks on mõistlik vähemalt projekti esimeses versioonis rakendada k-anonüümsust ja l-hajutust. Viimane oluline kitsendus on, et hierarhiate struktuurile ja formaadile ei panda eraldi nõudeid. Seda sellepärast, et Health Sense raames loodav anonüümimistööriist pakub võimalust lõppkasutajal ise hierarhiaid koostada ja neid võib koostada erinevatel viisidel. Näiteks võib hierarhilisel üldistamisel olla postiindeksi

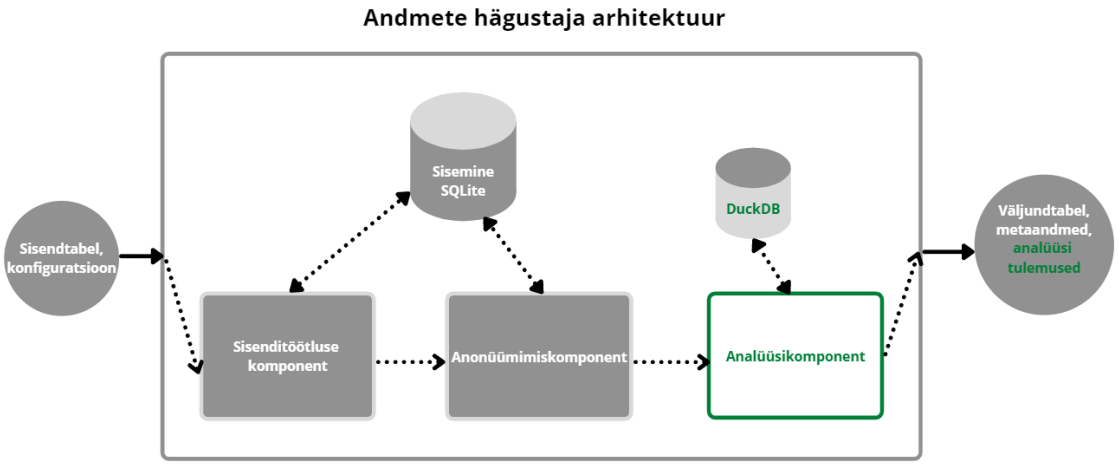
[15] https://akit.cyber.ee/term/1932-diferentsiaalprivaatsus
[16] https://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=2691&context=cstech
[17] https://akit.cyber.ee/term/3537-t-lahedus

50110 üldistamissammuks olla nii [50110-50119] kui ka 5011X ja selliste võimaluste vahel valimist otsustati mitte piirata. Sellest tulenevalt ei saa analüüsikonponendis informatsioonikao mõõdikute valikul kasutada mõõdikuid, mis eeldavad hierarhiate tundmist.

Võttes arvesse eelnevalt mainitud piiranguid, koostatakse vajalike ja sobivate mõõdikute tuvastamiseks maatriks, mille ridades on küsimused, mille vastused on lõppkasutajale olulised anonüümimiskomponendi töö valideerimisel ning veergudes on mõõdikute tüübid. Vastavalt paika pandud piirangutele ja formuleeritud küsimustele tehakse täpsemad otsused juba konkreetseid arvulisi või graafilisi tulemusi andvate mõõdikute kasutuselevõtu osas.

3.3 Mõõdikute arvutuse implementatsiooni põhimõtted

Pärast mõõdikute valiku tegemist implementeeriti nende arvutusalgoritmid Python programmina ja seejärel integreeriti Health Sense projekti raames loodavasse andmete hägustajasse. Andmete hägustaja arhitektuur ja selle bakalaureusetöö tulemusena lisanduv analüüsikomponent on illustreeritud joonisel 17.



Joonis 17. Analüüsikomponendi roll andmete hägustajas.

Analüüsikomponent implementeeritakse keeles Python, kuna eelnevad komponendid on realiseeritud keeles Python, ning sama programmeerimiskeele kasutamine hõlbustab komponentide liidestamist ja nende vahel toimuvat andmevahetust. Analüüsikomponent saab sisendiks anonüümimiskomponendi CSV (Comma Separated Values) formaadis väljundtabeli, töötlemata kujul sisendtabeli ja konfiguratsioonifaili. Põhiliste arvutusfunktsioonide realiseerimisel kasutatakse sisend- ja väljundtabeli omavahelisel võrdlusel

DuckDB[18] andmebaasiteeki, mis võimaldab CSV formaadis failidele lihtsasti esitada SQL keeles päringuid. Analüüsikomponendi põhiliseks oodatavaks väljundiks on nn "võti-väärtus" formaadis kirjed, kus võtmeks on mõõdiku nimetus ja väärtuseks komponendi käitamisel saadud tulemus. Sääraste "võti-väärtus" tüüpi andmestruktuuride praktikas kasutamiseks on Python-is võimalik kasutada sõnastikku, i.k *dictionary*. Vahetulemused talletataksegi Python-i sõnastikuna. Lisaks põhilisele väljundile luuakse ka mõningane graafiline väljund, näiteks visualiseerimaks anonüümimise tagajärjel tekkinud jaotuste erinevust mingis tunnuses. Graafiline väljund genereeritakse esialgu PNG[19] (Portable Network Graphics) formaadis ja talletatakse kuni analüüsikomponendi järgmise käitamiseni.

Health Sense projekti hilisemasse faasi on planeeritud põhjaliku kasutajaliidese loomine, selle valmimisel on võimalik graafiline väljund kuvada kasutajaliideses ka interaktiivsena, kuid see jääb selle bakalaureusetöö skoobist välja. Kuna siiski on ette teada, et luuakse põhjalik kasutajaliides mõnel JavaScript-il põhineval raamistikul, siis viiakse põhiosas tulemustena koostatud Python-i sõnastik JSON[20] (JavaScript Object Notation) formaati, sest selles formaadis on tulemused kasutajaliideses JavaScript-i abil lihtsasti parsitavad. Lisaks on Python-i sõnastik JSON-ile väga sarnane formaat ning Python-il on sisseehitatud meetodid sõnastike JSON formaati viimiseks.

