Anonimizacija podataka iz mnoštva prikupljenih uporabom mobilnih platformi

Damjanović, Martina

SVEUČILIŠTE JOSIPA JURJA STROSSMAYERA U OSIJEKU
FAKULTET ELEKTROTEHNIKE, RAČUNARSTVA I INFORMACIJSKIH TEHNOLOGIJA

Sveučilišni studij računarstva

ANONIMIZACIJA PODATAKA IZ MNOŠTVA PRIKUPLJENIH UPORABOM MOBILNIH PLATFORMI

Završni rad

Martina Damjanović

Osijek, 2021.

FERIT
FAKULTET ELEKTROTEHNIKE, RAČUNARSTVA I INFORMACIJSKIH TEHNOLOGIJA OSIJEK

Obrazac Z1P - Obrazac za ocjenu završnog rada na preddiplomskom sveučilišnom studiju

Osijek, 13.09.2021.

Odboru za završne i diplomske ispite

Prijedlog ocjene završnog rada na preddiplomskom sveučilišnom studiju

Ime i prezime studenta:	Martina Damjanović
Studij, smjer:	Preddiplomski sveučilišni studij Računarstvo
Mat. br. studenta, godina upisa:	R4190, 23.07.2018.
OIB studenta:	58035752058
Mentor:	Doc.dr.sc. Zdravko Krpić
Sumentor:	Doc. dr. sc. Bruno Zorić
Sumentor iz tvrtke:	
Naslov završnog rada:	Anonimizacija podataka iz mnoštva prikupljenih uporabom mobilnih platformi
Znanstvena grana rada:	**Obradba informacija (zn. polje računarstvo)**
Predložena ocjena završnog rada:	Izvrstan (5)
Kratko obrazloženje ocjene prema Kriterijima za ocjenjivanje završnih i diplomskih radova:	Primjena znanja stečenih na fakultetu: 3 bod/boda Postignuti rezultati u odnosu na složenost zadatka: 3 bod/boda Jasnoća pismenog izražavanja: 3 bod/boda Razina samostalnosti: 3 razina
Datum prijedloga ocjene mentora:	13.09.2021.
Datum potvrde ocjene Odbora:	22.09.2021.

Potpis mentora za predaju konačne verzije rada u Studentsku službu pri završetku studija:	Potpis:
	Datum: 24.9.2021.

FERIT
FAKULTET ELEKTROTEHNIKE, RAČUNARSTVA I INFORMACIJSKIH TEHNOLOGIJA OSIJEK

IZJAVA O ORIGINALNOSTI RADA

Osijek, 24.09.2021.

Ime i prezime studenta:	Martina Damjanović
Studij:	Preddiplomski sveučilišni studij Računarstvo
Mat. br. studenta, godina upisa:	R4190, 23.07.2018.
Turnitin podudaranje [%]:	3

Ovom izjavom izjavljujem da je rad pod nazivom: **Anonimizacija podataka iz mnoštva prikupljenih uporabom mobilnih platformi**

izrađen pod vodstvom mentora Doc.dr.sc. Zdravko Krpić

i sumentora Doc. dr. sc. Bruno Zorić

moj vlastiti rad i prema mom najboljem znanju ne sadrži prethodno objavljene ili neobjavljene pisane materijale drugih osoba, osim onih koji su izričito priznati navođenjem literature i drugih izvora informacija.
Izjavljujem da je intelektualni sadržaj navedenog rada proizvod mog vlastitog rada, osim u onom dijelu za koji mi je bila potrebna pomoć mentora, sumentora i drugih osoba, a što je izričito navedeno u radu.

Potpis studenta:

M. Damj

SADRŽAJ

1. UVOD

U današnje vrijeme gdje su ljudi okruženi brojnim tehnologijama koje im olakšavaju svakodnevni život, u isto vrijeme ugrožena im je privatnost. Pametni mobilni uređaji postali su svakodnevica, pa čak i starijim ljudima, a nude osim poziva i poruka, korištenje mnogih aplikacija. Aplikacije poput navigacije ili onih koje pronalaze mjesta od interesa u blizini, zahtijevaju trenutnu lokaciju osobe kako bi aplikacija uopće mogla služiti svrsi. Budući da se svaka aplikacija pokreće na nekom uređaju koji ima vlastitu IP adresu te je instalirana putem nečijeg računa elektroničke pošte, dolazi do rizika otkrivanja identiteta te saznanja o dodatnim informacijama osobe povezivanjem podataka iz drugih baza podataka. Nadalje, većina podataka poput medicinskih kartona, u velikoj mjeri se digitaliziraju kako bi se olakšao i ubrzao rad s velikom količinom podataka te kako ne bi bili dostupni samo u papirnatom obliku kod liječnika opće prakse. Prilikom tog postupka nastaje problem jer osjetljivi podaci postoje u digitalnim bazama podataka, kojima određeni broj ljudi ima pristup te su oni potencijalni napadači pri zloupotrebi podataka. Kako do toga ne bi došlo, primjenjuju se tehnike anonimizacije na prikupljenim podacima poput generalizacije, maskiranja, perturbacije i sl. Nakon tog procesa podaci trebaju zadovoljiti određena mjerila privatnosti kao što su *k*-anonimizacija , *l*-raznolikost te t-bliskost kako bi se smatrali dovoljno zaštićenima. Rezultat toga je veća sigurnost i privatnost pojedinca čiji se podaci nalaze u digitalnom obliku.

U ovom radu istražit će se kako mobilne platforme mogu pomoći u prikupljanju podataka te koje se tehnike anonimizacije primjenjuju na te podatke i na koji način. Također, istražit će se koja mjerila anonimizacije podaci moraju zadovoljiti kako bi se očuvala privatnost. Tehnike i mjerila anonimizacije bit će primijenjene na podatke iz javnog izvora u sklopu praktičnog dijela rada.

U drugom poglavlju rada opisana je nabava iz mnoštva, njeni oblici, primjena nabave iz mnoštva na mobilnim platformama u svrhu prikupljanja podataka te primjer postojećih alata za tu namjenu. Tema trećeg poglavlja vezana je uz pojam anonimizacije, tehnike i mjerila anonimizacije i na kraju opis rada postojećih aplikacija za anonimizaciju. Četvrto poglavlje predstavlja praktični dio rada u kojem se javni skup podataka anoninimizira primjenom različitih tehnika te se uspoređuju rezultati primjena. Posljednje poglavlje donosi zaključak i smjernice za budući rad.

1.1. Zadatak završnog rada

U teorijskom dijelu rada potrebno je opisati mogućnosti prikupljanja senzorskih podataka na mobilnim platformama i slučaje korištenja širokog prikupljanja takvih podataka iz mnoštva (engl. *mobile crowdsourcing*). Naglasak staviti na postojeće biblioteke i rješenja koji to omogućuju,

probleme koje donose s aspekta privatnosti te mehanizme anonimizacije podataka. U praktičnom dijelu rada ugraditi programsko rješenje koje omogućuje primjenu nekih od mehanizama anonimizacije opisanih u teorijskom dijelu rada.

2. NABAVA IZ MNOŠTVA

Pojam nabava iz mnoštva (engl. *crowdsourcing*) predstavlja model nabave sredstava u kojem svaki pojedinac daje potrebne usluge putem Interneta (posredstvom informacijske tehnologije). Sredstva su posljedica aktivnosti (na primjer, praćenje kretanja, nuđenje računalne snage, uplata novca, davanje ideja za rješavanje problema i sl.) traženih od strane tvrtke ili pojedinca, a koje se obavljaju koristeći Internet. Na taj način skupina ljudi koja je geografski raštrkana može zajedno doći do rješenja nekog problema, donositi odluke te pomoći tvrtkama da poboljšaju svoje usluge i proizvode. Daren C. Brahbaum pružio je definiciju: „Ja definiram *crowdsourcing* kao mrežno, distribuirano rješavanje problema i model koji, koristeći kolektivnu inteligenciju zajednice na mreži, služi zadanim organizacijskim ciljevima“ [1]. Između ostalog, nabava iz mnoštva predstavlja efektivan način prikupljanja velikog broj podataka na brži i jeftiniji način. Podaci prikupljeni metodom nabave iz mnoštva nazivaju se podaci iz mnoštva (engl. *crowdsourced data*). U zadnje se vrijeme čak i ustanove poput muzeja i knjižnica koriste ovom metodom u svrhu digitalizacije podataka. Najpoznatiji oblik nabave iz mnoštva jest *Wikipedia*, online enciklopedija koju ljudi popunjavaju bez ikakvog financijskog dohotka. Porijeklo naziva *crowdsourcing* potječe iz 2005. godine, kada su ga urednici časopisa *Wired* iskoristili za opis korištenja Interneta od strane kompanija kako bi „radile nabavu putem gomile (engl. *outsource work to the crowd*)“. Naziv je složenica nastala od riječi *crowd* (gomila) i *outsourcing* (izdvajanje posla, korištenje vanjskih poduzeća i pojedinaca za obavljanje pojedinog posla). *Crowdsourcing* se rabi kao metoda za *outsourcing*, no danas se koristi i za prikupljanje podataka putem, primjerice javnih mreža senzora, IoT-a i sl.

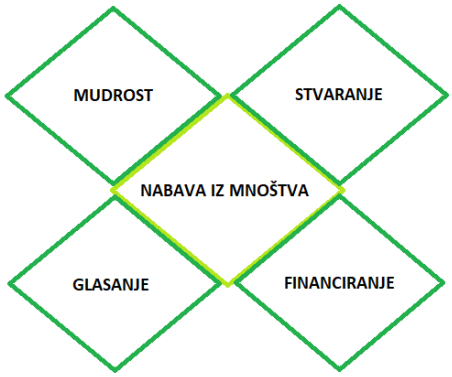
2.1. Oblici nabave iz mnoštva

Osoba koja je i prvi put upotrijebila izraz *crowdsourcing*, Jeff Howe, dijeli ga na 4 različita oblika na osnovu problema koji treba riješiti [1], a prikaz ove podjele dan je slikom 2.1.

Mudrost gomile (engl. *crowdwisdom*) predstavlja uzimanje u obzir mišljenje mnoštva radije nego samo mišljenje jednog stručnjaka. Smatra se da je grupa pametnija nego li pojedinac te da su rezultati u slučaju rješavanja problema, donošenja odluka ili inovacija bolji ukoliko je tome pridonijela grupa ljudi. Primjer toga gdje grupa inteligentno nadjačava pojedinca je televizijska emisija *Who wants to be a Millionaire?* u situaciji kada natjecatelj traži pomoć od publike koja kao cjelina ima više znanja nego li pojedinac koji je ekspert u određenom području.

Stvaralaštvo gomile (engl. *crowdcreation*) je prikupljanje ideja kako bi se kreiralo nešto ili riješilo

neki zadani problem. Ideje dolaze iz mnoštva, najčešće putem Interneta. *LEGO* je tako zatražio od mnoštva da predlože nove ideje za *LEGO* setove igračaka.



Slika 2.1. Oblici nabave iz mnoštva

Nabava glasova iz mnoštva (engl. *crowdvoting*) predstavlja takav oblik nabave iz mnoštva gdje pojedinac svojim glasom pridonosi rješavanju nekog problema. Ono je nešto s čime je čovjek imao susret puno puta samo nije bio svjestan toga. Najbolji primjeri su televizijske emisije poput *American Idol*, odnosno hrvatska adaptacija iste, *Super talent*. Komercijalne web stranice poput *Coca-Cole* i *Heineken* iskoristili su mnoštvo kako bi dobili povratnu informaciju o svojim proizvodima.

Skupno financiranje (engl. *crowdfunfing*) jest financiranje projekta ili pothvata na način da se prikupljaju male svote novca od strane velikog broj ljudi najčešće putem Interneta. Samo prikupljanje novca i transakcije obavljaju se preko specijaliziranih platformi za skupno financiranje. Primjeri takvih platformi su *Kickstarter, GoFundMe, Indiegogo*.

2.2. Mobilne platforme kao izvori podataka iz mnoštva

Mobilni uređaji postali su svakodnevnica ljudi, od onih najmlađih pa do onih najstarijih koji se hvataju u koštac s tehnologijom. Mobiteli su obogaćeni senzorima (na primjer, GPS), ali i komunikacijskim tehnologijama poput *Bluetooth-a* i WiFi-a koji omogućuju prijenos podataka. Kombinacijom mobilnih uređaja i nabave iz mnoštva nastala je mobilna nabava iz mnoštva (engl. *mobile crowdsourcing*, MCS) koja omogućava prihod podataka u stvarnom vremenu.

MCS se obzirom na svojstva zadatka i uključenosti ljudske pomoći dijeli na [2]:

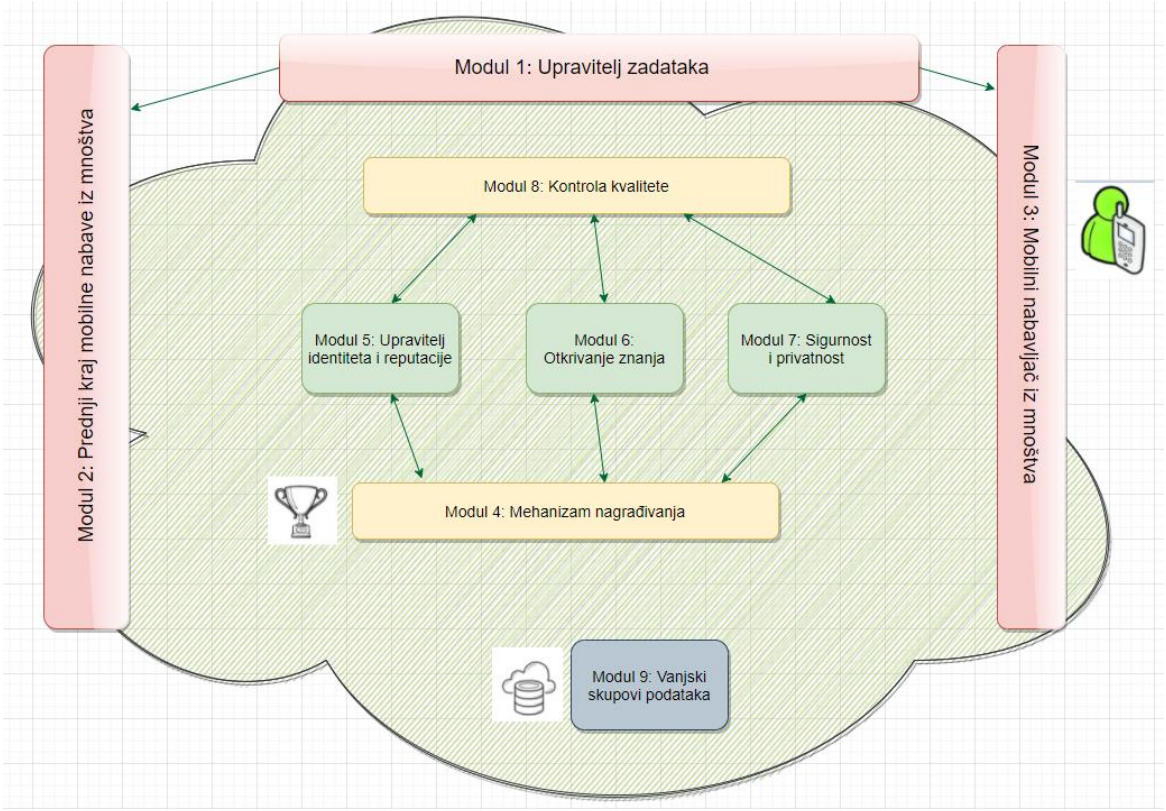
- *Mobile crowd computing* – zadatak se prosljeđuje na mobilne uređaje, a rezultati se prikupljaju raznim mrežama
- *Mobile crowd sensing* – mobilni uređaji upotrebljavaju se kao senzori

- *Human-assisted crowdsourcing* – korištenje ljudske inteligencije u svrhu rješavanja zadanog zadatka

2.2.1. Opis rada MCS aplikacije

MCS sustav sastoji se od tri sudionika: nabavljač iz mnoštva, odnosno onaj koji zadaje zadatke (engl. *crowdsourcers*), radnici iz mnoštva (engl. *crowdworkers*), te platforma za nabavu iz mnoštva (engl. *crowdsourcing platform*). Nabavljač iz mnoštva preko platforme postavlja zadatak, obavlja validaciju dobivenih podataka te nagrađuje radnike. Radnici iz mnoštva putem platforme odlučuju se za sudjelovanje u određenom zadatku i šalju podatke. Međusobna komunikacija između nabavljača iz mnoštva i radnika iz mnoštva ne mora nužno postojati, no ukoliko postoje pitanja vezana uz nagradu i detalje oko izvršavanja zadatka, oni mogu komunicirati na bilo koji način (na primjer telefonski, elektroničkom poštom). Sustav se sastoji i od devet modula [3] [4], prikazanih na slici 2.2. te opisanih u nastavku.

Upravitelj zadataka (engl. *task managment*) odgovoran je za prikupljanje podataka i prikazivanja istih. Prednji kraj (engl. *front-end*) daje radnicima iz mnoštva korisničko sučelje za slanje prikupljenih podataka.



