K-anonymity Menggunakan *Simple Distribution of Sensitive Values* dan *Aggregation of Sensitive Values*

http://dx.doi.org/10.28932/jutisi.v10i2.8798

Riwayat Artikel
Received: 26 April 2024 | *Final Revision*: 28 Juli 2024 | *Accepted*: 28 Juli 2024

Widodo✉ #1, Muhammad Ficky Duskarnaen#2, Murien Nugraheni*3

#*Pendidikan Teknik Informatika dan Komputer, Universitas Negeri Jakarta*
Jl. Rawamangun Muka, Jakarta Timur, 13220, Indonesia
[1]widodo@unj.ac.id
[2]duskarnaen@unj.ac.id

**Sistem dan Teknologi Informasi, Universitas Negeri Jakarta*
Jl. Rawamangun Muka, Jakarta Timur, 13220, Indonesia
[3]muriennugraheni@unj.ac.id
✉Corresponding author: widodo@unj.ac.id

Abstrak — **Penganoniman microdata sangat diperlukan jika microdata tersebut akan dipublikasikan keluar maupun akan berbagi dengan pihak lain. Tujuan penganoniman tersebut adalah agar data yang bersifat sensitif tidak akan dapat diketahui oleh pihak yang tidak berkepentingan baik secara langsung maupun tidak langsung. Pada penelitian yang berkembang saat ini teknik yang banyak digunakan adalah dengan generalisasi dan supresi pada model *k-anonymity*, namun teknik ini memiliki kelemahan yaitu tingkat *information loss* yang dihasilkan cukup tinggi. Selain itu, representasi *microdata* yang dihasilkan akibat penganoniman tersebut terlalu besar, sehingga perlu disederhanakan. Pada penelitian ini akan dibangun model *anonymity* dengan menggunakan teknik distribusi atribut sensitif yaitu *Simple Distribution of Sensitive Values* (SDSV). Tujuan utama metode ini adalah mengurangi probabilitas pihak yang tidak terotorisasi dalam menebak pemilik data sensitif. Sedangkan untuk menyederhanakan representasi dari *microdata* tersebut, teknik *aggregative of sensitive value* (ASENVA) akan diterapkan. Hasil dari penelitian ini, metode SDSV memiliki tingkat *information loss* yang lebih minimal, sedangkan penggunaan ASENVA menyederhanakan representasi tabel anonim menjadi rata-rata 13.67% untuk agregat tabel *quasi-identifier* dan 6.35% untuk tabel sensitif.**

Kata kunci— **aggregative of sensitive value; k-anonymity; privacy; simple distribution of sensitive values.**

The k-anonymity using Simple Distribution of Sensitive Values and Aggregation of Sensitive Values

Abstract — **Anonymizing microdata is a matter while the microdata is published or shared. The purpose of anonymization is so that sensitive data will not be known by unauthorized parties either directly or indirectly. The technique that is widely used is generalization and suppression in the k-anonymity model, however, this technique has the disadvantage that the level of information loss is quite high. In addition, the generated microdata representation due to anonymization is too large, thus it needs**

to be simplified. In this research, an anonymity model is built using a sensitive attribute distribution technique, namely Simple Distribution of Sensitive Values (SDSV). The main purpose of this method is to reduce the probability of unauthorized parties guessing the owner of sensitive data. Meanwhile, to simplify the representation of the microdata, the aggregative of sensitive value (ASENVA) technique is applied. The result shows that the SDSV metho has less information loss compared to others, while the use of ASENVA simplifies the representation of anonymized tables to an average of 13.67% for aggregated quasi-identifier tables and 6.35% for sensitive tables.

Keywords— **aggregative of sensitive value; k-anonymity; privacy; simple distribution of sensitive values.**

I. PENDAHULUAN

Data Anonymity merupakan kondisi data yang menjadikan pihak yang tidak berkepentingan dengan data tersebut sulit untuk mengetahui identitas pemiliknya. Perkembangan bidang *data anonymity* terutama untuk microdata dimulai dengan adanya model *k-anonymity* [1]. Pembentukan model *k-anonymity* pun sudah sangat beragam metodenya, diantaranya yang cukup baru adalah dengan *blackhole algorithm* [2]. Bidang ilmu yang terkait dengan *data anonymity* ini termasuk ke dalam bidang yang dinamakan *Privacy Preserving Data Publishing* (PPDP) dan *Privacy Preserving Data Mining* (PPDM). Metode maupun teknik-teknik yang digunakan baik pada PPDP maupun pada PPDM adalah sama. Perbedaan keduanya adalah pada penggunaannya, jika PPDP lebih banyak digunakan untuk publikasi data maupun berbagi data, sedangkan PPDM lebih mengarah pada *privacy* pada proses di Data Mining [3], [4].

