Submitted: 2022-11-25 Accepted: 2022-12-30 Published: 2023-06-06
Sebatik Vol. 27 No. 1 Juni 2023
ISSN: 1410-3737(p) 2621-069X(e)

DOI:10.46984/sebatik.v27i1.2157

COMPARATIVE ANALYSIS OF ACCURACY OF RANDOM FOREST AND GRADIENT BOOSTING CLASSIFIER ALGORITHM FOR DIABETES CLASSIFICATION

Sahat Pandapotan Nainggolan[1)] Ardiles Sinaga[2)]

[1,2]Software Engineering Technology, Faculty of Vocational, Del Institute of Technology
[1,2]Sitoluama, Kec. Balige, Toba, Sumatera Utara, 22381
E-mail: sahat.nainggolan@del.ac.id[1)], ardiles.sinaga@del.ac.id[2)]

ABSTRACT

Diabetes is a disease characterized by high blood sugar (glucose) levels. If blood sugar is not controlled properly, it can cause various critical diseases, one of which is diabetes. The purpose of this study was to determine the results of a comparison of the accuracy values of the *Random Forest* Algorithm and the *Gradient Boosting Classifier* Algorithm in the classification of diabetes which will be tested for accuracy, *Precision*, *Recall*, and *F1 score* performance. The method used in this study was descriptive and the data source used the Pima Indians Diabetes Dataset from Kaggle. Based on data analysis using a ratio of 80:20, the *Random Forest* Algorithm has an accuracy of 79% obtained from the results of the *confusion matrix*. From the *confusion matrix* results, the results obtained were AUC 0.835, *Recall* 78%, and *Precision* 90%. Based on the results of *Recall* and *Precision*, an *F1 score* of 83% was obtained. Whereas the Boosting Classifier Algorithm has an accuracy result obtained from the results of the *confusion matrix* which is 81%. From the *confusion matrix* results, the AUC results were 0.877, *Recall* 83%, and *Precision* 67%. Based on the results of *Recall* and *Precision*, an *F1 score* of 74% was obtained. In this study, the accuracy evaluation results obtained were through the results of the *Confusion matrix* and the AUC value. These results indicate that the *Gradient Boosting Classifier* Algorithm has a more excellent accuracy evaluation result compared to the *Random Forest* Algorithm.

Keywords: *Classification, Random Forest, Gradient boosting, Accuracy, Precision, Recall and, F1-score.*

1. PENDAHULUAN

Diabetes adalah penyakit kronis yang ditandai dengan kadar glukosa yang tinggi atau di atas normal. Diabetes adalah penyakit metabolik di mana kadar glukosa tinggi. Glukosa sangat penting bagi kesehatan manusia karena merupakan sumber energi penting bagi sel dan jaringan. Jika tidak ditangani dengan baik, dapat menyebabkan penyakit serius seperti penyakit jantung, obesitas, penyakit ginjal, stroke, penyakit mata, dan diabetes (Argina, 2020).

Diabetes adalah penyakit manusia yang paling umum. Setiap tahun, jumlah kematian akibat penyakit ini meningkat secara signifikan. Menurut Organisasi Kesehatan Dunia (WHO), hampir 350 juta orang menderita diabetes. WHO memprediksi bahwa diabetes akan menjadi salah satu penyebab utama kematian pada tahun 2030 (Hairani dkk., 2018). Berdasarkan data *International Diabetes Federation* (IDF), Indonesia berada dalam status waspada diabetes dan Indonesia sendiri berada di urutan ke-7 dari 10 negara dengan jumlah penderita diabetes terbanyak di dunia (Kementerian Kesehatan RI., 2020).

Pada penelitian-penelitian sebelumnya mengenai diabetes sudah banyak dilakukan penelitian seperti *K-Nearest Neighbor* (KNN) untuk klasifikasi diabetes (Indrayani, Sugianti and Al Karomi, 2019), diagnosis diabetes menggunakan *Bayesian Regularization Neural Network* (Rahman *dkk.*, 2017), klasifikasi diabetes menggunakan algoritma *Backpropagation* (Brian, 2017), perbandingan algoritma C4.5, KNN dan *Naïve Bayes* dalam klasifikasi penanggung jawab *BSI Entrepreneur Center* (Hasan, Hikmah and Utami, 2018) dan penelitian yang membandingkan beberapa algoritma penambangan data menggunakan data pasien diabetes yang tersedia dari *Kaggle* dalam format .csv dan diterbitkan oleh *Indian Pima*. Data berisi 768 data dan 9 kolom.

