JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 1, Maret 2025, Pp. 150-162

ANALISIS PERBANDINGAN METODE LOGISTIC REGRESSION, RANDOM FOREST, GRADIENT BOOSTING UNTUK PREDIKSI DIABETES

Nanda Hendra Setyawan*[1)], Nur Wakhidah [2)]
1. Universitas Semarang, Semarang, Indonesia
2. Universitas Semarang, Semarang, Indonesia

Article Info
Kata Kunci: Gradient Boosting; Kesehatan; Logistic Regression; Perbandingan; Random Forest

Keywords: Comparison; Gradient Boosting; Health; Logistic Regression; Random Forest

Article history:
Received 12 Oktober 2024
Revised 19 November 2024
Accepted 1 Maret 2025
Available online 1 Maret 2025

DOI :
https://doi.org/10.29100/jipi.v10i1.5743

* Corresponding author.
Corresponding Author
E-mail address:
nandahendrasetyawan@gmail.com

ABSTRAK
Salah satu penyakit yang paling umum pada manusia adalah diabetes. Hampir 350 juta orang menderita diabetes dan jumlah kematian akibat penyakit ini meningkat setiap tahun. Karena kurangnya pengetahuan dasar tentang diabetes, banyak orang awalnya tidak tahu mereka mengidap diabetes. Saat ini, metode untuk mendeteksi diabetes menggunakan tes laboratorium, yang memakan waktu yang lama. Salah satu metode yang dapat digunakan untuk mengembangkan sistem untuk memprediksi diabetes sejak dini adalah data mining dengan prinsip analitik. Dalam penelitian ini menggunakan tiga metode Logistic Regression, Random Forest, Gradient Boosting untuk membedakan karakteristik pemodelan dan mengetahui metode mana yang tepat untuk prediksi penyakit diabetes dengan cara membuat perbandingan antara ketiga metode tersebut. Tujuan penelitian ini adalah untuk membandingkan nilai akurasi terbaik dari tiga metode Logistic Regression, Random Forest, Gradient Boosting dalam memprediksi diabetes. Model ini kemudian dievakuasi menggunakan metrik kinerja seperti presisi, perolehan, dan skor F1 untuk membandingkan efektivitas masing-masing metode dalam memprediksi diabetes. Berdasarkan pengujian ketiga metode tersebut kita bisa tau bahwa metode Random Forest dengan accuracy 77%, precision 74%, recall 83%, dan F1 score 79%.

ABSTRACT
Diabetes is the most common human disease. Every year the number of deaths from this disease increases significantly with nearly 350 million people suffering from diabetes. Many people initially do not know they have diabetes due to lack of basic knowledge about diabetes. The current method of diabetes detection is by using laboratory tests and this method takes a long time. To detect diabetes early, a system can be developed to predict the disease using various methods, one of the methods that can be used is the data mining method with analytic principles. In this study using three methods Logistic Regression, Random Forest, Gradient Boosting to distinguish modeling characteristics and find out which method is appropriate for predicting diabetes by making a comparison between the three methods. The purpose of this study is to compare the best accuracy value of the three methods of Logistic Regression, Random Forest, Gradient Boosting in predicting diabetes. The model is then evacuated using performance metrics such as precision, recall, and F1 score to compare the effectiveness of each method in predicting diabetes. Based on the testing of the three methods, we can know that the Random Forest method with accuracy 77%, precision 74%, recall 83%, and F1 score 79%.

.

I. PENDAHULUAN

DIABETES merupakan suatu kelainan metabolisme parah pada tubuh manusia yang ditandai dengan tingginya kadar gula darah dan disertai gangguan metabolisme karbohidrat, lipid, dan protein akibat ketidakmampuan insulin untuk melakukan fungsinya secara penuh.[1]

Diabetes adalah penyakit yang sangat umum pada manusia. Jumlah kematian akibat penyakit ini meningkat secara signifikan setiap tahun. Organisasi Kesehatan Dunia (WHO) mengatakan bahwa hampir 350 juta orang mengalami diabetes. Karena menyediakan sumber vital bagi sel dan jaringan, glukosa sangat penting untuk

kesehatan manusia. Penyakit serius seperti diabetes, penyakit ginjal, stroke, penyakit mata, dan penyakit jantung dapat muncul jika tidak ditangani dengan baik.[2]

Karena kurangnya pengetahuan dasar tentang diabetes, banyak orang pertama kali tidak tahu mereka mengidap diabetes. Saat ini, pendekatan untuk mendeteksi diabetes menggunakan pemeriksaan lab seperti gula darah dan toleransi glukosa oral, yang memerlukan waktu lama. Pendekatan konvensional untuk deteksi dan manajemen diabetes termasuk biaya tinggi untuk perawatan medis, obat-obatan, dan perangkat monitoring, yang dapat menjadi tantangan bagi mereka yang tidak memiliki asuransi Kesehatan yang memadai. Selain itu, manajemen diabetes memerlukan pengukuran glukosa darah, pengaturan diet yang sehat, dan konsultasi medis teratur, yang dapat menantang bagi orang yang memiliki jadwal yang padat atau tinggal di daerah terpencil. Aksesibilitas ke perawatan kesehatan yang baik juga dapat menjadi masalah, terutama bagi mereka yang tinggal di daerah pedesaan atau di negara berkembang dimana layanan kesehatan mungkin tidak tersedia atau tidak terjangkau secara finansial. Namun, suatu metode untuk menganalisis penyakit diabetes dapat dikembangkan dengan menggunakan data mining dengan prinsip analitik.[3]

Diabetes tipe 1 dan 2 adalah dua macam yang paling sering ditemui. a) Diabetes tipe 1 disebabkan sama kekurangan produksi insulin; penderita diabetes tipe 1 memerlukan insulin melalui suntikan atau pompa insulin, dan mereka sering mengalami rasa pengen minum yang lebih besar (polidipsia), buang air kecil yang jauh lebih sering (poliuria), dan rasa pengen makan yang lebih besar (kolik). b) Diabetes tipe 2 disebabkan oleh resistensi insulin, adalah ketika sel tidak berhasil menggunakan insulin dengan baik. Pengobatan dengan cara diet, olahraga, obat oral, atau salah satunya. Diabetes tipe 2 mungkin saja tidak memiliki gejala sama sekali. Pengukuran glukosa darah naik dapat membantu diagnosis atau memerlukan pengukuran tambahan. Jika diabetes tidak diobati dengan cepat, itu berdampak buruk pada kesehatan seseorang. Dampak yang paling berbahaya adalah kemungkinan komplikasi kesehatan jangka panjang yang dapat muncul sebagai akibat dari diabetes yang tidak terkontrol dengan baik. Misal, jika retinopati diabetic tidak diobati dan diobati segera, dapat menyebabkan kerusakan mata yang permanen dan bahkan kebutaan. Kerusakan saraf, yang menyebabkan kehilangan sensasi atau nyeri jangka panjang, terutama pada kaki dan tangan, dapat disebabkan oleh neuropati diabetic. Selain itu diabetes yang tidak dikontrol meningkatkan resiko stroke, serangan jantung, gagal ginjal, dan masalah sirkulasi, yang dapat menyebabkan luka yang sulit sembuh dan bahkan amputasi. Oleh karena itu, untuk mencegah atau mengurangi kemungkinan komplikasi ini, deteksi dini dan manajemen yang baik sangat penting. Pembelajaran mesin muncul sebagai hasil dari kemajuan dalam teknologi informasi dan komunikasi, terkhususnya dalam bidang kecerdasan buatan.[4]

Salah satu karakteristik utama kecerdasan buatan adalah pembelajaran mesin, yang membantu dalam pengembangan sistem komputer yang dapat belajar dari pengalaman sebelumnya tanpa memerlukan pemrograman kasus per kasus. Dalam kondisi saat ini, pembelajaran mesin dianggap sebagai kebutuhan untuk mendukung otomatisasi dan mengurangi ketidaknyamanan. Ini memerlukan model prediksi yang menggunakan pembelajaran mesin dan teknik data mining untuk memprediksi kemungkinan diabetes.[5]

