P-ISSN: 1979-5661
E-ISSN: 2622-321X

Prediksi Dini Penyakit Diabetes pada Wanita dengan Algoritma Klasifikasi Berbasis Pohon

Zidan Ali Zaqi[a1], Mochammad Anshori[a2], Wahyu Teja Kusuma[a3]

[a]Informatika, Institut Teknologi, Sains, dan Kesehatan RS.DR. Soepraoen Kesdam V/BRW
Jl. S. Supriadi No.22, Sukun, Kec. Sukun, Kota Malang, Jawa Timur, Indonesia
[1]zidanalizaqi51@gmail.com
[2]moanshori@itsk-soepraoen.ac.id
[3]wtkusuma@itsk-soepraoen.ac.id

Abstrak

Salah satu penyakit kronis yang terus meningkat didunia adalah Diabetes. Berdasarkan data dari Pusat Data dan Informasi Kementerian Kesehatan Republik Indonesia dan Badan Federasi Diabetes Dunia pada tahun 2020, Indonesia berada di peringkat kelima dalam jumlah penderita diabetes terbanyak terutama pada wanita. Diagnosis dini menjadi penting untuk mengurangi risiko komplikasi serius, namun sering kali terhambat oleh keterlambatan identifikasi gejala. Dalam mendukung diagnosis dini diabetes pada wanita, diusulkan dengan Algoritma klasifikasi berbasis pohon yaitu Decision Tree, Random Forest dan XGBoost. Decision Tree bekerja dengan membentuk struktur pohon untuk menentukan keputusan berdasarkan atribut data. Random Forest merupakan pengembangan dari Decision Tree yang membangun banyak pohon untuk meningkatkan akurasi dan mengurangi overfitting. Sedangkan XGBoost merupakan algoritma gradient boosting yang melakukan proses pembelajaran secara berulang untuk meminimalkan kesalahan prediksi. Normalisasi dengan min-max diterapkan pada data dan Stratified cross validation dengan k=10-fold diterapkan saat pembuatan model. Hasil penelitian menunjukkan bahwa XGBoost memiliki performa terbaik dengan akurasi, presisi, recall, dan F1-Score sebesar 0,992, diikuti oleh Random Forest dan Decision Tree dengan nilai evaluasi sebesar 0,991. Penelitian ini mengindikasikan bahwa XGBoost memiliki keunggulan dalam akurasi prediksi, ditunjukkan pada confusion matrix dengan tingkat ketepatan prediksi masing-masing sebesar 99,5% dan 98,7%. Kesalahan klasifikasi yang terjadi sangat kecil, menjadikannya solusi yang andal untuk mendukung sistem diagnosis dini diabetes. Berdasarkan hasil temuan ini, XGBoost terbukti lebih handal digunakan dalam prediksi dini diabetes pada Wanita dibandingkan algoritma pembanding lainnya, yaitu Decision Tree dan Random Forest. Studi ini memberikan kontribusi signifikan terhadap teknologi prediktif berbasis data Kesehatan dengan potensi penerapan yang lebih luas.

Kata Kunci: *Klasifikasi, Decision Tree, Random Forest, XGBoost, Diabetes*

1. Pendahuluan

Diabetes adalah suatu gangguan metabolisme dalam tubuh yang terjadi akibat tingginya kadar gula darah yang bertahan dalam jangka waktu lama[1]. Berdasarkan data dari Pusat Data dan Informasi Kementerian Kesehatan Republik Indonesia dan Badan Federasi Diabetes Dunia pada tahun 2020, tercatat sebanyak 463 juta orang di dunia berusia 20 hingga 79 tahun mengidap diabetes[2]. Dari jumlah tersebut, wanita menjadi penyumbang terbesar dengan angka mencapai 199 juta jiwa dengan rasio rasio 2 dari 5 wanita menderita diabetes[3]. Wanita memiliki risiko lebih tinggi untuk mengidap diabetes karena secara fisik cenderung lebih mudah mengalami peningkatan indeks massa tubuh. Secara umum, Jumlah penderita diabetes diperkirakan akan meningkat menjadi 578 juta orang pada tahun 2030 dan terus bertambah hingga mencapai 700 juta orang pada tahun 2045[3]. Dan diperkirakan jumlah wanita yang menderita diabetes diprediksi akan meningkat signifikan hingga mencapai 313 juta jiwa pada tahun 2040[3]. Indonesia menempati peringkat kelima di antara sepuluh negara dengan jumlah penderita diabetes pada wanita tertinggi[4][5].

Pengetahuan awal tentang diabetes pada wanita sangat penting untuk mendukung upaya penanganan, pemulihan, dan pencegahan penyakit ini[6]. Salah satu kendala yang sering terjadi adalah keterlambatan diagnosis, yang dapat menyebabkan penanganan kurang optimal dan

