Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Edo Ljubijankić

Vzorčna anonimizacija umetne zdravstvene podatkovne baze

DIPLOMSKO DELO
UNIVERZITETNI ŠTUDIJSKI PROGRAM PRVE STOPNJE
RAČUNALNIŠTVO IN INFORMATIKA

Miha Mraz
MENTOR

Ljubljana, 2018

Tematika naloge:

Kandidat naj v svojem delu predstavi področje anonimizacije podatkov in opravi pregled metod in programskih orodij, ki so aktualni na tem področju. Na osnovi analize umetne zdravstvene podatkovne baze naj metode med seboj primerja po uspešnosti izvedene anonimizacije in zmogljivosti njihove izvedbe.

IZJAVA O AVTORSTVU DIPLOMSKEGA DELA

Spodaj podpisani izjavljam, da sem avtor dela, da slednje ne vsebuje materiala, ki bi ga kdorkoli predhodno že objavil ali oddal v obravnavo za pridobitev naziva na univerzi ali drugem visokošolskem zavodu, razen v primerih kjer so navedeni viri.
S svojim podpisom zagotavljam, da:

- sem delo izdelal samostojno pod mentorstvom prof. dr. Mihe Mraza,
- so elektronska oblika dela, naslov (slov., angl.), povzetek (slov., angl.) ter ključne besede (slov., angl.) identični s tiskano obliko in
- soglašam z javno objavo elektronske oblike dela v zbirki "Dela FRI".

— Edo Ljubijankić, Ljubljana, junij 2018.

Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Edo Ljubijankić

Vzorčna anonimizacija umetne zdravstvene podatkovne baze

POVZETEK

Z anonimizacijo podatkov zaščitimo identito posameznika pred razkritjem občutljivih osebnih podatkov. Bolnišnice hranijo podatke o svojih pacientih v podatkovnih bazah, precejšen del teh podatkov pa se uporablja v raziskovalne namene. V skladu z zakoni so morajo osebni podatki pacientov anonimizirati, saj bi v nasprotnem primeru ogrozili zasebnost pacientov. V diplomskem delu smo primerjali nabor anonimizacijskih metod na umetno generirani zdravstveni podatkovni bazi. Bazo so sestavljali slovenski pacienti, kjer je imel vsak pacient za občutljiv osebni podatek določeno količino holesterola v krvi. Za izvajanje anonimizacijskih metod smo si pomagali s programskim orodjem ARX. Pri primerjavi metod smo merili hitrost izvajanja algoritma in kvaliteto anonimnih podatkov.

Ključne besede: anonimizacija podatkov, anonimizacijske metode, avtomatizacija anonimizacije

University of Ljubljana
Faculty of Computer and Information Science

Edo Ljubijankić

Comparative study of anonymization methods on synthetically generated health database

ABSTRACT

By anonymizing data, we protect the identity of an individual from disclosing sensitive personal information. Hospitals store data on their patients in databases. Many of stored data is used for research purposes. According to the laws, personal data of patients must be anonymised in order to ensure the patient's privacy. In this thesis we compared a set of anonymization methods on the synthetically generated database. The database was composed of Slovenian patients, where sensitive personal information of each patient is their cholesterol level. We used the ARX software tool to perform anonymization methods. When comparing the methods, we measured the speed of the algorithm and the quality of anonymous data.

Key words: data anonymization, anonymization methods, automatisation of anonymization

ZAHVALA

Zahvaljujem se družini za podporo in vzpodbudo. Zahvaljujem se tudi mentorju prof. dr. Mihi Mrazu za odlične napotke in strokovno pomoč pri izdelavi diplomskega dela.

— Edo Ljubijankić, Ljubljana, junij 2018.

KAZALO

1 Uvod

V današnjem času je skorajda nemogoče vzpostaviti kontakt z nepoznano osebo brez osebne predstavitve. Pri predstavitvi ne izdamo vseh osebnih podatkov o sebi. Javne ustanove so zelo pazljive pri javni objavi osebnih podatkov, ker slednja mnogokrat ni v skladu z zakoni in bi v primeru tožb privedlo do resnih finančnih izgub, pravnih problemov in osebnih neprijetnosti [1]. Da bi osebne podatke posameznikov čimbolje zaščitili in ohranili njihovo zasebnost, uporabljamo anonimizacijo podatkov. Na anonimiziranih podatkih lahko raziskovalci opravljajo analize brez kršitve zasebnosti posameznikov od katerih so podatki pridobljeni.

Anonimizacija podatkov je proces šifriranja ali odstranjevanja osebno prepoznavnih informacij iz podatkov z namenom preprečevanja identifikacije posameznika, na katerega se podatki nanašajo. V okviru zdravstvenih podatkov se npr. anonimizirani podatki nanašajo na podatke bolnikov. Da bi varovali zasebnost bolnika je potrebno ime, naslov in celotno poštno številko odstraniti skupaj z vsemi drugimi informacijami, ki identificirajo bolnika [2].

Ogromne količine podatkov o posameznikih so v današnjem času zabeležene v sistemih

baz podatkov. Varnost teh podatkov je ključna, zato je iskanje izboljšav na področju anonimizacije še vedno aktualna tema. V nadaljevanju omenimo dva bolj znana incidenta, kjer podatki niso bili ustrezno anonimizirani pred njihovo objavo. Leta 2006 je podjetje Netflix najavilo milijon dolarjev vreden izziv, ki bi pripomogel k izboljšanju njihovega sistema za priporočanje filmov. Za pomoč tekmovalcem je Netflix javno objavil 10 milijonov filmskih ocen 500.000 uporabnikov. Podatkom so odstranili osebne informacije, ki bi identificirale posameznike, šifre pa so nadomestila imena. Raziskovalca iz Texasa sta kljub temu identificirala nekaj uporabnikov s primerjavo ocen filmov in časovnih žigov z javnimi informacijami objavljenimi v Internet Movie Database (IMDb) [3]. Istega leta je ameriški ponudnik interneta AOL na svoji spletni strani delil anonimne podatke o internetnem iskanju 650.000 uporabnikov. Vendar so naredili napako, saj so bile osebno prepoznavne informacije prisotne v številnih poizvedbah. Novinarji New York Timesa so lahko v nekaj dneh identificirali posameznike glede na njihovo zgodovino iskanj [4].

1.1 Motivacija

Po raziskavah iz leta 2016 ima ogromno ljudi težave s prekomerno težo. Kar dve tretjini Slovencev je pretežkih, v Evropski uniji pa Slovenija zaseda sedmo mesto po deležu debelih [5]. Debelost je eden od dejavnikov, ki vpliva na raven holesterola v krvi. Priporočena vrednost skupnega holesterola je manj kot 5 mmol/l, vendar ima kar 60% odraslih Slovencev povišan holesterol v krvi [6]. Bolnišnice in mnoge organizacije, ki posedujejo osebne podatke posameznikov se zavedajo, da morajo podatke anonimizirati, preden jih lahko objavijo ali delijo z raziskovalci. Da bi zaščitili zasebnost posameznikov, je zato potrebno uporabiti primerno anonimizacijsko metodo.
Odločili smo se, da bomo v okviru diplomske naloge pridobili podatkovno bazo slovenskih pacientov z njihovimi vrednostmi holesterola v krvi in jo anonimizirali z različnimi anonimizacijskimi metodami. Nato bomo primerjali uporabljene metode in ugotovili, katera bi bila najprimernejša za našo bazo podatkov.

1.2 Zakonske podlage

Tehnološki napredek in digitalizacija poslovanja so omogočile lažjo dostopnost, hranjenje in izmenjavo osebnih podatkov. Vzpostavil se je zakon za varovanje osebnih podatkov pred zlorabo identitet in vdiranja v zasebnost posameznikov. V Sloveniji je v veljavi

Zakon o varstvu osebnih podatkov (kratica ZVOP-1) [7].
V Evropski uniji je z 25.5.2018 začela veljati nova uredba GDPR (Splošna uredba EU o varstvu podatkov), ki zagotavlja nov sistem varstva osebnih podatkov. Novost posameznikom omogoča nadzor nad svojimi osebnimi podatki. Podatki se upoštevajoč uredbo zbirajo in obdelujejo samo ob podpisu soglasja oziroma privolitvi posameznika. Prav tako ima posameznik pravico za preklic privolitve. V Sloveniji je bil podan predlog o uveljavi ZVOP-2 zakona, ki bi vsebinsko sledil uredbi GDPR. ZVOP-2 bo dokončna prilagoditev Slovenije na GDPR [8].

