Anonimizacija podataka pohranjenih u relacijske baze podataka

Jerešić, Helena

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Helena Jerešić

ANONIMIZACIJA PODATAKA POHRANJENIH U RELACIJSKE BAZE PODATAKA

Diplomski rad

Voditelj rada:
dr. sc. Ognjen Orel

Zagreb, 2025.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Sadržaj

Sadržaj **iii**

Uvod **1**

1 Osnovni pojmovi i koncepti **3**
1.1 Osobni podaci i anonimizacija 4

2 Anonimizacijske tehnike **7**
2.1 Supstitucija 7
2.2 Miješanje 8
2.3 Dodavanje šuma 9
2.4 Poništavanje 10
2.5 Maskiranje simbolom 11
2.6 Kriptografija 12
2.7 Generalizacija 13

3 Alati za anonimizaciju podataka **17**
3.1 ARX 17
3.2 μ-ARGUS 18
3.3 SDCMicro 19
3.4 Amnesia 21

4 Aplikacija za anonimizaciju podataka u relacijskim bazama podataka **23**
4.1 Demonstracijska baza podataka 25
4.2 Prikaz mogućnosti aplikacije 26
4.3 Tehnička implementacija aplikacije 33
4.4 Mogućnosti unaprjeđenja aplikacije 37

Zaključak **39**

Bibliografija **41**

Uvod

Ovaj diplomski rad bavi se anonimizacijom podataka, s posebnim naglaskom na relacijske baze podataka. Rad započinje detaljnim objašnjenjem osnovnih pojmova i koncepta vezanih uz anonimizaciju, uz jasno definiranu razliku između anonimizacije i pseudonimizacije podataka. Nadalje, analiziraju se postojeće tehnike anonimizacije, dostupni alati koji se koriste u praksi te se na kraju predstavlja vlastita implementacija aplikacije za anonimizaciju relacijskih baza podataka.

Prvo poglavlje usmjereno je na temeljne definicije osobnih podataka i pojašnjenje ključnih pojmova poput anonimizacije i pseudonimizacije. Kako bi se ove metode bolje razumjele, prikazana je njihova razlika na konkretnom primjeru, uz objašnjenje kako pseudonimizacija smanjuje rizik reidentifikacije, ali je ne uklanja u potpunosti, dok anonimizacija nudi visoku razinu zaštite privatnosti, ali može dovesti do gubitka korisnih informacija.

Drugo poglavlje donosi pregled najčešće korištenih tehnika anonimizacije u današnjoj praksi. Objašnjene su metode poput supstitucije, miješanja, dodavanja šuma, poništavanja, maskiranja simbolima, primjene kriptografskih tehnika te generalizacije podataka. Svaka od ovih metoda ima svoje prednosti i ograničenja, ovisno o kontekstu u kojem se primjenjuje i razini privatnosti koju je potrebno osigurati.

Treće poglavlje istražuje trenutno dostupne alate za anonimizaciju podataka. Analizirani su popularni alati poput ARX, μ-ARGUS, SDCMicro i Amnesia, uz detaljniji opis njihovih funkcionalnosti, prednosti i nedostataka. Ovi alati omogućuju različite stupnjeve anonimizacije, a njihova primjena ovisi o specifičnim potrebama korisnika i vrsti podataka koji se obrađuju.

Četvrto poglavlje fokusira se na praktični dio diplomskog rada, a odnosi se na razvoj vlastite aplikacije za anonimizaciju relacijskih baza podataka pod nazivom AnonyDB. U ovom poglavlju detaljno se opisuje proces razvoja aplikacije, korištene tehnologije te se na primjerima prikazuju rezultati anonimizacije.

Poglavlje 1

Osnovni pojmovi i koncepti

U suvremenom svijetu, gdje se svakodnevno razmjenjuju goleme količine podataka, zaštita osobnih podataka postala je jedna od ključnih odgovornosti. Anonimizacija podataka, posebice u relacijskim bazama podataka, igra bitnu ulogu u očuvanju privatnosti korisnika i sprječavanju zloupotrebe osjetljivih informacija.

Jedan od najvećih incidenta curenja podataka u povijesti dogodio se 2013. godine, kada je američka tvrtka Yahoo! doživjela masovni sigurnosni propust. Tada su osjetljivi podaci više od milijardu korisnika, uključujući imena, e-mail adrese, telefonske brojeve, datume rođenja, adrese, kriptirane lozinke, postali dostupni neovlaštenim stranama [1]. Ovaj događaj podsjeća nas koliko je važno implementirati učinkovite mjere zaštite.

Kao odgovor na sve veće sigurnosne prijetnje i potrebu za regulacijom, Europska unija je 2018. godine uvela Opću uredbu o zaštiti podataka, poznatiju kao GDPR. Cilj ove uredbe je zaštititi osobne podatke građana Europske unije, omogućiti im veću kontrolu nad vlastitim podacima te osigurati visoku i ujednačenu razinu zaštite podataka unutar svih zemalja članica. Uredba određuje koja su prava pojedinaca, a u skladu s tim i koje su obveze organizacija koje obrađuju osobne podatke. Zaštita osjetljivih podataka važna je obveza nametnuta regulativama kako bi se spriječio pristup neovlaštenim osobama, odnosno programerima, testerima te bilo kojim drugim osobama kojima poslovni procesi ne zahtijevaju pristup tim podacima.

Kako je u ovom diplomskom radu naglasak na anonimizaciji baze podataka, potrebno je najprije definirati pojam baze i anonimizacije podataka. Baza podataka je skup međusobno povezanih podataka pohranjenih u vanjskoj memoriji računala. Posebno, relacijski model je zasnovan na matematičkom pojmu relacije. I podaci i veze među podacima prikazuju se tablicama koje se sastoje od redaka i stupaca [2]. Anonimizacija podataka je proces uklanjanja ili skrivanja identifikacijskih informacija u osjetljivim podacima, dok se istovremeno čuva njihov format i tip podataka [3]. Proces anonimizacije pruža dodatnu sigurnosnu razinu, jer i u slučaju krađe anonimiziranih podataka, oni ne mogu biti iskorišteni ili

zloupotrijebljeni.

1.1 Osobni podaci i anonimizacija

Osobni podaci su svi podaci koji se odnose na pojedinca čiji je identitet utvrđen ili se može utvrditi. Pojedinac čiji se identitet može utvrditi jest osoba koja se može identificirati izravno ili neizravno, osobito uz pomoć identifikatora kao što su ime, identifikacijski broj, podaci o lokaciji, mrežni identifikator ili uz pomoć jednog ili više čimbenika svojstvenih za fizički, fiziološki, genetski, mentalni, ekonomski, kulturni ili socijalni identitet tog pojedinca [4]. Upravo tako definirani osobni podatak u Općoj uredbi o zaštiti podataka članku 4. daje nam jasniju sliku o tome koje podatke moramo zaštiti.

Osjetljivi osobni podaci su podskup osobnih podataka koji zahtijevaju strožu zaštitu jer bi njihova zloupotreba mogla ozbiljno narušiti privatnost ili prava pojedinca. U to spadaju rasa ili etnička pripadnost, politički stavovi, vjerska uvjerenja, sindikalno članstvo, podaci o zdravlju ili spolnom životu te osobni podaci o kaznenim i prekršajnim postupcima. Takvi podaci podložni su strožim pravilima obrade i zaštite kako bi se spriječila njihova zloupotreba.

Ista uredba u razlogu 26. kaže kako podaci koji su potpuno anonimizirani tako da više ni na koji način nije moguće utvrditi identitet pojedinca ne podliježu pravilima te uredbe. Tako se omogućava obrada anonimiziranih podataka, u statističkim ili istraživačkim svrhama, bez ugrožavanja prava pojedinaca. Važno je napomenuti da se uredba ne fokusira na to kako se anonimizacija podataka mora ili treba provesti, već na krajnji rezultat anonimizacije. Ključno je da rezultat bude nepovratan, odnosno da anonimizacija osigura da identitet pojedinca više nije moguće utvrditi ni na koji način.

Kako bi u nastavku bolje razumjeli razliku između pseudoanonimizacije i anonimizacije postupke ćemo primijeniti na podatke iz tablice 1.1 koja sadrži osobne podatke studenata.

ID	JMBAG	IME	PREZIME	EMAIL ADRESA
1	12345678910	Ana	Jurić	anajur.math@pmf.hr
2	12345678911	Marko	Horvat	markhorv.math@pmf.hr
3	12345678912	Ivan	Kovačić	ivakova.math@pmf.hr

Tablica 1.1: Tablica *Studenti*

Pseudoanonimizacija

Pseudoanonimizacija je postupak kojim osobne podatke mijenjamo tako kasnije nije moguće direktno odrediti identitet pojedinca bez korištenja dodatnih informacija. Dodatne

informacije trebalo bi čuvati odvojeno i zaštićeno kako bi se smanjio rizik od otkrivanja identiteta pojedinca. Međutim, takav postupak ne daje potpunu zaštitu podataka. Ako dođe do neovlaštenog pristupa dodatnim informacija, moguća je reidentifikacija pojedinca. Iako korištenjem postupaka pseudoanonimizacije smanjujemo rizik reidentifikacije, njima ne eliminiramo rizike u potpunosti.

ID	JMBAG	PSEUDOIME	PSEUDOPREZIME	EMAIL ADRESA
1	21098765434	ime1	prezime1	dqdmxu.pdwk@spi.ku
2	21098765435	ime2	prezime2	pdunkruy.pdwk@spi.ku
3	21098765436	ime3	prezime3	lydnryd.pdwk@spi.ku

Tablica 1.2: Tablica *Studenti* nakon pseudoanonimizacije

PSEUDOIME	IME
ime1	Ana
ime2	Marko
ime3	Ivan

Tablica 1.3: Dodatne informacije o imenima

PSEUDOIME	PREZIME
prezime1	Jurić
prezime2	Horvat
prezime3	Kovačić

Tablica 1.4: Dodatne informacije o prezimenima

U pseudoanonimizaciji, podaci poput JMBAG-a i email adrese izmijenjeni su pomoću Cezarove šifre s pomakom od tri. To znači da su znamenke i slova u tim podacima pomaknuta za tri mjesta, stvarajući tako nove kriptirane vrijednosti. Dodatnim tablicama povezali smo pseudoimena i pseudoprezimena s pravim podacima. Iako smo promijenili izvorne podatke, podaci uz nekoliko dodatnih informacija poput metode kriptografije ili dodatnih tablica lako mogu dovesti do reidentifikacije identiteta pojedinca.

Anonimizacija

Anonimizacija je postupak čiji je cilj osobne podatke promijeniti tako da kasnije nije moguće ni na koji način niti bilo kojim dodatnim informacija odrediti identitet pojedinca. Iako anonimizacija teži potpunoj zaštiti osobnih podataka, veliki njezin nedostatak je gubitak informacija o pojedincu koji može dovesti do gubitka točnosti i iskoristivosti podataka za određene svrhe. Postupak anonimizacije složen je i vremenski zahtjevan zadatak, osobito kod velikih količina podataka, te treba uzeti u obzir faktor ljudske pogreške. Iako je mogućnost za reidentifikaciju mala, ona i dalje postoji.

U postupku anonimizacije podataka uklonili smo identifikacijske informacije poput imena i prezimena, dok smo JMBAG i email adrese ostavili u izvornom formatu, ali maskirali sve vrijednosti s generičkim znakom '*'.

ID	JMBAG	IME	PREZIME	EMAIL ADRESA
1	**********	-	-	******.****@***.**
2	**********	-	-	********.****@***.**
3	**********	-	-	*******.****@***.**

Tablica 1.5: Tablica *Studenti* nakon anonimizacije

Ovaj pristup osigurava visoku zaštitu privatnosti, jer onemogućuje bilo kakvu reidentifikaciju pojedinca, no istovremeno dolazi do gubitka korisnih informacija. Kao rezultat, podaci su postali neiskoristivi za analize ili određene svrhe.

Poglavlje 2

Anonimizacijske tehnike

Kako bismo ispravno odabrali tehnike anonimizacije koje ćemo primijeniti, potrebno je razumjeti prednosti i nedostatke svake od tehnika, jer one imaju različite karakteristike, a izbor ovisi o kompromisima na koje smo spremni pristati. Ove tehnike mogu biti statističke, algoritamske ili izrađene po našim potrebama, a moraju osigurati da tip podataka ostane nepromijenjen, odnosno da maskirana vrijednost i izvorna vrijednost pripadaju istom tipu podataka. Također važno je zadržati korisnost podataka pritom ne ugrožavajući prava na privatnost pojedinca.