Slika 2.2. Moduli MCS aplikacije, izrađeno prema [3]

Mobilni nabavljač iz mnoštva (engl. *mobile crowdsourcer*) ima zadatak objaviti zadatke (poslove) na platformu na način da komunicira s modulom upravitelja zadataka te isplatiti određene naknade radnicima iz mnoštva te za korištenje platforme za nabavu iz mnoštva. Mehanizam nagrađivanja (engl. *reward mechanism*) koristi se za motivaciju i nagrađivanje radnika iz mnoštva i nabavljača iz mnoštva novcem ili određenim pravima. Također, može i dodatno nagraditi ili kazniti ovisno o dobro, odnosno loše odrađenom poslu. Upravitelj identiteta i reputacije (engl. *identity and reputation managment*) upravlja identitetima radnicima i nabavljačima iz mnoštva te im gradi reputaciju na osnovu njihova prijašnja ponašanja kako bi poboljšali kvalitetu poslanih podataka.

Modul otkrivanja znanja (engl. *knowledge discovery*) obavlja obradu podataka u smislu izvlačenja onih korisnih, obzirom da poslani podaci mogu biti falsificirani ili nestrukturirani. Modul sigurnosti i privatnosti (engl. *security and privacy*) štiti privatnost korisnika te pokušava da se ista ne otkrije. Modul kontrole kvalitete (engl. *quality controler*) na osnovu analiziranja povratne informacije o prikupljenim podacima, prilagođava parametre modula pet, šest i sedam kako bi se postigla kvalitetna usluga. Modul vanjskih poslova podataka (engl. *external datasets*) uključuje vanjske skupove podataka u sustav, koji su otvoreni i postavljeni od strane vlasti ili agencija.

Nadalje, postoji još jedna podjela načina rada MCS aplikacija koja se tiče same arhitekture, a to je podjela na centraliziranu te decentraliziranu arhitekturu [2].

2.2.2. Centralizirana arhitektura mobilne nabave iz mnoštva

Centralizirana arhitektura podrazumijeva obradu svih sakupljenih podataka na centralnom serveru na koji sudionici mogu i slati zahtjeve. Ovakva struktura sastoji se od četiri sloja: sloj mobilnih senzora i prikupljanja (engl. *mobile sensing and gathering layer*), sloj povezanosti i mreže (engl. *connectivity and network layer*), sloj obrade mnoštva (engl. *crowd processing* layer) te sloj krajnjeg korisnika (engl. *end-user layer*).

Zadaća prvog sloja jest generiranje i prosljeđivanje podataka senzora ili zadataka na glavni server bežičnim putem. Sloj povezanosti i mreže osigurava mrežnu povezanost senzorskih uređaja i glavnog servera. Mrežna povezanost može se ostvariti putem Wi-Fi, Bluetooth, 3G/4G/5G. Treći sloj obrađuje i pohranjuje zadatke i podatke. Zadnji sloj predstavlja korisnike koji zahtijevaju uslugu nabave iz mnoštva. Omogućeno im je slanje zadatka te naposljetku uvid u rezultate putem korisničkog sučelja. Zadaci mogu biti prijavljivanje podataka iz stvarnog vremena, davanje povratne informacije o proizvodu ili usluzi, reklamiranje i sl.

2.2.3. Decentralizirana arhitektura mobilne nabave iz mnoštva

Svo računanje i komunikaciju obavlja svaki čvor te su na taj način korisnici uključeni u cijeli proces nabave iz mnoštva. Imaju mogućnost komunicirati s ostalim čvorovima te razmjenjivati podatke. Obzirom da ovdje nema slojeva već je sve na jednakoj razini, svatko može jednako doprinijeti. U isto vrijeme to može biti i velika mana iz razloga što je nužno da svaki čvor surađuje.

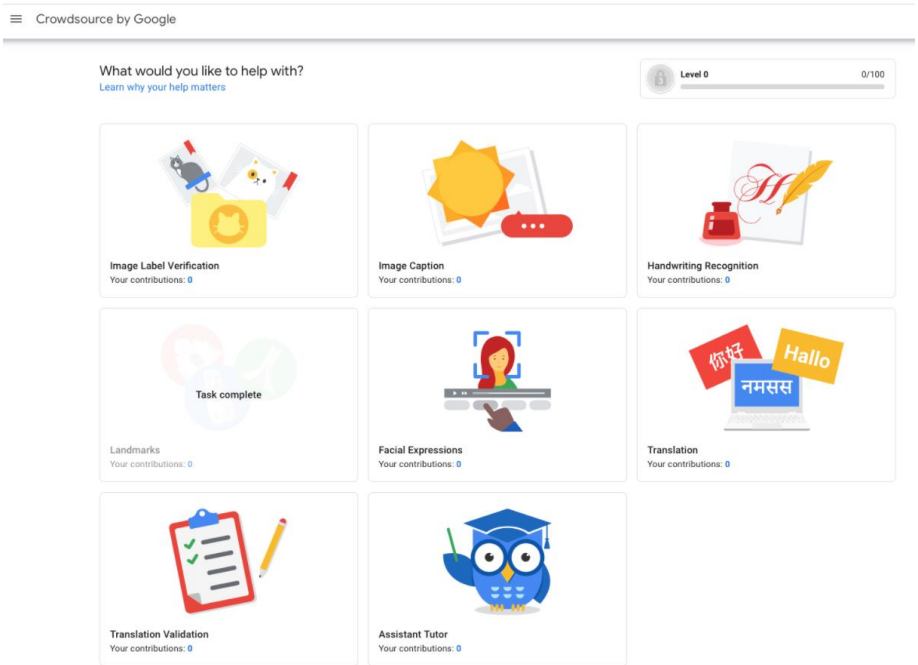
2.3. Postojeći alati za prikupljanje podataka

U ovom potpoglavlju opisan je rad nekoliko aplikacija koje koriste princip rada nabave iz mnoštva te njihove mogućnosti.

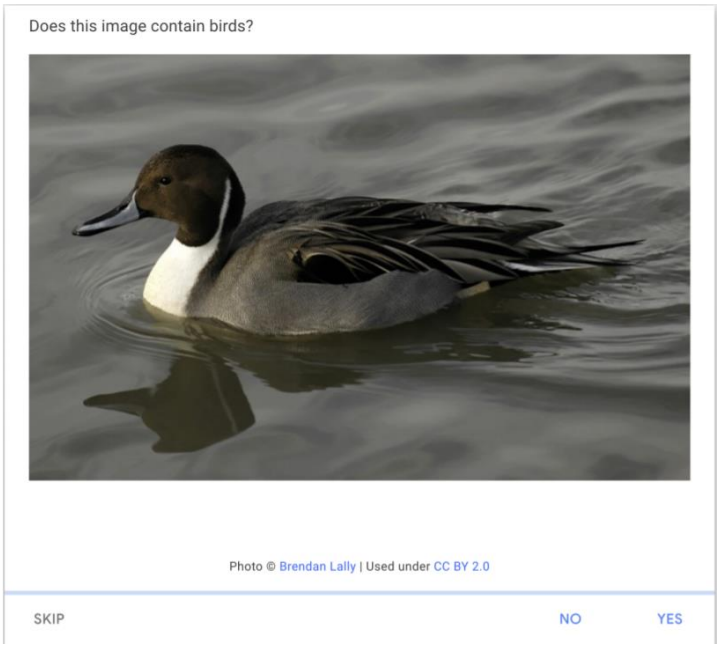
2.3.1. Crowdsource by Google

Crowdsource by Google je web i mobilna aplikacija kreirana 2016. godine, koja prikuplja podatke rješavanjem zadataka iz različitih kategorija u svrhu strojnog učenja te stvaranja proizvoda umjetne inteligencije (engl. *Artificial Intelligence*, AI).

Na samome početku otvara se početna stranica, čiji je izgled dan slikom 2.3. na kojoj korisnik odabire kategoriju zadatka. Jedna od njih je i potvrđivanje oznaka za slike (engl. *image label verification*) gdje je potrebno odgovoriti klikom na *Da* ili *Ne* ovisno o tome nalazi li se neki objekt na slici ili ne. Primjer takvog zadataka dan je slikom 2.4. Ova aplikacija ima i sustav nagrađivanja koji uključuje razine (engl. *levels*), značke (engl. *badges*) i ploču sa postignutim rezultatima (engl.



Slika 2.3. Početna stranica mrežne aplikacije



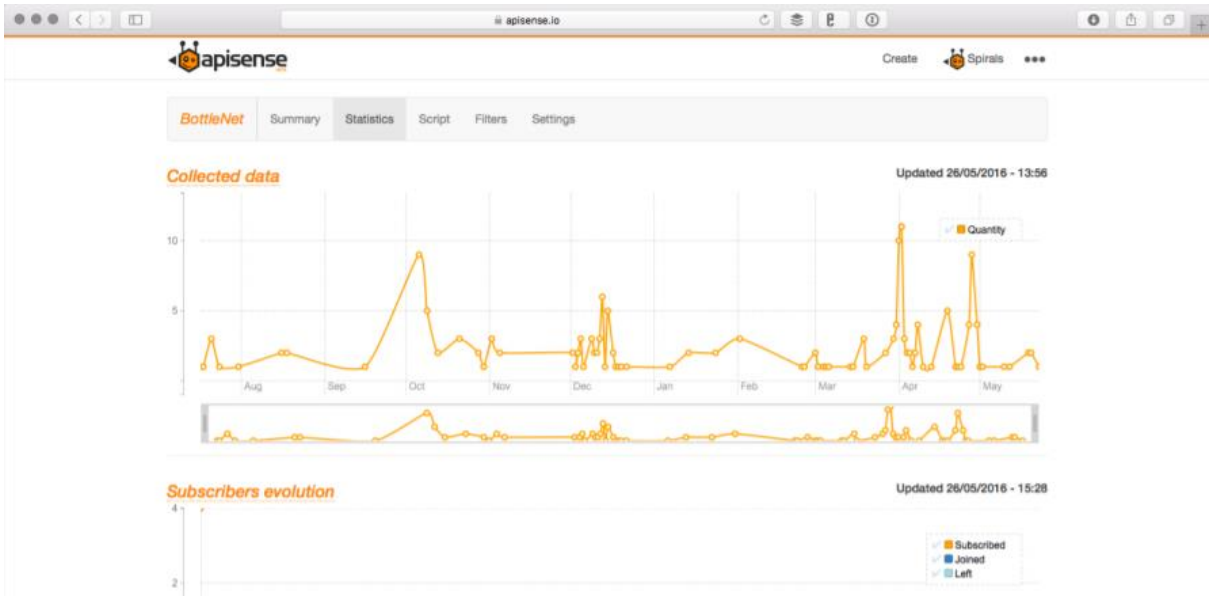
Slika 2.4. Potvrđivanje oznake za slike

leaderboard). Što više doprinosa korisnik napravi, odnosno što više zadataka riješi, time prelazi na višu razinu i dobiva sve više znački. U trenutku kada bude uvršten na ploču s rezultatima može vidjeti svoje postignuće u odnosu na ostale te može sudjelovati u prijateljskim natjecanjima s ostalim sudionicima. Postoji i zadatak koji uključuje da korisnik svojom kamerom na mobilnom uređaju uslika sliku i učita ju. Kako bi se zaštitila privatnost korisnika i njegove lokacije, sve informacije vezane uz lokaciju fotografiranja se brišu, jedino se koristi IP adresa vezana uz zemlju iz koje fotografija dolazi. Također, korisnik u bilo kojem trenutku može sliku izbrisati iz baze podataka [5].

2.3.2. APISENSE

APISENSE je platforma pomoću koje se prikupljaju podaci mnoštva putem mobilnih senzora. Posebno je kreiran kako bi pomogao znanstvenicima u primjeni eksperimenata koji uključuju senzore. Potrebno je napisati skriptu (za prikupljanje podataka) koja će se izvoditi na mobilnim uređajima sudionika putem mobilne aplikacije *Bee*. *APISENSE* biblioteka produžetak je *JavaScript, CoffeeScript* i *Python* programskih jezika koja omogućava opis zadatka bez potrebe za znanjem razvoja mobilnih programskih tehnologija poput *Android SDK* [6]. U svrhu praćenja, kontroliranja i ažuriranja projekta u stvarnom vremenu postoji mrežno sučelje te ono još omogućava i vizualizaciju prikupljenih podataka nakon što se prenesu na poslužitelj, a primjer toga prikazan je na slici 2.5. Kako bi osoba postala sudionik u određenim eksperimentima, nakon instalacije i registracije, iste može pronaći u trgovini eksperimenata (engl. *experiment store*).

Sudionici mogu urediti postavke privatnosti kako bi odredili uvjete pod kojima su voljni podijeliti svoje podatke. Podaci se šalju s mobilnog uređaja tek kada je on priključen na punjač iz razloga što komunikacija s poslužiteljem može biti veliki potrošač energije. Također, sudionik može sam odrediti prag postotka baterije nakon kojeg će se isključiti svi eksperimenti, odnosno samo izvođenje skripte na mobilnom uređaju.



Slika 2.5. APISENSE mrežno sučelje

3. ANONIMIZACIJA PODATAKA I PRIVATNOST

U drugome poglavlju opisan je princip rada nabave iz mnoštva te kako mobilni uređaji, odnosno korisnici mobilnih uređaja mogu značajno doprinijeti nabavi podataka. Postupak je to kojim se može na relativno brz i jeftin način doći do rješenja nekog problema, novih ideja te povratnih informacija. Unatoč tomu, kao što sve ima dobre strane, jednako tako ima i loše strane, što je u ovom slučaju narušavanje korisnikove privatnosti. Budući da MCS zahtjeva ljudsku uključenost, time se i korisnički podaci, podaci senzora s mobilnog uređaja prenose na platformu. Posebno veliki rizik postoji kod usluga koje zahtijevaju lokaciju (engl. *location based services,* LBS). Dodatan problem je što je topologija korisnika koji su odabrali određeni zadatak dinamična, u smislu da se stalno mogu pojavljivati novi sudionici, a ostali mogu odustati, što dodatno otežava očuvanje sigurnosti njihovih podataka [3]. Nadalje, pri odabiru zadatka, potrebni su i osobni podaci korisnika, pogotovo ako zadatak nosi sa sobom određenu nagradu. Dakle, sve osjetljive informacije nalaze na platformi, koja zatim ima zadatak zaštiti iste. Kako bi se to postiglo, ti podaci se anonimiziraju. Anonimizacija podataka je pretvaranje osobnih podataka u anonimizirane podatke primjenjujući tehnike anonimizacije [7]. Anonimizacija podrazumijeva uklanjanje direktnih i indirektnih osobnih identifikatora i primjenu tehničke zaštite [8]. Anonimizirani podaci ne bi trebali imati mogućnost vraćanja u originalni oblik, ali isto tako osoba koja je u posjedu tih podataka ne bi trebala moći obzirom na ostale podatke i neke druge izvore doći do identiteta i ostalih osjetljivih informacija. Tehnike anonimizacije najčešće uključuju uklanjanje jednog dijela podataka ili enkripciju istih. Cilj je da podaci nakon primjene tehnika i dalje budu korisni u svrhu analize. Na početku procesa anonimizacije nalaze se osobni podaci iz kojih se jasno vide sve informacije o nekoj osobi. Tada slijedi anonimizacija, odnosno primjena neke od tehnika anonimizacije. Na kraju, podaci uopće ne moraju biti potpuno anonimizirani, to jest, i dalje može postojati mogućnost da se indirektno identificira osoba. Takvi podaci nazivaju se *pseudonimizirani* podaci. Razine anonimizacije podataka dane su na slici 3.1.

Slika 3.1. Razine anonimizacije podataka

Dakle, primjenom tehnika anonimizacije ne mora značiti da će i krajnji podaci biti anonimizirani, ali će privatnost biti ipak malo više zaštićena.

3.1. Tehnike anonimizacije

Prije ulaska u detalje vezane uz tehnike potrebno je napomenuti da se podaci se najčešće pohranjuju u obliku tablica, gdje atribut predstavlja jedan stupac, jednu semantičku vrijednost (ime, prezime, adresa, poštanski broj i sl.). Atributi se dijele u slijedeće kategorije [9]:

1. Identifikatori (atributi koji imaju svojstvo da jedinstveno identificiraju pojedinca)
2. Kvazi-identifikatori (atributi koji uz podatke iz drugih izvora omogućuju identifikaciju)
3. Osjetljivi atributi (atributi koji sadrže osjetljive informacije poput bolesti pojedinca)

Pseudonimizacija

Atribut se zamjenjuje *pseudonimom*, što je izmišljena vrijednost koja može biti ireverzibilna ili reverzibilna. Ukoliko je reverzibilna, ona podrazumijeva čuvanje tablice koja povezuje pseudonim s originalnom vrijednosti atributa i nalazi se pod sigurnosnom kontrolom. Tehnike koje se koriste za stvaranje pseudonima su: ispremetanje (engl. *scrambling*), brojač (eng. *counter*), generator slučajnih brojeva (eng. *random number generator*, RNG), enkripcija, funkcija raspršivanja (engl. *hash function*), kod za provjeru autentičnosti poruke (engl. *message authentiacion code*, MAC) [10]. Primjer upotrebe generatora slučajnih brojeva u svrhu pseudonimizacije dan je na slici 3.2.