1.3 Pregled dela

V diplomski nalogi bomo anonimizirali podatkovno bazo s pomočjo programskega orodja. Potek anonimizacije je shematsko prikazan na sliki 1.1. Orodje nam bo omogočalo, da izvajamo različne anonimizacijske metode na testni bazi podatkov in primerjamo rezultate. S pomočjo metrik bomo ocenili uspešnost uporabljenih metod.
V drugem poglavju so opisane različne anonimizacijske metode ter njihove slabosti in prednosti. Programsko orodje in metrike so podrobneje opisane v tretjem poglavju. V četrtem poglavju opišemo našo testno bazo, postopek izvajanja anonimizacije in pridobljene rezultate. Nato rezultate še povzamemo v zaključku.

Slika 1.1 Programsko orodje izvaja anonimizacijo nad podatkovno bazo (PB).

2 Osnove metod anonimizacije

S pojavom potreb po anonimizaciji podatkov so se oblikovale metode za njihovo učinkovito anonimizacijo. Cilj metod je anonimizirane podatke uporabiti za obdelavo ali analizo, istočasno pa preprečiti tretji osebi reidentifikacijo originalnih podatkov [9].
Podatki so običajno shranjeni v podatkovnih bazah. Za prikaz teh podatkov ponavadi uporabljamo tabele (npr. glej tabelo 2.1). Posamezna vrstica tabele predstavlja podatkovni zapis oz. n-terico atributov, medtem ko posamezen stolpec predstavlja en atribut [10]. Vloge atributov so z vidika anonimizacije lahko sledeče:

- identifikator ali ključ podatkovne baze,
- kvazi identifikator (angl. $quasi - identifier$),
- občutljiv atribut,
- neobčutljiv atribut.

Identifikatorji neposredno identificirajo posameznika in so pred javno objavo podatkov odstranjeni. Takšni atributi so npr. ime, priimek, EMŠO itd. *Kvazi identifikatorji* so

Ime	Starost	Kraj	Mesečna plača v evrih
Janez	23	Ljubljana	900,00
Matej	31	Berlin	1100,00
Ana	44	Pariz	950,00
Mojca	27	Maribor	900,00

Tabela 2.1 Primer relacijske tabele. Vrstice predstavljajo zapise, stolpci pa atribute.

atributi, ki se navezujejo na osebne podatke in s kombinacijo le teh je možno identificirati posameznika z veliko verjetnostjo. Takšni atributi so npr. rojstni datum, spol, naslov itd. Teh atributov ne moremo odstraniti, sicer anonimizacija ne bi imela smisla, saj so potrebni za kasnejšo obdelavo podatkov. *Občutljivi atributi* vsebujejo zasebne podatke, ki jih ne želimo javno objaviti (npr. vrsta bolezni) [11]. Ostali atributi spadajo pod *neobčutljive atribute*. Slednji so pri anonimizaciji ignorirani.

Atribute lahko na splošno delimo po njihovih vrednostih. Atribute, ki imajo številčne vrednosti, poimenujemo *numerični atributi*. Poleg numeričnih atributov imamo tudi *kategorične atribute*. Primer kategoričnega atributa je atribut spol (posameznik spada v eno izmed dveh kategorij: moški ali ženska). Včasih so vrednosti kategoričnega atributa kodirane kot številke (npr. spol se lahko kodira kot 0 za moške in 1 za ženske). Kljub temu velja, da je atribut kategoričen.

2.1 Osnovne metode anonimizacije

Poznamo kar nekaj osnovnih, a zelo uspešnih metod za dosego anonimizacije podatkov. Njihovo uporabo predstavimo na primeru tabele 2.1 iz relacijske podatkovne baze. V nadaljevanju predstavimo štiri najbolj razširjene metode anonimizacije.

2.1.1 Prikrivanje podatkov

Prikrivanje (angl. *suppression*) oz. odstranitev podatkov je najenostavnejša metoda anonimizacije. S to metodo vrednosti določenih atributov zamenjamo z znakom '*', ki predstavlja poljubno vrednost. V anonimizirani tabeli 2.2 so zamenjane vse vrednosti atributa `Ime` z znakom '*'.

Ime	Starost	Kraj	Mesečna plača v evrih
*	23	Ljubljana	900,00
*	31	Berlin	1100,00
*	44	Pariz	950,00
*	27	Maribor	900,00

Tabela 2.2 Prikaz prikrivanja podatkov po atributu `Ime`.

Ime	Starost	Kraj	Mesečna plača v evrih
*	[21-30]	Ljubljana	900,00
*	[31-40]	Berlin	1100,00
*	[41-50]	Pariz	950,00
*	[21-30]	Maribor	900,00

Tabela 2.3 Prikaz posplošitve podatkov po atributu `Starost`.

2.1.2 Posploševanje podatkov

Posploševanje (angl. *generalization*) je metoda anonimizacije, ki se izvede tako, da se prvotno vrednost atributa zamenja z bolj splošno, a semantično ustrezno vrednostjo. Eden od načinov je uvedba velikostnih razredov npr. država namesto mesta, starostni interval namesto starosti [12] itd. Primer posploševanja vrednosti atributa `Starost` je prikazan v tabeli 2.3.

2.1.3 Dodajanje šuma

Dodajanje šuma pomeni dodajanje naključnih števil k vrednostim atributov kvantitativne vrste [13]. Metoda je uporabna, ker s šumom zmanjšamo natančnost vrednosti atributov, ki lahko škodijo posamezniku. Če uvedemo dovolj dober šum, potem tretja oseba ne more zaznati spremembe na podatkih ali podatkov povrniti v prvotno stanje. Dodajanje šuma ponazorimo z izrazom

$$X + \epsilon = Z, \tag{2.1}$$

kjer Z predstavlja nabor vrednosti, ki ga dobimo, ko naboru originalnih vrednosti X dodamo šum. Npr. če atributu `Višina`, ki ima nabor vrednosti $X = \{175, 180, 185\}$, dodamo šum ϵ=10, dobimo anonimiziran nabor vrednosti $Z = \{185, 190, 195\}$.

Ime	Starost	Kraj	Mesečna plača v evrih
*	23	Ljubljana	1100,00
*	31	Berlin	900,00
*	44	Pariz	900,00
*	27	Maribor	950,00

Tabela 2.4 Prikaz permutacije podatkov, prvi in drugi zapis sta zamenjala vrednost atributa `Mesečna plača v evrih`, prav tako tretji in četrti zapis.

2.1.4 Permutacije

Permutacija je metoda, pri kateri se vrednosti atributov ne spremenijo, se pa vrednosti atributov med zapisi zamenjajo. Tabela 2.4 prikazuje primer tabele po permutaciji na tabeli 2.2. V tem primeru sta si prvi in drugi zapis ter tretji in četrti zapis v tabeli 2.2 med seboj zamenjala vrednost atributa `Mesečna plača v evrih`.

2.2 k-anonimnost

V devetdesetih letih prejšnega stoletja je Latanya Sweeney preučevala javno objavljene anonimizirane podatke s področja zdravstva, ki niso imeli identifikatorjev, so pa bi bili poleg zdravstvenih podatkov objavljeni atributi `Spol`, `Poštna številka` in `Rojstni datum`. Prišla je do ugotovitve, da je 87% prebivalstva Združenih držav enolično določljiva na osnovi naštetih atributov in jih je možno reidentificirati iz javnih evidenc samo s kombinacijo teh treh kvazi identifikatorjev, ter dokazala, da anonimizacija ni bila izvedena ustrezno [14].

Da bi se preprečila identifikacija posameznikov le s kombinacijo kvazi identifikatorjev, se je razvila metoda k-anonimnost, ki temelji na prikrivanju in posploševanju podatkov. Koncept metode sta prvi predstavili Latanya Sweeney in Pierangela Samarati, njen opis pa najdemo v virih [15, 16]. Cilj metode je, da v tabeli vrednosti kvazi identifikatorjev prikrivamo in posplošujemo do te mere, da dobimo skupine k podatkovnih zapisov, ki imajo iste vrednosti kvazi identifikatorjev [11].

Primer uporabe metode si oglejmo na naslednjem primeru povzetem po viru [17]. Tabela 2.5 vsebuje podatke o pacientih. Tabela ne vsebuje identifikacijskih atributov, kot so ime, številka socialnega zavarovanja itd. Atribute delimo na kvazi identifikatorje in občutljive atribute. Nabor atributov `Poštna številka`, `Starost` in `Spol` pacienta so v

Poštna številka	Starost	Spol	Bolezen
13053	28	moški	težave s srcem
13068	29	ženska	težave s srcem
13068	21	ženska	virusna okužba
13053	23	moški	virusna okužba
14853	50	moški	rak
14853	55	ženska	težave s srcem
14850	47	moški	virusna okužba
14850	49	moški	virusna okužba
13053	31	ženska	rak
13053	37	moški	rak
13068	36	ženska	rak
13068	35	moški	rak

Tabela 2.5 Primer neanonimizirane tabele s podatki pacientov.

tem primeru kvazi identifikatorji. Občutljivi atribut je v tem primeru `Bolezen` pacienta. Nad to tabelo izvajamo k-anonimizacijo tako, da prikrijemo vrednosti atributa `Spol` in posplošimo vrednosti atributoma `Poštna številka` in `Starost`. Rezultat je anonimizirana tabela 2.6, ki je 4-anonimna. To pomeni, da za vsak posamezen zapis v tej tabeli obstajajo še vsaj trije zapisi z enakimi vrednostmi kvazi identifikatorjev.

