Metode *Random Forest* untuk Klasifikasi Penyakit Diabetes

Dhea Agustina Hadi*, Dwi Agustin Nuriani Sirodj

Prodi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Bandung, Indonesia.

*dheahadi3@gmail.com, dwi.agustinnuriani@unisba.ac.id

Abstract. Random Forest is a supervised learning algorithm developed from decision trees with the application of boostrap aggregating (bagging). This method grows trees from decision trees to produce a forest or the best model called the random forest model. Tree growth is done with randomly selected data with returns through the bagging process. Random forest is considered to provide better performance results for diabetes data among other supervised learning methods, because random forest and has the lowest error rate compared to other methods. Random forest is also an important technique for medical data classification, especially for diagnosing diabetics. In this study, classification was carried out using Pima Indian Diabetes data, which is an American tribe that lives in Arizona and Mexico. Classification analysis was carried out using an algorithm to see the level of accuracy in random forest classification on Pima Indian diabetes data. The results show that the accuracy value of random forest classification is 74.78%, this value is in the accuracy category at the fair classification level. In this random forest classification, there are three main variables that become importance variables, namely glucose then BMI, and age.

Keywords: Accuration, Pima Indian Diabetes, Random Forest.

Abstrak. *Random Forest* adalah sebuah algoritma *supervised learning* yang dikembangkan dari pohon keputusan (*decision tree*) dengan penerapan *boostrap aggregating* (*bagging*). Metode ini menumbuhkan pohon – pohon dari *decision tree* hingga menghasilkan sebuah *forest* atau model terbaik yang disebut sebagai model *random forest*. Penumbuhan pohon dilakukan dengan data yang dipilih secara acak dengan pengembalian melalui proses *bagging*. *Random forest* dianggap memberikan hasil performa yang lebih baik untuk data diabetes diantara metode *supervised learning* lain, karena *random forest* dan memiliki tingkat *error* paling rendah dibandingkan dengan metode lainnya. *Random forest* juga merupakan teknik yang penting untuk klasifikasi data medis, khususnya untuk diagnosa penderita diabetes. Pada penelitian ini dilakukan klasifikasi menggunakan data Pima Indian Diabetes, yaitu suku Amerika yang tinggal di Arizona dan Meksiko. Analisis klasifikasi dilakukan dengan menggunakan algoritma untuk melihat tingkat akurasi pada klasifikasi *random forest* terhadap data Pima Indian diabetes. Hasilnya diketahui bahwa nilai akurasi pada klasifikasi *random forest* adalah sebesar 74,78%, nilai tersebut berada pada kategori akurasi di tingkat *fair classification*. Pada klasifikasi *random forest* ini terdapat tiga variabel utama yang menjadi variabel *importance* yaitu *glucose* kemudian BMI, dan *age*.

Kata Kunci: *Akurasi, Pima Indian Diabetes, Random Forest*.

Corresponding Author
Email: dwi.agustinns@gmail.com

A. Pendahuluan

Random Forest adalah sebuah algoritma *supervised learning* yang dikembangkan dari pohon keputusan (*decision tree*) dengan penerapan *boostrap aggregating* (*bagging*). Metode ini menumbuhkan pohon – pohon dari *decision tree* hingga menghasilkan sebuah *forest* atau model terbaik yang disebut sebagai model *random forest*. Penumbuhan pohon dilakukan dengan data yang dipilih secara acak dengan pengembalian melalui proses *bagging* (Breiman, 2001). *Random Forest* dapat memberikan nilai akurasi yang tinggi dari algoritma *supervised learning* lainnya (Budianti & Suliadi, 2022) dan merupakan salah satu algoritma yang cocok untuk data besar.

