

## BAB 6

### KESIMPULAN DAN SARAN

Bab ini berisi kesimpulan dari awal hingga akhir penelitian beserta saran untuk penelitian selanjutnya.

#### 6.1 Kesimpulan

Kesimpulan yang dapat ditarik dari penelitian ini adalah sebagai berikut.

- Teknik *Random Rotation Perturbation* mengacak data dengan cara merotasi setiap titik yang merepresentasikan sebuah objek data pada bidang Euclidean sehingga nilai setiap objek data berubah tetapi jarak Euclidean antara setiap titik dengan titik lainnya tidak berubah. Oleh karena itu, data yang telah diacak tersebut masih dapat digunakan untuk teknik penambangan data yang hanya memanfaatkan jarak Euclidean.
- Teknik *Random Projection Perturbation* mengacak data dengan cara memproyeksikan data yang berdimensi cukup besar ke dimensi yang lebih kecil. Teknik ini memanfaatkan *Johnson-Lindenstrauss Lemma* yang menyatakan bahwa titik-titik pada bidang Euclidean  $d$ -dimensi dapat diproyeksikan ke bidang Euclidean yang berdimensi lebih kecil dari  $d$  tetapi jarak Euclidean antara setiap titik tetap terjaga dengan distorsi yang terkontrol tetapi dengan syarat  $d$  harus cukup besar. Oleh karena itu, data yang telah diacak tersebut masih dapat digunakan untuk teknik penambangan data yang hanya memanfaatkan jarak Euclidean.
- Perangkat lunak diimplementasikan dengan bahasa pemrograman Python. Antarmuka perangkat lunak dibuat dengan bantuan *framework* antarmuka bernama Kivy. Fungsi-fungsi pada perangkat lunak yang berkaitan dengan masukan, keluaran, dan implementasi algoritma dibantu oleh berbagai *library* seperti Numpy, Scipy, Pandas, dan Scikit-learn. Perangkat lunak hanya dapat menerima masukan dataset berupa dokumen *comma-separated values* yang setiap fiturnya bersifat numerik saja.
- Teknik *Random Rotation Perturbation* diimplementasikan dengan cara mengolah dataset yang ingin diacak menjadi matriks, lalu melakukan transformasi translasi pada matriks dataset tersebut dengan cara mengkalikannya dengan matriks translasi acak yang dibuat mengikuti distribusi *uniform* dengan rentang nilai  $[0, 100]$ . Kemudian transformasi rotasi dilakukan pada matriks dataset yang telah ditranslasi dengan cara mengkalikannya dengan matriks rotasi acak yang dibuat mengikuti distribusi Haar dengan bantuan *library* Scipy.
- Teknik *Random Projection Perturbation* diimplementasikan dengan cara mengolah dataset yang ingin diacak menjadi matriks, lalu melakukan pemeriksaan persyaratan terlebih dahulu dan menentukan target dimensi yang menjadi dimensi dataset hasil pengacakan. Kemudian apabila persyaratan terpenuhi maka proyeksi dapat dilakukan dengan cara mengkalikan matriks dataset tersebut dengan matriks proyeksi acak yang dibuat mengikuti distribusi normal.
- Teknik *Random Rotation Perturbation* terbukti menjaga jarak Euclidean dengan sempurna tanpa ada distorsi dan teknik *Random Projection Perturbation* terbukti menjaga jarak Euclidean dengan distorsi yang sesuai dengan keinginan pengguna.
- Berdasarkan pengujian eksperimental yang telah dilakukan, metode *Randomization* mengacak

data dan berbagai properti yaitu rata-rata, standar deviasi, nilai terkecil, nilai terbesar, kuartil bawah, kuartil tengah, dan kuartil atas setiap kolom tanpa merusak jarak Euclidean pada data tersebut.

- *Dataset* yang diacak dengan teknik *Random Rotation Perturbation* terbukti masih dapat digunakan untuk penambangan data klasifikasi menggunakan teknik *k-nearest neighbors* dengan akurasi model klasifikasi yang sama persis. Sedangkan untuk teknik *Random Projection Perturbation* memiliki akurasi model klasifikasi yang sangat mirip dengan akurasi model yang dilatih dengan *dataset* asli dan jumlah tetangga (nilai variabel  $k$ ) pada model yang memiliki akurasi tertinggi juga berbeda dengan model yang dilatih dengan *dataset* asli.
- *Dataset* yang diacak dengan teknik *Random Rotation Perturbation* terbukti masih dapat digunakan untuk penambangan data *clustering* menggunakan teknik *k-means* dengan hasil *cluster* yang sama persis dengan model yang dilatih dengan *dataset* asli. Sedangkan untuk teknik *Random Projection Perturbation* memiliki hasil *cluster* yang sangat mirip dengan model yang dilatih menggunakan *dataset* asli.
- Waktu eksekusi dalam proses penambangan data klasifikasi dan *clustering* menggunakan *dataset* yang diacak dengan teknik *Random Rotation Perturbation* tidak memiliki perbedaan yang signifikan dibandingkan penambangan data dengan *dataset* asli. Sedangkan untuk waktu eksekusi dalam proses penambangan data klasifikasi dan *clustering* menggunakan *dataset* yang diacak dengan teknik *Random Projection Perturbation* lebih cepat dibandingkan penambangan data dengan *dataset* asli karena jumlah fitur yang ada pada *dataset* lebih sedikit.