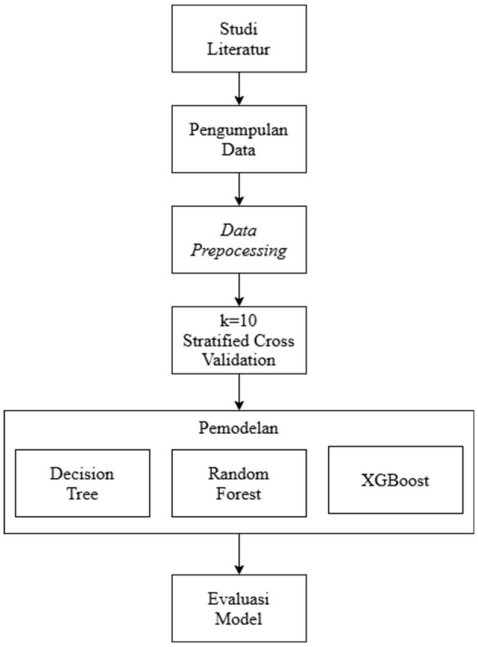
P-ISSN: 1979-5661
E-ISSN: 2622-321X

meningkatkan risiko komplikasi yang lebih serius[7]. Gejala awal diabetes sering kali tidak terlihat jelas, sehingga banyak penderita tahap awal menyadari penyakit ini pada tahap lanjut. Rendahnya kesadaran terhadap risiko dan tanda-tanda awal diabetes memperburuk kondisi penderita, yang akhirnya berisiko mengalami komplikasi berat seperti mengalami kesulitan proses persalinan, penyakit kardiovaskular, dan gangguan reproduksi bahkan menjadi penyebab kematian[8]. Oleh karena itu, diperlukan pendekatan diagnosis dini yang efektif untuk mendeteksi diabetes lebih awal dan memungkinkan pengobatan yang cepat serta tepat sasaran.

Pendekatan diagnosis dini diabetes pada wanita dapat dilakukan dengan memanfaatkan sistem berbasis pembelajaran mesin. Salah satu fungsi pembelajaran mesin adalah klasifikasi data, di mana metode seperti *Decision Tree, Random Forest, dan XGBoost* merupakan pilihan populer dalam *tree-based classification*. *Random Forest* dikenal luas karena kesederhanaan dan efektivitasnya, dengan fitur unggulan seperti pengukuran pentingnya variabel dan evaluasi kesalahan, sehingga dianggap cukup baik untuk melakukan klasifikasi karena akurasi yang dihasilkan lebih tinggi dibandingkan algoritma Decision Tree[9]. Disisi lain, *Decision Tree* merupakan algoritma yang sederhana namun kuat dalam mengklasifikasi dan memprediksi data, dengan kemampuan memetakan data ke dalam struktur pohon keputusan yang mudah dipahami secara visual, sehingga mampu memetakan berbagai kondisi menjadi cabang dalam sebuah struktur pohon keputusan[10][11]. Sementara itu, *XGBoost* adalah algoritma klasifikasi yang mengembangkan model prediksi secara bertingkat dengan menambahkan pohon keputusan satu per satu secara berurutan guna menawarkan akurasi tinggi melalui pendekatan berbasis rangkaian pohon keputusan, dengan optimasi agresif untuk meminimalkan kesalahan dan mencegah *overfitting*[12]. Selain itu, *XGBoost* juga dilengkapi fitur evaluasi pentingnya fitur dalam model prediksi[13].

Penelitian ini bertujuan mengevaluasi keefektifan ketiga metode klasifikasi berbasis pohon untuk menemukan model prediksi yang paling optimal. Faktor yang terdiri akurasi, presisi, *F1 Score* dan *recall* akan menjadi pertimbangan utama dalam mengukur performa algoritma, sehingga dapat mendukung pengembangan sistem diagnosis dini diabetes pada Wanita yang lebih akurat dan andal. Diharapkan hasil penelitian ini dapat memberikan kontribusi dalam meningkatkan kesadaran masyarakat tentang pentingnya prediksi dini penyakit diabetes khususnya pada wanita. Sehingga memungkinkan penanganan yang lebih cepat dan tepat serta mengurangi dampak komplikasi serius bagi penderita.

2. Metode Penelitian



Gambar 1. Alur Penelitian

Metode penelitian yang digunakan pada penelitian ditunjukkan pada Gambar 1. Berdasarkan pada gambar, terdapat beberapa tahap utama, yaitu studi literatur, pengumpulan data, *data preprocessing, stratified cross validation*, pemodelan prediksi dengan algoritma pembelajaran mesin berbasis pohon, dan terakhir adalah evaluasi model. Tahapan secara lebih mendalam akan dijelaskan berikut.

2.1 Studi Literatur

Studi literatur dilakukan untuk mencari penelitian sebelumnya yang membahas implementasi algoritma klasifikasi pohon, seperti *Decision Tree, Random Forest, dan XGBoost*, dalam memprediksi diabetes. Proses ini bertujuan menggali berbagai teori yang relevan dengan topik penelitian sebagai acuan dalam pembahasan. Kegiatan ini meliputi pengumpulan data dari berbagai sumber referensi, membaca dan mencatat informasi penting, serta mengolah bahan yang akan digunakan dalam penelitian.

2.2 Pengumpulan Data

Pada pengumpulan data, penelitian ini menggunakan dataset yang diperoleh dari platform Mendeley Data yang berjudul Diabetes[14]. Dataset berisi prediksi risiko diabetes pada wanita dirilis pada 21 Agustus 2024 dan berisi 768 sampel penyakit diabetes dengan berbagai atribut kesehatan yang dikumpulkan melalui prosedur ketat dan melibatkan izin dengan ahli medis. Data tersebut disediakan dalam format CSV dan bersifat akses terbuka. Data yang didapat terdapat dua buah kelas, yaitu kelas 0 dan kelas 1. Kelas 0 menandakan pasien tidak terindikasi diabetes, sedangkan kelas 1 yang menandakan pasien terindikasi diabetes.

2.3 Data Prepocessing

Tahap *data preprocessing* dalam penelitian ini dilakukan dengan menerapkan normalisasi *Min-Max* untuk memastikan nilai setiap atribut berada dalam rentang tertentu, biasanya antara 0 dan 1[15]. Teknik ini digunakan untuk mengurangi skala perbedaan antar atribut sehingga algoritma dapat bekerja lebih optimal dalam memproses data. Normalisasi ini penting untuk menghindari dominasi atribut dengan nilai besar terhadap atribut lainnya dalam proses pembelajaran algoritma[16]. Dengan rentang nilai yang seragam dan rendah, hal ini menjadikannya kelebihan tersendiri karena dapat meningkatkan waktu komputasi saat model melakukan proses pembelajaran [17]. Normalisasi *Min-Max* dapat dijabarkan dalam persamaan (1).

