Jurnal Multidisiplin Teknik, Sains, Pendidikan dan Teknologi

E-ISSN: 3064-6103
Volume 3, **Number** 3, Nopember 2026
https://doi.org/10.62671/suliwa.v3i3.366

IMPLEMENTASI K-ANONYMITY UNTUK MENGURANGI RISIKO IDENTIFIKASI PADA DATASET DATA PRIBADI

Fathimah Az Zahra[1], Taufiq[2] , Yanita[3]
[1,2,3] Universitas Malikussaleh, Indonesia
[1*]fathimah.240170199@mhs.unimal.ac.id, [2]taufiq.te@unimal.ac.id, [3]yanita@unimal.ac.id

ABSTRACT

Sharing personal data for research and analysis creates a risk of identifying individuals, because removing direct identifiers alone is not enough. This study implements k-anonymity on 16,007 records of the UCI Adult dataset using five quasi-identifiers (age, sex, race, marital status, native country), with income as the sensitive attribute. The method is full-domain generalization with record suppression (5% limit), evaluating all 180 combinations of generalization levels for k = 2, 3, 5, and 10 in Python. The results show that unique records fell from 1,352 (8.45%) to zero for all values of k, maximum re-identification risk fell to 1/k, and average risk fell from 0.1442 to 0.0341 (k = 2) and down to 0.0045 (k = 10). This protection came with a rise in total information loss from 0.0922 to 0.2471, with 4.2% to 4.7% of records suppressed and age fully hidden at k = 10. In this configuration, k = 5 was a balanced choice between privacy and utility. However, 1.51% to 12.65% of records still sit in classes that are homogeneous in income, leaving them vulnerable to homogeneity attacks, so k-anonymity should be complemented with l-diversity or t-closeness.

Keywords:
k-anonymity, data privacy, anonymization, personal data, re-identification risk.

PENDAHULUAN

Perkembangan teknologi digital membuat data pribadi dikumpulkan, disimpan, dan dibagikan dalam jumlah yang semakin besar, baik oleh instansi pemerintah, lembaga kesehatan, maupun perusahaan. Pembagian data tersebut penting untuk penelitian, penyusunan kebijakan, dan analisis, tetapi menimbulkan risiko bagi privasi individu yang datanya terkandung di dalamnya. Di Indonesia, Undang-Undang Nomor 27 Tahun 2022 tentang Pelindungan Data Pribadi (UU PDP) mendefinisikan data pribadi sebagai data tentang orang perseorangan yang teridentifikasi atau dapat diidentifikasi secara tersendiri atau dikombinasi dengan informasi lainnya (Republik Indonesia, 2022). Definisi ini penting karena menunjukkan bahwa data tanpa nama pun tetap dapat tergolong data pribadi apabila individunya dapat dikenali melalui kombinasi atributnya.

Anggapan bahwa menghapus identitas langsung seperti nama dan nomor identitas sudah cukup untuk menjaga privasi telah lama dibantah. Atribut seperti usia, jenis kelamin, dan kode pos, yang disebut quasi-identifier, dapat digabungkan dan ditautkan dengan data publik lain untuk mengenali seseorang (Olatunji et al., 2024). Rocher et al. (2019) memperkirakan bahwa 99,98% penduduk Amerika Serikat dapat dikenali kembali dengan benar pada dataset apa pun yang memuat 15 atribut demografis. Pengenalan kembali individu juga terbukti mungkin pada dataset yang hanya berupa sampel atau tidak lengkap, sehingga praktik de-identifikasi sederhana dinilai tidak memadai (Rocher et al., 2019).

Salah satu model yang dirancang untuk mengatasi masalah ini adalah k-anonymity (Domingo-Ferrer et al., 2016; Majeed & Lee, 2021). Suatu data memenuhi k-anonymity apabila setiap kombinasi nilai quasi-identifier dimiliki oleh paling sedikit k record, sehingga peluang mengenali satu individu tidak lebih dari 1/k. Tingkat k yang lebih tinggi memberikan perlindungan yang lebih kuat tetapi menuntut perubahan data yang lebih besar sehingga kegunaan data menurun (Olatunji et al., 2024; Arbelaez & Climent, 2025). Menemukan anonimisasi yang optimal juga merupakan masalah yang sulit secara komputasi (Arbelaez & Climent, 2025). Di sisi lain, k-anonymity memiliki kelemahan karena tidak melindungi atribut sensitif apabila seluruh anggota suatu kelompok memiliki nilai sensitif yang sama (Olatunji et al., 2024).

Survei Majeed dan Lee (2021) menunjukkan bahwa teknik anonimisasi perlu dipilih dengan mempertimbangkan karakteristik data, model ancaman, dan tujuan penggunaan. Templ dan Sariyar (2022) juga membahas beragam metode perlindungan data untuk kebutuhan analisis, sedangkan panduan NIST oleh Garfinkel et al. (2023) menekankan pentingnya evaluasi risiko pengungkapan dan tata kelola dalam proses de-identifikasi.

Berdasarkan hal tersebut, penelitian ini mengimplementasikan k-anonymity pada dataset UCI Adult dengan lima quasi-identifier, menggunakan generalisasi dan supresi, kemudian mengevaluasi hasilnya pada nilai k = 2, 3, 5, dan 10. Rumusan masalah penelitian ini adalah bagaimana k-anonymity diimplementasikan pada dataset data pribadi, seberapa besar penurunan risiko identifikasi pada berbagai nilai k, dan bagaimana dampaknya terhadap kegunaan data (data

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Jurnal Multidisiplin Teknik, Sains, Pendidikan dan Teknologi

E-ISSN: 3064-6103
Volume 3, **Number** 3, Nopember 2026
https://doi.org/10.62671/suliwa.v3i3.366

utility). Tujuan penelitian ini adalah mengimplementasikan k-anonymity pada dataset data pribadi menggunakan generalisasi dan supresi, mengukur penurunan risiko identifikasi pada berbagai nilai k, serta menganalisis dampak anonimisasi terhadap kegunaan data. Secara teoretis, penelitian ini memberikan gambaran kuantitatif mengenai trade-off antara privasi dan utilitas pada k-anonymity. Secara praktis, penelitian ini menjadi acuan bagi pengelola data dalam memilih nilai k dan memahami keterbatasan k-anonymity sebelum membagikan data yang mengandung data pribadi.

TINJAUAN PUSTAKA

Data Pribadi dan Klasifikasi Atribut

Menurut UU Nomor 27 Tahun 2022, data pribadi adalah data tentang orang perseorangan yang teridentifikasi atau dapat diidentifikasi secara tersendiri atau dikombinasi dengan informasi lainnya, baik secara langsung maupun tidak langsung (Republik Indonesia, 2022). Dalam anonimisasi data tabular, atribut dikelompokkan berdasarkan perannya dalam risiko pengungkapan (Olatunji et al., 2024; Domingo-Ferrer et al., 2016). Identifier dapat mengidentifikasi seseorang secara langsung, misalnya nama, nomor identitas, dan nomor telepon, sehingga harus dihapus sebelum data dirilis. Quasi-identifier, misalnya usia, jenis kelamin, ras, dan kode pos, tidak mengidentifikasi sendiri tetapi bila digabungkan dan ditautkan dengan data lain dapat mengenali individu. Sensitive attribute, misalnya penyakit dan pendapatan, adalah informasi yang ingin dilindungi dari penyingkapan.

Risiko Re-identifikasi dan Linkage Attack

Linkage attack adalah serangan yang menautkan data yang dirilis dengan data publik lain melalui quasi-identifier yang sama sehingga identitas individu dapat diketahui (Olatunji et al., 2024). Contoh klasik adalah penautan data medis yang identitasnya telah dihapus dengan daftar pemilih menggunakan kombinasi kode pos, jenis kelamin, dan tanggal lahir (Olatunji et al., 2024). Pengenalan kembali juga dimungkinkan pada dataset yang hanya berupa sampel atau tidak lengkap, sehingga sampling saja tidak menjamin anonimitas (Rocher et al., 2019).