Untuk mengidentifikasi diabetes dapat dibuat klasifikasi diabetes yang kemudian dapat digunakan untuk mendukung penatalaksanaan diabetes. Salah satu kemungkinan untuk klasifikasi adalah klasifikasi dalam *machine learning*, dengan bantuan *machine learning* klasifikasi dapat dilakukan dengan *data mining*. Pesatnya perkembangan teknologi pengumpulan dan penyimpanan data di berbagai bidang telah menyebabkan *database* yang terlalu besar. Salah satu bidang pengolahan data adalah *data mining* (Larose, 2004).

Penelitian tentang *data mining* telah banyak dilakukan seperti pemetaan penyebaran covid-19 dengan menggunakan algoritma *K-means clustering* (Gayatri and Hendry, 2021), identifikasi jenis ikan menggunakan model *hybird deep learning* (Azis, 2020) , analisa perbandingan metode *naïve bayes* dan KNN terhadap klasifikasi data (Indriani, 2020), peramalan jumlah produksi air dengan algoritma *backpropagation* (Yanto, Mulyani and Mayola, 2019), pola penentuan status peminjaman dengan algoritma *perceptron* (Arlis *dkk.*, 2018) dan masih banyak lagi penelitian yang berkaitan.

Salah satu teknik dalam *data mining* adalah *supervised learning*. Teknik *supervised learning* sendiri membutuhkan *dataset training* yang nantinya dengan menentukan nilai input akan menghasilkan *output* sesuai target (Sullivan, 2012).

Ada banyak algoritma yang termasuk dalam teknik *supervised learning* seperti Algoritma *Random Forest* dan *Gradient Boosting Classifier*.

Penulis menyadari bahwa sudah banyak penelitian yang menggunakan Algoritma *Random Forest* dan algoritma lain seperti Algoritma *Artificial Neural Network*. Pada tahun 2020, Ali Murtadho, dan Dwi Harini Sulistyawati membuat jurnal berjudul "*Machine Learning* Untuk Perbandingan Tingkat Akurasi Metode *Supervised Learning* Prediksi Diabetes"(Murtadho and Sulistyawati, 2020), dalam tulisannya Ali Murtadho, dan Dwi Harini Sulistyawati menggunakan 5 algoritma yang berbeda untuk menentukan klasifikasi dataset *Pima Indians Diabetes*, dan hasilnya menunjukkan bahwa Algoritma *Gradient Boosting Classifier* menghasilkan akurasi sebesar 81,3%, Algoritma *Decision Tree* menghasilkan akurasi sebesar 72%, Algoritma *Random Forest* menghasilkan akurasi sebesar 72%, Algoritma *Logistic Regression* menghasilkan akurasi sebesar 70%, dan Algoritma K-NN menghasilkan akurasi sebesar 64%. Pada tahun 2021, Shamriz Nahzat, dan Mete Yağanoğlu membuat jurnal berjudul "Prediksi Diabetes Menggunakan Algoritma Klasifikasi Pembelajaran Mesin" (NAHZAT and YAĞANOĞLU, 2021), dalam tulisannya Shamriz Nahzat, dan Mete Yağanoğlu Menggunakan 5 Algoritma berbeda yaitu KNN, *Random Forest*, *SVM*, *Artificial Neural Network* , dan *Decision Tree* untuk menentukan klasifikasi dataset *Pima Indians Diabetes*. Pada penelitian ini Algoritma *Random Forest* menghasilkan nilai akurasi sebesar 88,31% dan Algoritma *Artificial Neural Network* menghasilkan nilai akurasi sebesar 86%. Dalam studi ini, Shamriz Nahzat dan Mete Yağanoğlu menambahkan beberapa fitur ke dataset diabetes *Pima Indians*. Sementara itu, penulis hanya fokus pada fitur original dari dataset *Pima Indians Diabetes* tanpa ada tambahan fitur di dalamnya.