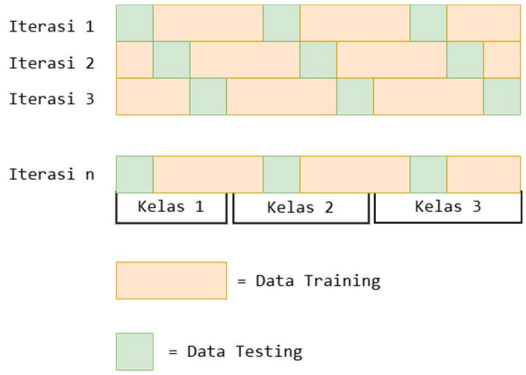
$$x' = \frac{x_i - \min(x)}{\max(x) - \min(x)} \quad (1)$$

Persamaan (1) menjelaskan mengubah setiap nilai data xi menjadi nilai baru *x'* dengan cara mengurangi nilai minimum dari dataset *min(x)* dari nilai data tersebut, kemudian membaginya dengan selisih antara nilai maksimum dan nilai minimum dalam dataset. Hasilnya adalah nilai *x'* yang berada dalam rentang 0 hingga 1, sehingga memungkinkan perbandingan yang lebih seimbang antar nilai data dengan skala yang berbeda.

2.4 Pemodelan dengan *Stratified Cross Validation*

Pengolahan model dilakukan dengan menggabungkan teknik *Stratified Cross Validation* dengan nilai k=10, yang bertujuan untuk mengevaluasi performa model secara sistematis dengan membagi data ke dalam beberapa *subset* secara berulang[16]. *Stratified Cross Validation* memungkinkan penilaian yang lebih akurat terhadap dampak nilai parameter yang berbeda pada kinerja model. Sehingga menghasilkan model yang lebih kuat dan dapat digeneralisasi. Selain itu, *Stratified Cross Validation* diterapkan untuk mempertahankan distribusi label yang seimbang di setiap *fold*. Hasilnya adalah evaluasi model lebih representatif, khususnya dalam mengatasi masalah dataset yang tidak seimbang dan mencegah *overfitting*[18].

P-ISSN: 1979-5661
E-ISSN: 2622-321X



Gambar 2. Ilustrasi *Stratified Cross Validation*

Ilustrasi *Stratified Cross Validation* ditunjukkan pada Gambar 2 yang menjelaskan bahwa setiap iterasi dalam *stratified cross-validation* mempertahankan proporsi kelas yang sama seperti pada dataset asli. Dengan kata lain, memungkinkan model dilatih dan diuji pada sampel yang representatif untuk setiap kelas, sehingga dapat meminimalkan bias dan meningkatkan performa secara keseluruhan[18].

2.5 *Decision Tree*

Pembuatan model prediksi penderita penyakit diabetes menggunakan algoritma *Decision Tree. Decision Tree* adalah metode klasifikasi yang sering digunakan dalam prediksi data. *Decision Tree* memiliki konsep seperti halnya pohon yang memiliki *node* yang berisi *data testing*, dan cabang yang berisi hasil dari *data testing* serta *leaf node* yang berisi kelas[19][20]. *Decision Tree* sangat membantu dalam pengambilan dan analisis informasi dari dataset yang sesuai berdasarkan nilai *gain*. *Node* dengan nilai *gain* terbesar akan berkembang[21]. *Node* yang memiliki atribut tidak akan dihitung lagi. Dan proses akan berhenti jika *leaf node* ditemukan.

GenDecTree(Sample S, Features F)
Steps:

1. ***If****stopping_condition(S, F) = true* ***then***
   a. *Leaf = createNode()*
   b. *leafLabel = classify(s)*
   c. ***return*** *leaf*
2. *root = createNode()*
3. *root.test_condition =findBestSpilt(S,F)*
4. *V = {v | v a possible outcomecfroot.test_condition}*
5. ***For each*** *value v Є V:*
   a. S_v = *{s | root.test_condition(s) = v and s Є S };*
   b. *Child = TreeGrowth (*S_v*,*F*);*
   c. *Add child as descent of root and label the edge {root → child} as v*
6. ***return*** *root*

Gambar 3. *Pseudocode Decision Tree*

Gambar 3 adalah pseudocode dari model *Decision Tree*. Proses dimulai dengan memeriksa apakah kondisi penghentian telah terpenuhi; jika ya, maka simpul daun dibuat dan diberikan label klasifikasi. Jika tidak, simpul akar dibentuk, kemudian kondisi pemisahan yang ditentukan berdasarkan dataset dan fitur yang tersedia. Selanjutnya, untuk setiap kemungkinan nilai hasil

dari kondisi pemisahan, dataset dibagi menjadi subset yang sesuai, dan algoritma dijalankan secara rekursif untuk membangun subpohon. Akhirnya, setiap simpul anak ditambahkan sebagai turunan dari simpul akar, dengan setiap cabang diberi label sesuai dengan nilai yang diwakilinya. Proses ini berlanjut hingga seluruh pohon keputusan terbentuk[22].