Risiko re-identifikasi dikaji berdasarkan skenario penyerang. Pada skenario prosecutor, penyerang ingin mengenali individu tertentu yang diketahui ada di dalam dataset, sedangkan pada skenario marketer penyerang ingin mengenali sebanyak mungkin record (Prasser et al., 2020). Pada data yang dibagi menjadi kelas-kelas ekuivalensi berdasarkan quasi-identifier, risiko sebuah record dapat dihitung sebagai 1 dibagi ukuran kelasnya. Record unik (kelas berukuran 1) memiliki risiko tertinggi, yaitu pasti dapat dikenali oleh penyerang yang mengetahui nilai quasi-identifier-nya.

Konsep K-Anonymity

Menurut Domingo-Ferrer et al. (2016), k-anonymity adalah syarat bahwa informasi setiap individu dalam data yang dirilis tidak dapat dibedakan dari paling sedikit k-1 individu lain. Dalam data tabular, artinya setiap kombinasi nilai quasi-identifier harus dimiliki oleh paling sedikit k record, sehingga peluang mengenali satu individu tidak lebih dari 1/k. Syarat ini dipenuhi dengan dua operasi utama (Domingo-Ferrer et al., 2016). Pertama, generalisasi, yaitu mengganti nilai atribut dengan nilai yang lebih umum menurut suatu hierarki, misalnya usia 31 menjadi rentang 30-39 atau kode pos 24351 menjadi 2435*. Kedua, supresi, yaitu menghapus record atau nilai yang masih terlalu unik setelah generalisasi. Generalisasi sebaiknya tidak dilakukan lebih dari yang diperlukan untuk mencapai k-anonymity karena generalisasi berlebihan menurunkan mutu analisis data (Olatunji et al., 2024). Selain generalisasi dan supresi, terdapat teknik lain seperti bucketization, perturbation, dan microaggregation (Olatunji et al., 2024). Contoh penerapan k-anonymity dengan k = 2 ditunjukkan pada Tabel 1.

Tabel 1. Contoh penerapan k-anonymity dengan k = 2

Usia	Jenis kelamin	Kode pos	Usia (hasil)	Jenis kelamin (hasil)	Kode pos (hasil)
31	L	24351	30-39	*	2435*
34	P	24353	30-39	*	2435*
52	L	24412	50-59	*	2441*
57	P	24415	50-59	*	2441*

Pada pendekatan full-domain, suatu atribut digeneralisasi ke level yang sama untuk seluruh record, sehingga kombinasi level generalisasi seluruh atribut membentuk kisi (lattice) yang dapat ditelusuri (Prasser et al., 2020; Olatunji et al., 2024).

Kelemahan K-Anonymity dan Pengembangannya

K-anonymity melindungi identitas individu, tetapi tidak melindungi atribut sensitif. Olatunji et al. (2024) merangkum dua serangan utama yang dapat menembus k-anonymity. Pada homogeneity attack, seluruh record dalam satu kelas memiliki nilai sensitif yang sama sehingga nilai tersebut dapat diketahui tanpa perlu mengenali record yang

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Jurnal Multidisiplin Teknik, Sains, Pendidikan dan Teknologi

E-ISSN: 3064-6103
Volume 3, **Number** 3, Nopember 2026
https://doi.org/10.62671/suliwa.v3i3.366

tepat. Pada background knowledge attack, penyerang memanfaatkan pengetahuan tambahan untuk mempersempit kemungkinan nilai sensitif.

Untuk mengatasinya telah diusulkan beberapa pengembangan. L-diversity mensyaratkan setiap kelas ekuivalensi memuat setidaknya l nilai sensitif yang berbeda atau terwakili dengan baik (Olatunji et al., 2024). T-closeness mensyaratkan distribusi nilai sensitif dalam setiap kelas tidak jauh berbeda dari distribusi pada seluruh data, diukur dengan jarak tertentu seperti Earth Mover's Distance (Olatunji et al., 2024). Differential privacy memberikan jaminan privasi dengan menambahkan derau (noise) pada hasil kueri atau data (Dwork et al., 2019; Olatunji et al., 2024), tetapi dapat mengorbankan ketelitian analisis (Olatunji et al., 2024). Pendekatan-pendekatan tersebut dapat dikombinasikan. Robinson et al. (2020), misalnya, mengusulkan algoritma anonimisasi aliran data yang menjamin k-anonymity, l-diversity, dan differential privacy sekaligus.

Algoritma K-Anonymity

Menemukan k-anonymity yang optimal, yaitu dengan kehilangan informasi minimum, dikenal sebagai masalah NP-hard (Arbelaez & Climent, 2025). Karena itu berbagai algoritma dikembangkan. Salah satu pendekatan memandang k-anonymity sebagai masalah pengelompokan (clustering), yaitu mencari kelompok beranggota minimal k record yang kemudian digeneralisasi bersama (Olatunji et al., 2024). Liang dan Samavi (2020) mengusulkan heuristik l-greedy dan Split & Carry, tetapi Split & Carry hanya berjalan pada atribut numerik dan nilai k yang kecil ($k \leq 5$). Algoritma k-members dan l-greedy termasuk heuristik terdepan untuk dataset besar, dan Arbelaez dan Climent (2025) menunjukkan bahwa kerangka local search yang mereka usulkan mampu menurunkan information loss hingga 54% dan 43% dibandingkan keduanya. Untuk pendekatan berbasis kisi generalisasi dengan hierarki yang telah ditentukan, perangkat ARX menyediakan berbagai algoritma pencarian, termasuk breadth-first dan algoritma genetika (Prasser et al., 2020; Arbelaez & Climent, 2025). Untuk data dengan banyak quasi-identifier, Prasser et al. (2016) mengembangkan algoritma Lightning untuk anonimisasi data berdimensi tinggi yang berorientasi pada utilitas. Untuk jumlah quasi-identifier yang kecil dan hierarki yang sederhana, pencarian menyeluruh terhadap seluruh kisi masih layak dilakukan sehingga solusi optimal dapat dijamin.

Metrik Evaluasi

Evaluasi k-anonymity menimbang dua hal: seberapa besar risiko berkurang dan seberapa besar informasi hilang. Risiko re-identifikasi dihitung dari ukuran kelas ekuivalensi. Risiko maksimum bergantung pada kelas terkecil (1/k), sedangkan risiko rata-rata menggambarkan peluang rata-rata seluruh record dikenali (Prasser et al., 2020). Generalization height mengukur tingkat generalisasi relatif terhadap tinggi hierarki masing-masing atribut, kemudian dirata-ratakan (Olatunji et al., 2024; Arbelaez & Climent, 2025). Discernibility Metric (DM) memberi penalti pada setiap record sebesar ukuran kelasnya, dan record yang disupresi dikenai penalti sebesar ukuran dataset; DM termasuk ukuran utilitas yang lazim dipakai untuk menilai kehilangan informasi (Olatunji et al., 2024). C_AVG membandingkan ukuran kelas rata-rata dengan k (lihat Persamaan 7), dengan nilai yang mendekati 1 menunjukkan kelas yang efisien. Terakhir, persentase record homogen, yaitu proporsi record yang berada di kelas dengan hanya satu nilai sensitif, menggambarkan kerentanan terhadap homogeneity attack karena kelas tersebut tidak memenuhi l-diversity dengan $l = 2$.

Penelitian Terkait

Beberapa penelitian terkait diringkas pada Tabel 2.

