Implementasi Algoritma *Mondrian Multidimensional K-Anonymity* pada Biodata Calon Legislatif

Adam Akbar[1], Rivanda Putra Pratama[2], Nur Aini Rakhmawati[3]
Departemen Sistem Informasi
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
e-mail: [1]adamakbar.id@gmail.com, [2]rivanda.19052@mhs.its.ac.id, [3]nur.aini@is.its.ac.id

Diajukan: 16 Januari 2021; Direvisi: 19 Maret 2021; Diterima: 26 Maret 2021

Abstrak

Uni Eropa menerbitkan sebuah peraturan yang bernama General Data Protection Regulation (GDPR) untuk menjaga privasi warga. Peraturan ini meregulasi penyebaran data-data pribadi seperti nama, nomor telepon atau alamat yang mungkin akan digunakan untuk tujuan tertentu. Salah satu teknik yang dapat digunakan untuk menyebarkan data tanpa melanggar privasi dari subjek pemilik data adalah K-Anonymity. K-Anonymity memodifikasi nilai quasi-identifier hingga subjek tidak dapat dikenali lagi tetapi dataset tetap mengandung informasi yang diperlukan. Artikel ini telah mengimplementasikan K-Anonymity pada data Calon Legislatif untuk Pemilihan Umum Calon Legislatif tahun 2019 yang dihimpun dari laman resmi Komisi Pemilihan Umum. Dengan algoritma Mondrian Multidimensional K-Anonymity hasil anonimisasi menunjukkan bahwa masih terdapat data yang unik. Namun, dari hasil visualisasi terlihat hampir semua data memiliki anonimitas sama, yang dimungkinkan karena jumlah data partisi yang kurang banyak ataupun kurangnya keberagaman data.

Kata kunci: *Anonimisasi, Privasi data, K-Anonymity, Mondrian, Calon legislatif.*

Abstract

The European Union issued a regulation called the General Data Protection Regulation (GDPR) to protect the privacy of its citizens. The GDPR regulates the distribution of personal data such as name, telephone number or address which may be used for certain purposes. One of the techniques that can be used to disseminate data without violating the privacy of general data is K-Anonymity. K-Anonymity modifies the quasi-identifier value until the subject can no longer be recognized but the dataset still contains the necessary information. This article has implemented K-Anonymity in data on Indonesian Legislative Candidates for The 2019 General Election compiled from the official website of the General Election Commission. The anonymization results show that there are still unique data by using the Mondrian Multidimensional K-Anonymity algorithm. However, from the visualization results, it can be seen that almost all data has the same anonymity, which is possible due to the insufficient number of partition data or the lack of data diversity.

Keywords*: Anonymization, Data privacy, K-Anonymity, Mondrian, Legislative candidate.*

1. Pendahuluan

General Data Protection Regulation (GDPR) merupakan peraturan yang berlaku di Uni Eropa sejak 25 Mei 2018 [1], [2]. GDPR diterbitkan untuk menggantikan peraturan sebelumnya yang bernama *Data Protection Acts* 1988-2003. GDPR memuat beberapa ketentuan yang sebelumnya tidak terdapat pada *Data Protection Acts* 1988-2003. GDPR merupakan jawaban atas kebutuhan Uni Eropa untuk proteksi modern di era internet saat ini. Secara sederhana, GDPR mengatur tentang privasi dan keamanan atas data masyarakat Uni Eropa yang dikumpulkan oleh siapa pun. Walau dengan batasan, GDPR mengizinkan pemberdayaan data subjek untuk tujuan tertentu misalnya penelitian. Beberapa metode seperti *Pseudonymization* atau *Anonymization* dapat menghasilkan kumpulan data (*dataset*) yang memuat banyak informasi dengan privasi yang tetap terjaga sesuai dengan ketentuan GDPR.

