Institut Riset dan Publikasi Indonesia (IRPI)
**MALCOM: Indonesian Journal of Machine Learning and Computer Science**
Journal Homepage: https://journal.irpi.or.id/index.php/malcom
Vol. 5 Iss. 4 October 2025, pp: 1423-1436
ISSN(P): 2797-2313 | ISSN(E): 2775-8575

# *Comparison of Diabetes Prediction Data Using Machine Learning*

# Perbandingan Data Prediksi Diabetes Menggunakan *Machine Learning*

**Mikhail Claudio Ibrahim[1*], Fachruddin[2], Nurhadi[3]**

[1,2,3]Program Studi Magister Sistem Informasi, Universitas Dinamika Bangsa, Indonesia

E-Mail: [1]mikhailibrahim53@gmail.com, [2]fachruddin.stikom@gmail.com, [3]nurhadi@unama.ac.id

*Received Sep 14th 2025; Revised Oct 15th 2025; Accepted Oct 25th 2025; Available Online Oct 31th 2025*
*Corresponding Author: Corresponding Author*


**Abstract**

*Diabetes is a long-term metabolic disease that affects the human body by converting blood glucose into energy. Individuals diagnosed with diabetes are unable to control their blood sugar levels, which leads to increased blood sugar and blood pressure. This study aims to analyze and compare data visualization from four machine learning models* (*Random Forest, Logistic Regression, SVC, and Gradient Boosting*) *in predicting diabetes. The results of this study are expected to provide guidance in selecting the optimal visualization technique to support the interpretation of machine learning-based diabetes prediction results, as well as improve the effectiveness of communicating prediction results to medical practitioners and patients. The results show that Gradient Boosting ranks first with an ROC AUC score of* 0.97*, followed by Random Forest* (0.94) *and Logistic Regression* (0.93)*, while SVC records a value of* 0.84*. Although the overall accuracy of the models was quite high* (≈97%)*, the recall value for the positive class remained low* (0.39–0.49)*, indicating challenges in identifying minority cases in an imbalanced dataset. Physiological variables, such as BMI and HbA1c, were found to play a significant role as predictors*; *however, model performance is expected to improve further by integrating behavioral factors and more detailed medical history*.

*Keyword: Diabetes, Gradient Boosting, Logistic Regression, Random Forest, Support Vector Classifier*

**Abstrak**

Diabetes merupakan penyakit metabolisme jangka panjang yang berdampak pada tubuh manusia dengan cara mengubah glukosa darah menjadi energi. Individu yang telah didiagnosis dengan diabetes tidak dapat mengontrol level gula dalam darah, yang akan menyebabkan peningkatan kadar gula darah dan tekanan darah. Penelitian ini bertujuan untuk menganalisis dan membandingkan visualisasi data dari empat model *machine learning* (*Random Forest, Logistic Regression, SVC,* dan *Gradient Boosting* ) dalam prediksi penyakit diabetes. Hasil penelitian ini diharapkan dapat memberikan panduan dalam pemilihan teknik visualisasi yang optimal untuk mendukung interpretasi hasil prediksi diabetes berbasis *machine learning*, serta meningkatkan efektivitas komunikasi hasil prediksi kepada praktisi medis dan pasien. Hasil penelitian memperlihatkan bahwa *Gradient Boosting* menempati posisi teratas dengan skor *ROC AUC* sebesar 0,97, disusul oleh *Random Forest* (0,94) dan *Logistic Regression* (0,93), sementara SVC mencatatkan nilai 0,84. Walaupun tingkat akurasi keseluruhan model cukup tinggi (≈97%), nilai *recall* pada kelas positif masih rendah (0,39–0,49), yang menunjukkan adanya kendala dalam mengidentifikasi kasus minoritas pada *dataset* yang tidak seimbang. Variabel fisiologis, seperti *BMI* dan HbA1c, terbukti berperan signifikan sebagai prediktor, namun kinerja model diperkirakan dapat meningkat lebih jauh dengan mengintegrasikan faktor perilaku dan riwayat medis yang lebih terperinci.

