

BAB 6

KESIMPULAN DAN SARAN

Pada bab ini akan dijelaskan kesimpulan dari awal hingga akhir penelitian beserta saran untuk penelitian selanjutnya.

6.1 Kesimpulan

Kesimpulan yang dapat diambil dari penelitian ini adalah sebagai berikut:

- Pada penelitian ini, telah dipelajari dua algoritma *k-anonymity* dengan teknik *clustering*, yaitu Algoritma *One Pass k-Means* (OKA) dan *Grading, Centering, Clustering, Generalization* (GCCG). Dalam penelitian ini juga dipelajari cara kerja Algoritma OKA dan GCCG dalam menganonimisasi data. Untuk menganonimisasi data dengan Algoritma OKA dan GCCG, diperlukan beberapa tahapan, yaitu mengurutkan data, melakukan *clustering*, dan generalisasi.
- Perangkat lunak yang dibangun telah mengimplementasikan dua algoritma *k-anonymity* dengan teknik *clustering*, yaitu Algoritma OKA dan GCCG. Pengurutan data pada perangkat lunak diimplementasikan dengan *library* yang disediakan Python, sedangkan tahapan anonimisasi lainnya diimplementasikan sesuai dengan algoritma yang dipakai. Pada perangkat lunak ini, nilai k (konstanta *k-anonymity*) dapat diatur oleh pengguna. Selain itu, tabel hasil anonimisasi juga dapat disimpan ke dalam format `.csv`.
- Selain perangkat lunak untuk anonimisasi, telah dibangun juga perangkat lunak yang mengimplementasikan teknik *data mining* (*clustering* dan klasifikasi) untuk menguji hasil anonimisasi. Algoritma *clustering* yang diimplementasikan adalah *k-Means*, sedangkan algoritma klasifikasi yang diimplementasikan adalah *Decision Tree*. Algoritma diimplementasikan dengan menggunakan *library* Python.
- Pengujian fungsional telah dilakukan pada perangkat lunak anonimisasi. Perangkat lunak anonimisasi sudah menghasilkan hasil anonimisasi yang sesuai dengan perhitungan manual.
- Selain pengujian fungsional, telah dilakukan juga pengujian eksperimental untuk mendapatkan relasi antara nilai k , *information loss*, jenis atribut (kategori, numerik atau campuran), dan waktu eksekusi Algoritma OKA dan GCCG. Dari pengujian ini, disimpulkan bahwa kedua algoritma memiliki kelebihan dan kekurangan masing-masing. Algoritma GCCG memiliki kelebihan dalam hal waktu eksekusi. Waktu eksekusi Algoritma GCCG lebih cepat daripada waktu eksekusi Algoritma OKA. Sedangkan Algoritma OKA memiliki kelebihan dalam hal *information loss*. Algoritma OKA dapat menghasilkan *information loss* yang lebih rendah dari Algoritma GCCG. Hal ini bergantung dari *centroid* yang diambil secara acak pada Algoritma OKA.
- Pengujian eksperimental lainnya dilakukan untuk mendapatkan relasi antara jumlah atribut, tingkat akurasi hasil klasifikasi, koefisien silhouette hasil *clustering*, serta perbedaan hasil *clustering* dan klasifikasi data setelah anonimisasi dengan hasil *clustering* dan klasifikasi data

sebelum anonimisasi. Dari pengujian ini dapat disimpulkan, hasil *clustering* data setelah anonimisasi menghasilkan *cluster* yang lebih baik (koefisien silhouette lebih tinggi) dari data sebelum anonimisasi. Algoritma GCCG menghasilkan *cluster* paling baik. Selain itu, perbedaan hasil *clustering* data yang dianonimisasi Algoritma GCCG dengan data sebelum anonimisasi lebih besar daripada perbedaan hasil *clustering* data yang dianonimisasi Algoritma OKA dengan data sebelum anonimisasi. Dari hasil *clustering*, diketahui bahwa Algoritma GCCG lebih baik daripada Algoritma OKA.

