BAB 5

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan pembahasan yang telah dilakukan pada bagian-bagian sebelumnya, dapat diambil beberapa kesimpulan terkait prediksi risiko terdiagnosis penyakit deiabetes menggunakan metode *Random Forest* sebagai berikut:

1. Prediksi terdiagnosis diabetes menggunakan metode *Radom Forest* dilakukan menggunakan algoritma *Bootstrap Aggregating* dan Pohon Keputusan terhadap dua jenis skala *dataset*. Skala kecil dengan 768 observasi dan 8 variabel menggunakan *Pima Indian Diabetes Dataset*. Data skala besar yang digunakan adalah *130 US Hospital Dataset* dengan 50 variabel serta 101.766 observasi. Data skala besar terbagi lagi kedalam dua kasus yang menggunakan *Feature Selection* dan tidak.

2. Setiap *dataset* diproses menggunakan dua algoritma, tanpa *resampling* dan menggunakan *resampling*. Data skala kecil yakni PIDD menggunakan *oversampling* dalam penanganannya. Sebaliknya, *130 US Hospital Dataset* menggunakan *undersampling*.

3. Faktor-faktor atau variabel bebas yang memiliki pengaruh terhadap model dapat diketahui dengan melakukan *Variabel Importance* terhadap masing-masing model. Pada PIDD, variabel *glucose* menjadi variabel yang paling berpengaruh dengan tingkat pengaruh sebesar 22,22%. Pada model yang menggunakan *130 US Hospital Dataset*, terdapat 16 variabel yang berkaitan dengan pengobatan tidak memiliki pengaruh sama sekali. Pengaruh tertinggi terdapat pada variabel *diag_2* untuk data yang tidak menggunakan *Feature Selection*. Variabel *age* memberikan pengaruh sebesar 91,90% pada data yang telah dilakukan *Feature Selection*.

4. Metode *Random Forest* berhasil melakukan pemodelan dengan hasil yang evaluasi yang baik terhadap berbagai skala data. Model yang menggunakan data skala kecil (PIDD) memberikan tingkat akurasi sebesar 75%. Sedangkan skala besar (*130 US Hospital Dataset*) berhasil memperoleh akurasi sebesar 94%.

5. Metode *Random Forest* tidak dapat diandalkan ketika menghadapi jenis data yang memiliki banyak kelas tidak seimbang. Hal ini ditunjukkan pada nilai *Recall* pada setiap dataset yang tidak dilakukan *resampling*. Nilai *Recall* yang selalu dibawah 0,5 menunjukkan model hanya mampu mengenali kelas yang benar sebanyak 50% dari keseluruhan data yang memiliki kelas terdiagnosis diabetes.

6. Akurasi dan Presisi pada model yang menggunakan PIDD mengalami peningkatan setelah dilakukan *resampling*. Nilai *Recall* dan *F1-Score* juga menunjukkan hasil yang bagus dan mengindikasikan model dapat melakukan prediksi dengan baik terhadap kedua kelas. Hal yang berbeda terjadi pada model yang menggunakan *130 US Hospital Dataset*. Baik model tanpa *feature selection* ataupun menggunakan *feature selectiom*, keduanya memberikan penurunan performa pada data yang telah dilakukan *resampling*. Hal ini disebabkan besarnya persentase *missing value* pada data sehingga mengharuskan penggantian *missing value* dengan nilai modus data.

7. Model prediksi yang menggunakan data PIDD dengan *resampling* hanya dapat diterapkan pada data asli untuk kasus diabetes gestasional saja. Hal ini disebabkan karena PIDD hanya berisi sampel dengan jenis kelamin perempuan dan berusia lebih dari 21 tahun serta memiliki variabel *pregnancies* yang terkait dengan banyaknya kejadian kehamilan. Oleh karena itu, model ini tidak cocok diterapkan pada kasus diabetes type 2 yang memiliki penderita lebih luas.

8. Model klasifikasi *Random Forest* yang menggunakan data *130 US Hospital Dataset* tidak dapat digunakan untuk memprediksi pasien terdiagnosis diabetes. Meskipun data ini memiliki skala yang besar, namun data ini memiliki banyak variabel yang tidak memiliki pengaruh sama sekali. Selain itu, data ini juga mengandung terlalu banyak *missing value* pada variabel-variabel yang cukup penting dalam melakukan diagnosis penyakit diabetes.

5.2 Saran

Penelitian ini memiliki beberapa kekurangan yang belum teratasi dengan baik. Beberapa percobaan masih dapat dilakukan dan dikembangkan pada penelitian berikutnya. Beberapa saran yang dapat diberikan untuk penelitian selanjutnya antara lain:

1. Menggunakan *dataset* yang lebih baik dan relevan. *Dataset* yang memiliki banyak kelas yang cukup seimbang. Selain itu, diperlukan *dataset* yang memiliki *Missing Value* sedikit, terutama pada variabel-variabel yang berkaitan untuk mendiagnosis penyakit diabetes.