Penelitian yang dilakukan Rony, dkk. Mengenai perbandingan metode random forest, regresi logistik, naïve bayes, dan multilayer perceptron dalam klasifikasi biaya kuliah tunggal, penelitian ini membandingkan 4 model pembelajaran mesin, dan mencakup 9 variabel. Model random forest memiliki akurasi tertinggi sebesar 97,9% sedangkan regresi logistic memiliki rata-rata akurasi sebesar 89,68%. Penelitian sebelumnya ini memiliki tingkat akurasinya tinggi, seleksi metode yang digunakan ini tidak hanya berdasarkan rerata akurasi maupun karakteristik fungsi rerata akurasi dalam penerapannya jug dapat mempertimbangakan waktu komputasi, distribusi UKT, dan nilai ekspetasi UKT dari metode yang dipilih, maka dari itu penelitian ini akan ada perbedaan mengenai objek dan variabelnya .[6]

Selanjutnya penelitian sebelumnya yang dilakukan oleh Ahmad Maulid Ridwan, dkk. Tentang perbandingan berbagai model machine learning untuk mendeteksi diabetes. Pada penelitian sebelumnya ini membandingkan 7 model pembelajaran mesin dan memiliki 9 variabel dalam data tersebut. Terdapat model rando forest memberikan hasil yang baik dengan akurasi 79% dan Adapun keterbatasannya Memiliki ukuran dataset yang kurang maka dari itu penelitian ini memiliki banyak dataset supaya lebih stabil dan menggambarkan lebih baik kinerja model.[7]

Kemudian penelitian sebelumnya yang dilakukan oleh Muhammad Salsabil, dkk. Yaitu tentang Implementasi data mining dalam melakukan prediksi penyakit diabetes menggunakan metode random forest dan Xgboost. Penelitian ini memiliki 8 variabel/atribut, penelitian sebelumnya melihatkan tingkat akurasi keseluruhan dalam penggunaan random forest sebesar 76%, penelitian ini belum cukup terakurasi karena hanya memiliki dua motode dan sebab itu dipenelitian ini akan manambah menjadi tiga metode.[8]

Penelitian terdahulu selanjutnya dilakukan oleh Cecep Wahyu, dkk. Dengan penelitian tentang analisis performa logistik regression, naïve bayes, dan random forest sebagai algoritma pendeteksi kanker payudara. Penelitian ini memiliki 9 variabel, dan melihatkan akurasi logistik regression data testing sebesar 80% dan akurasi data training sebesar 76%, akurasi random forest memperoleh hasil akurasi data testing sebesar 70% dan akurasi data training sebesar 100%. Namun saya memilih beberapa metode yang belum dilakukan atau digunakan di penelitian ini seperti

Gradient Boosting dan meiliki perbedaan objeknya.[9]

Kemudian penelitian terdahulu selanjutnya dilakukan Sahat Pandapotan Nainggolan, dkk. Tentang comparative analysis of accuracy of random forest and gradient boosting classifier algorithm for diabetes classification. Dengan memiliki 9 variabel kemudian melihatkan akurasi gradient boosting sebesar 81% sedangkan random forest sebesar 79%. Dengan penelitian sebelumnya ini saya bisa mempertimbangkan dengan beberapa metode yang berbeda, secara mendalam parameter yang dihasilkan menggunakan kedua algoritma sebelumnya dan menambahkan satu algoritma lainnya.[10]

Dalam penelitian ini menggunkan tiga metode yaitu metode logistic regression, random forest, gradient forest untuk untuk membedakan karakteristik pemodelan dan mengetahui metode mana yang tepat untuk prediksi penyakit diabetes dengan cara membuat perbandingan antara ketiga metode tersebut. Meskipun sederhana, logistic regression sangat membantu dalam memprediksi diabetes karna memberikan gambaran yang jelas tentang hubungan antara variabel input dan kemungkinan terjadinya diabetes. Dalam situasi ini dimana asumsi linieritas antara variabel target dan predictor masih kuat, ini akan sesuai. Random Forest dapat menangani interaksi variabel yang kompleks dan toleran terhadap overfitting. Sementara Gradient Boosting, dengan pendekatan yang mengoptimalkan model secara bertahap, cenderung memberikan prediksi diabetes yang lebih akurat, bahkan dalam dataset yang besar dan kompleks, dengan menemukan pola yang kompleks diantara variabel input. Dengan demikian, sambal mempertimbangkan asumsi model dan kompleksitas dataset, pemilihan model untuk prediksi diabetes akan tergantung pada kebutuhan interpretabilitas, ketangguhan terhadap overexposure, dan kebutuhan untuk mengidentifikasi pada pola yang rumit. Dari ketiga metode tersebut dibandingkan untuk mencari skor akurasi yang terbaik.[11]

Penelitian ini bertujuan untuk membandingkan nilai akurasi terbaik dari tiga pendekatan untuk memprediksi diabetes: Logistic Regression, Random Forest, dan Gradient Boosting. Data yang dikumpulkan dari pasien dengan berbagai gejala dan riwayat kesehatan digunakan dalam penelitian ini. Selanjutnya, model ini dievakuasi menggunakan metrik kinerja seperti skor F1, presisi, dan perolehan untuk membandingkan seberapa efektif masing-masing pendekatan dalam memprediksi diabetes.[12]

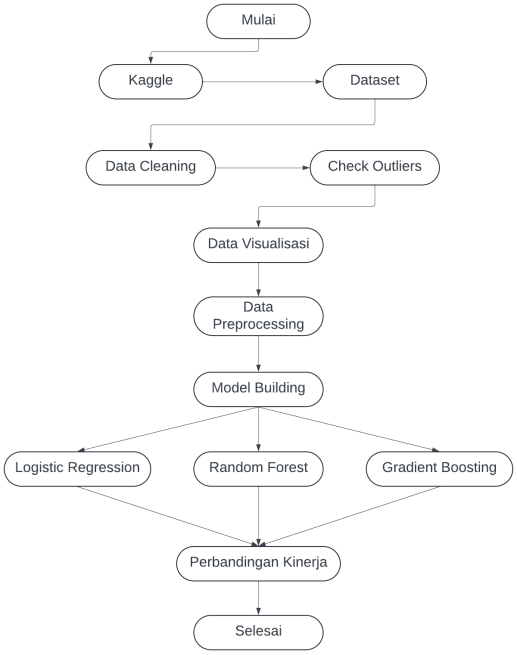
Selanjutnya, dengan mempertimbangkan dua scenario data yang berbeda, perbandingan dan analisis akurasi prediksi ketiga metode klasifikasi data diabetes dilakukan. Logistic Regression menonjol dibandingkan dengan metode klasifikasi lainya seperti k-Nearest Neighbors (k-NN), terutama dalam hal interpretasi hasil yang mudah dipahami dan efisiensi komputasi yang tinggi, terutama dalam hal dataset yang besar. Namun, mungkin kurang mampu menangani pola data yang kompleks dan non-linear. Sebaliknya, Random Forest dapat menangani interaksi variabel yang kompleks dan memiliki toleransi yang tinggi terhadap overfitting. Ini membuat pilihan yang bagus untuk situasi dimana hubungan antara variabel predictor dan target tidak linier. Karena kompleksitas algoritmanya, interpretasi model Random Forest seringkali lebih sulit. Semetara setiap metode memiliki keunggulan dan kekurangannya sendiri, pemilihan metode harus mempertimbangkan keseimbangan antara interpretasi, kinerja, dan kompleksitas komputasi yang dipahami. Gradient Boosting, disisi lain cenderung memberikan kinerja prediksi yang lebih baik, terutama dalam dataset besar dan kompleks, dan membutuhkan banyak waktu dan sumber daya komputasi untuk melatih modelnya. Tujuan utamanya adalah untuk menemukan metode klasifikasi terbaik berdasarkan hasil evakuasi dan pertandingan, dengan penekanan khusus pada metode yang memberikan akurasi tertinggi. Dengan mencapai tujuan tersebut, penelitian ini memberikan wawasan mendalam tentang bagaimana taksonomi berguna untuk mengatasi masalah prediksi data diabetes dan memberikan pemahaman yang lebih baik tentang kondisi ini dan cara yang lebih efektif untuk mengobatinya. Hal ini diharapkan akan memungkinkan penggunaan terapi.[13]