OSOBA	REZULTATI ISPITA	SATI UČENJA
Joe Hang	5	20
Zack Lim	4	26
Eu Cheng	3	30
Linnie Mok	2	29
Jeslyn Tan	4	32
Chan Lee	5	25

→

OSOBA	REZULTATI ISPITA	SATI UČENJA
416765	5	20
562396	4	26
964825	3	30
873892	2	29
239976	4	32
943145	5	25

Slika 3.2. Pseudonimizacija

Ukidanje zapisa (engl. *record suppression*)

Uklanja se cijeli redak u tablici, što razlikuje ovu tehniku od ostalih jer zahvaća sve atribute, a podrazumijeva potpuni gubitak informacija kao i uklanjanje atributa.

Maskiranje znaka

Promjena znaka u podatku s nekim simbolom, na primjer s ' * '. Obično se maskira dio podatka poput zadnjih nekoliko brojeva broja kartice, poštanskog broja. Na slici 3.3. dan je primjer kako izgleda rezultat maskiranja poštanskog broja, gdje su maskirane zadnje četiri znamenke.

POŠTANSKI BROJ	VRIJEME DOSTAVE	BROJ NARUDŽBI
100111	8:00 - 9:00	2
200222	10:00 - 11:00	8
300333	11:00 - 12:00	3

→

POŠTANSKI BROJ	VRIJEME DOSTAVE	BROJ NARUDŽBI
10****	8:00 - 9:00	2
20****	10:00 - 11:00	8
30****	11:00 - 12:00	3

Slika 3.3. Maskiranje znaka

Uklanjanje atributa

Podrazumijeva uklanjanje cijelog stupca iz podatkovne tablice, odnosno atributa koji može biti, na primjer: ime, prezime, adresa i sl. Ovo je najsnažnija tehnika jer ne postoji način da se podaci iz stupca vrate, izbrisani su zauvijek što je vidljivo na slici 3.4.

OSOBA	TRENER	REZULTATI ISPITA
Joe Hang	Tina	67
Zack Lim	Tina	80
Eu Cheng	Tina	55
Linnie Mok	Huang	29
Jeslyn Tan	Huang	81
Chan Lee	Huang	58

→

TRENER	REZULTATI ISPITA
Tina	67
Tina	80
Tina	55
Huang	29
Huang	81
Huang	58

Slika 3.4. Uklanjanje atributa

Generalizacija

Smanjuje se preciznost informacija na način da se grupiraju – datumi rođenja ili starost osobe se predstavlja rasponom, a ne točnom godinom, lokacija je predstavljena nekom manje preciznom, na primjer uklanjanje kućnog broja adrese ili ostaviti samo naziv mjesta. Primjer rezultata prikazivanja starosne dobi rasponom gdje je korak jednak devet, prikazano je na slici 3.5.

OSOBA	GODINE
357703	24
233121	31
9386637	44
591493	29
202626	23
888948	75
175878	28
312304	50
214025	30
271174	37
341338	22
529057	25
390438	39

→

OSOBA	GODINE
357703	21-30
233121	31-40
9386637	41-50
591493	21-30
202626	21-30
888948	>60
175878	21-30
312304	41-50
214025	21-30
271174	31-40
341338	21-30
529057	21-30
390438	31-40

Slika 3.5. Generalizacija

Zamjena ili permutacija

Razmještanje vrijednosti jednog ili više atributa tako da više ne odgovaraju originalnom zapisu. Dakle, podaci su i dalje tu, ali ova metoda nije pogodna ukoliko se podaci koriste za analiziranje veza između atributa. Ukoliko se želi odrediti broj zaposlenih osoba srednje dobi u nekom gradu ili broj medicinskih sestara starije dobi, tada se ova metoda ne primjenjuje. Primjer permutacije dan je na slici 3.6.

OSOBA	POSAO	DATUM ROĐENJA	VRSTA ČLANSTVA	MJESEČNI POSJETI
A	dekan	3.1.1970.	silver	0
B	prodavač	5.2.1972.	platinum	5
C	odvjetnik	7.3.1985.	gold	2
D	programer	10.4.1990.	silver	1
E	medicinska sestra	13.5.1995.	silver	2

↓

OSOBA	POSAO	DATUM ROĐENJA	VRSTA ČLANSTVA	MJESEČNI POSJETI
A	odvjetnik	10.4.1990.	silver	1
B	medicinska sestra	7.3.1985.	silver	2
C	prodavač	13.5.1995.	platinum	5
D	programer	3.1.1970.	silver	2
E	dekan	5.2.1972.	gold	0

Slika 3.6. Zamjena

Perturbacija podataka

Vrijednosti podataka se neznatno promjene od onih originalnih. U primjeru na slici 3.7. nove vrijednosti postignute su primjenom običnog zaokruživanja na bazu x (eng. *conventional rounding*

OSOBA	VISINA [cm]	TEŽINA [kg]	GODINE	PUŠAČ	BOLEST A	BOLEST B
A	160	50	30	ne	ne	ne
B	177	70	36	ne	ne	da
C	158	46	20	da	da	ne
D	173	75	22	ne	ne	ne
E	169	82	44	da	da	da

↓

OSOBA	VISINA [cm]	TEŽINA [kg]	GODINE	PUŠAČ	BOLEST A	BOLEST B
A	160	51	30	ne	ne	ne
B	175	69	36	ne	ne	da
C	160	45	18	da	da	ne
D	175	75	21	ne	ne	ne
E	170	81	42	da	da	da

Slika 3.7. Perturbacija podataka

to base x). To je metoda koja zaokružuje broj na najbliži višekratnik broja x. Tako je za visinu baza jednaka 5 pa se vrijednost od 173 zaokružuje na 175. Primjenjuje se na kvazi-identifikatore ukoliko njihova promjena ne utječe na analizu, to jest, gdje točnost tih vrijednosti nije nužna.

Agregacija podataka

Lista vrijednosti podatkovnog skupa se agregira (na primjer, zbrajanjem). Ovisno o tome koji podaci su potrebni, odlučuje se što će se agregirati. U primjeru na slici 3.8. tvrtki su bile dovoljne informacije o uplatama i u agregiranom obliku, a ne za svaku osobu pojedinačno.

DONATOR	MJESEČNI PRIHODI [$]	DONIRANO [$]
A	4000	210
B	4900	420
C	2200	150
D	4200	110
E	5500	260
F	2600	40
G	3300	130
H	5500	210
I	1600	380
J	3200	80
K	2000	440
L	5800	440
M	4600	390
N	1900	480
O	1700	320
P	2400	330
Q	4300	390
R	2300	260
S	3500	80
T	1700	290

→

MJESEČNI PRIHODI [$]	BROJ DONACIJA	SUMA DONACIJA [$]
1000-1999	4	1470
2000-2999	5	1220
3000-3999	3	290
4000-4999	5	1520
5000-6000	3	870
UKUPNO	**20**	**5370**

Slika 3.8. Agregacija podataka

3.2. Mjerila anonimizacije

U svrhu zaštite podataka koriste se kriteriji koji podaci moraju zadovoljiti kako bi se smatrali anonimiziranima. Tehnike za mjerenje anonimiziranosti, odnosno jesu li kriteriji zadovoljeni, koriste se kao smjernice tokom procesa anonimizacije, dakle, tokom procesa primjenjivanja tehnika anonimizacije i nakon primjene istih.

3.2.1. K-anonimnost

K-anonimnost je mjera zaštite protiv napada povezivanja (engl. *linkage attack*). Kako bi se zadovoljila k-anonimnost, potrebno je da za svaki zapis postoji još najmanje *k*-1 sličnih ili istih zapisa. U tome slučaju teško je izdvojiti pojedinca iz grupe, ali je i teže identificirati pojedinca povezujući podatke iz drugih izvora jer postoji *k* zapisa koji su jednaki po vrijednosti u identificirajućim atributima. Dakle, ako je $k = 5$, to znači da u tablici postoji još četiri takva zapisa. Ova metoda primjenjuje se na kvazi-identifikatore, što znači da se svaki kvazi-identifikator mora

pojaviti u identičnim kombinacijama u tablici u najmanje k zapisa [9]. Nad zapisima koji ne zadovolje kriterij, vrši se ukidanje zapisa . Što je veći *k*, veća je i mjera zaštite, ali samim time veća je i distorzija podataka [11]. Najveća vjerojatnost reidentifikacije svakog podatka ekvivalentnog razreda iznosi 1/*k*. Ekvivalentni razred predstavlja skup zapisa koji su međusobno nerazlučivi. Na slici 3.9. primijenjene su tehnike anonimizacije: generalizacija i ukidanje zapisa (gdje nije zadovoljena *k*-anonimnost) te nakon njih *k*-anonimnost (*2*-anonimnost).

GODINE	SPOL	ZANIMANJE	PROSJEČAN BROJ PUTOVANJA U TJEDNU
21	Ž	pravni savjetnik	15
38	M	policajac	2
25	Ž	bankar	8
44	Ž	administrator baze podataka	3
25	Ž	asistent	1
31	M	policajac	5
42	Ž	prgoramer	3
22	Ž	asistent	4
30	Ž	pravni savjetnik	2

→

GODINE	SPOL	ZANIMANJE	PROSJEČAN BROJ PUTOVANJA U TJEDNU
21-30	Ž	pravni savjetnik	15
31-40	M	policajac	2
~~21-30~~	~~Ž~~	~~bankar~~	~~8~~
41-50	Ž	IT	3
21-30	Ž	asistent	1
31-40	M	policajac	5
41-50	Ž	IT	3
21-30	Ž	asistent	4
21-30	Ž	pravni savjetnik	2

Slika 3.9. *K*-anonimizacija

3.2.2. L-raznolikost (engl. *l-diversity*)

Budući da *k*-raznolikost nije dovoljna, l-raznolikost je proširenje koje pruža zaštitu od napada izvođenih zaključivanjem (engl. *inference attack*). Skup podataka (koji je prethodno *k*-anonimiziran) smatra se da ima *l*-raznolikost ako:

- Svaki ekvivalentni razred sadrži *l* različitih vrijednosti za osjetljivi atribut → *distinct l-diversity*
- Entropija vrijednosti koja se pojavljuje jednaka je ili veća od *log(l)* → entropijska *l*-raznolikost
- Najčešća vrijednost ne pojavljuje se previše puta, a ona manje češća ne pojavljuje se prerijetko → rekurzivna *l*-raznolikost

U primjeru na slici 3.10. tablica s lijeve strane je *2*-anonimizirana. Problem je što zadnja dva zapisa imaju istu imaju bolest, isti osjetljivi atribut.

POŠTANSKI BROJ	GODINE	DIJAGNOZA
421*	30-39	prehlada
421*	30-39	zdrav
47**	20-39	upala uha
47**	20-39	upala uha

→

POŠTANSKI BROJ	GODINE	DIJAGNOZA
4***	20-39	prehlada
4***	39	zdrav
4***	39	upala uha
4***	20-39	upala uha

Slika 3.10. *L*-raznolikost

Stoga, na tablicu s desne strane primijenjena je *2*-raznolikost te se može primijetiti da za kombinacije vrijednosti kvazi-identifikatora postoje različite bolesti. Na primjer, ako napadač cilja na osobu za koju zna da ima 22 godine i da je njen poštanski broj 4723, više ne može sa sigurnošću znati da je njena dijagnoza zasigurno upala uha.

3.2.3. T-bliskost (engl. *t-closeness*)

T-bliskost jest nadogradnja prethodnih dvaju kriterija. Ekvivalentni razred ima *t*-bliskost ukoliko udaljenost distribucije osjetljivih atributa u tom razredu i distribucija atributa u cijeloj tablici nije veća od praga *t*. Kaže se da tablica ima *t*-bliskost ako ju imaju svi ekvivalentni razredi [12]. Udaljenost se računa koristeći *Earth Moverovu* udaljenost (engl. *Earth Mover's Distance,* EMD). Također, *t* je realni broj za razliku od ranije spomenutih *k* i *l*. Kako bi se izračunala udaljenost između atributa, potrebno je prvo odrediti jesu li njegove vrijednosti brojčane (na primjer, godine) ili semantičke (na primjer, naziv bolesti).

Za brojčane vrijednosti udaljenost između *P* i *Q* računa se prema

$$\boldsymbol{E(P,Q) = \frac{1}{m-1} \sum_{i=1}^{m} \left| \sum_{j=1}^{i} (p_j - q_j) \right|}, \tag{3.1.}$$

gdje su P i Q distribucije za koje se pretpostavlja da su normalne razdiobe te da vrijedi $1 \leq i, j \leq m$, a p_i i q_i pripadne su vjerojatnosti distribucija.

Uzme li se da je $Q_1' = \{1, 72, 99, 101\}$ te da je $P_1' = \{1, 99\}$ te da je pravi podskup i ujedno ekvivalentni razred od Q_1', tada je $Q_1 = \left\{\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right\}$ jer se svaka vrijednost pojavljuje točno jednom. Budući da P_1 također mora biti veličine $m = 4$ kao i Q_1, umjesto vrijednosti koje se ne pojavljuju (72, 101) pridružuje se vjerojatnost jednaka nuli te je $P_1 = \left\{\frac{1}{2}, 0, \frac{1}{2}, 0\right\}$. Izračun *t*-bliskosti između P_1 i Q_1 dan je izrazom 3.2.

$$\boldsymbol{E(P_1, Q_1) = \frac{1}{4-1} \left[\left|\frac{1}{2} - \frac{1}{4}\right| + \left|\left(\frac{1}{2} - \frac{1}{4}\right) + \left(0 - \frac{1}{4}\right)\right| + \left|\left(\frac{1}{2} - \frac{1}{4}\right) + \left(0 - \frac{1}{4}\right) + \left(\frac{1}{2} - \frac{1}{4}\right)\right| + \left|\left(\frac{1}{2} - \frac{1}{4}\right) + \left(0 - \frac{1}{4}\right) + \left(\frac{1}{2} - \frac{1}{4}\right) + \left(0 - \frac{1}{4}\right)\right| \right] \approx 0.1667} \tag{3.2.}$$

Što se tiče semantičkih atributa, koristi se

$$\boldsymbol{E'(P,Q) = \frac{1}{2} \sum_{i=1}^{m} |p_i - q_i|} \tag{3.3.}$$

Za primjer priložena je tablica na slici 3.11. te izraz 3.4. koji prikazuje izračun *t*-bliskosti između Q_2 i $P_{2,1}$, gdje je Q_2 = {nestanak struje, krađa, požar, prometna nesreća, popravak kolnika, kontrola

štetočine, sadnja drveća} te je $P'_{2,1}$ = {nestanak struje, nestanak struje, nestanak struje} s pripadajućom distribucijom $P_{2,1} = \left\{\frac{3}{3}, 0, 0, 0, 0, 0, 0\right\}$.

$$E'(P_{2,1}, Q_2) = \frac{1}{2}\left[\left|\frac{3}{3} - \frac{5}{14}\right| + \left|0 - \frac{1}{14}\right| + \left|0 - \frac{2}{14}\right| + \left|0 - \frac{1}{14}\right| + \left|0 - \frac{3}{14}\right| + \left|0 - \frac{1}{14}\right| + \left|0 - \frac{1}{14}\right|\right] \approx 0.6429 \quad (3.4.)$$

Nalik tomu, ekvivalentni razred $P'_{2,2}$ = {krađa, požar, prometna nesreća, požar} ima distribuciju $P_{2,2} = \left\{0, \frac{1}{4}, \frac{2}{4}, \frac{1}{4}, 0, 0, 0\right\}$. *T*-bliskost ostalih ekvivalentnih razreda iznosi $E'(P_{2,2}, Q_2) \approx 0.7143$, $E'(P_{2,3}, Q_2) \approx 0.4286$ i $E'(P_{2,4}, Q_2) \approx 0.4429$. Iznos *t*-bliskosti za cijelu tablicu jednak je najvećoj vrijednosti, dakle, 0.7143.

