Jurnal Informatika Universitas Pamulang **ISSN: 2541-1004**
Penerbit: Program Studi Teknik Informatika Universitas Pamulang **e-ISSN: 2622-4615**
Vol. 6, No. 2, Juni 2021 (240-245) **10.32493/informatika.v6i2.9161**

Anonimisasi Data Penjualan Pakaian di Toko Online Menggunakan Metode *K-Anonymity*, *L-Diversity*, dan *T-Closeness*

Rahmat Hidayat[1], I Gusti Agung Premananda[2], Nur Aini Rakhmawati[3]

[1,2,3]Departmen Sistem Informasi, Institut Teknologi Sepuluh Nopember, Kampus ITS Keputih, Sukolilo, Surabaya, Jawa Timur, Indonesia, 60111
e-mail: [1]rahmat.19052@mhs.its.ac.id, [2]gustiagungpremananda@gmail.com, [3]nur.aini@is.its.ac.id

Submitted Date: January 17th, 2021 Reviewed Date: June 10th, 2021
Revised Date: June 13th, 2021 Accepted Date: July 24th, 2021

Abstract

The development of information technology will have an impact on the development of data. The existence of data might contain sensitive elements that are not intended to become public consumers. Anonymization is a technique that can be applied in publishing data with a different identity or anonymously. K-anonymity is an approach that can anonymization data. Besides that, the l-diversity and t-closeness approaches are also one of the advanced alternatives in data anonymization. The Mondrian algorithm can be implemented in k-anonymity. Therefore, we use the Mondrian algorithm for anonymizing clothing online transaction. The application of these methods can overcome the problem of data privacy contained in the dataset. The results obtained are that the application of the Mondrian algorithm in k-anonymity, l-diversity, and t-closeness has successfully performed data anonymization so that data cannot be consumed freely by other users.

Keywords: Anonymization; *K-anonymity*; *l-diversity*; *t-closeness*; Mondrian

Abstrak

Pesatnya teknologi informasi akan berdampak pada perkembangan data. Adanya data-data di dalam sebuah instansi terdapat unsur-unsur yang dianggap sensitif yang tidak untuk menjadi konsumsi publik. Anonimisasi merupakan sebuah teknik yang dapat diterapkan dalam mempubikasikan data dengan identitas yang berbeda atau anonim. *K-anonymity* merupakan salah satu pendekatan yang dapat melakukan anonimisasi data. Selain itu pendekatan *l-diversity* dan *t-closeness* juga menjadi salah satu alternatif lanjutan dalam anonimisasi data. Algoritma Mondrian dapat diimplementasikan dalam *k-anonymity*. Penelitian ini akan memanfaatkan penggunaan algoritma Mondrian untuk penerapan anonimisasi data pada penjualan pakaian melalui media sosial. Penerapan metode-metode tersebut dapat mengatasi permasalahan privasi data yang terdapat pada dataset tersebut. Hasil yang diperoleh adalah dalam penerapan algoritma Mondrian dalam *k-anonimity, l-diversity,* dan *t-closeness* telah berhasil melakukan anonimisasi data, sehingga data tidak dapat dikonsumsi secara bebas oleh pengguna lain.

Kata kunci: Anonimisasi; *K-anonymity*; *l-diversity*; *t-closeness*; Mondrian

1 Pendahuluan

Dalam kemajuan dunia digital saat ini tentu saja tidak terlepas dari data. Data akan terus terhimpun dengan pesatnya teknologi informasi. Di dalam suatu perusahaan, organisasi, maupun lembaga masyarakat tentu saja memiliki data yang disajikan. Adanya data-data tersebut tentu saja terdapat nilai-nilai yang dapat dianggap sensitif, yaitu tidak semua orang dapat mengetahuinya atau data-data tersebut tidak bisa dijadikan konsumsi publik secara bebas.

Salah satu teknik yang dapat diterapkan dalam mempublikasikan data dengan identitas yang berbeda atau anonim yang terkenal adalah anonimisasi. Anonimisasi adalah sebuah solusi untuk mencegah serangan *linkage* dalam kumpulan data yang tidak teridentifikasi (Nergiz & Atzori, 2009). *K-anonymity* merupakan pendekatan anonimisasi data yang diperkenalkan oleh

(Samarati, 2001) dan (Sweeney, 2002). *K-anonymity* dikatakan sebagai "keamanan dalam angka" yang memastikan bahwa setiap entitas dalam tabel tidak dapat dibedakan dari k −1 entitas lainnya (Nergiz & Atzori, 2009).