II. METODOLOGI PENELITIAN

A. Tahap Penelitian

Penelitian ini melalui berbagai tahapan proses. Langkah pertama adalah mengumpulkan data dari Kaggle, Mengambil dataset dari Kaggle adalah tindakan yang sering dilakukan dalam penelitian data science karena Kaggle menyediakan berbagai dataset berkualitas tinggi yang dapat digunakan untuk berbagai tujuan penelitian. Ada beberapa alasan yang mendukung pemilihan dataset ini. Sebenarnya, dataset ini relevan dengan tujuan saya. Semenjak dataset berdasarkan pada ketegasan, hal ini memberikan penerapan sederhana dan mudah dari datset yang dipilih berdasarkan ketepatan kebutuhan dan hipotesis yang diuji. Kualitas dan kelengkapan versi data adalah prioritas, dan saya cenderung menggunakan dataset yang memiliki kualitas yang baik, seperti data lengkap. Semakin kecil dan jumlah minimal data yang hilang, semakin baik dan lebih bermanfaat analisis terlaksana. Langkah yang dilakukan adalah Setelah data di kumpulkan terjadi tahap prapemrosesan, yaitu tahap dimana data diproses kembali. Setelah data diolah terlebih dahulu, maka data tersebut diklasifikasi menggunakan Teknik Logistic Regression, Random Forest, dan Gradient Boosting. Berikut Langkah-langkah yang dilakukan dalam

penelitian ini, yaitu:



Gambar. 1. Variabel Dataset

1. *Pengambilan Data*
   Fase akuisisi data disediakan oleh Kaggle. Dataset ini diperoleh dari diabetes_prediction_dataset dalam format CSV dengan menggunakan atribut atau variabel karakteristik yang ada pada dataset tersebut. Ini akan muncul dilabel 1.

TABEL I
VARIABEL DATASET

Variabel	Deskripsi
Gender	Penelitian ini menunjukan bahwa ada perbedaan dalam manifestasi diabetes antara laki-laki dan perempuan, misalnya laki-laki mungkin memiliki resiko yang lebih tinggi untuk mengembangkan diabetes pada usia yang lebih muda.
Age	Resiko diabetes meningkat dengan bertambahnya usia.
Hypertension	Hipertensi adalah kondisi umum pada orang dengan dibetes.
Heart_disease	Ada hubungan yang kuat antara penyakit jantung dan diabetes. Diabetes adalah factor resiko utama untuk pengembangan penyakit jantung coroner.
Smoking_history	Merokok dapat meningkatkan resiko banyak komplikasi Kesehatan, termasuk penyakit jantung dan stroke yang juga terkait penyakit diabetes.
Body mass index (BMI)	Orang yang BMI dikategori obesitas memiliki resiko yang lebih tinggi untuk mengembangkan diabetes.
HbA1c_level	Mengukur rata-rata kadar glukosa darah anda selama 2 hingga 3 bulan terakhir.
Blood_Glucose_level	Kadar glukosa darah yang tinggi pada tes berulang adalah indicator utama diabetes.

2. *Data Cleaning dan Check Outliers*
   Proses selanjutnya setelah pengumpulan data adalah pembersihan data. Ini adalah Langkah yang bertujuan untuk memahami data atau variabel. Pada titik ini, kumpulan data diambil melalui situs Kaggle dengan memeriksa data duplikat dan memeriksa saluran yang hilang. Table 2 menunjukkan nilai kosong dan duplikat:

TABEL II
Duplicates Data

Duplikat Data	3,854

Setelah memeriksa data yang kosong dan duplikat, Langkah selanjutnya adalah mencari outliers. Outliers ini berguna untuk analisis data eksplorasi. Mengidentifikasi dan memahami distribusi outliers merupakan langkah penting dalam prapemrosesan data untuk banyak analisis statistik dan model prediktif. Nilai abnormal pengukuran ini (BMI, HbA1c, gula darah) merupakan indikator kuat diabetes.

3. *SMOTE*

SMOTE (Teknik Oversampling Minoritas Sintetis): SMOTE adalah salah satu teknik oversampling yang paling banyak digunakan. Berdasarkan sampel yang sudah ada, sampel sintetis baru untuk kelas minoritas dibuat untuk mencapai tujuan ini. SMOTE adalah metode yang relatif mudah dan efektif untuk mengatasi ketidakseimbangan kelas. Pada penelitian ini mempertimbangkan evaluasi accuracy, precision, recall, f1-score. Metrik yang relevan termasuk akurasi, yang mengukur proporsi prediksi yang benar dari total prediksi; presisi, yang mengukur proporsi prediksi positif yang benar, dan penting ketika biaya kesalahan positif palsu tinggi; recall (sensitivitas), yang mengukur proporsi kasus positif yang benar-benar terdeteksi oleh model, memastikan bahwa tidak ada kasus diabetes yang terlewatkan; dan F1-Score, yang merupakan harmonisasi antara presisi dan recall, berguna ketika semuanya seimbang. Menggunakan metrik ini akan membantu mengevaluasi model secara menyeluruh dan memastikan bahwa model tidak hanya akurat tetapi juga efektif dalam menemukan kasus diabetes, mengurangi kesalahan positif palsu, dan mengurangi kesalahan negatif palsu.

4. *Data Preprocessing*

Fase ini melakukan pemrosesan data untuk mempersiapkan model pembelajaran mesin termasuk berbagi data, Salah satu langkah penting dalam proses pembangunan model pembelajaran mesin adalah membagi data menjadi set pelatihan dan pengujian. Tujuan utama dari pebangian ini adalah untuk menilai kinerja model pada data yang belum pernah dilihat sebelumnya. Ini akan memungkinkan untuk membuat perkiraan yang lebih masuk akal tentang kemampuan model untuk melakukan generalisasi pada data yang belum diketahui. Selain itu, bagian data dibagi menjadi 70/30, rasio yang cukup umum dan memberikan keseimbangan yang baik. Ini memastikan bahwa model memiliki cukup data untuk mempelajari berbagai pola dan variasi dalam dataset, sementara tetap memiliki cukup data untuk validasi. Ini juga memastikan bahwa rasio yang dipilih dibenarkan, termasuk ukuran dataset, variabilitas dan kompleksitas data, keseimbangan kelas, dan tujuan eksperimen, yang mencakup pengujian awal, validasi, dan optimasi. Mengatasi ketidakseimbangan kelas, dan menyiapkan kemupulan data yeng seimbang. Dengan menggunakan metode ini, anda dapat kumpulan data yang lebih seimbang dan meningkatkan performa model pembelajaran mesin, terutama dalam masalah klasifikasi yang pada awalnya ketidakseimbangan kelas terlihat jelas.[14]

5. *Modelling*

Pada titik ini, permodelan dilakukan pada algoritma. Penelitian ini akan menggunakan model Logistic Regression, Random Forest, dan Gradient Boosting. Dalam penelitian prediksi diabetes, pilihan antara Logistic Regression, Random Forest, dan Gradient Boosting didasarkan pada beberapa pertimbangan utama. Yang pertama adalah bahwa model dengan variasi kompleksitas yang berbeda (dari sederhana ke kompleks) memungkinkan untuk menilai performa prediksi dari berbagai sudut pandang. Yang kedua adalah bahwa Logistic Regression memberikan interpretabilitas yang tinggi, sedangkan Random Forest dan Gradient Boosting menawarkan akurasi yang lebih tinggi dengan kemampuan untuk kemampuan menangani kompleksitas data, yang ketiga algoritma ini mampu menangani karakteristik unik dari data medis, seperti variabilitas yang tinggi dan hubungan fitur yang kompleks. Tiga metode ini memiliki kelebihan dan kekurangan, seperti Logistic Regression: cocok untuk situasi di mana kecepatan dan interpretabilitas sangat penting, tetapi mungkin tidak cukup untuk menangkap kompleksitas penuh dari data medis; Random Forest menawarkan keseimbangan yang baik antara akurasi dan fleksibilitas, mampu menangani kompleksitas data dengan baik, tetapi mungkin memerlukan lebih banyak sumber daya komputasi; dan Gradient Boosting menawarkan lebih banyak fleksibilitas daripada akurasi, tetapi dengan biaya komputasi yang lebih tinggi dan risiko overfitting jika tidak diatur dengan benar. pada penelitian ini mempertimbangkan evaluasi accuracy, precision, recall, f1-score.