2.6 *Random Forest*

Metode kedua yang digunakan dalam prediksi dini penyakit diabetes adalah *Random Forest. Random Forest* adalah teknik dalam klasifikasi yang menggabungkan hasil dari banyak *Decision Tree* untuk meningkatkan akurasi prediksi. *Random Forest* dikenal sebagai "hutan" karena menghasilkan banyak pohon keputusan yang dibentuk berdasarkan data dan fitur yang dipilih secara acak[23]. Proses diulang untuk setiap atribut hingga seluruh *instance* dalam data memiliki kelas yang seragam[24].

Input: *N - Quantitative amount of bootstrap samples*
M - Total number of features
m - Sample size
k - Next node
Output: A Random Forest (RF)
Steps:
1. ***Creates*** *N bootstrap samples from the dataset.*
2. Every node (sample) takings a feature randomly of size m where m<M.
3. Builds a split for the m features selected in Step 2 and detects the k node by using the best split point.
4. Split the tree iteratively until one leaf node is attained and the tree remains completed.
5. The algorithm is trained on each bootstrapped independently.
6. Using trees classification voting predicted data is collected from the trained trees (n).
7. The final RF model is build using the peak voted features.
8. ***return*** *RF*
End.

Gambar 4. *Pseudocode Random Forest*

Gambar 4 menjelaskan *Pseudocode Random Forest*. Algoritma dimulai dengan membuat sejumlah *bootstrap samples* dari dataset. Selanjutnya, pada setiap node, fitur dipilih secara acak dengan ukuran yang lebih kecil dari jumlah total fitur. Kemudian, pemisahan dilakukan berdasarkan fitur yang telah dipilih dengan menentukan node terbaik. Proses pemisahan berlangsung secara iteratif hingga setiap cabang mencapai simpul daun. Setelah itu, setiap pohon hasil dilatih secara independen. Prediksi dilakukan dengan metode pemungutan suara dari seluruh pohon yang telah dilatih. Akhirnya, model *Random Forest* akhir dibentuk berdasarkan fitur dengan suara terbanyak, yang kemudian digunakan untuk prediksi[25].

2.7 *XGBoost*

Metode ketiga yang digunakan dalam prediksi diabetes adalah *XGBoost. XGBoost* atau *Extreme Gradient Boosting* adalah metode *tree boosting* yang efektif dan banyak digunakan dalam *machine learning* untuk memprediksi nilai data. Algoritma ini menggunakan pohon keputusan sebagai model dasar dan menerapkan teknik *boosting* untuk meningkatkan kinerja model secara keseluruhan. *XGBoost* dilengkapi fitur seperti regularisasi, pemrosesan paralel, dan penanganan nilai yang hilang[26]. Selain itu, *XGBoost* dirancang dengan memperhatikan efisiensi akses *cache,* kompresi data, dan *sharding* agar mampu bekerja secara efisien pada data skala besar. Dengan kombinasi teknik, *XGBoost* dapat mengolah data besar dengan sumber daya yang lebih minimal dibandingkan metode tradisional. Berbeda dengan model *boosting* tradisional yang hanya memanfaatkan informasi turunan pertama*, XGBoost* melakukan ekspansi *Taylor orde* kedua pada *loss fuction*. Hal ini memungkinkan penggunaan *multithreading CPU* untuk komputasi paralel. *XGBoost* juga menerapkan berbagai metode untuk mengurangi risiko *overfitting*[26][27].

Input:
$F_{\text{normalized}}$ ($X_1^{\text{norm}}, .., X_n^{\text{norm}}$: set of financial ratios)

Output: F_{optimal} : Set of selected financial ratios
Step 1: Load the normalized financial ratios
Step 2: Create an empty dictionary E to save the scores of financial ratios
Step 3: Instantiate an Extreme Gradient Boosting Classifier as EGB
Step 4: fit EGB
Step 4: Generate Fls
Step 5: Determine the Fl threshold FI_h
Step 6:
for t from $F_{\text{normalized}}$ **do**
if (FI (f^i) ≥ Fl_h **then**
add FI (f^i) into E
end if
end for
Step 5: Utilize the scores in E to produce F_{optimal}

Gambar 5. *Pseudocode* Model *XGBoost*

Pseudocode XGBoost dapat dilihat pada Gambar 5. Dimulai dengan memuat data yang telah dinormalisasi, lalu membuat wadah kosong untuk menyimpan skor kepentingan fitur. Setelah itu, model *XGBoost* dikonfigurasi dan dilatih menggunakan data tersebut. Setelah model terbentuk, skor kepentingan setiap fitur dihitung. Kemudian, fitur yang memiliki skor di atas ambang batas tertentu dipilih dan disimpan. Terakhir, fitur yang terpilih digunakan untuk membangun set fitur optimal yang siap digunakan untuk analisis atau prediksi lebih lanjut[28].

2.8 Evaluasi

Berikutnya adalah evaluasi model. Pada tahap ini dilakukan eksperimen dan mengevaluasi model *Decision Tree, Random Forest dan XGBoost*. Penilaian performa model dilihat dari evaluasinya, yaitu akurasi, presisi, *recall* dan *F1 score*. Akurasi, rasio antara total data yang diprediksi benar dari total semua data; presisi, rasio antara total data yang diprediksi positif dan total data yang diprediksi benar; *recall* yakni rasio antara total data yang sebenarnya positif dengan total data yang diprediksi benar; dan *F1-Score,* yaitu kombinasi perhitungan dari *recall* dan presisi[29]. F1-score menunjukkan bahwa model mampu mendeteksi sebagian besar *instance* yang relevan, namun tidak memperhitungkan kesalahan yang disebabkan oleh *false positives*. Secara berurutan formula dari akurasi, presisi*, recall* dan *F1 score* ditunjukkan pada persamaan (2), (3), (4) dan (5).

$$\text{Akurasi} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (2)$$

$$\text{Presisi} = \frac{\text{TP}}{\text{TP+FP}} \quad (3)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (4)$$

$$\text{F1-Score} = 2 * \frac{\text{Presisi*Recall}}{\text{Presisi+Recall}} \quad (5)$$

Dimana TP atau *True Positive* mengacu pada kondisi di mana hasil prediksi dan nilai aktual sama-sama positif. TN atau *True Negative* adalah situasi ketika model memprediksi negatif sesuai dengan nilai sebenarnya. FP atau *False Positive* terjadi ketika model memprediksi positif namun bertolak belakang dengan nilai sebenarnya. Sementara itu, FN atau *False Negative* terjadi saat hasil prediksi menunjukkan negatif, padahal nilai aktualnya positif.