Model PPDP yang telah dikembangkan untuk membangun sebuah microdata menjadi lebih privat tersebut dimulai dengan lahirnya sebuah model data anonymization yang disebut k-anonymity [1], [5]. Pengembangan dari model dasar tersebut sudah cukup banyak variannya, seperti model *p-sensitive* [6], *l-diversity* [7], dan beberapa model lainnya yang dikembangkan untuk menyempurnakan kekurangan-kekurangan model sebelumnya. *k-anonymity* adalah model PPDP yang dikembangkan dengan cara melakukan generalisasi, sehingga dalam sebuah kelompok atribut *quasi-identifier* sejumlah k, akan dianonimkan.

Tabel microdata adalah tabel yang berisi empat unsur atribut (kolom) yaitu *explicit identifier*, *quasi-identifier*, *sensitive attribute*, dan *non sensitive attribute* [8]. *Explicit identifier* adalah atribut yang berisi data yang bersifat identitas, seperti nama, NIK, NIP, NIM. *Quasi-identifier* adalah atribut yang bisa menjadi kunci. Atribut ini jika dihubungkan dengan data lain yang sama bisa mengungkap identitas pemilik data atau *explicit identifier*-nya. Contoh atribut ini adalah umur, kodepos. *Sensitive attribute* adalah atribut yang memiliki tingkat sensitivitas tertentu. Contoh atribut ini adalah gejala penyakit, jenis kejahatan. Atribut *non sensitive* adalah atribut selain dari tiga atribut sebelumnya. Dari atribut-atribut tersebut, atribut *quasi-identifier*-lah yang akan digeneralisasi untuk mendapatkan model *k-anonymity*. Tabel 1 dan Tabel 2 mengilustrasikan *k-anonymity*.

TABEL 1
DATA AWAL SEBELUM DIANONIMKAN

No.	Nama	Umur	JenisKel	KodePos	Gejala
1	Asrul	23	L	13168	Flu
2	Darma	23	L	13179	Kanker
3	Arifin	26	L	13133	HIV
4	Dian	31	P	17166	Flu
5	Esri	37	P	17446	Flu
6	Sari	38	P	17801	Sakit Kepala
7	Adi	45	L	31876	Sakit Kepala
8	Yuni	41	P	35553	Diare
9	Fariz	44	L	31434	Kanker
10	Neli	47	P	31579	Flu

Tabel 1 yang merupakan tabel microdata yang belum dianonimkan, diasumsikan sebagai tabel data pasien dari sebuah rumah sakit. Tabel 1 terdiri dari sebuah atribut *explicit identifier* yaitu Nama, karena nama sebagai identitas yang sudah jelas dalam record/baris tersebut. Kemudian, atribut Umur, JenisKel, dan KodePos adalah atribut *quasi-identifier*, karena atribut-atribut ini yang bisa digeneralisasi supaya data menjadi anonim. Atribut terakhir yaitu Gejala merupakan atribut sensitif,

karena merupakan atribut yang memiliki nilai sensitivitas tertentu bagi seseorang. Sementara atribut nomor memang tidak digunakan dalam PPDP.