U ovom poglavlju ćemo pobliže objasniti samo neke od osnovnih anonimizacijskih tehnika koje se danas koriste, navesti njihove karakteristike i na primjeru pokazati njihovo korištenje. Većina definicija anonimizacijskih tehnika preuzeta je iz [3].

2.1 Supstitucija

U ovoj tehnici izvorni podaci se zamjenjuju zamjenskim podacima. Tip ulaznih podataka može biti numerički, alfanumerički ili čak datum. Zamjena se može izvršiti na svim podacima ili samo na nasumično odabranim te uz dodatne uvjete. U primjeru nasumične supstitucije u tablici 2.1 vidimo kako je svaki znak zamijenjen nekim nasumičnim znakom istog tipa. Također, primjećujemo kako je očuvana izvorna duljina podatka. Znak 'O', koji se u ulaznom podatku pojavljuje dva puta, zamijenjen je različitim znakovima, 'B' i 'Q', dok je slično učinjeno i za znak 'N'.

ulazni podatak	LONDON01
izlazni podatak	ABMEQD12

Tablica 2.1: Nasumična supstitucija

Prednost supstitucije je što čuva format i tip podatka, ali mana što ne daje realistične podatke, što je posebno važno kod imena. Postoje vrste supstitucija koje daju realistične podatke kao izlaz, to su ”supstitucija riječi” i ”supstitucija prema popisu”. U ovim tehnikama vanjska datoteka ili repozitorij sadrži popis vrijednosti. Podaci koji se trebaju maskirati zamjenjuju se vrijednostima iz te vanjske datoteke ili repozitorija.

Na primjer, ako želimo maskirati ime ’Ana’, iz vanjske datoteke koja sadrži popis imena nasumično odaberemo ime, poput ’Petra’. U ovom slučaju duljina izvornog podatka nije nužno očuvana. Veliki nedostatak ove tehnike na velikim količinama podataka je potreba za velikim vanjskim datotekama ili repozitorijima.

2.2 Miješanje

Ova tehnika podrazumijeva nasumičnu preraspodjelu podataka unutar istog stupca kroz različite retke u tablici. Tako izvorne vrijednosti atributa ostaju u skupu podataka, ali su povezane sa drugim zapisom. U primjeru prikazanom u Tablicama 2.2 i 2.3 vidimo kako je tehnika primijenjena na stupac ’BROJ RAČUNA’ time smo zadržali izvorne podatke o brojevima računa, ali oni više nisu vezani na izvorne datume i iznose transakcija.

ID	BROJ RAČUNA	DATUM TRANSAKCIJE	IZNOS TRANSAKCIJE
1	3498271645	2025-01-03	150.00
2	8712635490	2025-01-03	75.50
3	5647389201	2025-01-02	200.00

Tablica 2.2: Tablica *Korisničke transakcije* prije korištenja tehnike miješanja

ID	BROJ RAČUNA	DATUM TRANSAKCIJE	IZNOS TRANSAKCIJE
1	8712635490	2025-01-03	150.00
2	5647389201	2025-01-03	75.50
3	3498271645	2025-01-02	200.00

Tablica 2.3: Tablica *Korisničke transakcije* nakon korištenja tehnike miješanja

Verzija ove tehnike koja se također koristi je i grupno miješanje. U grupnom miješanju umjesto jednog stupca podatke grupe stupaca zajedno preraspodjeljujemo. Ova tehnika se često koristi u slučajevima kada je važno očuvati međusobnu povezanost podataka, kao što je odnos između poštanskog broja i mjesta u adresama. Primjerice, iako je adresa izmišljena, važno je da poštanski broj odgovara određenom gradu, što omogućuje zadržavanje realističnosti podataka. To upravo prikazuju tablice 2.4 i 2.5. Cilj grupnog miješanja je stvoriti realistične podatke, pri čemu adresa ostaje izmišljena.

ID	ADRESA	POŠTANSKI BROJ	MJESTO
1	Trg Stjepana Radića 4	10000	Zagreb
2	Ulica Zrinsko Frankopanska 25	21000	Split
3	Lorenzov prolaz 1	51000	Rijeka
4	Obala kneza Trpimira 24	23000	Zadar
5	Trg Svetog Trojstva 6	31000	Osijek

Tablica 2.4: Tablica *Adresa* prije korištenja tehnike grupnog miješanja

ID	ADRESA	POŠTANSKI BROJ	MJESTO
1	Trg Stjepana Radića 4	21000	Split
2	Ulica Zrinsko Frankopanska 25	31000	Osijek
3	Lorenzov prolaz 1	23000	Zadar
4	Obala kneza Trpimira 24	51000	Rijeka
5	Trg Svetog Trojstva 6	10000	Zagreb

Tablica 2.5: Tablica *Adresa* nakon korištenja tehnike grupnog miješanja

Nažalost ova tehnika ne daje potpunu zaštitu podataka jer uvijek postoji mogućnost da se ponovnom preraspodjelom podataka dobiju izvorne vrijednosti. Iz tog razloga preporučuje se ovu tehniku koristiti na velikim skupovima podataka kako bi se smanjio rizik za otkrivanjem izvornih podataka te uz kombinaciju drugih tehnika anonimizacije.

2.3 Dodavanje šuma

Dodavanje šuma jedna je od najčešće korištenih anonimizacijskih tehnika, koju koriste brojne tvrtke, a *Google* je jedna od njih [5]. Tehnika se primjenjuje na podatke čiji je tip broj ili datum, zato ovu tehniku zovu i varijacija broja ili datuma. Ideja je izvorne podatke malo modificirati kako bi oni postali manje precizni.

Najprije moramo odrediti donju i gornju granicu za raspon unutar kojeg želimo zadržati naše podatke. Tehnike ćemo pokazati na primjeru tablice 2.6. Za stanje računa odredili smo donju granicu 1200, a gornju 2000. Također definirali smo varijaciju tako da dodajemo svakoj od vrijednosti 100, a u slučaju kada bi vrijednost prelazila gornju granicu za vrijednost uzimamo baš gornju granicu. Za dodatnu sigurnost, varijacija broja može biti pojačana odabirom aritmetičkog operatora ili funkcije koji ovise o generiranju slučajnog broja. Na primjer, ako se za ulazni podatak generira slučajni broj između 0 i 5, stanje računa može biti smanjeno za 100, dok bi u slučaju broja između 6 i 9, stanje bilo povećano za 100. Analogno, za primjer datuma kao donju granicu odredili smo 1.1.2023., a gornju 31.12.2025. te pomak od 30 dana.

ID	STANJE RAČUNA	DATUM
1	1720	2023-01-03
2	1200	2024-05-14
3	2000	2024-03-26
4	1950	2025-12-18

Tablica 2.6: Tablica *Stanje računa*

ID	STANJE RAČUNA	DATUM
1	1820	2023-01-03
2	1300	2024-05-14
3	2000	2024-03-26
4	2000	2025-12-18

Tablica 2.7: Tablica *Stanje računa* nakon varijacije broja

ID	STANJE RAČUNA	DATUM
1	1720	2023-02-02
2	1200	2024-06-13
3	2000	2024-04-25
4	1950	2025-12-31

Tablica 2.8: Tablica *Stanje računa* nakon varijacije datuma

U bankarskim sustavima i sustavima ljudskih resursa, tehnika dodavanja šuma koristi se za anonimizaciju osobnih podataka poput plaća zaposlenika, stanja bankovnih računa, datuma rođenja i slično. Iako šum može pomoći u zaštiti identiteta pojedinca, važno je pažljivo upravljati razinom modifikacija podataka kako bi se očuvala njihova kvaliteta i vjerodostojnost.

2.4 Poništavanje

U ovoj tehnici stupac s osobnim podacima anonimizira se tako da se sve vrijednosti u stupcu zamijene vrijednošću NULL. Ova metoda ima ograničenu primjenu, jer se ne može koristiti na podacima za koje NULL vrijednost nema smisla. Na primjer, podatak o krvnoj grupi osobe pripada samo jednoj od vrijednosti: A, B, AB ili 0. Zbog toga se tehnika zamjene vrijednosti s NULL ne može primijeniti na ovakve podatke koji imaju unaprijed definirani skup mogućih vrijednosti.

Postoje tri varijacije ove tehnike.

Tehnika zamjene znakova gdje umjesto zamjene s NULL vrijednošću zamjenu radimo s nekim od znakova. Na primjer, u bazi podataka gdje su pohranjeni podaci poput ”POČINIO KAZNENO DJELO”, takvi bi podaci bili smatrani osjetljivim, te bi se na njih primijenila ova tehnika. S obzirom na to da jedine vrijednosti koje ovaj podatak može imati su ”DA” ili ”NE”, u svrhu anonimizacije, sve bi vrijednosti bile postavljene na ”NE”.

Supresija je varijacija tehnike poništavanja koja se koristi za podatke koji nisu relevantni ili nužni za daljnju analizu ili kad je nemoguće bilo kojom drugom tehnikom anonimizirati te podatke. U njoj podatke određenog stupca uklanjamo iz baze podataka, što omogućuje anonimizaciju podatka jer se ni na koji način ne mogu odrediti izvorni podaci. Na primjer, prilikom analize kvalitete trenera na nekom natjecanju, fokusiramo se na rezultate koje su natjecatelji postigli pod vodstvom određenog trenera, dok imena samih natjecatelja nisu relevantna za analizu te ih možemo ukloniti.

NATJECATELJ	TRENER	BODOVI
Ana	Nikola	156
Petra	Filip	123
Marija	Filip	137
Gabrijela	Nikola	184

Tablica 2.9: Tablica *Rezultati*

TRENER	BODOVI
Nikola	156
Filip	123
Filip	137
Nikola	184

Tablica 2.10: Tablica *Rezultati* nakon supresije

Također, postoji tehnika uvjetnog poništavanja ili supresije u kojoj vrijednosti podataka zamjenjujem s NULL vrijednošću ili ga uklanjamo samo kad je ispunjen određeni uvjet. Na primjer, uklanjamo podatke iz stupca 'KOMENTAR' ako stupac 'ŽALBA KLIJENTA' poprima vrijednost 'DA'.

2.5 Maskiranje simbolom

Maskiranje simbolom je još jedna često korištena tehnika anonimizacije. Ideja ove tehnike je zamijeniti sve ili određene znakove unutar vrijednosti unaprijed definiranim simbolima poput '*', '#', 'X' ili drugih, uz zadržavanje duljine izvorne vrijednosti. Tehnika se koristi u anonimizaciji brojeva kreditnih kartica, osobnih identifikacijskih brojeva, poštanskih brojeva i slično.

ulazni podatak	4532 8947 1234 5678
izlazni podatak	**** **** **** *678

Tablica 2.11: Maskiranje simbolom

2.6 Kriptografija

Kriptografija je znanstvena disciplina koja se bavi proučavanjem metoda za slanje poruka u takvom obliku da ih samo onaj kome su namijenjene može pročitati.

Osnovni zadatak kriptografije je omogućiti dvjema osobama komuniciranje preko nesigurnog komunikacijskog kanala na način da treća osoba, koja može nadzirati komunikacijski kanal, ne može razumjeti njihove poruke. Poruku koju pošiljatelj želi poslati primatelju zvat ćemo otvoreni tekst. Pošiljatelj transformira otvoreni tekst koristeći unaprijed dogovoreni ključ. Taj postupak se naziva šifriranje, a dobiveni rezultat šifrat ili kriptogram. Nakon toga pošiljatelj pošalje šifrat preko nekog komunikacijskog kanala. Protivnik prisluškujući može doznati sadržaj šifrata, ali ne može odrediti otvoreni tekst. Za razliku od njega, primatelj koji zna ključ kojim je šifrirana poruka može dešifrirati šifrat i odrediti otvoreni tekst [6].

Kriptografske tehnike mogu se podijeliti na tehnike simetričnog ključa, tehnike javnog ključa i tehnike sažetih poruka.