Preprost koncept metode je prinesel veliko popularnost. Kljub temu je k-anonimnost izpostavljena nevarnosti, ker ne zaščiti vrednosti občutljivih atributov. Znani sta dve vrsti nevarnosti ali dve vrsti napada tj. homogenostni napad in napad s poznavanjem ozadja [17]. Homogenostni napad velja v primerih, ko so vse vrednosti za občutljive atribute v nizu zapisov dolžine k enake. Napad si ogledamo na naslednjem zgledu. Predstavljajmo si osebo, ki ima soseda v bolnišnici. Zaradi radovednosti bi ta oseba rada odkrila, katero bolezen ima njen enaintridesetletni sosed. Predpostavimo, da oseba pozna poštno številko svojega soseda in njegovo starost. S pogledom na 4-anonimno tabelo 2.6 trenutnih pacientov bolnišnice bi oseba ugotovila, da so podatki o njenem sosedu v enem izmed zadnjih štirih zapisov. Ker pa imajo vsi štirje zapisi enako vrednost po atributu `Bolezen`, bi oseba ugotovila, da ima njen sosed raka. V takih primerih, čeprav so bili podatki k-anonimizirani, je vrednost občutljivih atributov za niz zapisov k na-

Poštna številka	Starost	Spol	Bolezen
130**	< 30	*	težave s srcem
130**	< 30	*	težave s srcem
130**	< 30	*	virusna okužba
130**	< 30	*	virusna okužba
1485*	≥ 40	*	rak
1485*	≥ 40	*	težave s srcem
1485*	≥ 40	*	virusna okužba
1485*	≥ 40	*	virusna okužba
130**	3*	*	rak
130**	3*	*	rak
130**	3*	*	rak
130**	3*	*	rak

Tabela 2.6 Primer 4-anonimne tabele. Najmanj štirje podatkovni zapisi si delijo isti kvazi identifikator.

tančno predvidena. Pri napadu s poznavanjem ozadja tretja oseba z uporabo dodatnih informacij iz ozadja sklepa, kateri občutljiv atribut pripada posamezniku. Kot zgled predpostavimo, da imamo osebo, ki išče podatke o svoji 21-letni prijateljici japonske narodnosti v 4-anonimni tabeli 2.6. Ker ve, da je njena poštna številka enaka 13068, ugotovi, da je njen zapis med prvimi štirimi v 4-anonimni tabeli 2.6. Tej osebi je znano, da imajo Japonci izjemno malo srčnih bolezni. Na podlagi te informacije iz ozadja sklepa, da prijateljica osebe nima težav s srcem, ampak z virusno okužbo.

2.3 l-raznolikost

V prejšnjem razdelku smo pokazali, da je k-anonimnost ranljiva na homogenostni napad in napad s poznavanjem ozadja, zato je potrebna močnejša opredelitev zasebnosti. To nam omogoča metoda l-raznolikosti, ki je izboljšana različica metode k-anonimnosti. Cilj metode je, da se občutljivi atributi v vsaki množici k zapisov razlikujejo na vsaj l načinov [16, 17]. Napadalca tako pustimo v negotovosti glede atributov, ker bo zaradi raznolikosti občutljivih atributov težje reidentificirati posameznika v množici k zapisov. V tabeli 2.7 je prikazan primer 3-raznolikosti z enakomerno porazdelitvijo vrednosti občutljivih atributov. Vendar l-raznolikosti ni enostavno doseči in je včasih nepotrebna. Prav tako

Poštna številka	Starost	Spol	Bolezen
476**	[21-30]	*	težave s srcem
476**	[21-30]	*	rak
476**	[21-30]	*	virusna okužba
479**	[41-50]	*	virusna okužba
479**	[41-50]	*	težave s srcem
479**	[41-50]	*	rak
476**	[31-40]	*	virusna okužba
476**	[31-40]	*	rak
476**	[31-40]	*	težave s srcem

Tabela 2.7 Primer 3-anonimne 3-raznolike tabele o pacientih.

napadalcu omogoča, da razbere informacije iz občutljivih atributov, če so le ti semantično podobni [11, 16, 18]. Npr. skupina zapisov z atributom `Bolezen` vsebuje tri vrednosti in sicer pljučni rak, jetrni rak in rak želodca. Napadalec iz te skupine zapisov sklepa, da ima pacient raka.

2.4 t-podobnost

Glede na obstoj napadov na metodo l-raznolikosti, kjer je mogoče sklepati in razbrati informacije iz občutljivih atributov, se je ustvarila metoda t-podobnost. Osnovna ideja metode teži k temu, da je porazdelitev občutljivih atributov v ekvivalenčnem razredu približno podobna porazdelitvi občutljivih atributov v celotni podatkovni tabeli. Podatkovni zapisi, ki imajo enake vrednosti kvazi identifikatorjev, predstavljajo ekvivalenčni razred. Po definiciji iz vira [18] velja, da ekvivalenčni razred zadošča t-podobnosti, če je razdalja med porazdelitvijo vrednosti občutljivega atributa v tem razredu in porazdelitvijo vrednosti občutljivega atributa v celotni tabeli manjša od t. Za tabelo velja t-podobnost, če vsak ekvivalenčni razred zadošča t-podobnosti.

Definicija t-podobnosti ne predpisuje nobene posebne razdalje med porazdelitvami. Vendar je EMD (angl. *Earth Mover Distance*) običajna razdalja, ki se uporablja za opredelitev t-podobnosti. Glavna prednost EMD je, da lahko zajame semantično razdaljo med vrednostmi. EMD(P,Q) meri najmanjšo količino dela, ki je potrebna za pretvorbo porazdelitve P v porazdelitev Q [18, 19]. Predpostavimo, da imamo niz vrednosti $\{v_1,....,v_r\}$ in

dve verjetnostni porazdelitvi P=(p_1,....,p_r) in Q=(q_1,....,q_r), kjer sta p_i in q_i verjetnosti, ki jih P in Q dodelita v_i. Razdaljo med vrednostima v_i in v_j označimo z d_{ij}, pretok med istima vrednostima pa s f_{ij}. EMD(P,Q) potem izrazimo z enačbo

$$EMD(P,Q) = \min_{f_{ij}} \sum_{i=1}^{r} \sum_{j=1}^{r} d_{ij} f_{ij}. \tag{2.2}$$

Enačba upošteva pogoj $f_{ij} \geq 0$, kjer $1 \leq i,j \leq$ r.

Prvi korak pri izračunu EMD razdalje je določitev merjenja razdalje d. Način je odvisen od tipa podatkov, ki jih uporabljamo. Pri numeričnih atributih (npr. `Starost`) se razdalja med dvema vrednostima izračuna na osnovi izraza

$$d(v_i, v_j) = \frac{|i-j|}{r-1}. \tag{2.3}$$

Iz tega izpeljemo EMD enačbo za numerične atribute po izrazu

$$EMD(P,Q) = \frac{1}{r-1} \sum_{i=1}^{r} | \sum_{j=1}^{i} (p_i - q_i)|. \tag{2.4}$$

Za kategorične atribute (npr. `Bolezen`) lahko pri merjenju EMD kot osnovno razdaljo izberemo hiearhično razdaljo (angl. *hierarchical distance*) ali enakomerno razdaljo (angl. *equal distance*). Pri enakomerni razdalji se predpostavlja, da so vse vrednosti kategoričnih atributov enako oddaljene med seboj, kar pomeni, da je d enak 1. V tem primeru se EMD razdalja izračuna po izrazu

$$EMD(P,Q) = \frac{1}{2} \sum_{i=1}^{r} |p_i - q_i|. \tag{2.5}$$

3 Programska orodja

Za potrebe anonimizacije podatkov se uporabljajo različna programska orodja. Splet ponuja množico odprtokodnih orodij, ki se sprotoma nadgrajujejo in so enostavna za uporabo. Uporabnikom in raziskovalcem omogočajo uporabo in ovrednotenje obstoječih metod ali razvoj novih metod za anonimizacijo. Poleg tega imajo uporabniki nadzor nad postopki izvajanja metod anonimizacije. Večina orodij ne ponuja vseh metod za anonimizacijo, načeloma pa so vsa orodja zmožna prikrivanja in posploševanja podatkov. V tabeli 3.1 primerjamo različna odprtokodna programska orodja za anonimizacijo podatkov. Predzadnja vrstica nam pove, v kakšnem formatu lahko uvozimo bazo podatkov, zadnji stoplec pa v kakšnem formatu lahko izvozimo bazo podatkov.

