PENERAPAN METODE RANDOM FOREST UNTUK KLASIFIKASI DAN PREDIKSI PENYAKIT DIABETES

Siti Mahmuda[1*], Sri Wahyuningsih[1], Rito Goejantoro[1], Indriasri Raming[1], Abdul Azis[1], Sherlina Sukma Ayu[1]

[1]Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Mulawarman, Indonesia

[*]*Corresponding author:* sitimahmuda24@gmail.com

Abstrak. Penyakit diabetes merupakan salah satu masalah kesehatan global yang memerlukan deteksi dini untuk mencegah komplikasi serius. Penelitian ini bertujuan untuk membangun model klasifikasi penyakit diabetes menggunakan metode *random forest* dengan data *Pima Indians Diabetes* dari *UCI Machine Learning Repository*. Dataset terdiri dari 768 pengamatan dengan tujuh variabel prediktor, antara lain glukosa plasma, tekanan darah diastolik, ketebalan kulit, kadar insulin, indeks masa tubuh, riwayat keluarga diabetes, dan usia. Data dibagi menjadi data latih (80%) dan data uji (20%) untuk pelatihan dan pengujian model. Hasil pengujian menunjukkan bahwa model mampu mengklasifikasikan status diabetes dengan akurasi sebesar 79,87%, sensitivitas 84,54%, dan spesifisitas 68,18%, yang mencerminkan performa klasifikasi dan prediksi yang cukup baik secara keseluruhan. Analisis tingkat kepentingan variabel berdasarkan *nilai mean decrease gini* mengindikasikan bahwa glukosa plasma, indeks massa tubuh, dan usia merupakan variabel paling berpengaruh dalam proses klasifikasi. Temuan ini menunjukkan bahwa metode *random forest* efektif digunakan dalam mendeteksi risiko diabetes serta memberikan wawasan penting terkait variabel yang paling relevan dalam diagnosis berbasis data.

Kata Kunci: *random forest, diabetes, klasifikasi, prediksi, machine learning.*

1 PENDAHULUAN

Diabetes merupakan salah satu masalah kesehatan global yang terus meningkat setiap tahunnya. Menurut IDF *Diabetes Atlas* (2025), sebanyak 11,1% atau 1 dari 9 orang dewasa (usia 20–79 tahun) di dunia hidup dengan diabetes, dan lebih dari 40% di antaranya tidak menyadari bahwa mereka mengidap kondisi tersebut [1]. Tingginya prevalensi diabetes, disertai dengan banyaknya penderita yang tidak menyadari kondisinya, menunjukkan ancaman serius bagi kesehatan masyarakat global. Kondisi ini dapat menyebabkan komplikasi berat dan beban ekonomi yang besar jika tidak diantisipasi melalui deteksi dini dan pengelolaan yang tepat. Untuk mengurangi dampak diabetes, solusi umum yang ditawarkan meliputi perubahan gaya hidup sehat, edukasi masyarakat, dan deteksi dini untuk mencegah komplikasi. Kini, kemajuan teknologi dan pemanfaatan *machine learning* juga mulai dimanfaatkan untuk memprediksi risiko diabetes, memantau kondisi pasien, dan membantu pengambilan keputusan medis secara lebih cepat dan akurat.

Random forest merupakan salah satu metode dalam *machine learning* yang digunakan untuk klasifikasi dan prediksi. Metode ini bekerja dengan membentuk sejumlah pohon keputusan (*decision trees*) pada data latih, kemudian menggabungkan hasil dari masing-masing pohon untuk menghasilkan prediksi yang lebih akurat dan stabil [2]. Dalam *machine learning*, klasifikasi bertujuan untuk mengelompokkan data ke dalam kategori berdasarkan yang telah ditentukan berdasarkan pola yang dipelajari dari data historis, sementara prediksi bertujuan untuk memperkirakan nilai atau kejadian di masa depan berdasarkan variabel-variabel prediktor yang tersedia [3]. Metode *random forest* efektif digunakan untuk kedua tujuan tersebut karena mampu menangkap pola kompleks dalam data dan menghasilkan prediksi yang andal melalui proses agregasi dari banyak pohon keputusan.

Klasifikasi penyakit diabetes memiliki peran krusial dalam mendeteksi kondisi pasien sejak tahap awal, sehingga memfasilitasi pengambilan keputusan medis yang lebih cepat dan akurat [4]. Selain itu, prediksi penyakit diabetes juga bermanfaat dalam mengidentifikasi individu dengan risiko tinggi terhadap diabetes mellitus, sehingga dapat dilakukan intervensi dini untuk mencegah atau menunda perkembangan penyakit tersebut [5]. Pendekatan ini sangat relevan dalam pengembangan sistem pendukung keputusan berbasis data yang bertujuan meningkatkan kualitas layanan kesehatan secara keseluruhan.

Deteksi dini terhadap penyakit diabetes menjadi sangat krusial mengingat komplikasi yang dapat ditimbulkannya jika tidak ditangani sejak awal. Penyakit ini berkembang secara perlahan dan seringkali tanpa gejala yang jelas pada tahap awal, sehingga diperlukan sistem prediksi dengan akurasi tinggi untuk membantu proses identifikasi awal. Semakin cepat penyakit ini dikenali, semakin besar peluang untuk melakukan intervensi medis dan perubahan gaya hidup yang dapat mencegah atau memperlambat progresivitasnya. Oleh karena itu, peningkatan akurasi dalam sistem

klasifikasi dan prediksi sangat dibutuhkan guna mendukung penanganan pasien yang lebih efektif.

Untuk menjawab kebutuhan tersebut, metode *random forest* dapat digunakan sebagai salah satu pendekatan yang menjanjikan dalam meningkatkan akurasi klasifikasi dan prediksi penyakit diabetes. *Random forest* merupakan metode ensembel yang menggabungkan banyak pohon keputusan untuk menghasilkan prediksi yang lebih akurat dan andal, terutama pada data yang kompleks [6]. Keunggulan lain dari *random forest* adalah kemampuannya dalam menangani variabel-variabel yang saling berinteraksi serta menyediakan estimasi pentingnya setiap variabel dalam proses klasifikasi, yang sangat berguna dalam konteks diagnosis medis [7]. Dengan demikian, penggunaan random forest dapat memperkuat sistem klasifikasi penyakit diabetes dan mendukung pengambilan keputusan yang lebih tepat berbasis data.

