EARLY DETECTION OF DIABETES MELLITUS USING RANDOM FOREST ALGORITHM

Andri Triyono*; Rahmawan Bagus Trianto; Dhika Malita Puspita Arum

ABSTRACT

Diabetes mellitus is a deadly disease. Patients with this disease often do not realize that they are improving their diabetes mellitus. It is necessary to do early prevention in order to reduce the sudden death rate of people with diabetes mellitus. In addition, during the COVID-19 pandemic, which increases the risk of death for people with comorbid diabetes mellitus. A system model for the prediction of diabetes mellitus is needed for early diagnosis of this disease. By using machine learning techniques using the Random Forest algorithm and Information Gain can be used to predict diabetes mellitus. This model has a fairly high level of accuracy, which is 98.27%, precision is 97.69% and recall is 98%.

Keywords: *diabetes mellitus; random forest; information gain; machine learning;*

Correspondence:
Andri Triyono
Universitas An Nuur, Email; andritriyono1@gmail.com

PENDAHULUAN

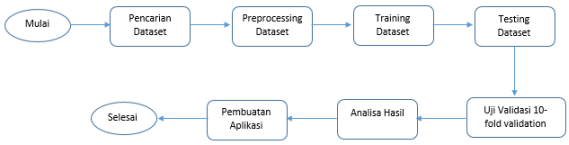
Diabetes mellitus adalah suatu penyakit yang memiliki banyak tanda, salah satunya tingginya kadar glukosa di dalam darah yang melebihi normal. Federasi Diabetes Internasional mengatakan jika penderita diabetes mellitus di dunia sebesar 1.9% dari total populasi manusia dan juga menempati urutan ke tujuh sebagai penyebab kematian tertinggi di dunia (Hestiana, 2017). Di masa pandemi covid-19 tingkat resiko para penderita diabetes mellitus untuk tertular virus ini semakin tinggi (Susilo et al., 2020). Pneumonia atau infeksi yang terjadi pada paru-paru, di mana covid-19 juga menyerang paru-paru, terbukti meningkatkan resiko kematian pada penderita diabetes mellitus (Simanjuntak, Simamora, & Sinaga, 2020). Di sisi lain, kebanyakan penderita diabetes mellitus mengesampingkan pengecekan kadar gula dalam darahnya dan lebih fokus pada pencegahan covid-19 saja (Simanjuntak et al., 2020). Padahal penyakit ini sering disebut sebagai penyakit yang silent killer, hal ini karena penderitanya tidak sadar kalau sedang menderita diabetes mellitus, dan setelah dilakukan pengecekan ternyata sudah pada tingkat komplikasi (Hestiana, 2017). Hal ini sangat berbahaya, karena jika penderita sudah mengalami komplikasi, maka harapan hidup akan menurun drastis dibandingkan dengan sebelumnya (Tantona, 2019).

Untuk mengurangi resiko terkena diabetes mellitus, terlebih di masa pandemi covid-19 dapat dilakukan dengan cara pengecekan secara dini penyakit ini. Dengan menggunakan data lampau tanda-tanda yang dialami oleh para penderita diabetes mellitus, teknik klasifikasi data mining bisa digunakan untuk mengelompokkan data baru pada seseorang. Dengan kata lain, dengan memanfaatkan data tanda-tanda penyakit diabetes melitus yang sudah ada, dapat digunakan untuk mendeteksi secara dini seseorang terkena diabetes mellitus atau tidak.

Penelitian ini menggunakan metode Random Forest dengan Information Gain sebagai criterion karena memiliki keunggulan menghasilkan akurasi yang tinggi, memiliki kesalahan yang rendah serta dapat menangani dataset training yang besar dan missing value pada dataset (Primajaya & Sari, 2018). Dengan adanya aplikasi deteksi dini diabetes mellitus maka informasi yang diterima seseorang dapat menjadikan mereka lebih memperhatikan kesehatannya, baik bagi yang sudah terkena maupun yang tidak terkena diabetes mellitus.