3. Hasil dan Pembahasan

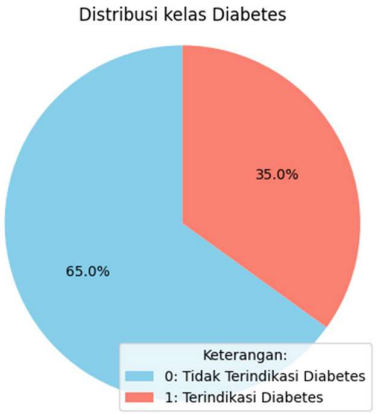
Hasil dan pembahasan dari penelitian yang dilakukan akan dipaparkan pada tahap ini. Mula-mula data yang akan digunakan dalam penelitian ini diperoleh dari Mendeley Data mengenai Penyakit Diabetes pada Wanita. Diketahui pada dataset terdapat 2251 data dan 9 fitur.

P-ISSN: 1979-5661
E-ISSN: 2622-321X

Tabel 1. Fitur Dataset

Fitur	Tipe Data	Rentang Nilai
Pregnancies	Numerikal	0 – 17
Glucose	Numerikal	0 – 200
Blood Pressure	Numerikal	0 – 122
Skin Thickness	Numerikal	0 – 110
Insulin	Numerikal	0 – 744
Body Mass Index	Numerikal	0 – 8.6
Diabetes Predigree Function	Numerikal	0.078 – 2.42
Age	Numerikal	21 – 81
Outcome	Numerikal	0 – 1

Tabel 1 menyajikan informasi tentang jumlah fitur, tipe data, dan rentang nilai dari dataset. Berdasarkan tabel, data yang digunakan mencakup berbagai variabel yang relevan untuk analisis diabetes. Semua fitur dalam dataset memiliki tipe data *numerikal*, sehingga tidak diperlukan proses *encoding* untuk mengubah data *kategorikal* menjadi *numerikal*. Untuk memahami lebih lanjut karakteristik setiap fitur dalam dataset, berikut dijelaskan masing-masing variabel beserta rentang nilainya secara rinci. *Pregnancies* menunjukkan jumlah kehamilan pasien dengan rentang nilai 0 – 17. *Glucose*, yang mencerminkan konsentrasi glukosa plasma (mg/dL) setelah 2 jam tes toleransi glukosa *oral*, memiliki rentang nilai 0 – 200. *Blood Pressure*, yang menunjukkan tekanan darah *diastolik* (mm Hg), memiliki rentang nilai 0 – 122. *Skin Thickness*, yang mengukur ketebalan lipatan kulit *triceps* (mm), memiliki rentang nilai 0 – 110. *Insulin*, yang merepresentasikan kadar serum insulin selama 2 jam (mu U/ml), memiliki rentang nilai 0 – 744. *Body Mass Index* (BMI), atau indeks massa tubuh yang dihitung dari berat badan (kg) dibagi tinggi badan kuadrat (m²), memiliki rentang nilai 0.0 – 80.6. *Diabetes Pedigree Function*, yang menghitung kemungkinan diabetes berdasarkan riwayat keluarga, memiliki rentang nilai 0.078 – 2.42. *Age*, yang menunjukkan usia pasien (tahun), memiliki rentang nilai 21 – 81. Terakhir, fitur *Outcome*, yang mengindikasikan pasien didiagnosis diabetes (1) atau tidak (0), memiliki rentang nilai 0 – 1. Data ini memberikan gambaran komprehensif tentang variabel yang relevan untuk analisis diabetes.



Gambar 6. Distribusi Kelas Data

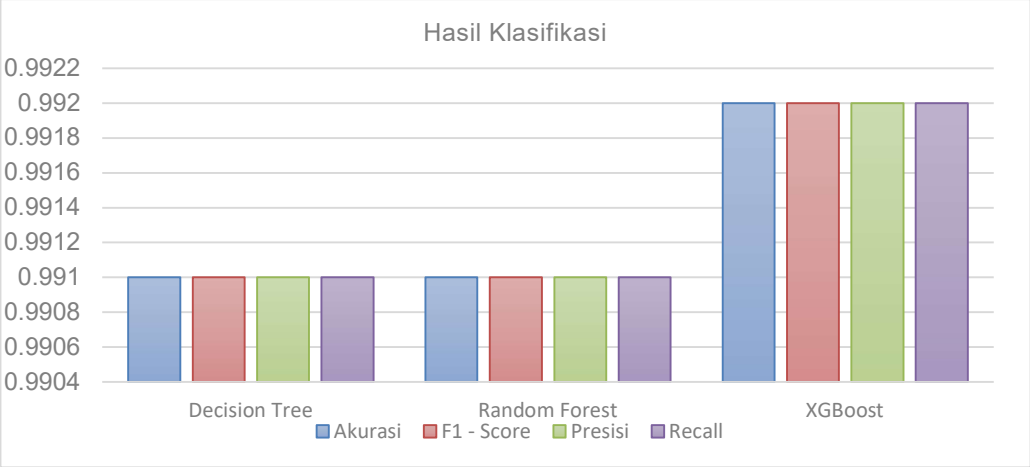
Gambar 6 menggambarkan distribusi atau persentase kelas dari dataset penelitian terkait diabetes. Tujuan visualisasi distribusi kelas adalah untuk menunjukkan proporsi pasien yang terindikasi dan tidak terindikasi diabetes dalam suatu kelompok data. Sebanyak 65 persen individu yang diteliti yaitu 1.464 orang tidak menunjukkan gejala diabetes berdasarkan kriteria

P-ISSN: 1979-5661
E-ISSN: 2622-321X

yang digunakan dalam dataset, sementara 35 persen lainnya yaitu 787 orang menunjukkan gejala diabetes.