Beberapa penelitian sebelumnya telah dilakukan untuk memprediksi penyakit, termasuk diabetes, sebagai upaya pencegahan dini. Salah satu pendekatan yang digunakan adalah metode regresi linear. Misalnya, penelitian oleh Suwaryo dkk [8] menggunakan regresi linear untuk memprediksi kemungkinan seseorang mengidap diabetes berdasarkan faktor risiko seperti kadar glukosa dan kadar insulin pada tubuh. Namun, pendekatan ini memiliki keterbatasan karena model regresi linear bersifat sederhana dan kurang mampu menangkap pola-pola kompleks, terutama pada data medis yang cenderung non-linear dan terdapat interaksi yang rumit antar variabel.

Di sisi lain, pendekatan *machine learning* mulai banyak digunakan untuk meningkatkan akurasi prediksi. Penelitian oleh Aditya dkk [9] menerapkan algoritma *decision tree* untuk prediksi diabetes melitus tipe 2. Meskipun model ini mampu memberikan interpretasi yang mudah dan hasil yang cukup baik, penggunaan satu algoritma klasifikasi tunggal membuat model ini cenderung kurang stabil dan memiliki risiko *overfitting* terhadap data latih [10]. Untuk mengatasi masalah tersebut, pendekatan ensembel seperti *random forest* menjadi pilihan yang lebih unggul. Sebagai contoh, penelitian oleh Aji, Suprianto, dan Dijaya [11] berhasil mencapai akurasi hingga 99%. Hasil penelitian menunjukkan bahwa metode *random forest* efektif dalam klasifikasi prediksi penyakit berkat kemampuannya dalam menangani keragaman data (data variability) dan meningkatkan generalisasi model, serta menyediakan peringkat kepentingan variabel yang informatif dalam konteks diagnosis medis.

Penelitian perbandingan antara algoritma *decision tree* dan *random forest* juga telah dilakukan untuk menilai efektivitas kedua metode dalam konteks klasifikasi penyakit. Salah satu peneltian yang dilakukan oleh Putra dan Hadayani [12] berjudul "Perbandingan Algoritma Decision Tree dan Random Forest dalam Pengklasifikasian Penyakit Tiroid" menunjukkan bahwa *random forest* memberikan hasil yang lebih unggul dibandingkan *decision tree*. Dalam penelitian tersebut, *random forest* berhasil mencapai akurasi sebesar 94,81%, sedangkan

decision tree hanya memperoleh akurasi sebesar 93,51%. Pencapaian akurasi yang lebih tinggi mendukung temuan sebelumnya bahwa metode ensembel seperti *random forest* memiliki keunggulan dibandingkan dengan model klasifikasi tunggal.

Selanjutnya, penelitian yang dilakukan oleh Sriyanto dan Supriyatna [13] dengan judul "Prediksi Penyakit Diabetes Menggunakan Algoritma Random Forest" juga memberikan kontribusi penting dalam bidang ini. Dalam studi tersebut, peneliti menggunakan dataset dari platform Kaggle dengan jumlah responden sebanyak 520 dan 17 variabel prediktor. Studi ini memiliki kesamaan dengan penelitian yang sedang dilakukan, yaitu sama-sama berfokus pada klasifikasi dan prediksi penyakit diabetes. Namun, terdapat beberapa perbedaan mendasar, antara lain sumber data yang digunakan, jumlah responden, dan jumlah variabel prediktor yang dilibatkan. Selain itu, dalam penelitian ini dilakukan eksplorasi lebih lanjut terhadap tingkat kepentingan variabel (*variable importance*), yaitu penelusuran variabel yang berkontribusi paling besar dalam proses klasifikasi. Informasi ini sangat berguna untuk mendukung proses diagnosis berbasis data dan membantu tenaga medis dalam menentukan faktor risiko utama yang perlu diperhatikan. Dengan demikian, penelitian ini tidak hanya fokus pada akurasi model, tetapi juga memberikan wawasan interpretatif yang lebih dalam terhadap variabel-variabel yang berpengaruh.

Berdasarkan kajian dan temuan sebelumnya, penelitian ini bertujuan untuk mengembangkan model prediksi diabetes menggunakan metode *random forest*, yang mampu memberikan hasil akurat dengan menggabungkan banyak pohon keputusan dan menangani data yang kompleks. Tujuan utamanya adalah mendapatkan prediksi secara cepat agar penanganan pasien bisa dilakukan lebih dini dan tepat sasaran. Deteksi dini sangat penting untuk mencegah komplikasi serius akibat keterlambatan diagnosis. Selain itu, penelitian ini juga menelusuri variabel-variabel yang paling berpengaruh dalam proses klasifikasi, sehingga bisa membantu tenaga medis mengenali faktor risiko utama dan menentukan prioritas intervensi. Dengan begitu, hasil penelitian ini diharapkan tidak hanya meningkatkan akurasi prediksi, tetapi juga mendukung pelayanan kesehatan yang lebih efektif dengan bantuan teknologi.

2 TINJAUAN PUSTAKA

2.1 Random Forest

Metode *random forest* diperkenalkan oleh Breiman pada tahun 2001 [2]. Metode *random forest* memiliki dua fungsi utama, yaitu klasifikasi dan prediksi. algoritma dasar yang digunakan dalam *random forest* adalah pohon keputusan (*decision tree*). Dengan kata lain, *random forest* merupakan kumpulan dari pohon-pohon keputusan yang digunakan untuk mengklasifikasikan dan memprediksi data dengan cara memberikan data masukan (variabel prediktor) ke akar pohon (*root*

node) pada bagian atas, kemudian diproses hingga mencapai daun (*leaf node*) pada bagian bawah [14]. Hasil analisis *random forest* untuk klasifikasi diperoleh dari *majority voting*, yaitu kelas dari variabel respon yang paling banyak muncul dari setiap pohon keputusan terbentuk, sedangkan hasil prediksi nilai numerik (regresi) diperoleh dari nilai rata-rata keluaran masing-masing pohon keputusan [15].