Salah satu model untuk *Anonymization* yang cukup populer adalah *K-Anonymity* yang diperkenalkan oleh Latanya Sweeney melalui artikelnya pada tahun 2002 [3]. Pada artikel tersebut, Sweeney menggunakan istilah *quasi-identifier* yang diperkenalkan oleh Dalenius [4] yaitu kumpulan

atribut yang dapat menjadi pengenal identitas seperti nama, alamat, NIK, atau nomor telepon. Selain itu, beberapa atribut yang jika dikombinasikan dapat menjadi pengenal identitas, maka atribut-atribut tersebut termasuk *quasi-identifier*, sebagai contoh kombinasi tanggal lahir, jenis kelamin, dan kode pos dapat mengidentifikasi satu orang yang spesifik.

Secara sederhana, *K-Anonymity* dapat dicapai dengan melakukan *generalization* dan *suppression* pada *quasi-identifier* hingga nilai-nilai *quasi-identifier* pada setiap baris tidak unik dan minimal muncul sebanyak *k*. Sebagai penggambaran, Tabel 1 adalah *dataset* awal sebelum dilakukan *K-Anonymity* dan Tabel 2 merupakan *dataset* hasil *K-Anonymity* dari Tabel 1. Seperti yang dapat dilihat pada Tabel 2, *supression* dilakukan pada atribut nama dan *generalization* dilakukan pada atribut usia dan kota. Kolom nama diubah menjadi "*" untuk setiap nilainya. Sedangkan, kolom usia diubah menjadi bentuk rentang nilai, serta kolom kota yang menunjukkan wilayah digeneralisasi menjadi kolom provinsi. Dari Tabel 2 juga diketahui bahwa kolom {nama, usia, provinsi} yang bernilai {*, 21 s/d 30, Jawa Timur} telah muncul sebanyak 2 kali dan nilai {*, 31 s/d 40, Jawa Timur} muncul sebanyak 3 kali, yang artinya nilai *k* dari Tabel 2 adalah 2, karena diambil dari nilai kemunculan terkecil.

Tabel 1. Contoh *dataset* pasien rumah sakit.

Nama	Usia	Usia	Penyakit
Wyatt Montes	23	Surabaya	Asma
Kimberley Archer	40	Sidoarjo	Kanker
Hermione Goodwin	40	Sidoarjo	Osteoporosis
Alessandro Raymond	27	Surabaya	Asma
Ameera Wheeler	34	Malang	Kanker

Tabel 2. *Generalization* dan *suppression* pada contoh *dataset* pasien rumah sakit.

Nama	Usia	Usia	Penyakit
*	21 s/d 30	Jawa Timur	Asma
*	31 s/d 40	Jawa Timur	Kanker
*	31 s/d 40	Jawa Timur	Osteoporosis
*	21 s/d 30	Jawa Timur	Asma
*	31 s/d 40	Jawa Timur	Kanker

Beberapa penelitian terhadap *K-Anonymity* telah dilakukan sebelumnya seperti pada artikel [5] yang telah mengimplementasikan dua algoritma *K-Anonymity* yaitu *paritioning-based* dan *k-means-clustering-based* pada data atlet pemain bola yang sering digunakan para peneliti untuk analisa peringkat atlet pemain bola. Selain data atlet pemain bola, artikel [6] menggunakan data medis pasien untuk menerapkan *K-Anonymity* yang berbasis *bilinier pairings* yang dapat melakukan pencarian data yang terenkripsi untuk menjamin privasi dari pasien. Tidak hanya data terstruktur yang perlu dilakukan anonimisasi, seperti pada artikel [7] yang menerapkan *K-Anonymity* berbasis *MapReduce* untuk melakukan anonimisasi pada *big data* dan berhasil mengurangi durasi pemrosesan pada *big data*. Sedangkan artikel [8] telah mengembangkan *K-Anonymity* berbasis *analytic hierarchy process* yang berhasil menekan tingkat informasi yang hilang (*information loss)* pada hasil anonimisasi.