METODE PENELITIAN

Penelitian ini termasuk pada penelitian eksperimental dengan menggunakan dataset sekunder dan terbuka yang dapat diakses melalui laman https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset. Dataset ini berkaitan dengan pasien diabetes di Rumah Sakit Diabetes Sylhet di Sylhet Bangladesh dan divalidasi oleh dokter. Dataset tersebut didapatkan melalui kuesioner yang diberikan kepada pasien. Dataset ini memiliki 520 baris dengan 17 atribut, di mana 16 atribut reguler dan 1 atribut lainnya merupakan label.



Gambar 1. Metode Penelitian

Penelitian ini termasuk pada penelitian eksperimental dengan menggunakan dataset sekunder dan terbuka yang dapat diakses melalui laman https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset. Dataset ini berkaitan dengan pasien diabetes di Rumah Sakit Diabetes Sylhet di Sylhet Bangladesh dan divalidasi oleh dokter. Dataset tersebut didapatkan melalui kuesioner yang diberikan kepada pasien. Dataset ini memiliki 520 baris dengan 17 atribut, di mana 16 atribut reguler dan 1 atribut lainnya merupakan label.

Penelitian dimulai dengan pencarian dan pengambilan dataset melalui laman yang telah disebutkan sebelumnya. Selanjutnya dilakukan tahap preprocessing, yaitu tahap untuk mempersiapkan dataset agar siap untuk diolah menggunakan algoritma data mining. Tahap ini dilakukan proses penghilangan data yang missing value. Proses selanjutnya adalah training dataset, yaitu proses melatih mesin untuk melakukan pembelajaran berdasarkan data latih. Setelah dilakukan training dataset, selanjutnya dilakukan proses testing menggunakan data baru untuk mengetahui performa dari hasil klasifikasinya. Untuk mengetahui performa model yang ada, pada penelitian ini menggunakan k-fold validation.

Adapun pengujian pada penelitian ini menggunakan k-fold validation dengan nilai k=10, yang berarti data akan dibagi menjadi 10 bagian dan setiap 9 bagian akan dipakai sebagai data learning atau pelatihan, dan 1 bagian dipakai sebagai data testing atau pengujian (Saputri, Mahendra, & Adriani, 2019). Proses ini dilakukan sebanyak 10 kali (Wahono, Herman, & Ahmad, 2014). Tahap ini dapat dilihat pada tabel 1.

Tabel 1. Pembagian 10-Fold Validation

Validasi ke-n	Pembagian dataset									
1	■									
2		■								
3			■							
4				■						
5					■					
6						■				
7							■			
8								■		
9									■	
10										■

Tanda warna hitam menandakan dataset yang dipakai untuk data pelatihan. Sedangkan pada bagian yang lain menandakan dataset yang dipakai untuk data pelatihan. Untuk mengetahui performa klasifikasi, diperlukan sebuah pengukuran, yaitu akurasi, presisi dan recall (Deolika, Kusrini, & Luthfi, 2019). Pengukuran ini sering dikenal dengan sebutan confusion matrix. Confusion matrix biasa dipakai pada proses pembelajaran terarah. Pada confusion matrix, data pada kolom mewakili data yang diharapkan, sedangkan data pada baris mewakili data yang diprediksi (Dhande & Patnaik, 2014). Berbeda dengan klasifikasi dengan 2 buah kelas, untuk menghitung akurasi, presisi dan recall pada klasifikasi dengan kelas lebih dari 2 menggunakan rata-rata (Sokolova & Lapalme, 2009). Untuk

menghitung akurasi, presisi dan recall dapat dilihat pada persamaan 1, 2 dan 3.