Random Forest (RF) banyak digunakan pada klasifikasi dan regresi, dari penelitian diperoleh hasil bahwa model RF lebih akurat dari akurasi biomass gandum dibandingkan dengan algoritma lain seperti *Support Vector Regression* (SVR) dan *Artificial Neural Network* (ANN) dan kekokohan sama baiknya dengan SVR , namun lebih baik dari ANN. Klasifikasi RF digunakan pada penyakit getah bening, dengan memadukan fitur seleksi algoritma genetik, didapatkan akurasi sebesar 92,2 % (Wang *et al.*, 2016), (Azar *et al.*, 2014). Penggunaan RF dalam mengidentifikasi 23 panjang gelombang yang berkaitan dengan struktur tumbuhan dan konten air (Vitrack-Tamam *et al.*, 2020). RF juga digunakan untuk memprediksi mutasi kanker dari dataset *genomic* (Agajanian, Oluyemi and Verkhivker, 2019). Akurasi RF sangat akurat sebesar 96,57% dan AUC besar dari 98% saat mengenali host tropis dari protein influenza individu (Eng, Tong and Tan, 2014).

Gradient Boosting Machine (GBM) memiliki beberapa keunggulan dari metode *machine learning* lain. Penelitian didapatkan bahwa GBM meningkatkan akurasi prediksi R kuadrat dan RMSE lebih dari 80 persen dibandingkan dengan model terbaik industri yakni algoritma *Random Forest* dan regresi linier (Natekin and Knoll, 2013). GBM juga digunakan pada prediksi waktu pergi dan kedatangan, dimana GBM memiliki kelebihan untuk prediksi waktu keberangkatan yang bebas memilih (Zhang and Haghani, 2015). Beberapa keluarga dari algoritma *gradient boosting* diuji pada kecepatan dan akurasi. Uji dilakukan pada *CatBoost*, *eXtreme Gradient Boosting* (XGBoost), *Random Forests*, *LightGBM*, dan *gradient boosting*. Hasil komparasi, diindikasikan bahwa *CatBoost* merupakan hasil terbaik untuk akurasi dan AUC, walaupun perbedaannya kecil. Light *Gradient Boosting Machine* (LightGBM) tercepat dari semua metode. *XGBoost* menempati posisi kedua untuk akurasi dan kecepatan *training* (Bentéjac, Csörgő and Martínez-Muñoz, 2019). *Light* GBM mempercepat waktu proses 20 kali lipat dari fase pelatihan *Gradient Boosting Decision Tree* (GBDT) konvensional dengan akurasi yang sama (Ke *et al.*, 2017).

Berdasarkan penelitian sebelumnya terdapat berbagai nilai akurasi dari algoritma yang diuji oleh penulis, pada penelitian ini penulis akan melakukan perbandingan algoritma yang berfokus pada penggunaan algoritma *Random Forest* dan *Gradient Boosting Classifier* dengan melakukan inputan mean dan median. Algoritma yang digunakan penulis termasuk dalam teknik *supervised learning*, dengan tujuan untuk mengetahui tingkat akurasi dalam memprediksi masalah diabetes.

2. RUANG LINGKUP

Rumusan masalah dalam penelitian ini adalah melakukan *preprocessing* data untuk penderita diabetes, menerapkan algoritma *Random Forest* dan *Gradient Boosting Classifier* pada penyakit diabetes dan menguji sistem klasifikasi penyakit diabetes.

Dalam penelitian ini, ada beberapa batasan masalah yang diangkat sebagai parameter, antara lain sebagai berikut: data yang digunakan untuk penelitian ini diperoleh dari data penderita diabetes yang dapat diakses melalui *Kaggle.com*. Semua data berjumlah 768 dan terdiri atas 9 kolom. Parameter yang digunakan terdiri atas 9 variabel, diantaranya 8 variabel bebas (X) yaitu *Pregnancies*, *Glucose*, *Blood Pressure*, *Skin Thickness*, *Insulin*, *BMI*, *Diabetes Pedigree Function*, *Age* dan 1 variabel bebas (Y) yaitu *Outcome*. Yang terakhir adalah perbandingan nilai akurasi antara Algoritma *Random Forest* dan *Gradient Boosting Classifier*.

3. BAHAN DAN METODE

Metode yang digunakan dalam penelitian ini adalah deskriptif dan diolah menggunakan aplikasi Python. Dataset diambil dari data pasien diabetes yang tersedia di *Kaggle* dalam format .csv yang diterbitkan oleh Pima Indians sebanyak 768 data dan 9 field/kolom.