Cara yang paling sederhana dalam membuat sebuah microdata menjadi privat adalah dengan menghilangkan atribut *explicit identifier*, yaitu Nama. Namun jika seseorang memiliki data yang beririsan dengan data dari rumah sakit tersebut, maka akan bisa ditebak pemilik atribut sensitifnya. Biasanya data yang beririsan tersebut adalah atribut-atribut *quasi-identifier*. Misalnya, Tabel 1 dihilangkan data nama-nya, kemudian seseorang memiliki data *quasi- identifier*-nya dari sumber lain, contoh umur='23', JenisKel='L', dan KodePos='13179', maka orang tersebut bisa menebak data dari rumah sakit yang menyembunyikan identitas nama. Orang tersebut akan dengan mudah menebak bahwa yang menderita gejala Kanker adalah Darma. Tabel 2 menunjukkan hasil *k-anonymity* dari Tabel 1. Tabel 2. merupakan hasil *k-anonymity* yang terbentuk dari penganoniman tiga buah atribut *quasi-identifier*. Atribut umur di-generalisasi sesuai jangkauan tertentu, isi KodePos dilakukan supresi (menutup sejumlah karakter dengan '*'), dan JenisKel jika sama dalam satu grup tidak perlu ditutup, namun jika berbeda ditutup karakternya. Nilai-nilai *quasi-identifier* yang sama tersebut membentuk sebuah kelompok yang dinamakan kelompok *quasi-identifier* atau *equivalence class*. Dari Tabel 2. tersebut terlihat bahwa dalam satu kelompok *quasi-identifier* tidak bisa dibedakan record-nya, sehingga jika seperti contoh sebelumnya seseorang mengetahui umur='26', JenisKel='L', dan KodePos='13179', maka orang tersebut tidak bisa langsung menebak dengan tepat karena umurnya ada pada jangkauan tertentu, kode_pos juga tidak lengkap. Peluang untuk menebak dengan benar hanya 1/3 karena dalam kelompok *quasi-identifier* tersebut terdapat 3 record.

TABEL 2
MODEL K-ANONYMITY TABEL 1 DENGAN K=1

Grup_id	Umur	JenisKel	KodePos	Gejala
1	[20-29]	L	131**	Flu
	[20-29]	L	131**	Kanker
	[20-29]	L	131**	HIV
2	[30-39]	P	17***	Flu
	[30-39]	P	17***	Flu
	[30-39]	P	17***	Sakit Kepala
3	[40-49]	*	3****	Sakit Kepala
	[40-49]	*	3****	Diare
	[40-49]	*	3****	Kanker
	[40-49]	*	3****	Flu

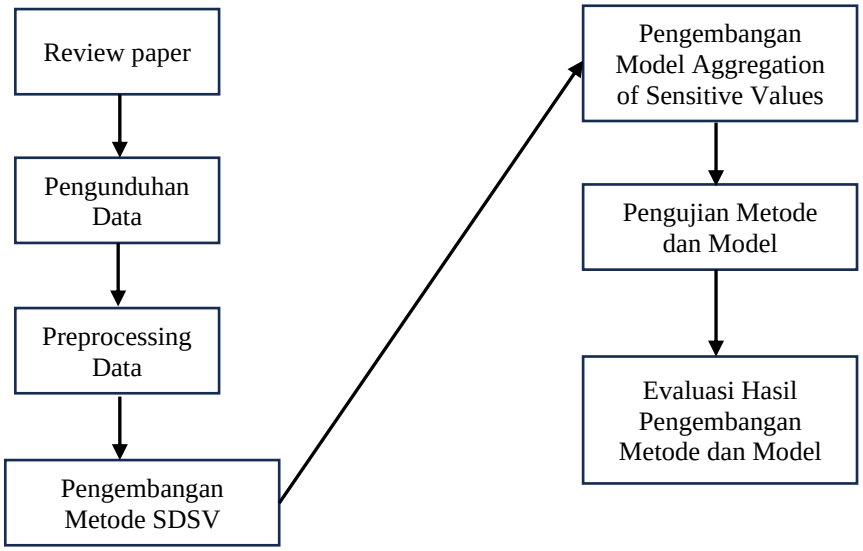
Dalam perkembangannya, salah satu permasalahan dalam membentuk model *k-anonymity* adalah pendistribusian atribut sensitif [9]. Distribusi mulai atribut sensitif yang tepat ke setiap grup *quasi-identifier* akan mengurangi peluang dari pihak yang tidak berkepentingan untuk bisa menebak dengan baik pemilik data, terutama data yang bersifat sensitif. Hanya beberapa penelitian yang sangat memperhatikan permasalahan ini. Sebagian besar penelitiannya adalah hanya mendistribusikan saja, tanpa melihat tingkat kesensitifan nilai atribut sensitif. Pada penelitian yang sebelumnya sendiri [10] metode atau algoritma yang digunakan sudah cukup baik dalam mendistribusikan nilai atribut sensitif tersebut, namun tingkat *information loss* dari data masih cukup tinggi [11]. Oleh karena itu, metode SDSV yang pernah diusulkan akan diperbaiki supaya metode SDSV bisa mengurangi tingkat *information loss* tersebut. Selain itu untuk merepresentasikan data, jika datanya sangat banyak akan cukup menyulitkan dalam pembacaan data, sehingga perlu diringkas.