Pima Indian adalah suku Amerika yang tinggal di Kawasan Arizona dan Meksiko. Diabetes pada suku Pima Indian lebih dikenal oleh dunia dibandingkan dengan kebudayaan dan sejarahnya (Smith-Morris, 2004). Lebih dari setengah penduduk berusia sekitar 35 tahun adalah penderita diabetes. Diabetes ialah penyakit dengan tanda peningkatan kadar gula dalam darah. Hal ini disebabkan oleh adanya gangguan dalam tubuh, sehingga glukosa menumpuk dalam darah karena tubuh tidak mampu menggunakan glukosa darah di dalam sel. Menurut *International Diabetes Federation* atau IDF, diperkirakan orang dewasa yang mengalami diabetes pada tahun 1980 berjumlah 108 juta orang, kemudian meningkat jauh hingga 415 juta orang pada tahun 2015. Tahun 2021, Penderita diabetes semakin bertambah menjadi 537 juta orang pada tahun 2021 dan diperkirakan akan terus meningkat hingga 783 juta orang pada tahun 2045.

Diabetes adalah penyakit menahun yang dapat diderita seumur hidup (Sihotang, 2017) serta dapat menyebabkan komplikasi jika penyakit tersebut tidak segera diidentifikasi secara akurat. Untuk melakukan identifikasi secara akurat maka dilakukan penelitian dengan berbagai macam metode, salah satunya dengan metode *supervised learning*, yakni klasifikasi dengan *random forest*. *Random forest* dianggap memberikan hasil performa yang lebih baik untuk data diabetes diantara metode *supervised learning* lain, karena *random forest* dan memiliki tingkat *error* paling rendah dibandingkan dengan metode lainnya (Suryanegara *et al*., 2021). *Random forest* juga merupakan teknik yang penting untuk klasifikasi data medis, khususnya untuk diagnosa penderita diabetes (Benbelkacem & Atmani, 2019). Salah satu contohnya adalah Saxena *et al*. (2022) yang melakukan penelitian dengan berbagai jenis algoritma terhadap data diabetes. Hasilnya, *random forest* memiliki akurasi paling tinggi dibandingkan dengan algoritma lainnya. Kemudian Nahzat dan Yağanoğlu (2021) juga melakukan penelitian menggunakan berbagai macam algoritma, yaitu *k-nearest neigbors* (KNN), *random forest*, *super vector machine* (SVM), *artificial neural network* (ANN) dan *decision tree* untuk mengklasifikasi diabetes. Dari penelitian tersebut didapatkan hasil bahwa *random forest* lebih unggul dibandingkan dengan algoritma lainnya. Berdasarkan latar belakang tersebut, terdapat tujuan sebagai berikut.

1. Melakukan klasifikasi *random forest* terhadap data Pima Indian Diabetes.
2. Mengetahui tingkat akurasi dari klasifikasi *random forest* terhadap data Pima Indian Diabetes.

B. Metodologi Penelitian

Data yang digunakan pada penelitian ini adalah set data dari Pima Indian Diabetes yang berasal dari National Institute of Diabetes and Kidney Diseases dengan variabel respon atau label outcome, 0 untuk diagnosa tidak menderita diabetes dan 1 untuk diagnosa menderita diabetes. Faktor – faktor yang dianggap mepengaruhi variabel respon outcome adalah pregnancies (jumlah kehamilan semasa hidup), glucose (kadar gula darah mg/dL), blood pressure (tekanan darah diastolik mm/Hg), skin thickness (lemak tubuh triceps mm), insulin (tingkat serum inulin), body mass index atau BMI (indeks massa tubuh), diabetes pedigree function (indikator riwayat diabetes dalam keluarga), dan age (umur). Metode dalam penelitian ini adalah algoritma random forest yang akan digunakan untuk meng-klasifikasikan data diabetes.