Adapun algoritma yang digunakan algoritma *Random Forest* dan *Gradient Boosting Classifier* pada *RapidMiner*, dan melakukan pengujian akurasi, *Precision*, dan *Recall* dengan melihat hasil klasifikasi menggunakan *Confusion matrix*.

4. PEMBAHASAN

Penelitian ini dilakukan dengan cara mencari dataset secara *online* melalui website *Kaggle.com*. Dataset yang di gunakan adalah dataset *Pima Indians diabetes* memiliki total 768 *row data* dan 9 *attribut*, *attribut* tersebut diantaranya adalah *Pregnancies*, *Glucose*,*Blood Pressure*, *SkinThickness*, *Insulin*, *BMI*, *Diabetes Pedigree Function*, *Age*, *Outcome*, isi atribut tersebut dapat dilihat pada tabel 1.

Tabel 1. *Pima Indians Diabetes Dataset*

No	*Pregnancies*	*Glucose*	*Blood Pressure*	*Skin Thickness*	*Insulin*	*BMI*	*Diabetes Pedigree Functions*	*Age*	*Outcome*
0	6	148	72	35	0	33,6	0,63	50	1
1	1	85	66	29	0	26,6	0,35	31	0
2	8	183	64	0	0	23,3	0,63	32	1
3	1	89	66	23	94	28,1	0,62	21	0
4	0	137	40	35	168	43,1	2,29	33	1
5	5	116	74	0	0	25,6	0,2	30	0
6	3	78	50	32	88	31	0,25	26	1
7	10	115	0	0	0	35,3	0,13	29	0
8	2	197	70	45	543	30,5	0,16	53	1
9	8	125	96	0	0	0	0,23	54	1

Dari tabel 1 dapat diketahui bahwa, *Dataset Pima Indians Diabetes* memiliki 9 Variabel yang terdiri 8 variabel independent dan 1 variabel dependent. Tabel 2 dan tabel 3 adalah daftar dari masing-masing Variabel Independen (X) dan Variabel Dependen (Y)

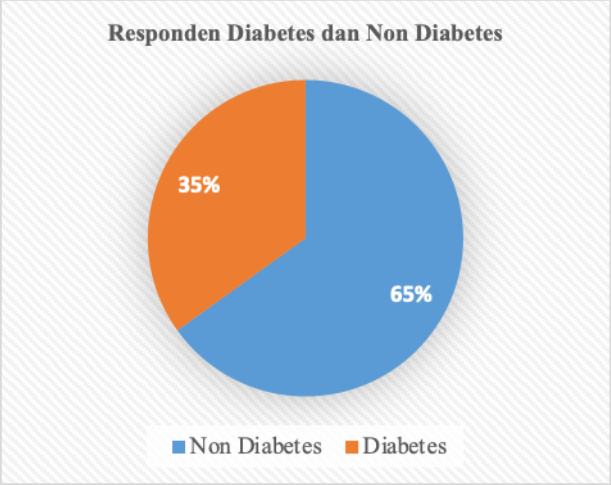
Tabel 2. Variabel Independent (X)

Kolom Dataset	Tipe Data	Deskripsi
Pregnancies	*Int64*	Variabel ini menunjukkan tentang seberapa banyak Wanita tersebut hamil semasa hidup
Glucose	*Int64*	Variabel ini menunjukkan tentang konsentrasi glukosa plasma pada 2 jam dalam tes toleransi glukosa
Blood Pressure	*Int64*	Variabel ini menunjukkan tentang tekanan darah yang dimiliki oleh seseorang
Skin Thickness	*Int64*	Variabel ini menunjukkan tentang perkiraan lemak tubuh yang dimiliki seseorang yang diukur pada lengan kanan setengah antara proses olecranon dari siku dan proses akromial skapula
Insulin	*Int64*	Variabel ini menunjukkan tentang tingkat Insulin seseorang
BMI	*Float64*	Variabel ini menunjukkan tentang indeks masa tubuh seseorang
Diabetes Pedigree Function	*Float64*	Variabel ini menunjukkan tentang indikator Riwayat diabetes di dalam keluarga
Age	*Int64*	Variabel ini menunjukkan tentang umur seseorang

Tabel 3. Variabel Dependent (Y)

Kolom Dataset	Tipe Data	Deskripsi
Outcome	*Int64*	Variabel ini menunjukkan hasil apakah seseorang mengidap diabetes (ditulis dengan angka 1) atau tidak mengidap diabetes (ditulis dengan angka 0)

Penulis selanjutnya melakukan tahapan *preprocessing* data agar data dapat digunakan secara tepat dalam proses klasifikasi. Di tahap ini ditemukan bahwa 35% dari responden mempunyai gejala diabetes seperti yang ditampilkan di gambar 1.