Metode *random forest* merupakan pengembangan dari metode *classification and regression tree* (CART), yang menerapkan metode *bagging* atau *bootstrap aggregating* serta penentuan variabel pemilah dari variabel prediktor secara acak. *Bagging* merupakan metode yang dapat meningkatkan kinerja algoritma klasifikasi. Metode ini didasarkan pada pendekatan ensembel [16]. Langkah-langkah metode *random forest* adalah sebagai berikut [17]:

1) Menetukan beberapa hal yaitu:
   a. Banyaknya pohon keputusan yang akan dibentuk, dinotasikan dengan B.
   b. Ukuran sampel *bootstrap*, dinotasikan q.
2) Mengulangi langkah-langkah berikut secara berurutan seebanyak B kali:
   a. Mengambil sampel secara *bootstrap* dari data latih L dengan ukuran q, dan hasil dari langkah ini adalah gugus data L^0.
   b. Menggunakan gugus data L^0 untuk memperoleh pohon keputusan, namun penentuan partisi di setiap simpul dipilih yang terbaik dari m variabel prediktor yang dipilih secara acak dari p variabel prediktor yang tersedia dan tidak ada proses pemangkasan.
3) Menyimpan B buah pohon yang diperoleh.
4) Menentukan hasil klasifikasi dari *majority voting* B buah pohon keputusan dan hasil prediksi nilai numerik (regresi) diperoleh dari nilai rata-rata nilai keluaran dari setiap pohon keputusan.

Metode *random forest* memiliki nilai m yang dapat bervariasi. Nilai m merupakan jumlah variabel prediktor yang digunakan sebagai pemilah (*splitters*) dalam pembentukan pohon keputusan. Hastie dkk [18] menyatakan bahwa jika m diperbesar maka *diversity* pohon menjadi berkurang karena korelasi antar pohon keputusan semakin tinggi. James dkk [19] menyatakan bahwa $m = \sqrt{p}$ adalah pilihan yang dapat digunakan pada metode *random forest*, meskipun dalam kasus tertentu dapat dilakukan pencarian m yang optimum. Penelitian Mahmuda, Nohe, dan Leonardo [20] menggunakan sebanyak 2 variabel prediktor sebagai pemilah dan menghasilkan tingkat kesalahan klasifikasi yang terkecil.

2.2 Gini Impurity

Setiap kali melakukan partisi atau pemisahan simpul (*node*) pada suatu iterasi algoritma pohon keputusan diupayakan menghasilkan simpul yang kelasnya lebih homogen atau lebih *pure*. Ukuran yang dapat digunakan sebagai nilai kuantitatif untuk melihat kehomogenan kelas adalah nilai *gini impurity,* dinotasikan gini(D). Pada variabel respon yang memiliki kategori 2 kelas (kelas 0 dan kelas 1), nilai gini(D) dirumuskan sebagai berikut:

$$gini(D) = 1 - p_0^2 - p_1^2 \quad (1)$$

Proporsi pengamatan kelas 0 dinotasikan p_0 dan proporsi pengamatan kelas 1 dinotasikan p_1. Nilai *gini impurity* berada pada rentang $0 \leq gini(D) \leq 0{,}5$. Nilai *gini impurity* yang mendekati 0 menunjukkan bahwa sebuah simpul terdiri atas pengamatan dari satu kelas saja (homogen), sedangkan nilai mendekati 0,5 menunjukkan bahwa simpul tersebut mengandung pengamatan dari kedua kelas dalam proporsi yang hampir seimbang (tidak homogen). Oleh karena itu, dalam proses pembentukan pohon keputusan, partisi terbaik adalah yang menghasilkan penurunan nilai *gini impurity* secara signifikan, karena hal ini mencerminkan peningkatan kehomogenan kelas pada simpul hasil pemisahan.

Perubahan nilai *gini impurity* dapat diukur melalui nilai *gini decrease* (GD) yaitu selisih antara nilai *gini impurity* pada data lengkap dengan rata-rata terboboti *gini impurity* pada subset data hasil partisi, dapat dinyatakan sebagai:

$$GD = gini(D) - \left[\frac{n_1}{n} gini(D_1) + \frac{n_2}{n} gini(D_2)\right] \quad (2)$$

Jika nilai *gini decrease* yang dihasilkan oleh suatu variabel prediktor pemilah cukup besar, maka variabel tersebut berperan penting dalam menghasilkan partisi yang meningkatkan kehomogenan kelas, sehingga dapat dikatakan bahwa variabel tersebut memiliki daya pisah yang baik dan berkontribusi signifikan dalam pembentukan struktur pohon keputusan [17].

2.3 Ukuran Ketepatan Klasifikasi

Terdapat beberapa ukuran yang digunakan untuk menyatakan besar-kecilnya tingkat ketepatan klasifikasi. Setelah dilakukan prediksi tergadap semua amatan dalam sebuah gugus data yang telah diketahui kelasnya, maka dapat disusun sebuah tabel tabulasi silang antara kelas aktual dan kelas prediksi yang bentuknya disajikan pada Tabel 1. Tabel ini juga disebut matriks konfusi.

Tabel 1. Matriks Konfusi pada Klasifikasi 2 Kelas

	Kelas Prediksi	
Kelas Aktual	0	1
0	a	b
1	c	d

Jika *n* adalah banyaknya pengamatan yang digunakan dalam proses evaluasi ketepatan klasifikasi, maka jelas bahwa $a + b + c + d = n$ karena setiap pengamatan hanya bisa masuk ke satu dan hanya satu kategori jenis pengamatan yang ada pada matriks konfusi.