Tabel 2. Ringkasan penelitian terkait

Peneliti	Fokus dan data	Temuan utama
Arbelaez & Climent (2025)	Kerangka iterated local search untuk k-anonymization dengan meminimalkan information loss, diuji pada dataset IPUMS, Adult, dan YouTube	Menurunkan information loss hingga 54% dan 43% dibandingkan heuristik k-members dan l-greedy, dua algoritma terdepan untuk dataset besar.
Liang & Samavi (2020)	Formulasi k-anonymity sebagai masalah optimasi matematis, dengan heuristik l-greedy dan Split & Carry	l-greedy sedikit lebih baik daripada k-member pada dataset besar, sedangkan Split & Carry hanya berjalan pada atribut numerik dan $k \leq 5$.
Robinson et al. (2020)	CASTLEGUARD, anonimisasi aliran data	Menjamin k-anonymity, l-diversity, dan differential privacy sekaligus sambil menjaga utilitas untuk machine learning.
Olatunji et al. (2024)	Tinjauan teknik anonimisasi (k-anonymity, l-diversity, t-closeness) pada data kesehatan, disertai serangan	Anonimisasi diperlukan tetapi tidak cukup untuk melindungi privasi pasien.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Jurnal Multidisiplin Teknik, Sains, Pendidikan dan Teknologi

E-ISSN: 3064-6103
Volume 3, **Number** 3, Nopember 2026
https://doi.org/10.62671/suliwa.v3i3.366

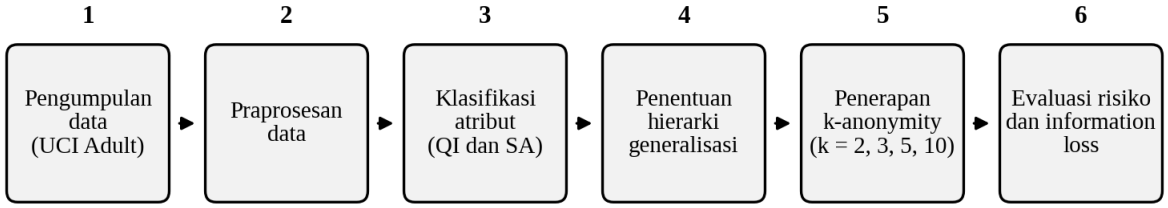
Ronal et al. (2026)	rekonstruksi pada dataset MIMIC-III k-anonymity ($k \geq 5$) dengan evaluasi l-diversity dan t-closeness pada 55.500 data medis (quasi-identifier: usia, jenis kelamin, golongan darah)	Seluruh kelas memenuhi $k \geq 5$ dengan supresi 1,15%, tidak ada kelas dengan $l < 2$, dan NGIL sebesar 0,079.

Berbeda dengan penelitian-penelitian tersebut, penelitian ini memusatkan perhatian pada gambaran kuantitatif yang transparan mengenai pengaruh peningkatan nilai k terhadap risiko dan kegunaan data pada dataset demografis publik UCI Adult. Penelitian ini menggunakan lima quasi-identifier, pencarian menyeluruh pada kisi generalisasi, serta metrik risiko, kehilangan informasi, dan kerentanan terhadap homogeneity attack dalam satu kerangka evaluasi, dan mengaitkan temuannya dengan konteks UU PDP.

METODE PENELITIAN

Pendekatan dan Alur Penelitian

Penelitian ini menggunakan pendekatan eksperimen komputasional, yaitu menerapkan algoritma anonimisasi pada sebuah dataset publik, kemudian mengukur dampaknya terhadap risiko identifikasi dan kualitas data. Karena data diperoleh dari dataset yang sudah tersedia untuk umum dan bukan dari responden, penelitian ini tidak menggunakan kuesioner maupun wawancara. Penelitian dilakukan melalui enam tahap yang ditunjukkan pada Gambar 1. Tahap pertama adalah pengumpulan data, yaitu mengunduh dataset UCI Adult. Tahap kedua adalah praprosesan data, yaitu merapikan format data dan membuang record yang bernilai kosong. Tahap ketiga adalah klasifikasi atribut untuk menentukan quasi-identifier dan sensitive attribute. Tahap keempat adalah penentuan hierarki generalisasi, yaitu menyusun tingkat generalisasi untuk setiap quasi-identifier. Tahap kelima adalah penerapan k-anonymity, yaitu mencari kombinasi tingkat generalisasi terbaik untuk setiap nilai k, lalu menekan (supresi) record yang berada di kelompok yang terlalu kecil. Tahap keenam adalah evaluasi, yaitu menghitung risiko re-identifikasi, kehilangan informasi, dan kerentanan terhadap homogeneity attack, sebelum dan sesudah anonimisasi.



Gambar 1. Alur penelitian

Dataset dan Praprosesan

Dataset yang digunakan adalah UCI Adult (Census Income) dari UCI Machine Learning Repository (https://archive.ics.uci.edu/dataset/2/adult), yaitu data hasil ekstraksi dari basis data sensus penduduk Amerika Serikat tahun 1994 yang banyak dipakai dalam penelitian privasi data dan machine learning, termasuk sebagai dataset yang paling sering digunakan untuk menguji algoritma k-anonymization (Arbelaez & Climent, 2025). Penelitian ini memakai berkas data uji (adult_test.csv) yang berisi 16.281 record dan 15 atribut. Dataset ini tidak memuat identitas langsung seperti nama atau nomor identitas, dan bersifat publik sehingga tidak menimbulkan persoalan etika terkait persetujuan responden.

Berkas yang digunakan memiliki beberapa ketidaksesuaian format sehingga dilakukan praprosesan. Baris kedua yang berisi teks bukan data dihapus, spasi di awal nilai teks dihapus (misalnya " Private" menjadi "Private"), dan nama kolom disamakan agar konsisten dalam program. Tanda titik di akhir nilai income dihapus (misalnya "<=50K." menjadi "<=50K"), tanda tanya dan sel kosong diperlakukan sebagai data kosong, dan record yang memiliki nilai kosong pada atribut yang dipakai dibuang. Dari 16.281 record awal, 274 record dibuang karena nilai native-country kosong, sehingga tersisa 16.007 record (N = 16.007) yang menjadi dasar seluruh analisis. Pada data akhir ini, sebanyak 12.235 record berpendapatan <=50K dan 3.772 record berpendapatan >50K.

Quasi-Identifier dan Sensitive Attribute

Atribut dikelompokkan berdasarkan perannya dalam risiko identifikasi, seperti pada Tabel 3. Kelima quasi-identifier dipilih karena merupakan karakteristik demografis dasar yang sering tersedia di data publik lain, sehingga berpotensi dipakai dalam serangan penautan (linkage attack) seperti yang ditunjukkan oleh Rocher et al. (2019).

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Jurnal Multidisiplin Teknik, Sains, Pendidikan dan Teknologi

E-ISSN: 3064-6103
Volume 3, **Number** 3, Nopember 2026
https://doi.org/10.62671/suliwa.v3i3.366

Tabel 3. Klasifikasi atribut

Kategori	Atribut	Keterangan
Identifier langsung	Tidak ada	Dataset tidak memuat nama atau nomor identitas.
Quasi-identifier	age, sex, race, marital-status, native-country	Atribut demografis yang lazim bisa diperoleh dari sumber eksternal dan, bila digabungkan, dapat membedakan individu.
Sensitive attribute	income	Informasi yang ingin dilindungi dari penyingkapan.
Tidak digunakan	workclass, fnlwgt, education, education-num, occupation, relationship, capital-gain, capital-loss, hours-per-week	Dikeluarkan dari analisis untuk membatasi ruang lingkup penelitian.