Selanjutnya Proses *Preprocessing Data* dilakukan menggunakan normalisasi *Min-Max* untuk memastikan bahwa semua fitur memiliki skala yang sama dan berada dalam rentang nilai tertentu, yaitu antara 0 hingga 1. Pendekatan ini digunakan untuk mengurangi bias yang mungkin terjadi akibat perbedaan rentang nilai antar fitur. Sehingga algoritma klasifikasi dapat lebih optimal dalam mempelajari pola dari data. Normalisasi *Min-Max* sangat penting terutama ketika fitur memiliki rentang nilai yang sangat berbeda, karena algoritma seperti *Random Forest* dan *XGBoost* dapat lebih sensitif terhadap data yang tidak ternormalisasi.

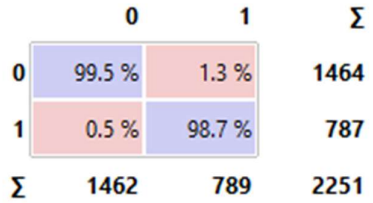
Setelah dilakukan normalisasi, kemudian dilakukan pembuatan model klasifikasi yang berbasis pohon. Pada proses pembuatan modelnya, *stratified cross validation* diterapkan disini. Nilai *k-fold* yang digunakan adalah k=10. Penerapan *Stratified Cross Validation* dengan nilai k = 10 memastikan bahwa evaluasi dilakukan secara lebih representatif dalam kinerja model untuk mendeteksi risiko diabetes secara dini. Berikut adalah evaluasi hasil pembuatan model klasifikasi yang berbasis pohon.



Gambar 7. Hasil Pemodelan Kombinasi *Stratified Cross Validation*

Gambar 7 menyajikan hasil evaluasi dari tiga model berbeda yang digunakan dalam suatu proses pembelajaran mesin dalam konteks klasifikasi. Ketiga model yang dievaluasi adalah *Random Forest, Decision Tree, dan XGBoost. Random Forest* digunakan untuk membangun model prediksi dini penyakit diabetes, dengan evaluasi kinerja yang dilakukan melalui *Stratified Cross Validation* dengan nilai k = 10. Selain itu, nilai akurasi klasifikasi, *F1-Score*, presisi, dan *recall* semuanya mencapai 0,991. Selanjutnya *decision tree, Decision tree* diterapkan untuk memprediksi dini penyakit diabetes memperoleh hasil evaluasi dengan nilai akurasi klasifikasi, *F1-Score*, presisi, dan *recall* sebesar 0,991. Serta *XGBoost* menunjukkan hasil yang konsisten dengan nilai akurasi klasifikasi, *F1-Score*, presisi, dan *recall* yang masing-masing mencapai 0,992. Berdasarkan hasil evaluasi, dapat disimpulkan bahwa *XGBoost* adalah model yang paling baik di antara ketiga model yang diuji. Hal ini didasarkan pada nilai akurasi klasifikasi, *F1-Score*, presisi, dan *recall* yang mencapai 0,992, lebih unggul daripada model *Random Forest* dan *Decision Tree*, yang masing-masing memperoleh nilai 0,991 untuk metrik yang sama. Kinerja algoritma *XGBoost* membuktikan performa yang lebih baik dalam mengatasi data yang kompleks, memastikan prediksi yang lebih akurat dan andal dalam konteks klasifikasi dini penyakit diabetes pada wanita. Untuk mendukung analisis lebih lanjut, dilakukan evaluasi melalui *confusion matrix* pada model *XGBoost* seperti yang tertera pada Gambar 8.

P-ISSN: 1979-5661
E-ISSN: 2622-321X



Gambar 8. *Confusion Matrix XGBoost*

Gambar 8 menunjukkan *confusion matrix* dari model *XGBoost*, di mana *True Negative* (TN) atau prediksi tidak terindikasi diabetes yang benar sebesar 99,5 persen, *False Positive* (FP) atau prediksi terindikasi diabetes yang salah sebesar 1,3 persen, *False Negative* (FN) atau prediksi tidak terindikasi diabetes yang salah sebesar 0,5 persen, dan *True Positive* (TP) atau prediksi terindikasi diabetes yang benar sebesar 98,7 persen. Hasil ini menunjukkan *XGBoost* memiliki performa yang sangat baik dengan dominasi nilai *True Positive* dan *True Negative* yang tinggi serta tingkat kesalahan yang sangat kecil dalam prediksi dini penyakit diabetes.