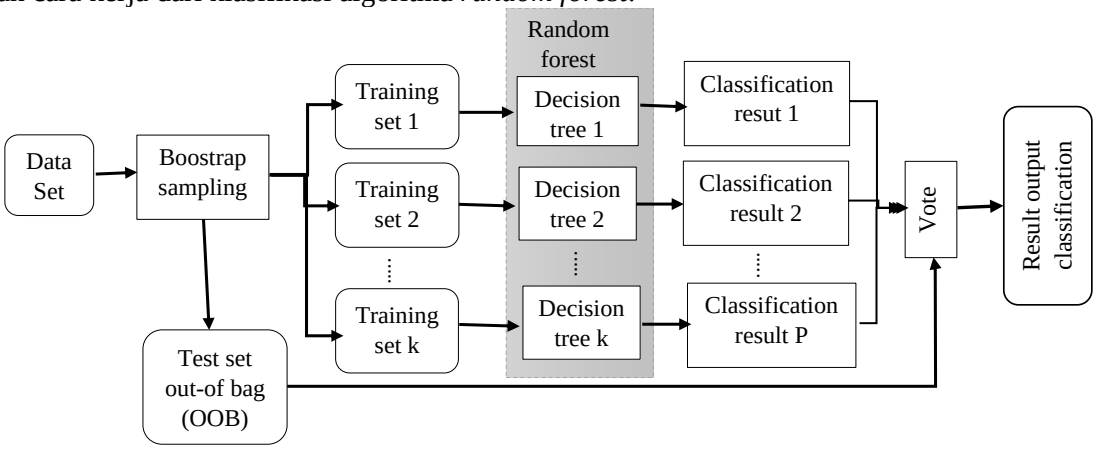
Pre-processing

Pre-processing adalah proses dalam *machine learning* yang dilakukan untuk mempersiapkan

data agar data siap digunakan. Proses ini sangat berpengaruh untuk performa dari *machine learning* karena *machine learning* hanya mempelajari data yang relevan (Budiarti, 2006). Salah satu permasalahan data yang harus dilakukan *pre-processing* adalah *missing data* atau data hilang. *Missing data* dapat ditangani dengan menghapus baris data yang memiliki *missing data*. Namun hal tersebut dapat mengurangi ketepatan dalam pendugaan karena jumlah data yang diambil akan berkurang (Hendrawati, 2015). Terdapat solusi lain untuk *missing data* tanpa menghapus baris data, salah satunya dengan melakukan imputasi data. Metode imputasi data ialah metode menentukan nilai dari data hilang menggunakan nilai pengganti yang lebih konstan (Arfarisi *et al*., 2013). Imputasi data dilakukan berdasarkan jenis pada set data, seperti nilai rata – rata digunakan untuk mengganti data numerik dan nilai modus digunakan untuk mengganti data kategorik.

Random Forest

Seperti namanya, *random forest* merupakan sekumpulan pohon (*tree*) yang membentuk sebuah hutan (*forest*) atau sebuah model baru dan menerapkan *boostrap aggregating* (*bagging*) serta *random feature selection* (Breiman, 2001). Misal terdapat variabel prediktor $X_1, X_2, \dots, X_p$ dengan variabel respon yaitu Y, diambil secara bersamaan dari $\mathcal{X} \times \mathcal{y}$. Diasumsikan bahwa distribusi bersama tidak diketahui, yaitu $P_{XY}(X, Y)$ dan X menunjukkan *random vector* $X = (X_1, X_2, \dots, X_p)$. $P(X = x, Y = y)$ adalah peluang variabel acak saat pengambilan nilai X dari $\mathcal{X}$ dan Y dari $\mathcal{y}$. Setelahnya akan dibentuk pohon – pohon untuk menemukan prediksi dari fungsi $f(X)$ untuk memprediksi Y. Pohon tersebut ditumbuhkan dengan menggunakan metode *classification and regression tree* atau yang dikenal dengan CART (Breiman, 2001). Berikut adalah cara kerja dari klasifikasi algoritma *random forest*.



Gambar 1. Klasifikasi dengan Random Forest

Untuk melakukan analisis dengan menggunakan algoritma *random forest*, data dibagi menjadi data *training* dan data *testing*. Data *training* digunakan untuk membangun model *random forest* sedangkan data *testing* dimasukkan ke dalam model *random forest* untuk mengetahui keakuratan dalam model. Dalam penelitian ini, data *training* diambil sebesar 70% dan 30% lainnya digunakan untuk data *testing*. Proses pembagian data ke dalam data *training* dan data *testing* tidak memiliki ketentuan khusus, tetapi pengujian akan lebih baik jika data *training* lebih banyak dari data *testing* (Budiman & Ramadina, 2015).