Penelitian pada artikel ini akan melakukan implementasi *K-Anonymity* pada *dataset* biodata Calon Legislatif (Caleg) untuk Pemilihan Umum Anggota Legislatif (Pileg) di Indonesia pada tahun 2019.

2. Metode Penelitian

2.1. *Dataset*

Penelitian ini menggunakan *dataset* Caleg milik Nur Aini Rakhmawati [9] yang dihimpun menggunakan metode *crawling* pada laman resmi Komisi Pemilihan Umum untuk penelitian terdahulu. *Dataset* tersebut memiliki sekitar 36.000 baris data, namun dari semua baris tersebut, sekitar 26.000 baris ditemukan tidak lengkap karena tidak dilengkapi oleh Caleg yang bersangkutan. *Dataset* ini memiliki 15 kolom antara lain dapil, partai, nama, jk, lokasi, tahun_lahir, agama, nikah, pasangan, pendidikan, pekerjaan, status, motivasi, dan target.

2.2. Pra-proses *Dataset*

Beberapa atribut pada *dataset* dikodekan dalam bentuk desimal untuk memudahkan pemetaan program dan visualisasi. Daftar atribut dan detail pengodean dapat dilihat pada Tabel 3.

Tabel 3. Daftar pengodean.

Atribut	Nilai	Kode Nilai
jenis_kelamin	Laki-Laki	0
	Perempuan	1
agama	Islam	1
	Kristen Protestan	2
	Katolik	3
	Hindu	4
	Budha	5
	Konghucu	6
status_nikah	Belum Menikah	1
	Sudah Menikah	2
	Pernah Menikah	3
pendidikan	Tidak Sekolah	1
	SD	2
	SMP/Sederajat	3
	SMA/Sederajat	4
	D1	5
	D2	6
	D3	7
	D4/S1	8
	S2	9
	S3	10
pekerjaan	Tidak Bekerja	1
	PNS	2
	Wiraswasta	3
	Pengusaha	4
	Pegawai Pemerintah	5
	Pensiunan	6

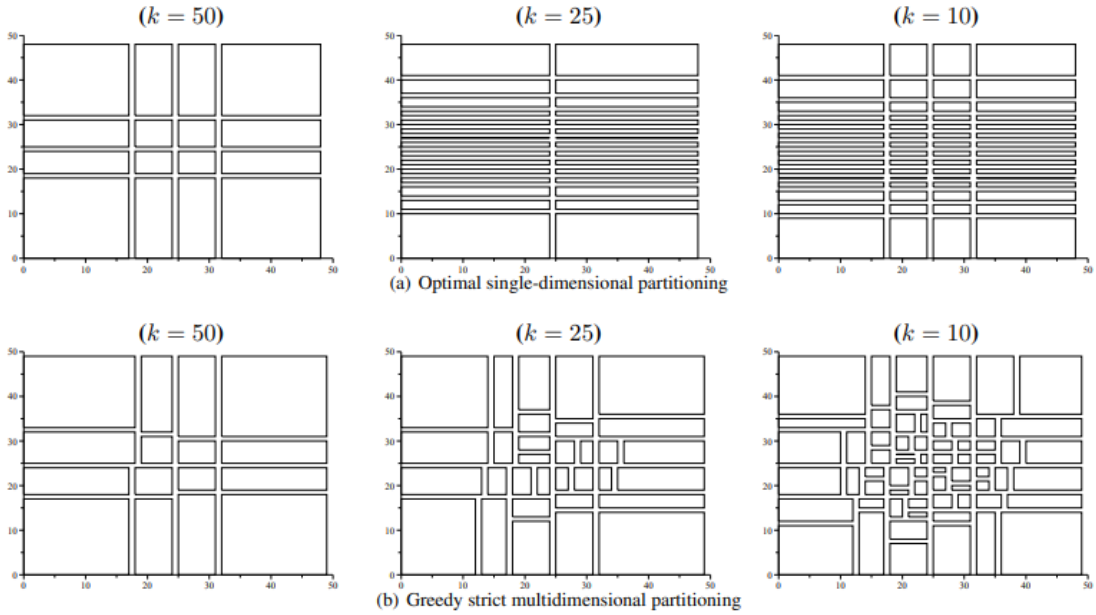
2.3. *Mondrian Multidimensional K-Anonymity*