4. Kesimpulan

Penelitian ini meneliti tentang diabetes pada wanita. Metode yang diusulkan dengan algoritma berbasis pohon, yaitu *Decision Tree, Random Forest,* dan *XGBoost*. Proses pelatihannya menerapkan *preprocessing data* menggunakan normalisasi *min-max* dilanjutkan *stratified cross validation* dengan nilai *k=10-fold*. Hasil penelitian menunjukkan bahwa *XGBoost* memiliki performa terbaik dengan evaluasi sebesar 0,992 untuk akurasi, *F1-Score*, presisi, dan *recall*. Sedangkan *Decision Tree* dan *Random Forest* menghasilkan evaluasi yang lebih rendah, dengan nilai 0.991 untuk setiap metrik evaluasinya. Berdasarkan hasil tersebut, *XGBoost* terbukti lebih unggul dalam hal ketepatan prediksi. Dibuktikan dengan ketepatan prediksi benar dan salah pada *confusion matrix* sebesar 99.5% dan 98.7%, sedangkan sisanya adalah kesalahan dalam klasifikasi yang sangat minor. Hal ini membuktikan bahwa *XGBoost* menghasilkan model prediksi yang lebih baik dari pada algoritma pembanding lainnya yang berbasis pohon, yaitu *Decision Tree* dan *Random Forest*. Penelitian ini mengindikasikan bahwa *XGBoost* dapat diterapkan sebagai model yang handal dalam mendukung diagnosis dini penyakit diabetes pada wanita. Sehingga dapat diaplikasikan untuk pengambilan keputusan medis dan dapat memberikan tindakan yang tepat. Untuk penelitian selanjutnya, disarankan untuk menerapkan teknik reduksi dimensi, guna mengurangi kompleksitas data yang tinggi. Sehingga dapat membantu mengurangi kompleksitas dan mempercepat waktu proses pelatihan model.

References

[1] R. P. Kurniadi, R. R. Saedudin, and V. P. Widartha, “Perbandingan Akurasi Algoritma K-Nearest Neighbor Dan Logistic Regression Untuk Klasifikasi Penyakit Diabetes,” *e-Proceeding Eng.*, vol. 8, no. 5, pp. 9757–9764, 2021.

[2] M. K. Nasution, R. R. Saedudin, and V. P. Widartha, “Perbandingan Akurasi Algoritma Naïve Bayes Dan Algoritma Xgboost Pada Klasifikasi Penyakit Diabetes,” *e-Proceeding Eng.*, vol. 8, no. 5, pp. 9765–9772, 2021, [Online]. Available: https://journal.ubpkarawang.ac.id/mahasiswa/index.php/ssj/article/view/424/338%0Ahttps://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/15759

[3] M. Faisal Fahrul and W. Hadikurniawati, “KLASIFIKASI DIABETES PADA WANITA MENGGUNAKAN METODE NAIVE BAYES CLASSIFIER,” *J. Ilm. Inform.*, vol. 10, no. 01, pp. 70–73, Mar. 2022, doi: 10.33884/jif.v10i01.4705.

[4] S. T. Siridion and B. Siregar, “ANALISIS KLASIFIKASI DIAGNOSA PENYAKIT DIABETES MELITUS BERDASARKAN KOMPARASI ALGORITMA SUPERVISED LEARNING Universitas Matana , Banten , Indonesia Email : sherly.taurin@student.matanauniversity.ac.id Analisi,” vol. 2, no. 3, pp. 1006–1014, 2024.

[5] B. R. Pramananditya, “PERANCANGAN APLIKASI SISTEM PAKAR PENYAKIT DIABETES MELLITUS MENGGUNAKAN ALGORITMA RANDOM FOREST,” *J. Ilmu Komput.*, vol. 17, no. 2, p. 8, Sep. 2024, doi: 10.24843/JIK.2024.v17.i02.p05.

[6] A. W. Mucholladin, F. A. Bachtiar, and M. T. Furqon, “Klasifikasi Penyakit Diabetes menggunakan Metode Support Vector Machine,” *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 2, pp. 622–633, 2021, [Online]. Available: http://j-ptiik.ub.ac.id

[7] R. P. Fadhillah, R. Rahma, A. Sepharni, R. Mufidah, B. N. Sari, and A. Pangestu, “Klasifikasi Penyakit Diabetes Mellitus Berdasarkan Faktor-Faktor Penyebab Diabetes menggunakan Algoritma C4.5,” *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 7, no. 4, pp. 1265–1270, 2022, doi: 10.29100/jipi.v7i4.3248.

[8] Fatmawati and N. A. K. Rifai, “Klasifikasi Penyakit Diabetes Retinopati Menggunakan Support Vector Machine dengan Algoritma Grid Search Cross-validation,” *J. Ris. Stat.*, pp. 79–86, 2023, doi: 10.29313/jrs.v3i1.1945.

[9] D. P. Sinambela, H. Naparin, M. Zulfadhilah, and N. Hidayah, “Implementasi Algoritma Decision Tree dan Random Forest dalam Prediksi Perdarahan Pascasalin,” *J. Inf. dan Teknol.*, vol. 5, no. 3, pp. 58–64, 2023, doi: 10.60083/jidt.v5i3.393.

[10] D. Septhya *et al.*, “Implementasi Algoritma Decision Tree dan Support Vector Machine untuk Klasifikasi Penyakit Kanker Paru,” *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 3, no. 1, pp. 15–19, 2023, doi: 10.57152/malcom.v3i1.591.

[11] A. Naseh Khudori and M. Syauqi Haris, “Implementasi Decision tree Untuk Prediksi Kanker Paru-Paru,” *J. Ris. Sist. Inf. Dan Tek. Inform. (JURASIK*, vol. 9, no. 1, pp. 94–106, 2024, [Online]. Available: https://tunasbangsa.ac.id/ejurnal/index.php/jurasik

[12] M. Salsabil, N. Lutvi, and A. Eviyanti, “Implementasi Data Mining Dalam Melakukan Prediksi Penyakit Diabetes Menggunakan Metode Random Forest Dan Xgboost,” *J. Ilm. Komputasi*, vol. 23, no. 1, pp. 51–58, 2024, doi: 10.32409/jikstik.23.1.3507.