Gambar 1. Deskripsi Pie chart: 35% dari responden mempunyai diabetes

Tahapan yang dilakukan dimulai dari data *cleansing* dengan untuk memeriksa apakah ada data *null* (*missing value*) dari data yang digunakan. Setelah di temukannya nilai *null* (*missing value*) langkah selanjutnya mencari nilai median dari setiap kolom untuk diisi ke dalam data yang bernilai *null*. Selanjutnya memisahkan data variabel X dan Y serta melakukan *splitting data*. Dan tahap terakhir yang dilakukan adalah *Feature Scaling* untuk menyetarakan skala dari nilai data yang digunakan. Setelah data dilakukan *preprocessing* dataset bisa di implementasi-kan dengan Algoritma *Random Forest* dan Algoritma *Gradient Boosting Classifier*.

1. *Random Forest* adalah metode klasifikasi yang melibatkan kumpulan pohon keputusan yang nantinya akan dijadikan vote untuk mendapatkan sebuah output (Alita and Isnain, 2020). *Random Forest* adalah sebuah classifier model yang berbentuk seperti pohon keputusan untuk mendapatkan hasil prediksi yang lebih baik (Kumari, Kumar and Mittal, 2021).
2. *Gradient Boosting Classifier* adalah teknik *supervised learning* berbasis decision tree. Algoritma dimulai dari menghasilkan pohon klasifikasi awal dan terus menyesuaikan pohon baru melalui minimalisasi fungsi kerugian (Natekin and Knoll, 2013). *Gradient Boosting Classifier* memerlukan konfigurasi hyperparameter pada tahap awal, kombinasi hyperparameter yang optimal mempengaruhi hasil prediksi gradient boosting. *Hyperopt* adalah library *Python* yang menunjang tuning hyperparameter dengan konsep optimasi Bayesian.

Pada Algoritma *Random Forest* dan *Gradient Boosting Classifier*, penulis akan membagi *data training* dan *data testing* dari keseluruhan data yang memiliki total 768 data. Data tersebut akan di uji menggunakan perbandingan rasio 80:20. Dengan dilakukannya pengetesan pada kedua Algoritma, diperoleh hasil yiatu akurasi Algoritma *Random Forest* dan Algoritma *Gradient Boosting Classifier* dengan rasio 80:20 menghasilkan akurasi yang hampir sama bahkan hanya sedikit perbedaan yaitu 80% untuk akurasi Algoritma *Random Forest* dan 79% untuk Algoritma Gradient Boosting.

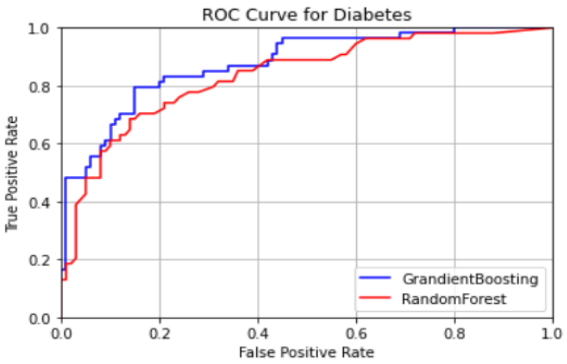
Tahap selanjutnya adalah mengukur ulang akurasi dengan menggunakan data features dengan Target (*Outcome*) . Dengan dilakukan nya inputan berdasarkan nilai korelasi yang lebih tinggi, diperoleh hasil akurasi dengan rasio 80:20 untuk Algoritma *Random Forest* sebesar 78% dan Algoritma *Gradient Boosting Classifier* menghasilkan akurasi sebesar 81%. Tahap selanjutnya adalah melakukan evaluasi prediksi menggunakan *Confusion matrix*. *Confusion matrix* adalah suatu metode yang sering sekali digunakan untuk melakukan perhitungan akurasi pada data *mining* (Rahman *dkk.*, 2017). Hasil evaluasi setelah di terapkan ke dalam *Confusion matrix* bisa di lihat pada gambar 2.

```
# Confusion Matrix (threshold = 0.25)
confusion_matrix(y_test, y_pred_class[:,1])
```

```
array([[78, 22],
       [ 9, 45]])
```