ADRESA	ZONA	INCIDENT
*	2C	nestanak struje
*	2C	nestanak struje
*	2C	nestanak struje
*	4F	krađa
*	4F	požar
*	4F	prometna nesreća
*	4F	požar
*	9A	popravak kolnika
*	9A	nestanak struje
*	3B	popravak kolnika
*	3B	sadnja drveća
*	3B	popravak kolnika
*	3B	kontrola štetočine
*	3B	nestanak struje

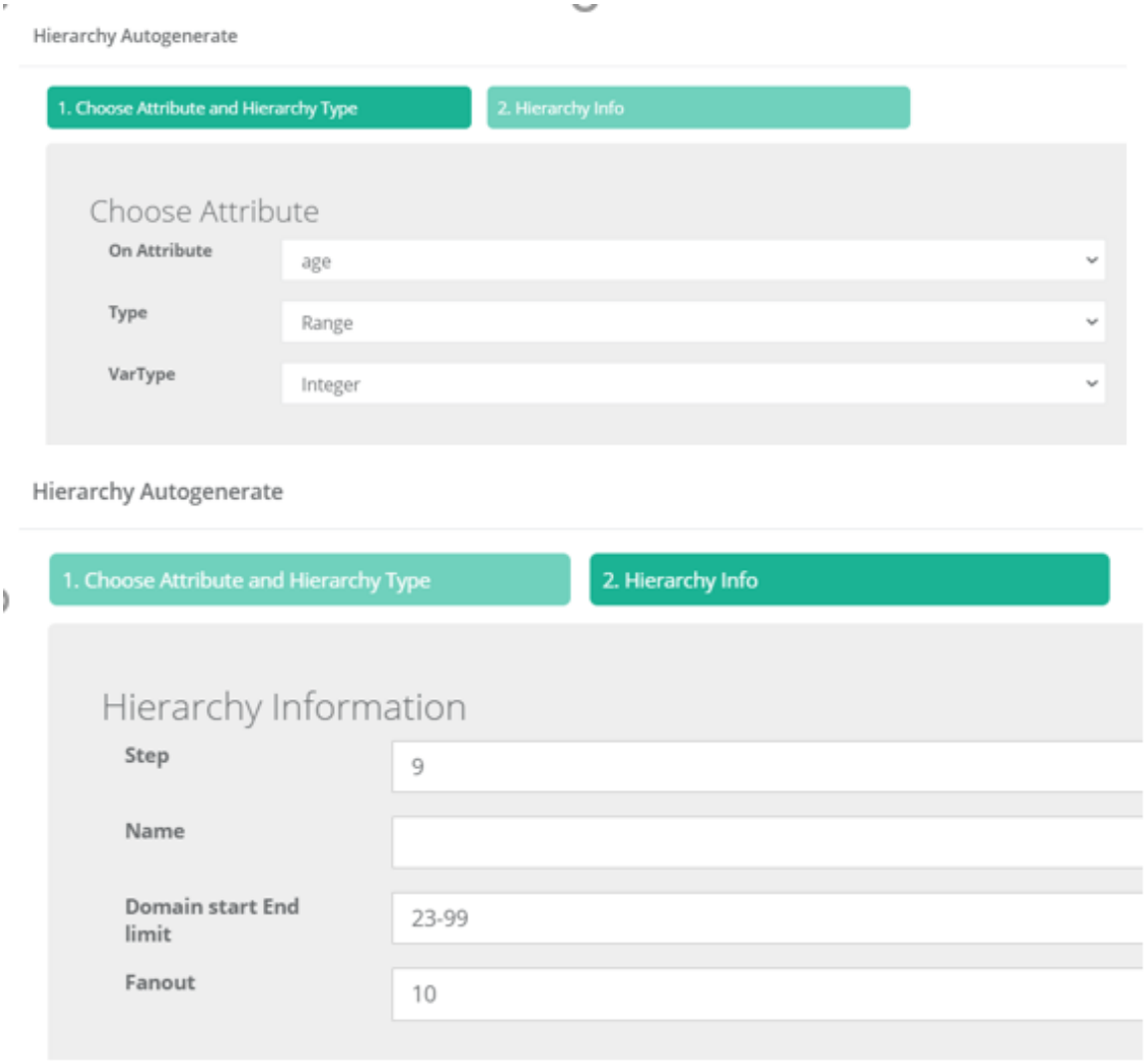
Slika 3.11. *T*-bliskost

3.3. Postojeći alati za anonimizaciju

Opća uredba o zaštiti podataka (eng. *General Data Protection Regulation*, GDPR) zakon je o zaštiti privatnosti i osobnih podataka koji se primjenjuje unutar Europske Unije. Dakle, tvrtke koje se bave obradom podataka koje potječu iz Europske Unije moraju poštivati ovaj zakon. Iz tog razloga alati za anonimizaciju olakšavaju rad tim tvrtkama. U ovome potpoglavlju opisani su neki od alata koji omogućavaju anonimizaciju podataka.

3.3.1. Amnesia

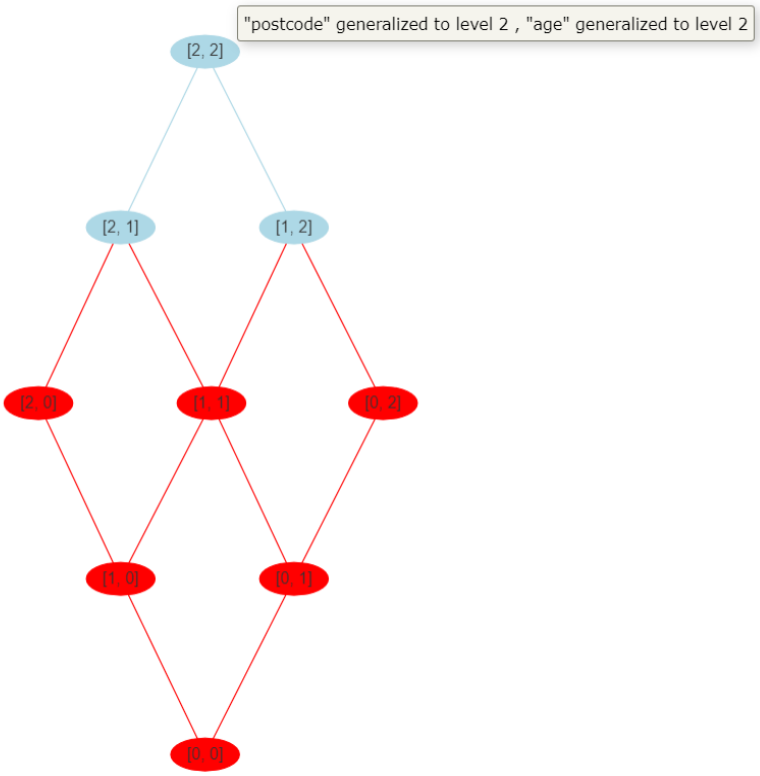
Postupak anonimizacije skupa podataka pomoću ovog alata sastoji se od tri koraka. Na početku potrebno je skup podataka uvesti u program. Ti podaci trebaju biti u obliku tekstualne datoteke gdje su atributi odvojeni nekim razdjelnikom, na primjer crticom. Tada program pretpostavlja tip podatka atributa, a korisnik ju potvrdi ili ispravi. Zatim je potrebno odrediti nad kojim atributom će se provesti generalizacija. Na primjer za atribut koji predstavlja godine, može se odrediti kolike će veličine biti intervali, odnosno raskorak između pojedinih, a prikaz sučelja nalazi se na slici 3.12. Tada se iz grafičkog prikaza može vidjeti rezultat generalizacije, odnosno hijerarhija vrijednosti. Zadnji korak jest odabrati tehniku anonimizacije koja će se primijeniti na podatke.



Slika 3.12. Amenisa generalizacija atributa

U slučaju *k*-anonimizacije potrebno je odrediti *k* te program primjeni algoritam. Rezultat je graf u kojem plavi čvorovi predstavljaju sigurno rješenje, a crveni nesigurno koji se mogu pretvoriti u sigurno daljnjom primjenom ukidanja atributa. Primjer rezultata anonimizacije ovim alatom

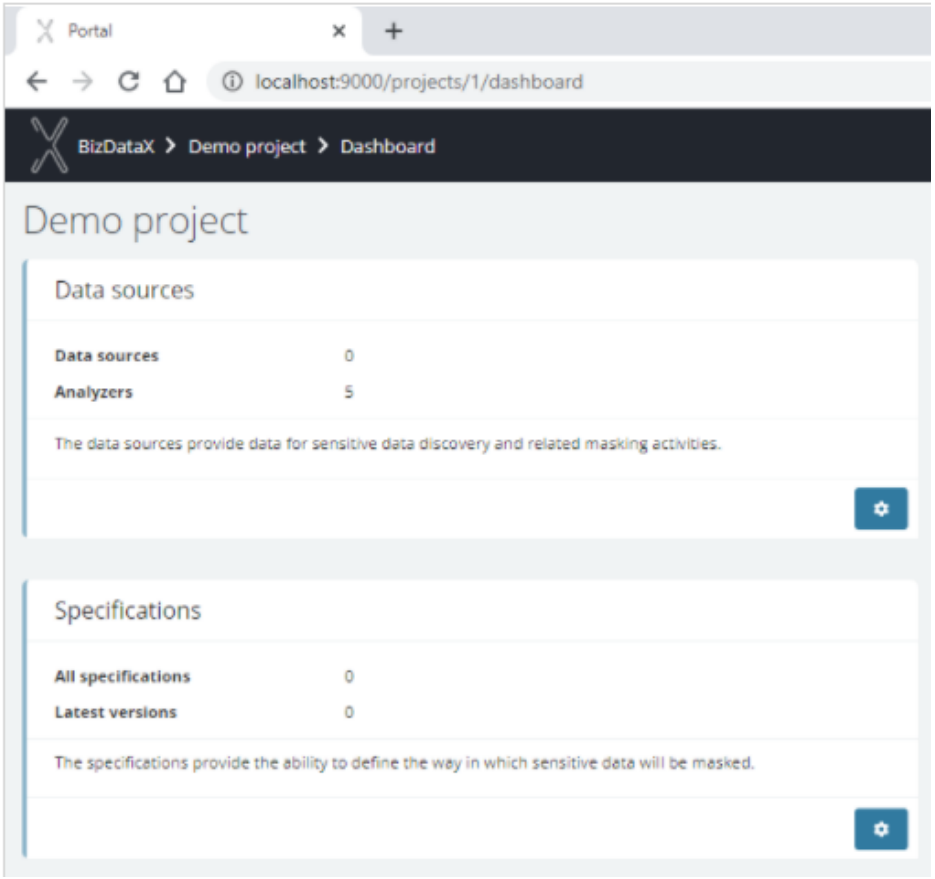
prikazan je na slici 3.13. Brojevi označavaju razinu generalizacije kvazi-identifikatora te ukoliko se klikne na određeno rješenje dobije se prikaz podataka.



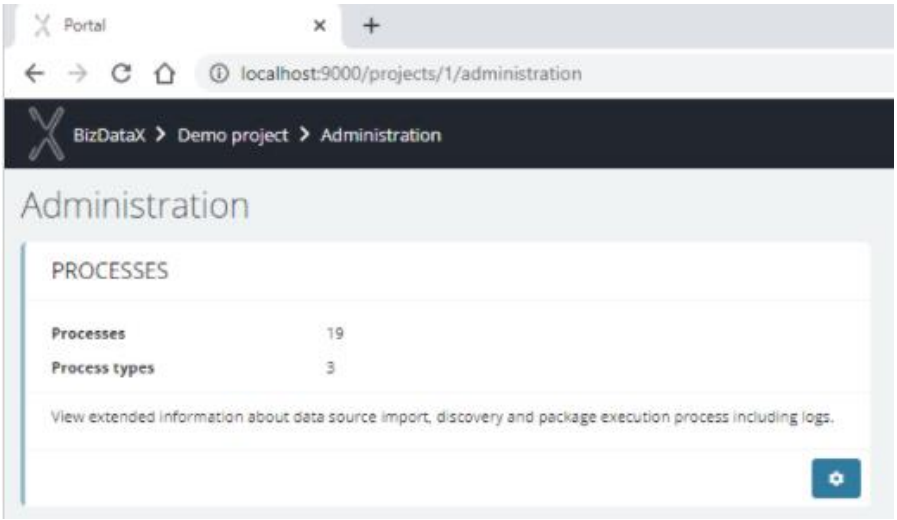
Slika 3.13. Amnesia rezultat *2*-anonimizacije

3.3.2. BizDataX Portal

Ova web aplikacija osim što nudi maskiranje te anonimizaciju, nudi i otkrivanje osjetljivih podataka, kreiranje sintetičkih podataka te stavljanje podataka u podskup. Za uvoz podataka moguće je spojiti se na bazu podataka te će aplikacija automatski učitati podatke kako bi se mogli dalje koristiti. Na samome početku kreira se projekt i uvezu se podaci. Na slici 3.15. prikazan je dio kontrolne ploče gdje izvori podataka opisuju stanje izvora podataka u projektu te broj dostupnih analizatora izvora podataka. Specifikacije omogućuju definiranje načina maskiranja podataka. Ova web aplikacija ima omogućava korisniku definiranje parametara, pravila koji će biti primijenjeni pri pronalasku specifičnih tipova podataka u specifičnom izvoru podataka. Nalazi otkrivanja (engl. *discovery findings*) predstavljaju kandidate za osjetljive podatke otkrivene korištenjem spomenutih pravila, a oni se tada mogu označiti kao osjetljivi ili ne. Paketi i povijest njihova izvođenja nalazi se pod paketima (engl. *packages*). Oni sadrže definicije za maskiranje podataka. Zadnji dio kontrolne ploče jest administracija pod kojom se mogu pronaći sve informacije vezane uz procese koji su se izvodili ili izvode nad podacima, što je prikazano na slici 3.16. *BizDataX* dizajner dodatak je za *Microsoft Visual Studio* koji omogućuje dizajniranje paketa.



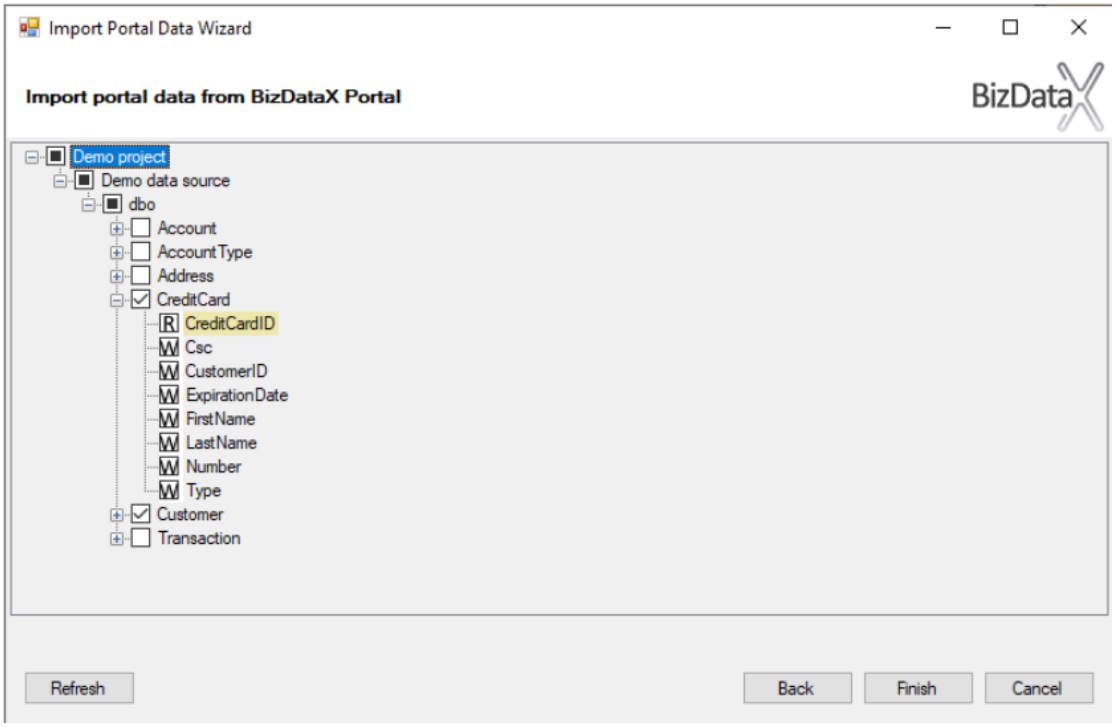
Slika 3.14. BizDataX Portal kontrolna ploča



Slika 3.15. BizDataX Portal administracija

Potrebno je tablice iz baze podataka prenijeti iz *BizDataX Portal*-a u *BizDataX* projekt *Visual Studio*, ali je dovoljno prenijeti samo one koje je potrebno maskirati. Atributi tablice se mogu prenijeti sa svojstvom da su samo za čitanje (engl. *read-only)* ili za čitanje i pisanje (engl. *read-write*). Slika 3.18. prikazuje odabiranje svojstva za atribute. Nakon što se paket kreira potrebno ga

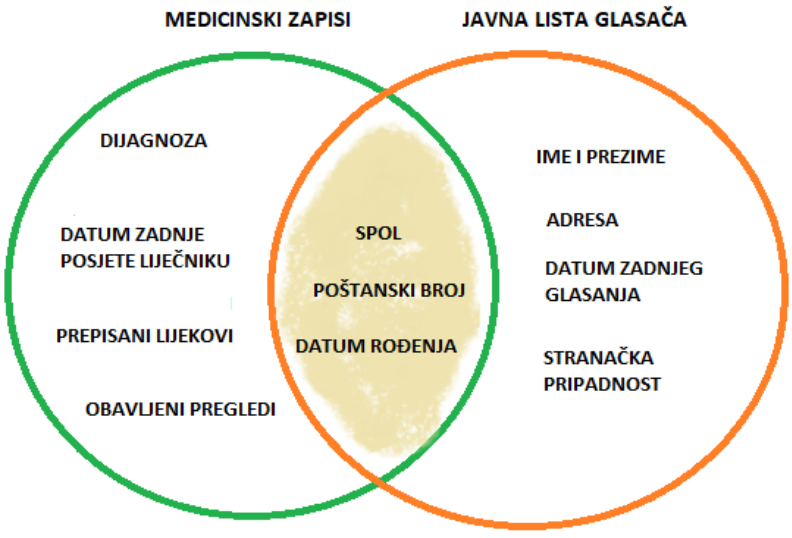
je objaviti na *BizDataX Portal*-u te će biti vidljiv pod paketima. Tada se paket može pokrenuti i putem naredbenog retka (engl. *command prompt*), web aplikacije ili iz *Microsoft Visual Studio*.