Matriks konfusi dapat digunakan untuk beberapa ukuran ketepatan klasifikasi, antara lain:

1) Akurasi (*accuracy*), yaitu persentase pengamatan yang tergolong tepat kalsifikasinya oleh model. Nilai akurasi ini diperoleh sebagai berikut:

$$akurasi = \frac{a+d}{n} \quad (3)$$

2) Sensitivitas (*sensitifity*), yaitu tingkat ketepatan prediksi pada pengamatan-pengamatan yang aktualnya berasal dari kelas 0, diperoleh menggunakan rumus:

$$sensitivitas = \frac{a}{a+b} \quad (4)$$

3) Spesifisitas (*specificity*), yaitu tingkat ketepatan prediksi pada pengamatan-pengamatan yang aktualnya berasal dari kelas 1. Nilai spesifisitas ini diperoleh dengan rumus:

$$spesifisitas = \frac{d}{c+d} \quad (5)$$

4) Presisi (*precision*), yaitu ukuran yang menyatakan persentase pengamatan-pengamatan yang diprediksi kelas 0 adalah benar-benar pengamatan kelas 0. Nilai presisi diperoleh menggunakan rumus:

$$presisi = \frac{a}{a+c} \quad (6)$$

Ukuran lainnya untuk mengukur ketepatan klasifikasi adalah *statistics area under the curve* (AUC). AUC bernilai tunggal untuk sebuah model klasifikasi. Sesuai dengan namanya AUC merupakan luas daerah di bawah sebuah kurva yang disebut sebagai ROC *curve* (kurva ROC, ROC = *receiver operating characteristic*). Kurva ROC adalah kurva yang dibentuk dengan menghubung-hubungka titik pada diagram pencar terhadap nilai sensitifitas pada sumbu vertical dan (1-spesifisitas) pada sumbu horizontal.

Semakin luas area di bawah kurva ROC, semakin baik kemampuan model dalam membedakan antara kelas positif dan negatif. Nilai AUC berada pada rentang 0 hingga 1, di mana nilai AUC mendekati 1 menunjukkan performa klasifikasi yang sangat baik, sedangkan nilai AUC mendekati 0,5 mengindikasikan bahwa model tidak lebih baik dari tebakan acak. Secara umum, AUC di atas 0,9 dianggap sangat baik, antara 0,8 hingga 0,9 dikategorikan baik, antara 0,7 hingga 0,8 cukup, dan di bawah 0,7 menunjukkan performa model yang kurang memadai [17].

2.4 Tingkat Kepentingan Variabel

Tingkat kepentingan variabel pada pohon keputusan berbasis ensembel dapat didasarkan pada perubahan nilai *gini*, ukuran ini disebut dengan *mean decrease gini* (MDG). Seandainya terdapat p variabel prediktor dengan $h = 1,2,\dots,q$, maka MDG mengukur tingkat kepentingan variabel prediktor x_h dengan rumus sebagai barikut [21]:

$$MDG_h = \frac{1}{k}\sum_t [d(h,t)I(h,t)] \quad (7)$$

Dengan,
k = banyaknya pohon keputusan dalam *random forest*
$d(h,t)$ = penurunan nilai *gini impurity* untuk variabel prediktor x_h pada simpul t
$I(h,t) = 1$ jika x_h memilah simpul t; 0 selainnya

Nilai MDG yang lebih tinggi menunjukkan bahwa suatu variabel memberikan kontribusi lebih besar terhadap pemisahan data yang meningkatkan homogenitas

dalam simpul-simpul pohon keputusan. Artinya, semakin besar penurunan gini impurity yang dihasilkan oleh suatu variabel pada banyak pohon dalam *random forest*, maka semakin penting peran variabel tersebut dalam proses klasifikasi. Sebaliknya, nilai MDG yang rendah mengindikasikan bahwa variabel tersebut jarang digunakan dalam pembelahan simpul atau tidak memberikan dampak yang signifikan terhadap pengurangan ketidakmurnian data. Oleh karena itu, MDG dapat digunakan sebagai dasar untuk seleksi fitur, dengan mempertimbangkan hanya variabel-variabel yang memiliki nilai kepentingan tinggi [17].

2.5 Diabetes

Diabetes mellitus adalah penyakit metabolik kronis yang ditandai dengan meningkatnya kadar glukosa dalam darah akibat gangguan pada produksi atau fungsi insulin. Penyakit ini dibagi menjadi beberapa tipe, yang paling umum adalah diabetes melitus tipe 1 dan tipe 2. Diabetes tipe 1 umumnya terjadi karena kerusakan autoimun pada sel beta pankreas yang memproduksi insulin, sedangkan diabetes tipe 2 disebabkan oleh resistensi insulin dan penurunan produksi insulin secara progresif. Menurut World Health Organization (WHO), prevalensi global diabetes meningkat secara signifikan dalam beberapa dekade terakhir, menjadikannya salah satu masalah kesehatan masyarakat utama di dunia [22].

Diabetes yang tidak dikelola dengan baik dapat menyebabkan berbagai komplikasi jangka panjang, termasuk penyakit jantung, gagal ginjal, gangguan penglihatan, dan kerusakan saraf. Faktor risiko utama diabetes tipe 2 meliputi obesitas, kurangnya aktivitas fisik, pola makan tidak sehat, serta faktor genetik. Pencegahan dan pengelolaan diabetes membutuhkan pendekatan komprehensif yang mencakup perubahan gaya hidup, pengendalian berat badan, dan pemeriksaan kesehatan secara berkala. Edukasi kepada masyarakat mengenai pentingnya deteksi dini dan manajemen penyakit sangat penting untuk menekan angka komplikasi dan meningkatkan kualitas hidup penderita diabetes [1].

3 DATA

Data yang digunakan dalam penelitian ini adalah data *Pima Indians Diabetes* dari *UCI Machine Learning Repository*. Dataset mencakup 768 pengamatan yang terdiri dari variabel prediktor dan variabel target. Variabel prediktor sebanyak 7 variabel, diantaranya glukosa plasma (*glucose*), tekanan darah diastolik (*blood pressure*), ketebalan kulit (*skin thickness*), kadar insulin (*insulin*), indeks massa tubuh (*Body Mass Index*), riwayat keluarga diabetes (*diabetes pedigree function*), dan usia (*age*). Variabel target adalah variabel dengan betuk data kategori berupa status penyakit diabetes. Deskripsi masing-masing variabel disajikan dalam Tabel 2.