$$Akurasi = \frac{\sum_{i=1}^{c} \frac{TPi + TNi}{TPi + TNi + FPi + FNi}}{c} * 100\% \quad (1)$$

$$Presisi = \frac{\sum_{i=1}^{c} TPi}{\sum_{i=1}^{c} (FPi + TPi)} * 100\% \quad (2)$$

$$Recall = \frac{\sum_{i=1}^{c} TPi}{\sum_{i=1}^{c} (TPi + FNi)} * 100\% \quad (3)$$

Dimana:

- c = jumlah kelas
- TPi = jumlah *True Positive* pada kelas ke i, jumlah data positif yang diklasifikasikan benar oleh sistem pada kelas ke i
- TNi = jumlah *True Negative* pada kelas ke i, jumlah data negatif yang diklasifikasikan benar oleh sistem pada kelas ke i
- FNi = jumlah *False Negative* pada kelas ke i, jumlah data negatif yang diklasifikasikan salah oleh sistem pada kelas ke i
- FPi = jumlah *False Positive* pada kelas ke i, jumlah data positif yang diklasifikasikan salah oleh sistem pada kelas ke i

HASIL DAN PEMBAHASAN

Proses pencarian dataset yang bersifat open access telah dilakukan pada laman yang telah disebutkan sebelumnya. Tahap selanjutnya dilakukan proses preprocessing, yaitu untuk menghilangkan dataset yang tidak lengkap isiannya, atau sering disebut juga dengan missing value. Langkah selanjutnya dilakukan proses training dan testing dataset menggunakan algoritma Random Forest dengan criterion sebagai criterion yang diuji performanya menggunakan 10-fold validation. Adapun pengukuran performa menggunakan parameter akurasi, presisi dan recall. Adapun nilai akurasi, presisi dan recall dapat dilihat pada tabel 2.

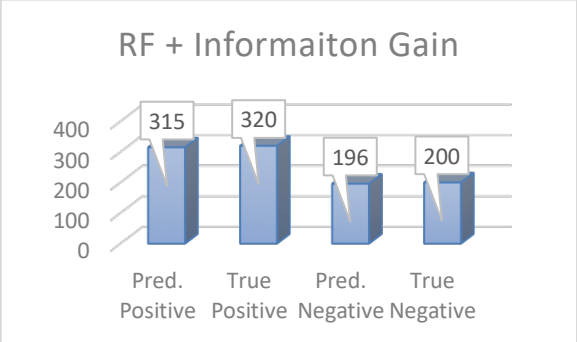
Tabel 1; Hasil Pengujian Prediksi

Model	Akurasi	Presisi	Recall
Random Forest + Information Gain	**98.27%**	**97.69%**	**98.00%**
Random Forest + Gain Ration	98,08%	97.69%	97.50%
Random Forest + Gini Index	98,08%	97.26%	98.00%
Random Forest + Accuracy	97,69%	96.31%	98.00%

Tabel 2 merupakan hasil uji coba pada penelitian ini. Dapat dilihat bahwa akurasi, presisi dan *recall* tertinggi terdapat pada model yang menggunakan *Random Forest* dan *Information Gain* sebagai criterion nya. Masing-masing nilai akurasi, presisi dan *recall* dari model ini adalah sebesar 98.27%, 97.69% dan 98.00%.

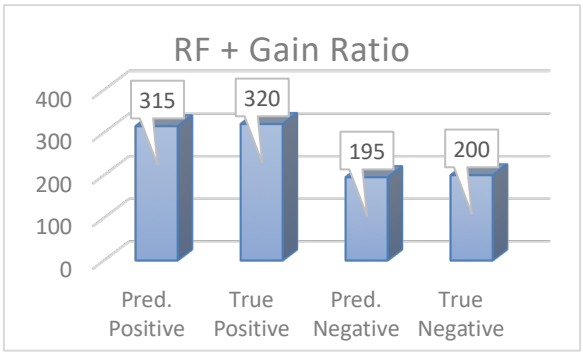
Berikut ini contoh salah satu *Tree* atau pohon yang terbentuk dari algoritma *Random Forest* dan juga *Information Gain*.