Slika 3.16. *BizDataX Designer* uvoz podataka

3.4. Napadi na anonimizirane podatke

Mjerila i modeli privatnosti stvoreni su radi očuvanja privatnosti podataka. Do potrebe za time došlo je iz razloga što se sve češće dijele podatkovni skupovi u analitičke svrhe. Problem je što takvi podatkovni skupovi sadrže osjetljive informacije.



Slika 3.17. Napad povezivanja

Kako bi podaci bili korisni za analizu, u velikom broju slučaja upravo su potrebne te osjetljive informacije. Modeli privatnosti temelje se na tome da je gubitak informacija što manji, ali samim time i dalje postoji mogućnost otkrivanja identiteta.

- **Napad povezivanja (engl. *linkage attack*)**

Ovaj napad sprječava se modelom *k*-anonimnosti. Podrazumijeva otkrivanje osjetljivih informacija ili identiteta iz dvaju ili više podatkovnih skupova. Iz javne liste glasača gdje se nalaze osnovne informacije o osobi poput spola, poštanskog broja i datuma rođenja, povezivanjem tih vrijednosti s vrijednostima iz medicinskog zapisa, napadač može vrlo lako otkriti od čega određena osoba boluje. Dakle, ovaj napad izvodi se pomoću povezivanja vrijednosti kvazi-identifikatora. Napadi na *k*-anonimizirane podatke opisani su u nastavku [13].

- **Napad na homogenost (engl. *homogeneity attack*)**

Temelji se na tome što ekvivalentni razred sadrži *k* istih zapisa, odnosno na tome što nedostaje raznolikosti u vrijednostima osjetljivih atributa. Na primjer, osoba A zna poštanski broj osobe B koji je 13053 i zna da ima 35 godina. Prema medicinskom zapisu na slici 3.17. osoba A može zaključiti da osoba B pripada ekvivalentom razredu čije su vrijednosti osjetljivog atributa bolest rak.

BOLEST	NACIONALNOST	GODINE	POŠTANSKI BROJ
bolest srca	*	<30	130**
bolest srca	*	<30	130**
virus	*	<30	130**
virus	*	<30	130**
rak	*	>40	1485*
bolest srca	*	>40	1485*
virus	*	>40	1485*
virus	*	>40	1485*
rak	*	3*	130**
rak	*	3*	130**
rak	*	3*	130**
rak	*	3*	130**

Slika 3.18. *4*-anonimizirana tablica

- **Napad zaključivanja (engl. *background knowledge attack*)**

Ovaj napad odnosi se na slučaj kada osoba A poznaje osobu B. Ukoliko osoba A zna da je poštanski broj osobe B 14853 te da ima 50 godina, prema medicinskom zapisu na slici 3.17. ostaje mogućnost da osoba B ima rak, virus ili bolest srca. Međutim, osoba A također zna da osoba B ima nizak tlak te da izbjegava masnu hranu, stoga zaključuje da osoba B ima bolest srca. Nadalje su opisani napadi koji su prijetnja *l*-raznolikosti.

- **Napad na iskrivljenost (engl. *skewness attack*)**

L-raznolikost može biti zadovoljena u ekvivalentnim razredima, no ovaj napad se može izvesti ukoliko se učestalost pojavljivanja osjetljivog atributa unutar ekvivalentnog razreda razlikuje od učestalosti pojavljivanja u cijeloj tablici. Na primjer, postotak HIV pozitivnih osoba iznosi 1%. Ukoliko ekvivalentni razred sadrži 50% pozitivnih i negativnih, ono zadovoljava raznolikost međutim, narušava privatnost. *L*-raznolikost ne uzima u obzir cjelokupnu distribuciju vrijednosti osjetljivih atributa.

- **Napad na sličnost (engl. *similarity attack*)**

Da bi se zadovoljiva *l*-raznolikost, dovoljno je da osjetljivi atribut ima različite vrijednosti za istu kombinaciju kvazi-identifikatora. Unatoč tomu, vrijednosti koje imaju semantički sličnu vrijednost mogu narušiti privatnost. Na primjer, ako osjetljivi atribut označava vrstu ovisnosti (ovisnost o marihuani, heroinu, kokainu i sl.) tada je napadaču koji može biti diler droge ili službenik policije koji radi na odjelu za droge dovoljna informacija da je osoba ovisnik.

4. PRIMJENA ALGORITAMA ANONIMIZACIJE U SUSTAVIMA ZA PRIKUPLJANJE PODATAKA IZ MNOŠTVA

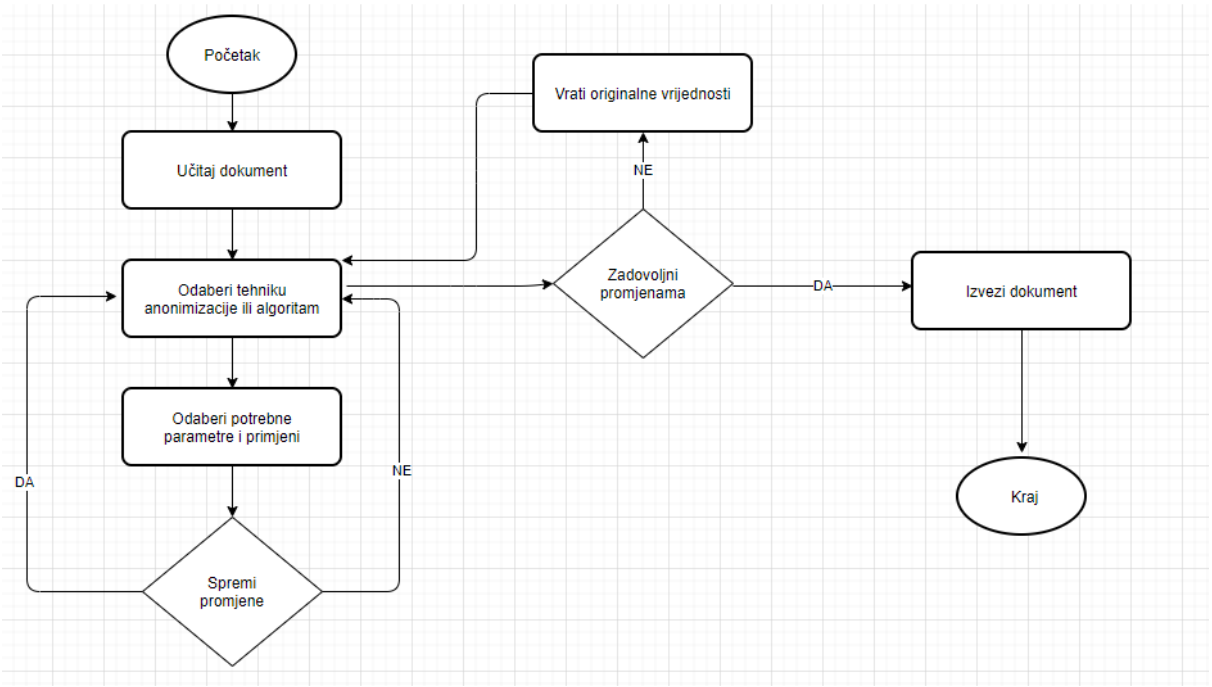
Praktični dio ovog rada ostvaren je korištenjem programskog jezika C# te okvira za stvaranje korisničkog sučelja za Windows naziva *Windows Presentation Foundation* koji je dio .NET radnog okvira. Cilj aplikacije imena *Simple Anonymizer* jest omogućiti primjenu tehnika anonimizacije na proizvoljnim skupovima podataka te njihov izvoz u obliku anonimiziranih podatkovnih skupova. U tablici 4.1. nalazi se popis funkcionalnosti, naziv te opis funkcionalnosti koje program nudi.

Tablica 4.1. Funkcionalnosti alata *Simple Anonymizer*

	FUNKCIONALNOST	NAZIV	OPIS
F1	učitavanje podataka u obliku CSV dokumenta	učitaj dokument	učitavanje CSV dokumenta pri kojem korisnik bira razdjelnik (',' ili ';') te njegov prikaz u tabličnom obliku
F2	izvoz anonimizarnih podataka u obliku CSV dokumenta	izvezi CSV dokument	spremanje anonimiziranih podataka u obliku CSV dokumenta pri kojem korisnik bira razdjelnik (',' ili ';')
F3	primjena tehnike anonmizacije	generalizacija	stvaranje intervala brojčanih vrijednosti s korakom između vrijednosti kojeg korisnik sam odabire
F4	primjena tehnike anonmizacije	uklanjanje atributa	korisnik odabire koji atribut (stupac) želi ukloniti
F5	primjena tehnike anonmizacije	maskiranje	zamjena *n* znakova vrijednosti retka sa znakom '*' ili '#' korisnik odabire znak maskiranja te početnu poziciju maskiranja koja može biti od početka ili kraja vrijednosti
F6	primjena tehnike anonmizacije	zamjena (permutacija)	vrijednosti u stupcu se nasumično razmještaju
F7	primjena tehnike anonmizacije	uklanjanje zapisa	korisnik klikom na zapis u tabličnom prikazu odabire zapis koji želi ukloniti
F8	primjena tehnike anonmizacije	perturbacija	zaokruživanje brojčanih vrijednosti stupca na višekratnik baze x koja može biti 0.5, 3, 5 ili 10
F9	primjena tehnike anonmizacije	pseudonimizacija	vrijednosti stupca mijenjaju se novom vrijednosti (pseudonimom)
F10	metoda stvaranja pseudonima	izmiješaj (engl. *scramble*)	razmještanje znakova vrijednosti

F11	metoda stvaranja pseudonima	enkripcija	primjena enkripcije na vrijednost
F12	metoda stvaranja pseudonima	generator slučajnih brojeva	stvaranje slučajnih brojeva primjenom kriptografskog generatora slučajnih brojeva
F13	metoda stvaranja pseudonima	funkcija raspršivanja (engl. *hash function*)	primjena funkcije raspršivanja na vrijednosti stupca
F14	primjena mjerila anonimizacije	k-anonimizacija	korisnik odabire broj *k* te za rezultat dobiva *k*-anonimiziran skup podataka
F15	primjena mjerila anonimizacije	l-raznolikost	provjera zadovoljava li skup podataka *l*-raznolikost pri čemu korisnik odabire *l*
F16	poništavanje anonimizacije na stupcu	vrati originalne vrijednosti	dohvaćanje originalnih vrijednosti za odabrani stupac
F17	pretpregled promjena nad podacima	primjeni	pretpregled promjena nad podacima nastalih kao rezultat primjena metoda anonimizacije
F18	spremi izmijenjene podatke	spremi	pohrana promjena nad podacima

Dijagram tijeka koji opisuje tijek rada programa dan je na slici 4.1. Na početku se učitava dokument s podacima koji se žele anonimizirati te koji moraju biti u CSV obliku. Prvi redak treba sadržavati nazive atributa.



Slika 4.1. Dijagram tijeka *Simple Anonymizer*-a

Zatim, odabire se tehnika anonimizacije koja se želi primijeniti na podatke od kojih *Simple Anonymizer* nudi one navedene tablici 4.1. Također, korisnik može primijeniti algoritam *k*-anonimizacije ili provjeriti zadovoljavaju li podaci *l*-raznolikost. Ukoliko se na atribut primjeni

neka tehnika, a korisnik se predomisli, postoji mogućnost dohvaćanja originalnih vrijednosti, odnosno, poništavanje primjene tehnike na odabranom atributu. Kada je korisnik zadovoljan sa stupnjem anonimnosti, takve anonimizirane podatke može spremiti u obliku CSV dokumenta.

4.1. Prikaz i priprema podatkovnog skupa

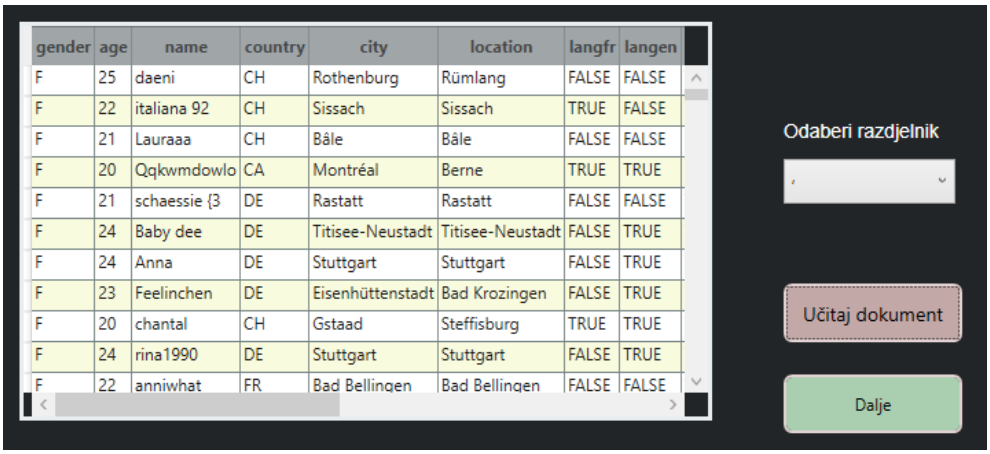
Podatkovni skup koji se koristi u svrhu prikaza primjene metoda anonimizacije sadrži informacije o korisnicima mrežne i mobilne aplikacije za upoznavanje, *Loovoo*. Podaci su prikupljeni tokom travnja i svibnja 2015. godine te se sastoji od 3974 zapisa, dostupnih na mrežnoj stranici *data.world*. Skup sadrži i poneke atribute poput broja posjeta profila, broja slika, ID profilne fotografije i sl. koji nisu nužno potrebni za prikaz anonimizacije pa su oni uklonjeni. Kvazi-identifikatori su: *gender, age, country, city, lang_fr, lang_en, lang_de, lang_it, lang_es i lang_pt*. Osjetljivi atribut je *location*. U tablici 4.2. dani su opisi spomenutih atributa.

Tablica 4.2. Atributi podatkovnog skupa

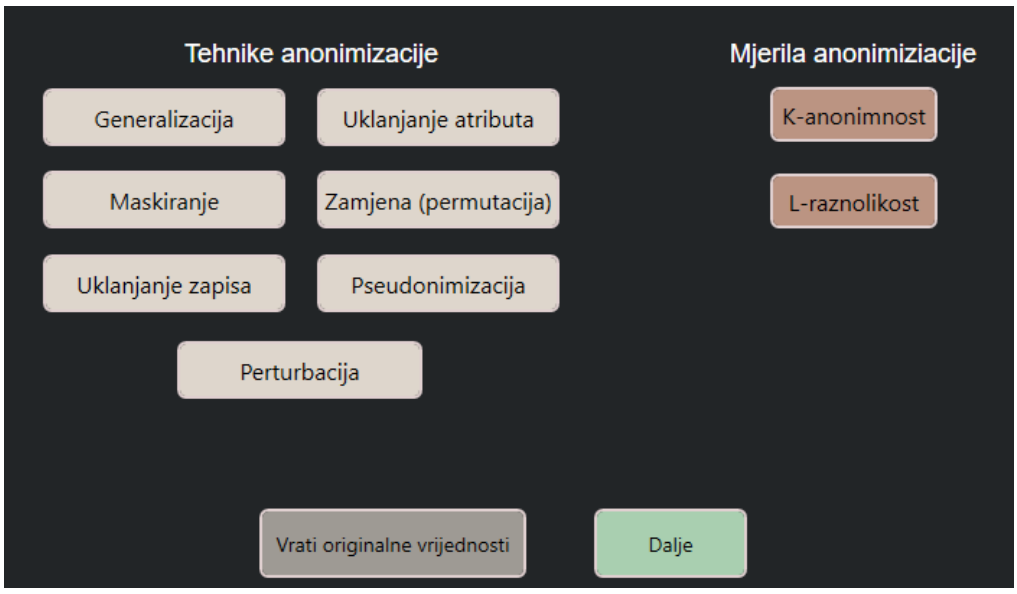
NAZIV	OPIS
gender	spol korisnika
age	starost korisnika
country	država u kojoj se korisnik nalazi
city	grad u kojem se korisnik nalazi
location	preciznija lokacija osobe
langfr	ukoliko korisnik zna francuski jezik, vrijednost je *TRUE* (1) u suprotnome *FALSE* (0)
langen	ukoliko korisnik zna engleski jezik, vrijednost je *TRUE* (1) u suprotnome *FALSE* (0)
langes	ukoliko korisnik zna španjolski jezik, vrijednost je *TRUE* (1) u suprotnome *FALSE* (0)
langde	ukoliko korisnik zna njemački jezik, vrijednost je *TRUE* (1) u suprotnome *FALSE* (0)
langit	ukoliko korisnik zna talijanski jezik, vrijednost je *TRUE* (1) u suprotnome *FALSE* (0)
langpt	ukoliko korisnik zna portugalski jezik, vrijednost je *TRUE* (1) u suprotnome *FALSE* (0)
name	korisničko ime
lastOnlineDate	datum kad je korisnik zadnji puta bio na mreži
lastOnlineTime	vrijeme kad je korisnik zadnji puta bio na mreži
userId	korisnički ID

4.2. Implementacija i usporedba različitih metoda anonimizacije

Recimo da je svrha analize ovog podatkovnog skupa odrediti koliki postotak ljudi iz određene države govori neki od ponuđenih stranih jezika. Cilj ovog dijela rada jest te podatke anonimizirati, no da i dalje mogu poslužiti analizi. Podatkovni skup koji je u obliku CSV dokumenta razdijeljen je zarezima, stoga je zarez odabran kao razdjelnik u kombiniranome okviru (engl. *combobox*). Pritiskom na gumb (engl. *button*) *Učitaj dokument*, u podatkovnoj mreži (engl. *data grid*) dobije se tablični prikaz podataka, a primjer toga dan je slikom 4.2. Pritiskom na gumb *Dalje*, otvara se novi prozor s anonimizacijskim tehnikama, mjerilima anonimizacije i gumbom koji nud vraćanje originalnih vrijednosti stupaca, odnosno atributa te gumb za dalje što je vidljivo na slici 4.3. Za početak, korisnički ID će se ukloniti iz podatkovnog skupa iz razloga što se u bazi podataka aplikacije sve informacije o korisniku vežu uz taj atribut koji je jedinstven za svakog korisnika.