Saat dilakukan klasifikasi, nilai akurasi hasil klasifikasi kedua algoritma tidak menentu, dan Algoritma GCCG menghasilkan akurasi paling tinggi. Sementara itu, Algoritma OKA menghasilkan akurasi yang paling rendah. Hasil klasifikasi data yang dianonimisasi Algoritma GCCG selalu sama dengan hasil klasifikasi data sebelum anonimisasi, sedangkan hasil klasifikasi data yang dianonimisasi Algoritma OKA memiliki sedikit perbedaan dengan hasil klasifikasi data sebelum anonimisasi. Dari hasil klasifikasi, diketahui bahwa Algoritma GCCG lebih baik daripada Algoritma OKA.

- Pada penelitian ini, pada saat pengujian eksperimental dapat disimpulkan juga bahwa *information loss* dan waktu eksekusi dapat diatur dengan menyesuaikan nilai k .
- Pada penelitian ini, pada saat pengujian eksperimental jumlah atribut memengaruhi perbedaan hasil *clustering* dengan data sebelum anonimisasi. Selain itu, jumlah atribut juga memengaruhi koefisien silhouette hasil *clustering*.

6.2 Saran

Saran untuk penelitian selanjutnya adalah sebagai berikut:

- Perangkat lunak yang dibuat pada penelitian ini hanya mengimplementasikan dua algoritma *k-anonymity*, yaitu Algoritma OKA dan GCCG. Algoritma GCCG yang diimplementasikan belum dioptimisasi. Untuk penelitian selanjutnya, Algoritma GCCG yang diimplementasikan dapat disertai dengan metode optimisasinya agar waktu eksekusi algoritma yang diimplementasikan lebih cepat dari yang sebelumnya.
- Pengujian yang dilakukan penelitian ini masih sangat sedikit karena hanya menggunakan set data yang berisi 25 baris. Meskipun perangkat lunak yang dibangun dapat menganonimisasi hingga 100 baris, namun set data yang terlalu besar rawan menyebabkan perangkat lunak tidak merespon. Oleh karena itu, pengujian dilakukan hanya dengan set data berisi 25 baris. Untuk penelitian selanjutnya, pengujian yang dilakukan dapat menggunakan set data yang lebih besar dibandingkan pengujian pada penelitian ini.
- Pada penelitian ini, perhitungan jarak antar setiap baris pada data disimpan dalam sebuah *array* dua dimensi dan dihitung seluruhnya sebelum melakukan proses anonimisasi. Hal ini membuat waktu eksekusi seluruh proses anonimisasi lebih lambat. Untuk penelitian selanjutnya, implementasi perhitungan jarak antar setiap baris pada data dapat dioptimalkan kembali agar waktu eksekusi seluruh proses anonimisasi lebih cepat. Perhitungan dapat dioptimalkan dengan menghitung jarak saat jarak akan dipakai saja, tidak perlu melakukan perhitungan jarak antar setiap baris pada data sekaligus.

DAFTAR REFERENSI

- [1] Moore, A. D. (2002) Privacy: Its meaning and value. *American Philosophical Quarterly*, **40**, 215–227.
- [2] McCallister, E., Grance, T., dan Scarfone, K. A. (2010) Guide to protecting the confidentiality of personally identifiable information (pii). Technical Report SP 800-122. National Institute of Standards & Technology, Gaithersburg, MD, United States.
- [3] Han, J. (2012) *Data Mining: Concepts and Techniques*, 3rd edition. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
- [4] Qian, Q., Ni, S., dan Xie, M. (2017) Clustering based k-anonymity algorithm for privacy preservation. *International Journal of Network Security*, **19**, 1062–1071.
- [5] Torra, V. (2017) *Data Privacy: Foundations, New Developments and the Big Data Challenge*. Springer Nature, Switzerland.
- [6] Faherty, V. E. (2008) Compassionate statistics: Applied quantitative analysis for social services (with exercises and instructions in spss). *Correlation: Spearman's rho and Pearson's r*. SAGE Publications, Inc, Los Angeles.
- [7] Shafranovich, Y. (2005) Common format and mime type for comma-separated values (csv) files. RFC 4180. RFC Editor, <http://www.rfc-editor.org>.