Berikut adalah klasifikasi dengan algoritma *random forest* (Breiman, 2001).

1. Boostrap sample yaitu pengambilan sample dengan random sampling with replacement berukuran n dari kumpulan data training.
2. Membangun pohon hingga pohon berukuran maksimum. Jumlah variabel diambil pada setiap simpul dengan menggunakan $mtry = \sqrt{p}$, dimana $mtry$ adalah ukuran peubah prediktor dan p adalah variabel yang akan diambil untuk membentuk pohon. Setelahnya dilakukan pembentukan pohon menggunakan CART dengan tahapan sebagai berikut:
    a. Memecah dan melakukan pelabelan pada simpul (memilah simpul).

Simpul dipilih berurutan berdasarkan nilai indeks gini yang paling besar. Banyaknya pemecahan dapat diketahui dengan menghitung $2^{M-1}-1$, dimana M adalah banyak kategori dari suatu variabel prediktor. Indeks gini didapat dari persamaan berikut.

$$GINI(t) = 1 - \sum_{i \neq j}[p(j|t)]^2 = \sum_{j=1}^{j-1}\sum_{j'=j+1}^{j}\frac{n_j(t)}{n(t)}\frac{n'_j(t)}{n(t)} \quad \ldots(1)$$

Setelah pemilahan simpul, dilakukan maka akan dicari nilai *goodness of split* dengan persaamaan berikut.

$$\phi(s,t) = \Delta i(s,t) = GINI(t) - p_L(t).GINI_L(t) - p_R(t).GINI_R(t) \quad \ldots(2)$$

Goodness of split pemilah s pada simpul ke-t adalah $\phi(s,t)$, indeks gini pada simpul ke-t adalah $GINI(t)$, frekuensi relatif pada simpul kiri adalah $p_L(t)$, frekuensi relatif pada simpul kanan adalah $p_R(t)$, indeks gini simpul t ke simpul kiri adalah $GINI_L(t)$, indeks gini simpul t ke simpul kanan adalah $GINI_R(t)$. Untuk melakukan pemecahan simpul t, maka akan dicari simpul dengan keheterogenan yang tinggi, yaitu dengan mencari nilai terbesar dari *goodness of split* dengan persamaan berikut.

$$\Delta i(s^*,t) = max_{s \epsilon S}\Delta i(s^*,t) \quad \ldots(3)$$

3. Menghentikan proses pembentukan pohon.

Proses pembentukan pohon dihentikan ketika pohon dinyatakan homogen yaitu tiap data dapat dimasukkan pada kelas dengan kategori yang sama dan sudah tidak ada kemungkinan pemilahan.

4. Langkah ke-1 sampai ke-2 diulang hingga *k* kali untuk membangun hutan dengan sejumlah *k* pohon.
5. Berdasarkan *k* buah pohon akan dilakukan penggabungan dengan *majority vote* pada persamaan berikut.

$$f(x) = argmax_Y \sum_{k=1}^{k} I(h_k(x) = Y) \quad \ldots(4)$$

Hasil akhir prediksi adalah $f(x)$, kelas dengan perhitungan maksimum dari seluruh pohon keputusan adalah argument the maxima atau $argmax_Y$, fungsi indikator dilambangan dengan $I(...)$, prediksi dari variabel respon x dengan model pohon ke-k adalah $h_k(x)$, dan *output* variabel adalah Y.

Confusion matrix

Confusion matrix adalah sebuah matriks berukuran $N \times N$ untuk mengevaluasi performa dalam sebuah model klasifikasi, dimana *N* adalah nomor dari kelas targetnya. Akurasi dapat ditentukan dari model dengan mengobservasi jumlah klasifikasi secara diagonal melalui *confusion matrix* (Ting, 2017).