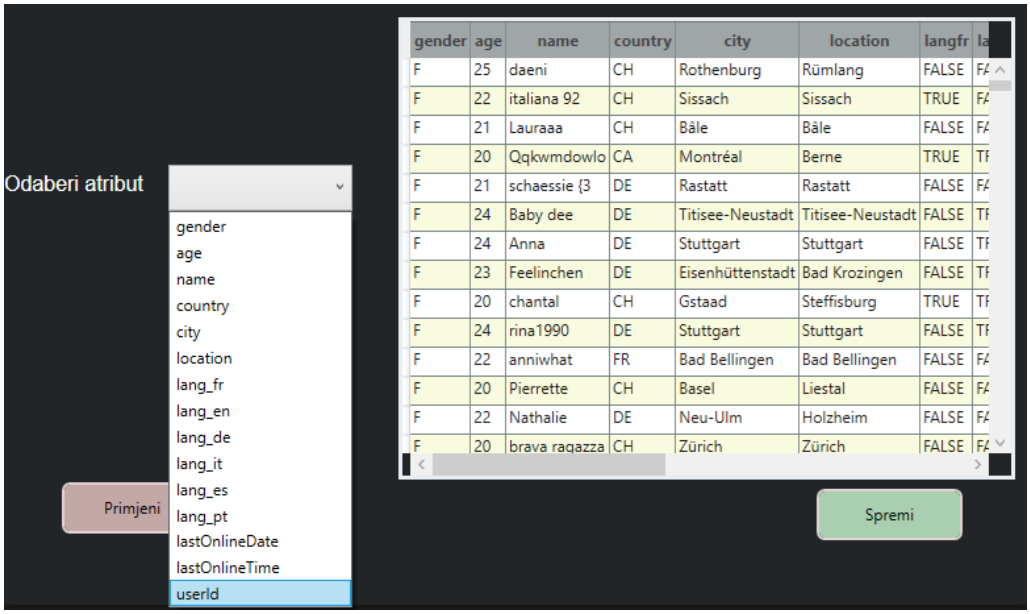


Slika 4.2. Početni prozor



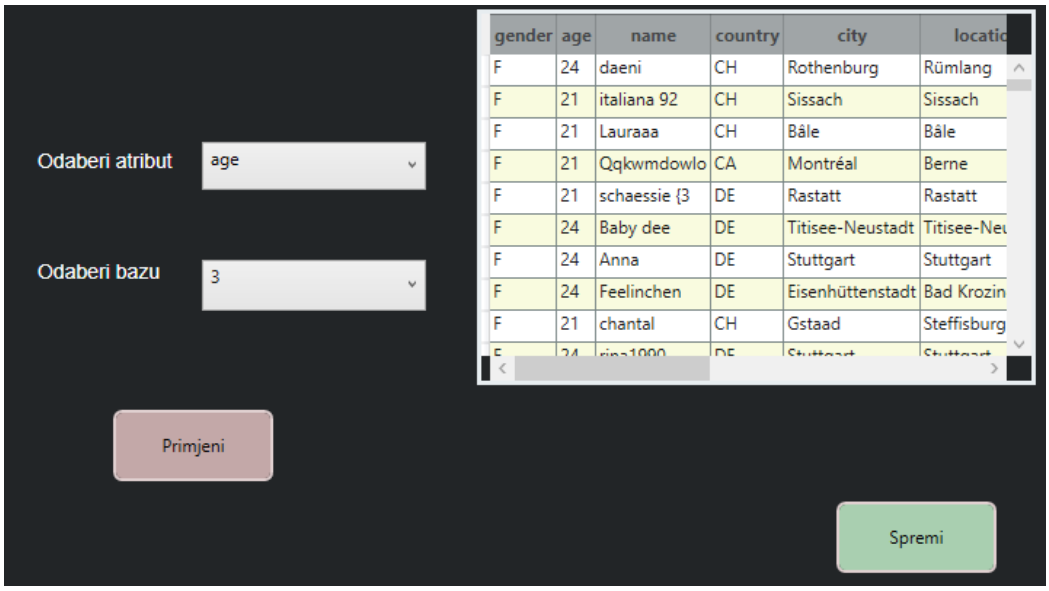
Slika 4.3. Prozor za odabir anonimizacijskih tehnika i mjerila

Kako bi se to postiglo, klikom na gumb *Uklanjanje atributa*, otvara se novi prozor, gdje se zatim u kombiniranome okviru bira naziv atributa kojeg se želi ukloniti. Pritiskom na gumb *Primjeni*, dobiva se pretpregled nove tablice vidljiv na slici 4.4. te ako se klikne gumb *Spremi*, te promjene se trajno spremaju i vrijednosti atributa se više ne mogu vratit čak ni klikom na gumb *Vrati originalne vrijednosti*. Što se tiče same metode uklanjanja atributa, ona je najsigurnija jer su podaci nepovratno izbrisani, ali je i time gubitak informacija stopostotni.



Slika 4.4. Prozor za uklanjanje atributa

Starost korisnika može se anonimizirati i primjenom perturbacije, koja neće dati točnu vrijednost, već približnu, što je vidljivo na slici 4.5. Nova vrijednost se dobiva na osnovu baze x, a baza se bira na osnovu raspona vrijednosti, a budući da je raspon veličine 10, odabrana je baza 3.



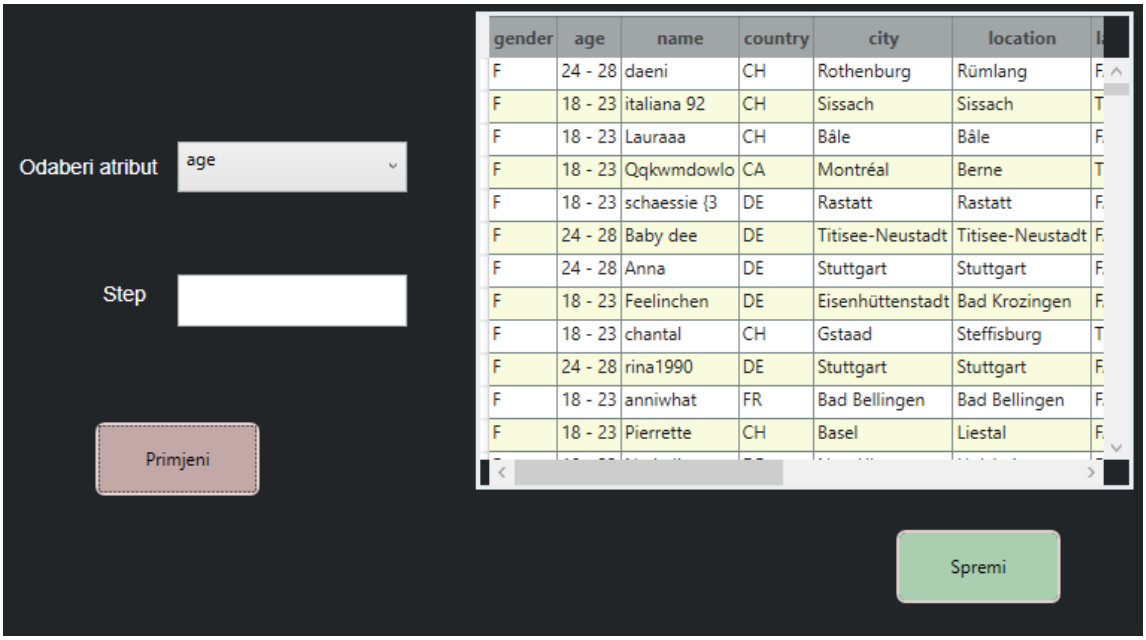
Slika 4.5. Prozor za perturbaciju

Obzirom da ovdje i dalje postoji konkretna vrijednost te da napadač može otkriti kako je primijenjena ova tehnika jer postoje samo višekratnici nekog broja, sigurnije bi bilo primijeniti generalizaciju. Prije nego što se primjeni, potrebno je vrijednosti atributa vratiti na početne, a prozor u kojem se to odrađuje te vraćene vrijednosti atributa dane su na slici 4.6. Generalizirat će se starost korisnika s korakom vrijednosti 5 iz razloga što je raspon godina od 18 do 28.

gender	age	name	country	city	location
F	25	daeni	CH	Rothenburg	Rümlang
F	22	italiana 92	CH	Sissach	Sissach
F	21	Lauraaa	CH	Bâle	Bâle
F	20	Qqkwmdowlo	CA	Montréal	Berne
F	21	schaessie {3	DE	Rastatt	Rastatt
F	24	Baby dee	DE	Titisee-Neustadt	Titisee-Neustad
F	24	Anna	DE	Stuttgart	Stuttgart
F	23	Feelinchen	DE	Eisenhüttenstadt	Bad Krozingen
F	20	chantal	CH	Gstaad	Steffisburg
F	24	rina1990	DE	Stuttgart	Stuttgart
F	22	anniwhat	FR	Bad Bellingen	Bad Bellingen
F	20	Pierrette	CH	Basel	Liestal
F	22	Nathalie	DE	Neu-Ulm	Holzheim

Slika 4.6. Prozor za vraćanje originalnih vrijednosti atributa

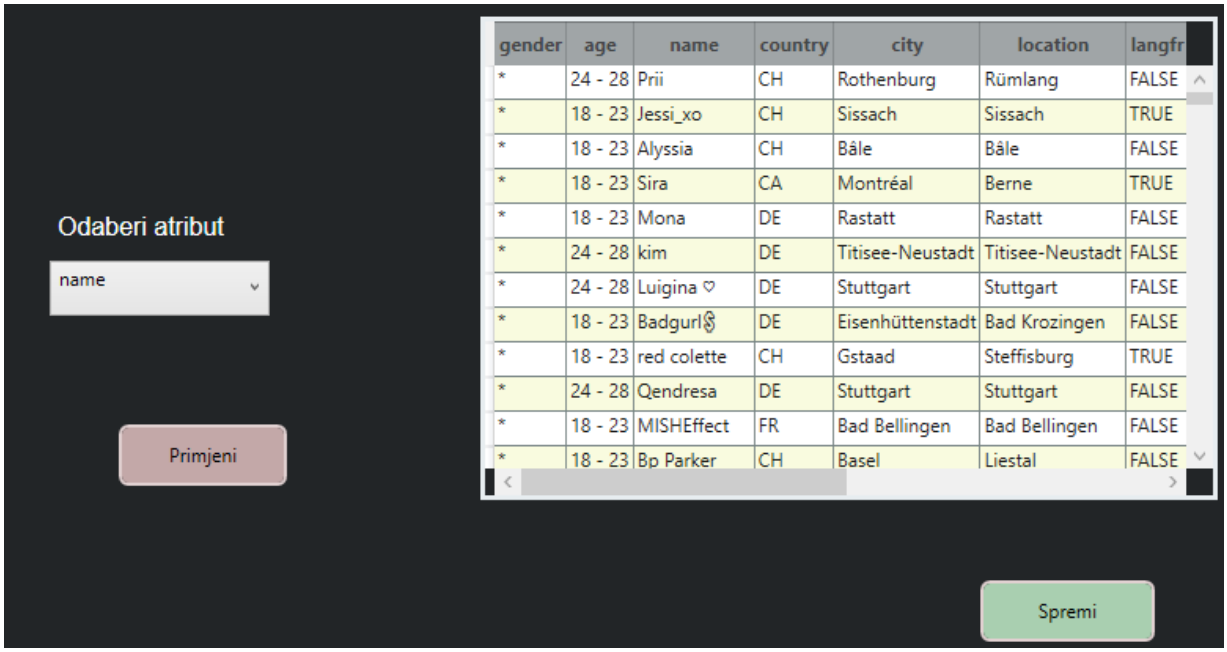
Na slici 4.7. vidljiv je rezultat generalizacije gdje su dobivena dva raspona, prvi od 18 do 23, a drugi od 24 do 28. Drugi raspon sadrži samo četiri vrijednosti, a ne pet jer je najveća vrijednost 28. Generalizacija, kao i perturbacija, izvodi se samo nad brojčanim atributima.



gender	age	name	country	city	location
F	24 - 28	daeni	CH	Rothenburg	Rümlang
F	18 - 23	italiana 92	CH	Sissach	Sissach
F	18 - 23	Lauraaa	CH	Bâle	Bâle
F	18 - 23	Qqkwmdowlo	CA	Montréal	Berne
F	18 - 23	schaessie {3	DE	Rastatt	Rastatt
F	24 - 28	Baby dee	DE	Titisee-Neustadt	Titisee-Neustadt
F	24 - 28	Anna	DE	Stuttgart	Stuttgart
F	18 - 23	Feelinchen	DE	Eisenhüttenstadt	Bad Krozingen
F	18 - 23	chantal	CH	Gstaad	Steffisburg
F	24 - 28	rina1990	DE	Stuttgart	Stuttgart
F	18 - 23	anniwhat	FR	Bad Bellingen	Bad Bellingen
F	18 - 23	Pierrette	CH	Basel	Liestal

Slika 4.7. Prozor za generalizaciju

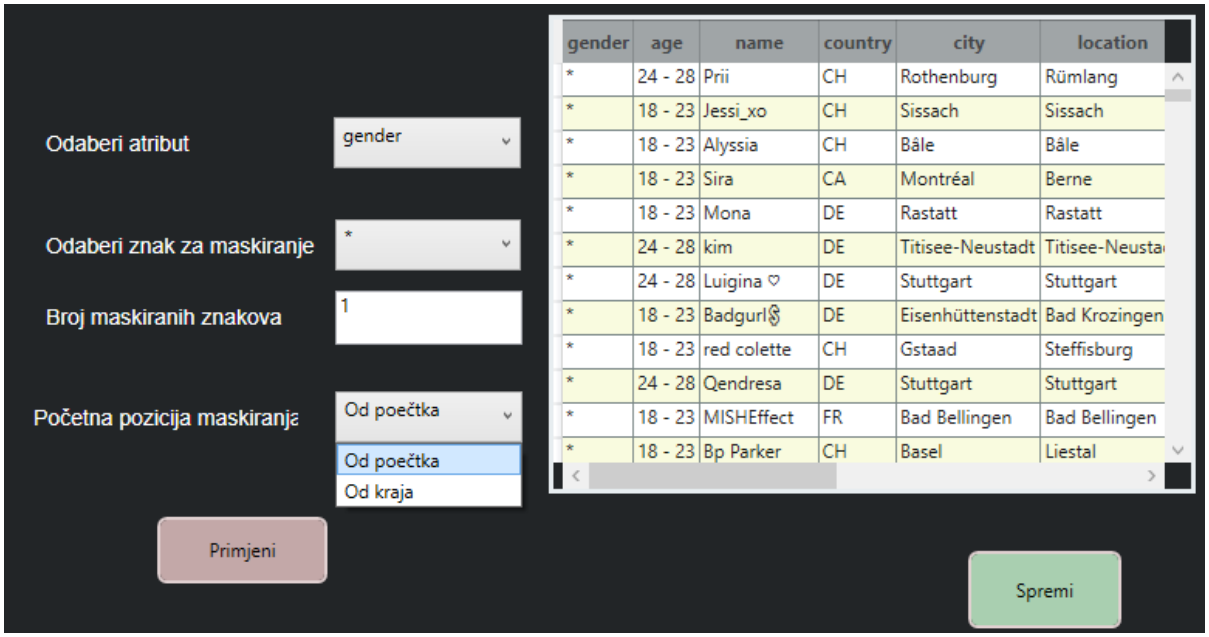
Nadalje, nad korisničkim imenima izvedena je permutacija jer korisničko ime može koristiti u reidentifikaciji, a za potrebnu analizu nije potrebno. Rezultat primjene nalazi se na slici 4.8.