3.1 ARX

Za potrebe vzorčne analize v pričujoči diplomski nalogi smo izbrali orodje Data Anonymization Tool ARX. ARX je odprtokodno orodje za izvajanje anonimizacijskih metod na podatkih [20]. Ima svoj grafični uporabniški vmesnik.
Orodje ARX ima vgrajeno funkcionalnost za uvoz podatkov iz relacijskih baz podatkov

Orodje / Metode	ARX	TIAMAT	UTD Toolbox	PARAT	sdcMicro
k-anonimnost	DA	DA	DA	DA	DA
l-raznolikost	DA	DA	DA	NE	DA
t-podobnost	DA	DA	DA	NE	NE
Format uvoza	CSV, Excel(XLSX), Database(JDBC)	SQL	ASCII, TXT	CSV, JDBC	.r
Format izvoza	CSV	/	TXT	CSV	.r

Tabela 3.1 Primerjava orodij za anonimizacijo podatkov.

(MS SQL, DB2, SQLite, MySQL), MS Excel-a in CSV datotek (vse oblike delujejo s samodejnim zaznavanjem). Podpira različne vrste podatkov kot so nizi, datumi, cela števila in realna števila. Vrste podatkov in oblike spremenljivk se samodejno zaznajo med uvozom podatkov. Mankajoče ali nepravilne podatke pri uvozu detektira in jih samodejno popravi oz. prikrije (ali odstrani) iz tabele. ARX je zelo prilagodljiv in omogoča obdelavo zelo velikih podatkovnih baz (več milijonov podatkovnih zapisov) [21].

3.2 Opis dela z ARX

Največji izziv pri anonimizaciji podatkov je doseči ustrezno ravnovesje med informacijsko koristnostjo podatkov in zasebnostjo nosilcev podatkov. ARX uporabnikom omogoča nastavljanje metod, dokler se rezultati ne ujemajo z njihovimi zahtevami. Osnovni koraki uporabe orodja so sledeči:

1. konfiguriranje procesa transformacije;
2. analiza anonimizirane baze;
3. izvoz anonimizirane baze.

V fazi konfiguracije uporabnik uvozi podatke v obliki tabele. Orodje nam pokaže uvoženo bazo na levi strani grafičnega vmesnika. Uporabnik določi tipe atributov za relacijsko tabelo. V glavi tabele so tipi atributov prikazani z različnimi barvami:

- rdeča označuje *identifikator*; identifikatorji so tisti atributi, ki neposredno identificirajo posameznika, zato se jih pri anonimizaciji odstrani iz tabele; tipični primeri so npr. imena ali številke socialnega zavarovanja;
- rumena označuje *kvazi identifikator*; kvazi identifikatorji se navezujejo na osebne podatke, zato jih preoblikujemo z metodami za anonimizacijo; tipični primeri so spol, datum rojstva in poštne številke;
- vijolična označuje *občutljiv atribut*; občutljivi atributi kodirajo lastnosti, s katerimi posamezniki ne želijo biti povezani; če bi bili razkriti, lahko povzročijo škodo posameznikom, na katere se podatki nanašajo; tipični primeri so zdravstvene diagnoze;
- zelena označuje *neobčutljiv atribut*; neobčutljivi atributi ne razkrivajo ničesar pomembnega in bodo ostali nespremenjeni, saj nad njimi ne izvajamo anonimizacijskih metod;

Za vsak atribut ARX specificira tip vrednosti (niz, datum, celo število, realno število). Uporabnik nato izbira anonimizacijske metode. Če vemo, da bomo posploševali določene atribute, se je najbolje pripraviti vnaprej. Pri posloševanju vrednosti kvazi identifikatorjov si sami nastavimo intervale, ki se nam zdijo smiselni. ARX ponuja izbor metrik, ki so uporabnikom koristne za analizo. Uporabnik si izbere metriko in njeno vrednost dobi izpisano po končani anonimizaciji v spodnjem delu grafičnega vmesnika. Po vseh želenih nastavitvah zaženemo anonimizacijo. Na desni strani grafičnega uporabniškega vmesnika se prikaže anonimizirana tabela, medtem ko je na levi strani originalna tabela. V fazi analize primerjamo anonimizirano tabelo z originalno tabelo. Pri analizi uporabljamo različne metrike. Na koncu imamo možnost izvoza anonimizirane tabele v formatu CSV datoteke.

3.3 Metrike za oceno anonimizacije

Za oceno uspešnosti anonimizacijskega algoritma Flash, ki ga uporablja ARX, smo izbrali tri metrike za merjenje kvalitete anonimnih podatkov:

- informacijska izguba,
- razločljivost,
- normalizirana povprečna velikost ekvivalenčnega razreda.

S prvo metriko merimo količino informacij, ki smo jih izgubili v anonimizacijskem procesu. Metrika za merjenje izgube informacij LM (angl. *general information loss metric*), podrobneje opisana v [22–24], nam bo pripomogla pri analizi eksperimentov v nadaljevanju diplomske naloge. Rezultat metrike je med 0 in 1. Rezultat 0 bi pomenil, da nismo izgubili informacij (originalni podatki), medtem ko 1 pomeni popolno anonimizacijo podatkov.

Metrika za razločljivost DM (angl. *discernibility metric*) [25] meri, kako nerazpoznaven je podatkovni zapis od drugih, tako da vsakemu zapisu dodeli kazen, ki je enaka velikosti ekvivalenčnega razreda v katerega spada. Če ima podatkovni zapis vse vrednosti prikrite, mu je dodeljena kazen enaka velikosti celotne tabele. Metriko lahko matematično navedemo v naslednji obliki

$$DM(T^*) = \sum_{|EQ| \geq k} |EQ|^2 + \sum_{|EQ| < k} |T| * |EQ|, \tag{3.1}$$

kjer je T prvotna tabela, $|T|$ je število podatkovnih zapisov in $|EQ|$ je velikost ekvivalenčnih razredov ustvarjenih po anonimizaciji. Ideja te metrike je, da večji ekvivalenčni razredi predstavljajo več informacij o izgubi, zato so zaželene nižje vrednosti za to metriko. Za ilustracijo delovanja metrike bomo izračunali DM tabele 2.6. Anonimizirana tabela 2.6 ima tri ekvivalenčne razrede velikosti 4, zato je rezultat metrike DM(2.6) enak $4^2+4^2+4^2 = 48$.

Normalizirana povprečna velikost ekvivalenčnega razreda (angl. *normalized average equivalence class size metric*) [23] je po ideji podobna zgornji metriki. Vrednost 1 bi nakazala idealno anonimizacijo, pri kateri je velikost ekvivalenčnega razreda enaka določeni vrednosti k. Normalizirano povprečno velikost ekvivalenčnega razreda (C_{avg}) za anonimizirano tabelo T^* podaja enačba

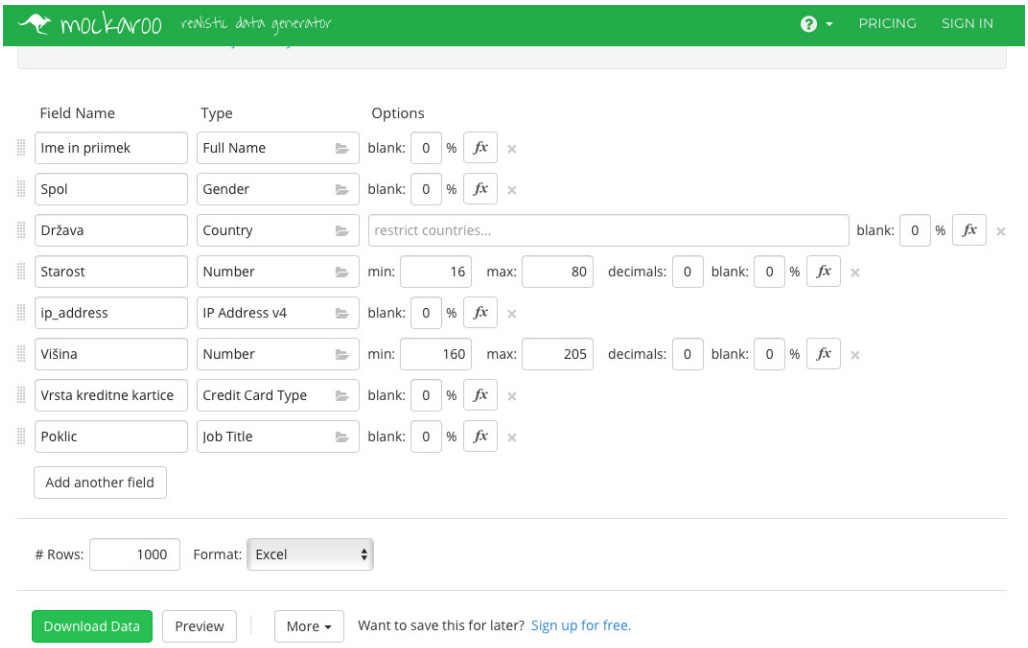
$$C_{avg}(T^*) = \frac{|T|}{|EQs| * k}, \tag{3.2}$$

kjer je T prvotna tabela, $|T|$ je število podatkovnih zapisov, $|EQs|$ je skupno število razvitih ekvivalenčnih razredov in k je določena vrednost pri k-anonimizaciji. Povprečna velikost ekvivalenčnega razreda tabele 2.6 je enaka 4, normalizirana povprečna velikost pa je enaka 12/(3*4) = 1.