3.4 Mõõdikute testimine

Kuna mõõdikute arvutusalgoritmide implementatsioonid luuakse definitsioonide põhjal, siis on selgelt vaja ka põhjalikku funktsionaalsuse testimist, et veenduda tarkvara õigesti töötamises. Testjuhud formuleeritakse käsitsi eraldi CSV formaadis tabelitena, kuna anonüümimiskomponendi sisendiks ja väljundiks on üldjuhul CSV formaadis failid. Testimisel kasutatavad andmed ei sisalda tegelikke isikuandmeid, küll aga on mõned näidised koostatud otseselt isikuandmete põhjal ja omavad piisavat sarnasust tegelike isikuandmetega.

Testjuhtude koostamisel luuakse 10..50-realised andmestikud, milles on mingid spetsiifilisemad olukorrad ning kaastakse tekstifail oodatavate väärtuste ja testi sisulise eesmärgi kirjeldusega. Testidega püütakse saavutada 100%-line koodikate, i.k *code coverage* ning simuleerida võimalikult paljusid oodatavaid äärejuhtumeid. Äärejuhtumeid testides eeldab analüüsikomponent ka mõningaid väliseid faile, mis on saadaval Health Sense raames loodava tarkvara teistes komponentides, nagu näiteks konfiguratsioonifail. Testimisel luuakse vajalikud failid simuleerimaks integreeritust ning erinevaid olukordi. Lisaks testitakse tarkvara käitumist ka potentsiaalselt ootamatute väärtustega konfiguratsioonifailis

[18] https://duckdb.org/2021/05/14/sql-on-pandas.html
[19] https://www.iso.org/standard/29581.html
[20] https://www.json.org/json-en.html

eesmärgiga vähendada sõltuvust välistest komponentidest. Graafilise väljundi loomist ei testita automaatselt, vaid visuaalselt ilma protsessi dokumenteerimata, kuna graafilise väljundi automaatne testimine on antud juhul ebamõistlik. Testimisel kasutatud failid tuuakse välja töö lisas I olevas repositooriumis.

Testide realiseerimiseks ja automatiseerimiseks kasutatakse raamistikku PyTest[21], mis võimaldab luua klassi- ja meetodipõhiseid teste. Teine võimalik variant oli raamistik unittest[22], kuid kuna sisenditöötluse komponendis on juba kasutusel PyTest, siis otsustati liigsete sõltuvuste, i. k *dependencies* vältimiseks Health Sense projektis testimiseks kasutada PyTest-i tarkvara. Testid luuakse klassipõhiselt, igale analüüsikomponendi alammoodulile hakkab vastama üks Python-i klass ning igale klassile üks PyTest-i abil realiseeritud testklass, testimaks vastavas klassis realiseeritud meetodeid.

Selles peatükis selgitati töö tulemuste saavutamiseks kasutatavaid lähenemisi, vastu võetud otsuseid ja skoobikitsendusi. Järgmises peatükis kirjeldatakse täpsemalt planeeritud lähenemiste kasutamise tulemusi ja tekkinud probleeme.

[21] `https://docs.pytest.org/en/7.1.x/`
[22] `https://docs.python.org/3/library/unittest.html`

4 Tulemused ja arutelu

Selles peatükis kirjeldatakse küsimuste formuleerimise, analüüsikomponendi loomise ja testimise tulemusi täpsemalt ning arutletakse tulemuste kvaliteedi üle. Täpsustatakse komponendi seesmist struktuuri ning analüüsikomponendi repositooriumi asukohta. Peatüki lõpus esitatakse ka analüüsikomponendi Health Sense raames loodavasse andmete hägustajasse integreerimise tulemused.

4.1 Küsimuste ja mõõdikute maatriks

Küsimuste ja potentsiaalselt kasutusele võetavate mõõdikute vahelise seose paremaks mõistmiseks formuleeriti viis küsimust, mille vastused võiksid anda lõppkasutajale piisava hulga informatsiooni, otsustamaks, kas anonüümimiskomponendi väljund on väljastamiskõlbulik, s.t kas soovitud privaatsus on tagatud ning kas andmed on endiselt kasutatavad.

Formuleeritud küsimused on alljärgnevad:

1. Kui suurt hulka andmetest muudeti?
2. Kui suur hulk andmeid kaotati täielikult?
3. Kas väljundtabel vastab valitud privaatsusmudelite nõuetele?
4. Kui palju väheneb põhilistest ründevektoritest tulenev risk?
5. Kuidas muutuvad üldised statistilised näitajad ja mitmekesisus sisend- ja väljundandmestikus?

Nendele küsimustele toetudes teostati seejärel analüüsikomponendi jaotamine veidi vähem abstraktseteks alammooduliteks ning formuleeriti maatriks, mille veergudes alammoodulite spetsiifilisemad nimetused ning ridades on küsimused. Maatriksi põhieesmärgiks oli luua visuaalne representatsioon vastust vajavate küsimuste ja implementeeritavate moodulite vahelisest suhtest. Formuleeritud maatriks on lõplikul kujul joonisel 18. Maatriksit vaadates on näha, et igale küsimusele vastab vähemalt üks alammoodul ning esineb ka ülekatet. Alammoodulites implementeeritavad mõõdikud valiti välja juba spetsiifiliselt joonisel 18 oleval joonisel olevale maatriksile toetudes, ning jaotati need sobivatesse alammoodulitesse.