gender	age	name	country	city	location	langfr
*	24 - 28	Prii	CH	Rothenburg	Rümlang	FALSE
*	18 - 23	Jessi_xo	CH	Sissach	Sissach	TRUE
*	18 - 23	Alyssia	CH	Bâle	Bâle	FALSE
*	18 - 23	Sira	CA	Montréal	Berne	TRUE
*	18 - 23	Mona	DE	Rastatt	Rastatt	FALSE
*	24 - 28	kim	DE	Titisee-Neustadt	Titisee-Neustadt	FALSE
*	24 - 28	Luigina ♡	DE	Stuttgart	Stuttgart	FALSE
*	18 - 23	Badgurl$	DE	Eisenhüttenstadt	Bad Krozingen	FALSE
*	18 - 23	red colette	CH	Gstaad	Steffisburg	TRUE
*	24 - 28	Qendresa	DE	Stuttgart	Stuttgart	FALSE
*	18 - 23	MISHEffect	FR	Bad Bellingen	Bad Bellingen	FALSE
*	18 - 23	Bp Parker	CH	Basel	Liestal	FALSE

Slika 4.8. Prozor za permutaciju

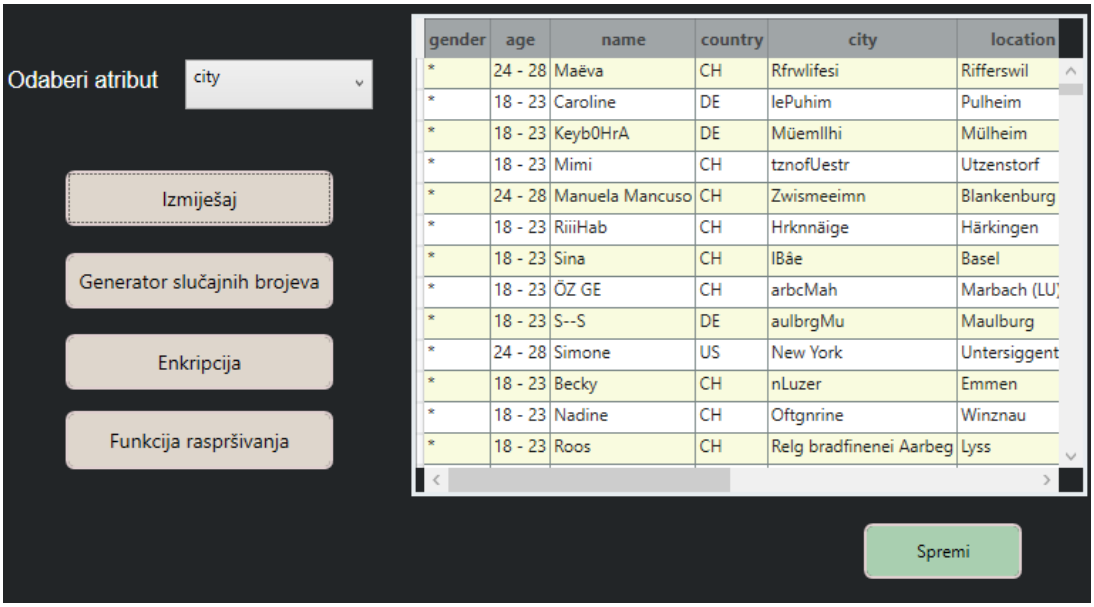
Daljnji korak u anonimizaciji jest maskiranje spola što je dano na slici 4.9. Budući da spol sadrži samo jedan znak, nije važno hoće li se početna pozicija maskiranja biti od početka ili od kraja. Također, ukoliko se broj maskiranih znakova unese veći nego li što vrijednost sadrži znakova, program tada postavlja sve znakove vrijednosti na znak maskiranja, kako bi duljina vrijednosti i dalje ostala ista, ali maskirana.



gender	age	name	country	city	location
*	24 - 28	Prii	CH	Rothenburg	Rümlang
*	18 - 23	Jessi_xo	CH	Sissach	Sissach
*	18 - 23	Alyssia	CH	Bâle	Bâle
*	18 - 23	Sira	CA	Montréal	Berne
*	18 - 23	Mona	DE	Rastatt	Rastatt
*	24 - 28	kim	DE	Titisee-Neustadt	Titisee-Neusta
*	24 - 28	Luigina ♡	DE	Stuttgart	Stuttgart
*	18 - 23	Badgurl$	DE	Eisenhüttenstadt	Bad Krozingen
*	18 - 23	red colette	CH	Gstaad	Steffisburg
*	24 - 28	Qendresa	DE	Stuttgart	Stuttgart
*	18 - 23	MISHEffect	FR	Bad Bellingen	Bad Bellingen
*	18 - 23	Bp Parker	CH	Basel	Liestal

Slika 4.9. Prozor za maskiranje

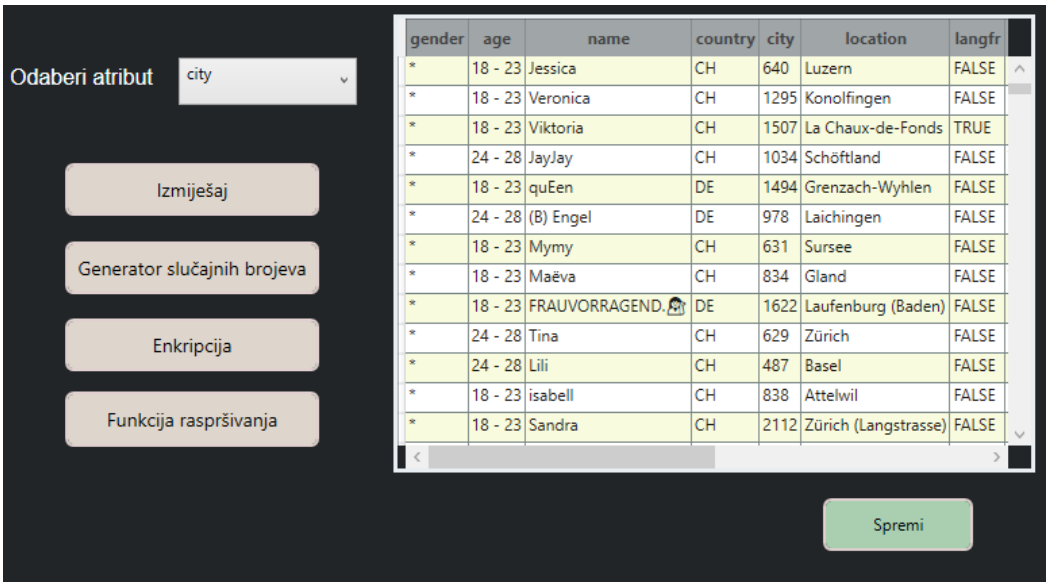
Nadalje, grad je atribut čija vrijednost može znatno naštetiti privatnosti korisnika. Na grad je primijenjena pseudonimizacija. *Simple Anonymizer* nudi opciju *izmiješaj* kao način dobivanja pseudonima, ali obzirom na to da se u podatkovnom skupu nalaze nazivi država, napadaču neće biti teško otkriti o kojem se gradu radi. Rezultat bi izgledao kao na slici 4.10. Od ostalih opcija za stvaranje pseudonima, u ovome slučaju svejedno je koja će se odabrati.



gender	age	name	country	city	location
*	24 - 28	Maëva	CH	Rfrwlifesi	Rifferswil
*	18 - 23	Caroline	DE	IePuhim	Pulheim
*	18 - 23	Keyb0HrA	DE	MüemlIhi	Mülheim
*	18 - 23	Mimi	CH	tznofUestr	Utzenstorf
*	24 - 28	Manuela Mancuso	CH	Zwismeeimn	Blankenburg
*	18 - 23	RiiiHab	CH	Hrknnäige	Härkingen
*	18 - 23	Sina	CH	IBåe	Basel
*	18 - 23	ÖZ GE	CH	arbcMah	Marbach (LU)
*	18 - 23	S--S	DE	aulbrgMu	Maulburg
*	24 - 28	Simone	US	New York	Untersiggent
*	18 - 23	Becky	CH	nLuzer	Emmen
*	18 - 23	Nadine	CH	Oftgnrine	Winznau
*	18 - 23	Roos	CH	Relg bradfinenei Aarbeg	Lyss

Slika 4.10. Pseudonimizacija - izmiješaj

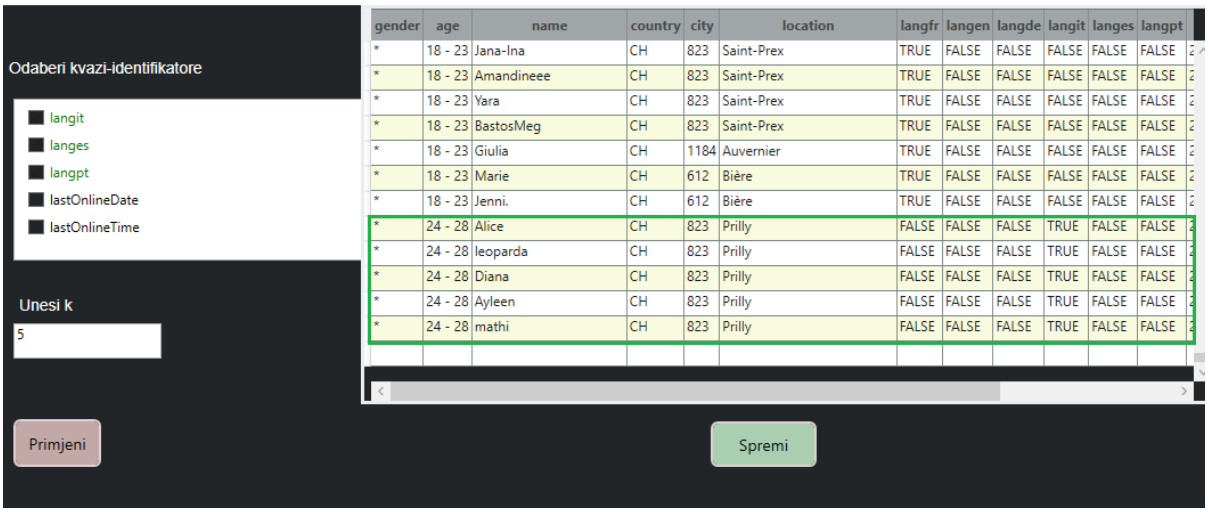
Enkripcija je nešto nesigurnija jer postoji mogućnost da napadač otkrije ključ enkripcije. Odabrana tehnika je funkcija raspršivanja, a kako to izgleda primijenjeno, vidljivo je na slici 4.11. Idući korak je primijeniti *k*-anonimizaciju. *Simple Anonymizer* radi tako da za odabrani *k*, zapise kojih ima *k* ili više ostavlja u tablici, a one kojih ima manje, izbriše.



gender	age	name	country	city	location	langfr
*	18 - 23	Jessica	CH	640	Luzern	FALSE
*	18 - 23	Veronica	CH	1295	Konolfingen	FALSE
*	18 - 23	Viktoria	CH	1507	La Chaux-de-Fonds	TRUE
*	24 - 28	JayJay	CH	1034	Schöftland	FALSE
*	18 - 23	quEen	DE	1494	Grenzach-Wyhlen	FALSE
*	24 - 28	(B) Engel	DE	978	Laichingen	FALSE
*	18 - 23	Mymy	CH	631	Sursee	FALSE
*	18 - 23	Maëva	CH	834	Gland	FALSE
*	18 - 23	FRAUVORRAGEND.	DE	1622	Laufenburg (Baden)	FALSE
*	24 - 28	Tina	CH	629	Zürich	FALSE
*	24 - 28	Lili	CH	487	Basel	FALSE
*	18 - 23	isabell	CH	838	Attelwil	FALSE
*	18 - 23	Sandra	CH	2112	Zürich (Langstrasse)	FALSE

Slika 4.11. Pseudonimizacija - funkcija raspršivanja

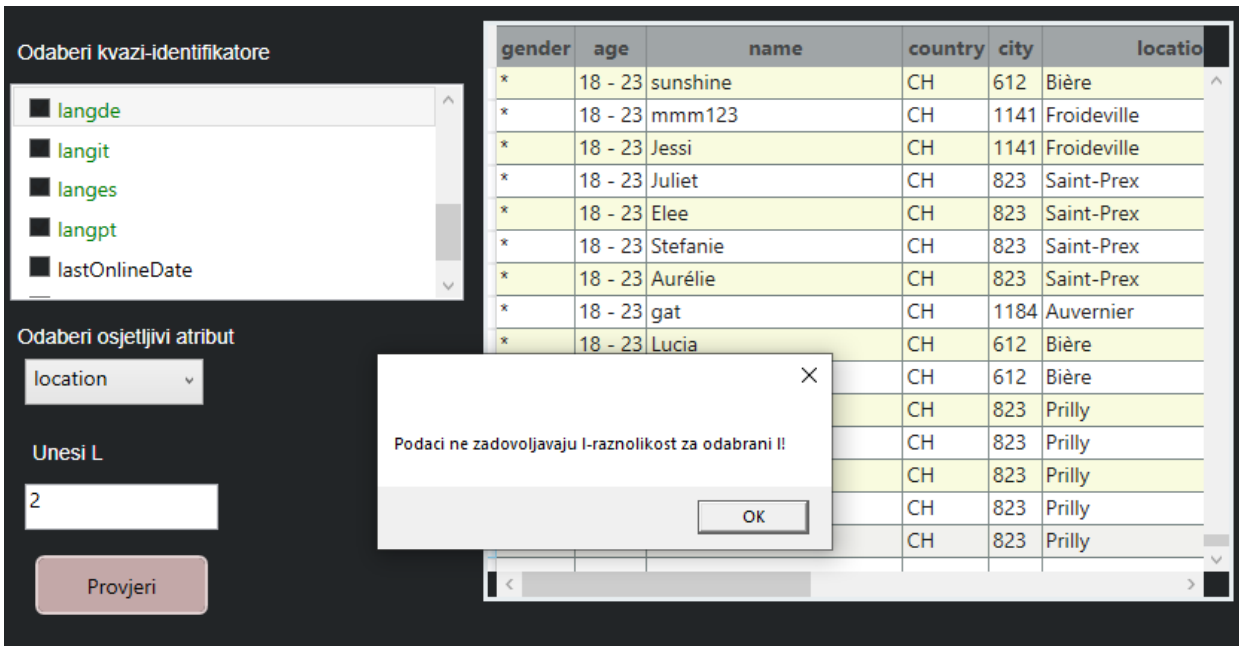
Budući da je spol maskiran te na naziv grada primijenjena funkcija raspršivanja, više ih ne moramo uzimati u obzir kao kvazi-identifikatore. Na slici 4.12. odabrani *k* je 5 te je zeleno uokviren jedan ekvivalentni razred koji se sastoji od 5 zapisa. Obzirom na to da je većina korisnika iz Njemačke, a samim time je i jezik koji govore njemački, izabran je manji *k* iz razloga što je potrebno da zapisi sadrže i što raznolikije vrijednosti.



gender	age	name	country	city	location	langfr	langen	langde	langit	langes	langpt
*	18 - 23	Jana-Ina	CH	823	Saint-Prex	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
*	18 - 23	Amandineee	CH	823	Saint-Prex	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
*	18 - 23	Yara	CH	823	Saint-Prex	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
*	18 - 23	BastosMeg	CH	823	Saint-Prex	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
*	18 - 23	Giulia	CH	1184	Auvernier	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
*	18 - 23	Marie	CH	612	Bière	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
*	18 - 23	Jenni.	CH	612	Bière	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
*	24 - 28	Alice	CH	823	Prilly	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
*	24 - 28	leoparda	CH	823	Prilly	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
*	24 - 28	Diana	CH	823	Prilly	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
*	24 - 28	Ayleen	CH	823	Prilly	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
*	24 - 28	mathi	CH	823	Prilly	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE

Slika 4.12. Prozor za *k*-anonimizaciju

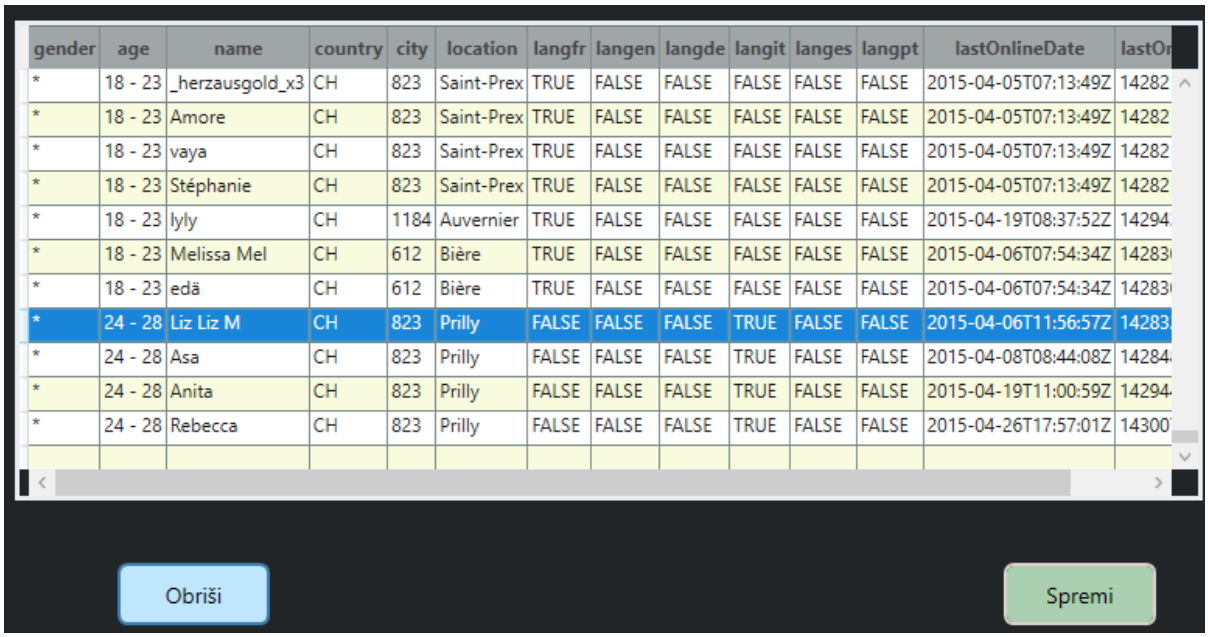
Slijedi provjera *l*-raznolikosti. Najmanje potrebno jest da *l* barem bude 2, što se na slici 4.13. vidi da podatkovni skup ne zadovoljava. Kako bi zadovoljio, potrebno je pronaći ekvivalentni razred koji ne zadovoljava raznolikost. Taj razred uokviren je na slici 4.12. te ga je potrebno ukloniti. To se čini tako da se odabere opcija *Uklanjanje zapisa* koja se može vidjeti na slici 4.3.