Penelitian mengenai *k-anonymity* memiliki permasalahan mendasar dalam hal pendistribusian atribut sensitif. Atribut sensitive yang berisi nilai sensitif ini harus didistribusikan secara merata pada seluruh grup *quasi-identifier* supaya tidak terjadi penumpukan pada beberapa grup saja. Pendistribusian secara merata ini juga akan meminimalkan tingkat *information loss* pada saat data sudah dianonimkan. Hal lain yang menjadi permasalahan adalah representasi data yang lebih sederhana pada saat sudah dianonimkan.

Penelitian ini mempunyai kontribusi keilmuan yaitu, sebuah metode atau algoritma pendistribusian nilai atribut sensitif yang sekaligus untuk mengurangi tingkat *information loss* dan mengurangi probabilitas terungkapnya pemilik data sensitif. Kontribusi lainnya adalah pengagregasian atribut sensitif untuk menyederhanakan representasi *microdata* yang sudah dianonimkan..

II. METODE PENELITIAN

Penelitian ini merupakan penelitian yang bersifat algoritmis, sehingga metode penelitiannya adalah metode eksperimen yang dilakukan dengan menggunakan program dan simulasi komputer. Dataset diambil dari situs *open data* yaitu di website *UCI Machine Learning*. Dengan memperhatikan karakteristik *microdata*, maka data yang diunduh pada *UCI Machine Learning* adalah *adult datasets*. Tahapan penelitian dilakukan secara terstruktur seperti terlihat pada Gambar 1.



Gambar 1. Tahap Penelitian model *k-anonymity*

A. *Review Paper*

Pada tahap ini, akan dikaji beberapa paper yang berkaitan dengan penelitian yang dilakukan, yaitu paper yang berkaitan dengan *privacy preserving data publishing* dan paper mengenai distribusi atribut sensitif. Hasil yang diharapkan adalah ditemukannya *state of the art* yang mengarahkan pada fokus penelitian.

B. *Pengunduhan Data*

Pada tahap ini dilakukan pengunduhan data, data yang diunduh berasal dari repositori terbuka pada internet. Data yang diunduh adalah data yang memiliki karakteristik sebagai *microdata*. Data tersebut adalah *adult datasets* yang berasal dari repositori *online UCI Machine Learning*.

C. *Preprocessing Data*

Preprocessing data diperlukan agar data sudah memiliki kesiapan untuk diproses. Dalam penelitian ini *preprocessing* dilakukan dengan mengambil atribut-atribut yang relevan untuk digunakan dan menghilangkan *missing values*.

D. *Pengembangan Metode Simple Distribution of Sensitive Values (SDSV)*

Tahap ini melengkapi tahap sebelumnya, yaitu mengembangkan sebuah metode atau algoritma untuk mendistribusikan atribut sensitif. Pendistribusian atribut sensitif ini diperlukan agar data dengan tingkat sensitivitas tinggi tidak menumnpuk pada beberapa kelompok *quasi-identifier* saja.

E. *Pengembangan Model Aggregation of Sensitive Values*

Pada tahap ini dilakukan pengembangan model *k-anonymity* dengan pengagregasian atribut sensitif. Hal ini dilakukan supaya representasi data lebih efisien. Metode yang digunakan adalah ASENVA (*Aggregation of Sensitive Values*).

F. *Pengujian Model dan Metode*

Model dan metode yang dikembangkan akan diuji dengan cara diimplementasikan menggunakan *adult datasets*. Benchmark metode yang digunakan adalah systematic clustering, karena dianggap cukup baik dalam k-anonymity dengan menghasilkan *minimum information loss*.

G. Evaluasi Hasil Pengembangan Model dan Metode

Hasil eksperimen dan simulasi yang dilakukan dengan *adult datasets*, dievaluasi pada tahap ini. Evaluasi ini bertujuan untuk mendapatkan kinerja algoritma yang dibangun. Selain menggunakan matriks evaluasi yang sudah ada, sebuah metrics evaluasi juga diusulkan untuk menormalkan nilainya.