## 6.2 Saran

Saran untuk penelitian selanjutnya adalah sebagai berikut.

- Pada penelitian ini tidak menguji apakah data yang telah diacak berpotensi dapat dikembalikan ke aslinya. Untuk penelitian selanjutnya, metode *Randomization* dapat diuji untuk mengetahui apakah ada potensi data yang telah diacak dengan metode *Randomization* dapat dikembalikan ke aslinya.
- Pada penelitian ini metode *Randomization* diuji dengan menggunakan penambangan data biasa, bukan dalam lingkungan *big data*. Untuk penelitian selanjutnya, metode *Randomization* dapat diuji untuk diimplementasikan dalam lingkungan *big data* dan mengukur kecepatan dan kualitas teknik *Random Rotation Perturbation* dan *Random Projection Perturbation*.
- Untuk penelitian selanjutnya dapat dilakukan pengujian lebih lanjut lagi terhadap teknik *Random Projection Perturbation* menggunakan data yang lain untuk mengetahui nilai variabel *epsilon* yang tepat sehingga kualitas model penambangan data yang dilatih menggunakan *dataset* yang telah diacak masih dapat ditoleransi perbedaannya dari yang asli dan bagaimana menanggulangi masalah nilai variabel  $k$  (dimensi minimal) pada teknik *Random Projection Perturbation* semakin menaik seiring model penambangan data dilatih dengan data yang baru.
- Untuk penelitian selanjutnya dapat dilakukan eksperimen mengenai perbedaan teknik *Principal Component Analysis* dengan teknik *Random Projection Perturbation* dalam mereduksi dimensi data. Eksperimen dapat bertujuan untuk menganalisa kualitas hasil dari kedua teknik tersebut dan apakah ada potensi sebuah data yang telah direduksi dengan kedua teknik tersebut dapat dikembalikan ke aslinya.

## DAFTAR REFERENSI

- [1] Keyvanpour, M. dan Moradi, S. S. (2011) Classification and evaluation the privacy preserving data mining techniques by using a data modification–based framework. *International Journal on Computer Science and Engineering*, **3**, 862–870.
- [2] NIST Special Publication 800-122 (2010) *Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)*. National Institute of Standards and Technology, U.S. Department of Commerce, Erika McCallister, Tim Grance, Karen Scarfone. Gaithersburg, Maryland.
- [3] Oliveira, S. R. M. dan Zaiane, O. R. (2004) Towards standardization in privacy-preserving data mining. *ACM SIGKDD 3rd Workshop on Data Mining Standards*, **3**, 862–870.
- [4] MENDES, R. dan VILELA, J. P. (2017) Privacy-preserving data mining: Methods, metrics, and applications. *IEEE Access*, **5**, 10562–10582.
- [5] Han, J., Kamber, M., dan Pei, J. (2012) *Data Mining: Concepts and Techniques*, 3rd edition. Morgan Kaufmann, Waltham.
- [6] Torra, V. (2017) *Data Privacy: Foundations, New Developments and the Big Data Challenge*, 1st edition. Springer, Warsaw.
- [7] Chen, K. dan Liu, L. (2005) A random rotation perturbation approach to privacy preserving data classification. Technical Report GIT-CC-05-12. Georgia Institute of Technology, Georgia.
- [8] Agrawal, R. dan Srikant, R. (2000) Privacy preserving data mining. *In Proceedings of the ACM SIGMOD*, **3**, 439–450.
- [9] STEWART, G. W. (1980) The efficient generation of random orthogonal matrices with an application to condition estimators. *SIAM Journal on Numerical Analysis*, **17**, 403–409.
- [10] Johnson, W. B. dan Lindenstrauss, J. (1984) Extensions of lipschitz mappings into a hilbert space. *Contemporary Mathematics*, **26**, 189–206.
- [11] Kun Liu, J. R., Hillol Kargupta (2006) Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, **18**, 92–106.
- [12] Ella Bingham, H. M. (2001) Random projection in dimensionality reduction: Applications to image and text data. *In Proceedings of the ACM SIGKDD*, **7**, 245–250.
- [13] Rossum, G. V. (1995) Python tutorial. Technical Report CS-R9526. Centrum voor Wiskunde en Informatica (CWI), Amsterdam.
- [14] Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Jarrod Millman, K., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C., Polat, İ., Feng, Y., Moore, E. W., Vand erPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa,

- F., van Mulbregt, P., dan Contributors, S. . . (2020) SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, **17**, 261–272.
- [15] Wes McKinney (2010) Data Structures for Statistical Computing in Python. Bagian dari Stéfan van der Walt dan Jarrod Millman (ed.), *Proceedings of the 9th Python in Science Conference*, pp. 56 – 61.
- [16] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., dan Duchesnay, E. (2011) Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.
- [17] Hunter, J. D. (2007) Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, **9**, 90–95.