Gambar 2. Hasil nilai akurasi menggunakan data *Confusion matrix*

Setelah didapatkannya hasil perhitungan pada *Confusion matrix* tahapan selanjutnya adalah membuat *ROC Curve* berdasarkan antara nilai *false* positif dengan *true* positif. ROC adalah grafik yang menjadikan hasil dari *false* positif sebagai garis horizontal dan hasil dari *true* positif untuk mengukur perbedaan performansi metode yang digunakan, dan ROC biasanya digunakan untuk mengekspresikan *Confusion matrix* (Santosa and Yuliantara, 2017). Pada *ROC Curve* dapat menghasil kan nilai untuk melihat kinerja model menggunakan *Area Under The Curve* (AUC) (Memprediksi and Diabetes, 2017). Untuk hasil grafik dari ROC dan hasil skor AUC dalam Algoritma *Random Forest* dan Algoritma *Gradient Boosting Classifier* dapat dilihat pada gambar 3.



Gambar 3. Kurva ROC untuk penyakit Diabetes

Pada gambar 3 menunjukkan Kurva ROC yang dihasilkan dari Algoritma *Random Forest* menghasilkan nilai AUC sebesar 0.83, sedangkan Kurva ROC yang dihasilkan dari Algoritma Gradient Boosting menghasilkan nilai AUC sebesar 0.87. Berdasarkan tabel klasifikasi performa nilai yang dihasilkan adalah baik karena berada di atas 80% (Suwarno & AA Abdillah, 2017). Setelah di dapatkan hasil dari *Confusion matrix*, tahap selanjutnya penulis akan melakukan perhitungan terhadap nilai *Recall*, *Precision*, dan *F1-score*. *Recall* memiliki fungsi untuk mengevaluasi seberapa besar cakupan dari sebuah model dalam melakukan prediksi suatu kelas tertentu (Hanifa, Adiwijaya and Al-Faraby, 2017). Langkah selanjutnya adalah mencari nilai *Precision*, *Precision* dihitung untuk melakukan evaluasi seberapa baik ketepatan model dapat memprediksi suatu kelas (Hanifa, Adiwijaya and Al-Faraby, 2017). Langkah terakhir yang akan penulis lakukan adalah melakukan perbandingan hasil dari kedua Algoritma *Random Forest* dan Algoritma *Gradient Boosting Classifier*. Untuk melihat hasil dari kedua Algoritma dapat di dilihat pada tabel 4.

Tabel 4. *Precision, Recall and F1-score*

	Precision	*Recall*	*F1-score*	*Support*
0	*0.90*	*0.78*	*0.83*	*100*
1	*0.67*	*0.83*	*0.74*	*54*
Accuracy			*0.80*	*154*
Macro avg	*0.78*	*0.81*	*0.79*	*154*
Weighted avg	*0.82*	*0.80*	*0.80*	*154*

Berdasarkan analisis yang telah dilakukan, dengan membandingkan akurasi dan ROC, *Gradient Boosting Classifier* merupakan model terbaik untuk memprediksi seseorang terkena diabetes. Setelah mengubah parameter referensi yang berkorelasi kuat dengan penyakit diabetes, akurasi prediksi model yang digunakan menjadi naik. Akurasi prediksi model *Gradient Boosting Classifier* mencapai 0.81%.

5. KESIMPULAN

Berdasarkan hasil penelitian dengan menggunakan rasio 80:20, Algoritma *Random Forest* memiliki hasil akurasi sebesar 79%. Sementara dengan menggunakan Algoritma *Gradient Boosting Classifier*, hasil akurasi yang didapat sebesar 81%. Dari hasil tersebut menunjukkan bahwa Algoritma *Gradient Boosting Classifier* memiliki hasil evaluasi akurasi yang lebih besar dibandingkan dengan Algoritma *Random Forest*. Dengan membandingkan akurasi dan ROC, disimpulkan bahwa algoritma *Gradient Boosting Classifier* merupakan model terbaik untuk memprediksi seseorang terkena diabetes.