III. HASIL DAN PEMBAHASAN

A. Deskripsi Data

Penelitian ini merupakan penelitian yang mengimplementasikan model dan algoritma. Pada penelitian ini data yang digunakan adalah data yang diambil dari repositori *online UCI Machine Learning*, yaitu *Adult Dataset*. Dataset ini memiliki 14 atribut, namun dalam penelitian ini data yang diambil berasal dari 6 atribut yang relevan saja. Tabel 3. menjelaskan struktur dataset yang digunakan. Pada Tabel 3 terdapat 3 (tiga) buah atribut sensitif, yaitu *Education*, *WorkClass*, dan *Occupation*. Sesuai dengan model ASENVA, pada saat diagregasikan, maka ketiga atribut sensitif tersebut digabung menjadi satu saja, dengan nama atribut *Sensitive*.

TABEL 3
STRUKTUR ADULT DATASETS

No.	Nama Atribut	Jumlah isi yang unik	Jenis Atribut
1	*Age*	72	Quasi Identifier
2	*Sex*	2	Quasi Identifier
3	*MaritalStatus*	7	Quasi Identifier
4	*Education*	16	Sensitive Attribute
5	*WorkClass*	7	Sensitive Attribute
6	*Occupation*	14	Sensitive Attribute

B. Deskripsi SDSV+ dan ASENVA

Penelitian ini menghasilkan model *data anonymity* yang menggabungkan algoritma SDSV+ dan *Aggregation of Sensitive Values* (ASENVA). Algoritma SDSV+ merupakan turunan dari algoritma SDSV dengan tambahan untuk *multiple attribute sensitive*-nya digabung menjadi satu. Penggabungan ini bertujuan untuk lebih meng-efisienkan komputasi pada tabel *microdata*. Penggabungan ini tidak memperhitungkan *Primary Sensitive Attribute* (PSA) dan *High-Sensitive Values* (HSV) lagi karena atribut sensitif digabung menjadi satu. Tingkat sensitivitas nilai atribut menjadi satu kesatuan tersendiri yaitu berapa jumlah nilai sensitif pada sebuah baris pada sebuah grup *quasi-identifier*. Gambar 2.a berikut ini adalah algoritma SDSV yang original, tanpa menggabungkan atribut sensitif [12].Sedangkan Gambar 2.b., merupakan algoritma SDSV+.

1. *Create sub tables, each sub table has QI attributes and at most two sensitive attributes which have less correlation.*
2. *Form a QI group in each sub table.*
3. *Create PSA and CSA and put its HSVs on top order*
4. *Distribute evenly tuples that contain HSV in PSA to each QI group:*
    a. *If all tuples contain HSV in PSA is distributed evenly, but there are still QI group then put tuples contain HSV in CSA to next QI group, otherwise tuples randomly.*
    b. *If all groups have been filled tuples that is contained HSV in PSA, but there are still more HSV in PSA, then repeat to distribute them.*
5. *Check each sub-table for privacy guarantee to meet p-sensitive. If a sub table does not satisfy p-sensitive, then exchange a non LSV tuple in a QI group with other groups.*

Gambar 2.a. Algoritma SDSV

1. *Create sub tables, each sub table has QI attributes and at most two sensitive attributes which have less correlation.*
2. *Form a QI group in each sub table.*
3. *Create PSA and CSA and put its HSVs on top order*
4. *Distribute evenly tuples that contain HSV in PSA to each QI group:*
    a. *If all tuples contain HSV in PSA is distributed evenly, but there are still QI groups then put tuples containing HSV in CSA to next QI group, otherwise tuples randomly.*
    b. *If all groups have been filled with tuples that contain HSV in PSA, but there are still more HSV in PSA, then repeat to distribute them.*
5. *Disassociate table into a sensitive table and QI-table*
    a. *In sensitive tables, perform sensitive attribute aggregation by summing the same data with the number of its category in one record.*
    b. *In the QI table, adjust aggregately to the same QI.*
6. *Check each sub-table for privacy guarantee to meet p-sensitive. If a sub-table does not satisfy p-sensitive, then exchange a non LSV tuple in a QI group with other groups*

Gambar 2.b. Algoritma SDSV+

Sedangkan algoritma ASENVA merupakan turunan dari model anatomi. ASENVA memecah atribut sensitif dan *quasi-identifier* sehingga memperkecil kemungkinan pihak luar untuk menebak korelasi antara keduanya. Konsep yang ditambahkan ke ASENVA dari model anatomi adalah pengagregasian atribut sensitif. Konsep ini disamping semakin memperkecil korelasi antara *quasi-identifier*, nilai atribut sensitif, dan *identifier attribute*, juga meng-efisien-kan representasi microdata [13].