Alammoodulites olevate mõõdikute valikul võeti arvesse metoodikas kirjeldatud piiranguid. Lõpliku otsusena välistati valikul kõik mõõdikud, mis ei ole relevantsed hierarhilise üldistamise abil anonüümitud ühe relatsioonideta staatilise tabeli raamistikus ning mis

ei ole informatiivsed k-anonüümsuse, l-hajutuse ja (X, Y)-anonüümsuse privaatsusmudelite mõistes. Lisaks võttes arvesse teadmist, et kasutajalt ei eeldata valdkonna põhjalikke teadmisi ning tööriist ei ole mõeldud mitmekordse katsetamise analüüsiks, välistati mõõdikud, mille tulemused sõltuvad omakorda mingitest parameetritest (sh näiteks ennustusmudelid) või mis on kasulikud peamiselt erinevate parameetritega saadud anonüümimistulemuste võrdlemiseks. Kuna hierarhiad sõltuvad andmestikust ja antakse edasi kasutajasisendina, jäeti välja ka mõõdikud, mis kasutavad näiteks informatsioonikao hindamiseks hierarhiaid. Nende asemel võeti kasutusele mõõdikud, mis annavad tagasisidet sisend- ja väljundtabeli otseste erinevuste ja muudetud või kaotatud väärtuste kohta. See hõlbustab ka tulemuste sisulist mõistmist laiemale kasutajaskonnale.

	Statistikamoodul	Ekvivalentsiklasside moodul	Jaotuste moodul	Privaatsusmudelite verifitseerimise moodul	Ründemudelite moodul
Kui suurt hulka andmetest muudeti?	✓				
Kui suur hulk andmeid kaotati täielikult?	✓		✓		
Kas väljundtabel vastab valitud privaatsusmudelite nõuetele?		✓		✓	
Kui palju väheneb põhilistest ründevektoritest tulenev risk?					✓
Kuidas muutuvad üldised statistilised näitajad ja mitmekesisus sisend- ja väljundandmestikus?	✓	✓	✓		

Joonis 18. Küsimuste ja alammoodulite maatriks

Lõpuks valiti välja mõõdikud moodulite kaupa alljärgnevalt:

- **Statistikamoodul** teeb arvutusi nii sisend- kui väljundandmestikul ning väljundandmestiku tulemustesse kirjutab ka kahe andmestiku vahelise võrdluse tulemusel leitud muudetud ja kaotatud väärtuste kogused. Sisendandmestikul arvutatakse järgnevad väärtused: unikaalsete väärtuste arv veerukaupa, informatiivsete väärtuste arv (s. t mitte täielikult kaotatud väärtuste koguarv) veerukaupa, moodid

veerukaupa. Väljundandmestikul arvutatakse kõik väärtused, mis arvutati ka sisendandmestikul, kuid lisanduvad järgnevad: täielikult kaotatud väärtuste osakaal muudetud väärtustest (%), muudetud või kaotatud väärtuste veerukaupa koguarv, kaotatud väärtuste koguarv ja osakaal veerus, kaotatud või muudetud väärtuste koguarv väljundandmestikus ning kaotatud väärtuste koguarv väljundandmestikus.

- **Ekvivalentsiklasside moodul** teeb arvutusi nii sisend- kui väljundandmestikul. Kuna mõlema andmestiku raames tehakse samad arvutused, siis saab väärtusi kirjeldada ilma alajaotuseta. Ekvivalentsiklasside moodul arvutab andmestike peal järgnevad väärtused: keskmine ekvivalentsiklassi suurus koos täielikult kaotatud QID väärtuste ekvivalentsiklassiga, keskmine ekvivalentsiklassi suurus ilma täielikult kaotatud QID väärtuste ekvivalentsiklassita, suurima ekvivalentsiklassi suurus, täielikult kaotatud QID väärtuste ekvivalentsiklassi suurus, ekvivalentsiklasside koguarv, andmestikus olevate kirjete koguarv ja väikseima ekvivalentsiklassi suurus.

- **Jaotuste moodul** ei arvuta otseselt midagi, vaid genereerib jaotuste graafikud veerukaupa enne ja pärast anonüümimist, eesmärgiga visualiseerida muudetavates veergudes jaotuse muutumist.

- **Privaatsusmudelite verifitseerimise moodul** teeb arvutusi ainult väljundandmestikul. Arvutatakse k-anonüümsuse, l-hajutuse ja (X, Y)-anonüümsuse tegelikud vähimad parameetrid ning juhul kui tegelik parameeter on väiksem kui konfiguratsioonis nõutud (näiteks konfiguratsioonis on nõutud 5-anonüümsust, aga andmestik on tegelikult 4-anonüümne), tuvastatakse kõik nõudmisi rikkuvad ekvivalentsiklassid.

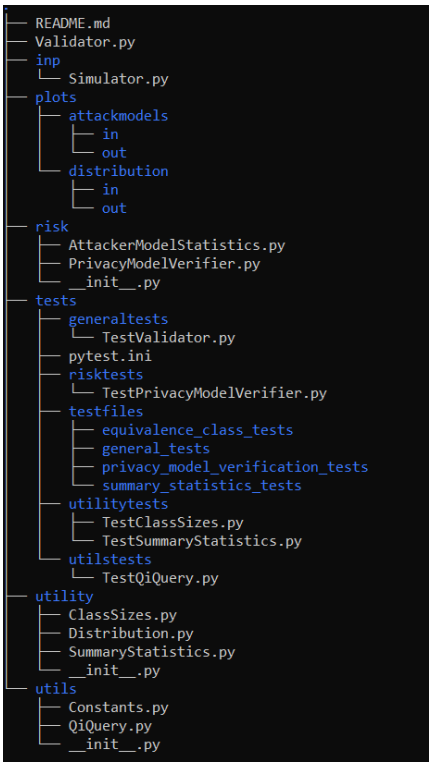
- **Ründemudelite moodul** teeb arvutusi nii sisend- kui väljundandmestikul ning arvutab samuti mõlema andmestiku jaoks samu väärtusi, seega saab väärtusi kirjeldada ilma alajaotuseta. Ründemudelite moodul arvutab andmestike peal järgnevad väärtused: keskmine prokuröri ründe õnnestumise risk, kõrgeim prokuröri ründe õnnestumise risk, vähim prokuröri ründe õnnestumise risk, ligikaused ajakirjaniku ja turundaja rünnete õnnestumise riskid, suurima riski mõjusfääris olevate kirjete osakaal, vähima riski mõjusfääris olevate kirjete osakaal. Lisaks genereeritakse graafikud visualiseerimaks riske enne ja pärast anonüümimist.