Tabel 3 membagi lima belas atribut dataset ke dalam empat kategori menurut perannya terhadap risiko identifikasi. Kategori identifier langsung kosong karena dataset UCI Adult tidak memuat nama maupun nomor identitas, sehingga tidak ada atribut yang perlu dihapus pada tahap ini. Lima atribut, yaitu age, sex, race, marital-status, dan native-country, ditetapkan sebagai quasi-identifier. Masing-masing atribut tidak mengidentifikasi seseorang secara tunggal, tetapi kombinasinya dapat membedakan individu dan lazim tersedia pada sumber data lain, misalnya data kependudukan atau media sosial. Atribut income ditetapkan sebagai sensitive attribute karena informasi penghasilan termasuk informasi finansial pribadi yang perlu dilindungi dari penyingkapan; pada data akhir terdapat 12.235 record (76,4%) berpendapatan <=50K dan 3.772 record (23,6%) berpendapatan >50K. Sembilan atribut lainnya tidak digunakan agar ruang lingkup penelitian tetap terfokus, karena setiap quasi-identifier tambahan akan melipatgandakan jumlah kombinasi generalisasi yang harus dievaluasi.

Pembagian ini menjadi dasar seluruh perhitungan berikutnya: risiko re-identifikasi dihitung dari kombinasi kelima quasi-identifier, sedangkan kerentanan terhadap homogeneity attack diukur terhadap nilai income. Apabila atribut lain seperti occupation atau relationship turut dijadikan quasi-identifier, jumlah record yang unik berpotensi lebih banyak, sehingga angka risiko pada penelitian ini sebaiknya dipahami sebagai hasil pada skenario lima quasi-identifier.

Hierarki Generalisasi

Generalisasi dilakukan dengan mengganti nilai quasi-identifier menjadi nilai yang lebih umum. Level 0 adalah nilai asli, dan semakin tinggi level, semakin umum nilainya. Level tertinggi selalu berupa tanda bintang (*) yang menyembunyikan nilai sepenuhnya. Hierarki yang digunakan disajikan pada Tabel 4.

Tabel 4. Hierarki generalisasi

Atribut	Level 0	Level 1	Level 2	Level 3	Level 4
age	Usia asli (27)	Rentang 5 tahun (25-29)	Rentang 10 tahun (20-29)	Rentang 20 tahun (20-39)	*
sex	Nilai asli	*	-	-	-
race	Nilai asli	*	-	-	-
marital-status	7 kategori asli	3 kelompok	*	-	-
native-country	Negara asli	United-States / Non-US	*	-	-

Tabel 4 menunjukkan jenjang generalisasi setiap quasi-identifier. Setiap baris adalah satu atribut dan setiap kolom adalah level generalisasi; level 0 adalah nilai asli, level yang lebih tinggi berarti nilai yang lebih umum, dan tanda strip berarti atribut tersebut tidak memiliki level itu. Atribut age memiliki jenjang paling bertahap dengan lima level (0 sampai 4), yaitu usia asli, rentang 5 tahun, rentang 10 tahun, rentang 20 tahun, dan akhirnya disembunyikan (*). Atribut sex dan race hanya memiliki dua level, yaitu nilai asli atau disembunyikan, sehingga tidak ada bentuk antara. Atribut marital-status memiliki tiga level (tujuh kategori asli, tiga kelompok, lalu *), dan native-country juga memiliki tiga level (negara asli, United-States atau Non-US, lalu *).

Jumlah level pada tabel ini menentukan ruang pencarian dan besarnya biaya generalisasi. Dengan 5, 2, 2, 3, dan 3 level, terdapat 180 kombinasi yang dievaluasi. Biaya generalisasi suatu atribut dihitung dari level yang dipilih dibagi level maksimumnya (Persamaan 4). Sebagai contoh, usia pada level 1 bernilai 1/4 = 0,25, sedangkan menyembunyikan sex atau race sepenuhnya bernilai 1/1 = 1. Karena kelima atribut dirata-ratakan dengan bobot sama, usia pada level 1 menyumbang 0,05 terhadap loss generalisasi (0,25/5), dan menyembunyikan satu atribut sepenuhnya menyumbang 0,2 (1/5). Dengan demikian usia memberikan kontrol yang paling halus karena dapat dinaikkan secara bertahap, sedangkan sex dan race hanya dapat dipertahankan atau dihapus seluruhnya.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Jurnal Multidisiplin Teknik, Sains, Pendidikan dan Teknologi

E-ISSN: 3064-6103
Volume 3, **Number** 3, Nopember 2026
https://doi.org/10.62671/suliwa.v3i3.366

Pada level 1 atribut marital-status, kategori asli dikelompokkan menjadi Menikah (Married-civ-spouse, Married-spouse-absent, Married-AF-spouse), Belum menikah (Never-married), dan Pernah menikah (Divorced, Separated, Widowed). Dengan jumlah level 5, 2, 2, 3, dan 3 untuk kelima atribut, terdapat 5 x 2 x 2 x 3 x 3 = 180 kombinasi level generalisasi yang mungkin.

Penerapan K-Anonymity dan Skenario Pengujian

Suatu dataset memenuhi k-anonymity apabila setiap kombinasi nilai quasi-identifier dimiliki oleh paling sedikit k record (Domingo-Ferrer et al., 2016). Kelompok record dengan nilai quasi-identifier yang sama disebut kelas ekuivalensi. Penelitian ini menggunakan full-domain generalization dengan supresi record, yaitu setiap atribut digeneralisasi ke level yang sama untuk seluruh record (Prasser et al., 2020). Karena ruang pencariannya hanya 180 kombinasi, seluruh kombinasi dievaluasi secara menyeluruh (exhaustive search) tanpa teknik pemangkasan ruang pencarian. Untuk setiap nilai k, prosedurnya adalah sebagai berikut.

1. Untuk setiap kombinasi level, bentuk dataset hasil generalisasi dan kelompokkan record ke dalam kelas ekuivalensi.
2. Hitung rasio supresi, yaitu proporsi record yang berada di kelas berukuran kurang dari k.
3. Terima kombinasi tersebut bila rasio supresi tidak melebihi batas 5%.
4. Dari kombinasi yang diterima, pilih yang memiliki information loss total terkecil.
5. Terapkan kombinasi terpilih, lalu hapus (supresi) record yang berada di kelas berukuran kurang dari k.
6. Verifikasi bahwa seluruh kelas ekuivalensi yang tersisa berukuran minimal k.

Skenario pengujian menggunakan nilai k = 2, 3, 5, dan 10 dengan batas supresi 5%. Data asli tanpa anonimisasi (setara k = 1) digunakan sebagai pembanding. Implementasi dilakukan dengan bahasa pemrograman Python 3 pada lingkungan Google Colab, menggunakan pustaka pandas untuk pengolahan data, NumPy untuk perhitungan, Matplotlib untuk pembuatan grafik, dan modul itertools untuk membentuk kombinasi level generalisasi. Algoritma disusun sendiri tanpa memakai perangkat anonimisasi siap pakai.

Metrik Evaluasi

Kinerja anonimisasi dievaluasi pada data asli dan pada hasil setiap nilai k menggunakan dua kelompok metrik. Kelompok pertama menilai risiko. Ukuran dan jumlah kelas ekuivalensi, serta jumlah dan persentase record unik (record yang menjadi satu-satunya anggota kelasnya), menggambarkan struktur data. Risiko re-identifikasi sebuah record dihitung sebagai 1 dibagi ukuran kelasnya. Risiko maksimum (1 dibagi ukuran kelas terkecil) menggambarkan skenario prosecutor, yaitu penyerang mengetahui bahwa targetnya ada di dataset, sedangkan risiko rata-rata adalah rata-rata risiko seluruh record, yang juga dikenal sebagai marketer risk. Dataset diperlakukan sebagai populasi. Kerentanan terhadap homogeneity attack diukur dengan persentase record homogen, yaitu proporsi record yang berada di kelas dengan hanya satu nilai income. Pada record ini penyerang tetap dapat menebak nilai atribut sensitif walaupun identitas individu tidak pasti (Olatunji et al., 2024).