a. *Logistic Regression*

Logistic Regression merupakan algoritma klasifikasi yang digunakan untuk memprediksi probabilitas variabel dependen kategori. Variabel terikat dalam Logistic Regression adalah variabel biner yang memiliki nilai 1 (ya) atau 0 (tidak). Untuk klasifikasi biner, logistic regresi adalah algoritma statistik yang tujuan utamanya adalah memprediksi kemungkinan bahwa suatu contoh akan masuk ke salah satu dari dua belas. Dalam konteks prediksi diabetes, Logistic Regression mencoba memprediksi apakah seorang pasien memiliki diabetes (positif) atau tidak (negatif Dan ada beberapa cara bekerjanya. Yang pertama adalah fungsi logistik sigmoid atau logistik yang membatasi output menjadi nilai antara 0 dan 1. Fungsi ini diberikan oleh :

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Dimana $z = \beta o + \beta_1 \chi_1 + \beta_2 \chi_2 + . + \beta_n \chi_n . \beta$ adalah koefisien regresi yang dipelajari dari data pelatihan. Yang kedua adalah fungsi probabilitas sigmoid yang mengubah input linear menjadi probabilitas dan mengestimasi parameter. Untuk mengevaluasi ikatan antara sejumlah variabel dan variabel biner atau acak, logistic regression biner adalah teknik analisis data yang umum. Variabel respon biner (y) dan variabel prediktor (x) terdiri dari dua kategori sukses dan gagal, yang diwakili oleh nilai y=1 (sukses) dan nilai y=0 (gagal). [15] Formula dasar untuk regresi logistik adalah sebagai berikut:

$Y = \beta 0 + \beta 1 X + \varepsilon$ [16]

Diketahui:

Y : variabel dependen
β0: konstatnta
β1: koefisien regresi
X: variabel bebas
ε : kesalahan acak

b. *Random Forest*

Pendekatan Random Forest yang diusulkan oleh Breiman adalah pembelajaran mesin yang memiliki banyak pohon keputusan. Random Forest adalah hasil dari kombinasi teknik bagging dan sub-ruang rando. Dalam beberapa tahun terakhir, Metode ini telah terbukti berhasil dalam menangani masalah regresi dan klasifikasi., dan dianggap sebagai salah satu algoritma pembelajaran mesin yang paling populer di banyak industri. [17] Random Forest algoritma dasar menggunakan banyak pohon keputusan untuk membuat prediksi. Dalam Hutan Random, setiap pohon terdiri dari subset data yang berbeda dan subset fitur yang berbeda. Dalam konteks prediksi diabetes, Random Forest akan membuat n pohon keputusan dari subset data pelatihan yang berbeda dan kemudian menggabungkan hasil dari semua pohon untuk membuat prediksi akhir. Dalam proses pembentukan pohon, sampel bootstrap acak dari data pelatihan akan digunakan untuk membuat n pohon keputusan, dan setiap pohon akan memiliki subset fitur yang berbeda.

c. *Gradient Boosting*

Untuk regresi dan klasifikasi, gradien peningkatan adalah metode pembelajaran mesin. Model prediktif yang dibangun menggunakan algoritma peningkatan gradien merupakan ansambel dari model prediktif yang lebih lemah; ini biasanya dibentuk dalam bentuk pohom keputusan. Pohon keputusan ini dilatih dengan memberikan nilai yang sama untuk setiap observasi. Setelah evaluasi awal, Nilai observasi yang sulit diklasifikasikan ditingkatkan, sedangkan nilai observasi yang mudah diklasifikasikan dikurangi.[18] Gradient Boosting adalah algoritma ensemble yang membangun model prediksi secara bertahap. Dalam prediksi diabetes, Gradient Boosting akan membangun serangkaian pohon keputusan kecil (lemah) yang setiap pohonnya berusaha mengoreksi kesalahan prediksi dari pohon sebelumnya. dengan proses kerja, yaitu memulai model, berinteraksi dan belajar dari kesalahan, dan agresi terhadap hasilnya.

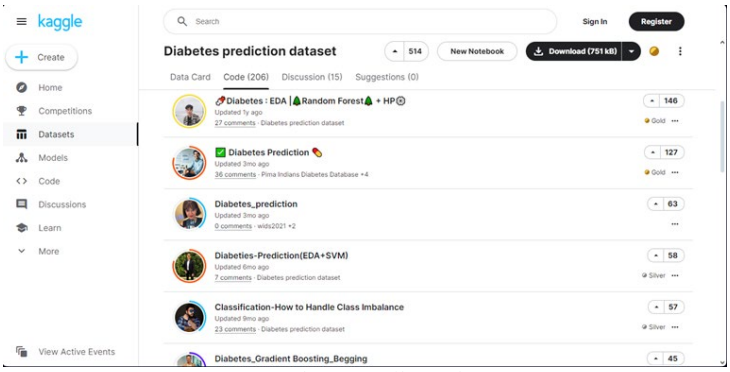
III. HASIL DAN PEMBAHASAN PENELITIAN

Metode prediksi yang digunakan dalam penelitian ini adalah Logistic Regression, Random Forest, dan Gradient Booting. Dataset yang digunakan berjumlah 10.0001 dan memiliki atribut 8 [19]. Kerangka kerja penelitian yang diusulkan pada bagian sebelumnya menghasilkan hasil berikut.

A. Pengambilan Dataset

Penelitian ini menggunakan dataset terbuka Kaggle. Data didownload secara langsung dari situs Kaggle, yang ditunjukkan pada Gambar 2.

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 1, Maret 2025, Pp. 150-162



Gambar. 2. Hasil Kaggle

B. Preprocessing Data

1. Mengecek Nilai yang Hilang

Dataset memiliki atribut yang tidak memiliki nilai, jadi perlu dilakukan proses penyiapan data. Dalam proses penyiapan data, nilai yang kosong di seluruh dataset dihilangkan atau diisi dengan nilai rata-rata dari setiap kolom [20]. Berikut tabel melihatkan hasilnya :

TABEL II
DATA YANG HILANG

Variabel	Data yang Hilang
Gender	0.0
Age	0.0
Hypertension	0.0
Heart_disease	0.0
Smoking_history	0.0
Body mass index (BMI)	0.0
HbA1c_level	0.0
Blood_Glucose_level	0.0

2. Jumlah Data Duplikat Sekarang

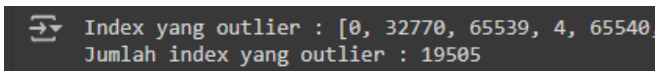
Duplikasi data adalah bagian penting dari proses pra-pemrosesan data karena duplikat data dapat mengganggu kinerja model pelatihan dan mengurangi kinerja model. Untuk memastikan bahwa data yang digunakan untuk melatih model pelatihan mesin bebas dari duplikat, proses ini disebut duplikasi data.

TABEL III
DATA DUPLIKAT SEKARANG

Data Duplikat	0

3. Hasil Visualisasi Outliers

Dari hasil Outliers ini jumlah index yang outliers adalah 19505, dengan jumlah ini akan dihapus agar tidak memperngaruhi hasil akhirnya. Dilihat pada Gambar 3.