Tabel 2. Deskripsi Variabel

No.	Variabel	Deskripsi
1	Glukosa Plasma	Kadar glukosa plasma 2 jam setelah pelaksanaan tes toleransi glukosa oral
2	Tekanan Darah Diastolik	Tekanan pada dinding arteri saat jantung berelaksasi (fase *diastole*), yang diukur dalam milimeter air raksa (mm Hg).
3	Ketebalan Kulit	Ukuran lemak bawah kulit di area trisep yang diukur dalam satuan milimeter (mm) sebagai indikator persentase lemak tubuh.
4	Kadar Insulin	Kadar insulin dalam serum darah yang diukur dua jam setelah konsumsi glukosa, dengan satuan mikro unit per mililiter (μU/mL).
5	Indeks Masa Tubuh	Indeks massa tubuh yang dihitung dari berat badan dalam kilogram dibagi kuadrat tinggi badan dalam meter (kg/m^2).
6	Riwayat Keluarga Diabates	Ukuran risiko genetik terhadap diabetes berdasarkan riwayat keluarga, dinyatakan dalam skala 0 hingga 1, di mana nilai lebih tinggi menunjukkan risiko yang lebih besar.
7	Usia	Usia responden yang dinyatakan dalam satuan tahun
8	Diabetes	Status penyakit diabetes dikategorikan ke dalam dua kelas, yaitu 0 untuk responden yang tidak terdiagnosis diabetes dan 1 untuk responden yang terdiagnosis diabetes.

4 HASIL DAN PEMBAHASAN

4.1 Statistika Deskriptif

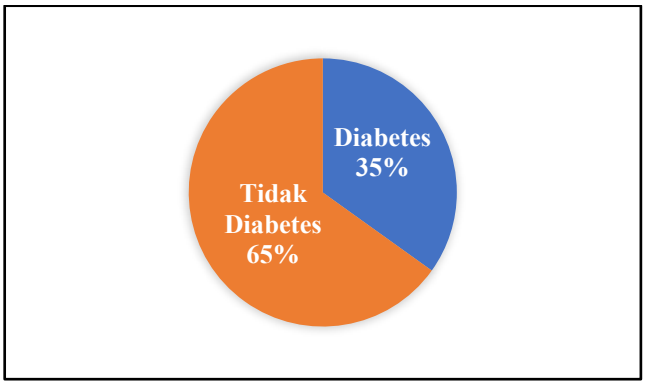
Statistika deskriptif digunakan untuk memberikan gambaran awal mengenai variabel prediktor dan variabel respon. Deskriptif varibel prediktor yang seluruhnya merupakan skala data rasio disajikan pada Tabel 3.

Tabel 3. Statistika Deskriptif Variabel Prediktor

Variabel Prediktor	Median	Rataan	Deviasi Standar
Glukosa Plasma	117,0	120,90	31,97
Tekanan Darah Diastolik	72,0	69,11	19,35
Ketebalan Kulit	23,0	20,54	15,95
Kadar Insulin	30,5	79,80	115,24
Indeks Masa Tubuh	32,0	31,99	7,88
Riwayat Keluarga Diabates	0,4	0,47	0,33
Usia	29,0	33,24	11,76

Tabel 3 menunjukkan bahwa variabel glukosa plasma memiliki nilai rataan tertinggi sebesar 120,90 dengan median 117,0 dan deviasi standar 31,97, menunjukkan kecenderungan hiperglikemia pada responden yang merupakan indikator utama diagnosis diabetes mellitus [22]. Kadar insulin memiliki deviasi standar tertinggi, yaitu 115,24 dengan rataan 79,80, mencerminkan tingkat variabilitas yang tinggi dan kemungkinan adanya gangguan metabolisme seperti resistensi insulin [23]. Sebaliknya, variabel riwayat keluarga diabetes menunjukkan variasi terendah dengan deviasi standar 0,33 dan rataan 0,47, mengindikasikan

mayoritas responden tidak memiliki riwayat genetik, meskipun faktor keturunan tetap menjadi prediktor penting [24]. Variabel indeks massa tubuh memiliki rataan senilai 31,99 menunjukkan prevalensi obesitas, yang merupakan faktor risiko signifikan untuk diabetes [25]. Usia responden berkisar luas dengan rataan 33,24, menunjukkan bahwa diabetes tidak hanya menyerang lansia, tetapi juga populasi usia produktif. Secara keseluruhan, temuan ini menunjukkan bahwa glukosa, insulin, dan faktor gaya hidup seperti obesitas berkontribusi besar dalam proses klasifikasi diabetes.



Gambar 1. Proporsi Kelas Variabel Respon

Berdasarkan visualisasi data dalam bentuk diagram lingkaran, diketahui bahwa proporsi pengamatan dengan status diabetes mencapai 35%, sedangkan sisanya, sebesar 65%, tidak mengalami kondisi tersebut. Temuan ini mengindikasikan bahwa meskipun mayoritas pengamatan dalam dataset tidak menderita diabetes, prevalensi kasus diabetes tetap cukup tinggi dan memerlukan perhatian khusus. Proporsi tersebut dapat mencerminkan pola risiko dalam populasi yang dianalisis, sehingga penting untuk dilakukan analisis lebih lanjut guna mengidentifikasi faktor-faktor utama yang berkontribusi terhadap kejadian diabetes serta merumuskan strategi intervensi yang efektif dalam upaya pencegahan dan pengendalian penyakit.

4.2 Data Latih dan Data Uji

Pada tahap awal penerapan metode *random forest*, dataset dibagi ke dalam dua subset utama, yaitu data uji dan data latih. Data latih digunakan untuk membangun model melalui proses pembelajaran algoritma *random forest*, sedangkan data uji digunakan sebagai dasar ukuran ketepatan klasifikasi dan prediksi, khususnya dalam mengukur akurasi dan kemampuan generalisasi.