Za dodatno metriko smo izbrali čas izvajanja anonimizacije. S to metriko merimo hitrost algoritma pri izvajanju različnih anonimizacijskih metod.

3.4 Generiranje umetne baze

Za potrebe diplomske naloge zaradi varovanja osebnih podatkov nismo mogli pridobiti podatkovne baze iz realnega sveta, zato smo delali z umetno. Umetno podatkovno bazo oseb smo generirali s spletne strani https://mockaroo.com. Spletna stran, prikazana na sliki 3.1, nam omogoča naključno generiranje do tisoč podatkovnih zapisov. Število stolpcev oz. atributov določimo sami in jih poimenujemo. Vsakemu atributu določimo tip podatkov za generiranje.



Slika 3.1 Spletna stran za generiranje podatkovne baze.

3.5 Eksperimenti anonimizacije

Z dvema eksperimentoma si bomo ogledali potek izvajanja anonimizacije na manjših umetno generiranih podatkovnih bazah. Z metrikami za merjenje kvalitete anonimnih podatkov bomo analizirali uspešnost izvedene anonimizacije.

Za začetek generiramo populacijo delovnih ljudi, starih med 16 in 80 let, različnih narodnosti. Atributi za generirano bazo so predstavljeni v tabeli 3.2.

Atribut	**Vrsta**
Ime in priimek	identifikator
Spol	kvazi identifikator
Starost	kvazi identifikator
Država	kvazi identifikator
IPv4 naslov	kvazi identifikator
Višina	kvazi identifikator
Vrsta kreditne kartice	občutljiv
Poklic	občutljiv

Tabela 3.2 Tipizacija atributov uporabljenih za eksperimente.

3.5.1 Eksperiment I

Pri prvem eksperimentu smo se omejili na tri kvazi identifikatorje in testirali parameter k pri k-anonimizaciji. Ogledali smo si razlike med manjšim in večjim parametrom k. Velikost baze je tisoč podatkovnih zapisov. Atributi v tem primeru so `Ime in priimek`, `Spol`, `Država`, `Starost` in `Poklic`.

Vrednosti kvazi identifikatorjev smo med anonimizacijo posplošili. Ena od oblik posploševanja je prikrivanje podatkov, kar smo uporabili pri atributu `Spol`, kjer smo vrednost Moški/Ženska zamenjali z znakom '*'. Za atributa `Država` in `Starost` smo uporabili velikostne razrede. Naša baza je generirala 131 različnih držav sveta, ki smo jih združevali po celinah. Afriške države so imele poslošeno vrednost 'Afrika', evropske države 'Evropa', enako pa velja tudi za ostale države sveta. Za atribut `Starost` smo imeli starostne intervale petih, desetih in dvajsetih let. Da bi programsko orodje ARX posploševalo po zgoraj opisanem postopku, smo morali za vsak kvazi identifikator posebej ustvariti CSV datoteko. V prvem stolpcu CSV datoteke je originalna vrednost atributa, v drugem stolpcu pa njena posplošena vrednost. Ustvarjene CSV datoteke uvozimo v ARX, zato da bi algoritem pri anonimizaciji kvazi identifikatorje posploševal po naših željah. Na ta način algoritmu določimo pogoje, ki jih mora upoštevati. Omenimo še, da pri anonimizaciji identifikatorji niso sodelovali oz. je njihova vrednost prikrita in da ne vplivajo na vrednosti metrik.

Rezultati anonimizacije prvega eksperimenta so predstavljeni v tabeli 3.3. Opazimo, da imamo z večanjem parametra k večjo izgubo informacij. Pri majhnem k algoritem ustvari

k	**LM**	C_{avg}	**št. razredov**	**DM**
3	0,366	5,05	66	28.690
4	0,375	4,03	62	40.654
5	0,381	3,33	60	48.622
10	0,427	3,45	29	105.551
15	0,443	3,70	18	116.427
20	0,478	3,33	15	171.381
50	0,616	2,86	7	386.088
100	0,719	2,00	5	550.563

Tabela 3.3 Eksperiment 1.

veliko ekvivalenčnih razredov vendar majhne velikosti. Z večjim k število razredov upada, vendar so večje velikosti, posledično pa to pomeni da bolj posplošuje atribute, zato da zadovolji kriterije za k-anonimnost. Pri $k = \{3, 4, 5\}$ je s posploševanjem osebe grupiral po petletnih intervalih, pri $k = \{10, 15\}$ je osebe grupiral po desetletnih intervalih, medtem ko je za $k = \{20, 50, 100\}$ osebe grupiral po dvajsetletnih intervalih. Z večanjem starostnega intervala je več oseb vključil v posamezen ekvivalenčni razred, s tem pa je zmanjšal razlike med njimi. Zato so temu primerne tudi vrednosti DM. Torej velikostni razredi atributov tako kot velikost ekvivalenčnega razreda vplivajo na vrednosti metrike LM.

Padajoče vrednosti pri metriki C_{avg} kažejo, da se povprečna velikost ustvarjenih ekvivalenčnih razredov približa idealnemu scenariju, pri čemer je velikost razreda enaka določenemu parametru k. Pri $k = 100$ je samo 5 ekvivalenčnih razredov s povprečno 200 podatkovnimi zapisi. Če bi bilo 10 razredov s povprečno 100 podatkovnimi zapisi, potem bi imeli idealno anonimizacijo, ki je v praksi redkost.

3.5.2 Eksperiment II

Drugi eksperiment temelji na podatkovni bazi z deset tisoč podatkovnimi zapisi. Deset tisoč zapisov smo dobili tako, da smo desetkrat generirali tabelo s tisoč zapisi in nato tabele zlepili skupaj. Pri eksperimentu smo postopoma zviševali število kvazi identifikatorjev in si ogledali razlike med metodami k-anonimnost, l-raznolikost in t-podobnost. Parametri za k, l in t so fiksni. V tem eksperimentu sta k in l enaka 5, parameter t pa

je enak 0,2. Pri merjenju t-podobnosti smo uporabili EMD z enakomerno razdaljo, ker je občutljiv atribut kategoričen. Za anonimizacije bomo uporabljali vse kvazi identifikatorje predstavljene v tabeli 3.2. Občutljiv atribut v eksperimentu bo `Vrsta kreditne kartice`.

Za lažjo primerjavo smo izračunali metrike pri vseh treh metodah in rezultate za vsako metodo predstavili v tabelah 3.4, 3.5 in 3.6. V tabelah kratica QID predstavlja kvazi identifikator. Prva meritev je imela dva kvazi identifikatorja in občutljiv atribut. Nato smo za vsak naslednjo meritev priključili dodaten kvazi identifikator.

št. QID	LM	C_{avg}	št. razredov	DM
2	0,123	29,41	68	2.962.641
3	0,149	7,75	258	3.914.282
4	0,246	4,81	416	9.877.949
5	0,370	4,81	416	9.877.949

Tabela 3.4 Eksperiment 2. Rezultati za k=5 pri različnem številu kvazi identifkatorjev.

št. QID	LM	C_{avg}	št. razredov	DM
2	0,124	30,30	66	3.062.591
3	0,178	9,66	207	7.052.260
4	0,294	8,26	242	11.090.115
5	0,411	8,26	242	11.090.115

Tabela 3.5 Eksperiment 2. Rezultati za k=5 in l=5 pri različnem številu kvazi identifkatorjev.