Evaluasi yang digunakan adalah menggunakan *diversity metrics* untuk mengukur keberagaman dalam satu grup *quasi-identifier*. Satu teknik evaluasi kinerja lainnya adalah dengan *information loss metrics* yang digunakan untuk melihat seberapa banyak informasi yang hilang akibat dilakukan anonimisasi [14]. Pengukuran dilakukan terhadap data yang dianonimisasi pada tingkat *k-anonymity*. Metode/algoritma *baseline* yang digunakan adalah sebagai *benchmark* adalah *systematic clustering* dan SDSV. Metode *baseline* digunakan sebagai pembanding kinerja algoritma yang diusulkan. *Systematic clustering* digunakan karena dianggap memiliki kinerja yang sangat baik terutama untuk *information loss* yang minimal, sedangkan SDSV digunakan karena merupakan akar dari SDSV+.

Nilai *information loss* dari sebuah tabel microdata yang dianonimkan dapat diukur dengan:

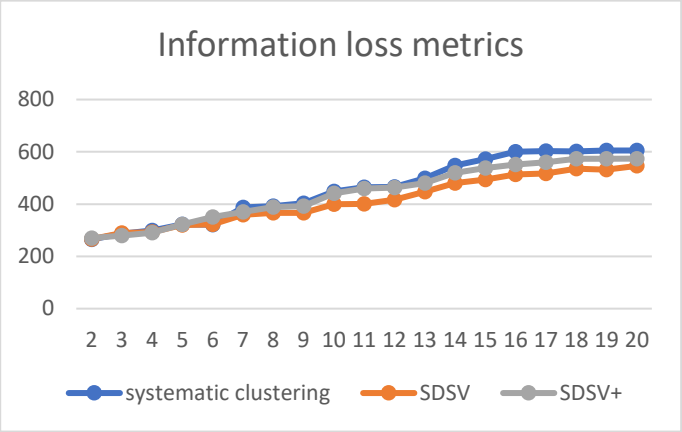
$$IL(e) = |e|.\left(\sum_{i=1\ldots n} \frac{max_{ni} - min_{ni}}{n}\right) + \left(\sum_{j=1\ldots n} \frac{H(\Lambda(\cup C_j}{H(TC_j)}\right) \qquad (1)$$

IL(e) adalah nilai *information loss* dari tabel *microdata*
|*e*| adalah jumlah grup *quasi-identifier*
Dalam kurung sebelah kiri adalah pengukuran untuk atribut numerik
Dalam kurung sebelah kanan adalah pengukuran untuk atribut non numerik

Pada pengujian yang dilakukan dengan model k-anonymity ini, parameter *k* yang digunakan adalah *k=2* sampai dengan *k=20*. *K=2* merupakan minimal parameter *k* karena *k* akan mengelompokkan baris, sehingga nilai minimal *k* adalah 2, semetara *k=20* diset menjadi nilai *k* tertinggi karena permasalahan komputasi dan mempertimbangkan jumlah data yang realistis.

Gambar 3 memperlihatkan grafik *information loss metrics* hasil eksperimen dengan metode SDSV+ dengan *baseline* yang digunakan adalah *systematic clustering* dan SDSV.

p-ISSN : 2443-2210
e-ISSN : 2443-2229



Gambar 3. Information loss metrics dari SDSV+, SDSV, dan *systematic clustering*

Langkah berikutnya adalah membangun model *anonymity* menggunakan model anatomi dan ASENVA. Baik anatomi maupun ASENVA akan memecah tabel menjadi dua, yaitu tabel *quasi-identifier* dan tabel sensitif. Perbedaannya, pada Anatomi, pemecahannya sesuai dengan *record* pada tabel *microdata*, sedangkan pada ASENVA tabel *quasi-identifier* maupun tabel sensitifnya diagregasikan. Dataset diambil dari *UCI machine learning repository* yang memiliki karakteristik sebagai microdata, yaitu *Adult Dataset* [15]. Jumlah record yang digunakan dari *Adult Datasets* adalah 7755. Tabel 4 menunjukkan hasil proses ASENVA.