Tabel 2 menunjukkan hasil pengujian prediksi atau deteksi dini diabetes mellitus menggunakan algoritma *Random Forest*. Hasil terbaik didapat pada pemakaian *Information Gain* pada metode *Random Forest*, yaitu dengan nilai akurasi sebesar 98.27%, nilai presisi sebesar 97.69% dan nilai *recall* sebesar 98.00%. Dapat dilihat pula bahwa penggunaan criterion lain memiliki performa yang tidak begitu jauh, bahkan recall dan presisi pada criterion *Information Gain* memiliki nilai yang sama dengan criterion yang lain. Namun demikian, *Information Gain* unggul pada nilai akurasinya bila dibandingkan dengan criterion lain.



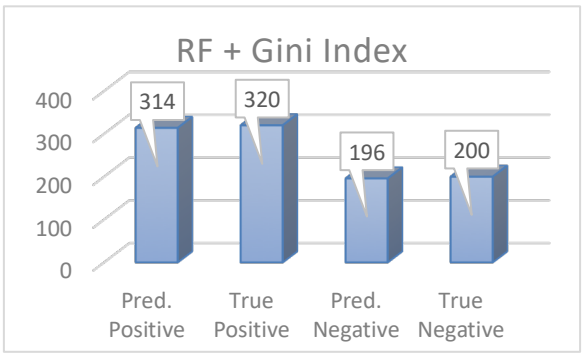
Gambar 2; Grafik sebaran prediksi *Random Forest + Information Gain*

Pada gambar 2 data sebaran penderita diabetes mellitus dan non penderita diabetes mellitus menggunakan *Random Forest + Information Gain*. Sebanyak 315 prediksi positif yang benar dari data yang benar sebanyak 320, serta 196 prediksi negative benar dari data yang benar negative diabetes mellitus sebanyak 201.



Gambar 3; Grafik sebaran prediksi *Random Forest + Gain Ratio*

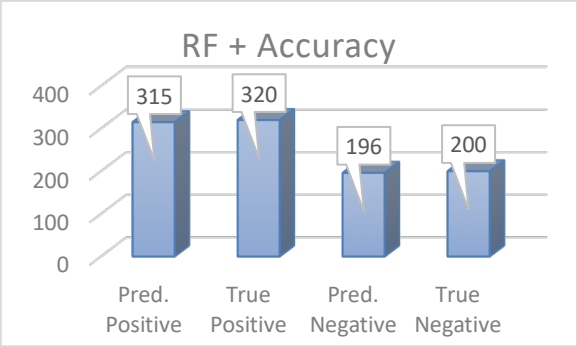
Pada gambar 2 data sebaran penderita diabetes mellitus dan non penderita diabetes mellitus menggunakan *Random Forest + Gain Ratio*. Sebanyak 315 prediksi positif yang benar dari data yang benar sebanyak 320, serta 195 prediksi negative benar dari data yang benar negative diabetes mellitus sebanyak 200.



Gambar 4; Grafik sebaran prediksi *Random Forest + Gini Index*

Gambar 4 data sebaran penderita diabetes mellitus dan non penderita

diabetes mellitus menggunakan *Random Forest + Gini Index*. Sebanyak 314 prediksi positif yang benar dari data yang benar sebanyak 320, serta 196 prediksi negative benar dari data yang benar negative diabetes mellitus sebanyak 200.



Gambar 5; Grafik sebaran prediksi *Random Forest + Accuracy*

Gambar 5 data sebaran penderita diabetes mellitus dan non penderita diabetes mellitus menggunakan Random Forest + Accuracy. Sebanyak 315 prediksi positif yang benar dari data yang benar sebanyak 320, serta 196 prediksi negative benar dari data yang benar negative diabetes mellitus sebanyak 200.