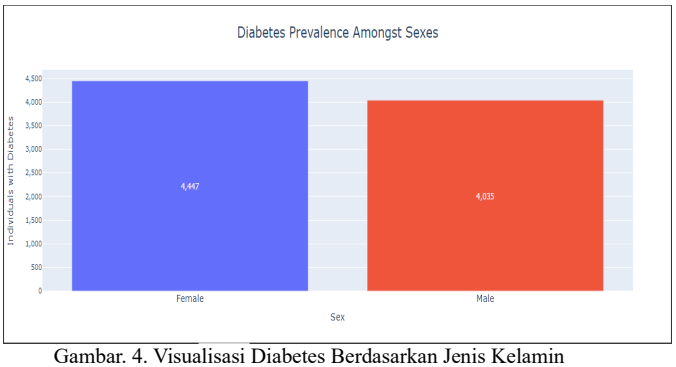
Gambar. 3. Visualisasi Outliers

4. Visualisasi Data

a. Prevalansi Diabetes Berdasarkan Jenis Kelamin

Tujuan dari pengelompokan diabetes berdasarkan jenis kelamin dan menghitung jumlah kasus diabetes untuk setiap jenis kelamin adalah untuk menyediakan analisis visual mengenai distribusi kasus diabetes berdasarkan gender. Dilihat pada Gambar 4.

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 1, Maret 2025, Pp. 150-162

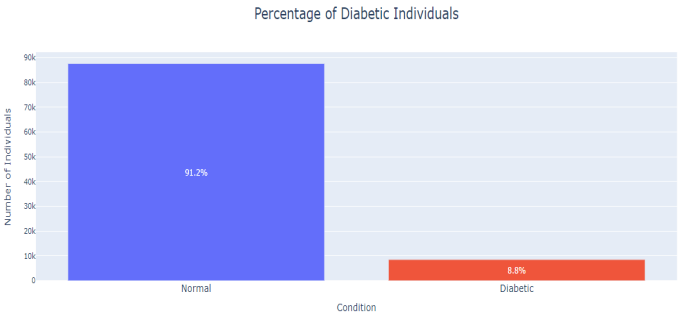


Gambar. 4. Visualisasi Diabetes Berdasarkan Jenis Kelamin

Berdasarkan jenis kelamin, bar chart ini menunjukkan prevalensi diabetes. Sumbu horizontal (X) menunjukkan kategori jenis kelamin, yaitu perempuan dan laki-laki, sedangkan sumbu horizontal (Y) menunjukkan jumlah orang yang menderita diabetes. Diabetes lebih umum pada perempuan ketika jumlah perempuan yang menderita lebih tinggi (4,447) dibandingkan dengan laki-laki (4,035), dengan sekitar 412 orang lebih banyak perempuan yang menderita diabetes dibandingkan laki-laki. Kesimpulan dari data ini menunjukkan bahwa prevalensi diabetes pada perempuan sedikit lebih tinggi dibandingkan laki-laki dalam populasi yang diteliti.

b. *Presentase Orang yang Menderita Diabetes*

Dengan menggunakan data dari kolom "diabetes" dalam DataFrame, seseorang dapat membuat visualisasi dengan Plotly Express untuk menunjukkan persentase individu yang diagnosa sebagai normal dan diabetik. Ini memberikan gambaran yang jelas tentang distribusi kondisi diabetes dalam dataset, yang memudahkan pemahaman visual tentang distribusi data berdasarkan kondisi kesehatan individu. Dilihat pada Gambar 5.

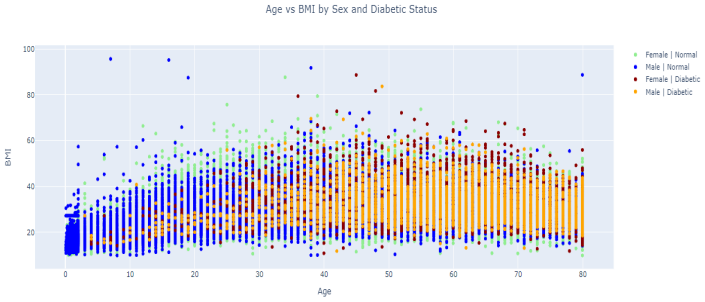


Gambar. 5. Presentase Orang yang Menderita Diabetes

Berdasarkan grafik ini, bar chart menunjukkan presentase individu dengan diabetes (normal) dan tanpa diabetes. Sumbu horizontal (X) menunjukkan kategori kondisi kesehatan normal dan diabetes, sedangkan sumbu Y menunjukkan jumlah individu. Temuan utamanya adalah bahwa sebagian besar orang dalam populasi yang diteliti adalah normal (tidak menderita diabetes), yaitu 91,2 persen, dan hanya 8,8 persen yang menderita diabetes. Kesimpulannya menunjukkan bahwa prevalensi diabetes dalam populasi yang diteliti sangat rendah, dengan sebagian besar orang (lebih dari 90 persen) tidak menderita diabetes.

c. *Usia vs BMI Berdasarkan Jenis Kelamin dan Status Diabetes*

Visualisasi ini secara efektif melakukan pengkategorian, visualisasi, dan analisis data terkait hubungan antara umur, BMI, gender, dan status diabetes. Ini membantu mendapatkan pemahaman tentang bagaimana variabel-variabel ini berinteraksi satu sama lain dan bagaimana kondisi diabetes didistribusikan. Dilihat pada Gambar 6 dan 7.



Gambar. 6. Usia vs BMI Berdasarkan Jenis Kelamin dan Status Diabetes

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 1, Maret 2025, Pp. 150-162

Sementara data menunjukkan bahwa individu individu dengan diabetes, yang diperlihatkan oleh titik merah dan orange cenderung tersebar di seluruh sumbu usia dan BMI. Individu tanpa diabetes, yang diperlihatkan oleh titik hijau dan biru, menunjukan distribusi BMI yang lebih terpusat khususnya pada rentang usia dewasa muda hingga menengah. Jenis kelamin di titik biru (laki-laki | normal) dan titik hijau muda (perempuan | normal) menunjukan bahwa titik biru hingga hijau tersebar di kisaran BMI yang lebih rendah dan tersebar di seluruh rentang usia, sedangkan titik oranye (laki-laki | diabetic) dan titik merah (perempuan | diabetic) menunjukan konsentrasi yang lebih tinggi dari BMI yang lebih tinggi, terutama dalam kelompok usia demasa menengah hingga tua.

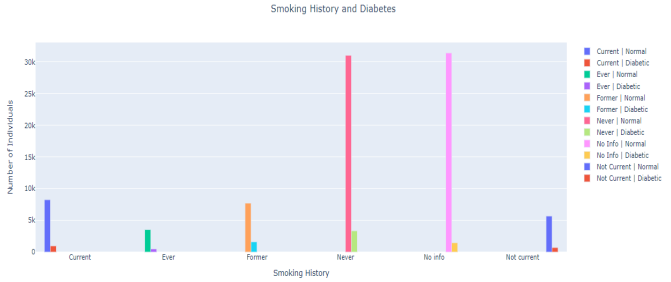
Category	Amount of Individuals	% of Individuals
Female | Normal	51,714	53.8
Male | Normal	35,932	37.4
Female | Diabetic	4,447	4.6
Male | Diabetic	4,035	4.2

Gambar. 7. Hasil Visualisasi

Tabel berikut menunjukkan jumlah total dan persentase dari masing-masing kategori berdasarkan jenis kelamin dan status diabetes. Data ini menunjukkan bahwa mayoritas orang yang tidak memiliki diabetes, dengan perempuan yang paling sering terkena. Meskipun prevalensi diabetes lebih rata antara laki-laki dan perempuan, itu masih merupakan bagian kecil dari populasi. Perbedaan gender Dibandingkan dengan laki-laki, lebih banyak perempuan dalam kategori normal dan diabetik.

d. *Visualisasi Sejarah Merokok dan Diabetes*

Mengawasi dan menampilkan data yang mencakup riwayat merokok dan diabetes. Tujuannya adalah untuk memberikan gambaran visual yang jelas tentang hubungan antara riwayat merokok dan prevalensi diabetes, tetapi Anda harus memastikan bahwa label dan warna benar-benar mencerminkan data yang dikumpulkan. Dilihat pada Gambar 8.