Kehilangan informasi akibat generalisasi dihitung sebagai rata-rata rasio level generalisasi terhadap level maksimum pada lima quasi-identifier. Kehilangan informasi total didefinisikan secara operasional sebagai jumlah loss generalisasi dan rasio supresi. Kedua komponen diberi bobot sama dalam fungsi objektif; oleh karena itu, metrik ini digunakan untuk membandingkan kandidat di dalam eksperimen dan tidak diperlakukan sebagai standar universal. Risiko rata-rata dihitung pada record yang masih tersisa setelah supresi. C_AVG didefinisikan sebagai ukuran kelas rata-rata dibagi k, sedangkan DM dihitung dari jumlah kuadrat ukuran kelas ditambah penalti N untuk setiap record yang disupresi.

Asumsi dan Batasan

Penelitian ini memiliki beberapa asumsi dan batasan. Risiko dihitung berdasarkan lima quasi-identifier yang dipilih dan tidak memodelkan seluruh informasi tambahan yang mungkin dimiliki penyerang. Hasil bergantung pada hierarki generalisasi, batas supresi 5%, dan fungsi kehilangan informasi yang digunakan. Homogenitas income diukur untuk menunjukkan potensi pengungkapan atribut sensitif; penelitian ini tidak menerapkan l-diversity atau t-closeness. Eksperimen hanya menggunakan satu dataset dan satu pendekatan anonimisasi sehingga hasilnya tidak dapat langsung digeneralisasi ke semua konteks.

Rumus dan Definisi Metrik

Misalkan N adalah jumlah record sebelum anonimisasi, N′ adalah jumlah record setelah supresi, n_i adalah ukuran kelas ekuivalensi yang memuat record ke-i, g adalah jumlah kelas ekuivalensi yang tersisa, dan m adalah jumlah quasi-identifier (m = 5). Seluruh risiko rata-rata dihitung hanya pada record yang dipertahankan setelah supresi.

$$(1)\quad r_i = 1 / n_i$$

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Jurnal Multidisiplin Teknik, Sains, Pendidikan dan Teknologi

E-ISSN: 3064-6103
Volume 3, **Number** 3, Nopember 2026
https://doi.org/10.62671/suliwa.v3i3.366

Risiko record r_i merupakan kebalikan dari ukuran kelas ekuivalensi tempat record tersebut berada. Ukuran kelas ekuivalensi yang lebih besar menghasilkan risiko identifikasi berbasis kelas yang lebih rendah.

$$(2)\ R_{max} = 1 / \min(n_i)$$

R_{max} adalah risiko maksimum pada dataset yang dievaluasi, yaitu kebalikan dari ukuran kelas ekuivalensi terkecil yang tersisa.

$$(3)\ R_{av}g = (1 / N') \times \Sigma_{i=1}{}^{N'} (1 / n_i)$$

$R_{av}g$ adalah rata-rata risiko seluruh record yang tersisa setelah supresi. Definisi ini mengikuti implementasi hitung_risiko pada kode penelitian; record yang disupresi tidak masuk ke rata-rata risiko hasil anonimisasi.

$$(4)\ Lgen = (1 / m) \times \Sigma_{j=1}{}^{m} (l_j / l_j max)$$

Lgen adalah loss generalisasi, l_j adalah level generalisasi yang dipilih untuk quasi-identifier ke-j, dan $l_j max$ adalah level maksimum yang tersedia untuk atribut tersebut. Setiap atribut diberi bobot sama dalam rata-rata ini.

$$(5)\ S = (N - N') / N$$

S adalah proporsi record yang disupresi. Untuk penyajian sebagai persentase, nilai S dikalikan 100%. Batas supresi dalam eksperimen ditetapkan sebesar 0,05 atau 5%.

$$(6)\ Ltotal = Lgen + S$$

Ltotal adalah metrik operasional kehilangan informasi total yang digunakan oleh kode untuk membandingkan kandidat. Metrik ini menjumlahkan loss generalisasi dan proporsi supresi dengan bobot sama; karena itu, nilainya merupakan definisi khusus penelitian ini, bukan satu-satunya standar universal untuk mengukur utilitas data.

$$(7)\ CAVG = (N' / g) / k$$

CAVG dihitung sebagai ukuran rata-rata kelas ekuivalensi yang tersisa dibagi nilai k. Nilai yang mendekati 1 berarti ukuran kelas rata-rata mendekati k. Pada data asli, perbandingan CAVG tidak sepenuhnya setara karena tidak ada persyaratan k-anonymity yang diterapkan.

$$(8)\ DM = \Sigma_{i=1}{}^{g} (n_i{}^2) + (N - N') \times N$$

DM merupakan Discernibility Metric pada implementasi ini: jumlah kuadrat ukuran kelas ekuivalensi yang tersisa ditambah penalti N untuk setiap record yang disupresi. Dengan definisi tersebut, nilai yang lebih rendah menunjukkan penalti yang lebih rendah.

$$(9)\ H = (\text{record pada kelas dengan satu nilai income} / N') \times 100\%$$

H adalah persentase record homogen terhadap atribut sensitif income. Suatu record dihitung homogen apabila kelas ekuivalensinya hanya memiliki satu nilai income berbeda. Metrik ini menunjukkan potensi pengungkapan atribut sensitif dan tidak sama dengan risiko identifikasi berbasis ukuran kelas.

Persamaan (1)–(9) digunakan untuk menyajikan hasil secara konsisten. Persamaan risiko berbasis kelas ekuivalensi mengasumsikan penyerang mengetahui kombinasi quasi-identifier yang digunakan; hasilnya tidak mencakup semua kemungkinan sumber informasi tambahan atau seluruh jenis serangan.

HASIL DAN PEMBAHASAN

Kondisi Dataset Sebelum Anonimisasi

Setelah praprosesan, dataset berisi 16.007 record dengan lima quasi-identifier (age, sex, race, marital-status, native-country). Berdasarkan kombinasi nilai kelima atribut tersebut, record terbagi ke dalam 2.308 kelas ekuivalensi. Ukuran kelas sangat timpang: kelas terkecil beranggotakan 1 record, kelas terbesar 197 record, dan nilai median ukuran kelas adalah 1.

Sebanyak 1.352 record (8,45%) bersifat unik, artinya kombinasi kelima atributnya tidak dimiliki orang lain di dataset, sehingga seseorang yang mengetahui kelima atribut tersebut dapat langsung menunjuk record yang dimaksud. Jumlah ini setara dengan sekitar satu dari dua belas record. Selain itu, 18,36% record berada di kelas yang beranggotakan lima record atau kurang. Dengan demikian, risiko re-identifikasi maksimum pada data asli adalah 1,0 (kepastian), dan risiko rata-ratanya 0,1442. Sebanyak 30,07% record berada di kelas yang seluruh anggotanya memiliki nilai income yang sama (homogen), sehingga nilai atribut sensitif pada record tersebut sudah dapat ditebak hanya dari keanggotaan kelasnya.

Temuan ini menunjukkan bahwa menghapus identitas langsung saja tidak cukup. Dataset ini tidak memuat nama maupun nomor identitas, tetapi lima atribut demografis dasar sudah cukup untuk membuat sebagian record dapat dikenali. Interpretasi risiko ini dibatasi pada model yang menggunakan kombinasi lima quasi-identifier tersebut dan tidak mencakup seluruh kemungkinan serangan atau sumber informasi eksternal.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Jurnal Multidisiplin Teknik, Sains, Pendidikan dan Teknologi

E-ISSN: 3064-6103
Volume 3, **Number** 3, Nopember 2026
https://doi.org/10.62671/suliwa.v3i3.366

Proses Implementasi

Untuk setiap nilai k, algoritma memilih kombinasi level generalisasi dengan information loss total terkecil. Kombinasi yang terpilih ditunjukkan pada Tabel 5.