TABEL 4
HASIL PENGAGREGASIAN DATA MENGGUNAKAN ASENVA (JUMLAH RECORD DATASETS = 7755)

K	Jumlah record Aggegat Tabel QI	Jumlah record Agregat Tabel Sensitif	% (2) terhadap jumlah record datasets	% (3) terhadap jumlah record datasets
(1)	(2)	(3)	(4)	(5)
2	3887	1630	50,12	21,02
3	2585	1311	33,33	16,91
4	1938	892	24,99	11,50
5	1551	710	20,00	9,16
6	1292	597	16,66	7,70
7	1107	494	14,27	6,37
8	969	443	12,50	5,71
9	861	401	11,10	5,17
10	775	366	9,99	4,72
11	705	346	9,09	4,46
12	646	322	8,33	4,15
13	596	280	7,69	3,61
14	553	263	7,13	3,39
15	517	249	6,67	3,21
16	484	235	6,24	3,03
17	456	220	5,88	2,84
18	430	211	5,54	2,72
19	408	202	5,26	2,60
20	387	186	4,99	2,40

Penurunan jumlah *record* yang yang diperlihatkan cukup signifikan. Penurunan tersebut seperti terlihat pada Tabel 4., jumlah record awal tanpa ASENVA adalah 7755, setelah dilakukan ASEVA, untuk *k=2* maka jumlah record menjadi 3887 untuk tabel *quasi-identifier* dan menjadi 1630 untuk tabel sensitif. Demikian juga untuk *k>2*, semua record menjadi lebih efisien dengan jumlah lebih kecil daripada 7755. Representasi yang seperti ini, selain lebih efisien, juga akan lebih aman karena banyak *record* yang diagregasikan sehingga keterkaitan dengan *identifier* menjadi semakin menurun. Sedangkan hasil untuk algoritma SDSV+ terlihat pada Gambar 3. Pada Gambar 3 tersebut, tingkat *information loss,* algoritma SDSV+ lebih kecil daripada *baseline* yaitu *systematic clustering*.

Dari hasil yang terlihat pada Gambar 3 dan Tabel 4, beberapa hal dapat disimpulkan. Algoritma SDSV+ yang merupakan varian dari Algoritma SDSV menunjukkan kinerja yang cukup baik. Pada Gambar 3. terlihat algoritma SDSV+ mengungguli algoritma *systematic clustering* dalam menghasilkan *information loss*. Keunggulan SDSV+ tersebut dapat terlihat pada nilai *information loss* yang lebih kecil yang berarti lebih baik. Namun jika dibandingkan dengan SDSV ternyata nilai *information loss*-nya masih lebih tinggi. Hal tersebut karena penggabungan sejumlah atribut sensitif menjadi satu, sehingga generalisasi maupun supresi-nya lebih banyak daripada SDSV. Keadaan seperti ini menyebabkan tingkat *information loss*-nya menjadi lebih tinggi.

Pada tahap berikutnya, penggunaan model ASENVA membuat representasi tabel menjadi lebih efisien. Hal ini terlihat pada persentase terhadap jumlah *record* pada dataset awal. Proses dengan menggunakan ASENVA selalu meringkas representasi tabel *microdata* tersebut, sehingga selain lebih efisien, agregasi tersebut juga akan lebih menjaga data-data yang sensitif dari probabilitas pihak luar mengaitkan dengan data *identifier*-nya. Rata-rata persentase jumlah agregat tabel *quasi-identifier* terhadap dataset secara keseluruhan adalah 13.67% dan rata-rata record agregat tabel sensitif adalah 6,35%.

IV. Simpulan

Salah satu permasalahan dalam data anonymity adalah distribusi atribut sensitif yang masih belum merata. Keadaan ini bisa mengakibatkan penumpukan atribut sensitif pada beberapa grup *quasi-identifier* saja. Penelitian ini bertujuan untuk dapat mendistribusikan nilai sensitif secara merata, mengurangi Tingkat information loss, dan membuat lebih efisien representasi dari microdata yang dianonimkan.

Berdasarkan hasil penelitian yang telah dilakukan, maka metode yang diusulkan yaitu SDSV+ dan model ASENVA bisa berjalan dengan baik menggunakan adult datasets sebagai microdata. Metode SDSV+ mengungguli *systematic clustering* sebagai *baseline method* dalam menghasilkan information loss. Sedangkan jika dibandingkan dengan SDSV, memang masih belum mengungguli. Hal ini disebabkan karena SDSV+ melakukan penggabungan atribut sensitif. Sedangkan dari representasi tabel microdata yang dianonimkan, dengan model ASENVA terbukti lebih efisien dibandingkan dengan tidak menggunakan ASENVA.