Tabel 1. Confusion Matriks

Kelas prediksi (*predicted class*)	Kelas sebenarnya (*true class*)		
		Yes	*No*
	Yes	TP	FP
	No	FN	TN

Kolom confusion matrix merepresentasikan hasil dari kelas yang sesungguhnya, sedangkan barisnya adalah kelas yang diprediksi oleh peneliti, dan sebaliknya. True positive (TP) adalah ketika prediksi peneliti adalah yes dan data sesungguhnya yes. False positive (FP) adalah ketika prediksi peneliti adalah yes, namun data sesunguhnya no, disebut juga sebagai kesalahan tipe I. False negative (FN) adalah ketika prediksi peneliti no dan data sesungguhnya yes, disebut juga

sebagai kesalahan tipe II. True negative (TN) adalah ketika prediksi peneliti no, dan data sesungguhnya no. Dari Tabel 1 dapat dilakukan perhitungan untuk mengukur nilai akurasi. Akurasi adalah teknik dasar yang dilakukan untuk mengukur ketepatan hasil prediksi dari klasifikasi.

$$\text{Akurasi} = \frac{TP+TN}{TP+FP+TN+FN} \quad \ldots(5)$$

Menurut Subarkah (2020), keakuratan klasifikasi dapat dilihat dengan tingkatan kategori sebagai berikut.

Tabel 2 Kategori Nilai Akurasi Klasifikasi

Nilai	Tingkatan kategori
0,90 – 1,00	*excellent classification*
0,80 – 0,90	*good classification*
0,70 – 0,80	*fair classification*
0,60 – 0,70	*poor classification*
0,50 – 0,60	*failure*

Variable Importance

Variable importance adalah ukuran untuk mengukur tingkat kepentingan suatu variabel. Semakin tinggi nilainya, maka variabel tersebut akan menjadi semakin penting. Perhitungan dari *variable importance* dapat memberi informasi yang lebih banyak saat menginterpretasikan masalah. *Variable importance* bekerja sesuai dengan seberapa besar penurunan akurasi dari model atau *error* suatu model, setelah dan sebelum dilakukan permutasi (Mercadier & Lardy, 2019). Proses dari pembuatan *variable importance* disebut dengan *mean decrease gini* (MDG). Misal terdapat peubah prediktor sebanyak *p*, maka tingkat kepentingan peubah prediktor x_j diukur dengan persamaan berikut.

$$MDG(x_j) = \frac{1}{k}\left[\sum_{k} Gini(j)^k\right] \quad \ldots(6)$$

Indeks gini untuk peubah prediktor adalah $Gini(j)^k$, banyaknya pohon dalam *random forest* adalah k, banyaknya peubah prediktor adalah j, j = (1,2,…,p).