Tabel 5. Kombinasi level generalisasi terpilih untuk setiap nilai k

k	Level (age, sex, race, marital-status, native-country)	Generalisasi yang terjadi
2	1, 0, 0, 0, 0	Usia menjadi rentang 5 tahun
3	2, 0, 0, 0, 0	Usia menjadi rentang 10 tahun
5	1, 0, 0, 0, 1	Usia menjadi rentang 5 tahun, negara menjadi United-States / Non-US
10	4, 0, 0, 0, 0	Usia disembunyikan sepenuhnya (*)

Tabel 5 menunjukkan kombinasi level yang dipilih algoritma dari 180 kombinasi untuk setiap nilai k. Kolom level berisi lima angka berurutan menurut age, sex, race, marital-status, dan native-country, dengan arti angka mengacu pada Tabel 4. Sebagai contoh, "1, 0, 0, 0, 1" pada k = 5 berarti usia pada level 1 (rentang 5 tahun), sex, race, dan marital-status tetap asli, dan native-country pada level 1 (United-States atau Non-US).

Pada k = 2 dan k = 3 hanya usia yang digeneralisasi, masing-masing ke rentang 5 tahun dan 10 tahun, sehingga loss generalisasi bernilai 0,05 dan 0,10. Pada k = 5 kombinasi yang terpilih berubah: usia kembali ke rentang 5 tahun, tetapi native-country ikut digeneralisasi sehingga loss generalisasi menjadi 0,15 ((0,25 + 0,5) / 5). Pada k = 10 usia disembunyikan sepenuhnya sehingga loss generalisasi mencapai 0,20. Hal ini menunjukkan bahwa kombinasi terbaik tidak selalu merupakan peningkatan bertahap dari kombinasi pada k sebelumnya, karena algoritma memilih kombinasi dengan information loss total terkecil yang masih memenuhi batas supresi 5%. Meskipun bentuk kombinasinya berbeda, loss generalisasi naik secara teratur sebesar 0,05 pada setiap langkah k.

Pada seluruh skenario, atribut sex, race, dan marital-status tetap pada nilai aslinya, dan usia menjadi atribut yang paling banyak digeneralisasi. Hal ini kemungkinan karena usia memiliki paling banyak nilai berbeda, sehingga menjadi penyumbang utama keunikan record. Sebagai ilustrasi, pada k = 5 usia 25 tahun menjadi rentang 25-29, sedangkan nilai native-country selain United-States diganti menjadi Non-US. Perlu dicatat pula bahwa pada setiap skenario, sekitar 4,2% hingga 4,7% record disupresi, mendekati batas 5% yang ditetapkan. Algoritma memanfaatkan jatah supresi ini untuk menghindari generalisasi yang lebih tinggi.

Hasil untuk Setiap Nilai k

Hasil pengujian pada nilai k = 2, 3, 5, dan 10 diringkas pada Tabel 6 (risiko re-identifikasi) dan Tabel 7 (kualitas data), serta divisualisasikan pada Gambar 2.

Tabel 6. Hasil penerapan k-anonymity terhadap risiko re-identifikasi (N = 16.007)

k	Record tersisa	Record disupresi	Kelas ekuivalensi	Record unik	Risiko maksimum	Risiko rata-rata
1 (asli)	16.007	0	2.308	1.352	1,0000	0,1442
2	15.331	676 (4,22%)	523	0	0,5000	0,0341
3	15.248	759 (4,74%)	282	0	0,3333	0,0185
5	15.291	716 (4,47%)	268	0	0,2000	0,0175
10	15.253	754 (4,71%)	68	0	0,1000	0,0045

Tabel 6 merangkum dampak anonimisasi terhadap risiko re-identifikasi. Kolom record tersisa adalah jumlah record setelah supresi (N′), kolom record disupresi adalah N dikurangi N′ beserta persentasenya, dan kolom kelas ekuivalensi adalah jumlah kelompok record dengan kombinasi quasi-identifier yang sama. Kolom record unik adalah jumlah record yang menjadi satu-satunya anggota kelasnya. Risiko maksimum adalah 1 dibagi ukuran kelas terkecil (Persamaan 2) dan menggambarkan skenario prosecutor, sedangkan risiko rata-rata (Persamaan 3) menggambarkan peluang rata-rata seluruh record tersisa dikenali. Baris k = 1 (asli) adalah data sebelum anonimisasi sebagai pembanding.

Pada seluruh skenario, jumlah record tersisa berkisar 15.248 hingga 15.331 (sekitar 95,3% hingga 95,8% dari 16.007 record), dan record yang disupresi sebanyak 676 hingga 759 (4,22% hingga 4,74%), sehingga tetap di bawah batas 5%. Jumlah kelas ekuivalensi turun dari 2.308 menjadi 523 (k = 2), 282 (k = 3), 268 (k = 5), dan 68 (k = 10) karena generalisasi menggabungkan kombinasi nilai yang semula berbeda. Record unik turun dari 1.352 menjadi 0 pada semua skenario, artinya tidak ada lagi record yang dapat langsung ditunjuk hanya dari kelima atributnya.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Jurnal Multidisiplin Teknik, Sains, Pendidikan dan Teknologi

E-ISSN: 3064-6103
Volume 3, **Number** 3, Nopember 2026
https://doi.org/10.62671/suliwa.v3i3.366

Risiko maksimum turun dari 1,0 pada data asli, yang berarti ada individu yang dapat dikenali dengan pasti, menjadi 0,5; 0,3333; 0,2; dan 0,1 pada k = 2, 3, 5, dan 10, sesuai batas teoretis 1/k. Risiko rata-rata turun dari 0,1442 menjadi 0,0341; 0,0185; 0,0175; dan 0,0045. Artinya, pada k = 10 peluang rata-rata mengenali satu record tinggal sekitar 0,45%, dibandingkan sekitar 14,4% pada data asli. Perubahan risiko rata-rata dari k = 3 ke k = 5 sangat kecil karena jumlah kelas hampir sama (282 dan 268); alasannya dibahas lebih lanjut setelah Gambar 2.

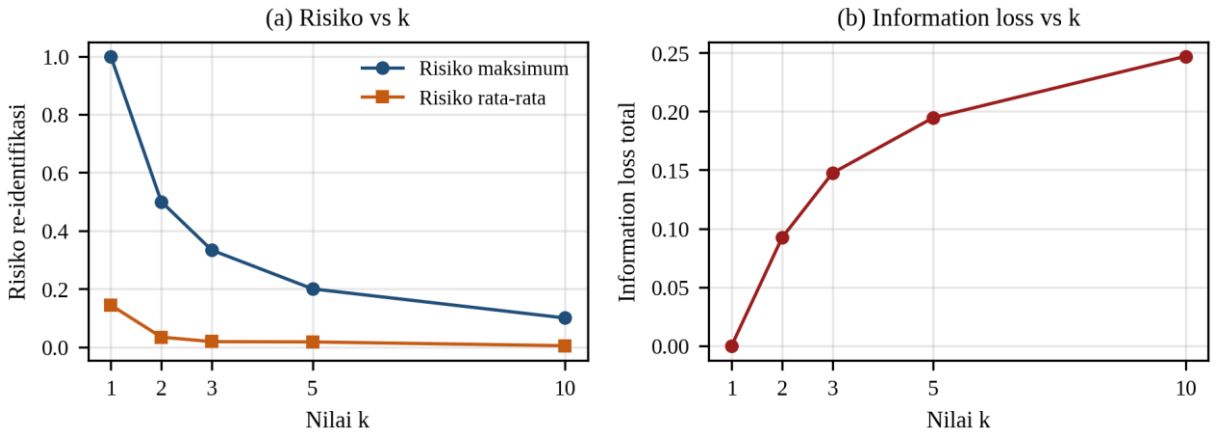
Tabel 7. Hasil penerapan k-anonymity terhadap kualitas data

k	Loss generalisasi	Information loss total	C_AVG	DM	Record homogen (%)
1 (asli)	0,0000	0,0000	6,94	1.163.151	30,07
2	0,0500	0,0922	14,66	16.356.039	12,65
3	0,1000	0,1474	18,02	22.636.899	8,60
5	0,1500	0,1947	11,41	17.020.375	10,17
10	0,2000	0,2471	22,43	53.574.807	1,51

Tabel 7 menunjukkan biaya yang harus dibayar untuk penurunan risiko pada Tabel 6. Loss generalisasi dihitung dengan Persamaan 4 dan information loss total dengan Persamaan 6, yaitu loss generalisasi ditambah proporsi supresi. C_AVG (Persamaan 7) membandingkan ukuran kelas rata-rata dengan k, dan nilai yang mendekati 1 berarti kelas-kelas tidak jauh lebih besar daripada yang diperlukan. DM (Persamaan 8) adalah jumlah kuadrat ukuran kelas ditambah penalti untuk record yang disupresi, dan nilai yang lebih rendah lebih baik. Kolom terakhir adalah persentase record homogen (Persamaan 9), yaitu record yang berada di kelas dengan hanya satu nilai income.