Penelitian ini menggunakan algoritma *Mondrian Multidimensional K-Anonymity* [10] yang merupakan salah satu variasi dari *K-Anonymity* yang dikembangkan oleh LeFevre dkk. untuk memperoleh hasil anonimisasi yang lebih baik. Pada *Mondrian Multidimensional K-Anonymity*, sebelum melakukan anonimisasi, *dataset* dipartisi secara multi-dimensi menjadi bagian-bagian kecil berdasarkan nilai dari satu atribut *quasi-identifier* (*sensitive column*). Bagian-bagian kecil tersebut dipetakan pada dua atribut *quasi-identifier* yang lain (*feature column*) sebagai batas untuk rentang partisi. Ilustrasi hasil partisi dapat dilihat pada Gambar 1, di mana bagian (a) merupakan ilustrasi hasil partisi dengan dimensi tunggal, dan bagian (b) adalah ilustrasi untuk hasil partisi secara multi-dimensi yang telah dipaparkan.

Hasil partisi yang berupa bagian-bagian kecil dari *dataset* tersebut yang kemudian dilakukan anonimisasi. Pada penelitian ini, dua atribut yang digunakan untuk proses partisi adalah atribut tahun_lahir dan pekerjaan sebagai *feature column*, atribut partai sebagai *sensitive column*, dan nilai *k* adalah 3.

2.4. Visualisasi

Setelah dilakukan anonimisasi dengan algoritma *Mondrian Multidimensional K-Anonymity*, hasil dari proses partisi disajikan dalam bentuk grafik *scatter plot* dengan sumbu x adalah atribut tahun_lahir dan sumbu y adalah atribut pekerjaan, hal ini dimaksudkan untuk melihat persebaran partisi yang dipetakan pada masing-masing rentang nilai dari *feature column*. Kemudian, hasil partisi yang telah dilakukan anonimisasi disajikan dalam bentuk tabel yang berisi nilai dari atribut tahun lahir, pekerjaan, partai dan jumlah data (*count*) pada partisi.



Gambar 1. Ilustrasi Hasil Partisi [10]

3. Hasil dan Pembahasan

3.1. Inisialisasi *Dataframe*

Penelitian ini menggunakan bahasa pemrograman Python. Untuk menggunakan algoritma *Mondrian Multidimensional K-Anonymity* yang selanjutnya disebut algoritma *Mondrian*, perlu untuk mengubah format *dataset* dari file CSV menjadi bentuk *dataframe*. *Dataframe* yang digunakan dalam penelitian ini menggunakan *dataframe* dari *library* Pandas. Serta, *dataset* yang digunakan dalam penelitian ini berjumlah 1000 data dan masing-masing data telah distandardisasi dan dibersihkan. Sampel *dataset* biodata Caleg yang digunakan pada penelitian ini dapat dilihat pada Tabel 4. *Dataset* secara lengkap dapat dilihat pada *file* data_penelitian.csv [11]. Dalam proses inisialisasi *dataframe* juga membagi kolom yang berjenis *categorical*. Kolom *categorical* tersebut adalah semua kolom dalam *dataset* yang bukan bertipe numerik. Kolom *categorical* pada *dataset* yang digunakan pada penelitian ini antara lain dapil, partai, dan lokasi_asal.

Tabel 4. Sampel *dataset* biodata caleg.