Kelemahan yang terjadi di *k-anonymity* menurut (Sya'airillah & Adhi, 2020) yang diacu dari (Machanavajjhala et al., 2007) adalah data sensitif yang ada bisa saja terungkap apabila terjadi *homogeneity attack* dan *background knowledge attack.* Kelemahan *k-anonymity* yang harus ditingkatkan yaitu dalam mempertahankan Batasan frekuensi paling tinggi dari nilai jarak dalam atribut kritis (Yousra & Mazleena, 2018). Kemudian (Machanavajjhala et al., 2007) mengembangkan *l-diversity,* yaitu sebuah kriteria privasi baru dan kuat yang disebut keragaman yang dapat bertahan dari serangan-serangan pada permasalahan *k-anonymity*.

Selanjutnya (Li et al., 2007) mendapatkan bahwa *l-diversity* juga memiliki kelemahan, yakni keterbatasan dalam asumsi *adversarial knowledge.* Selain itu, *l-diversity* tidak cukup baik dalam mencegah jenis serangan kesamaan (*similarity attack*) dan serangan *skewness* (Frikken & Zhang, 2008). Maka (Li et al., 2007) mengusulkan sebuah gagasan privasi baru disebut *t-closeness* yang mengharuskan distribusi atribut sensitif dalam setiap kelas ekivalen dekat dengan distribusi atribut dalam tabel keseluruhan (yaitu, jarak antara dua distribusi tidak boleh lebih dari ambang t).

Algoritma anonimisasi biasanya bertujuan untuk melindungi privasi individu, dengan dampak minimal pada kualitas data yang dihasilkan (LeFevre et al., 2006). Algoritma Mondrian dapat diimplementasikan dalam anonimisasi, yaitu sebuah algoritma yang berasal dari fakta bahwa sebuah gambar dibagi menjadi berbagai blok dengan ukuran berbeda (Williams, 2018).

Pada penelitian ini akan dilakukan penggunaan algoritma Mondrian untuk penerapan anonimisasi data ke dalam *k-anonimity, l-diversity*, dan *t-closeness.* Penerapan anonimisasi data menggunakan *Jupyter Notebook*, yang merupakan sebuah *notebook* komputasi sumber terbuka (gratis) yang populer dan dapat membuat pengguna untuk melakukan eksperimen kode, visualisasi, dan teks dalam satu dokumen yang struktur dasarnya berupa data JSON (Rule et al., 2018). Pengujian efektivitas algoritma Mondrian dalam *k-anonimity, l-diversity,* dan *t-closeness* menggunakan dataset penjualan pakaian salah satu toko yang menjual di sosial media facebook.

2 Metode Penelitian

Penelitian ini akan melakukan beberapa tahapan dalam anonimisasi data, yaitu pengumpulan dataset, *partitioning data, data agregarion, l-diversity, t-closeness*. Tahapan-tahapan dalam anonimisasi data pada penelitian ini akan dijelaskan lebih lanjut di bawah ini.

2.1 Pengumpulan Dataset

Penggunaan dataset pada penelitian ini berupa kumpulan data hasil perhitungan penjualan sebuah toko yang melakukan transaksi di salah satu media sosial (Hidayat, Premananda, & Rakhmawati, 2021). Dataset berisikan data penjual dan produk yang dijual. Data penjual seperti nama pembeli, jumlah pesanan, ukuran, alamat, sumber pembelian, no telepon dan tanggal. Sementara itu, data produk yang dijual berupa id produk, jumlah item, ukuran, dan tanggal pembelian produk. Dataset pada awalnya diperoleh sebanyak 2400 data, kemudian dilakukan *filtering data* berdasarkan data yang sesuai dengan kebutuhan anonimisasi data. Tabel 1 merupakan contoh isi dari dataset.

Tabel 1. Contoh isi dataset

Nomor	Nama	Alamat	Sumber
1	Bunga	Jln. Merdeka No. 11 Kota Surabaya	SMS
2	Putri	Jln. Kusumanegara No. 23 Kota Sidoarjo	WA
3	Melati	Jln. Sultan Agung No. 31 Kota Malang	BBM

2.2 Partitioning Data

Setelah melakukan pengumpulan *dataset*, hal yang dilakukan selanjutnya adalah melakukan *partitioning* pada data untuk dilakukan kategori data, kemudian dibentuk viasualisasi datanya berdasarkan kolom yang dipilih. Tabel 2 menunjukkan hasil dari *dataset* yang telah dilakukan *partitioning*.