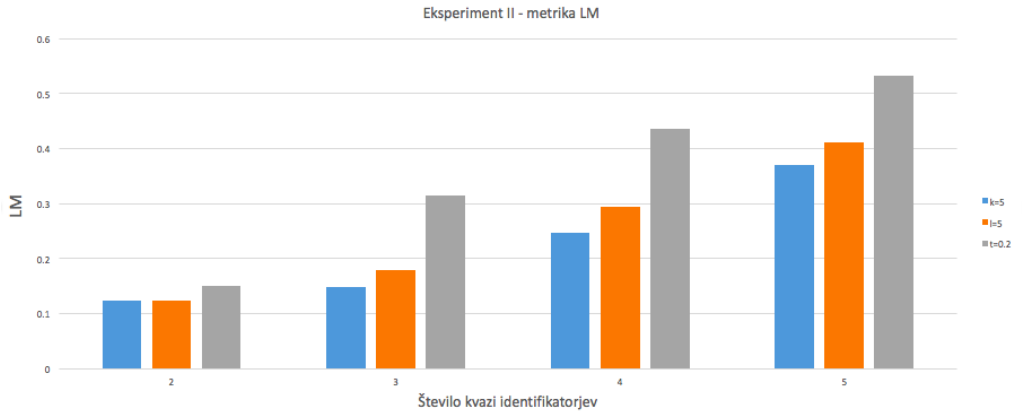
št. QID	LM	C_{avg}	št. razredov	DM
2	0,150	55,55	36	5.242.445
3	0,315	32,79	61	20.685.596
4	0,436	30,77	65	12.789.001
5	0,533	30,77	65	12.789.001

Tabela 3.6 Eksperiment 2. Rezultati za k=5 in t=0.2 pri različnem številu kvazi identifkatorjev.

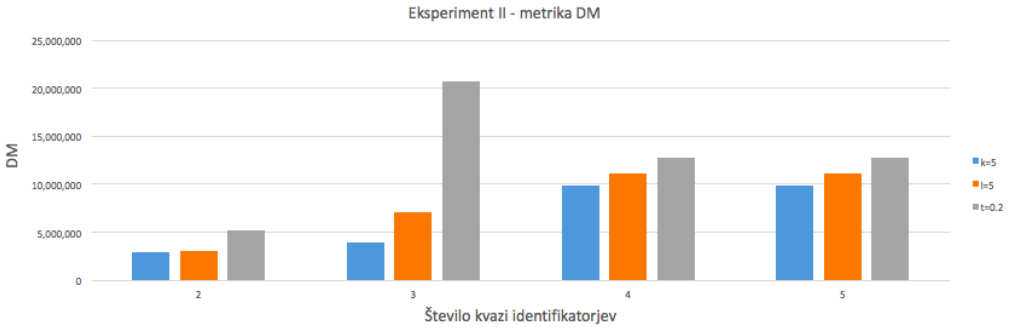
V vseh treh tabelah je opazen trend, da z večanjem števila atributov prihaja do večje izgube informacij. Graf na sliki 3.2 nakazuje, da je izguba informacij neposredno sorazmerna s številom atributov, ki jih je treba anonimizirati. Tudi med metodami so manjše razlike v izgubi informacij. Če primerjamo metode, kjer imajo vse enako število kvazi identifikatorjev vidimo, da ima t-podobnost večjo izgubo informacij kot l-raznolikost, l-raznolikost pa večjo kot k-anonimnost. Iz tega sklepamo, da v tem eksperimentu dobimo najbolj uporabne anonimne podatke z metodo k-anonimnost.

Graf na sliki 3.3 predstavi rezultate za kvaliteto podatkov glede na metriko DM. Pri metodah se z večanjem števila kvazi identifikatorjev povečuje vrednost metrike DM. Pri štirih in petih kvazi identifikatorjih imamo enako število razredov in enako vrednost DM metrike, a različno LM vrednost. Razlog je v tem, da smo z atributom `Spol` pridobili nove informacije, ki smo jih s posploševanjem izgubili. Vseeno se je atribut lepo vklopil v razrede, tako da se je število razredov ohranilo. Ker so bile vse posplošene vrednosti novega atributa enake, so bili podatkovni zapisi enako nerazpoznavni kot v primeru, ko novega atributa ni bilo zraven. Zato je tudi vrednost DM metrike pri štirih in petih kvazi identifikatorjih enaka pri vseh metodah.

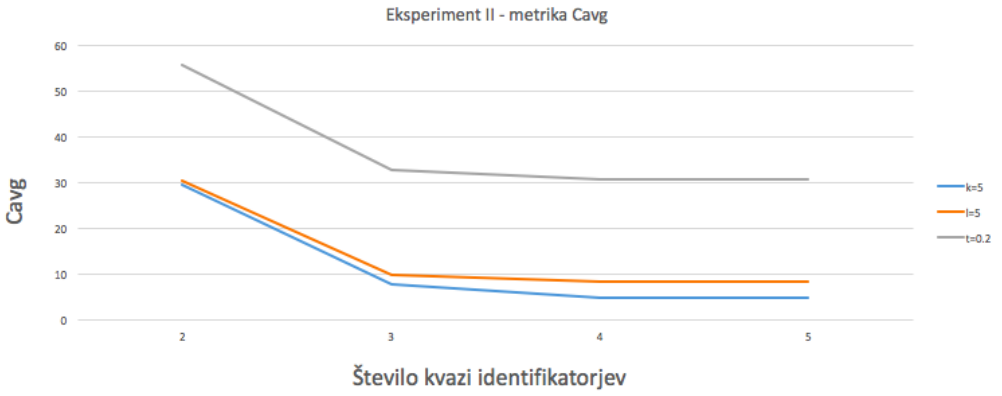
Graf na sliki 3.4 prikazuje vrednost metrike C_{avg}, kjer z večanjem števila kvazi identifikatorjev vrednost metrike upada. Najbolj se idealni anonimizaciji približa metoda k-anonimnost.



Slika 3.2 Izguba informacij pri različnem številu kvazi identifikatorjev.



Slika 3.3 Metrika za razločljivost pri različnem številu kvazi identifikatorjev.



Slika 3.4 Normalizirana povprečna velikost ekvivalenčnega razreda pri različnem številu kvazi identifikatorjev.

3.6 Umetna podatkovna baza slovenskih pacientov

Za glavni del diplomske naloge smo generirali podatkovno bazo z natanko dvemi milijoni podatkovnih zapisov. Ker ni bilo možno generirati tabele z dvemi milijoni podatkovnih zapisov, smo generirali več manjših tabel in jih nato združili skupaj. Posamezen zapis predstavlja osebne in zdravstvene podatke pacienta v Republiki Sloveniji. Vsakemu pacientu smo dodelili enoličen identifikator ID, težo v kilogramih, spol, datum rojstva in količino holesterola v krvi. Atributi baze so sledeči:

- ID pacienta;
- teža (kg);
- spol;
- datum rojstva;
- holesterol (mmol/L).

Za celotno populacijo pacientov smo poskušali zgenerirati podatke, ki se čim bolj prilegajo realnim podatkom. Zato smo si pri generiranju podatkov za bazo pomagali s statistiko, ki je objavljena na spletni strani statističnega urada Republike Slovenije. Izvedeli smo, da je 49,67% prebivalcev moških in 50,33% žensk [26]. Poleg tega smo izvedeli, da je 15% prebivalcev mlajših od 15 let, 65,9% prebivalcev je starih med 15 in 64 let ter 19,1% je starejših od 64 let [27]. Odločili smo se, da bomo najprej generirali podatke za mlajšo generacijo, nato za odrasle ljudi in na koncu za starejše prebivalce. Prva generirana tabela je imela 300.000 zapisov (15%), druga tabela je imela 1.318.000 zapisov (65,9%) in tretja 382.000 (19,1%). Poleg datuma rojstva smo za vsako tabelo generirali id, spol, težo primerno letom in količino holesterola v krvi. Vsaka tabela je bila ustvarjena v csv formatu. Z ukazno vrstico smo združili vse tri tabele in ohranili format csv.

Združena tabela ima natanko dva milijona podatkovnih zapisov in predstavlja končno bazo. Zgenerirana baza je podobna realni. Vsak id pacienta je enak številu vrstice v tabeli. Dobili smo 999913 zapisov, ki predstavljajo moške in 1000087 zapisov za ženske. Okrog 57% vseh zapisov ima visoke vrednosti holesterola v krvi, preostali zapisi pa imajo priporočljive vrednosti holesterola v krvi.

4 Vzorčna analiza

Bazo slovenskih pacientov iz razdelka 3.6 uvozimo v ARX-ov grafični vmesnik. Paciente ločimo med seboj po identifikatorju `ID`. Atributi `Spol`, `Datum rojstva` in `Teža` predstavljajo kvazi identifikatorje, `Holesterol` pa občutljiv atribut. Po označitvi tipa atributov izberemo anonimizacijsko metodo in nastavimo parameter. Pred anonimizacijo določimo način posploševanja kvazi identifikatorjev. Vse je odvisno od tega, kako želimo paciente grupirati, da bi zagotovili čim večjo zasebnost. Vrednosti atributa `Spol` smo prikrili, pri atributu `Datum rojstva` pa smo rojstne datume zamenjali z letnico rojstva. Za atribut `Teža` smo uporabili velikostni interval petih kilogramov. V nadaljevanju so prikazane opravljene meritve po anonimizaciji.

