BAB 5

PENUTUP

5.1 Kesimpulan

Berdasarkan hasil-hasil penelitian yang telah diperoleh sebelumnya, diperoleh kesimpulan-kesimpulan berikut.

1. Analisis komponen utama berhasil mereduksi dimensi data penyakit jantung koroner serta menghilangkan multikolinearitas antara variabel bebas dengan mempertahankan mayoritas informasi. Sebanyak 8 variabel numerik berhasil direduksi menjadi 5 komponen utama dengan menjaga minimal 80% keseluruhan informasi. Model dengan analisis komponen utama memberikan hasil yang tidak terlalu buruk dibandingkan model lainnya, tetapi performa model menjadi turun dibandingkan jika tidak dilakukan analisis komponen utama.

2. Penggantian nilai ambang batas berhasil meningkatkan nilai *F1-score* pada semua model untuk data penyakit jantung koroner. Akan tetapi, penggantian nilai ambang batas pada model untuk data penyakit kanker paru-paru menurunkan nilai *F1-score*. Hal yang dapat disimpulkan adalah dengan mengganti nilai ambang batas dapat diperoleh hasil prediksi dari model yang lebih baik, tetapi hal tersebut tidak berlaku untuk semua data.

3. Metode *undersampling* dapat memberikan hasil prediksi yang lebih baik dibandingkan model tanpa adanya modifikasi pada data *training*. Akan tetapi, metode *undersampling* melibatkan pembuangan data yang dapat mengakibatkan hilangnya informasi penting dalam data tersebut.

4. Regresi logistik berbobot menghasilkan model dengan performa yang sebanding atau lebih baik jika dibandingkan dengan metode lainnya sambil mempertahankan semua informasi data. Model terbaik untuk kedua data diperoleh pada model regresi logistik berbobot dengan bobot dari *hyperparameter tuning*.

5. *Confusion matrix* dapat digunakan untuk membandingkan hasil klasifikasi berbagai model dan menentukan model yang lebih baik.

5.2 Saran

Saran untuk penelitian selanjutnya adalah penggunaan teknik *resampling* lain, dengan KNN (*K-Nearest Neighbors*) [15], SMOTE [16], atau SMOTE-NC [16] yang merupakan teknik *oversampling*

sebagai pembanding lainnya. Penentuan model yang terbaik dapat dilakukan dengan cara lain, tidak hanya berdasarkan *confusion matrix* atau ukuran *precision*, *recall*, dan *F1-score*, seperti menggunakan nilai *Area Under Curve* (AUC) [17]. Selain itu, dapat digunakan data yang lebih tidak seimbang dibanding data yang digunakan pada skripsi ini untuk melihat efek dari teknik-teknik yang telah dijelaskan dalam mengatasi data tidak seimbang.

DAFTAR REFERENSI

[1] Finegold, J. A., Asaria, P., dan Francis, D. P. (2013) Mortality from ischaemic heart disease by country, region, and age: statistics from World Health Organisation and United Nations. *International Journal of Cardiology*, **168**, 934–945.

[2] Kashf, D. W. A., Okasha, A. N., Sahyoun, N. A., El-Rabi, R. E., dan Abu-Naser, S. S. (2018) Predicting DNA lung cancer using artificial neural network. *International Journal of Academic Pedagogical Research (IJAPR)*, **2**, 6–13.

[3] Choi, Y., An, J., Ryu, S., dan Kim, J. (2022) Development and evaluation of machine learning-based high-cost prediction model using health check-up data by the National Health Insurance Service of Korea. *International Journal of Environmental Research and Public Health*, **19**, 13672.

[4] Park, H.-A. (2013) An introduction to logistic regression: from basic concepts to interpretation particular attention to nursing domain. *Journal of Korean Academy of Nursing*, **43**, 154–164.

[5] Paelongan, P. L. dan Palupi, I. (2022) Lung cancer prediction model using logistic linear regression with imbalanced dataset. *Indonesia Journal on Computing (Indo-JC)*, **7**, 1–14.

[6] Rencher, A. C. (2002) *Methods of Multivariate Analysis*, 2nd edition. A John Wiley & Sons.

[7] Maalouf, M. dan Siddiqi, M. (2014) Weighted logistic regression for large-scale imbalanced and rare events data. *Knowledge-Based Systems*, **59**, 142–148.

[8] Derczynski, L. (2016) Complementarity, F-score, and NLP evaluation. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia, 23-28 May, pp. 261–266. European Language Resources Association (ELRA), Paris.

[9] Maloof, M. (2003) Learning when data sets are imbalanced and when costs are unequal and unknown. *Analysis*, **21**.

[10] Krawczyk, B. (2016) Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, **5**, 221–232.

[11] Hosmer, D. W. dan Lemeshow, S. (2000) *Applied Logistic Regression*, 2nd edition. John Wiley and Sons Inc.

[12] Klein, A. dan Mélard, G. (2023) An algorithm for the fisher information matrix of a VARMAX process. *Algorithms*, **16**, 364.

[13] Gill, J. dan King, G. (2004) What to do when your hessian is not invertible: alternatives to model respecification in nonlinear estimation. *Sociological Methods and Research*, **32**, 54–87.

[14] Chicco, D., Tötsch, N., dan Jurman, G. (2021) The Matthews correlation coefficient (mcc) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData mining*, **14**, 1–22.

[15] Beckmann, M., Ebecken, N. F., dan Pires de Lima, B. S. (2015) A KNN undersampling approach for data balancing. *Journal of Intelligent Learning Systems and Applications*, **7**, 104–116.

[16] 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.

[17] Salih, A. A. dan Abdulazeez, A. M. (2021) Evaluation of classification algorithms for intrusion detection system: a review. *Journal of Soft Computing and Data Mining*, **2**, 31–40.