Siinkohal võib tekkida küsimusi, miks on statistikamoodulist välja jäetud kõikvõimalikud väärtused nagu mediaan, miinimum, maksimum jms. Sellele otsusele jõuti töögrupi siseselt, kuna kõiki väärtusi koheldakse kategoorilistena ning sellest tulenevalt ei ole miinimum ja maksimum tihtipeale tähenduslikud. Mediaan võib siiski mõnes olukorras tähenduslik olla, ent selle tähenduslikkus eeldab sorteeritust. Näiteks võttes vaatluse

alla TEHIKu andmetes tihedalt esineva väärtuse EHAK-kood[23] (Eesti haldus- ja asustusjaotuse klassifikaatori kood), siis nende sorteeritud vorm ei ole informatiivne ja sellest tulenevalt ei anna mediaan väärtuslikku informatsiooni. Mood see-eest annab aga informatsiooni näiteks selle kohta, et millise piirkonna elanikke on andmestikus kõige rohkem. Järgnevalt kirjeldatakse eelkirjeldatud mõõdikute implementeerimist tehnilisemast vaatevinklist.

4.2 Mõõdikute arvutusfunktsioonide loomine

Alammoodulid realiseeriti Python-is ning jaotati kahte laiemasse paketti, milleks on riskianalüüsi pakett ja informatsioonikao (kasutatavuse) analüüsi pakett. Analüüsikomponent realiseeriti esialgu eraldiseisva tarkvarakomponendina, mille struktuur on näha joonisel 19. Siiski analüüsikomponendi integreerimise soodustamiseks on Simulator.py failis implementeeritud kaks meetodit, mis imiteerivad Health Sense raames loodava tarkvara sees implementeeritud funktsioone.



Joonis 19. Analüüsikomponendi kui eraldi tarkvarakomponendi failipuu.

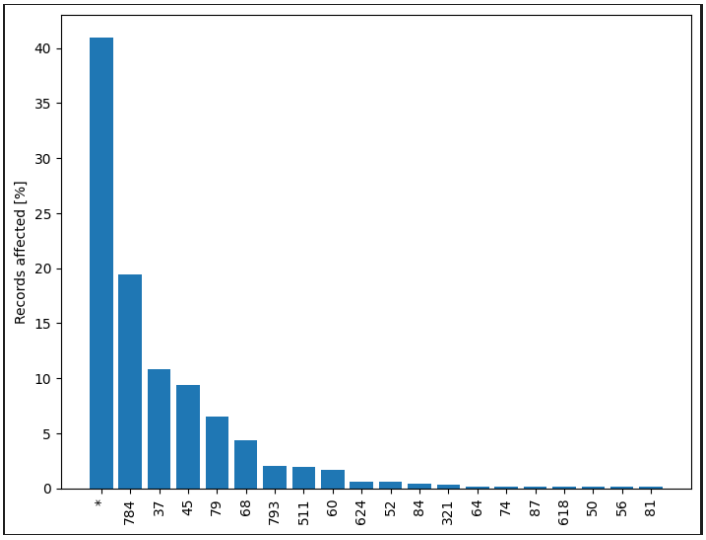
[23] https://klassifikaatorid.stat.ee/item/stat.ee/7fd0b185-4122-439b-bb27-d0f6c4d9c02b

Mõõdikute implementeerimiseks kasutati objektorienteeritud programmeerimise põhimõtteid. Kõiki küsimuste ja mõõdikute maatriksis olevaid alammooduleid esindab üks klass. Igas klassis on peameetod, mis tagastab vastava alammooduli väljundi alamsõnastiku Python-i sõnastiku kujul, mille võtmeks on alammooduli nimetus ning vajadusel on alammooduli väljund jaotatud sisend- ja väljundandmestiku jaoks veel eraldi alamsõnastikeks. Kõige sügavamal tasemel olevad väärtused vastavad juba konkreetsete mõõdikute väljunditele. Objektorienteeritud programmeerimist kasutati eesmärgiga hoida alammoodulid eraldiseisvatena ning vähendada programmis olevaid sõltuvusi. Nagu metoodikas mainiti, on sõnastike kasutamise eesmärk eelkõige väljundi kui terviku võimalikult lihtsasti edastatavasse formaati viimine ja väljundi pikkuse kasvamisel segaduse vältimiseks loogilise struktuuri säilitamine. Kuna Health Sense raames realiseeritava tarkvara kasutajaliides implementeeritakse JavaScript-il põhinevatel raamistikel, siis lihtsasti edastatava formaadi all on siinkohal mõeldud JSON-i. Väljund on terviklikult näitamiseks liiga pikk, ent joonisel 20 on näha väljundi esimesed 30 rida, kus on täielikult näha ründemudelite mooduli väljundit ning osaliselt näha ekvivalentsiklasside mooduli väljundit. Joonistel 22 ja 21 on näha ka kinnise testandmestikuga tarkvara käitamisel genereeritud graafikute näidiseid.


```
{
    "Attacker model risks": {
        "Input attacker model risks": {
            "Average prosecutor risk": "97.943 %",
            "Estimated journalist risk": "100.0 %",
            "Estimated marketer risk": "97.943 %",
            "Highest prosecutor risk": "100.0 %",
            "Lowest prosecutor risk": "33.333 %",
            "Records affected by highest risk": "96.016 %",
            "Records affected by lowest risk": "0.146 %"
        },
        "Output attacker model risks": {
            "Average prosecutor risk": "14.87 %",
            "Estimated journalist risk": "20.0 %",
            "Estimated marketer risk": "14.87 %",
            "Highest prosecutor risk": "20.0 %",
            "Lowest prosecutor risk": "4.348 %",
            "Records affected by highest risk": "25.632 %",
            "Records affected by lowest risk": "0.559 %"
        }
    },
    "Equivalence class statistics": {
        "Input equivalence class": {
            "Average equivalence class size (including suppressed)": 1.021,
            "Average equivalence class size (without suppressed)": 1.021,
            "Biggest equivalence class size": 3,
            "Completely suppressed class size": 0,
            "Number of classes": 4033,
            "Number of records": 4116,
            "Smallest equivalence class size": 1
        },
```


Joonis 20. JSON kujul väljundi esimesed 30 rida.