4.1 Izbira metod

Odločili smo se testirati in narediti primerjavo metod k-anonimnosti, l-raznolikosti in t-podobnosti. Najprej smo na bazi podatkov izvajali k-anonimnost. Parameter k zajema različne vrednosti. Rezultati k-anonimizacije so prikazani v tabeli 4.1.

k	**LM**	**št. razredov**	**povprečna velikost razreda**	C_{avg}
10	0,3468	1366	1464	146,41
100	0,3474	1336	1497	14,97
200	0,3494	1285	1551	7,78
300	0,3498	1284	1557	5,19
400	0,3531	1254	1594	3,99
500	0,3547	1243	1609	3,22

Tabela 4.1 Rezultati k-anonimizacije.

Sledilo je izvajanje metode l-raznolikosti. Parameter l je lahko manjši ali enak k in obvezno večji od 1. Za naše testiranje smo se odločili, da bo vsak parameter l enak parametru k. Rezultati za metodo l-raznolikosti so prikazani v tabeli 4.2.

k, l	**LM**	**št. razredov**	**povprečna velikost razreda**	C_{avg}
10	0,3468	1366	1464	146,41
100	0,3477	1330	1503	15,04
200	0,3513	1274	1569	7,85
300	0,3551	1243	1609	5,36
400	0,3624	1211	1651	4,13
500	0,4510	1071	1867	3,73

Tabela 4.2 Rezultati l-raznolikosti.

Nazadnje je potekalo izvajanje metode t-podobnost. Parameter t je enak 0,2. Pri računanju porazdelitve smo uporabili EMD enačbo za numerične atribute, saj je občutljiv atribut `holesterol` numeričen. Rezultati za metodo t-podobnosti so prikazani v tabeli 4.3. Časi izvajanja treh metod so prikazani v tabeli 4.4.

k	**LM**	**št. razredov**	**povprečna velikost razreda**	C_{avg}
10	0,3468	1366	1464	146,41
100	0,3474	1336	1497	14,97
200	0,3494	1285	1551	7,78
300	0,3498	1284	1557	5,19
400	0,3531	1254	1594	3,99
500	0,3547	1243	1609	3,22

Tabela 4.3 Rezultati t-podobnosti.

V nadaljevanju bomo podrobneje analizirali izvajanje uporabljenih metod za vsako metriko.

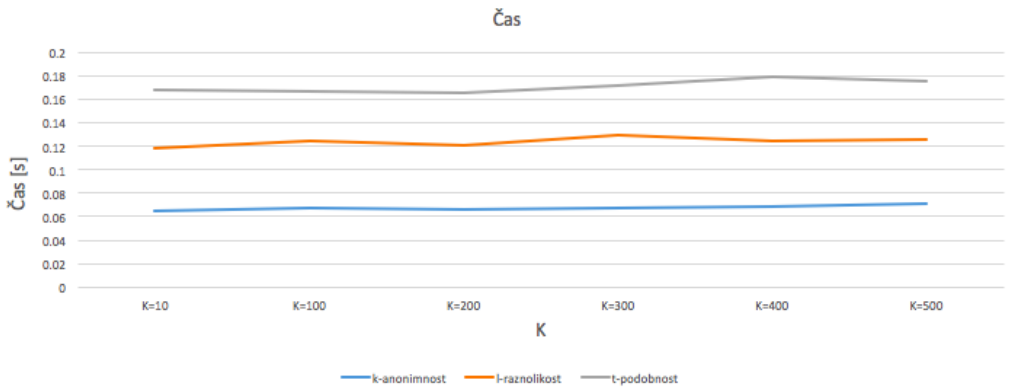
4.2 Analiza metrike časa

Vse tri metode so se izvedle v manj kot eni sekundi. Iz slike 4.1 je razvidno, da je metoda k-anonimnosti hitrejša od ostalih dveh metod. Metoda l-raznolikosti je nekoliko počasnejša, vseeno pa je metoda t-podobnosti najpočasnejša. Razlog je v tem, da metoda t-podobnost izgubi ogromno časa na računanju porazdelitev občutljivega atributa v celotni podatkovni tabeli.

Čas izvajanja posamezne metode je izmerilo orodje ARX. Takoj po končanem izvajanju anonimizacijske metode se je podatek o času izvajanja izpisal v zgornjem desnem kotu grafičnega uporabniškega vmesnika.

k	**k-anonimnost**	**l-raznolikost**	**t-podobnost**
10	0,065s	0,118s	0,168s
100	0,068s	0,124s	0,167s
200	0,066s	0,121s	0,166s
300	0,068s	0,129s	0,172s
400	0,069s	0,124s	0,179s
500	0,071s	0,126s	0,175s

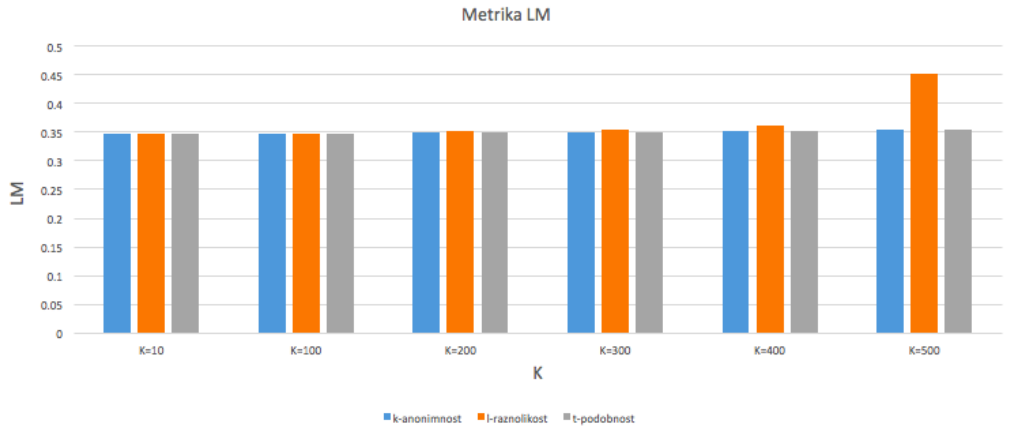
Tabela 4.4 Čas izvajanja metod v sekundah.



Slika 4.1 Čas izvajanja anonimizacijskih metod.

4.3 Analiza metrike LM

Slika 4.2 nakazuje, da imamo z manjšim k bolj kvalitetne anonimne podatke, saj med anonimizacijo izgubimo manj informacij. Med primerjavo metod vidimo, da imamo največjo izgubo informacij prav z l-raznolikostjo, ker pri anonimizaciji upošteva še dodaten pogoj o raznolikosti občutljivega atributa v vsakem ekvivalenčnem razredu.

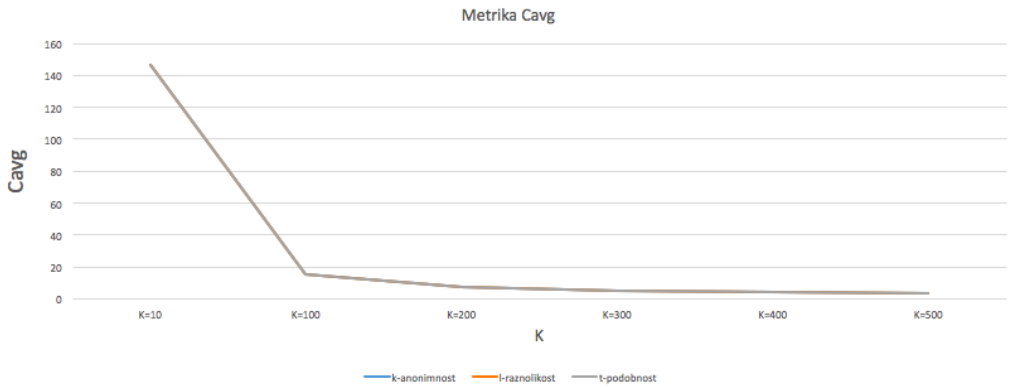


Slika 4.2 Graf za metriko LM.

4.4 Analiza metrike C_{avg}

Z metodami tvorimo med 1000 in 1400 ekvivalenčnih razredov. Manj kot je razredov, večja je povprečna velikost posameznega razreda. Metoda l-raznolikosti tvori večje ekvivalenčne razrede kot ostali dve metodi in ima zato večjo normalizirano povprečno velikost

ekvivalenčnega razreda. S povečevanjem parametra k vrednosti C_{avg} upadajo za vse tri metode (glej graf na sliki 4.3).