Loss generalisasi naik teratur dari 0,05 menjadi 0,20, dan information loss total naik dari 0,0922 menjadi 0,2471, sekitar 2,7 kali lipat. Selisih antara information loss total dan loss generalisasi pada setiap baris sama dengan proporsi supresi pada Tabel 6, yaitu 4,22% hingga 4,74%. Nilai C_AVG pada semua skenario (11,41 hingga 22,43) jauh di atas 1, artinya ukuran kelas rata-rata sekitar 11 hingga 22 kali lebih besar daripada k. Hal ini wajar karena pada full-domain generalization seluruh record digeneralisasi ke level yang sama, sehingga banyak kelas yang terbentuk berukuran jauh lebih besar daripada batas minimum k. Nilai C_AVG pada data asli (6,94) tidak sepenuhnya setara untuk dibandingkan karena tidak ada persyaratan k-anonymity pada data tersebut.

Nilai DM pada data asli (1.163.151) jauh lebih kecil daripada hasil anonimisasi karena belum ada generalisasi dan supresi. Pada hasil anonimisasi, DM mengandung penalti supresi sebesar N x (N − N′), misalnya 676 x 16.007 = 10.820.732 pada k = 2, atau sekitar 66% dari DM k = 2. Pada k = 10 penalti supresi hanya sekitar 23% dari DM, sedangkan sisanya berasal dari kelas-kelas yang sangat besar akibat usia disembunyikan, sehingga DM mencapai 53.574.807. Persentase record homogen turun dari 30,07% menjadi 12,65% (k = 2), 8,60% (k = 3), 10,17% (k = 5), dan 1,51% (k = 10). Penurunan ini menunjukkan bahwa kelas yang lebih besar cenderung mencampur kedua nilai income, tetapi persentasenya tidak pernah mencapai nol dan tidak turun secara teratur terhadap k, sehingga potensi pengungkapan atribut sensitif tetap ada.



Gambar 2. Risiko re-identifikasi (a) dan information loss (b) pada berbagai nilai k

Seluruh skenario berhasil menghilangkan record unik (dari 1.352 menjadi 0), dan ukuran kelas ekuivalensi terkecil pada setiap skenario sama dengan k sehingga syarat k-anonymity terpenuhi. Risiko maksimum turun sesuai batas teoretisnya, yaitu 1/k (0,5; 0,3333; 0,2; dan 0,1). Risiko rata-rata turun sebesar 76,4% pada k = 2, 87,2% pada k =

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Jurnal Multidisiplin Teknik, Sains, Pendidikan dan Teknologi

E-ISSN: 3064-6103
Volume 3, **Number** 3, Nopember 2026
https://doi.org/10.62671/suliwa.v3i3.366

3, 87,9% pada k = 5, dan 96,9% pada k = 10 dibandingkan data asli. Penurunan terbesar terjadi pada langkah pertama: hanya dengan k = 2, risiko rata-rata turun dari 0,1442 menjadi 0,0341.

Dari k = 3 ke k = 5, risiko rata-rata hampir tidak berubah (0,0185 menjadi 0,0175), sementara risiko maksimum turun dari 0,3333 menjadi 0,2. Hal ini terjadi karena risiko rata-rata sama dengan jumlah kelas ekuivalensi dibagi jumlah record tersisa, sehingga ia baru turun berarti bila jumlah kelas berkurang cukup banyak (282 kelas pada k = 3 dan 268 kelas pada k = 5). Pada k = 10, jumlah kelas turun tajam menjadi 68 sehingga risiko rata-rata menjadi 0,0045.

Information loss total naik dari 0,0922 (k = 2) menjadi 0,2471 (k = 10). Kenaikan ini terutama berasal dari generalisasi, yang bertambah 0,05 pada setiap langkah, sedangkan bagian supresi relatif konstan pada kisaran 4,2% hingga 4,7%. Nilai loss generalisasi 0,2 pada k = 10 berarti usia disembunyikan sepenuhnya, setara dengan kehilangan satu dari lima quasi-identifier.

Metrik C_AVG dan DM tidak naik secara teratur mengikuti k. DM pada k = 5 (17,0 juta) justru lebih kecil daripada k = 3 (22,6 juta), dan C_AVG pada k = 5 (11,41) adalah yang terendah di antara keempat skenario, walaupun tingkat generalisasinya lebih tinggi. Penyebabnya adalah kombinasi generalisasi yang terpilih berbeda sehingga menghasilkan kelas yang lebih seimbang ukurannya. Pada k = 10, DM mencapai 53,6 juta, sekitar tiga kali lipat k = 5, karena usia yang disembunyikan membentuk kelas-kelas besar, dengan kelas terbesar beranggotakan 5.509 record.

Analisis Trade-off Privasi dan Utilitas

Hasil penelitian menunjukkan trade-off yang jelas: semakin besar k, semakin kecil risiko re-identifikasi, tetapi semakin besar kehilangan informasi. Hubungan ini tidak linear. Keuntungan privasi terbesar diperoleh pada nilai k yang kecil, sedangkan biaya utilitas terus bertambah pada k yang lebih besar.

Berdasarkan data dan konfigurasi pada penelitian ini, k = 5 dapat dipandang sebagai pilihan yang seimbang. Pada k = 5, risiko maksimum turun menjadi 0,2, risiko rata-rata turun 87,9%, nilai DM lebih rendah daripada k = 3, nilai C_AVG terbaik di antara skenario teranonimisasi, dan usia masih tersedia dalam rentang 5 tahun. Nilai k = 3 tetap layak bila prioritasnya adalah menjaga detail data (information loss total 0,1474), dengan konsekuensi risiko maksimum 0,3333. Nilai k = 10 memberikan perlindungan terkuat, tetapi mengharuskan usia disembunyikan sepenuhnya dan menaikkan DM sekitar tiga kali lipat sehingga analisis yang bergantung pada usia tidak lagi dapat dilakukan.

Pilihan k yang tepat pada akhirnya bergantung pada tingkat sensitivitas data dan tujuan penggunaannya. Perlu diingat bahwa ukuran information loss yang dipakai bersifat berbasis level generalisasi dan memberi bobot yang sama pada setiap atribut. Bila suatu analisis sangat bergantung pada atribut tertentu (misalnya usia), dampak sebenarnya terhadap kegunaan data bisa lebih besar daripada yang tercermin pada angka tersebut.

Supresi juga perlu diperhatikan. Pada setiap skenario, 676 hingga 759 record dihapus karena berada di kelas yang terlalu kecil. Record ini adalah kombinasi atribut yang langka, sehingga penghapusannya dapat menggeser distribusi data, misalnya untuk kelompok yang jumlahnya sedikit. Analisis lanjutan terhadap distribusi record yang disupresi sebaiknya dilakukan bila data hasil anonimisasi akan dipakai untuk analisis statistik.

