BAB 5

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Pada skripsi ini telah diuraikan penelitian terkait model prediksi menggunakan *Random Forest* dengan metode reduksi dimensi analisis komponen utama terhadap data berdimensi besar dan juga timpang. Berdasarkan pembahasan yang diuraikan, diperoleh beberapa kesimpulan sebagai berikut:

1. Penerapan metode analisis komponen utama (PCA) untuk data berdimensi besar dengan banyak variabelnya yang saling berkorelasi lebih baik dalam mereduksi dimensi data, terlebih lagi apabila berkorelasi kuat.

2. Penerapan metode analisis komponen utama tidak terlalu memberikan dampak yang signifikan terhadap model *Random Forest* meskipun terdapat kemungkinan untuk meningkatkan jumlah benar terprediksi kelas minoritas (namun hal ini dalam jumlah yang kecil). Artinya, untuk data berdimensi besar terbukti model *Random Forest* memang sudah cukup baik melakukan prediksi. Akan tetapi, setelah penerapan, terdapat hal yang bisa dipastikan dan menjadi pertimbangan yaitu waktu pelatihan model lebih cepat dan juga dimensi data yang tentunya lebih kecil daripada data aslinya.

   (a) Untuk himpunan data kanker payudara, penerapan model *Random Forest* saja sudah cukup untuk digunakan dalam membuat prediksi. Data ini menampilkan hasil yang baik menggunakan *Random Forest* karena terdapat beberapa korelasi yang cukup kuat antara variabel bebasnya.

   (b) Untuk data kebangkrutan perusahaan, model *Random Forest* dengan adanya penerapan PCA dapat dipertimbangkan untuk digunakan karena memiliki dimensi data yang lebih kecil. Penerapan PCA mereduksi dimensi data dan menyederhanakan informasi terkait korelasi antar variabel bebasnya (yang disusun dalam bentuk kombinasi linear) sehingga dengan dimensi yang lebih kecil pun masih cukup baik dalam mempertahankan sebanyak mungkin informasi terkandung dalam data asli dan menghimpun variabilitas datanya. Adapun hasil prediksi yang optimal masih sulit dicapai pada kasus ini, terlihat dari nilai *F1 Score* yang kecil, karena berhadapan dengan data yang sangat timpang.

3. Adanya penerapan metode analisis komponen utama meringkas sebagian besar informasi yang terkandung dalam *variable importance* dari data asli, sehingga dapat diwakili oleh visualisasi lebih kecil, namun tetap dapat memberikan gambaran besar informasi dari *variable importance* seperti dalam model *Random Forest* dengan data asli.

5.2 Saran

Berdasarkan penelitian dan pembahasan yang telah dilakukan, terdapat saran yang dapat diimplementasikan untuk penelitian lebih lanjut, yaitu :

1. Tampaknya perlu penanganan lain untuk mengatasi karakteristik data timpang. Dalam hal ini dapat dicoba dan dilakukan teknik *oversampling* atau *undersampling* untuk pra-pemrosesan data berdimensi besar yang juga memiliki karakteristik data timpang, sebelum diterapkan metode-metode yang tertera pada penelitian ini. Untuk *oversampling* misalnya menggunakan *Random Oversampling* dan SMOTE (*Synthetic Minority Over-sampling Technique*). Untuk *undersampling* misalnya menggunakan *Random Undersampling* dan *Tomek Links*.

2. Dapat dilakukan pengembangan dengan menggunakan teknik lain seperti misalnya *Bayesian Additive Regression Tree* yang lebih akurat karena memiliki informasi utama (*prior*) yang dapat meningkatkan prediksi.

DAFTAR REFERENSI

[1] Kulkarni, V. dan Sinha, P. (2013) Random forest classifiers: A survey and future research directions. *International Journal of Advanced Computing*, **36**, 1144–1153.

[2] Zhu, C., Idemudia, C. U., dan Feng, W. (2019) Improved logistic regression model for diabetes prediction by integrating pca and k-means techniques. *Informatics in Medicine Unlocked*, **17**, 100179.

[3] Jamal, A., Handayani, A., Septiandri, A., Ripmiatin, E., dan Effendi, Y. (2018) Dimensionality reduction using pca and k-means clustering for breast cancer prediction. *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, **9**, 192–201.

[4] Saini, H., Sayal, R., Govardhan, A., dan Buyya, R. (2021) *Innovations in Computer Science and Engineering: Proceedings of 8th ICICSE*. Springer Nature Singapore, Singapore.