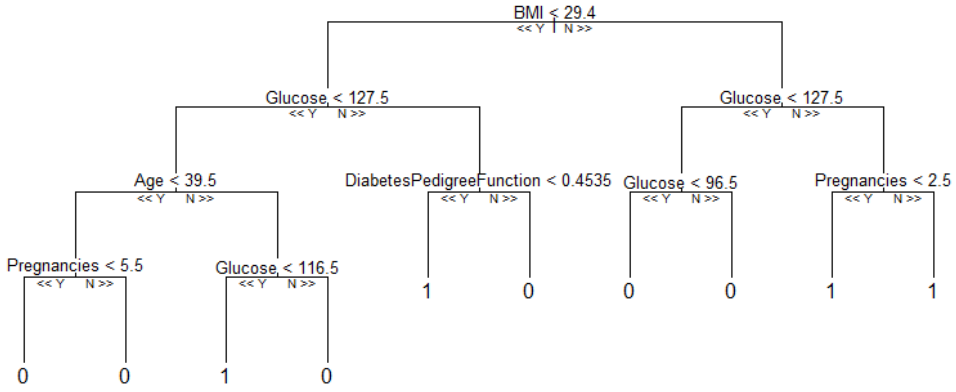
C. Hasil Penelitian dan Pembahasan

Penelitian dilakukan dengan menggunakan set data Pima Indian Diabetes dengan 8 variabel prediktor yaitu *pregnancies*, *glucose*, *blood pressure*, *skin thickness*, insulin, *body mass index* atau BMI, *diabetes pedigree function*, dan *age*, serta 1 variabel respon yaitu *outcome*. Data melalui *pre-processing* yaitu imputasi data untuk menangani *missing data*. Kemudian, data dibagi menjadi 70% data *training* dan 30% data *testing*. Setelah itu, data training masuk ke proses *random forest* hingga terdapat model dari klasifikasi. Data *testing* akan dimasukkan pada model klasifikasi sehingga didapatkan tabel *confussion matrix* dan dilakukan perhitungan akurasi.

Klasifikasi *Random Forest*

Untuk melakukan klasifikasi *random forest*, perlu menentukan pemilah yang diambil dari ukuran peubah predictor dan jumlah pohon yang akan dibangun. Maka, akan dibentuk sebanyak 1000 pohon dan pemilahnya akan dipilih sebanyak $mtry = \sqrt{8} = 2{,}8284 \approx 3$,

diambil dari p = 8. Pohon tersebut akan dibangun menggunakan data *training* dan nilai *mtry* akan digunakan untuk mencari pemilah terbaik untuk akar (*root*) dari masing – masing pohon. Setelah membangun 1000 pohon, kemudian *voting* dilakukan dan menghasilkan model dari *random forest*. Model *random forest* pada pengujian ini adalah sebagai berikut.



Gambar 2. Model Random Forest

Setelah model *random forest* didapatkan, selanjutnya adalah memasukkan data *testing* satu per-satu ke dalam model. Kemudian model akan memberikan hasil pengujian seperti pada Tabel 3. sedangkan evaluasinya didapatkan dari perhitungan nilai akurasi menggunakan persamaan (5).

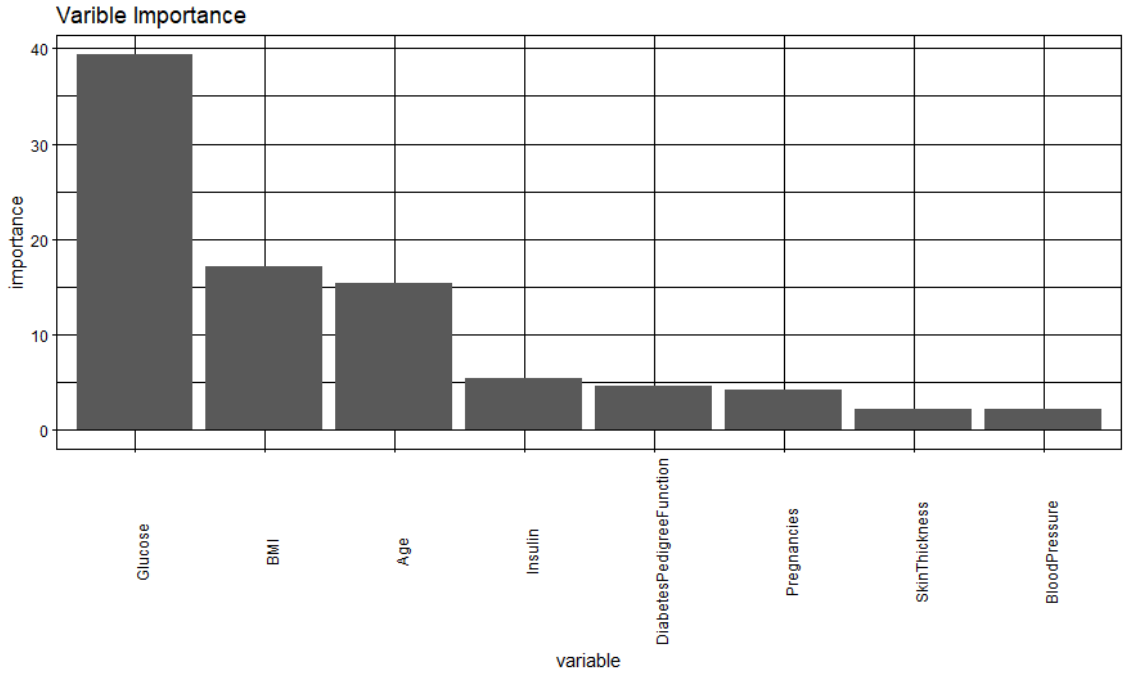
Tabel 3. Confussion Matrix Random Forest Tanpa Penerapan Normalisasi Min-max