6. SARAN

Penelitian selanjutnya diharapkan dapat menganalisis secara mendalam parameter yang dihasilkan menggunakan kedua algoritma (*Random Forest* dan *Gradient Boosting Classifier*) tersebut mengapa memberikan nilai baik atau buruk serta membandingkan algoritma klasifikasi lainnya dalam penyakit diabetes maupun penyakit lainnya yang sedang ramai menjadi perbincangan.

7. DAFTAR PUSTAKA

Agajanian, S., Oluyemi, O. and Verkhivker, G.M. (2019) 'Integration of *Random Forest* classifiers and deep convolutional neural networks for classification and biomolecular modeling of cancer driver mutations', *Frontiers in Molecular Biosciences*, 6(JUN). Available at: https://doi.org/10.3389/fmolb.2019.00044.

Alita, D. and Isnain, A.R. (2020) 'Pendeteksian Sarkasme pada Proses Analisis Sentimen Menggunakan *Random Forest* Classifier', *Jurnal Komputasi*, 8(2), pp. 50–58. Available at: https://doi.org/10.23960/komputasi.v8i2.2615.

Argina, A.M. (2020) 'Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes', *Indonesian Journal of Data and Science*, 1(2), pp. 29–33. Available at: https://doi.org/10.33096/ijodas.v1i2.11.

Arlis, S. *et al.* (2018) 'Pola Penentuan Status Peminjaman Dengan', pp. 619–623.

Azar, A.T. *et al.* (2014) 'A *Random Forest* classifier for lymph diseases', *Computer Methods and Programs in Biomedicine*, 113(2), pp. 465–473. Available at: https://doi.org/10.1016/j.cmpb.2013.11.004.

Azis, A. (2020) 'Identifikasi Jenis Ikan Menggunakan Model Hybrid Deep Learning Dan Algoritma Klasifikasi', *Sebatik*, 24(2), pp. 201–206. Available at: https://doi.org/10.46984/sebatik.v24i2.1057.

Bentéjac, C., Csörgő, A. and Martínez-Muñoz, G. (2019) 'A Comparative Analysis of XGBoost', pp. 1–20. Available at: https://doi.org/10.1007/s10462-020-09896-5.

Brian, T. (2017) 'Analisis Learning Rates Pada Algoritma Backpropagation Untuk Klasifikasi Penyakit Diabetes', *Edutic - Scientific Journal of Informatics Education*, 3(1), pp. 21–27. Available at: https://doi.org/10.21107/edutic.v3i1.2557.

Eng, C.L., Tong, J.C. and Tan, T.W. (2014) 'Predicting host tropism of influenza A virus proteins using *Random Forest*', *BMC Medical Genomics*, 7(3), pp. 1–11. Available at: https://doi.org/10.1186/1755-8794-7-S3-S1.

Gayatri, L. and Hendry, H. (2021) 'PemetaaPenyebaran Covid-19 Pada Tingkat Kabupaten/Kota Di Pulau Jawa Menggunakan Algoritma K-Means Clustering', *Sebatik*, 25(2), pp. 493–499. Available at: https://doi.org/10.46984/sebatik.v25i2.1307.

Hanifa, T.T., Adiwijaya and Al-Faraby, S. (2017) 'Analisis Churn Prediction pada Data Pelanggan PT. Telekomunikasi dengan Logistic Regression dan Underbagging', *e-Proceeding of Engineering* , 4(2), pp. 3210–3225.

Hasan, F.N., Hikmah, N. and Utami, D.Y. (2018) 'Perbandingan Algoritma C4.5, KNN, dan Naive Bayes untuk Penentuan Model Klasifikasi Penanggung jawab BSI Entrepreneur Center', *Jurnal Pilar Nusa Mandiri*, 14(2), p. 169. Available at: https://doi.org/10.33480/pilar.v14i2.908.

Indrayani, Sugianti, D. and Al Karomi, M.A. (2019) 'Optimasi Parameter K pada Algoritma K-Nearest Neighbour untuk Klasifikasi Penyakit Diabetes Mellitus', *Prosiding SNATIF ke-6 Tahun 2019*, (2007), pp. 96–101.