Tabel 2. Partitioning data

Nomor	Kabupaten	Sumber
1	Surabaya	SMS
2	Sidoarjo	WA
3	Malang	BBM

2.3 Data Agregation

Pada tahapan ini dilakukan penggabungan nilai-nilai *quasi-identifiers* dan atribut sensitif di setiap grup *k-anonymity*. Hal ini dilakukan dengan tujuan untuk menghasilkan *dataset* baru yang berisi

satu baris untuk setiap partisi dan nilai atribut sensitif. Dapat dilakukan penggabungan kolom di setiap partisi. Penggabungan itu seperti alamat menjadi kode kabupaten dan sumber pemesanan (melalui SMS, BBM, atau WA) menjadi kode sumber. Tabel 3 merupakan contoh *data agregation.*

Tabel 2. Partitioning data

Nomor	KodeKabupaten	KodeSumber
1	1	1
2	2	2
3	3	3

2.4 L-diversity

Tahapan implementasi *l-diversity* memastikan bahwa setiap grup *k-anonymity* berisi setidaknya l nilai yang berbeda dari atribut sensitif (Machanavajjhala et al., 2007). Pada tahapan ini dilakukan pengelompokan suatu baris, kemudian mengubah menjadi kode tertentu. Hal ini dimaksudkan agar tidak ditemukannya nilai atribut tertentu dengan pasti, meskipun telah mengetahui kode kelompoknya. Tabel 3 merupakan contoh implementasi *l-diversity.*

Tabel 3. Contoh implementasi *l-diversity*

Nomor	KodeKabupaten	KodeSumber
1	1	1.0
2	2	2.0
3	3	3.0

2.5 T-closeness

Pada implementasi *t-closeness* akan menghasilkan partisi yang berisi entri yang telah disempurnakan dari *l-diversity* yang memiliki tujuan untuk mengurangi perincian representasi data. Dalam implementasi *t-closeness* dilakukan pengelompokan dengan kode tertentu, sama halnya dengan anonimisasi sebelumnya akan tetapi dilakukan penyempurnaan secara meningkat dari anonimisasi *l-diversity*. Tabel 4 merupakan contoh implementasi anonimisasi *t-closeness.*

Tabel 5. Contoh implementasi *t-closeness*

Nomor	KodeKabupaten	KodeSumber
1	1.2	1.2
2	3.7	2.27
3	4.28	3.18

3 Hasil dan Pembahasan

3.1 Pengumpulan dataset

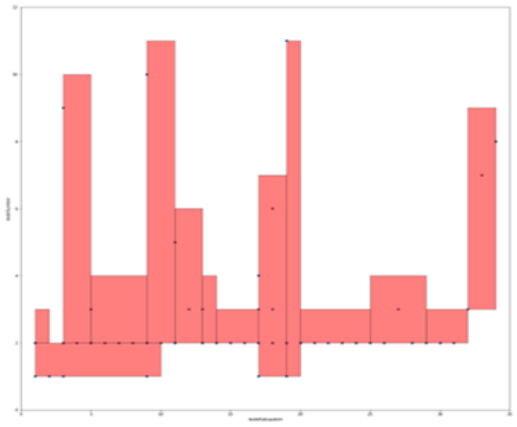
Data yang digunakan pada penelitian ini berupa data penjualan produk pakaian dalam oleh sebuah toko *online* yang melakukan pemasaran di media sosial facebook (Hidayat et al., 2021). Hasilnya didapatkan 100 data penjualan yang berisi nama pembeli, jumlah pesanan, ukuran, alamat, sumber pembelian, no telepon dan tanggal. Beberapa data telah disamarkan seperti nama nomor telepon dan alamat. Pada alamat hanya akan digunakan kota atau kabupatennya saja. Data pada kolom ukuran, alamat dan sumber penjualan akan dibuatkan kolom baru yang merupakan pengkategorian yang diterjemahkan dalam bentuk angka. Hasil dari pengumpulan data tersebut setelah dijalankan di Jupyter dapat dilihat pada Tabel 6.