Penelitian ini menerapkan skema pembagian data dengan proporsi 80:20, di mana 80% dari total data digunakan sebagai data latih dan 20% sisanya sebagai data uji. Proporsi ini juga digunakan dalam beberapa penelitian sebelumnya karena dianggap mampu memberikan keseimbangan antara kebutuhan pelatihan model dan pengujian, seperti yang ditunjukkan oleh penelitian Mahmuda, Nohe, dan

Leonardo [20] dan Mahmuda [26]. Dengan jumlah total pengamatan sebanyak 768, maka diperoleh:

$$data\ latih = \frac{80}{100} \times 768 = 614$$

$$data\ uji = \frac{20}{100} \times 768 = 154$$

Pemilahan data dilakukan secara acak guna menghindari bias dan memastikan representativitas sampel dalam proses pelatihan dan pengujian pada metode *random forest*.

4.3 Algoritma Random Forest

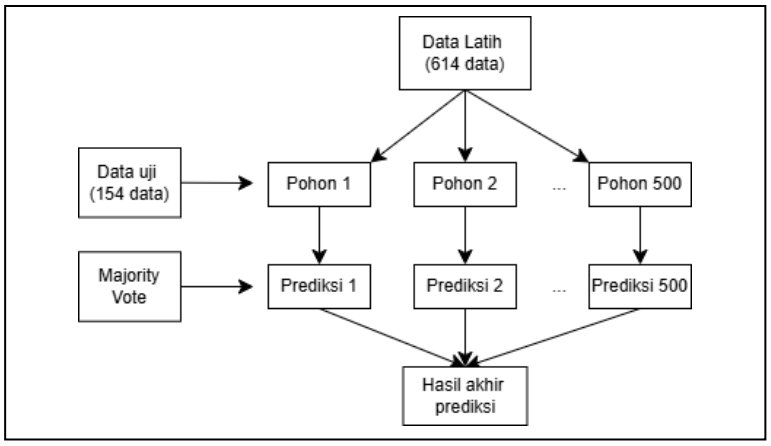
Proses pelatihan untuk membentuk model awal *random forest* dimulai dengan pemilihan data latih sebanyak 614 pengamatan yang diambil secara acak dari keseluruhan dataset. Data latih ini kemudian digunakan dalam tahap *bootstrapping*, yaitu proses pengambilan sampel ulang secara acak dengan pengembalian, yang dilakukan dengan bantuan komputasi untuk menghasilkan berbagai subset data pelatihan sebagai dasar pembentukan pohon keputusan.

Algoritma *random forest* dimulai dengan menetapkan beberapa nilai terlebih dahulu, yaitu jumlah pohon keputusan (B) yang akan dibangun dan jumlah variabel pemilah (m) yang dipertimbangkan pada setiap proses partisi simpul (*node*). Pada penelitian ini, ditetapkan jumlah pohon keputusan sebanyak B = 500, dengan m = 2 variabel yang secara acak dipilih pada setiap simpul dalam proses partisi. Penetapan nilai parameter tersebut didasarkan pada temuan studi yang dilakukan oleh Mahmuda, Nohe, dan Leonardo [20], yang menunjukkan bahwa dengan jumlah pohon keputusan yang dibetuk sebanyak 500 pohon dan penggunaan dua variabel prediktor untuk setiap partisi mampu menghasilkan tingkat kesalahan klasifikasi terendah.

Pada setiap simpul (*node*) pohon keputusan, proses pemilahan dilakukan berdasarkan nilai *gini decrease*. Nilai ini digunakan untuk menilai seberapa efektif suatu variabel dalam membagi data menjadi kelompok yang lebih homogen. Semakin besar penurunan nilai *gini*, semakin baik kualitas partisi yang dihasilkan. Dengan demikian, variabel dengan *gini decrease* terbesar akan dipilih sebagai pemilah pada simpul tersebut. Prinsip ini menjadi landasan dalam pembentukan struktur pohon keputusan, yang secara matematis dirumuskan dalam Persamaan (2).

Setelah seluruh proses pembentukan pohon selesai, terbentuk sebanyak 500 pohon keputusan yang secara bersama-sama menyusun model *random forest*. Pohon-pohon keputusan tersebut digunakan sebagai model untuk melakukan proses pengujian dengan data uji. Data uji yang berjumlah 154 diprediksi melalui model pohon keputusan 1 hingga pohon ke-500. Setiap pohon memberikan prediksi terhadap kelas masing-masing data uji (kelas tidak diabetes atau kelas diabetes), kemudian melalui *majority vote* ditentukan hasil akhir kategori kelas. Kelas yang

memperoleh suara terbanyak dari seluruh pohon akan menjadi hasil prediksi akhir untuk masing-masing data.



Gambar 2. Ilustrasi Algoritma Random Forest

Gambar 2 menujukkan ilustrasi algoritma *random forest* pada penelitian ini. Apabila mayoritas pohon dari model *random forest* memprediksi bahwa data uji termasuk dalam kategori diabetes, maka kesimpulan akhirnya adalah data tersebut diklasifikasikan sebagai diabetes. Sebaliknya, jika lebih banyak pohon memprediksi kelas bukan diabetes, maka data tersebut dikategorikan sebagai tidak diabetes.

4.4 Ukuran Ketepatan Klasifikasi

Setelah proses pengujian dilakukan menggunakan data uji, diperoleh hasil prediksi yang bisa ditabulasikan menggunakan matriks konfusi.