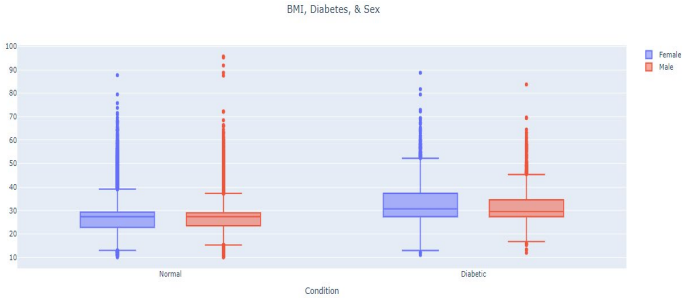


Gambar. 8. Visualisasi Sejarah Merokok dan Diabetes

Di grafik ini, bar chart menunjukkan hubungan antara riwayat merokok dan diabetes. dengan sumbu horizontal (X) menunjukkan kategori riwayat merokok (Sekarang, Selalu, Sebelumnya, Tidak pernah, Tidak ada informasi, Tidak saat ini), dan sumbu vertical (Y) menunjukkan jumlah individu. Mayoritas orang dalam penelitian adalah normal dan tidak merokok. Sebagian besar orang dalam populasi yang diteliti tidak pernah merokok dan tidak menderita diabetes. Ada sejumlah besar orang yang status merokoknya tidak diketahui, yang menunjukkan kemungkinan ketidaklengkapan data. Selain itu, prevalensi merokok individu yang merokok atau pernah merokok relatif lebih rendah dibandingkan dengan individu yang tidak merokok.

e. *Visualisasi Gender, BMI, dan Diabetes*

Ini memungkinkan visualisasi yang informatif mengenai distribusi BMI berdasarkan status diabetes dan gender, serta menyediakan analisis tambahan tentang nilai rata-rata BMI untuk masing-masing kategori tersebut. Ini meningkatkan pemahaman tentang hubungan antara diabetes, BMI, dan jenis kelamin. Dilihat dari Gambar 9 dan 10.



Gambar. 9. Visualisasi Gender, BMI, dan Diabetes

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 1, Maret 2025, Pp. 150-162

Gambar ini menunjukkan distribusi indeks massa tubuh (BMI) berdasarkan jenis kelamin dan diabetes. Kategori kondisi kesehatan diwakili oleh sumbu horizontal (X) dan nilai BMI diwakili oleh sumbu Y. Garis tengah kotak menunjukkan median BMI, dan kotak menunjukkan rentang interkuartil (IQR), yang terjadi dari kuartil pertama (Q1) hingga kuartil ketiga (Q3). Garis vertikal (whiskers) menunjukkan jangkauan data yang tidak termasuk outliers, titik diluar whiskers : outliers, nilai BMI yang jauh dari jangkauan normal. Hasilnya menunjukkan bahwa individu yang menderita diabetes, baik perempuan maupun laki-laki, memiliki BMI yang lebih tinggi dibandingkan individu normal. Di sisi lain, perbedaan gender dalam distribusi BMI terlihat pada laki-laki, yang cenderung memiliki BMI yang lebih tinggi dibandingkan perempuan, baik dalam kelompok normal maupun diabetik. dan pengaruh BMI pada diabetes didata ini menunjukan adanya hubuangan antara BMI yang lebih tinggi dengan kondisi diabetes, menyoroti pentingnya pengelolaan berat badan dalam pencegahan dan pengelolaan diabetes.

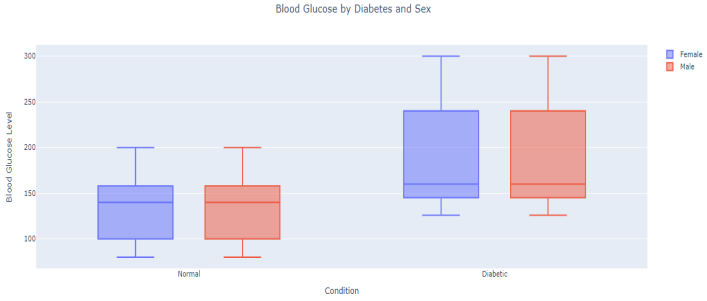
	Average BMI
Male | Normal	26.67
Male | Diabetic	31.29
Female | Normal	27.01
Female | Diabetic	32.64

Gambar. 10. Hasil Visualisasi

Dari data diatas menunjukan BMI rata-rata Laki-laki, Normal : rata-rata untuk laki-laki yang tidak menderita diabetes adalah 26.67, Diabetic : rata-rata BMI untuk laki-laki yang menderita diabetes adalah 31.29, Perbedaan : laki-laki yang menderita diabetes memiliki rata-rata BMI yang lebih tinggi (sekitar 4.62 poin) dibandingkan yang tidak menderita diabetes. BMI Rata-rata Perempuan, Normal : rata-rata BMI untuk perempuan yang tidak diabetes adalah 27.01, Diabetic : rata-rata BMI untuk perempuan yang menderita diabetes adalah 32.64, Perbedaan : perempuan yang menderita diabetes memiliki rata-rata BMI yang lebih tinggi (sekitar 5.63 poin) dibandingkan yang tidak menderita diabetes. Selanjutnya perbandingan BMI rata-rata antara jenis kelamin, Normal : rata-rata BMI perempuan (27.01) sedikit lebih tinggi dari pada laki-laki (26.67), sedangkan Diabetic : rata-rata BMI perempuan (32.64) juga lebih tinggi dari pada laki-laki (31.29).

f. Glukosa Darah Berdasarkan Gender dan Diabetes

Visualisasi data dan menghitung kadar glukosa darah rata-rata berdasarkan jenis kelamin dan diabetes. Setiap kategori memiliki rata-rata glukosa darah dalam DataFrame yang dibuat ('averages_glucose_df'), yang membantu memahami distribusi glukosa berdasarkan jenis kelamin dan status diabetes. Setiap kategori memiliki kadar glukosa darah yang berbeda, dan box plot yang dihasilkan memungkinkan analisis visual yang mudah. Dilihat dari Gambar 11 dan 12.



Gambar. 11. Glukosa Darah Berdasarkan Gender dan Diabetes

Tingkat glukosa darah berdasarkan jenis kelamin dan diabetes ditunjukkan dalam diagram kotak ini. Sumbu X (Kondisi) menunjukkan dua kondisi: normal dan diabetik. Sumbu Y (Tingkat Glukosa Darah) menunjukkan tingkat glukosa darah yang berkisar antara 100 dan lebih dari 300. Temuan utamanya adalah bahwa individu dengan diabetes memiliki tingkat glukosa darah yang lebih tinggi secara signifikan dibandingkan dengan individu normal. Namun, perbedaan tingkat glukosa darah antara laki-laki dan perempuan dengan diabetes tidak signifikan, tetapi laki-laki mengalami variasi yang lebih besar.

	Average Blood Glucose Level
Male | Normal	132.9
Male | Diabetic	194.2
Female | Normal	132.8
Female | Diabetic	193.8

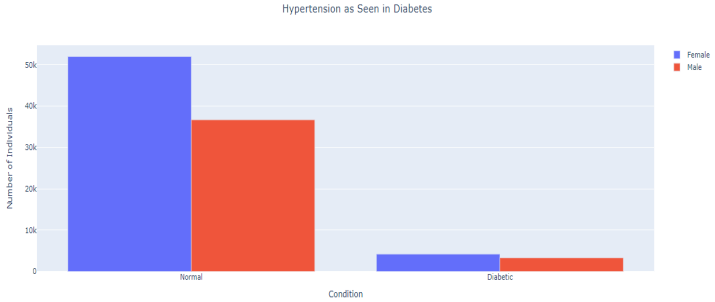
Gambar. 12. Hasil Visualisasi

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 1, Maret 2025, Pp. 150-162

Tabel ini menunjukkan tingkat glukosa darah rata-rata berdasarkan jenis kelamin dan kondisi diabetes. Secara umum, diabetes meningkatkan tingkat glukosa darah hampir dua kali lipat dibandingkan kondisi normal, tetapi antara laki-laki dan perempuan dalam kondisi normal dan diabetes tidak ada perbedaan yang signifikan.

g. *Hipertensi Pada Pasien Diabetes*

Menampilkan distribusi kasus diabetes dan hipertensi dikelompokkan menurut gender menggunakan Plotly Express. Namun, label dan istilah yang digunakan, terutama yang berkaitan dengan hipertensi, perlu diklarifikasi dan dikoreksi. Dilihat dari Gambar 13.