Penelitian ini masih dapat disempurnakan lagi. Penurunan nilai information loss bisa dicoba dengan memodifikasi SDSV+ dengan memperbaiki metode clustering-nya. Jika menggunakan metode clustering yang lebih bagus, maka distribusi nilai sensitif-nya bisa lebih merata.

Ucapan Terima Kasih

Ucapan terima kasih diberikan pada Fakultas Teknik Universitas Negeri Jakarta yang telah membiayai penelitian ini dengan dana BLU FT UNJ.

Daftar Pustaka

[1] P. Samarati and L. Sweeney, “Protecting privacy when disclosing information: k-anonymity and its inforcement through generalization and suppression,” Tech. rep. SRI-CSL-98-04, SRI, Computer Science Laboratory, Palo Alto, CA, USA, 1998.

[2] L Kacha, A. Zitouni, and Mahieddine Djoudi, “KAB: A new k-anonymity approach based on black hole algorithm,” *Journal of King Saud University – Computer and Information Sciences,* vol. 34, no. 7, pp. 4075-4088, 2022.

[3] B.C.M. Fung, K. Wang, R. Chen, and P.S. Yu, “Privacy Preserving Data Publishing: A Survey of Recent Development,” *ACM Computing Surveys,* vol. 42, no. 4, pp. 14:1-14:53, 2010.

[4] Y. Xiao and H. Li, “Privacy Preserving Data Publishing for Multiple Sensitive Attributes Based on Security Level,” *Information, 11,* p. 166, 2020.

[5] L. Sweeney, “k-anonymity: a model for protecting privacy,” *Fuzziness and Knowledge-based Systems,* vol. 10, no. 5, pp. 557-570, 2002.

[6] T.M. Truta, A. Campan, M. Abrinica. and J. Miller., “A Comparison Between Local and Global Recoding Algorithms for Achieving Microdata p-Sensitive k-Anonymity,” *Acta Universitatis Apulensis,* vol. 15, no. 15, pp. 213-233, 2008.

[7] A. Machanavajjhala, J. Gehrke, D. Kifer, & M.R. Venkitasubramaniam, “l-Diversity: Privacy Beyond k-Anonymity,” in *International Conference on Data Engineering (ICDE),* Seoul, Korea, 2006.

[8] V. S. Susan, “An Anonymization Approach for Dynamic Dataset with Multiple Sensitive Attributes,” *Intelligent Computing and Applications,* vol. 1172, pp. 731-739, 2021.

[9] Widodo, E.K. Budiardjo, and W.C. Wibowo, “Privacy Preserving Data Publishing with Multiple Sensitive Attributes based on Overlapped Slicing,” *Information,* vol.10 no. 12, p. 362, 2019.

[10] Widodo and Wahyu C. Wibowo, "A Distributional Model of Sensitive Values on p-Sensitive in Multiple Sensitive Attributes," in *International Conference on Informatics and Computational Sciences (ICICoS)*, Semarang, Indonesia, 2018.

[11] Widodo,E.K. Budiardjo, W.C. Wibowo, and Harry T.Y. Achsan, "An Approach for Distributing Sensitive Values in k-Anonymity," in *International Workshop on Big Data and Information Security (IWBIS)*, Nusa Dua, Bali, Indonesia, 2019.

[12] Widodo, M. Nugraheni, and I.P. Sari, "Simple Distribution of Sensitive Values for Multiple Sensitive Attributes in Privacy Preserving Data Publishing to Achieve Anatomy," in *2nd International Conference on Innovative and Creative Information Technology (ICITech)*, Salatiga, 2021.

[13] Widodo, I.P. Sari, and M. Nugraheni, "ASENVA: Summarizing Anatomy Model by Aggregating Sensitive Values," in *International Conference on Electrical Engineering and Informatics (ICELTICs)*, Banda Aceh, 2020.

[14] A.S.M. Touhidul Hasan, Q. Jiang, H. Chen, and Shengrui Wang, "A New Approach to Privacy-Preserving Multiple Independent Data Publishing," *Applied Science,* vol. 8, no. 5, 2018.

[15] Becker,Barry and Kohavi,Ronny, "Adult Dataset," UCI Machine Learning Repository, 1996.