Tabel 3. Matriks Konfusi Prediksi Data Uji

		Kelas Prediksi		Total
		Tidak Diabetes	Diabetes	
Kelas Aktual	Tidak Diabates	93	14	107
	Diabetes	17	30	47
	Total	110	44	154

Ukuran ketepatan klasifikasi yang dapat diperoleh berdasarkan matriks konfusi pada Tabel 3, yaitu berupa nilai akurasi (persamaan 3), sensitivitas (persamaan 4), dan nilai spesifisitas (persamaan 5) yang dirumuskan sebagai berikut:

$$akurasi = \frac{93 + 30}{154} \times 100\% = 79{,}87\%$$

$$sensitivitas = \frac{93}{93 + 17} \times 100\% = 84{,}54\%$$

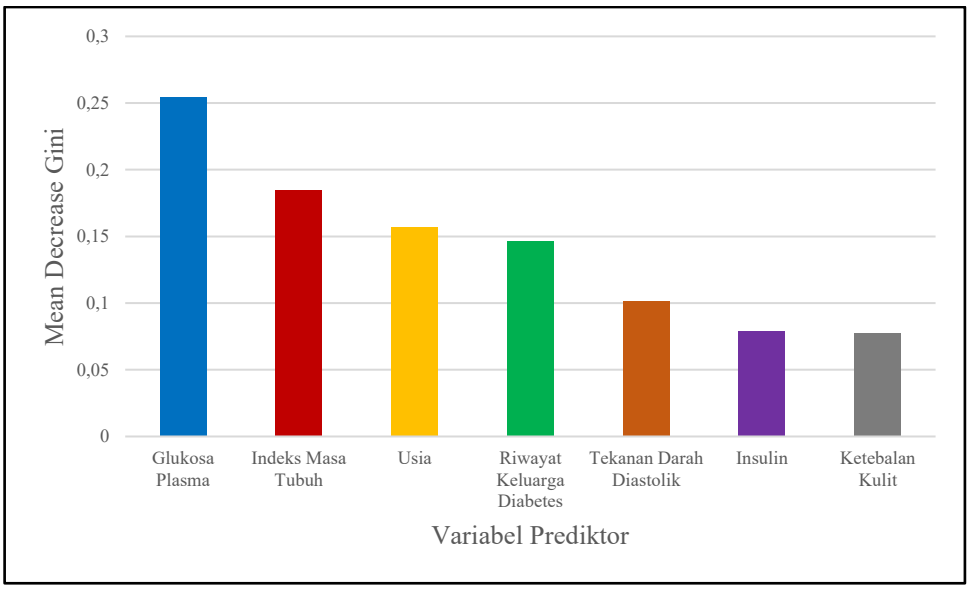
$$spesifisitas = \frac{30}{14 + 30} \times 100\% = 68{,}18\%$$

Nilai akurasi sebesar 79,87% menunjukkan persentase pengamatan yang tergolong tepat dikasifikasikan oleh model, baik untuk kelas diabetes maupun tidak diabetes. Sedangkan, tingkat ketepatan klasifikasi pada pengamatan-pengamatan yang aktualnya berasal dari kelas tidak diabetes dilihat dari nilai sensitivitas, yaitu sebesar 84,54%. Pengamatan-pengamatan yang aktualnya berasal dari kelas diabetes mampu diklasifikasikan 68,18% dengan tepat sesuai kelasnya, hal ini didasarkan pada nilai spesifisitas.

Berdasarkan hasil pengujian menggunakan data uji dan analisis matriks konfusi, model *random forest* menunjukkan performa yang cukup baik dalam mengklasifikasikan data secara umum. Kemampuan model dalam mengenali pengamatan yang bukan penderita diabetes tergolong tinggi. Namun demikian, ketepatan model dalam mengidentifikasi pengamatan yang merupakan penderita diabetes masih perlu ditingkatkan. Hal ini mengindikasikan bahwa meskipun model sudah cukup efektif secara keseluruhan, optimalisasi lebih lanjut diperlukan agar dapat memberikan hasil prediksi yang lebih seimbang antara kedua kelas.

4.5 Tingkat Kepentingan Variabel

Tingkat kepentingan variabel pada pohon keputusan berbasis ensembel dapat didasarkan pada perubahan nilai *gini*, ukuran ini disebut dengan *mean decrease gini* (MDG). Gambar 3 menunjukkan nilai MDG dari masing-masing variabel prediktor.



Gambar 3. Nilai Mean Decrease Gini Variabel Prediktor

Nilai MDG tertinggi yaitu pda variabel prediktor glukosa plasma, dengan MDG sebesar 0,2542, yang berarti variabel ini memberikan kontribusi paling besar dalam proses klasifikasi diabetes. Disusul oleh variabel indeks masa tubuh dengan MDG sebesar 0,1848 dan usia dengan MDG sebesar 0,1573, keduanya juga memiliki pengaruh yang cukup signifikan terhadap hasil prediksi model. Ketiga variabel ini dapat dianggap sebagai fitur utama yang memengaruhi keputusan

model dalam membedakan antara individu yang menderita dan tidak menderita diabetes.

Sementara itu, variabel seperti riwayat keluarga diabetes memiliki MDG sebesar 0,1468, tekanan darah diastolik sebesar 0,1013, insulin sebesar 0,0789, dan ketebalan kulit sebesar 0,0778. Nilai-nilai tersebut menunjukkan bahwa kontribusi variabel-variabel ini terhadap akurasi klasifikasi lebih kecil dibandingkan dengan variabel lainnya. Oleh karena itu, informasi MDG ini dapat dimanfaatkan untuk mengidentifikasi variabel-variabel penting dalam pemodelan, serta membantu dalam penyederhanaan model tanpa mengorbankan kinerja secara signifikan.

Berdasarkan hasil tersebut, dapat disimpulkan bahwa glukosa plasma, indeks massa tubuh, dan usia merupakan variabel yang paling berpengaruh dalam prediksi diabetes, sehingga perlu mendapatkan perhatian utama dalam proses analisis maupun pengambilan keputusan berbasis data.

5 KESIMPULAN

Berdasarkan hasil yang diperoleh dalam penelitian ini, dapat disimpulkan bahwa:

1) Klasifikasi dan prediksi dengan metode *random forest* pada studi kasus penyakit diabetes menunjukkan ketepatan yang cukup baik, hal ini dapat dilihat dari nilai akurasi sebesar 79,87%, spesifisitas sebesar 68,18%, dan sensitivitas sebesar 84,54%.
2) Variabel glukosa plasma menempati posisi teratas dalam hal kontribusi terhadap klasifikasi dan prediksi diabetes dengan nilai *mean decrease gini* (MDG) sebesar 0,25. Disusul oleh variabel indeks masa tubuh yang memiliki nilai MDG sebesar 0,19 dan usia sebesar 0,17. Ketiga variabel ini menjadi indikator yang paling berpengaruh dalam memprediksi status diabetes seseorang.