Dapil	Partai	Jenis Kelamin	Lokasi Asal	Tahun Lahir	Agama	Status Nikah	Pendidikan	Pekerjaan
aceh-besar	PARTAI GERAKAN INDONESIA RAYA	Laki-Laki	JAKARTA BARAT	1956	Islam	Sudah Menikah	S3	Wiraswasta
aceh-besar	PARTAI KEADILAN SEJAHTERA	Laki-Laki	ACEH BESAR	1967	Islam	Sudah Menikah	SMA/Sederajat	Pegawai Pemerintah
bandung-barat	PARTAI DEMOKRASI INDONESIA PERJUANGAN	Perempuan	KOTA BEKASI	1967	Kristen Protestan	Belum Menikah	SMA/Sederajat	Wiraswasta
aceh-besar	PERSATUAN INDONESIA	Perempuan	ACEH BESAR	1986	Islam	Sudah Menikah	D4/S1	Tidak Bekerja
bandung-barat	PARTAI GOLONGAN KARYA	Perempuan	KOTA BANDUNG	1972	Islam	Belum Menikah	SMA/Sederajat	Pengusaha

3.2. Mengubah ke Bentuk Kategori Angka

Mengubah kolom yang memiliki kategori bukan angka ke bentuk kategori angka diperlukan untuk memudahkan komputasi pembelajaran mesin. Kolom yang perlu diubah ke bentuk kategori angka pada

penelitian ini antara lain jenis_kelamin, agama, status_nikah, pendidikan, dan pekerjaan. Contoh hasil perubahan kategori angka dapat dilihat pada Tabel 5.

Tabel 5. Contoh hasil perubahan kategori angka.

Dapil	Partai	Jenis Kelamin	Lokasi Asal	Tahun Lahir	Agama	Status Nikah	Pendidikan	Pekerjaan
aceh-besar	PARTAI GERAKAN INDONESIA RAYA	0	JAKARTA BARAT	1956	1	2	10	3
aceh-besar	PARTAI KEADILAN SEJAHTERA	0	ACEH BESAR	1967	1	2	4	5
bandung-barat	PARTAI DEMOKRASI INDONESIA PERJUANGAN	1	KOTA BEKASI	1967	2	1	4	3
aceh-besar	PERSATUAN INDONESIA	1	ACEH BESAR	1986	1	2	8	1
bandung-barat	PARTAI GOLONGAN KARYA	1	KOTA BANDUNG	1972	1	1	4	4

3.3. Mendapatkan Nilai Rentang

Proses awal dalam algoritma *Mondrian* adalah mendapatkan nilai rentang. Fungsi yang diterapkan akan mengembalikan nilai rentang dari semua kolom untuk partisi *dataframe*. Nilai rentang untuk kolom numerik diambil dari nilai maksimum dikurangi dengan nilai minimum yang terdapat pada kolom. Serta, nilai rentang untuk kolom kategorikal menggunakan nilai unik. Perhitungan nilai rentang dapat dicontohkan dengan kolom Tahun Lahir yang memiliki nilai maksimum = 1997 dan nilai minimum = 1879, sehingga nilai rentang dari kolom Tahun Lahir = 1997 – 1879 = 118. Nilai rentang masing-masing kolom yang diperoleh dapat dilihat pada Tabel 6.

Tabel 6. Nilai rentang masing-masing kolom.