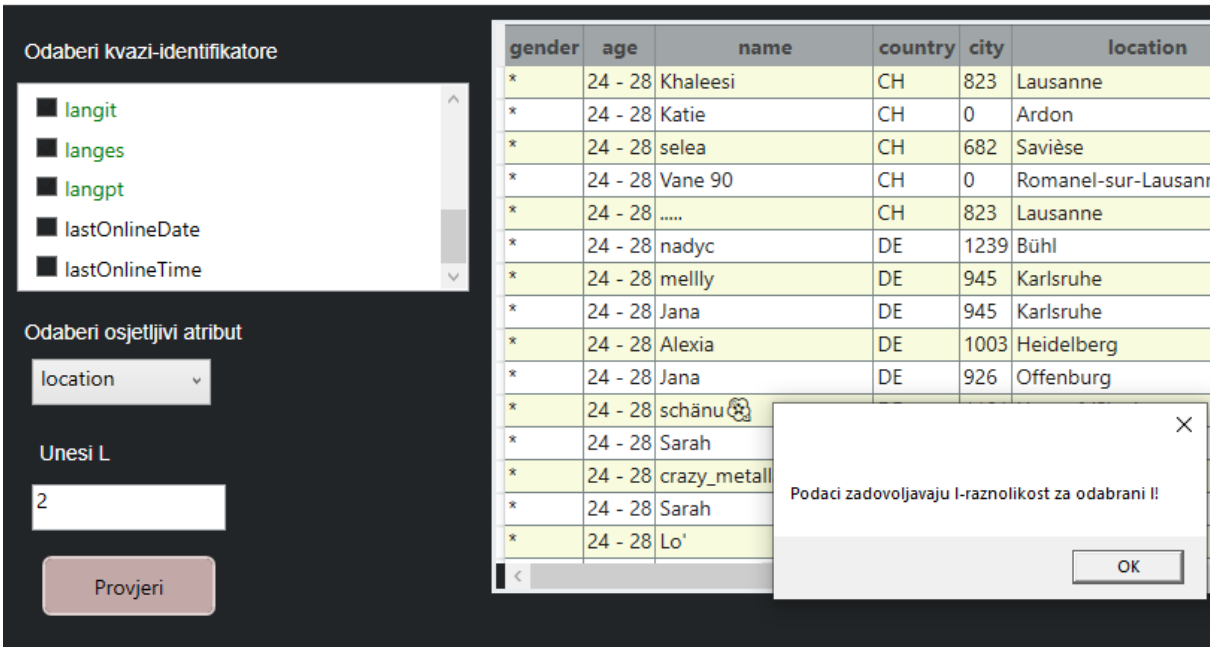
gender	age	name	country	city	locatio
*	18 - 23	sunshine	CH	612	Bière
*	18 - 23	mmm123	CH	1141	Froideville
*	18 - 23	Jessi	CH	1141	Froideville
*	18 - 23	Juliet	CH	823	Saint-Prex
*	18 - 23	Elee	CH	823	Saint-Prex
*	18 - 23	Stefanie	CH	823	Saint-Prex
*	18 - 23	Aurélie	CH	823	Saint-Prex
*	18 - 23	gat	CH	1184	Auvernier
*	18 - 23	Lucia	CH	612	Bière
			CH	612	Bière
			CH	823	Prilly
			CH	823	Prilly
			CH	823	Prilly
			CH	823	Prilly
			CH	823	Prilly

Slika 4.13. Prozor za *l*-raznolikost

Zatim se odabere zapis iz prikaza popisa (engl. *list view*) kako je prikazano na slici 4.14. te klikom na gumb *Obriši*, zapis se izbriše. Nakon što se na taj način uklone svi željeni zapisi, klikne se na gumb *Spremi*. Uklanjanje zapisa, kao i atributa, najefikasnija je tehnika, ali podrazumijeva potpuni gubitak informacija. Ukoliko se sada provjerava zadovoljavaju li podaci *l*-raznolikost, rezultat je prikazan na slici 4.15.

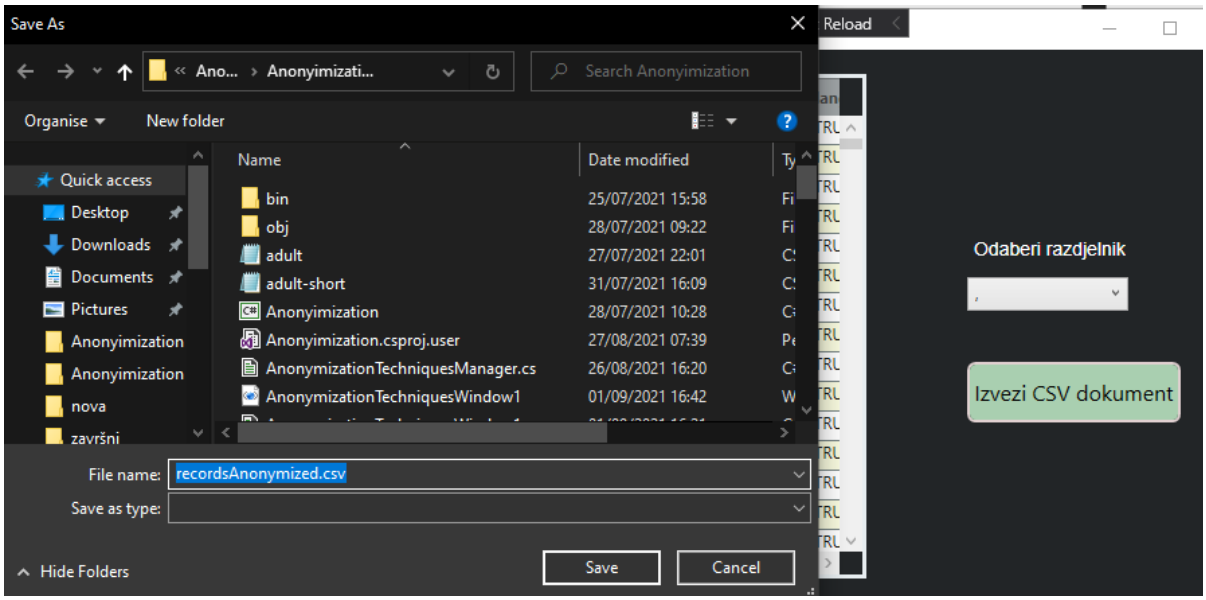


Slika 4.14. Prozor za uklanjanje zapisa



Slika 4.15. Prozor zadovoljena *2*-raznolikost

Za kraj, takvi se anonimirani podaci mogu spremiti. Klikom na gumb *Dalje* vidljivog na slici 4.3., otvara se novi prozor prikazan na slici 4.16. Odabire se razdjelnik koji će se upisivati u CSV datoteku te naziv datoteke u dijaloškom okviru koji se otvara na klik gumba Izvezi.



Slika 4.16. Prozor za spremanje anonimiziranih podataka

5. ZAKLJUČAK

Tehnike anonimizacije i modeli privatnosti, odnosno, mjerila anonimizacije uvelike pridonose privatnosti i zaštiti podataka. U radu je dan pregled najčešće korištenih tehnika anonimizacije, od kojih neke pružaju veću zaštitu u smislu nepovratnosti originalnih vrijednosti, a u drugu ruku pružaju veći ili potpuni nedostatak informacija. Podatkovni skupovi nad kojima se provodi anonimizacija, mogu biti prikupljani metodom nabave iz mnoštva. Isto tako, pojedinac može pridonijeti rješavanju nekog problema putem mobilne nabave iz mnoštva koja podrazumijeva uključenost mobilnih uređaja. U tom slučaju, korisnik daje svoje osobne podatke kako bi dobio određenu nagradu, a u pozadini dok daje svoje rješenje problema, mnogi drugi osobni podaci mogu biti prikupljani, a koji se kasnije mogu iskoristiti za ugrožavanje privatnosti korisnika. Opis rada mobilnih platformi kao izvora podataka, slojevi te oblici arhitekture opisani su u radu. Iako su mobilni uređaji idealna platforma za prikupljanje podataka, zbog pristupa Internetu i bogatstva senzora koji nude kvalitetne podatke, u isto vrijeme ti podaci mogu dovesti do reidentifikacije te saznanja o trenutnoj lokaciji pojedinca. Dakle, još u procesu nabave podataka potrebno je postaviti ograničenja. Na koji god način da su podaci prikupljeni, sve dok postoje negdje kao javno dostupan skup podataka, unatoč tome što su anonimizirani raznim tehnikama, napadači i dalje pronalaze načine za reidentifikaciju zajedno s vještinom hakiranja ili s pozadinskim znanjima. U tu svrhu, implementirano je programsko rješenje za anonimizaciju podataka koja nudi učitavanje podataka, primjenu tehnika anonimizacije navedenih u radu te izvoz anonimiziranog podatkovnog skupa. Primjena tehnika ne mora značiti da je podatkovni skup zaista anonimiziran, već može značiti da je pseudonimiziran, što znači da su podaci i dalje podložni reidentifikaciji. Niti zadovoljavanje mjerila anonimizacije u današnje vrijeme više ne osigurava anonimizaciju. Zaključak je da podaci ne mogu u isto vrijeme biti korisni i pružati potpunu zaštitu pojedinca uz dosad spomenute načine. Diferencijalna privatnost noviji je pristup koji je matematički temeljen i sve više se koristi od strane velikih tvrtki poput *Google-a* i *Apple-a* te je trenutno najučinkovitiji algoritam za očuvanje privatnosti. Unaprjeđenje izraženog programskog rješenja moguće je kroz programsku implementaciju diferencijalne privatnosti te načini za njeno pojednostavljivanje.

LITERATURA

[1] D. C. Brabham, Massachusetts, SAD, Crowdsourcing, MIT Press, Massachusetts, SAD, 2013.

[2] J. Phuttharak i S. W. Loke, A Review of Mobile Crowdsourcing Architectures and Challenges: Toward Crowd-Empowered Internet-of-Things, IEEE Access, svez. 7, str. 304-324, prosinac 2018.

[3] Y. Wang, X. Jia, J. Qun i J. Ma, Mobile crowdsourcing: framework, challenges, and solutions, Concurrency and Computation: Practice and Experience, svez. 29, br. 3, veljača 2016.

[4] B. Liu, W. Zhong, J. Xie, L. Kong, Y. Yang, C. Lin i H. Wang, Deep Learning for Mobile Crowdsourcing Techniques, Methods, and Challenges: A Survey, Mobile Information Systems, svez. 2021, siječanj 2021.

[5] S. Sarin, K. Pipatsrisawat, K. Pham, A. Batra i L. Valente, Crowdsource by Google: A Platform for Collecting Inclusive and Representative Machine Learning Data, u AAAI Conference on Human Computation and Crowdsourcing, Washington, SAD, 28.-30. listopad 2019.

[6] N. Haderer, R. Rouvoy i L. Seinturier, Dynamic Deployment of Sensing Experiments in the Wild Using Smartphones, u IFIP International Conference on Distributed Applications and Interoperable Systems, str. 43-56, Firenca, Italija, 3.-5. lipanj 2013.

[7] Personal Data Protection Commission, Guide to basic data anonymisation techniques [online], Bloomberg, Singapore, 2018., dostupno na: https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/Guide-to-Anonymisation_v1-(250118).pdf [15. lipanj 2021.]

[8] Bryan Cave Leighton Paisner, At A Glance: De-Identification, Anonymization, and Pseudonymization under the GDPR [online], BCLP, 2017., dostupno na: https://www.bclplaw.com/en-US/insights/at-a-glance-de-identification-anonymization-and-pseudonymization-1.html [12. lipanj 2021].

[9] A. Bradić-Martinović i A. Zdravković, Zaštita privatnosti – anonimizacija podataka, u V naučni skup USPON, str. 206-213, Beograd, Srbija, 28. studeni, 2013.

[10] The European Union Agency for Cybersecurity (ENISA), Pseudonymisation techniques and best practices [online], ENISA, 3. prosinca 2019., dostupno na: https://op.europa.eu/en/publication-detail/-/publication/8c53ec8c-170f-11ea-8c1f-01aa75ed71a1 [20. lipanj 2021.]

[11] K. E. Emam i F. K. Dankar, Protecting Privacy Using k-Anonymity, Journal of the American Medical Informatics Association, svez. 15, br. 5, str. 627-637, listopad 2008.

[12] N. Li, T. Li i S. Venkatasubramanian, t-Closeness: Privacy Beyond k-Anonymity and l-Diversity, u 2007 IEEE 23rd International Conference on Data Engineering, str. 106-115, Istanbul, Turska, 15.-16. travnja 2007.

[13] J. B. Abdo, T. Bourgeau, J. Demerjian i H. Chaouchi, Extended Privacy in Crowdsourced Location-Based Services Using Mobile Cloud Computing, Mobile Information Systems, svez. 2016., str. 1-13, 31. Srpanj 2016.

[14] E. Devaux, How anonymous is anonymous [online], Medium, 2020, dostupno na: https://medium.com/statice/how-anonymous-is-anonymous-c92ad265a3e3. [10. Lipanj 2021].

SAŽETAK

Mobilna nabava iz mnoštva sve više je korištena od strane tvrtki koje žele koristiti vanjska poduzeća i pojedinca za obavljanje posla. Razlog tomu je što mobilna nabava iz mnoštva pruža prijenos podataka u stvarnom vremenu te jednostavniji način prikupljanja istih. Podaci prikupljeni na taj način koriste se u svrhu rješavanja problema ili raznih analiza, ali problem je što ti podaci mogu sadržavati osjetljive informacije o pojedincu. Stoga, nad tim podacima primjenjuju se tehnike anonimizacije. U praktičnom dijelu ovog rada izrađen je alat koji nudi primjenu tih tehnika nad učitanim podacima. Alat implementira modele privatnosti poput *k*-anonimizacije i *l*-raznolikosti. Nakon što se nad podacima primjene tehnike anonimizacije te se utvrdi da zadovoljavaju modele privatnosti, korisnik i dalje ne može biti siguran od opasnosti reidentifikacije. Do problema dolazi kada se mora birati između očuvanja privatnosti pojedinca i kvalitete podataka. Sigurnije tehnike po pitanju zaštite privatnosti podrazumijevaju veći gubitak informacija.

Ključne riječi: anonimizacija, mobilna nabava iz mnoštva, privatnost, tehnike anonimizacije

Data anonymization approaches for mobile crowdsourced data

ABSTRACT

Mobile crowdsourcing is increasingly being used by companies that want to outsource work. The reason for that is that mobile crowdsourcing provides real-time data transmission. Data collected that way is used for solving a problem or for various types of analyses, but problem lies in the fact that data may contain sensitive information about an individual. Therefore, anonymization techniques can be applied on data. As practical part of this thesis, a tool is made which provides application of these techniques on uploaded data. Tool implements models of privacy, such as *k*-anonymizationa and *l*-diversity. After anonymization techniques are applied on the data and it is concluded that the data meets models of privacy, user still can not be protected from danger of reidentification. The problem arises when a choice has to be made between protecting privacy and ensuring the quality of the data. More secure techniques regarding privacy protection involve greater loss of information.

Key words: anonymization, mobile crowdsourcing, privacy, anonymization techniques

ŽIVOTOPIS

Martina Damjanović rođena je u Osijeku, 20. listopada, 1999. godine. Osnovnu školu pohađala je u Osijeku. Srednjoškolsko obrazovanje stekla je u III. gimnaziji u Osijeku sa završetkom u 2018. godini. Iste godine upisala je preddiplomski sveučilišni studij računarstva na fakultetu elektrotehnike, računarstva i informacijskih tehnologija Osijek. Odradila je back-end praksu u Factory-u.

PRILOZI

1. Završni rad u formatu .docx
2. Završni rad u formatu .pdf
3. Izvorni kod programskog rješenja