Joonis 21. Ajakirjaniku ründe edukuse tõenäosus anonüümitud testandmestikul.



Joonis 22. EHAK-koodide jaotus anonüümitud testandmestikul.

Väljundiga on täpsemalt võimalik tutvuda, kloonides lisas I toodud repositooriumis olev lähtekood ning käitades tarkvarakomponenti lokaalselt mõne kirjeldatud või vabalt loodud testjuhuga. Järgnevas alapeatükis kirjeldatakse lühidalt loodud testkomplekti.

4.3 Testimise tulemused

Testid on leitavad lisas I viidatud repositooriumi output_validation/tests alamkaustast, milles on alamkaustad riskianalüüsi teostavate alammoodulite ja informatsioonikao analüüsi teostatavate alammoodulite testimiseks. Igale loodud moodulile vastab üks testfail sama nimetusega, mis mooduli implementatsiooni fail, ainult prefiksiks on "Test". Lisaks on ka üks üldisem testifail nimega TestValidator.py, mis testib analüüsikomponenti

tervikuna. Testfailide sees on implementeeritud testfaili nimega testklassid, milles on meetoditena realiseeritud fundamentaalset funktsionaalsust ja oodatavaid äärejuhte katvad testid. Nagu metoodikas mainiti, on testkomplekt realiseeritud kasutades PyTest tarkvara. Alljärgnevaid käske tuleb jooksutada lisas I viidatud repositooriumi juurkaustas.

Testkomplekti on võimalik jooksutada järgneva käsuga:

```
$ pytest tests/
```

Kui on vaja ka koodikatte raportit, siis tuleks jooksutada järgnevat käsku:

```
$ coverage run --source . -m pytest tests/ && coverage report -m
```

Viimast mainitud käsku jooksutades saadakse koodikatte raport, mille väljund on näha joonisel 23.

```
jtavits@DESKTOP-O1CM3K7:~/projects/output_validation$ coverage run --source . -m pytest tests/ && coverage report -m
============================================================================================= test session starts
platform linux -- Python 3.8.10, pytest-6.2.5, py-1.11.0, pluggy-1.0.0
rootdir: /home/jtavits/projects/output_validation/tests, configfile: pytest.ini
collected 28 items

tests/generaltests/TestValidator.py .......
tests/risktests/TestPrivacyModelVerifier.py .......
tests/utilitytests/TestClassSizes.py ......
tests/utilitytests/TestSummaryStatistics.py ....
tests/utilstests/TestQiQuery.py ....

============================================================================================= 28 passed in 2.77s
Name                                                Stmts   Miss  Cover   Missing
---------------------------------------------------------------------------------
Validator.py                                           69      4    94%   114-118
inp/Simulator.py                                       22      0   100%
risk/AttackerModelStatistics.py                        80      0   100%
risk/PrivacyModelVerifier.py                          109      0   100%
risk/__init__.py                                        0      0   100%
tests/generaltests/TestValidator.py                    49      0   100%
tests/risktests/TestPrivacyModelVerifier.py            72      0   100%
tests/utilitytests/TestClassSizes.py                   75      0   100%
tests/utilitytests/TestSummaryStatistics.py            59      0   100%
tests/utilstests/TestQiQuery.py                        79      0   100%
utility/ClassSizes.py                                  81      0   100%
utility/Distribution.py                                42      0   100%
utility/SummaryStatistics.py                           77      0   100%
utility/__init__.py                                     0      0   100%
utils/Constants.py                                     53      0   100%
utils/QiQuery.py                                       56      0   100%
utils/__init__.py                                       0      0   100%
---------------------------------------------------------------------------------
TOTAL                                                 923      4    99%
jtavits@DESKTOP-O1CM3K7:~/projects/output_validation$
```

Joonis 23. Testkomplekti väljund ja koodikatte analüüs.

Raporti kohaselt jääb mulje, et ei ole saavutatud 100%-list koodikatet, vaid 99%. Siinkohal aga ainukesed katmata read on Validator.py failis olev analüüsikomponendi üldine peameetod, mida rakendatakse tarkvara käitamiseks. See tähendab, et vastav osa koodist

