# BAB 7

# KESIMPULAN DAN SARAN

Pada bab ini akan dijelaskan kesimpulan penelitian beserta saran untuk penelitian selanjutnya.

## 7.1 Kesimpulan

Kesimpulan yang dapat diambil dari penelitian ini adalah sebagai berikut:

1. Secara singkat tahapan *greedy k-member clustering* adalah mencari kandidat anggota sebuah kelompok data menggunakan nilai *information loss* paling kecil.

2. Secara singkat tahapan *k-anonymity* adalah mengambil sebuah *cluster* dan melakukan proses anonimisasi pada setiap kolom *cluster* yang bernilai unik menggunakan pohon DGH.

3. Secara singkat, implementasi algoritma *greedy k-member clustering* dilakukan pada perangkat lunak anonimisasi menggunakan *library* Spark SQL, karena banyak operasi pada algoritma ini yang dapat diimplementasikan dengan kueri SQL. Implementasi algoritma ini dapat menimbulkan permasalahan memori akibat prisip *lazy evaluation* pada Spark.

4. Performa algoritma *greedy k-member clustering* pada lingkungan *big data* sangat buruk (mencapai waktu 5 jam untuk pengelompokan 10.000 data), sedangkan performa dari algoritma *k-anonymity* pada lingkungan *big data* cukup baik (mencapai waktu kurang dari 1 jam untuk melakukan anonimisasi 10.000 data).

5. Performa pemodelan *data mining* menggunakan teknik *clustering* (*k-means*) dan teknik klasifikasi (*naive bayes*) pada lingkungan *big data* sangat baik karena menggunakan *library* bawaan Spark yaitu Spark MLlib.

6. Kualitas informasi pengelompokan data dengan *greedy k-member clustering* cukup baik jika menggunakan kolom campuran yang tidak menggunakan terlalu banyak kolom numerik dan menggunakan jumlah *quasi-identifier* secukupnya. Hal ini terbukti pada pengujian kualitas informasi, dimana kondisi tersebut memiliki *total information loss* yang paling rendah.

7. Kualitas hasil *data mining* menggunakan pemodelan *clustering k-means* cukup baik untuk diterapkan pada data anonimisasi, karena melalui pengujian hasil *data mining* yang dilakukan sebelumnya, diketahui bahwa *silhouette score* dan perbedaan hasil *clustering* sebelum dan setelah dilakukan anonimisasi cukup kecil.

8. Kualitas hasil *data mining* menggunakan pemodelan klasifikasi *naive bayes* cukup baik untuk diterapkan pada data anonimisasi, karena melalui pengujian hasil *data mining* yang dilakukan sebelumnya, diketahui bahwa tingkat akurasi dan perbedaan hasil klasifikasi sebelum dan setelah dilakukan anonimisasi cukup kecil.

## 7.2 Saran

Saran untuk penelitian selanjutnya adalah sebagai berikut:

- Pada penelitian ini, diketahui bahwa algoritma *greedy k-member clustering* memiliki waktu komputasi yang sangat lama jika merancang algoritma pengelompokan secara mandiri. Solusi dari permasalahan pengelompokan data pada lingkungan *big data* adalah menggunakan *library* KMeans pada Spark MLlib agar waktu eksekusinya menjadi lebih efisien.

- Pada penelitian ini, pernah terjadi *error* `java.lang.OutOfMemoryError` terkait *lazy evaluation* pada Spark yang diterapkan pada algoritma yang iteratif. Dikutip dari Medium[1], masalah ini terjadi ketika fungsi *transformation* `filter(),union()` dipanggil pada setiap iterasi. Fungsi *transformation* adalah fungsi dengan biaya komputasi yang mahal, terutama jika dipanggil beberapa kali dalam satu iterasi. Konsep *lazy evaluation* pada Spark mirip dengan konsep rekursif, dimana fungsi *transformation* akan dijalankan saat fungsi *action* dipanggil. *Error* ini terjadi ketika fungsi *action* dipanggil pada iterasi tertentu yang membuat fungsi *transformation* pada iterasi sebelumnya juga ikut dijalankan. Solusi yang dapat diterima adalah menyimpan dan membaca hasil komputasi pada sistem penyimpanan HDFS.

- Struktur data pada pemrosesan *big data* perlu diperhatikan. Diusahakan untuk memilih struktur data Dataframe/RDD, karena hanya operasi tersebut yang dapat berjalan secara paralel. Selain itu diusahakan untuk tidak terlalu banyak menggunakan operasi perulangan. Pada kasus tertentu, operasi perulangan dapat diganti dengan operasi kueri SQL.

[1] https://medium.com/@cyneo41/memory-issues-caused-by-lazy-evaluation-in-spark-a419517c1732

# DAFTAR REFERENSI


[1] Han, J., Kamber, M., dan Pei, J. (2012) *Data Mining: Concepts and Techniques*, 3rd edition. Morgan Kaufmann, Waltham, USA.

[2] Veronica S. Moertini, M. T. A. (2020) *Pengantar Data Science dan Aplikasinya bagi Pemula*, 1st edition. Unpar Press, Bandung, Indonesia.

[3] MENDES, R. dan VILELA, J. P. (2017) Privacy-preserving data mining: Methods, metrics, and applications. *IEEE Access*, **5**, 1–21.

[4] Vanessa Ayala Rivera, T. C., Patrick McDonagh (2014) A systematic comparison and evaluation of k-anonymization algorithms for practitioners. *Transaction on Data Privacy 7*, **1**, 337–370.

[5] Judith Hurwitz, M. K., Alan Nugent (2013) *Big Data For Dummies*, 1st edition. John Wiley and Sons, Inc., Hoboken, New Jersey.

[6] Holden Karau, P. W., Andy Konwinski (2015) *Learning Spark*, 1st edition. OReilly Media, Inc, Sebastopol, CA.

[7] Byun, J.-W., Kamra, A., Bertino, E., dan Li, N. (2007) Efficient k-anonymity using clustering techniques. *CERIAS Tech Report 2006-10*, **1**, 1–12.

[8] Martin Odersky, B. V., Lex Spoon (2008) *Programming in Scala*, 1st edition. Artima, Inc, Mountain View, California.