Keterbatasan dan Implikasi terhadap UU PDP

K-anonymity melindungi identitas individu, tetapi tidak menjamin perlindungan atribut sensitif. Setelah anonimisasi, persentase record homogen turun dari 30,07% menjadi 12,65% (k = 2), 8,60% (k = 3), 10,17% (k = 5), dan 1,51% (k = 10), tetapi tidak pernah mencapai nol. Pada record tersebut, penyerang tetap dapat mengetahui nilai income seseorang walaupun tidak tahu record mana yang miliknya, cukup dengan mengetahui bahwa orang tersebut berada di kelas yang homogen (Olatunji et al., 2024). Persentase ini juga tidak turun secara teratur terhadap k, karena bergantung pada kombinasi generalisasi yang terpilih. Pada k = 10 persentasenya sangat rendah terutama karena usia disembunyikan sehingga kelas menjadi sangat besar dan mencampur kedua nilai income. Temuan ini menjadi alasan untuk menggabungkan k-anonymity dengan l-diversity atau t-closeness pada penelitian lanjutan. Keterbatasan lain adalah bahwa dataset diperlakukan sebagai populasi, pengetahuan latar penyerang tidak dimodelkan, hanya satu dataset dan satu algoritma yang diuji, serta hasil bergantung pada pilihan quasi-identifier, hierarki, dan batas supresi 5%. Karena itu angka pada penelitian ini sebaiknya tidak dianggap berlaku umum untuk dataset lain.

UU Nomor 27 Tahun 2022 tentang Pelindungan Data Pribadi mencakup data yang dapat mengidentifikasi seseorang, termasuk bila dikombinasikan dengan informasi lain (Republik Indonesia, 2022). Hasil penelitian ini menunjukkan bahwa atribut yang secara terpisah tampak tidak berbahaya, seperti usia, jenis kelamin, ras, status pernikahan, dan negara asal, dapat menjadi data yang mengidentifikasi ketika digabungkan: pada data asli, 8,45% record dapat langsung dikenali. K-anonymity terbukti dapat mengurangi risiko tersebut secara terukur. Namun, UU PDP tidak mengatur nilai k tertentu, dan penerapan k-anonymity tidak otomatis menjadikan suatu data berada di luar cakupan undang-undang tersebut, karena masih terdapat risiko pengungkapan atribut sensitif seperti yang ditunjukkan oleh persentase record homogen. Dengan demikian, k-anonymity lebih tepat dipandang sebagai salah satu langkah teknis untuk mengurangi risiko, yang perlu dilengkapi dengan kontrol lain seperti pembatasan akses dan penggunaan data, dan bukan sebagai jaminan kepatuhan hukum.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Jurnal Multidisiplin Teknik, Sains, Pendidikan dan Teknologi

E-ISSN: 3064-6103
Volume 3, **Number** 3, Nopember 2026
https://doi.org/10.62671/suliwa.v3i3.366

KESIMPULAN

Penelitian ini berhasil menerapkan k-anonymity pada dataset UCI Adult (16.007 record) menggunakan full-domain generalization dengan supresi record (batas 5%) pada lima quasi-identifier, sehingga seluruh kelas ekuivalensi yang tersisa berukuran minimal k untuk k = 2, 3, 5, dan 10. Penerapan ini menurunkan record unik dari 1.352 (8,45%) menjadi 0, risiko maksimum menjadi 1/k, dan risiko rata-rata sebesar 76,4% hingga 96,9% (dari 0,1442 menjadi 0,0341 hingga 0,0045). Namun, perlindungan ini mengorbankan kegunaan data: information loss total naik dari 0,0922 menjadi 0,2471 dan sekitar 4,2% hingga 4,7% record disupresi, dengan usia disembunyikan sepenuhnya pada k = 10. Pada konfigurasi ini, k = 5 merupakan pilihan yang seimbang antara privasi dan utilitas. K-anonymity juga belum melindungi atribut sensitif, karena 1,51% hingga 12,65% record masih berada di kelas yang homogen terhadap nilai income sehingga rentan terhadap homogeneity attack. Oleh karena itu, penelitian selanjutnya disarankan menggabungkan k-anonymity dengan l-diversity atau t-closeness, membandingkannya dengan algoritma lain seperti k-members, l-greedy, atau local search, mengukur kegunaan data melalui tugas nyata seperti klasifikasi, serta menguji pada dataset yang lebih besar dan beragam. Dalam praktiknya, nilai k sebaiknya disesuaikan dengan sensitivitas data dan dilengkapi kontrol lain seperti pembatasan akses.

REFERENSI

Arbelaez, A., & Climent, L. (2025). Iterative local search for preserving data privacy. Applied Intelligence, 55, Article 189. https://doi.org/10.1007/s10489-024-05909-w

Domingo-Ferrer, J., Sánchez, D., & Soria-Comas, J. (2016). Database anonymization: Privacy models, data utility, and microaggregation-based inter-model connections. Morgan & Claypool. https://doi.org/10.2200/S00690ED1V01Y201512SPT015

Dwork, C., Kohli, N., & Mulligan, D. (2019). Differential privacy in practice: Expose your epsilons! Journal of Privacy and Confidentiality, 9(2). https://doi.org/10.29012/jpc.689

Garfinkel, S., Guttman, B., Near, J., Dajani, A. N., & Singer, P. (2023). De-identifying government datasets: Techniques and governance (NIST Special Publication 800-188). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-188

Liang, Y., & Samavi, R. (2020). Optimization-based k-anonymity algorithms. Computers & Security, 93, Article 101753. https://doi.org/10.1016/j.cose.2020.101753

Majeed, A., & Lee, S. (2021). Anonymization techniques for privacy preserving data publishing: A comprehensive survey. IEEE Access, 9, 8512–8545. https://doi.org/10.1109/ACCESS.2020.3045700

Olatunji, I. E., Rauch, J., Katzensteiner, M., & Khosla, M. (2024). A review of anonymization for healthcare data. Big Data. https://doi.org/10.1089/big.2021.0169

Prasser, F., Bild, R., Eicher, J., Spengler, H., Kohlmayer, F., & Kuhn, K. A. (2016). Lightning: Utility-driven anonymization of high-dimensional data. Transactions on Data Privacy, 9(2), 161–185. https://www.tdp.cat/issues16/abs.a223a15.php

Prasser, F., Eicher, J., Spengler, H., Bild, R., & Kuhn, K. A. (2020). Flexible data anonymization using ARX—Current status and challenges ahead. Software: Practice and Experience, 50(7), 1277–1304. https://doi.org/10.1002/spe.2812

Republik Indonesia. (2022). Undang-Undang Republik Indonesia Nomor 27 Tahun 2022 tentang Pelindungan Data Pribadi.

Robinson, A., Brown, F., Hall, N., Jackson, A., Kemp, G., & Leeke, M. (2020). CASTLEGUARD: Anonymised data streams with guaranteed differential privacy. In 2020 IEEE International Conference on Dependable, Autonomic and Secure Computing, International Conference on Pervasive Intelligence and Computing, International Conference on Cloud and Big Data Computing, International Conference on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech). IEEE. https://doi.org/10.1109/DASC-PICom-CBDCom-CyberSciTech49142.2020.00102

Rocher, L., Hendrickx, J. M., & de Montjoye, Y.-A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. Nature Communications, 10, Article 3069. https://doi.org/10.1038/s41467-019-10933-3

Ronal, Tambunan, D. E., & Yuliana. (2026). Evaluasi efektivitas teknik privacy-preserving: K-anonymity, l-diversity, t-closeness pada data sensitif. HOAQ (High Education of Organization Archive Quality): Jurnal Teknologi Informasi, 17(1), 102–111. https://doi.org/10.52972/hoaq.vol17no1.p102-111

Templ, M., & Sariyar, M. (2022). A systematic overview on methods to protect sensitive data provided for various analyses. International Journal of Information Security, 21, 1233–1246. https://doi.org/10.1007/s10207-022-00607-5

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.