BAB 5

PENUTUP

5.1 Kesimpulan

1. Pendeteksian transaksi *fraud* berdasarkan data berlabel yang memiliki fitur numerik dan kategori dapat dipandang sebagai masalah pembelajaran mesin tersupervisi. Sistem kecerdasan detektor *fraud* dibuat berupa *pipeline* pembelajaran mesin, dengan model yang linier (Regresi Logistik atau *Linear Support Vector Classifier*) atau non-linier (Pohon Keputusan atau *Random Forest*). Dari performa model-model non-linier yang lebih baik dibandingkan model-model linier, disimpulkan bahwa masalah klasifikasi *fraud* dalam kasus dataset BankSim [6] ini adalah masalah non-linier.
2. Secara umum, penggunaan SMOTE-NC dalam *pipeline* pembelajaran mesin dapat meningkatkan performa model – yang diukur dari skor F_1 – dan efisiensi data *training* dalam masalah klasifikasi *fraud*. Dari ANOVA, model-model yang terbukti mengalami peningkatan performa secara signifikan adalah Regresi Logistik, *Linear Support Vector Classifier*, dan *Random Forest*.
3. Peningkatan performa (skor F_1) model oleh SMOTE-NC terjadi karena adanya *trade-off* antara *precision* dan *recall*. Bila model diibaratkan sebagai entitas hidup, keberadaan SMOTE-NC menyebabkan model menjadi "lebih sensitif namun kurang presisi"; yaitu menjadi lebih "waspada" dan sering memprediksi suatu transaksi sebagai *fraud*, sehingga akan menemukan lebih banyak transaksi *fraud* yang sesungguhnya, tetapi juga lebih banyak "salah mengira" transaksi *non-fraud* sebagai *fraud*.

5.2 Saran

Untuk penelitian-penelitian selanjutnya yang berfokus menyelidiki efisiensi data *training* pada model, sebaiknya digunakan banyaknya sampel n – alih-alih fraksi data $T\%$ – sebagai faktor eksperimen. Eksperimen perlu diulang pada beragam dataset agar kajian penggunaan SMOTE-NC terhadap model-model pembelajaran mesin menjadi lebih menyeluruh.

Khusus untuk masalah klasifikasi transaksi *fraud*, perlu digunakan dataset yang mencakup lebih banyak fitur informatif, sehingga dapat meningkatkan performa detektor *fraud*. Misalnya, apabila *customer* biasanya hanya bertransaksi di lingkup kota tempat tinggalnya, maka fitur lokasi tempat tinggal *customer* dan fitur lokasi *merchant* mungkin dapat berguna sebagai indikator *fraud*, sehingga transaksi dengan *merchant* di luar kota atau negara *customer* tersebut menjadi anomali dan patut dicurigai sebagai *fraud*.

Kajian pengaruh keberadaan SMOTE-NC terhadap fungsi keputusan model perlu dilakukan secara lebih menyeluruh daripada kajian dalam subbab 4.3, dengan usulan menggunakan metode pengukuran *feature importance* yang tidak bergantung pada tipe modelnya (*model-agnostic*), seperti misalnya *permutation feature importance*. Berdasarkan perbandingan fungsi keputusan model Regresi Logistik dan Pohon Keputusan antara tanpa *oversampler* dan dengan SMOTE-NC, diyakini bahwa keberadaan SMOTE-NC mempengaruhi nilai bobot kepentingan fitur-fitur yang menjadi indikator *fraud*. Dalam model Regresi Logistik, keberadaan SMOTE-NC mempertegas signifikansi peran fitur-fitur indikator *fraud*. Dalam model Pohon Keputusan, keberadaan SMOTE-NC tidak

secara drastis mengubah bobot kepentingan fitur-fitur indikator *fraud*, karena fungsi keputusan dari model non-linier ini hanya bergantung pada satu fitur saja, sehingga hasil interpolasi pada fitur-fitur lainnya oleh SMOTE-NC menjadi kurang bermanfaat.

DAFTAR REFERENSI

[1] Weiss, G. M. (2013) Foundations of imbalanced learning. Bagian dari He, H. dan Ma, Y. (ed.), *Imbalanced Learning Foundations, Algorithms, and Applications*. John Wiley & Sons, Inc., New Jersey.

[2] Chawla, N. V., Bowyer, K. W., Hall, L. O., dan Kegelmeyer, W. P. (2002) Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, **16**, 321–357.

[3] Sifa, R., Hadiji, F., Runge, J., Drachen, A., Kersting, K., dan Bauckhage, C. (2015) Predicting purchase decisions in mobile free-to-play games. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, California, USA, 14-18 November, pp. 79–85. The AAAI Press, Palo Alto.

[4] Akyon, F. C. dan Kalfaoglu, M. E. (2019) Instagram fake and automated account detection. *2019 Innovations in Intelligent Systems and Applications Conference (ASYU)*, Izmir, Turkey, 31 October - 2 November, pp. 1–7. IEEE.

[5] Gök, E. C. dan Olgun, M. O. (2021) Smote-nc and gradient boosting imputation based random forest classifier for predicting severity level of covid-19 patients with blood samples. *Neural Computing and Applications*, **33**, 15693–15707.

[6] Lopez-Rojas, E. A. dan Axelsson, S. (2014) Banksim: A bank payment simulation for fraud detection research. *Inproceedings 26th European Modeling and Simulation Symposium*, Bordeaux, France, 10-12 September, pp. 144–152. CAL-TEK S.r.l., Rende.

[7] Trevor Hastie, R. T. dan Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd edition. Springer, New York.

[8] Branco, P., Torgo, L., dan Ribeiro, R. P. (2016) A survey of predictive modeling on imbalanced domains. *Association for Computing Machinery*, **49**, 1–50.

[9] Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., dan Herrera, F. (2018) *Learning from Imbalanced Data Sets*, 1st edition. Springer International Publishing AG, Cham.

[10] Han, J., Kamber, M., dan Pei, J. (2012) *Data mining : concepts and techniques*, 3rd edition. Morgan Kaufmann Publishers, Waltham.

[11] Domingos, P. (2000) A unified bias-variance decomposition and its applications. *Proceedings of the 17th International Conference on Machine Learning*, California, USA, 29 June-2 July, pp. 231–238. Stanford University.

[12] Mendenhall, W., Beaver, R. J., dan Beaver, B. M. (2011) *Introduction to Probability and Statistics*, 14th edition. Brooks/Cole, Boston.

[13] Dean, A., Voss, D., dan Draguljić, D. (2017) *Design and Analysis of Experiment*, 2nd edition. Springer International Publishing AG, Cham.