Slika 4.3 Graf za metriko C_{avg}.

4.5 Analiza metrike DM

Večja kot je velikost razreda, več zapisov med seboj je nerazločljivih. Za to metriko smo opravili primerjavo samo za k=100. Torej pri 100-anonimnosti smo dobili vrednost 7,53*10^9, za 100-raznolikost vrednost 9*10^9 in za 0,2-podobnost enako vrednost kot pri 100-anonimnosti. Največje razrede tvori metoda l-raznolikosti, zato ima tudi večjo DM vrednost kot ostali dve metodi.

4.6 Ugotovitve

Za anonimizacijo naše baze pacientov odločitev o izbiri primerne metode ni lahka. Časovno se najhitreje izvede metoda k-anonimnosti. Z l-raznolikostjo imamo manjšo kvaliteto anonimiziranih podatkov, a smo dosegli večjo zasebnost podatkov pacientov. Metoda t-podobnost vrne kvalitetnejše anonimne podatke in je v tem primeru, če zanemarimo čas izvajanja metode, najbolj primerna metoda za anonimizacijo naše baze pacientov. Ugotovili smo, da na anonimizacijo podatkovne baze vplivajo različni dejavniki. Če bi imeli pri posploševanju vrednosti kvazi identifikatorjev intervale drugačne velikosti, bi dobili drugačne rezultate metrik. Tudi pri izbiri metode je potrebno dobro razmisliti. Izbira parametra za metodo je pravzaprav odvisna od uporabnikovih zahtev. Pri t-podobnosti vrsta občutljivega atributa vpliva na izbiro primerne EMD razdalje itd.

5 Zaključek

V diplomskem delu smo opisali najbolj razširjene metode, ki se uporabljajo za anonimizacijo podatkov. Opisali smo njihovo delovanje, prednosti in slabosti. Opravili smo pregled programskih orodij, s katerimi lahko izvajamo različne anonimizacijske metode. Iskali smo orodje, ki podpira izvajanje metod k-anonimnosti, l-raznolikosti, t-podobnosti ter je prilagodljivo za delo z bazami. Odločili smo se za ARX, ki ponuja grafični uporabniški vmesnik in je odličen za analizo anonimizacijskih metod. Nato smo opravili dva manjša eksperimenta. Pri obeh eksperimetih smo generirali manjšo bazo posameznikov. Bazi smo določili občutljive atribute in atribute, s katerimi lahko identificiramo posameznika. Pri prvem eksperimentu smo analizirali razlike, ki jih dobimo s spreminjanjem vrednosti parametra k pri k-anonimizaciji. V drugem eksperimentu smo postopoma zviševali število kvazi identifikatorjev in si ogledali razlike pri izvajanju izbranih metod. Za glavni del diplomske naloge smo se odločili anonimizacijo opraviti na primeru s področja zdravstva. Ker ima velik odstotek ljudi v Sloveniji visoke vrednosti holesterola v krvi, smo generirali umetno podatkovno bazo slovenskih pacientov. Vsak pacient je imel v bazi osebne podatke in za občutljiv atribut vrednost holesterola. Izvedli smo vzorčno testi-

ranje s tremi metodami (k-anonimnosti, l-raznolikosti, t-podobnosti) in jih medsebojno primerjali. Za primerjavo metod smo si pomagali z različnimi metrikami. Poskušali smo ugotoviti, katera od izbranih metod je najprimernejša za praktično uporabo in katera bo nudila najboljšo zasebnost podatkov. Prišli smo do ugotovitev, da je izbira metode odvisna od načina posplošitve kvazi identifikatorjev, števila atributov, algoritma itd.

Analizo anonimizacijskih metod bi lahko izboljšali, če bi imeli na voljo več metrik s področja anonimizacije. Prav tako obstajajo številni algoritmi kot so Mondrian, Incognito, Datafly itd., ki izvajajo anonimizacijo in bi v prihodnosti lahko naredili primerjavo le teh algoritmov.

Literatura

[1] Introduction to anonymity, data protection and privacy. Dosegljivo: http://knowledgebasement.com/introduction-to-anonymity-data-protection-and-privacy/. [Dostopano: 15. 11. 2017].

[2] C. Cordess. *Confidentiality and Mental Health*. Jessica Kingsley, 2001.

[3] Arvind Narayanan and Vitaly Shmatikov. How to break anonymity of the Netflix prize dataset. *CoRR*, abs/cs/0610105, 2006.

[4] AOL removes search data on group of web users. Dosegljivo: http://www.nytimes.com/2006/08/08/business/media/08aol.html. [Dostopano: 25. 4. 2018].

[5] Več kot polovica v EU je pretežkih. Dosegljivo: http://www.rtvslo.si/zdravje/novice/vec-kot-polovica-v-eu-ju-je-pretezkih-najvecji-delez-debelih-ima-malta-slovenija-na-7-mestu/405854. [Dostopano: 25. 4. 2018].

[6] Uredimo holesterol. Dosegljivo: http://www.krka.biz/sl/v-skrbi-za-vase-zdravje/v-skrbi-za-vase-zdravje/srce-in-zilje/uredimo-holesterol/1590. [Dostopano: 25. 4. 2018].

[7] Uradni list, zakon o varstvu osebnih podatkov (uradno prečiščeno besedilo) (ZVOP-1-UPB1). Dosegljivo: http://www.uradni-list.si/glasilo-uradni-list-rs/vsebina?urlid=200794&stevilka=4690. [Dostopano: 25. 4. 2018].

[8] Osnutek zakona o varstvu osebnih podatkov (ZVOP-2). Dosegljivo: http://www.iusinfo.si/download/razno/171004_ZVOP-2_status.pdf. [Dostopano: 25. 4. 2018].

[9] Graham Cormode and Divesh Srivastava. Anonymized data: Generation, models, usage. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*, SIGMOD '09, pages 1015–1018, New York, NY, USA, 2009. ACM.

[10] Modeliranje podatkovnih baz, relacijski model. Dosegljivo: http://drenovec.tsckr.si/model/relac.htm. [Dostopano: 15. 11. 2017].

[11] J. Domingo-Ferrer and V. Torra. A critique of k-anonymity and some of its enhancements. In *2008 Third International Conference on Availability, Reliability and Security*, pages 990–993, March 2008.

[12] Latanya Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):571–588, October 2002.

[13] Kato Mivule. Utilizing noise addition for data privacy, an overview. *CoRR*, abs/1309.3958, 2013.

[14] Latanya Sweeney. Simple demographics often identify people uniquely. *Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh*, 2000.

[15] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, SRI International, 1998.

[16] Maruša Stanek, Miroslav Babić, and Aleksandar Jurišić. Anonimizacija podatkovnih baz. *Nilt-ekon organ inform zdrav 2009*, 2(25):53–59, 2009.

[17] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. In *22nd International Conference on Data Engineering (ICDE'06)*, pages 24–24, April 2006.

[18] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering*, pages 106–115, April 2007.

[19] Josep Domingo-Ferrer, David Sánchez, and Jordi Soria-Comas. *Database Anonymization: Privacy Models, Data Utility, and Microaggregation-based Inter-model Connections*, volume 8. Morgan & Claypool Publishers, 01 2016.

[20] ARX data anonymization tool. Dosegljivo: http://arx.deidentifier.org. [Dostopano: 15. 11. 2017].

[21] Overview of ARX data anonymization tool. Dosegljivo: http://arx.deidentifier.org/overview/. [Dostopano: 15. 11. 2017].

[22] Xiaokui Xiao and Yufei Tao. Personalized privacy preservation. In *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, SIGMOD '06, pages 229–240, New York, NY, USA, 2006. ACM.

[23] Vanessa Ayala-Rivera, Patrick McDonagh, Thomas Cerqueus, and Liam Murphy. A systematic comparison and evaluation of k-anonymization algorithms for practitioners. *Trans. Data Privacy*, 7(3):337–370, December 2014.

[24] Vijay S. Iyengar. Transforming data to satisfy privacy constraints. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 279–288, New York, NY, USA, 2002. ACM.

[25] Roberto J. Bayardo and Rakesh Agrawal. Data privacy through optimal k-anonymization. In *Proceedings of the 21st International Conference on Data Engineering*, ICDE '05, pages 217–228, Washington, DC, USA, 2005. IEEE Computer Society.

[26] Prebivalci po skupinah in spolu. Dosegljivo: http://www.stat.si/StatWeb/News/Index/6619. [Dostopano: 20. 4. 2018].

[27] Sestava prebivalstva. Dosegljivo: http://www.stat.si/StatWeb/Field/Index/17/104. [Dostopano: 20. 4. 2018].