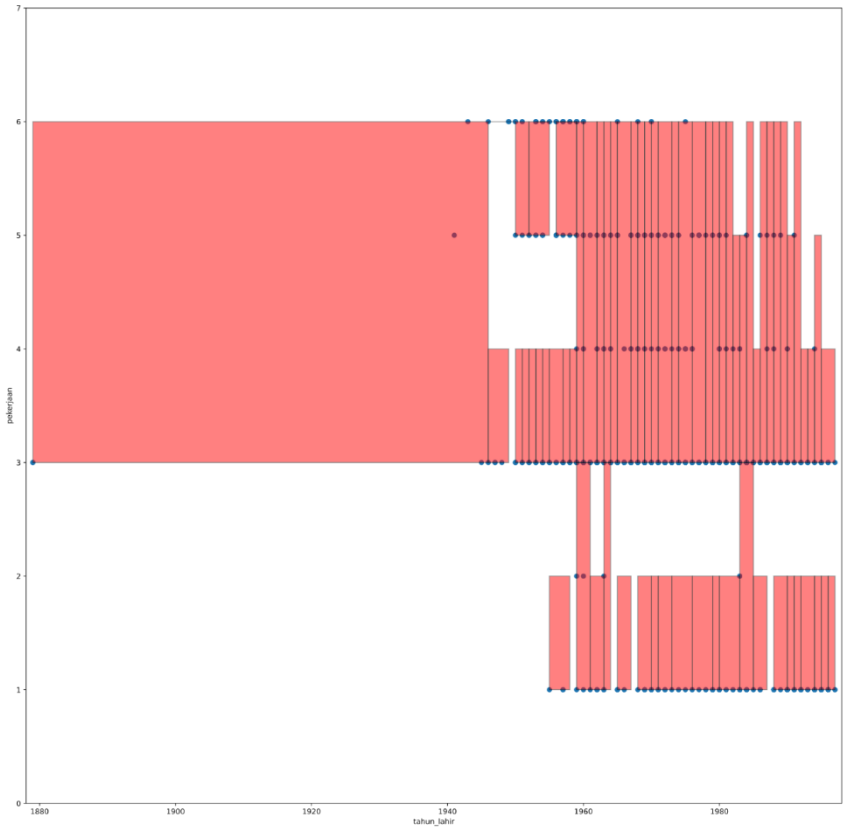
Kolom	Nilai Rentang
Dapil	7
Partai	19
Jenis Kelamin	1
Lokasi Asal	61
Tahun Lahir	118
Agama	4
Status Nikah	2
Pendidikan	7
Pekerjaan	5

3.4. Partisi *Dataset*

Pada tahapan ini *dataset* dipartisi menjadi 2 bagian, yaitu *left partition* dan *right partition*. Dalam proses partisi terdapat beberapa masukan, antara lain *dataframe*, *feature columns*, *sensitive column*, *full spans* yang didapat dari proses sebelumnya, dan *is_k_anonymous* yang menandakan apakah partisi dari *dataframe* telah memenuhi *K-Anonymity* atau tidak. *Feature columns* merupakan kumpulan kolom pada *dataframe* yang akan dipartisi berdasarkan anonimitasnya. Sedangkan, *sensitive column* merupakan kolom yang dapat menjadi kunci pengenal suatu data, sehingga kolom ini digunakan sebagai acuan dalam partisi. Nilai *k* pada penelitian ini sebesar 3 yang merupakan nilai bawaan dari fungsi algoritma *Mondrian*. *Feature columns* yang digunakan pada penelitian ini adalah tahun lahir dan pekerjaan, serta *sensitive column* adalah partai. Jumlah partisi yang diperoleh dari proses ini sebanyak 70 data.

3.5. Visualisasi Partisi

Setelah *dataframe* dipartisi, maka tahap selanjutnya adalah memvisualisasikannya ke dalam bentuk *scatter plot*. *Library* yang digunakan untuk memvisualisasi partisi *dataframe* adalah *Matplotlib*. Hasil visualisasi partisi *dataframe* dapat dilihat pada Gambar 2.



Gambar 2. Hasil visualisasi partisi *dataframe*.

Dari hasil visualisasi dapat dilihat bahwa kolom pekerjaan berada pada sumbu vertikal dan tahun lahir pada sumbu horizontal. Hasil visualisasi ini menunjukkan bagaimana sebuah data menjadi anonimitas. Semakin samar warna dari *scatter plot* maka menunjukkan semakin besar anonimitas data tersebut. Dari hasil visualisasi terlihat hampir semua data memiliki anonimitas yang sama. Hal ini dimungkinkan karena jumlah data partisi yang kurang banyak ataupun kurangnya keberagaman data.