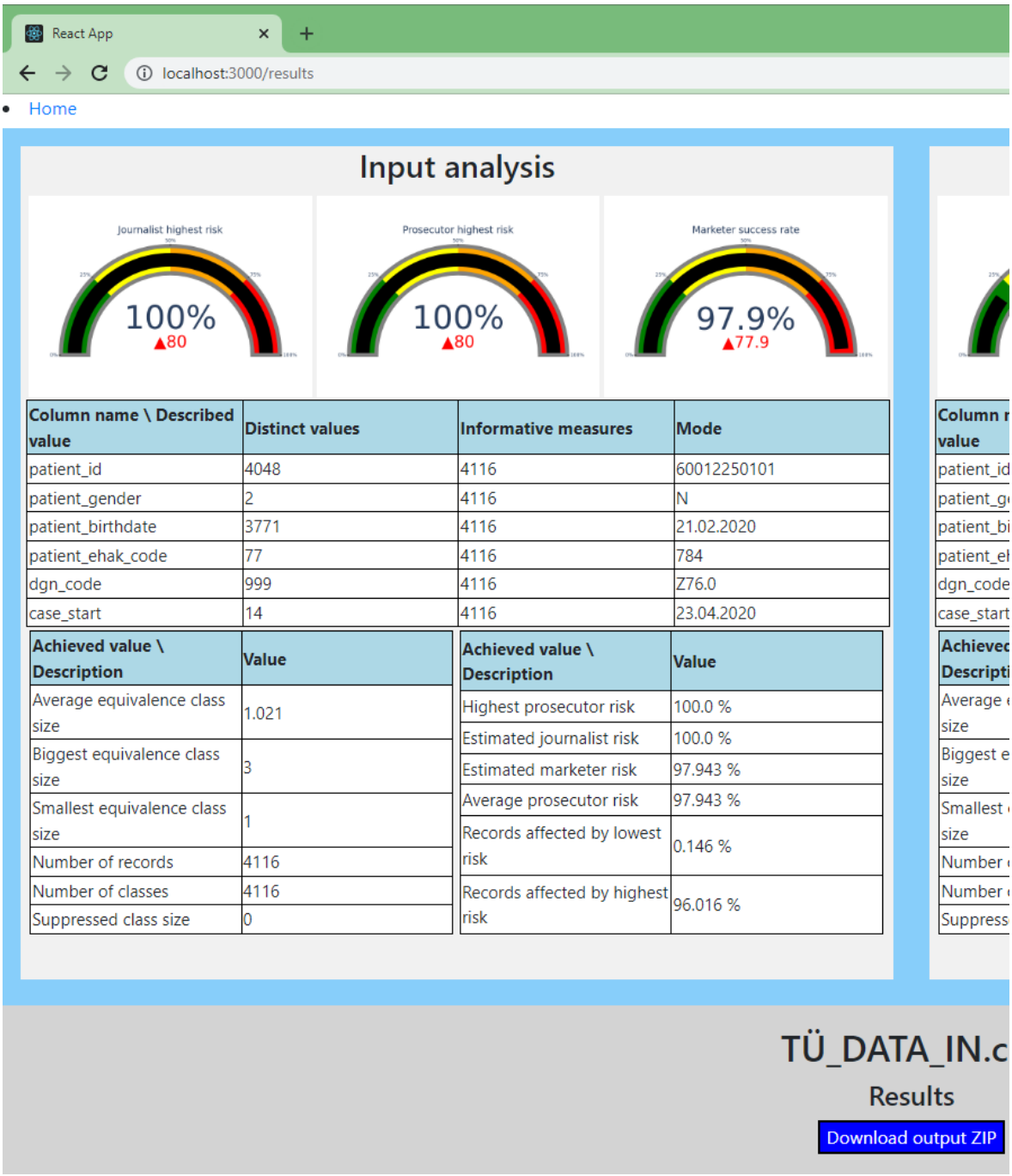
ei sisalda mingisugust seesmiselt vajalikku funktsionaalsust, äärmisel juhul sisaldavad need parameetrite puudumise korral rakendatavat veatöötlust. Sisulise funktsionaalsuse puudumise tõttu otsustati neid ridu mitte testida. Siiski mõõdikute arvutusalgoritmid ja abifunktsioonid on 100%-lise koodikatte ja lihtsamate äärejuhtudega testitud.

4.4 Tarkvara integreerimine

Valminud analüüsikomponent võeti kasutusele Health Sense raames loodavas andmete hägustajas. Selleks integreeriti lisas I olevas repositooriumis olev tarkvara Health Sense projekti lähtekoodi. Kuna andmete hägustaja tarkvara lähtekood ei ole hetkel avalik ning muud komponendid ei ole ka selle bakalaureusetöö raames otseselt olulised, siis andmete hägustaja suuremat repositooriumit selle tööga ei kaasata. Siiski on võimalik demonstreerida analüüsikomponendi väljundit andmete hägustaja viimase sammu tulemusel kasutajaliideses.

Joonistel 24 ja 25 on näha analüüsikomponendi väljundit andmete hägustaja tarkvara kasutajaliideses pärast anonüümimiskomponendi töö lõpetamist. Analüüsikomponent on tervikuna lisatud andmete hägustaja lähtekoodi ning liidestatud kasutajaliidesega läbi Flask[24] API. Analüüsikomponendi väljundi JSON saadetakse läbi API kasutajaliidesesse ning parsitakse seal JavaScript-i objektiks ja kuvatakse joonistel 24 ja 25 nähtavas formaadis. See formaat ei ole lõplik, kuna Health Sense raames loodav andmete hägustaja on endiselt arendusfaasis olev projekt. Kasutajaliideses oleva väljundi näol on tegemist pigem prototüübiga.