Tehnika simetričnog ključa uključuje korištenje istog ključa za anonimizaciju podataka kao i za deanonimizaciju podataka. Ova tehnika vrlo je korisna u scenarijima gdje je potrebno osigurati povjerljivost i integritet podataka tijekom prijenosa te omogućiti deanonimizaciju podataka kada je to potrebno, poput dinamičkog maskiranja i integracijskih testiranja. Prednost je njihova jednostavnost i učinkovitost u zaštiti podataka, no nedostatak je što zahtijevaju sigurnu pohranu i strogo upravljanje ključevima, što povećava složenost. Također, nisu prikladne za situacije u kojima anonimizirani podaci moraju biti realistični.

Tehnike javnog ili asimetričnog ključa, koje uključuju korištenje para javnog i privatnog ključa te algoritme poput RSA, nisu često korištene u svijetu anonimizacije podataka.

Tehnika sažetka poruka uključuje korištenje hash funkcija koje, za ulaz proizvoljne duljine, generiraju rezultati fiksne duljine koji nazivamo hash. Svaki algoritam za sažimanje poruka trebao bi imati ove osobine: nepovratnost, bez kolizija i deterministčnost. MD5 i SHA-2 su algoritmi koji se najčešće koriste za sažimanje poruka. Tehnike sažetka poruka češće se koriste u SSL protokolima i prijenosu poruka, a rijetko u scenarijima dinamičkog maskiranja podataka (koriste se samo kada je ključno osigurati da izvorni podaci nisu izmijenjeni).

2.7 Generalizacija

Još jedna vrsta tehnike anonimizacije koju koristi *Google* je generalizacija i sastoji se od generaliziranja podataka kako bi se promijenila odgovarajuća razina ili red veličine [7]. Ideju generalizacije najbolje možemo ilustrirati primjerom anonimizacije podataka, poput datuma koji uključuje dan, mjesec i godinu. Korištenjem ove tehnike, podaci bi se pojednostavili tako da ostane samo godina, dok bi dan i mjesec bili uklonjeni. Ovaj pristup anonimizaciji smanjuje rizik od reidentifikacije, ali ga ne uklanja u potpunosti.

Tehnike k-anonimnosti i l-raznolikosti ubrajamo među tehnike generalizacije, a njihovu primjenu prikazat ćemo na primjeru tablice 2.12. Proces anonimizacije ove tablice provest ćemo u dva koraka: prvo ćemo ukloniti stupac koji sadrži imena, a zatim ćemo podatke generalizirati.

IME	DOB	POŠTANSKI BROJ	BOLEST
Ivan	29	47677	aritmija
Ana	22	47602	aritmija
Marko	27	47678	aritmija
Petra	43	47905	gripa
Luka	42	47909	aritmija
Iva	47	47906	skolioza
Josip	30	47605	aritmija
Marija	36	47673	skolioza
Ivana	32	47607	skolioza

Tablica 2.12: Tablica *Bolesti*

DOB	POŠTANSKI BROJ	BOLEST
29	47677	aritmija
22	47602	aritmija
27	47678	aritmija
43	47905	gripa
42	47909	aritmija
47	47906	skolioza
30	47605	aritmija
36	47673	skolioza
32	47607	skolioza

Tablica 2.13: Tablica *Bolesti* nakon prvog koraka

Razlika između osobnih podataka koji direktno i indirektno identificiraju pojedinca ključna je za razumijevanje zaštite privatnosti u kontekstu k-anonimnosti. Osobni podaci poput imena, prezimena ili osobnog identifikacijskog broja izravno povezuju osobu s njezinim identitetom, dok podaci poput dobi, spola ili poštanskog broja sami po sebi nisu dovoljni da bi se identificirala osoba, ali mogu u kombinaciji s drugim informacijama omogućiti njezinu identifikaciju. Ti podaci koji indirektno omogućuju identifikaciju nazivaju se kvazi-identifikatori i upravo na njima ćemo primijeniti k-anonimnost. U našem primjeru, kako bismo povećali zaštitu privatnosti, podaci koji izravno identificiraju pojedinca, kao što je ime, uklonjeni su, a k-anonimnost se primjenjuje na preostale podatke koji mogu poslužiti za indirektnu identifikaciju, poput dobi i poštanskog broja.

Tehnika k-anonimnosti sastoji se od grupiranja zapisa k pojedinaca u kategorije koje imaju iste kombinacije. Tako, svaki zapis u skupu podataka "sličan je najmanje k-1 drugih zapisa". Pojedinačne vrijednosti atributa zamjenjuju se širim kategorijama, na primjeru godina 29 zamijeniti ćemo s rasponom 20-29. Ovi podaci imaju 3-anonimnost u odnosu na atribute dob i poštanski broj, jer za svaku kombinaciju tih atributa koja se nalazi u bilo kojem redu tablice, uvijek postoje barem 3 reda s tim točno istim atributima. Nakon primjene k-anonimnosti, vjerojatnost identificiranja pojedinca jednaka je ili manja od $1/k$. Stoga, što je k veći, to je manja vjerojatnost identifikacije.

DOB	POŠTANSKI BROJ	BOLEST
20-29	476**	aritmija
20-29	476**	aritmija
20-29	476**	aritmija
40-49	4790*	gripa
40-49	4790*	aritmija
40-49	4790*	skolioza
30-39	476**	aritmija
30-39	476**	skolioza
30-39	476**	skolioza

Tablica 2.14: Tablica *Bolesti* nakon 3-anonimnosti

Tehnika l-raznolikosti predstavlja nadogradnju k-anonimnosti, koja zahtijeva da u svakoj klasi ekvivalencije postoji najmanje l različitih vrijednosti osobnih atributa. To znači da svaki atribut unutar svake klase ekvivalencije mora imati barem l različitih vrijednosti. Glavni cilj ove tehnike je smanjiti pojavu klasa ekvivalencije s niskom raznolikošću atributa, čime se smanjuje rizik od reidentifikacije.

DOB	POŠTANSKI BROJ	BOLEST
40-49	4790*	gripa
40-49	4790*	aritmija
40-49	4790*	skolioza
30-39	476**	aritmija
30-39	476**	skolioza
30-39	476**	skolioza

Tablica 2.15: Tablica *Bolesti* nakon 3-anonimnosti i 2-raznolikosti

Na tablici 2.14 možemo vidjeti tri klase ekvivalencije. Druga klasa ekvivalencije, koja uključuje vrijednosti 40-49 za dob i 4790* za poštanski broj, je 3-raznolika, jer unutar nje postoje tri različite vrijednosti za atribut bolesti. S druge strane, posljednja klasa ekvivalencije je 2-raznolika, jer sadrži samo dvije različite vrijednosti za bolest (aritmija i skolioza). Kako bismo postigli 2-raznolikost naših podataka, prvu klasu ekvivalencije ćemo izbaciti jer je 1-raznolika, odnosno sadrži samo jednu vrijednost za bolest (aritmija). Konačni prikaz podataka iz tablice 2.12 nakon primjene 3-anonimnosti i 2-raznolikosti prikazan je u tablici 2.15.

Poglavlje 3

Alati za anonimizaciju podataka

U ovom poglavlju ćemo detaljnije istražiti postojeće alate za anonimizaciju podataka, s naglaskom na njihove funkcionalnosti i tehničke mogućnosti. Analizirat ćemo koje tehnike anonimizacije podržavaju i na koji način ih implementiraju u praksi. Također, osvrnut ćemo se na specifične prednosti i nedostatke alata, kao i na situacije u kojima je pojedini alat najprikladniji za korištenje.

Pregled tržišta pokazuje širok raspon alata za anonimizaciju podataka, no većina njih ima ograničenu primjenu jer su razvijeni za specifične svrhe ili unutar organizacija. U nastavku ćemo se fokusirati na alate za anonimizaciju podataka, poput ARX, μ-ARGUS, SDCMicro i Amnesia, pri čemu će primjeri prikazani u nastavku koristiti konkretne podatke o studentskoj depresiji preuzete s [8].

Korišteni podaci sadrže informacije o studentima, uključujući demografske podatke, akademsku izvedbu, životne navike i podatke o mentalnom zdravlju, s ciljem analize i predviđanja razina depresije među studentima. Varijable u ovom skupu podataka obuhvaćaju dob, spol, ocjene, trajanje sna, radni pritisak, zadovoljstvo učenjem i druge relevantne faktore.

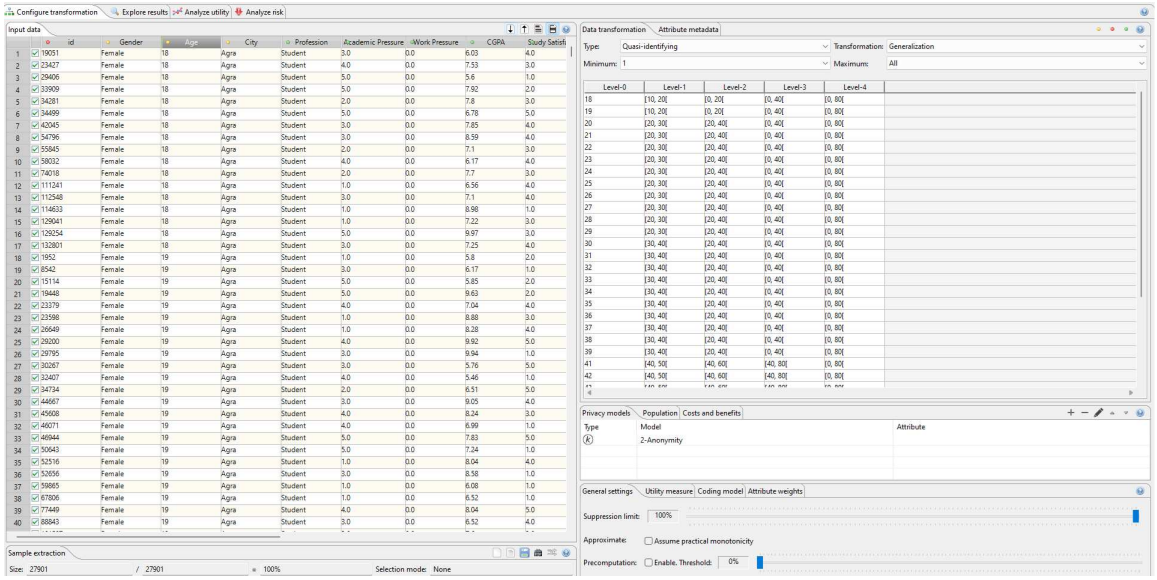
3.1 ARX

ARX je alat otvorenog koda namijenjen transformaciji strukturiranih (tabličnih) osobnih podataka pomoću tehnika iz područja anonimizacije podataka i kontrole statističkog otkrivanja. Njegova glavna svrha je omogućiti transformaciju podataka na način koji osigurava usklađenost s modelima privatnosti i korisnički definiranim pragovima rizika, čime se smanjuje vjerojatnost napada koji mogu ugroziti privatnost.

Napisan je u Javi, što ga čini prenosivim i kompatibilnim s različitim operativnim sustavima. Dostupan je kao desktop aplikacija koja korisnicima nudi intuitivno grafičko sučelje za upravljanje procesima anonimizacije. Osim toga, ARX podržava razne formate poda-

taka, uključujući CSV, XLS, XLSX, te integraciju putem JDBC-a za rad s bazama podataka.

ARX podržava širok raspon tehnika anonimizacije, uključujući supresiju, dodavanje šuma, k-anonimnost, l-raznolikost, t-bliskost i δ-privatnost. Neke od tih tehnika već smo detaljnije objasnili u prethodnom poglavlju. Iako smo se u tom pregledu fokusirali na najpoznatije i najčešće korištene tehnike, treba napomenuti da ARX nudi mnoge druge opcije koje omogućuju korisnicima veliku fleksibilnost u odabiru, ovisno o specifičnim potrebama i zahtjevima.