DAFTAR PUSTAKA

[1] International Diabetes Federation. (2025). *IDF diabetes atlas (11th ed.)*. International Diabetes Federation. https://diabetesatlas.org.
[2] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324.
[3] Tan, P.-N., Steinbach, M., & Kumar, V. (2019). *Introduction to data mining* (2nd ed.). Pearson. https://doi.org/10.5555/3208440.
[4] Musa, J., & Abdulazeez, A. M. (2024). A review on diabetes classification based on machine learning algorithms. *Indonesian Journal of Computer Science*, *13*(2). https://doi.org/10.33022/ijcs.v13i2.3886.
[5] Wang, S. (2023). Diabetes prediction using random forest in healthcare. *Highlights in Science, Engineering and Technology*. https://doi.org/10.54097/5ndh9a05.
[6] Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22. http://CRAN.R-project.org/doc/Rnews/.

[7] Chen, X., & Ishwaran, H. (2012). Random forests for variable selection in classification problems in bioinformatics. *BMC Bioinformatics, 13*, 91. https://doi.org/10.1186/1471-2105-13-91.

[8] Suwaryo, N., Rahman, A., Atmaja, D. M. U., & Basri, A. (2023). Prediksi penyakit diabetes untuk pencegahan dini dengan metode regresi linear. *Bulletin of Information Technology (BIT)*, *4*(2), 313–319. https://doi.org/10.47065/bit.v3i1.739.

[9] Aditya, M. F., Pramuntadi, A., Wijaya, D. P., & Wicaksono, Y. (2024). Implementation of decision tree method for diabetes mellitus type 2 prediction [Implementasi metode decision tree pada prediksi penyakit diabetes melitus tipe 2]. *MALCOM: Indonesian Journal of Machine Learning and Computer Science, 4*(3), 1104–1110. https://doi.org/10.57152/malcom.v4i3.1284.

[10] Dwyer, K., & Holte, R. (2007). Decision tree instability and active learning. In J. N. Kok, J. Koronacki, R. L. de Mántaras, S. Matwin, D. Mladenič, & A. Skowron (Eds.), *Machine Learning: ECML 2007. Lecture Notes in Computer Science* (Vol. 4701, pp. 128–139). Springer. https://doi.org/10.1007/978-3-540-74958-5_15.

[11] Aji, P. W. S., Suprianto, & Dijaya, R. (2023). Prediksi penyakit stroke menggunakan metode Random Forest. *KESATRIA: Jurnal Penerapan Sistem Informasi (Komputer & Manajemen)*, *4*(4), 916–924. https://doi.org/10.30645/kesatria.v4i4.242.g240.

[12] Putra, M. R. A., & Handayani, R. N. (2023). Perbandingan algoritma Decision Tree dan Random Forest dalam pengklasifikasian penyakit tiroid. *e-Proceeding Sistem Informasi*, *5*(2), 166–172. https://eprosiding.ars.ac.id/index.php/psi/article/view/1164.

[13] Sriyanto, & Supriyatna, A. R. (2023). Prediksi penyakit diabetes menggunakan algoritma Random Forest. *Teknika*, *17*(1), 163–172. https://doi.org/10.5281/zenodo.8051410.

[14] Wulansari, M. J. (2018). *Analisis faktor-faktor yang mempengaruhi seseorang terkena penyakit diabetes melitus menggunakan regresi Random Forest* (Skripsi tidak dipublikasikan). Universitas Islam Indonesia. https://dspace.uii.ac.id/handle/123456789/8015.

[15] Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random forest algorithm overview. *Babylonian Journal of Machine Learning, 2024*, 69–79. https://doi.org/10.58496/BJML/2024/007.

[16] Chen, Y., Cheung, K. C., Sun, R. Z., & et al. (2024). A user guide of CART and random forests with applications in FinTech and InsurTech. *Japanese Journal of Statistics and Data Science, 7*, 999–1038. https://doi.org/10.1007/s42081-024-00258-x.

[17] Sartono, B., & Dharmawan, H. (2023). *Pemodelan prediksi berbasis pohon klasifikasi*. IPB Press.

[18] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7.

[19] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* (2nd ed.). Springer. https://doi.org/10.1007/978-1-0716-1418-1.

[20] Mahmuda, S., Nohe, D. A., & Leonardo, A. M. (2024). Classification of the human development index in Kalimantan using random forest method. In *Proceedings of the International Seminar on Science and Technology* (pp. 231–239). https://doi.org/10.33830/isst.v3i1.2283.

[21] Sandri, M., & Zuccolotto, P. (2008). A bias correction algorithm for the Gini variable importance measure in classification trees. *Journal of Computational and Graphical Statistics, 17*(3), 611–628. https://doi.org/10.1198/106186008X344522.

[22] World Health Organization. (2023). *Diabetes*. https://www.who.int/news-room/fact-sheets/detail/diabetes.

[23] American Diabetes Association. (2022). Standards of medical care in diabetes—2022. *Diabetes Care, 45*(Supplement_1), S1–S264. https://doi.org/10.2337/dc22-Sint.

[24] Lyssenko, V., & Laakso, M. (2013). Genetic screening for the risk of type 2 diabetes: Worthless or valuable? *Diabetes Care, 36*(Suppl 2), S120–S126. https://doi.org/10.2337/dcS13-2011.

[25] Ng, M., Fleming, T., Robinson, M., Thompson, B., Graetz, N., Margono, C., ... Gakidou, E. (2014). Global, regional, and national prevalence of overweight and obesity in children and adults during 1980–2013: A systematic analysis. *The Lancet, 384*(9945), 766–781. https://doi.org/10.1016/S0140-6736(14)60460-8.

[26] Mahmuda, S. (2024). Implementasi metode Random Forest pada kategori konten kanal YouTube. *Jurnal Jendela Matematika, 2*(1), 21–31. https://www.ejournal.jendelaedukasi.id/index.php/JJM