[13] B. A. Candra Permana and I. K. Dewi Patwari, “Komparasi Metode Klasifikasi Data Mining Decision Tree dan Naïve Bayes Untuk Prediksi Penyakit Diabetes,” *Infotek J. Inform. dan Teknol.*, vol. 4, no. 1, pp. 63–69, 2021, doi: 10.29408/jit.v4i1.2994.

[14] S. Balaji, “Diabetes,” 2024, *Mendeley Data*. doi: 10.17632/pd64hfttwy.1.

[15] A. Oktaviana, D. P. Wijaya, A. Pramuntadi, and D. Heksaputra, “Prediksi Penyakit Diabetes Melitus Tipe 2 Menggunakan Algoritma K-Nearest Neighbor (K-NN),” *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 3, pp. 812–818, 2024, doi: 10.57152/malcom.v4i3.1268.

[16] I. Permana and F. N. S. Salisah, “Pengaruh Normalisasi Data Terhadap Performa Hasil Klasifikasi Algoritma Backpropagation,” *Indones. J. Inform. Res. Softw. Eng.*, vol. 2, no. 1, pp. 67–72, 2022, doi: 10.57152/ijirse.v2i1.311.

[17] H. Prasetyo Wibowo, M. Anshori, and M. Syauqi Haris, “the Discriminant Analysis Function Was Implemented To Predict the Presence of Diabetes,” *J. Enhanc. Stud. Informatics Comput. Appl.*, vol. 1, no. 2, pp. 47–55, 2024, doi: 10.47794/jesica.v1i2.10.

[18] M. A. I. Hutagalung and Sutarman, “Penalized Maximum Likelihood Estimation dengan Algoritma Gradient descent pada Model Regresi Logistik Multinomial,” *IJM Indones. J. Multidiscip.*, vol. 2, no. 6 SE-Articles, pp. 673–683, 2024, [Online]. Available: http://journal.csspublishing.com/index.php/ijm/article/view/951

[19] N. Nuradha, A. riski Ramadani, H. Hazriani, and Y. Yuyun, “Penerapan Algoritma C4. 5 dalam Mengidentifikasi Karakteristik Pasien Beresiko Diabetes,” *Pros. SISFOTEK*, pp. 325–331, 2023, [Online]. Available: http://seminar.iaii.or.id/index.php/SISFOTEK/article/view/412

[20] S. A. Pratiwi, A. Fauzi, S. Arum, P. Lestari, and Y. Cahyana, “KLIK: Kajian Ilmiah Informatika dan Komputer Prediksi Persediaan Obat Pada Apotek Menggunakan Algoritma Decision Tree,” *Media Online*, vol. 4, no. 4, pp. 2381–2388, 2024, doi: 10.30865/klik.v4i4.1681.

[21] A. Arista, “Comparison Decision Tree and Logistic Regression Machine Learning Classification Algorithms to determine Covid-19,” *Sinkron*, vol. 7, no. 1, pp. 59–65, 2022, doi: 10.33395/sinkron.v7i1.11243.

[22] E. Demirović *et al.*, “MurTree: Optimal Decision Trees via Dynamic Programming and Search,” *J. Mach. Learn. Res.*, vol. 23, pp. 1–47, 2022, [Online]. Available: http://jmlr.org/papers/v23/20-520.html

[23] Suci Amaliah, M. Nusrang, and A. Aswi, “Penerapan Metode Random Forest Untuk Klasifikasi Varian Minuman Kopi di Kedai Kopi Konijiwa Bantaeng,” *VARIANSI J. Stat. Its Appl. Teach. Res.*, vol. 4, no. 3, pp. 121–127, 2022, doi: 10.35580/variansiunm31.

[24] P. Handayani and A. Charis Fauzan, “KLIK: Kajian Ilmiah Informatika dan Komputer

Machine Learning Klasifikasi Status Gizi Balita Menggunakan Algoritma Random Forest," *Media Online)*, vol. 4, no. 6, pp. 3064–3072, 2024, doi: 10.30865/klik.v4i6.1909.

[25] M. Hambali, T. Oladele, A. S., A. Kumar, and W. Gao, "Feature selection and computational optimization in high-dimensional microarray cancer datasets via InfoGain-modified bat algorithm," *Multimed. Tools Appl.*, vol. 81, 2022, doi: 10.1007/s11042-022-13532-5.

[26] G. Abdurrahman, H. Oktavianto, and M. Sintawati, "Optimasi Algoritma XGBoost Classifier Menggunakan Hyperparameter Gridesearch dan Random Search Pada Klasifikasi Penyakit Diabetes," *INFORMAL Informatics J.*, vol. 7, no. 3, p. 193, 2022, doi: 10.19184/isj.v7i3.35441.

[27] M. A. Rayadin, M. Musaruddin, R. A. Saputra, and I. Isnawaty, "Implementasi Ensemble Learning Metode XGBoost dan Random Forest untuk Prediksi Waktu Penggantian Baterai Aki," *BIOS J. Teknol. Inf. dan Rekayasa Komput.*, vol. 5, no. 2, pp. 111–119, 2024.

[28] S. Smiti, M. Soui, and K. Ghedira, "Tri-XGBoost model improved by BLSmote-ENN: an interpretable semi-supervised approach for addressing bankruptcy prediction," *Knowl. Inf. Syst.*, vol. 66, pp. 1–38, 2024, doi: 10.1007/s10115-024-02067-w.

[29] A. M. Argina, "Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes," *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 29–33, 2020, doi: 10.33096/ijodas.v1i2.11.