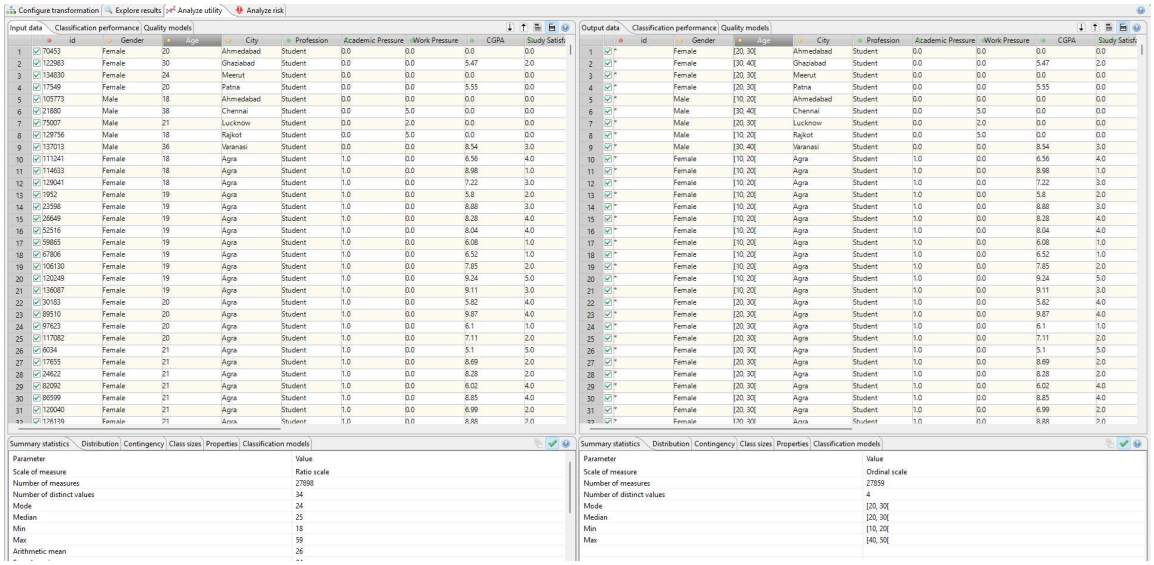
Slika 3.1: Podaci prije primjene tehnike 2-anonimnosti

Kako bismo bolje upoznali samu aplikaciju, primijenili smo 2-anonimnost nad kvazi-identifikatorom 'Dob'. Postavke prije pokretanja anonimizacije prikazane su na slici 3.1, dok su rezultati nakon primjene tehnike prikazani na slici 3.2.

Budući da je atribut 'ID' bio identificirajući, na njega je primijenjena supresija. Prije primjene 2-anonimnosti, atribut 'Dob' sadržavao je 34 različite vrijednosti, s rasponom od 18 do 59 godina. Nakon anonimizacije, broj različitih vrijednosti smanjen je na samo 4 intervala. Iako je ovom agregacijom podataka značajno smanjena preciznost, razina privatnosti je istovremeno značajno povećana.

3.2 μ-ARGUS

μ-Argus je softver za anonimizaciju mikropodataka. Naziv je akronim za "*Anti Re-identification General Utility System*", što se na hrvatski prevodi kao "*Opći sustav za*



Slika 3.2: Podaci nakon primjene tehnike 2-anonimnosti

sprječavanje reidentifikacije". Razvijen je kako bi osigurao sigurno upravljanje mikropodacima koristeći razne tehnike anonimizacije, uključujući dodavanje šuma i supresiju podataka.

Baziran je na R programskom jeziku, a proces anonimizacije uključuje definiranje osobnih varijabli, procjenu rizika od otkrivanja i reidentifikacije te primjenu metoda koje smanjuju taj rizik. μ-Argus se koristi u istraživačkim, statističkim i administrativnim okruženjima.

3.3 SDCMicro

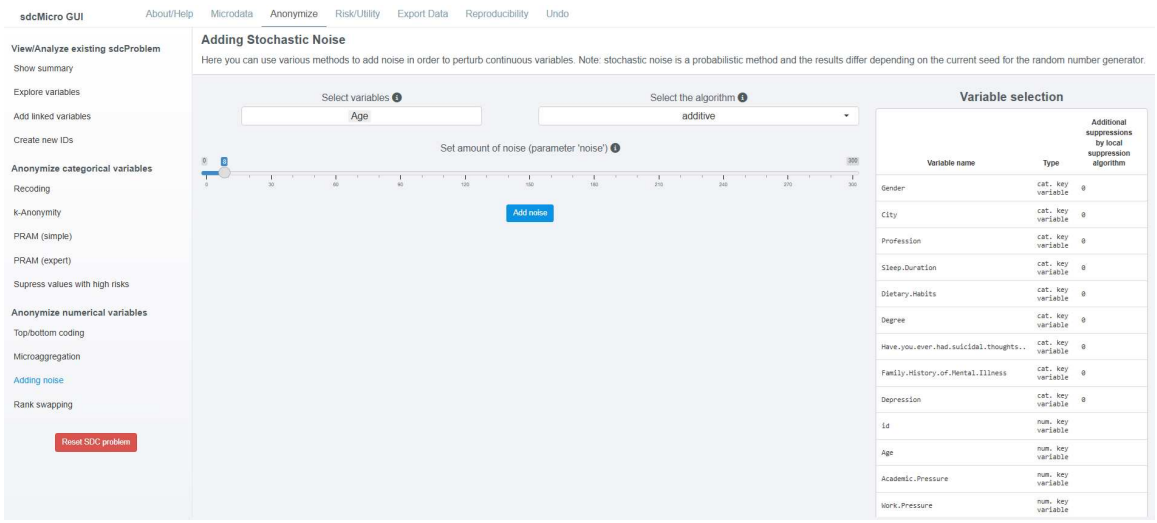
SDCMicro, što je skraćenica za *Statistical Disclosure Control for Micro-Data* (Statistička kontrola otkrivanja za mikropodatke), besplatni je R-paket otvorenog koda namijenjen znanstvenoj i javnoj upotrebi. Ovaj alat omogućuje primjenu raznih tehnika anonimizacije, poput supresije, dodavanja šuma, k-anonimnosti i miješanja, te uključuje funkcije za mjerenje rizika tijekom cijelog procesa.

Dostupan je kao korisnički prilagođeno grafičko sučelje sdcMicro GUI, u kojem korisnici koji nisu upoznati s R-om mogu provoditi metode anonimizacije. Podržava razne formate mikropodataka kao što su STATA, SAS, SPSS, CSV i R. Jedna od prednosti ovog alata je što uključuje funkcije za mjerenje, vizualizaciju i usporedbu rizika i korisnosti tijekom procesa anonimizacije, pomažući organizacijama u pripremi izvještaja.

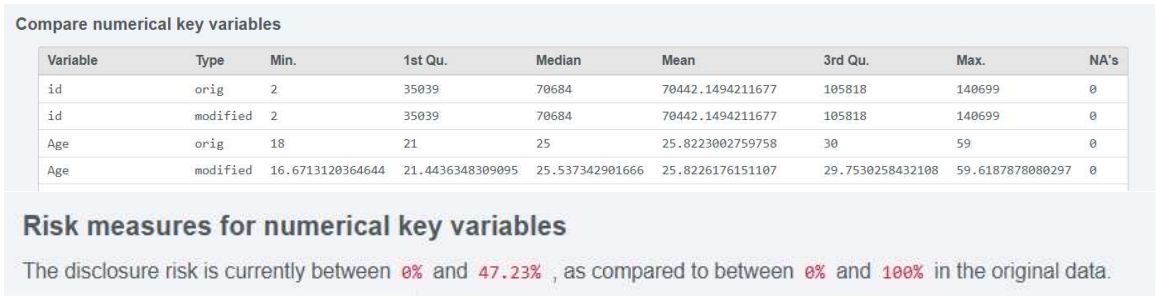
U ovom primjeru, na podatke o studentskoj depresiji, dodajemo šum s parametrom 8

na atribut 'Dob'.

Rezultate nakon dodavanja šuma možemo vidjeti na slici 3.4. Ovim postupkom smo značajno smanjili rizik od otkrivanja, s početnih 100% na 47.23%."



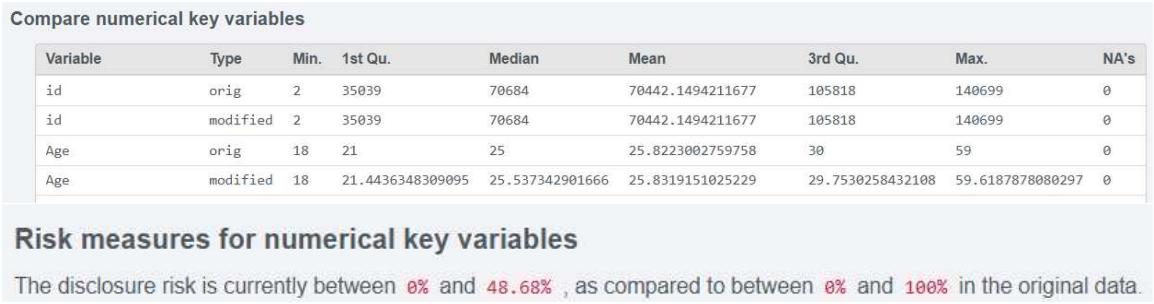
Slika 3.3: Izgled sučelja za dodavanje šuma u SDCMicro GUI



Slika 3.4: Podaci nakon dodavanja šuma

U našim podacima primjećujemo da minimalna vrijednost dobi sada iznosi 16 godina, što nije vjerodostojno jer su podaci o studentima, a studenti su obično stariji od 18 godina. Stoga smo odlučili primijeniti tehniku koja zamjenjuje sve vrijednosti ispod određene granice s novom vrijednošću. U ovom slučaju, sve dobi mlađe od 18 godina zamijenjene su s vrijednošću 18, čime smo poboljšali vjerodostojnost podataka.

Takvim postupkom povećali smo rizik od otkrivanja na 48.68%, ali smatramo da je to opravdano. Iako je došlo do blagog povećanja rizika, zamjena vrijednosti koje ne odgovaraju stvarnim okolnostima s novim, prihvatljivijim vrijednostima omogućuje precizniju analizu i donošenje boljih zaključaka, čime je kvaliteta podataka značajno poboljšana.



Compare numerical key variables

Variable	Type	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
id	orig	2	35039	70684	70442.1494211677	105818	140699	0
id	modified	2	35039	70684	70442.1494211677	105818	140699	0
Age	orig	18	21	25	25.8223002759758	30	59	0
Age	modified	18	21.4436348309095	25.537342901666	25.8319151025229	29.7530258432108	59.6187878080297	0

Slika 3.5: Podaci nakon zamjene vrijednosti

3.4 Amnesia

Alat za anonimizaciju Amnesia je softver napisan u Javi i JavaScriptu koji se koristi lokalno za anonimizaciju osobnih i osjetljivih podataka, pri čemu korisnici učitaju datoteku s osobnim podacima, a ona ih transformira u anonimizirani skup podataka koji se zatim može lokalno pohraniti. Transformacija se temelji na odabirima korisnika i pruža jamstvo anonimizacije za rezultirajući skup podataka.

Amnesia trenutačno podržava tehnike k-anonimnosti i km-anonimnosti. Također, omogućava uvoz podataka iz lokalnih datoteka, Zenodo, Dataverse i DICOM slika, uz opciju povezivanja s odgovarajućim računima i odabira željenih skupova podataka.

Sažeti pregled alata i pripadnih tehnika za anonimizaciju podataka

Ovaj pregled pokazuje da alati za anonimizaciju podataka nisu univerzalni već se razlikuju po svojim značajkama i primjenama. Izbor pravog alata ovisi o specifičnom zadatku, veličini skupa podataka i razini zaštite privatnosti koja je potrebna. Alati poput ARX-a i SDCMicro-a pogodniji su za složenije zadatke s naprednim zahtjevima, dok su μ-ARGUS i Amnesia prikladniji za scenarije u kojima su osnovne tehnike dovoljna zaštita.

Tablica 3.1 prikazuje najčešće korištene tehnike anonimizacije koje podržavaju navedeni alati za anonimizaciju podataka.

ALATI ZA ANONIMIZACIJU PODATAKA	TEHNIKE ANONIMIZACIJE
ARX	supresija dodavanje šuma k-anonimnost l-raznolikost
μ-ARGUS	supresija dodavanje šuma
SDCMicro	supresija dodavanje šuma k-anonimnost miješanje
Amnesia	k-anonimnost km-anonimnost

Tablica 3.1: Popis najčešće korištenih tehnika anonimizacije za različite alate

Poglavlje 4

Aplikacija za anonimizaciju podataka u relacijskim bazama podataka

U ovom poglavlju detaljnije ćemo se posvetiti praktičnom dijelu ovog diplomskog rada, koji je dostupan na [9]. Kao dio rada izrađena je aplikacija pod nazivom AnonyDB, čiji je osnovni cilj anonimizacija relacijskih baza podataka. Alati koji su predstavljeni u prethodnom poglavlju uglavnom su usmjereni na anonimizaciju pojedinačnih skupova podataka, dok se anonimizacija cijele baze podataka pokazuje kao znatno složeniji zadatak.