2. Menerapkan penggunaan metode yang lebih baik dalam penaganan *missing value* pada data.

3. Model yang dibangun pada penelitian ini memiliki kekurangan ketika terjadi pengulangan. Performa yang dihasilkan memiliki perbedaan setiap kali dilakukan pengulangan pembentukan model walaupun perbedaan tersebut tidak begitu berdampak terlalu besar. Hal ini disebabkan pada saat melakukan *bootstrap* dan membagi data latih serta data uji, terjadi pengambilan sampel secara acak. Masalah ini dapat diatasi misalnya dengan melakukan uji coba sebanyak n kali dan mengambil agregasi nilai dari keseluruhan uji coba. Masalah in juga dapat diatasi dengan menentukan metode untuk pengambilan sampel acak pada data.

4. Menerapkan algoritma pembelajaran mesin lainnya seperti *Naive Bayes*, *Logistic Regression*, *k-Nearest Neighbors* (KNN), dan *XGBoost* untuk membandingkan hasil serta performa model.

DAFTAR REFERENSI

[1] IDF Diabetes Atlas Committee (2021) *IDF Diabetes Atlas*, 10th edition. International Diabetes Federation.

[2] Alehegn, M., Joshi, R. R., dan Mulay, P. (2019) Diabetes analysis and prediction using random forest, KNN, Naïve Bayes, and J48: An ensemble approach. *International Journal of Scientific and Technology Research*, **8**, 1346–1354.

[3] Almadni, D. (2011) Comparative Analysis of Classification Models for Diagnosis Type 2 Diabetes. Thesis. King Abdul Aziz University, Saudi Arabia.

[4] Kumari, S. dan Singh, A. (2013) A data mining approach for the diagnosis of diabetes mellitus. *7th International Conference on Intelligent Systems and Control, ISCO 2013*, **120**, 373–375.

[5] Iyer, A., Jeyalatha, S., dan Sumbaly, R. (2015) Diagnosis of diabetes using classification mining technique. *International Journal of Data Mining Knowledge Management Process*, **5**, 1–14.

[6] Zhenga, T., Xie, W., Xu, L., He, X., Zhang, Y., You, M., Yang, G., dan Chen, Y. (2017) A machine learning-based framework to identify type 2 diabetes through electronic health records. *International Journal of Medical Informatics*, **97**, 120–127.

[7] Nai-Arun, N. dan Moungmai, R. (2015) Comparison of classifiers for the risk of diabetes prediction. *Procedia Computer Science*, **69**, 132–142.

[8] Liu, Q., Zhang, M., He, Y., Zhang, L., Zou, J., Yan, Y., dan Guo, Y. (2022) Predicting the risk of incident type 2 diabetes mellitus in chinese elderly using machine learning techniques. *Journal of Personalized Medicine*, **12**, 905–920.

[9] Daghistani, T. dan Alshammari, R. (2020) Comparison of statistical logistic regression and randomforest machine learning techniques in predicting diabetes. *Journal of Advances in Information Technology*, **11**, 78–83.

[10] Care, D. dan Suppl, S. S. (2021) Classification and diagnosis of diabetes: Standards of medical care in diabetes-2021. *Diabetes Care*, **44**, S15–S33.

[11] Swamynathan, M. (2017) *Mastering Machine Learning with Python in Six Steps*. Apress, Bangalore.

[12] Bruce, P., Bruce, A., dan Gedeck, P. (2021) *Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python*, 2nd edition. O'Reilly Media.

[13] James, G., Witten, D., Hastie, T., dan Tibshirani, R. (2013) *An Introduction to Statistical Learning with Application in R*. Springer.

[14] Breiman, L. (2001) Random forests. *Machine learning*, **45**, 5–32.

[15] Breiman, L. (1996) Bagging predictors. *Machine learning*, **24**, 123–140.

[16] Hull, J. C. (2019) *Machine Learning in Business : An Introduction to the World of Data Science*, 2nd edition. University of Toronto, Ontario.

[17] UCI Machine Learning Repository (2016) Pima Indians Diabetes Dataset. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database. 30 Mei 2023.

[18] Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., dan Johannes, R. S. (1988) Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *Proceedings - Annual Symposium on Computer Applications in Medical Care*, November, pp. 261–265. American Medical Informatics Association.

[19] Clore, C. K. D. J., John dan Strack, B. (2014) Diabetes 130-US hospitals for years 1999-2008. https://archive.ics.uci.edu/static/public/296/diabetes+130-us+hospitals+for+years+1999-2008.zip. 12 Juli 2023.

[20] Strack, B., Deshazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., dan Clore, J. N. (2014) Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. *BioMed Research International*, **2014**.