Indriani, A. (2020) 'Analisa Perbandingan Metode Naïve Bayes Classifier Dan K-Nearest Neighbor Terhadap Klasifikasi Data', *Sebatik*, 24(1), pp. 1–7. Available at: https://doi.org/10.46984/sebatik.v24i1.909.

Ke, G. *et al.* (2017) 'LightGBM: A highly efficient gradient boosting decision tree', *Advances in Neural Information Processing Systems*, 2017-December(Nips), pp. 3147–3155.

Kementerian Kesehatan RI. (2020) 'Infodatin tetap produktif, cegah, dan atasi Diabetes Melitus 2020', *Pusat Data dan Informasi Kementerian Kesehatan RI*, pp. 1–10.

Kumari, S., Kumar, D. and Mittal, M. (2021) 'An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier', *International Journal of Cognitive Computing in Engineering*, 2(January), pp. 40–46. Available at: https://doi.org/10.1016/j.ijcce.2021.01.001.

Larose, D.T. (2004) 'Discovering Knowledge in Data', *Discovering Knowledge in Data* [Preprint]. Available at: https://doi.org/10.1002/0471687545.

Memprediksi, U. and Diabetes, P. (2017) 'Penerapan Algoritma Bayesian Regularization Backpropagation Untuk Memprediksi Penyakit Diabetes', *Jurnal MIPA*, 39(2), pp. 150–158.

Murtadho, A. and Sulistyawati, D.H. (2020) 'Machine Learning Untuk Perbandingan Tingkat Akurasi Prediksi Penyakit Diabetes Dengan Supervised Learning', *Repository Untag Surabaya* [Preprint], (Ml).

NAHZAT, S. and YAĞANOĞLU, M. (2021) 'Makine Öğrenimi Sınıflandırma Algoritmalarını Kullanarak Diyabet Tahmini', *European Journal of Science and Technology*, (24), pp. 53–59. Available at: https://doi.org/10.31590/ejosat.899716.

Natekin, A. and Knoll, A. (2013) 'Gradient boosting machines, a tutorial', *Frontiers in Neurorobotics*, 7(DEC). Available at: https://doi.org/10.3389/fnbot.2013.00021.

Rahman, M.F. *et al.* (2017) 'Klasifikasi Untuk Diagnosa Diabetes Menggunakan Metode Bayesian Regularization Neural Network (RBNN)', *Jurnal Informatika*, 11(1), p. 36. Available at: https://doi.org/10.26555/jifo.v11i1.a5452.

Santosa, S. and Yuliantara, R. (2017) 'Model Prediksi Pola Loyalitas Pelanggan Telekomunikasi Menggunakan Naive Bayes Dengan Optimasi Particle Swarm Optimization', *Jurnal Teknologi Informasi*, 13(2), pp. 154–169. Available at: http://research.

Sullivan, R. (2012) 'Introduction to data *mining* for the life sciences', *Introduction to Data Mining for the Life Sciences*, 9781597452, pp. 1–635. Available at: https://doi.org/10.1007/978-1-59745-290-8.

Vitrack-Tamam, S. *et al.* (2020) '*Random Forest* algorithm improves detection of physiological activity embedded within reflectance spectra using stomatal conductance as a test case', *Remote Sensing*, 12(14). Available at: https://doi.org/10.3390/rs12142213.

Wang, L. *et al.* (2016) 'Estimation of biomass in wheat using *Random Forest* regression algorithm and remote sensing data', *Crop Journal*, 4(3), pp. 212–219. Available at: https://doi.org/10.1016/j.cj.2016.01.008.

Yanto, M., Mulyani, S.R. and Mayola, L. (2019) 'Peramalan Jumlah Produksi Air Dengan Algoritma Backpropagation', *Sebatik*, 23(1), pp. 172–177. Available at: https://doi.org/10.46984/sebatik.v23i1.465.

Zhang, Y. and Haghani, A. (2015) 'A gradient boosting method to improve travel time prediction', *Transportation Research Part C: Emerging Technologies*, 58, pp. 308–324. Available at: https://doi.org/10.1016/j.trc.2015.02.019.

UCAPAN TERIMA KASIH

Apresiasi dan terima kasih kepada LPPM Institut Teknologi Del yang telah mendukung dan menyediakan dana sehingga penelitian ini dapat diselesaikan.