[24] `https://flask.palletsprojects.com/en/2.1.x/`

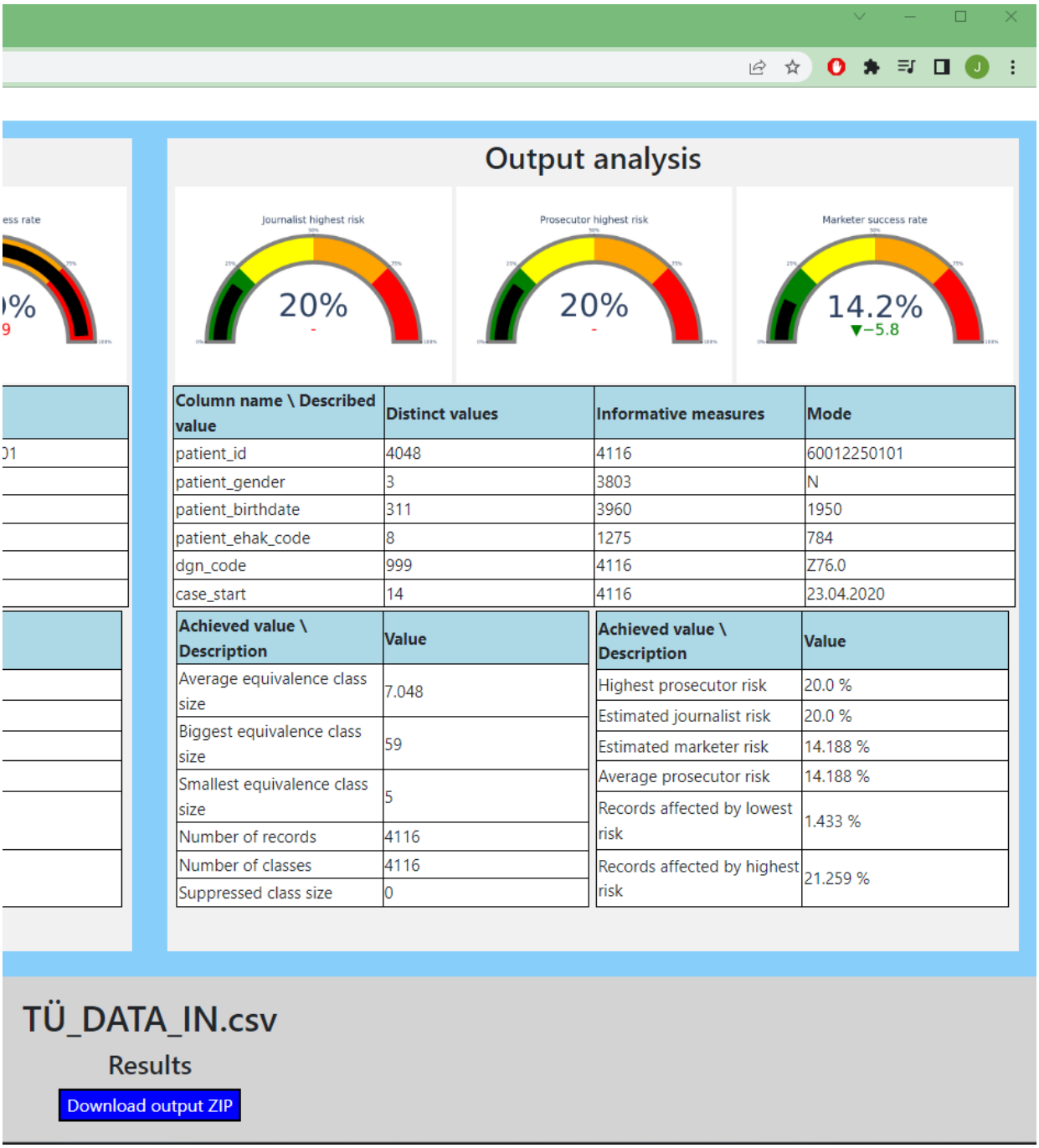


Column name \ Described value	Distinct values	Informative measures	Mode
patient_id	4048	4116	60012250101
patient_gender	2	4116	N
patient_birthdate	3771	4116	21.02.2020
patient_ehak_code	77	4116	784
dgn_code	999	4116	Z76.0
case_start	14	4116	23.04.2020

Achieved value \ Description	Value
Average equivalence class size	1.021
Biggest equivalence class size	3
Smallest equivalence class size	1
Number of records	4116
Number of classes	4116
Suppressed class size	0

Achieved value \ Description	Value
Highest prosecutor risk	100.0 %
Estimated journalist risk	100.0 %
Estimated marketer risk	97.943 %
Average prosecutor risk	97.943 %
Records affected by lowest risk	0.146 %
Records affected by highest risk	96.016 %

Joonis 24. Integreeritud tarkvara väljund teostatud analüüs sisendandmestikul Health Sense tarkvara kasutajaliideses.



Joonis 25. Integreeritud tarkvara väljundi teostatud analüüs anonüümitud andmestikul Health Sense tarkvara kasutajaliideses.

Joonistel 24 ja 25 olevad väljundid on tegelikult kuvatud samal lehel, ent on pildi selguse tagamiseks lisatud kahes osas. Võrreldes mõõdikute tulemusi sisendandmestikul ja anonüümitud andmestikul näeme, et privaatsusriskid on tunduvalt vähenenud. See on selge juba riskianalüüsi joonistelt. Keskmises tulbas on statistilise analüüsi tulemused veerukaupa. Alumises tulbas on andmestikuülesed väärtused. Kõiki selle bakalaureusetöö raames loodud mõõdikute väärtusi ei ole kuvatud, kuna tegemist on arendusfaasis tarkvaraga ja seetõttu ei ole väljundi disaini osas veel vastu võetud lõplikke otsuseid. Väljundi disain ei ka ole selle bakalaureusetöö raames oluline.

5 Kokkuvõte

Töös anti lühiülevaade Health Sense projektist ja TEHIKu vajadustest andmete anonüümimise raamistikus. Lisaks uuriti kaasaegseid meetodeid ja akadeemilisi hoiakuid anonüümimise suhtes andmekaitse valdkonnas ning tutvustati lühidalt enamlevinud privaatsusmudeleid koos nende tugevuste ja probleemidega. Töö põhieesmärgiks oli sobilike mõõdikute valimine üldistamisel põhineva anonüümimise kvaliteedi hindamiseks. Alameesmärgiks oli valitud mõõdikute realiseerimine ja integreerimine Health Sense projekti. Lisaks teostati ka loodud mõõdikute töötamise garanteerimiseks automaattestide komplekt.

Töös jagati mõõdikud viite abstraktsemasse kategooriasse ning võeti kasutusele neljast kategooriast kakskümmend viis täpsema eesmärgiga mõõdikut. Sealhulgas viis veerupõhist ja kakskümmend andmestikupõhist mõõdikut. Täielikult välja jäeti hierarhiapõhiste informatsioonikao mõõdikute kategooria ja selle alla kuuluvad mõõdikud. Pärast mõõdikute arvutusalgoritmide implementeerimist Python tarkvarakoodina läbi viidud testide tulemused kinnitavad, tarkvara töötab korrektselt. Lõpuks integreeriti loodud analüüsikomponent Health Sense projekti lähtekoodi.