Tabel 6 Hasil pengumpulan *dataset*

	nama	jumlahPesanan	kodeUkuran	ukuran	alamat	kodeKabupaten	sumber	kodeSumber	no telpon	tanggal
0	IBU X	1	5	XXL-CLASSIC	BOGOR	1	SMS	1	081XXXXXXXX	04/08/2016
1	IBU X	1	5	XXL-CLASSIC	CIBUBUR	2	SMS	1	081XXXXXXXX	04/08/2016
2	IBU X	1	8	L-PASTEL	BANDUNG	3	WA	2	081XXXXXXXX	04/08/2016
3	IBU X	1	5	XXL-CLASSIC	BENGKULU	4	WA	2	081XXXXXXXX	04/08/2016
4	IBU X	1	4	XL-CLASSIC	JAKARTA TIMUR	5	WA	2	081XXXXXXXX	04/08/2016
...	...	...	...	...	...	...	...	...	...	...
95	IBU X	1	9	XL-PASTEL	CIMAHI	33	WA	7	081XXXXXXXX	20/12/2016
96	IBU X	1	8	L-PASTEL	GARUT	34	WA	8	081XXXXXXXX	20/12/2016
97	IBU X	1	3	L-CLASSIC	BANDUNG	3	WA	9	081XXXXXXXX	20/12/2016
98	IBU X	1	15	xxL-SUMMER	JAKARTA BARAT	9	WA	10	081XXXXXXXX	21/12/2016
99	IBU X	1	3	L-CLASSIC	JAKARTA SELATAN	19	WA	11	081XXXXXXXX	21/12/2016

3.2 Partitioning data

Pada tahapan ini dipilih 2 kolom yaitu kolom kodeKabupaten dan kodeSumber sebagai kolom fitur, kemudian kolom ukuran sebagai kolom sensitif. Hasilnya didapatkan data dibagi menjadi 14 bagian. Gambar 1 merupakan hasil visualisasi partisi data.



Gambar 1. Hasil visualisasi *partitioning data*

3.3 Data agregation

Pada tahapan ini dilakukan penggabungan kolom pada setiap partisi untuk menciptakan *dataset k-anonymous*. Hasilnya didapat pengelompokan berdasarkan 14 pembagian *dataset* seperti yang terlihat pada Tabel 7 dan Tabel 8. Berdasarkan pengelompokan tersebut dapat dilihat masih ada kolom fitur yang tidak dilakukan annomisasi seperti pada bari 14, 15, dan 16 di mana kolom kodeKabupaten dan kode sumber masih sama dengan data aslinya.

Tabel 7 Hasil D*ata agregation*

	kodeKabupaten	kodeSumber	ukuran	count
14	1.000000	2.000000	L-CLASSIC	1
15	1.000000	2.000000	L-PASTEL	2
16	1.000000	2.000000	S-PASTEL	1
17	3.111111	2.777778	L-CLASSIC	2
18	3.111111	2.777778	L-PASTEL	1
19	3.111111	2.777778	M-CLASSIC	1
20	3.111111	2.777778	XL-CLASSIC	2
21	3.111111	2.777778	XL-PASTEL	1
22	3.111111	2.777778	XXL-CLASSIC	2
0	3.750000	1.000000	L-CLASSIC	1
1	3.750000	1.000000	XL-CLASSIC	1
2	3.750000	1.000000	XXL-CLASSIC	2
23	5.461538	2.153846	L-CLASSIC	3
24	5.461538	2.153846	L-PASTEL	1
25	5.461538	2.153846	M-CLASSIC	1
26	5.461538	2.153846	M-PASTEL	1
27	5.461538	2.153846	S-Classic	1
28	5.461538	2.153846	XL-CLASSIC	2
29	5.461538	2.153846	XL-PASTEL	2
30	5.461538	2.153846	XXL-CLASSIC	2
35	9.250000	4.000000	XXL-CLASSIC	3
36	9.250000	4.000000	xxL-SUMMER	1
39	11.142857	2.571429	L-CLASSIC	1
40	11.142857	2.571429	L-SUMMER	1
41	11.142857	2.571429	M-CLASSIC	1
42	11.142857	2.571429	XL-CLASSIC	1
43	11.142857	2.571429	XXL-CLASSIC	3
49	13.000000	2.333333	M-CLASSIC	1