Klasifikasi Random Forest		Prediksi		Total
		Tidak Diabetes	Diabetes	
Kelas sebenarnya	Tidak Diabetes	124	21	145
	Diabetes	37	48	85
Total		161	69	230

Hasil *confussion matrix* pada tabel di atas menunjukkan 230 kasus pada data *testing*, 145 orang didiagnosa tidak menderita diabetes dan 85 orang didiagnosa menderita diabetes, sedangkan menurut prediksinya ada 161 orang didiagnosa tidak menderita diabetes dan 69 orang didiagnosa menderita diabetes. Terdapat 124 orang didiagnosa tidak menderita diabetes diprediksi secara benar, dan 21 orang didiagnosa tidak menderita diabetes namun diprediksi menderita diabetes. Kemudian untuk penderita diabetes terdapat 48 orang yang diprediksi secara benar, dan 37 orang didiagnosa menderita diabetes namun diprediksi tidak menderita diabetes.

$$\text{Akurasi} = \frac{124+48}{124+21+37+48} = 0{,}7478 = 74{,}78\%$$

Dari perhitungan di atas, didapatkan nilai akurasi sebesar 0,7478, menyatakan bahwa 74,78% model dapat diprediksi secara tepat untuk mengelompokkan penderita diabetes maupun yang tidak menderita diabetes. Nilai tersebut berada pada kategori akurasi di tingkat *fair classification*, yaitu klasifikasi dinilai cukup baik dalam melakukan klasifikasi dan prediksi.



Gambar 3. Variabel Importance Pada Analisis Klasifikasi Random Forest

Gambar 3 adalah variabel *importance*, yaitu variabel yang mempengaruhi hasil prediksi dari kinerja *random forest*. Variabel *glucose* adalah variabel yang dinyatakan sangat penting terhadap prediksi dari klasifikasi *random forest* terhadap data Pima Indian Diabetes. Berdasarkan hasil variabel *importance* tersebut, dapat diketahui bahwa nilai *mean decreased gini* (MDG) yang dihitung berdasarkan persamaan (6) milik *glucose* adalah nilai MDG paling besar dibandingkan variabel lainnya. Setelah *glucose*, terdapat dua variabel teratas lainnya yang dianggap penting, yaitu BMI dan *age*.

D. Kesimpulan

Berdasarkan penelitian yang telah dilakukan, klasifikasi *random forest* terhadap data Pima Indian Diabetes memiliki nilai akurasi 74,78%, nilai tersebut berada pada tingkat *fair classification* atau dianggap cukup baik untuk melakukan klasifikasi dan prediksi terhadap data Pima Indian Diabetes. Pada klasifikasi *random forest* terdapat variabel *importance* yang berperan penting dalam diagnosa ini. Tiga variabel yang dianggap paling penting dalam diagnosa pada Pima Indian Diabetes adalah *glucose* kemudian BMI, dan *age*.

Acknowledge

Penelitian ini merupakan sebagian dari penelitian pada tugas akhir yang penulis lakukan. Penelitian ini dapat diselesaikan berkat Rahmat dari Allah SWT., dan dukungan dari berbagai pihak. Maka, penulis ingin mengucapkan terima kasih kepada kedua orang tua dan adik – adik, kepada Ibu Dwi Agustin Nuriani Sirodj S.Si., M.Stat. selaku pembimbing, jajaran dosen Statistika Unisba yang telah memberikan ilmunya, serta dukungan dari seluruh teman.