3.6. Menyusun *Anonymized Dataset*

Tahapan terkahir adalah menyusun *anonymized dataset* dan menyimpannya ke dalam format *file* CSV. Untuk menyusun *anonymized dataset* dapat menggunakan fungsi dari algoritma *Mondrian*, yaitu *build_anonymized_dataset.* Setelah *anonymized dataset* berhasil tersusun, selanjutnya dapat disimpan dengan menggunakan fungsi *df.to_csv* yang terdapat pada *library Pandas.* Sampel hasil *anonymized dataset* dapat dilihat pada Tabel 7. Hasil *anonymized dataset* secara lengkap dapat dilihat pada anonymized_dataset.csv [11].

Tabel 7. Sampel hasil *anonymized dataset*.

Tahun Lahir	Pekerjaan	Partai	*Count*
1956.333	1	PARTAI BULAN BINTANG	3
1959.5	1.5	PARTAI AMANAT NASIONAL	2
1963	1.25	PARTAI GERAKAN INDONESIA RAYA	1
1965.452	3.333333	PARTAI AMANAT NASIONAL	4
1969	3.48	PARTAI KEADILAN SEJAHTERA	3

4. Kesimpulan

Penelitian ini mengimplementasikan *K-Anonymity* menggunakan algoritma *Mondrian Multidimensional K-Anonymity* pada *dataset* biodata Calon Legislatif untuk Pemilihan Umum Anggota Legislatif di Indonesia pada tahun 2019. Proses diawali dengan inisialisasi *dataframe* hingga partisi *dataset*. Setelah didapatkan partisi maka *dataset* divisualisasikan ke dalam bentuk *scatter plot* menggunakan Matplotlib. Dari hasil visualisasi terlihat hampir semua data memiliki anonimitas sama, yang dimungkinkan karena jumlah data partisi yang kurang banyak ataupun kurangnya keberagaman data. Oleh karena itu,

penelitian selanjutnya disarankan untuk menggunakan *dataset* yang berjumlah banyak dan memiliki banyak keberagaman data. Serta, nilai *k* pada penggunaan fungsi disesuaikan dengan kebutuhan.

Daftar Pustaka

[1] GDPR.eu, "What is GDPR, the EU's new data protection law?," 2018. https://gdpr.eu/what-is-gdpr/ (accessed Jan. 10, 2021).
[2] P. Regulation, "General Data Protection Regulation," *Intouch*, 2018.
[3] L. Sweeney, "k-anonymity: A model for protecting privacy," *Int. J. Uncertainty, Fuzziness Knowlege-Based Syst.*, 2002, doi: 10.1142/S0218488502001648.
[4] T. Dalenius, "Finding a needle in a haystack-or identifying anonymous census record," *Journal of Official Statistics*. 1986.
[5] R. Li *et al.*, "K-anonymity model for privacy-preserving soccer fitness data publishing," 2018, doi: 10.1051/matecconf/201818903007.
[6] L. Meng, X. Hong, Y. Chen, Y. Ding, and C. Zhang, "K-Anonymous Privacy Preserving Scheme Based on Bilinear Pairings over Medical Data," 2020, doi: 10.1007/978-3-030-59016-1_32.
[7] B. B. Mehta and U. P. Rao, "Privacy preserving big data publishing: A scalable k-anonymization approach using MapReduce," *IET Softw.*, 2017, doi: 10.1049/iet-sen.2016.0264.
[8] K. Wang, W. Zhao, J. Cui, Y. Cui, and J. Hu, "A K-anonymous clustering algorithm based on the analytic hierarchy process," *J. Vis. Commun. Image Represent.*, 2019, doi: 10.1016/j.jvcir.2018.12.052.
[9] N. A. Rakhmawati, "The Biodata of Legislative Candidates for Indonesian General Election 2019," Oct. 2019, doi: 10.5281/ZENODO.3474543.
[10] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional K-anonymity," 2006, doi: 10.1109/ICDE.2006.101.
[11] R. P. Pratama, A. Akbar, and N. A. Rakhmawati, "Caleg 2019 Biodata Datasets For K-Anonymity Research," Jan. 2021, doi: 10.5281/ZENODO.4444802.