Tabel 8 Hasil D*ata agregation*

50	13.000000	2.333333	XXL-CLASSIC	2
51	14.750000	2.000000	XL-PASTEL	2
52	14.750000	2.000000	XXL-CLASSIC	2
3	17.388889	2.388889	L-CLASSIC	2
4	17.388889	2.388889	L-PASTEL	4
5	17.388889	2.388889	L-SUMMER	1
6	17.388889	2.388889	M-PASTEL	1
7	17.388889	2.388889	XL-CLASSIC	5
8	17.388889	2.388889	XXL-CLASSIC	5
9	19.000000	2.666667	L-CLASSIC	2
10	19.000000	2.666667	L-PASTEL	2
11	19.000000	2.666667	M-CLASSIC	1
12	19.000000	2.666667	XL-PASTEL	2
13	19.000000	2.666667	XXL-CLASSIC	5
31	21.272727	2.000000	M-PASTEL	1
32	21.272727	2.000000	XL-CLASSIC	2
33	21.272727	2.000000	XL-PASTEL	1
34	21.272727	2.000000	XXL-CLASSIC	7
37	26.400000	2.200000	L-PASTEL	1
38	26.400000	2.200000	XXL-CLASSIC	4
44	30.000000	2.000000	XL-CLASSIC	1
45	30.000000	2.000000	XL-PASTEL	1
46	30.000000	2.000000	XXL-CLASSIC	1
47	33.000000	6.000000	L-PASTEL	2
48	33.000000	6.000000	XL-PASTEL	1

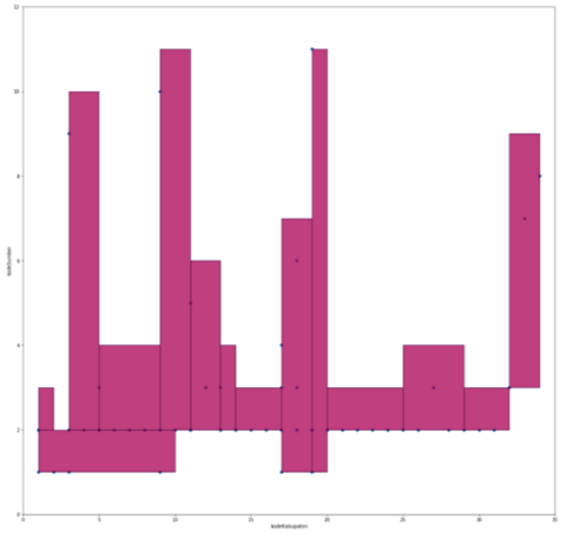
3.4 Implementasi L-diversity

Tahapan selanjutnya adalah menambahkan *l-diversity* untuk memperbaik hasil dari tahapan selanjutnya. Hasil dari *l-diversity* ternyata tidak memiliki pengaruh pada dataset ini. Pembagian data yang dilakukan sama persis dengan tahapan sebelumnya yaitu membagi data menjadi 14 bagian dan agregasi yang dilakukan sama persis seperti yang terlihat pada Gambar 2. Hasil *l-diversity* diwakilkan oleh warna biru, dan hasil dari k-Anonymous diwakli dengan warna merah. Namun pada Gambar 3 warna yang muncul adalah ungu yang merupakan penggabungan dari warna merah dan biru karena hasil agregasi dari dua metode ini sama persis.

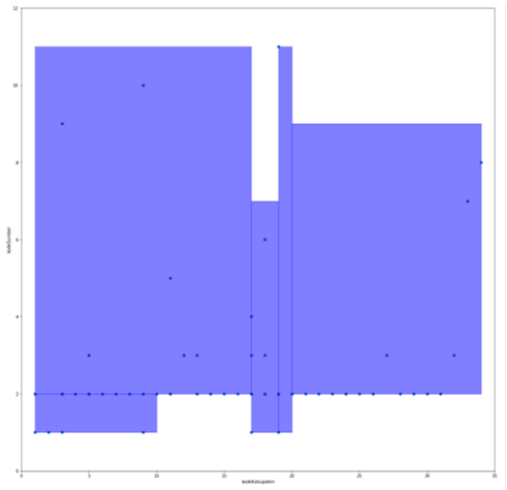
3.5 Implementasi t-closeness

Pada penerapan *t-closeness* didapatkan perubahan pembagian data yaitu hanya dibagi menjadi 5 bagian. Visualisasi dari pembagian *t-closeness* dapat dilihat pada Gambar 3.

Hasil dari penerapan *t-clossness* menunjukan semua data telah dianonimisasi dengan baik dengan seperti pada Tabel 8. Hal ini ditandakan dengan salah satu kolom fitur pada setiap baris selalu ada yang disamarkan tidak seperti pada hasil *k-annonimity* dan *l-diversity* yang masih ditemukan adanya baris yang semua kolom fiturnya masih sama persis dengan data awalnya.