SIMPULAN DAN SARAN

Prediksi dini diabetes mellitus dapat dilakukan dengan menggunakan metode *Random Forest* dengan berbagai criterion, namun hasil terbaik dengan menggunakan criterion *Information Gain*. Diawali dengan proses *preprocessing* untuk menghilangkan data yang tidak lengkap dengan teknik *missing value*. Hasil terbaik *Random Forest* dengan menggunakan *Information Gain* didapat akurasi sebesar 98.27%, presisi sebesar 97.69% dan recall sebesar 98%. Dengan hasil performa model yang telah dibuat cukup bagus dipakai untuk langkah awal dalam memprediksi masyarakat apakah berpotensi mengidap diabetes mellitus atau tidak. Dengan adanya langkah awal ini diharapkan dapat menurunkan resiko yang terjadi akibat dampak dari diabetes mellitus itu sendiri.

Penelitian ini menggunakan data publik yang disediakan secara umum. Penelitian ini juga belum dilakukan di kehidupan nyata, sehingga perlu ada penelitian lebih lanjut untuk menguji di objek penelitian seperti rumah sakit. Selain itu, dengan karakteristik metode *Random Forest* yang dapat berubah-ubah performanya, seperti tingkat akurasi, maka perlu dicari metode untuk meminimalkan performa yang tidak maksimal, atau dengan menggunakan metode machine learning yang lainnya.

DAFTAR PUSTAKA

Deolika, A., Kusrini, K., & Luthfi, E. T. (2019). Analisis Pembobotan Kata Pada Klasifikasi Text Mining. *Jurnal Teknologi Informasi*, *3*(2), 179. https://doi.org/10.36294/jurti.v3i2.1077

Dhande, L. L., & Patnaik, P. G. K. (2014). Analyzing Sentiment of Movie

Review Data using Naive Bayes Neural Classifier. *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, *3*(4), 313–320. Diambil dari www.ijettcs.org

Hestiana, D. W. (2017). Faktor-Faktor Yang Berhubungan Dengan Kepatuhan Dalam Pengelolaan Diet Pada Pasien Rawat Jalan Diabetes Mellitus Tipe 2 Di Kota Semarang. *Journal of Health Education*, *2*(2), 138–145. https://doi.org/10.15294/jhe.v2i2.14448

Primajaya, A., & Sari, B. N. (2018). Random Forest Algorithm for Prediction of Precipitation. *Indonesian Journal of Artificial Intelligence and Data Mining*, *1*(1), 27. https://doi.org/10.24014/ijaidm.v1i1.4903

Saputri, M. S., Mahendra, R., & Adriani, M. (2019). Emotion Classification on Indonesian Twitter Dataset. *Proceedings of the 2018 International Conference on Asian Language Processing, IALP 2018*, 90–95. IEEE. https://doi.org/10.1109/IALP.2018.8629262

Simanjuntak, G. V., Simamora, M., & Sinaga, J. (2020). Optimalisasi Kesehatan Penyandang Diabetes Melitus Tipe II Saat Pandemi Covid-19. *Journal of Community Engagement in Health*, *3*(2), 171–175. https://doi.org/10.30994/jceh.v3i2.59

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Susilo, A., Rumende, C. M., Pitoyo, C. W., Santoso, W. D., Yulianti, M., Herikurniawan, H., … Yunihastuti, E. (2020). Coronavirus Disease 2019: Tinjauan Literatur Terkini. *Jurnal Penyakit Dalam Indonesia*, *7*(1), 45. https://doi.org/10.7454/jpdi.v7i1.415

Tantona, M. D. (2019). Jurnal Penelitian Perawat Profesional. *Jurnal Penelitian Perawat Profesional*, *1*(November), 89–94. Diambil dari http://jurnal.globalhealthsciencegroup.com/index.php/JPPP/article/download/83/65

Wahono, R. S., Herman, N. S., & Ahmad, S. (2014). A comparison framework of classification models for software defect prediction. *Advanced Science Letters*, *20*(10–12), 1945–1950. https://doi.org/10.1166/asl.2014.5640