Gambar. 13. Visualisasi Hypertensi pada Pasien Diabetes

Jenis kelamin dan kondisi diabetes menentukan jumlah pasien hipertensi pada diagram batang ini. Sumbu X menunjukkan kondisi "Normal" dan "Diabetic", sedangkan sumbu Y menunjukkan jumlah orang yang menderita hipertensi, dengan skala dari 0 hingga 50.000. Temuan utamanya adalah sebagai berikut: Jumlah perempuan dengan hipertensi jauh lebih tinggi (sekitar 50 ribu) dibandingkan laki-laki (sekitar 35 ribu) pada kondisi "Normal", Jumlah laki-laki dengan hipertensi pada kondisi diabetes juga menurun, tetapi sedikit lebih tinggi dari jumlah perempuan (sekitar 10 ribu), Perbandingan Hipertensi dalam Kondisi Normal dan Diabetes: Untuk kedua jenis kelamin, hipertensi jauh lebih umum dalam kondisi normal daripada diabetes. Kesimpulan dari sebelumnya adalah bahwa hipertensi lebih umum pada perempuan dalam kondisi normal dibandingkan laki-laki; selain itu, jumlah orang dengan diabetes menurun drastis untuk kedua jenis kelamin, tetapi sedikit lebih banyak pada laki-laki dibandingkan perempuan.

5. *Modelling*

a. *Logistic Regression*

Pengujian dilakukan menggunakan metode regresi logistik di Google Collaboratory. Hasil pengujian yang dilakukan dengan menggunakan model ini untuk memprediksi penyakit diabetes untuk menentukan menunjukkan hasil yang cukup baik, dengan nilai akurasi sebesar 76%, precision sebesar 76%, recall sebesar 77%, f1 score 77%, seperti yang ditunjukkan pada Gambar 14 ini:

```
Logistic Regression:
              precision    recall  f1-score   support

           0       0.77      0.76      0.77        67
           1       0.76      0.77      0.77        66

    accuracy                           0.77       133
   macro avg       0.77      0.77      0.77       133
weighted avg       0.77      0.77      0.77       133

Accuracy: 0.7669172932330827
ROC AUC Score: 0.7669606512890095
[[51 16]
 [15 51]]
```

Gambar. 14. Hasil Akurasi Logistic Regression

b. *Random Forest*

Pengujian kedua menggunakan algoritma Random Forest. Hasilnya menunjukkan nilai yang bagus dengan akurasi sebesar 77%, precision sebesar 74%, recall sebesar 83%, f1 score 79%, pada data tes untuk prediksi penyakit diabetes, seperti yang ditunjukkan pada Gambar 15 ini:

```
Random Forest:
              precision    recall  f1-score   support

           0       0.81      0.72      0.76        67
           1       0.74      0.83      0.79        66

    accuracy                           0.77       133
   macro avg       0.78      0.77      0.77       133
weighted avg       0.78      0.77      0.77       133

Accuracy: 0.7744360902255639
ROC AUC Score: 0.7748756218905474
[[48 19]
 [11 55]]
```

Gambar. 15. Hasil Akurasi Random Forest

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 1, Maret 2025, Pp. 150-162

c. *Gradient Boosting*

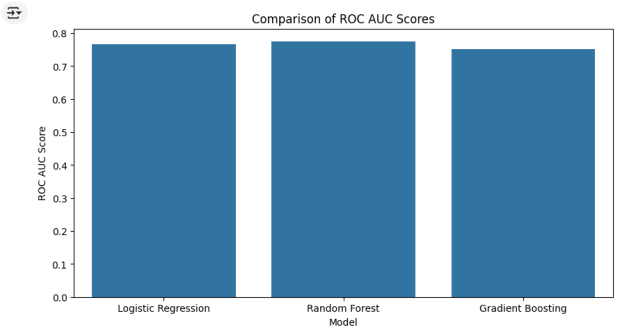
Pengujian sebelumnya memanfaatkan metode/algoritma Gradient Boosting. Pengujian yang dilakukan dengan model ini menghasilkan hasil yang cukup baik, dengan nilai akurasi data sebesar 75%, precision sebesar 74 %, recall sebesar 77%, f1 score 76%, seperti yang ditunjukkan dalam Gambar 16 berikut:[21]

```
Gradient Boosting:
              precision    recall  f1-score   support

           0       0.77      0.73      0.75        67
           1       0.74      0.77      0.76        66

    accuracy                           0.75       133
   macro avg       0.75      0.75      0.75       133
weighted avg       0.75      0.75      0.75       133

Accuracy: 0.7518796992481203
ROC AUC Score: 0.7520352781546812
[[49 18]
 [15 51]]
```

Gambar. 16. Hasil Akurasi Gradient Boosting

6. *Hasil Perbandingan*

Hasil perbandingan ini mendefinisikan fungsi untuk menghitung skor evaluasi model, Penjelasan tentang metrik yang digunakan adalah sebagai berikut: Akurasi: untuk mengukur seberapa banyak prediksi yang benar dari total prediksi, tetapi tidak membedakan antara jenis kesalahan (false positives dan false negatives), Precision: untuk memberikan informasi tentang kualitas prediksi positif, membantu mengurangi alarm yang tidak perlu, Recall: untuk memastikan bahwa model tidak melewatkan banyak kasus positif, yang sangat penting untuk penyakit serius seperti diabetes, diabetes mellitus, dan penyakit serius lainnya. F1-Score: memberikan analisis menyeluruh tentang kinerja model, terutama dalam kasus ketidakseimbangan antara precision dan recall. Kemudian, menggunakan fungsi ini untuk membandingkan tiga model pembelajaran mesin: Logistic Regression, Random Forest, dan Gradient Boosting. Model mana yang memberikan hasil terbaik berdasarkan metrik evaluasi yang dihitung? Gambar 17 berikut menunjukkan hasil perbandingan dari ketiga pendekatan tersebut:



Gambar. 17. Hasil Perbandingan Tiga Metode

Hasil menunjukkan bahwa Random Forest memiliki akurasi tertinggi, diikuti oleh Logistic Regression, dan Gradient Boosting memiliki akurasi paling rendah. Ini disebabkan oleh beberapa faktor. Random Forest menggabungkan banyak pohon keputusan dan mengurangi risiko overfitting melalui averaging. Dengan demikian, Random Forest mampu menangani variasi dan kompleksitas data dengan baik, Meskipun model linear, logistic regression dapat berfungsi dengan baik jika data memiliki hubungan yang cukup linear dan interaksi yang tidak terlalu kompleks. Dalam hal ini, ia menghasilkan akurasi yang lebih tinggi daripada Gradient Boosting. Meskipun biasanya sangat kuat, Gradient Boosting memerlukan pengaturan hiperparameter yang sangat tepat dan dapat overfitting jika tidak diatur dengan baik; dalam penelitian ini, mungkin ada kurangnya optimasi atau overfitting yang menyebabkan akurasi yang lebih rendah dibandingkan dua model lainnya. Tiga metode yang berbeda memiliki implikasi praktis yaitu Random Forest: model ini sangat cocok untuk prediksi diabetes karena sangat akurat dan dapat memberikan estimasi fitur penting yang membantu pengambilan keputusan medis yang lebih cerdas. Gradient Boosting: dapat digunakan ketika presisi sangat penting dan ada cukup waktu dan sumber daya untuk mengatur hyperparameter secara menyeluruh, Logistic Regression: Model ini dapat digunakan sebagai baseline atau dalam situasi di mana kecepatan dan interpretabilitas prediksi lebih penting daripada akurasi absolut, meskipun kurang akurat.