Gambar 2. Hasil visualisasi *l-diversity*



Gambar 3. Visualisasi Hasil *t-closeness*

Tabel 8 Hasil t-closeness

	kodeKabupaten	kodeSumber	ukuran	count
0	3.750000	1.000000	L-CLASSIC	1
1	3.750000	1.000000	XL-CLASSIC	1
2	3.750000	1.000000	XXL-CLASSIC	2
3	7.181818	2.500000	L-CLASSIC	7
4	7.181818	2.500000	L-PASTEL	4
5	7.181818	2.500000	L-SUMMER	1
6	7.181818	2.500000	M-CLASSIC	4
7	7.181818	2.500000	M-PASTEL	1
8	7.181818	2.500000	S-Classic	1
9	7.181818	2.500000	S-PASTEL	1
10	7.181818	2.500000	XL-CLASSIC	5
11	7.181818	2.500000	XL-PASTEL	5
12	7.181818	2.500000	XXL-CLASSIC	14
13	7.181818	2.500000	xxL-SUMMER	1
14	17.388889	2.388889	L-CLASSIC	2
15	17.388889	2.388889	L-PASTEL	4
16	17.388889	2.388889	L-SUMMER	1
17	17.388889	2.388889	M-PASTEL	1
18	17.388889	2.388889	XL-CLASSIC	5
19	17.388889	2.388889	XXL-CLASSIC	5
20	19.000000	2.666667	L-CLASSIC	2
21	19.000000	2.666667	L-PASTEL	2
22	19.000000	2.666667	M-CLASSIC	1
23	19.000000	2.666667	XL-PASTEL	2
24	19.000000	2.666667	XXL-CLASSIC	5
25	25.227273	2.590909	L-PASTEL	3
26	25.227273	2.590909	M-PASTEL	1
27	25.227273	2.590909	XL-CLASSIC	3
28	25.227273	2.590909	XL-PASTEL	3
29	25.227273	2.590909	XXL-CLASSIC	12

4 Kesimpulan

Berdasarkan keseluruhan yang telah dijabarkan pada penelitian yang telah dilakukan, peneliti berkesimpulan bahwa penerapan algoritma Mondrian dalam *k-anonimity, l-diversity,* dan *t-closeness* telah berhasil melakukan anonimisasi data pada *dataset* penjualan pakaian. Sehingga data privasi dari pembeli yang terdapat pada dataset penjualan pakaian melalui media sosial facebook tidak dapat dikonsumsi secara bebas oleh pengguna lainnya.

Referensi (*References*)

Hidayat, R., Premananda, I. G. A., & Rakhmawati, N. A. (2021). *Dataset Penjualan Pakaian Menggunakan K-Anonymity (1.0) [Computer*

software]. Zenodo. https://doi.org/10.5281/ZENODO.4445564.

Frikken, K. B., & Zhang, Y. (2008). Yet another privacy metric for publishing micro-data. *Proceedings of the 7th ACM Workshop on Privacy in the Electronic Society*, 117–122.

LeFevre, K., DeWitt, D. J., & Ramakrishnan, R. (2006). Workload-aware Anonymization. *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 277–286.*

Li, N., Li, T., & Venkatasubramanian, S. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. *2007 IEEE 23rd International Conference on Data Engineering*, 106–115.

Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD), Vol. 1, No. 1, Article 3*.

Nergiz, M. E., & Atzori, M. (2009). Towards Trajectory Anonymization: A Generalization-Based Approach. *SPRINGL '08: Proceedings of the SIGSPATIAL ACM GIS 2008 International Workshop on Security and Privacy in GIS and LBS, pages 52–61*.

Rule, A., Tabard, A., & Hollan, J. D. (2018). Exploration and explanation in computational notebooks. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–12.

Samarati, P. (2001). Protecting respondents Privacy in Microdata release. *IEEE TransacY Tions on Knowledge and Data Engineering*, *vol. 13, no. 6, pp. 1010-1027*.

Sweeney, L. (2002). K-Anonymity: A Model For Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(05), 557–570. https://doi.org/10.1142/S0218488502001648.

Sya'airillah, S., & Adhi, B. P. (2020). *Analisis Model L-Diversity Dengan Algoritma Systematic Clustering Dan Datafly*. PINTER: Jurnal Pendidikan Teknik Informatika Dan Komputer, 4(1), 43–48.

Williams, D. P. (2018). The Mondrian Detection Algorithm for Sonar Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, *56*(2), 1091–1102, https://doi.org/10.1109/TGRS.2017.2758808.

Yousra, S. A., & Mazleena, S. (2018). A new heuristic anonymization technique for privacy preserved datasets publication on cloud computing. *Journal of Physics: Conference Series*, *1003*(1), 012030.