Anonimizacija baza podataka zahtijeva ne samo uklanjanje ili prikrivanje osobnih podataka, već i očuvanje međusobnih odnosa između različitih tablica i zapisa u bazi podataka. To podrazumijeva očuvanje integriteta strukture podataka, konzistentnosti odnosa te istovremeno osiguravanje da privatnost pojedinaca ostane zaštićena.

OIB (PK)	IME	PREZIME	SPOL
12345678901	Ana	Anić	Ž
98765432109	Marko	Marić	M

Tablica 4.1: Tablica t1

OIB (PK, FK)	**AKADEMSKA GODINA (PK)**	SMJER
12345678901	2023/2024	Financijska matematika
98765432109	2022/2023	Matematička statistika

Tablica 4.2: Tablica t2

Kao primjer, možemo razmotriti tablice 4.1 i 4.2. Tablica 4.1 sadrži osobne podatke studenata, dok tablica 4.2 sadrži informacije o akademskoj godini i smjeru studija. Ove

tablice su međusobno povezane putem primarnog ključa (PK) i vanjskog ključa (FK) — atributa 'OIB'. Povezanost tablica omogućuje nam da izvodimo složene upite, poput onog prikazanog ispod:

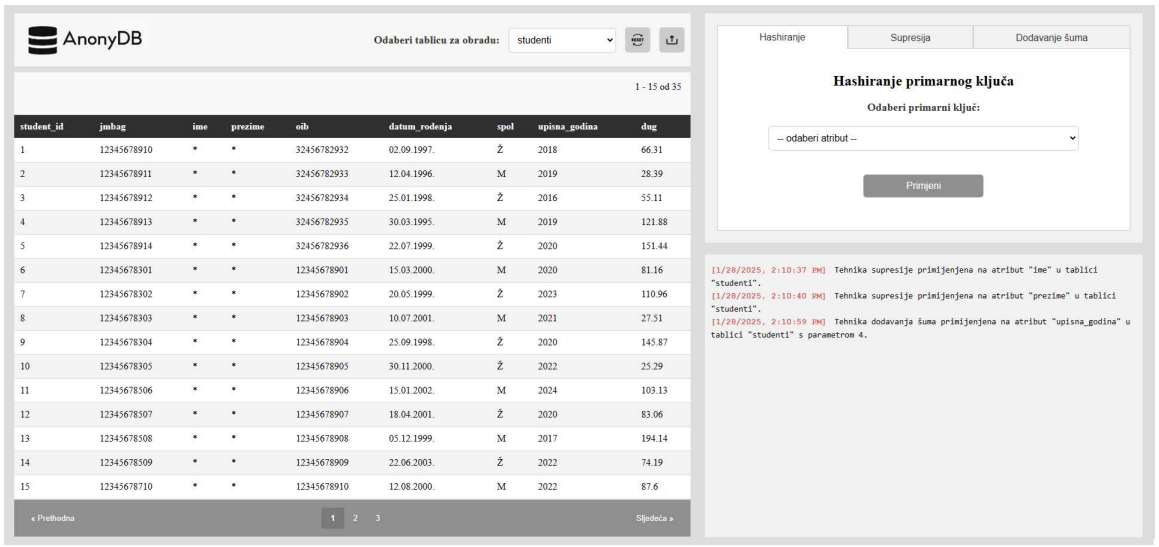
```
SELECT t1.SPOL, t2.SMJER
  FROM t1, t2
 WHERE t1.OIB = t2.OIB;
```

Ovaj upit trebao bi vraćati iste zapise prije i nakon anonimizacije, omogućujući smislene analize uz istovremenu zaštitu identiteta pojedinca.

Aplikacija razvijena u sklopu ovog rada fokusira se upravo na ove izazove. Cilj je bio stvoriti alat koji će omogućiti sigurnu i učinkovitu anonimizaciju podataka u bazama podataka, pri čemu će se pažljivo uravnotežiti potreba za zaštitom privatnosti i zadržavanje korisnosti podataka za daljnju analizu ili upotrebu.

U nastavku ćemo detaljno opisati korake razvoja aplikacije, korištene tehnologije, primijenjene metode anonimizacije te pružiti primjere koji demonstriraju njenu funkcionalnost. Uz to, opisat ćemo bazu podataka koja se koristi za testne primjere, kako bismo omogućili bolje razumijevanje same aplikacije i njenih mogućnosti.

Riječ je o jednostavnoj verziji aplikacije koja služi kao temelj za demonstraciju osnovne ideje i ključnih principa koje bi takva aplikacija trebala implementirati. Na kraju poglavlja osvrnut ćemo se na mogućnosti daljnjeg razvoja aplikacije, pri čemu ćemo istaknuti potencijalne smjerove unapređenja koji bi omogućili širu primjenu, poboljšanu funkcionalnost te veću učinkovitost i fleksibilnost alata.



Slika 4.1: Izgled sučelja aplikacije AnonyDB

Na slici 4.1 može se vidjeti izgled sučelja u slučaju kada je na atribute 'ime' i 'prezime' iz tablice *studenti* primijenjena tehnika supresije, dok je na atribut 'upisna_godina' primijenjena tehnika dodavanja šuma.

4.1 Demonstracijska baza podataka

Za potrebe razvoja ove aplikacije, izradili smo demonstracijsku bazu podataka koja nam omogućava prikazivanje funkcionalnosti aplikacije na realističnim podacima. Baza podataka je implementirana u PostgreSQL sustavu za upravljanje bazama podataka, koji je instaliran lokalno na računalu.

Na slici 4.2 je prikazana shema baze podataka koja sprema podatke o studentskim ocjenama za kolegije na fakultetu. Baza uključuje osobne podatke studenata (kao što su ime, prezime, OIB, datum rođenja, kontakt podaci), informacije o kolegijima koje studenti pohađaju, profesore koji predaju te kolegije, kao i ocjene koje studenti ostvaruju. U nastavku ćemo pobliže objasniti svaku od tablica unutar ove baze podataka.

Slika 4.2: Shema baze podataka

Tablica *studenti*

Tablica *studenti* pohranjuje osobne podatke studenata, uključujući JMBAG (jedinstveni matični broj akademskog građanina), ime, prezime, OIB (osobni identifikacijski broj), datum rođenja, spol, godinu upisa na fakultet te informacije o nepodmirenim školarinama

prema fakultetu. Primarni ključ ove tablice je atribut 'student_id', koji omogućava jedinstvenu identifikaciju svakog studenta unutar baze podataka.

Tablica *kontaktpodaci*

Tablica *kontaktpodaci* pohranjuje informacije o studentovim kontakt podacima, uključujući broj telefona, adresu, grad, poštanski broj, državu i email adresu. Povezana je s tablicom *studenti* putem stranog ključa 'student_id', koji upućuje na odgovarajući zapis u tablici studenata. Primarni ključ tablice je 'kontakt_id', koji osigurava jedinstvenost svakog zapisa u ovoj tablici.

Tablica *kolegiji*

Tablica *kolegiji* pohranjuje podatke o kolegijima, uključujući naziv kolegija i broj ECTS bodova. Primarni ključ tablice čini atribut 'kolegij_id'.

Tablica *profesori*

Tablica *profesori* pohranjuje osnovne podatke o profesorima, uključujući ime, prezime, email adresu i titulu. Primarni ključ tablice čini atribut 'profesor_id'.

Tablica *predaje*

Tablica *predaje* povezuje tablice *kolegiji* i *profesori* preko stranih ključeva, te prikazuje koji profesor predaje koji kolegij u određenoj akademskoj godini. Primarni ključ tablice čini atribut 'predaje_id'.

Tablica *ocjene*

Tablica *ocjene* povezuje tablice *studenti* i *predaje* putem stranih ključeva, pohranjujući podatke o ocjenama koje su studenti dobili za određene kolegije, koje predaju profesori, te datum kada je ocjena dodijeljena. Primarni ključ ove tablice je 'ocjena_id'.

4.2 Prikaz mogućnosti aplikacije

Aplikacija omogućuje korisnicima da odaberu tablice koje žele anonimizirati, pri čemu su dostupne tri tehnike anonimizacije: hashiranje, supresija i dodavanje šuma. Svaki atribut može biti anonimiziran isključivo jednom tehnikom, čime se osigurava jasnoća i dosljednost u obradi podataka.

Pored osnovnih funkcionalnosti anonimizacije, AnonyDB pruža dodatne mogućnosti poput resetiranja svih izmjena, izvoza anonimizirane baze podataka u CSV format te pregleda primijenjenih tehnika na pojedinim atributima.

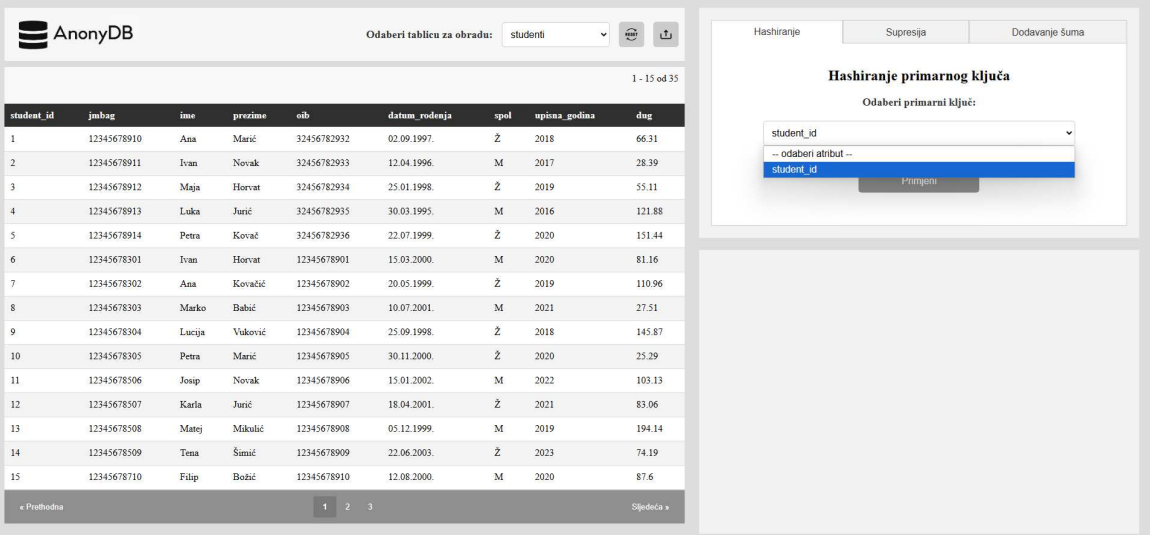
Za anonimizaciju podataka odabrali smo supresiju i dodavanje šuma, jer su to jedne od najčešće korištenih tehnika u postojećim alatima na tržištu. Hashiranje koristimo isključivo za anonimizaciju primarnih ključeva, čime osiguravamo očuvanje odnosa između tablica i zapisa u bazi podataka.

Jedno od ključnih ograničenja aplikacije je to da se na svaki atribut može primijeniti samo jedna tehnika anonimizacije. Na primjer, kombiniranje dodavanja šuma na već supresirane podatke nije logično ni korisno, stoga je aplikacija dizajnirana tako da spriječi takve slučajeve.

U nastavku će biti detaljnije opisane sve implementirane funkcionalnosti i način njihove primjene u AnonyDB-u.