IV. KESIMPULAN

Berdasarkan penelitian ini, dilakukan perbandingan antara Logistic Regression, Random Forest, Gradient Boosting pada dataset penyakit diabetes. Setelah dilakukan perhitungan pada masing-masing algoritma didapatkan

hasil menunjukan bahwa Random Forest memberikan performa yang terbaik dalam hal akurasi dengan nilai sebesar 77%, Diikuti oleh Logistic Regression dengan nilai sebesar 76% sedangkan Gradient Boosting memberikan akurasi sebesar 75%. Meningkatkan nilai akurasi algoritma di atas dapat membantu menentukan apakah seseorang menderita diabetes untuk memulai perawatan dan mencegah komplikasi sejak dini. Adapun sejumlah keterbatasan pada penelitian ini, yaitu keseimbangan data yang dapat mengontribusikan bias terhadap kelas, yang bersentuhan dengan resiko overfitting pada model yang kompleks, selain itu, Teknik validasi yang kurang baik seperti K-Fold Cross Validation. Hal tersebut dapat diperbaiki dengan metode penyeimbangan kelas yang lebih canggih untuk tema yang kompleks, serta menggunakan teknik regularisasi yang lebih efektif dan penggunaan cross-validation yang lebih ekstensif. Selain itu, perlunya penelitian lebih lanjut dengan menggunakan dataset yang lebih bervariasi untuk menguji generalization dari model pada berbagai jenis data. Sehingga penelitian lebih lanjut ini dapat memberikan wawasan yang lebih mendalam dalam applicasinya kepada machine learning di berbagai bidang.

DAFTAR PUSTAKA

[1] E. Cahya, P. Witjaksana, R. Rohmat Saedudin, and V. P. Widartha, "PERBANDINGAN AKURASI ALGORITMA RANDOM FOREST DAN ALGORITMA ARTIFICIAL NEURAL NETWORK UNTUK KLASIFIKASI PENYAKIT DIABETES."
[2] S. P. Nainggolan and A. Sinaga, "COMPARATIVE ANALYSIS OF ACCURACY OF RANDOM FOREST AND GRADIENT BOOSTING CLASSIFIER ALGORITHM FOR DIABETES CLASSIFICATION," *Sebatik*, vol. 27, no. 1, pp. 97–102, Jun. 2023, doi: 10.46984/sebatik.v27i1.2157.
[3] J. Elektronik *et al.*, "Implementasi Logistic Regression dalam Sistem Diagnosa Penyakit Diabetes dengan KNN," vol. 11, no. 4, pp. 2654–5101.
[4] "JEPIN (Jurnal Edukasi dan Penelitian Informatika) Peningkatan Kinerja Akurasi Prediksi Penyakit Diabetes Mellitus Menggunakan Metode Grid Seacrh pada Algoritma Logistic Regression".
[5] W. Apriliah *et al.*, "SISTEMASI: Jurnal Sistem Informasi Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest," 2021. [Online]. Available: http://sistemasi.ftik.unisi.ac.id
[6] R. Susetyoko, W. Yuwono, E. Purwantini, and N. Ramadijanti, "Perbandingan Metode Random Forest, Regresi Logistik, Naïve Bayes, dan Multilayer Perceptron Pada Klasifikasi Uang Kuliah Tunggal (UKT)," vol. 7, no. 1.
[7] A. M. Ridwan and G. D. Setyawan, "PERBANDINGAN BERBAGAI MODEL MACHINE LEARNING UNTUK MENDETEKSI DIABETES," *TEKNOKOM*, vol. 6, no. 2, pp. 127–132, Aug. 2023, doi: 10.31943/teknokom.v6i2.152.
[8] "Implementasi Data Mining Dalam Melakukan Prediksi Penyakit Diabetes Menggunakan Metode Random Forest Dan Xgboost," *Jurnal Ilmiah Komputasi*, vol. 23, no. 1, Mar. 2024, doi: 10.32409/jikstik.23.1.3507.
[9] C. W. Cahyana and A. Nurlayli, "ANALISIS PERFORMA LOGISTIC REGRESSION, NAÏVE BAYES, DAN RANDOM FOREST SEBAGAI ALGORITMA PENDETEKSI KANKER PAYUDARA," *INSERT: Information System and Emerging Technology Journal*, vol. 4, no. 1, 2023.
[10] S. P. Nainggolan and A. Sinaga, "COMPARATIVE ANALYSIS OF ACCURACY OF RANDOM FOREST AND GRADIENT BOOSTING CLASSIFIER ALGORITHM FOR DIABETES CLASSIFICATION," *Sebatik*, vol. 27, no. 1, pp. 97–102, Jun. 2023, doi: 10.46984/sebatik.v27i1.2157.
[11] A. P. Wicaksono, T. B. #2, and A. Basuki, "Data Mining Studi Perbandingan Prediksi Penyakit Diabetes dengan menggunakan Logistic Regression dan Decision Trees".
[12] Dwipa Jaya Made Krisna, "Perbandingan Random Forest, Decision tree, Gradient Boosting, Logistic Regression untuk Klasifikasi Penyakit Jantung," *JNATIA*, vol. 2, 2023.
[13] D. Nasien *et al.*, "Perbandingan Implementasi Machine Learning Menggunakan Metode KNN, Naive Bayes, Dan Logistik Regression Untuk Mengklasifikasi Penyakit Diabetes," 2024.
[14] J. Elektronik *et al.*, "Implementasi Logistic Regression dalam Sistem Diagnosa Penyakit Diabetes dengan KNN," vol. 11, no. 4, pp. 2654–5101.
[15] E. Oktavia, A. Id Hadiana, and F. Rahmat Umbara, "Penerapan Metode Regresi Logistik Dalam Prediksi Risiko Diabetes Melitus Gestasional," vol. 5, no. 2, pp. 177–185, doi: 10.52661.
[16] T. Zulhaq Jasman, E. Hasmin, C. Susanto, and W. Musu, "Perbandingan Logistic Regression, Random Forest, dan Perceptron pada Klasifikasi Pasien Gagal Jantung," *Oktober*, vol. 14, no. 3, pp. 271–286, 2022, doi: 10.22303/csrid.14.3.2022.271-286.
[17] W. Apriliah *et al.*, "SISTEMASI: Jurnal Sistem Informasi Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest," 2021. [Online]. Available: http://sistemasi.ftik.unisi.ac.id
[18] M. K. Dwipa Jaya, "Perbandingan Random Forest, Decision Tree, Gradient Boosting, Logistic Regression untuk Klasifikasi Penyakit Jantung," *JNATIA*, vol. 2, pp. 1–5, 2023.
[19] F. Handayani *et al.*, "JEPIN (Jurnal Edukasi dan Penelitian Informatika) Komparasi Support Vector Machine, Logistic Regression Dan Artificial Neural Network dalam Prediksi Penyakit Jantung".
[20] Gde Agung Brahmana Suryanegara, Adiwijaya, and Mahendra Dwifebri Purbolaksono, "Peningkatan Hasil Klasifikasi pada Algoritma Random Forest untuk Deteksi Pasien Penderita Diabetes Menggunakan Metode Normalisasi," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 1, pp. 114–122, Feb. 2021, doi: 10.29207/resti.v5i1.2880.
[21] C. W. Cahyana and A. Nurlayli, "ANALISIS PERFORMA LOGISTIC REGRESSION, NAÏVE BAYES, DAN RANDOM FOREST SEBAGAI ALGORITMA PENDETEKSI KANKER PAYUDARA," *INSERT: Information System and Emerging Technology Journal*, vol. 4, no. 1, 2023.