Daftar Pustaka

[1] Arfarisi, A. R., Tjandrasa, H., & Arieshanti, I. (2013). Perbandingan Performa antara Imputasi Metode Konvensional dan Imputasi dengan Algoritma Mutual Nearest Neighbor. *JURNAL TEKNIK POMITS*, *2*(1), 1–4.

[2] Benbelkacem, S., & Atmani, B. (2019). Random forests for diabetes diagnosis. *2019 International Conference on Computer and Information Sciences, ICCIS 2019*, 1–4. https://doi.org/10.1109/ICCISci.2019.8716405

[3] Breiman, L. (2001). *Random Forest* [University of California Berkeley]. https://doi.org/10.14569/ijacsa.2016.070603

[4] Budianti, L., & Suliadi. (2022). Metode Weighted Random Forest dalam Klasifikasi Prediksi Kelangsungan Hidup Pasien Gagal Jantung. *Bandung Conference Series: Statistics*, *2*(2), 103–110. https://doi.org/10.29313/bcss.v2i2.3318

[5] Budiarti, A. (2006). Bab 2 landasan teori. *Aplikasi Dan Analisis Literatur Fasilkom UI*, *Dm*, 4–25.

[6] Budiman, I., & Ramadina, R. (2015). Penerapan Fungsi Data Mining Klasifikasi untuk Prediksi Masa Studi Mahasiswa Tepat Waktu pada Sistem Informasi Akademik Perguruan Tinggi. *Ijccs*, *x*, *No.x*(1), 1–5.

[7] Hendrawati, T. (2015). Kajian Metode Imputasi Dalam Menangani Missing Data. *Prosiding Seminar Nasional Matematika Dan Pendidikan Matematika UMS*, 637–642.

[8] Mercadier, M., & Lardy, J. P. (2019). Credit spread approximation and improvement using random forest regression. *European Journal of Operational Research*, *277*(1), 351–365. https://doi.org/10.1016/j.ejor.2019.02.005

[9] Nahzat, S., & Yağanoğlu, M. (2021). Diabetes Prediction Using Machine Learning Classification Algorithms. *European Journal of Science and Technology*, *24*, 53–59. https://doi.org/10.31590/ejosat.899716

[10] Saxena, R., Sharma, S. K., Gupta, M., & Sampada, G. C. (2022). A Novel Approach for Feature Selection and Classification of Diabetes Mellitus: Machine Learning Methods. *Computational Intelligence and Neuroscience*, *2022*. https://doi.org/10.1155/2022/3820360

[11] Sihotang, H. T. (2017). Perancangan Aplikasi Sistem Pakar Diagnosa Diabetes Dengan Metode Bayes. *Jurnal Manik Penusa*, *1*(1), 36–41.

[12] Smith-Morris, C. M. (2004). Reducing Diabetes in Indian Country: Lessons from the Three Domains Influencing Pima Diabetes. *Human Organization*, *63*(1), 34–46.

[13] Subarkah, P. (2020). *Penerapan Algoritma Klasifikasi Classification And Regression Trees (CART) untuk Diagnosis Penyakit Diabetes Retinopathy*. *19*(2), 294–301.

[14] Suryanegara, G. A. B., Adiwijaya, & Purbolaksono, M. D. (2021). Peningkatan Hasil Klasifikasi pada Algoritma Random Forest untuk Deteksi Pasien Penderita Diabetes Menggunakan Metode Normalisasi. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *5*(1), 114–122. https://doi.org/10.29207/resti.v5i1.2880

[15] Ting, K. M. (2017). Confusion Matrix. *Encyclopedia of Machine Learning and Data Mining*, *October*, 260–260. https://doi.org/10.1007/978-1-4899-7687-1_50