Hashiranje primarnog ključa

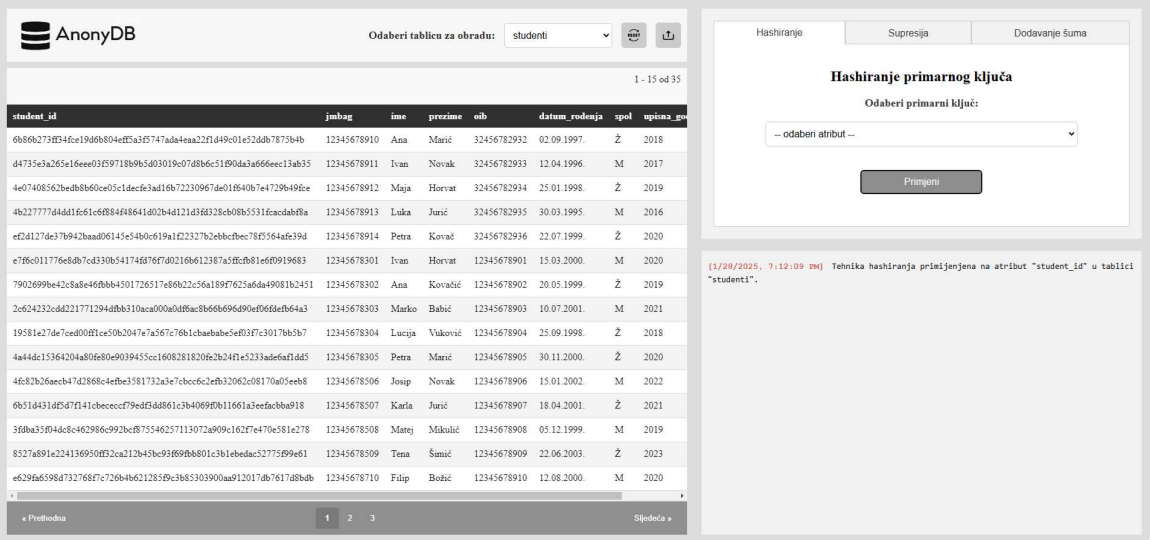
Kao što smo već ranije spomenuli, tehniku hashiranja koristimo isključivo za anonimizaciju primarnih ključeva. Stoga, na slici 4.3 možemo vidjeti da je u padajućem izborniku za tablicu *studenti* dostupan jedino izbor 'student_id'.



Slika 4.3: Izgled tablice *studenti* prije hashiranja primarnog ključa

Budući da postoji mnogo različitih algoritama za hashiranje, za razvoj ove aplikacije odabrali smo algoritam SHA-256 zbog njegove sigurnosti i široke primjene u industriji.

SHA-256 je dio obitelji SHA algoritama i nudi visok nivo zaštite podataka zahvaljujući svojoj otpornosti na kolizije i kriptografsku sigurnost.

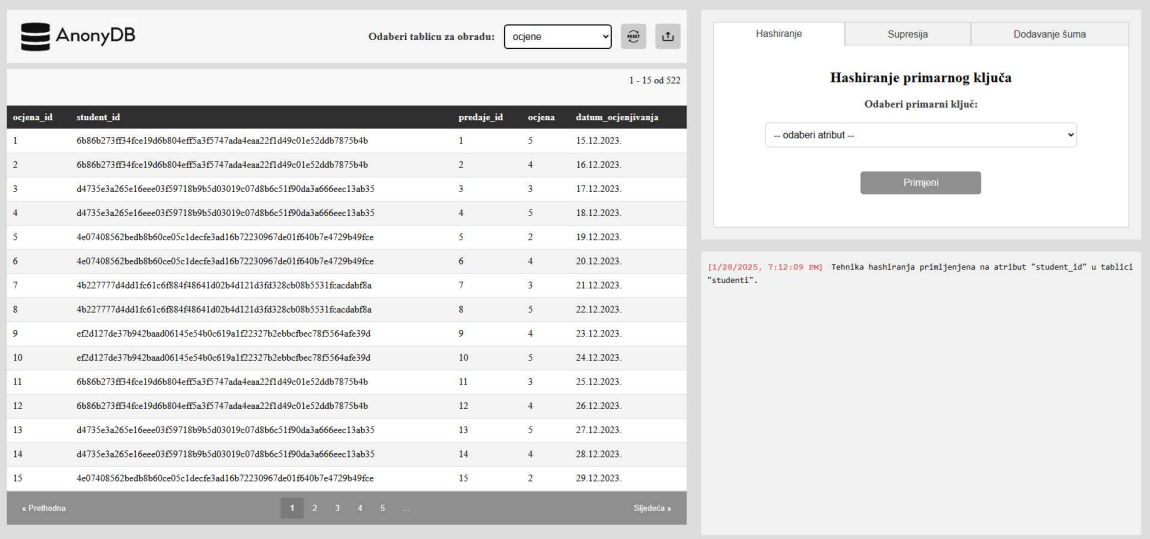


Slika 4.4: Izgled tablice *studenti* nakon hashiranja primarnog ključa

Na slici 4.4 prikazana je tablica *studenti* nakon primjene hashiranja na njen primarni ključ. Ova tehnika osigurava zaštitu identiteta pojedinaca dok istovremeno omogućuje očuvanje integriteta podataka unutar baze.

Važno je istaknuti da se tehnika anonimizacije hashiranje ne može izravno primijeniti na strane ključeve u povezanim tablicama. Umjesto toga, anonimizacija stranih ključeva odvija se kaskadno, kao posljedica anonimizacije primarnog ključa u pripadajućoj originalnoj tablici. To znači da se, umjesto zasebne obrade stranih ključeva, njihove vrijednosti automatski prilagođavaju promjenama primarnog ključa, čime se osigurava konzistentnost odnosa između tablica.

Primjer ovog procesa može se vidjeti na slici 4.5, gdje tablica *ocjene* sadrži atribut student_id, koji predstavlja strani ključ prema tablici *studenti*. Nakon primjene hashiranja na primarni ključ student_id u tablici *studenti*, isti algoritam se koristi za ažuriranje pripadajućih vrijednosti stranog ključa u tablici *ocjene*. Na taj način, integritet referencijalnih veza u bazi podataka ostaje očuvan, unatoč tome što su originalne vrijednosti ključeva zamijenjene njihovim hash ekvivalentima.



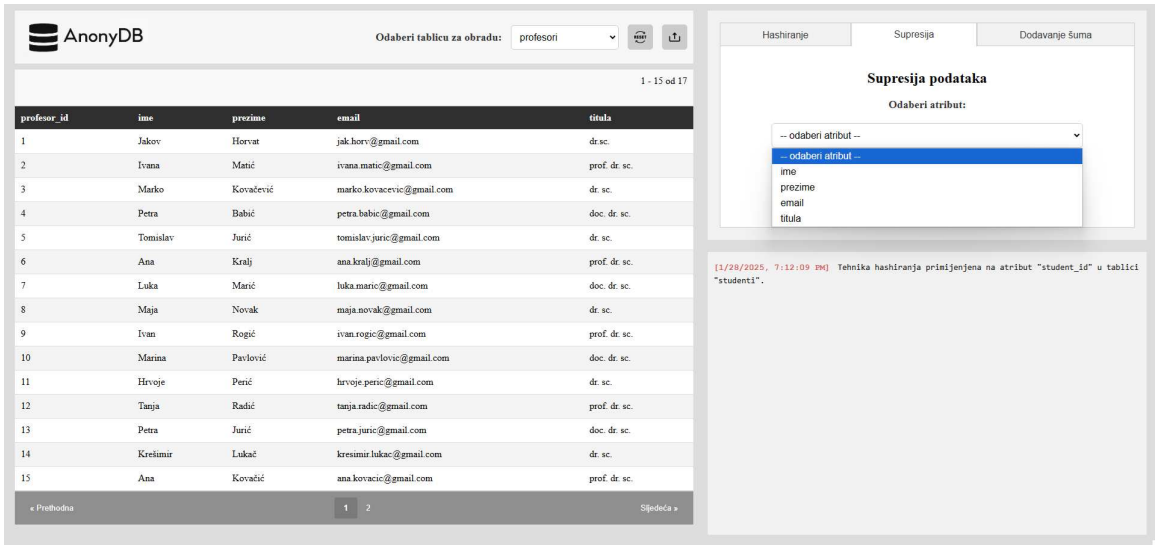
Slika 4.5: Izgled tablice *ocjene* nakon hashiranja primarnog ključa 'student_id' iz tablice *studenti*

Supresija

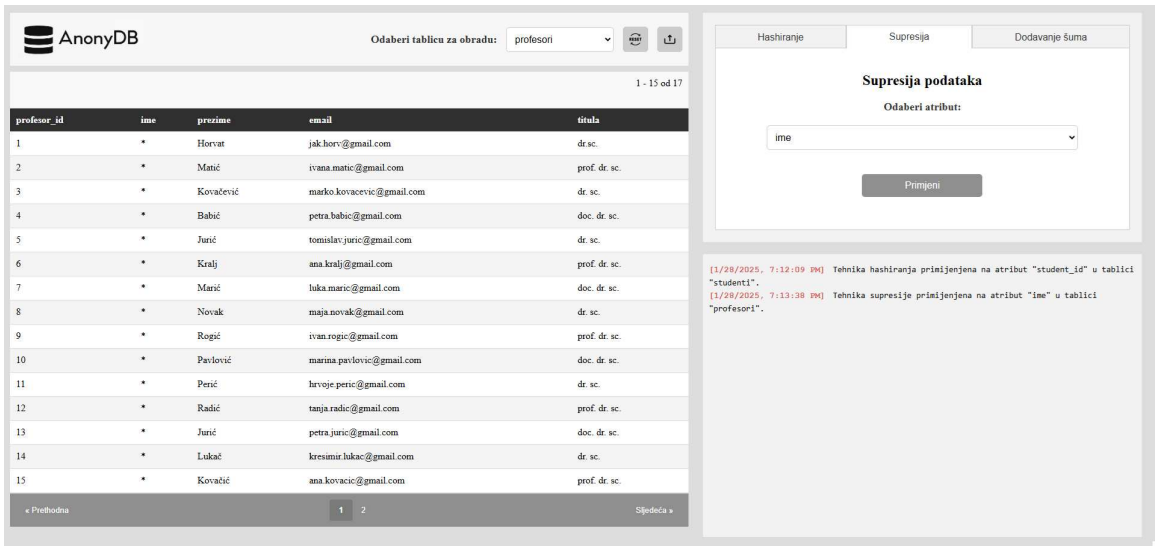
Supresija je jedna od ključnih tehnika anonimizacije podataka koju smo odabrali za zaštitu svih atributa u tablici, osim primarnih i stranih ključeva.

Jedna od najvećih prednosti supresije je njezina široka primjenjivost, budući da se može koristiti na različitim vrstama podataka, neovisno o njihovoj strukturi ili formatu. Kako aplikacija omogućuje korisnicima da samostalno definiraju identifikatore i kvazi-identifikatore, supresija se može selektivno primijeniti na atribute koji predstavljaju rizik od reidentifikacije. Na taj način korisnik dobiva potpunu kontrolu nad razinom zaštite podataka, prilagođavajući anonimizaciju specifičnim potrebama.

Na slikama 4.6 i 4.7 prikazana je tablica *profesori* prije i nakon primjene supresije na atribut 'ime'. U ovom primjeru, sve originalne vrijednosti unutar tog atributa zamijenjene su znakom '*', čime je postignuta potpuna anonimizacija ovog podatka. Ovim postupkom osigurava se da osobni podaci profesora više nisu prepoznatljivi, dok se istovremeno očuva struktura tablice i njena funkcionalnost unutar baze podataka.