Jätkutööna saaks täiendada analüüsikomponendi integreeritust andmete hägustaja kasutajaliidesega ja teha graafiline väljund interaktiivseks.

Viidatud kirjandus

[AS11] Cybernetica AS. *Andmekaitse ja infoturbe leksikon*, 2011. https://akit.cyber.ee/.

[Bam11] Jane Bambauer. Tragedy of the data commons. *SSRN Electronic Journal*, 03 2011.

[BCMFY10] Rui Chen Benjamin C. M. Fung, Ke Wang and PhilipŠ. Y. Privacy-preserving data publishing: A survey on recent developments. ACM Computing Surveys 42(4), 2010. https://www.researchgate.net/publication/220566406_Privacy-Preserving_Data_Publishing_A_Survey_of_Recent_Developments.

[Bog22] Dan Bogdanov. Lecture notes in privacy enhancing technologies course, February 2022.

[Col19] Liane Colonna. Privacy, risk, anonymization and data sharing in the internet of health things. Journal of Technology Law and Policy, Volume XX – 2019-2020 ISSN 2164-800X (online), 2019. https://tlp.law.pitt.edu/ojs/tlp/article/view/235/230.

[Dal77] T. Dalenius. Towards a methodology for statistical disclosure control. *Statistik Tidskrift*, 15(429-444):2–1, 1977.

[DE10] Fida Dankar and Khaled Emam. A method for evaluating marketer re-identification risk, 03 2010.

[DFSSC16] Josep Domingo-Ferrer, David Sánchez, and Jordi Soria-Comas. *Database Anonymization: Privacy Models, Data Utility, and Microaggregation-based Inter-model Connections*. Synthesis Lectures on Information Security, Privacy, Trust. Morgan and Claypool Publishers, 2016.

[ED08] Khaled Emam and Fida Dankar. Protecting privacy using k-anonymity. *Journal of the American Medical Informatics Association : JAMIA*, 15:627–37, 07 2008.

[EED08] Khaled El Emam and Fida Kamal Dankar. Protecting privacy using k-anonymity. *Journal of the American Medical Informatics Association : JAMIA*, 15(5):627–637, 2008. 18579830[pmid].

[Gar15] Simson Garfinkel. De-identification of personal information, 2015-10-22 2015.

[Gol06] Philippe Golle. Revisiting the uniqueness of simple demographics in the us population. In *2006 Workshop on Privacy in the Electronic Society*, pages 77–80. ACM Press, 2006.

[JDF01] V. Torra J. Domingo-Ferrer, J. M. Mateo-Sanz. Comparing sdc methods for microdata on the basis of information loss and disclosure risk. ResearchGate, 2001. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.2.3707&rep=rep1&type=pdf.

[LwWFP06] Jiuyong Li, Raymond Chi wing Wong, Ada Wai-Chee Fu, and Jian Pei. Achieving k-anonymity by clustering in attribute hierarchical structures. In *DaWaK*, 2006.

[MKGV07] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data*, 1(1):3–es, mar 2007.

[MMD20] Clémence Mauger, Gaël Le Mahec, and Gilles Dequen. Modeling and evaluation of k-anonymization metrics, 2020.

[Mä22] Margit Laurits Männiste. Tehik andmete väljastamise protsess. Jagatud e-kirja teel, kinnine allikas, 2022.

[Ohm09] Paul Ohm. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review, Vol. 57, p. 1701, 2010*, 2009.

[OMSG+21] Huda O. Mansour, Maheyzah M. Siraj, Fuad A. Ghaleb, Faisal Saeed, Eman H. Alkhammash, and Mohd A. Maarof. Quasi-identifier recognition algorithm for privacy preservation of cloud data based on risk reidentification. *Wireless Communications and Mobile Computing*, 2021:7154705, Aug 2021.

[PotEU16] European Parliament and Council of the European Union. *General Data Protection Regulation*, 2016. https://gdpr-info.eu/.

[SCDF15] Jordi Soria-Comas and Josep Domingo-Ferrer. Big data privacy: Challenges to privacy principles and models. Data Sci. Eng. (2016) 1(1):21–28, 2015. https://link.springer.com/content/pdf/10.1007/s41019-015-0001-x.pdf.

[SH16] Rubinstein Ira S. and Woodrow Hartzog. Anonymization and risk. *Wash. L. Rev.*, 91:703, 2016.

[SS98] Pierangela Samarati and Latanya Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. In *Proceedings of the IEEE Symposium on Research in Security and Privacy*, 1998.

[SS13] Paul Schwartz and Daniel Solove. Reconciling personal information in the united states and european union. *SSRN Electronic Journal*, 102, 05 2013.

[Swe00] Latanya Sweeney. Simple demographics often identify people uniquely. *Health (San Francisco)*, 671:1–34, 2000.

[TEH22] TEHIK. Meist, 2022. `https://www.tehik.ee/meist` (28.03.2022).

[WF06] Ke Wang and Benjamin Fung. Anonymizing sequential releases, 01 2006.

[ZWB+21] Z. Zuo, M. Watson, D. Budgen, R. Hall, C. Kennelly, and N. A. Moubayed. Data anonymization for pervasive health care: Systematic literature mapping study. JMIR Publications, 2021. `https://medinform.jmir.org/2021/10/e29871`.

Lisad

I. Repositoorium

Analüüsikomponendi ja automaattestide lähtekood ning testfailid, mis selle bakalaureusetöö raames realiseeriti on avalikud ja kättesaadavad GitHub-i repositooriumist:

`https://github.com/joosepgit/output_validation`

Kasutusjuhised on failis README.md

II. Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, **Joosep Tavits**,

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

   Andmete üldistamisel põhineva anonüümimise kvaliteedi hindamine,

   mille juhendaja on Sulev Reisberg,
   (juhendaja nimi)

   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.

Joosep Tavits
10.05.2022