[5] Johnson, R. dan Wichern, D. (2007) *Applied Multivariate Statistical Analysis*, 6th edition. Pearson, New Jersey.

[6] Deisenroth, M., Faisal, A., dan Ong, C. (2020) *Mathematics for Machine Learning*. Cambridge University Press, Cambridge.

[7] Berry, M. W., Mohamed, A., dan Yap, B. W. (2019) *Supervised and Unsupervised Learning for Data Science*. Springer Cham, Cham, Switzerland.

[8] James, G., Witten, D., Hastie, T., dan Tibshirani, R. (2013) *An Introduction to Statistical Learning*, 2nd edition. Springer, New York.

[9] Zelterman, D. (2015) *Applied Multivariate Statistics with R*. Springer Cham, New York.

[10] Rizalde, F. A. (2022) Metode Principal Component Analysis pada Faktor-Faktor yang Mempengaruhi Kemiskinan Pulau Sumatera. Skripsi. Universitas Riau, Indonesia.

[11] Shalev-Shwartz, S. dan Ben-David, S. (2014) *Understanding Machine Learning: From Theory to Algorithms*. Cambridge university press, New York.

[12] Serrano, L. (2021) *Grokking Machine Learning*. Simon and Schuster, New York.

[13] Zheng, A. dan Safari, a. O. M. C. (2015) *Evaluating Machine Learning Models*. O'Reilly Media, Incorporated, California.

[14] Robles, E., Zaidouni, F., Mavromoustaki, A., dan Refael, P. (2020) Threshold optimization in multiple binary classifiers for extreme rare events using predicted positive data. *AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering (1)*, Stanford University, Palo Alto, California, USA, 23-25 Maret.

[15] Lin, X., Hou, Z. J., Chen, Y., Rose, S., Ma, Y., dan Pan, F. (2020) Probabilistic forecasting of generators startups and shutdowns in the miso system based on random forest. *2020 IEEE Power & Energy Society General Meeting (PESGM)*, Montreal, Quebec, Canada, 2-6 Agustus, pp. 1–5. IEEE.

[16] Rais, M., Goejantoro, R., dan Prangga, S. (2021) Optimalisasi k-means cluster dengan principal component analysis pada pengelompokan kabupaten/kota di Pulau Kalimantan berdasarkan indikator tingkat pengangguran terbuka. *EKSPONENSIAL*, **12**, 129–136.

[17] Rencher, A. C. dan Christensen, W. F. (2012) *Methods of Multivariate Analysis*. John Wiley & Sons, Provo, Utah.

[18] Hanif, I. (2016) Pendekatan analisis biplot dan swot untuk menganalisis daya saing ekonomi indonesia menghadapi masyarakat ekonomi asean. *Prosiding Seminar Nasional Matematika dan Statistika (SEMASTAT) 2016*, Padang, Indonesia, 25-26 Februari, pp. 651–656. FORSTAT.

[19] Mulugeta, G., Zewotir, T., Tegegne, A. S., Juhar, L. H., dan Muleta, M. B. (2023) Classification of imbalanced data using machine learning algorithms to predict the risk of renal graft failures in Ethiopia. *BMC Medical Informatics and Decision Making*, **23**, 1–17.

[20] Esposito, C., Landrum, G. A., Schneider, N., Stiefl, N., dan Riniker, S. (2021) Ghost: adjusting the decision threshold to handle imbalanced data in machine learning. *Journal of Chemical Information and Modeling*, **61**, 2623–2640.

[21] Lu Gao, Y. R., Pan Lu (2021) A deep learning approach for imbalanced crash data in predicting highway-rail grade crossings accidents. *Reliability Engineering & System Safety*, **216**, 108019.

[22] Boixaderas, I., Zivanovic, D., Moré, S., Bartolome, J., Vicente, D., Casas, M., Carpenter, P. M., Radojković, P., dan Ayguadé, E. (2020) Cost-aware prediction of uncorrected dram errors in the field. *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, Virtual Event, 9-19 November, pp. 1–15. IEEE.

[23] Deron Liang, C. F. T., Chia Chi Lu dan Shih, G. A. (2016) Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. *European Journal of Operational Research*, **252**, 561–572.

[24] Chen, C. dan Breiman, L. (2004) Using random forest to learn imbalanced data. Technical Report 666. University of California, Berkeley.