Slika 4.6: Izgled tablice *profesori* prije supresije

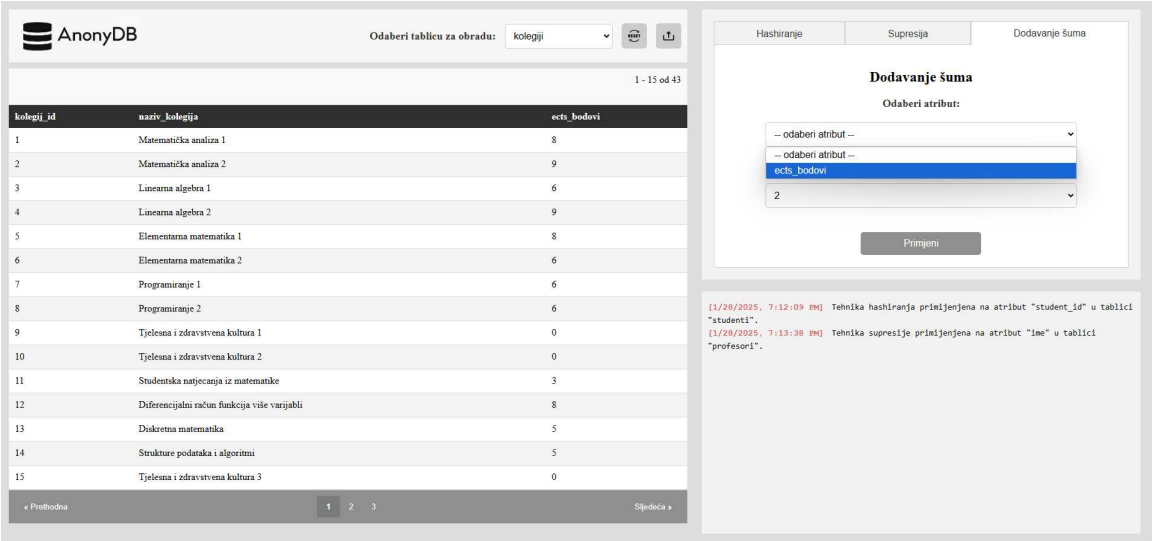


Slika 4.7: Izgled tablice *profesori* nakon supresije atributa 'ime'

Dodavanje šuma

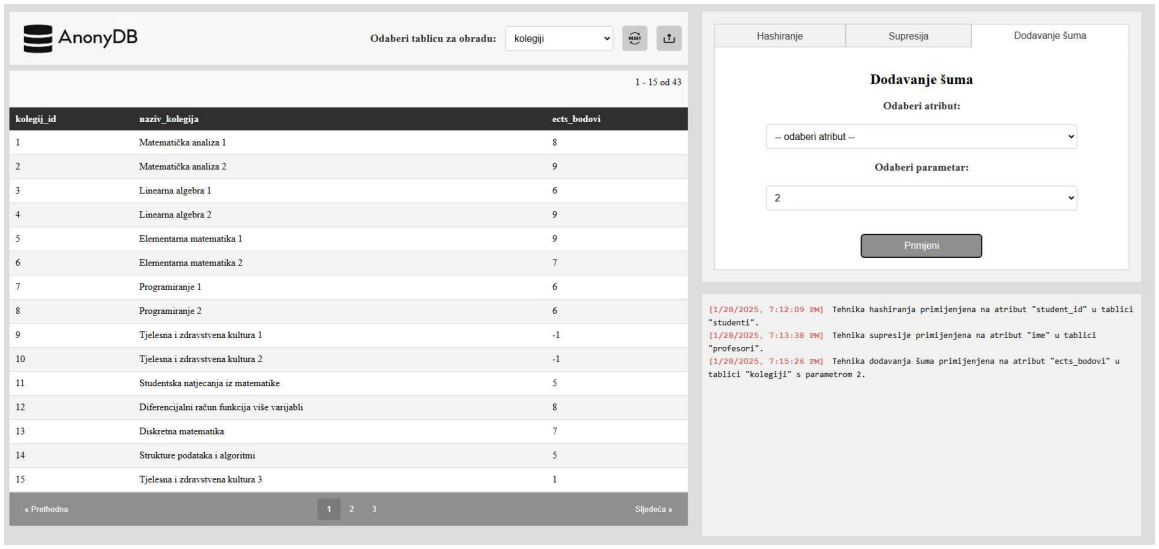
Dodavanje šuma je tehnika anonimizacije koju smo koristili isključivo za numeričke atribute u bazi podataka koji nisu primarni ili strani ključevi. Ova metoda funkcionira tako da smo na postojeće numeričke vrijednosti dodavali nasumične varijacije, čime smo osigurali anonimizaciju podataka, ali bez gubitka njihove osnovne strukture ili obrasca.

U našem slučaju, primijenili smo dodavanje šuma na atribut 'ects_bodovi' u tablici kolegiji, budući da je to bio jedini numerički atribut u toj tablici pogodan za ovu tehniku, što je prikazano na slici 4.8. Kako bismo omogućili veću fleksibilnost korisnicima aplikacije, dodali smo mogućnost unosa parametra u rasponu od 1 do 10, čime se određuje intenzitet varijacije. Viši parametar dovodi do većih odstupanja od originalnih vrijednosti, dok niži parametar omogućuje blaže promjene, čime se postiže veća sličnost s izvornim podacima. Ovaj pristup korisnicima pruža kontrolu nad razinom anonimizacije, prilagođavajući je specifičnim zahtjevima i kontekstu korištenja baze podataka.



Slika 4.8: Izgled tablice *kolegiji* prije dodavanja šuma

Kao što je prikazano na slici 4.9, primjenom tehnike s parametrom 2 na atribut 'ects_bodovi' dobili smo nove vrijednosti koje su slične originalnim. Budući da su one generirane nasumično uočavamo problem: za neki od atributa koji predstavljaju ECTS bodove dobili smo negativne vrijednosti. Iako u određenim slučajevima to možda ne bi predstavljalo problem, u stvarnom svijetu ECTS bodovi uvijek imaju nenegativne vrijednosti.



Slika 4.9: Izgled tablice *kolegiji* nakon dodavanja šuma na atribut 'ects_bodovi'

Kao rješenje ovog problema i za stvaranje realističnijih podataka, bilo bi korisno implementirati tehniku zamjene vrijednosti, sličnu onoj koja je omogućena u SDCMicro alatu (vidi 3.3). Također, budući da je dodavanje šuma trenutno omogućeno samo za numeričke atribute, a znamo da se ova tehnika može primijeniti i na datume, smatramo da bi to moglo biti jedno od unaprjeđenja ove aplikacije. U odjeljku 4.4 navodimo još nekoliko mogućih unaprjeđenja koja bi mogla biti implementirana u budućnosti.

Ostale mogućnosti aplikacije

Uz izbornik za odabir tablice u bazi podataka nalaze se dva gumba koji nude dodatne mogućnosti.

Lijevi gumb omogućuje resetiranje svih primijenjenih tehnika anonimizacije, čime se baza podataka vraća u svoje izvorno stanje prije obrade. Ova opcija korisnicima pruža fleksibilnost da eksperimentiraju s različitim tehnikama anonimizacije, bez rizika trajnog gubitka originalnih podataka. Resetiranjem se poništavaju sve promjene, omogućujući ponovnu primjenu anonimizacije s novim parametrima ili tehnikama.

Desni gumb služi za izvoz baze podataka u CSV format, što omogućuje jednostavno dijeljenje i daljnju obradu podataka izvan same aplikacije. Klikom na ovaj gumb, svaka tablica u bazi podataka sprema se kao zasebna CSV datoteka, a zatim se sve te datoteke automatski komprimiraju u ZIP arhivu, olakšavajući preuzimanje i organizaciju podataka.

4.3 Tehnička implementacija aplikacije

AnonyDB je web aplikacija razvijena u programskom jeziku Java, koja koristi RESTful arhitekturu za komunikaciju između klijenta i poslužitelja. Kao sustav za upravljanje bazama podataka koristi PostgreSQL te je aplikacija dizajnirana tako da omogućuje jednostavnu integraciju s bilo kojom drugom bazom podataka bez gubitka funkcionalnosti. Za prikaz korisničkog sučelja koristi se kombinacija standardnih web tehnologija HTML, CSS i JavaScript, čime se osigurava pregledan i intuitivan dizajn.

Jedna od ključnih značajki aplikacije je način na koji upravlja anonimizacijom podataka. Važno je istaknuti da se sve primijenjene tehnike anonimizacije pohranjuju isključivo u memoriji aplikacije, a ne u samoj bazi podataka. Ovakav pristup predstavlja značajnu prednost jer omogućuje očuvanje izvornih podataka u bazi, bez trajnih izmjena, što osigurava veću kontrolu i fleksibilnost pri anonimizaciji. Međutim, s obzirom na to da se sve tehnike anonimizacije izvršavaju i pohranjuju u memoriji aplikacije, kod obrade velikih količina podataka to bi moglo negativno utjecati na performanse sustava.

Analiza ključnih komponenti projekta

Za razvoj ovog projekta koristimo nekoliko biblioteka, uključujući Hibernate ORM, Panache, Quarkus ARC, Resteasy, JSON-B i Qute, koje nam omogućuju učinkovito upravljanje podacima, razvoj RESTful servisa, serializaciju podataka i dinamičko generiranje sadržaja. Hibernate ORM i Panache omogućuju jednostavno mapiranje objekata u relacijsku bazu podataka, dok Quarkus ARC olakšava ubrizgavanje ovisnosti u aplikaciji. Resteasy zajedno s JSON-B omogućuje izradu RESTful servisa i konverziju podataka u JSON format za komunikaciju s klijentima. Za dinamičko generiranje sadržaja koristimo Qute, koji omogućuje izradu prilagođenih HTML stranica temeljenih na podacima aplikacije.

Projekt se sastoji od tri kontrolerske klase, svaka povezana s odgovarajućom servisnom klasom. *AnonymizationController* upravlja HTTP zahtjevima vezanim uz anonimizaciju podataka, dok *AnonymizationService* implementira specifične anonimizacijske tehnike i logiku vezanu uz anonimizaciju podataka. Na sličan način, *ExportController* obrađuje zahtjeve za izvoz podataka, dok *ExportService* implementira logiku izvoza. *TablesController* odgovara na zahtjeve za dohvat podataka iz baze, dok *TablesService* upravlja procesom dohvaćanja i obrade podataka.

U nastavku ćemo detaljnije razmotriti neke metode unutar servisnih klasa.

```
1     public void hashPrimaryKey(String tableName,
2         String primaryKeyColumn) throws Exception {
3         if (!tablesService.isPrimaryKey(tableName, primaryKeyColumn)) {
4             throw new Exception("Column is not a primary key.");
5         }
```

```
6
7         if (tablesService.isColumnAnonymized(tableName,
8             primaryKeyColumn)) {
9             throw new Exception("Technique already applied
10                to this column.");
11        }
12
13        List<Map<String, Object>> tableData =
14            tablesService.getTableData(tableName);
15
16        Map<Object, String> hashedValues = new HashMap<>();
17        for (Map<String, Object> row : tableData) {
18            Object primaryKeyValue = row.get(primaryKeyColumn);
19            String hashedValue = hashValue(primaryKeyValue.toString());
20            hashedValues.put(primaryKeyValue, hashedValue);
21            row.put(primaryKeyColumn, hashedValue);
22        }
23
24        tablesService.setColumnAnonymizationTechnique(tableName,
25            primaryKeyColumn, "hash");
26
27        updateForeignKeys(tableName, hashedValues);
28    }
29
```

Obje prikazane metode nalaze se u klasi *AnonymizationService*. Prva metoda, *hashPrimaryKey*, implementira tehniku hashiranja primarnih ključeva. Metoda prima ime tablice i naziv atributa koji predstavlja primarni ključ. Prvo se provodi provjera je li navedeni atribut zaista primarni ključ te tablice. Slijedi provjera je li na tom atributu već primijenjena tehnika anonimizacije. Nakon toga, metoda dohvaća sve podatke iz trenutnog stanja u memoriji jer već može biti primijenjena neka tehnika anonimizacije nad tom tablicom. Za svaki zapis u memoriji, vrijednost primarnog ključa se hashira, a originalna vrijednost se sprema u pomoćnu mapu zajedno s hashiranim vrijednostima. Na kraju, postavlja se oznaka da je tehnika anonimizacije primijenjena na taj atribut, a zatim se poziva metoda koja kaskadno ažurira sve strane ključeve u povezanim tablicama.

Druga metoda, *applyNoise*, implementira tehniku dodavanja šuma na vrijednosti određenog stupca u tablici. Ova metoda prima ime tablice, naziv atributa na kojem će se primijeniti šum, te parametar koji definira raspon šuma. Kao i prethodna metoda, prvo se provodi provjera je li atribut primarni ili strani ključ, te se osigurava da tehnika još nije primijenjena. Zatim se dohvaćaju podaci iz trenutnog stanja u memoriji, a za svaki zapis provodi se dodavanje šuma na numeričke vrijednosti atributa. Ako vrijednost nije numerička, baca se iznimka. Nakon što se šum uspješno primijeni, označava se da je tehnika anonimizacije primijenjena na taj atribut.

```
    public void applyNoise(String tableName, String columnName,
        String noiseParameter) throws Exception {
        if (tablesService.isPrimaryKey(tableName, columnName)) {
            throw new Exception("Column is a primary key
                or foreign key.");
        }

        if (tablesService.isForeignKey(tableName, columnName)) {
            throw new Exception("Column is a primary key
                or foreign key.");
        }

        List<Map<String, Object>> tableData =
            tablesService.getTableData(tableName);
        if (tablesService.isColumnAnonymized(tableName, columnName)) {
            throw new Exception("Technique already applied
                to this column.");
        }

        double noiseRange = Double.parseDouble(noiseParameter);

        for (Map<String, Object> row : tableData) {
            Object originalValue = row.get(columnName);

            if (originalValue instanceof Number) {
                double noisyValue =
                    applyNoiseToNumericValue(originalValue, noiseRange);
                row.put(columnName,
                    preserveDecimalFormat(originalValue, noisyValue));
            } else {
                throw new Exception("Unsupported numeric type.");
            }
        }

        tablesService.setColumnAnonymizationTechnique(tableName,
            columnName, "noise");
    }

```

Metoda *exportAllData* koja se nalazi u *ExportService* odgovorna je za izvoz svih podataka iz tablica u CSV format te pakiranje tih CSV datoteka u zip arhivu. Ova metoda uzima sve tablice iz trenutnog stanja podataka u memoriji i generira zip datoteku koja sadrži po jednu CSV datoteku za svaku tablicu.

Na početku, metoda postavlja putanju za zip datoteku AnonyDB.zip. Zatim se za svaku tablicu dohvaćaju podaci i provjerava je li tablica prazna. Ako tablica nije prazna, kreira se novi *ZipEntry* za tu tablicu, čije ime odgovara imenu tablice s ekstenzijom .csv. Unutar

svake tablice, prvo se generiraju zaglavlja za CSV datoteku koristeći ključeve prvog zapisa. Zatim se za svaki zapis generira redak u CSV formatu, pri čemu se svaka vrijednost atributa stavlja u odgovarajući redak, uz osiguranje da su svi specijalni znakovi poput ';' ili novog reda pravilno zamijenjeni. Na kraju, svi podaci za tu tablicu upisuju se u zip datoteku. Nakon što su svi podaci iz svih tablica obrađeni, zip datoteka se zatvara i vraća putanja do nje.

```
    public Path exportAllData() throws IOException, SQLException {
        Path zipFilePath = Paths.get("AnonyDB.zip");

        try (FileOutputStream fos =
            new FileOutputStream(zipFilePath.toFile());
             ZipOutputStream zipOut =
                new ZipOutputStream(fos, StandardCharsets.UTF_8)) {

            for (String tableName : tablesService.getTables()) {
                List<Map<String, Object>> tableData =
                    tablesService.getTableData(tableName);

                if (tableData == null || tableData.isEmpty()) {
                    continue;
                }

                ZipEntry zipEntry = new ZipEntry(tableName + ".csv");
                zipOut.putNextEntry(zipEntry);

                Map<String, Object> firstRow = tableData.getFirst();
                String headers = String.join(";", firstRow.keySet());
                zipOut.write(headers.getBytes(StandardCharsets.UTF_8));
                zipOut.write("\n".getBytes());

                for (Map<String, Object> row : tableData) {
                    StringBuilder rowData = new StringBuilder();
                    for (Object value : row.values()) {
                        if (value != null) {
                            String valueStr =
                                value.toString().replace(";", "\\;")
                                .replace("\n", " ").replace("\r", " ");
                            rowData.append(valueStr).append(";");
                        } else {
                            rowData.append(";");
                        }
                    }
                    if (!rowData.isEmpty()) {
                        rowData.deleteCharAt(rowData.length() - 1);
                    }
```

```
40                      zipOut.write(rowData.toString().getBytes(
    StandardCharsets.UTF_8));
41                      zipOut.write("\n".getBytes());
42                  }
43
44                  zipOut.closeEntry();
45              }
46
47          } catch (IOException e) {
48              e.printStackTrace();
49              throw e;
50          }
51
52          return zipFilePath;
53      }
54
```

4.4 Mogućnosti unaprjeđenja aplikacije

Iako je trenutna verzija aplikacije osmišljena prvenstveno kao demonstracijski alat, postoji značajan potencijal za njezin daljnji razvoj i proširenje funkcionalnosti. Jedan od ključnih smjerova nadogradnje uključuje omogućavanje izravne promjene podataka unutar baze podataka, umjesto dosadašnjeg pristupa gdje se promjene pohranjuju u memoriji aplikacije. Takav pristup smanjio bi opterećenje memorije i omogućio bržu obradu podataka, osobito u velikim bazama podataka. Uz to, implementacija naprednih algoritama za obradu velikih baza podataka te prilagodba za rad u distribuiranim sustavima omogućila bi obradu većih količina podataka u stvarnom vremenu, što je od velike važnosti za organizacije koje rade s dinamičnim i opsežnim bazama podataka.

Također, aplikacija bi se mogla unaprijediti dodavanjem podrške za primjenu više tehnika anonimizacije na istom atributu, čime bi se omogućile raznovrsnije i naprednije mogućnosti zaštite podataka. Osim toga, aplikacija bi mogla ponuditi i širi spektar tehnika anonimizacije, ne ograničavajući se samo na tri trenutno implementirane (hashiranje, supresija i dodavanje šuma). Na primjer, kombinacija dodavanja šuma i generalizacije mogla bi dodatno povećati razinu privatnosti, a istovremeno zadržati korisnost podataka za analitičke svrhe. Ovakva fleksibilnost omogućila bi korisnicima da odaberu optimalnu kombinaciju tehnika ovisno o specifičnim zahtjevima i ciljevima zaštite podataka.

Još jedno ključno područje razvoja bilo bi uvođenje funkcionalnosti za analizu rizika od otkrivanja identiteta. Ovakva analiza, dostupna u nekim postojećim alatima, korisnicima bi omogućila procjenu razine zaštite podataka nakon primjene anonimizacijskih tehnika. Osim toga, alat bi mogao prepoznati potencijalne ranjivosti i ponuditi prijedloge za optimizaciju anonimizacije podataka.

Sve ove nadogradnje učinile bi AnonyDB ne samo korisnim demonstracijskim alatom već i snažnim rješenjem za anonimizaciju baza podataka, prilagođenim kako akademskim, tako i industrijskim potrebama. Proširenjem funkcionalnosti, aplikacija bi mogla značajno unaprijediti standarde zaštite privatnosti i sigurnosti podataka u suvremenom digitalnom okruženju.

Zaključak

Ovaj diplomski rad bavio se anonimizacijom podataka, s posebnim naglaskom na relacijske baze podataka. U teorijskom dijelu rada detaljno su objašnjeni osnovni pojmovi vezani uz anonimizaciju, uključujući razliku između anonimizacije i pseudonimizacije. Analizirane su najčešće korištene tehnike anonimizacije, poput supstitucije, miješanja, dodavanja šuma, poništavanja, maskiranja simbolima, primjene kriptografskih tehnika i generalizacije podataka, uz njihova ograničenja i prednosti u kontekstu zaštite privatnosti. Također su istraženi popularni alati za anonimizaciju podataka, poput ARX, μ-ARGUS, SDCMicro i Amnesia, koji omogućuju različite pristupe anonimizaciji u skladu s potrebama korisnika i vrstom podataka koji se obrađuju.

Praktični dio diplomskog rada obuhvatio je razvoj vlastite aplikacije za anonimizaciju relacijskih baza podataka pod nazivom AnonyDB, koja implementira različite tehnike anonimizacije s ciljem zaštite privatnosti korisničkih podataka. U radu je detaljno opisan proces razvoja aplikacije, od odabira tehnologija do implementacije konkretnih funkcionalnosti, uz primjere koji prikazuju rezultate anonimizacije na demonstracijskoj bazi podataka. Također, predstavljene su mogućnosti unaprjeđenja aplikacije, uključujući izravnu promjenu podataka u bazi podataka, podršku za primjenu više tehnika anonimizacije na istom atributu te implementaciju analize rizika od reidentifikacije. Daljnjim razvojem, AnonyDB ima potencijal postati konkurentan alat na tržištu anonimizacije relacijskih baza podataka.

Bibliografija

[1] Yahoo: *Yahoo Security Notice December 14, 2016*. https://help.yahoo.com/kb/previously-announced-company-december-sln27925.html.

[2] Manger, Robert: *Baze podataka*. Element, 2014.

[3] Raghunathan, Balaji: *The Complete Book of Data Anonymization: From Planning to Implementation*. Infosys Press, 2013.

[4] European Union: *Uredba (EU) 2016/679 Europskog parlamenta i Vijeća od 27. travnja 2016. o zaštiti pojedinaca u vezi s obradom osobnih podataka i o slobodnom kretanju takvih podataka*, 2016. https://eur-lex.europa.eu/legal-content/HR/TXT/PDF/?uri=CELEX:02016R0679-20160504.

[5] Google: *How Google Anonymises Data*. https://policies.google.com/technologies/anonymization, Retrieved from Google Privacy&Terms.

[6] Marcel Maretić, Andrej Dujella i: *Kriptografija*. Element, 2007.

[7] Jorge Bernardino, Joana Ferreira Marques i: *Analysis of Data Anonymization Techniques*. KOED, 2020.

[8] Hopesb: *Student Depression Dataset*. https://www.kaggle.com/datasets/hopesb/student-depression-dataset/data.

[9] Jerešić, Helena: *AnonyDB*, 2025. https://github.com/helenajeresic/AnonyDB.

Sažetak

Zaštita osobnih podataka postala je ključna u današnjem digitalnom društvu, osobito u kontekstu relacijskih baza podataka koje svakodnevno obrađuju velike količine osjetljivih informacija. Anonimizacija podataka omogućuje zaštitu privatnosti korisnika smanjujući rizik od reidentifikacije, što omogućava sigurno pohranjivanje, obradu i razmjenu podataka, te pomaže u ispunjavanju pravnih zahtjeva poput Opće uredbe o zaštiti podataka.

U prvom poglavlju rada definirani su ključni pojmovi poput pseudoanonimizacije i anonimizacije, te razlika između njih, uz objašnjenje kako svaka od tih tehnika doprinosi zaštiti privatnosti.

Drugo poglavlje fokusira se na najčešće korištene tehnike anonimizacije u praksi, poput supstitucije, miješanja podataka, dodavanja šuma, poništavanja, maskiranja simbolom, kriptografskih tehnika i generalizacije.

U trećem poglavlju analizirani su postojeći alati za anonimizaciju podataka, kao što su ARX, μ-ARGUS, SDCMicro i Amnesia, koji omogućuju različite pristupe i razine zaštite privatnosti.

Posljednje, četvrto poglavlje detaljno opisuje praktični dio diplomskog rada, u kojem je razvijena aplikacija za anonimizaciju podataka pohranjenjih u relacijskim bazama podataka pod nazivom AnonyDB. U ovom poglavlju detaljno je opisan cijeli proces razvoja aplikacije, uključujući odabir tehnologija, implementaciju ključnih funkcionalnosti i demonstraciju rezultata anonimizacije na demonstracijskoj bazi podataka. Aplikacija omogućava korisnicima da primjenjuju različite tehnike zaštite privatnosti, kao što su hashiranje, supresija i dodavanje šuma.

Summary

The protection of personal data has become crucial in today's digital society, especially in the context of relational databases that process large amounts of sensitive information on a daily basis. Data anonymization ensures the protection of users' privacy by reducing the risk of re-identification, enabling the secure storage, processing, and exchange of data, while also helping to comply with legal requirements such as the General Data Protection Regulation.

The first chapter of this paper defines key terms such as pseudonymization and anonymization, and the differences between them, explaining how each technique contributes to privacy protection.

The second chapter focuses on the most commonly used anonymization techniques in practice, such as substitution, data mixing, noise addition, suppression, symbol masking, cryptographic techniques, and data generalization.

The third chapter analyzes existing data anonymization tools, such as ARX, μ-ARGUS, SDCMicro, and Amnesia, which offer various approaches and levels of privacy protection.

Finally, the fourth chapter provides a detailed description of the practical part of this thesis, in which an application for anonymizing data stored in relational databases, called AnonyDB, was developed. This chapter describes the entire development process of the application, including the selection of technologies, the implementation of key functionalities, and the demonstration of anonymization results on a sample database. The application allows users to apply various privacy protection techniques, such as hashing, suppression, and noise addition.

Životopis

Rođena sam 2. prosinca 1999. godine u Varaždinu. Osnovnu školu Trnovec završila sam 2014. godine, nakon čega sam upisala Prvu gimnaziju Varaždin koju sam završila 2018. godine. Iste godine upisujem preddiplomski studij *Matematika* na matematičkom odsjeku Prirodoslovno-matematičkog fakulteta u Zagrebu koji sam završila 2022. godine. Svoje školovanje nastavila sam na istom odsjeku, upisavši diplomski studij *Računarstvo i matematika*.