Kata Kunci: Diabetes, *Gradient Boosting, Logistic Regression, Random Forest, Support Vector Classifier*

## 1. PENDAHULUAN

Diabetes merupakan penyakit metabolisme jangka panjang yang berdampak pada tubuh manusia dengan cara mengubah glukosa darah menjadi energi [1]. Individu yang telah didiagnosis dengan diabetes tidak dapat mengontrol level gula dalam darah, yang akan menyebabkan peningkatan kadar gula darah dan tekanan darah [2]. Jika diabetes tidak dikenali, didiagnosis, dan tidak diobati dengan tepat sejak awal, Diabetes akan berdapampak pada kondisi jangka panjang yang serius dan berdampak besar besar yang berdampak pada kehidupan dan kesejahteraan individu, keluarga, dan masyarakat di seluruh dunia [3]. Ini adalah salah satu dari 10 penyebab utama kematian pada orang dewasa. dan diperkirakan telah menyebabkan

empat juta kematian secara global pada tahun 2017 dan pada tahun 2017 pengeluaran kesehatan global untuk diabetes diperkirakan mencapai *United States Dollar* (USD) 727 miliar [4].

Tantangan utama dalam pengendalian diabetes adalah keterbatasan akses terhadap layanan kesehatan dan kurangnya kesadaran masyarakat tentang faktor resiko diabetes. Lebih dari 50% kasus diabetes tidak terdiagnosis karena minimnya *screening* dan deteksi dini [5]. Kondisi ini diperburuk dengan terbatasnya sumber daya kesehatan, terutama di daerah rural dan remote.

Seiring dengan perkembangan teknologi, *machine learning* telah menjadi alat yang potensial dalam mendukung diagnosis dan prediksi penyakit diabetes. Berbagai model *machine learning* seperti *Random Forest, Logistic Regression, Support Vector Classification* (SVC), dan *Gradient Boosting* telah menunjukkan hasil yang menjanjikan dalam memprediksi diabetes dengan tingkat akurasi yang tinggi [6]. Namun, salah satu tantangan utama dalam penerapan model *machine learning* di bidang kesehatan adalah bagaimana mengkomunikasikan hasil prediksi secara efektif kepada para praktisi medis dan pasien.

Berbagai penelitian terkait prediksi diabetes berbasis *machine learning* telah berkembang pesat dalam beberapa tahun terakhir. Tasin et al. merancang sistem prediksi otomatis dengan menggabungkan data pasien perempuan Bangladesh dan *dataset* Pima Indian, menggunakan metode semi-*supervised* dan algoritma *XGBoost* yang mencapai akurasi 81% dan *Area Under the Curve* (AUC) 0.84. Namun, studi ini belum membandingkan secara luas dengan algoritma lain seperti *Random Forest* atau *Gradient Boosting*. Sementara itu, El-Sofany et al. menguji sepuluh teknik klasifikasi dalam aplikasi *mobile* untuk prediksi diabetes, menggunakan data sukarelawan dari rumah sakit di Arab Saudi dan Mesir. Penelitian mereka belum mengintegrasikan visualisasi data yang mendalam untuk mendukung interpretasi medis. Di sisi lain, Modak dan Jha mengevaluasi berbagai algoritma termasuk *ensemble methods* seperti *CatBoost* dan *LightGBM*, dengan *CatBoost* mencatat akurasi tertinggi sebesar 95.4% dan AUC 0.99. Meski hasilnya unggul, tantangan seperti ketidakseimbangan kelas dalam *dataset* masih belum ditangani secara komprehensif [7], [8], [9].

Sebagai peneliti di bidang sistem informasi kesehatan, peneliti menemukan bahwa *Random Forest* adalah metode klasifikasi dalam statistika yang berbasis komputasi. Metode klasifikasi digunakan untuk pembelajaran fungsi-fungsi berbeda yang memetakan masing-masing data terpilih ke dalam salah satu dari kategori kelas yang telah ditetapkan [10]. SVC memberikan pendekatan yang unik dalam klasifikasi diabetes melalui konsep *hyperplane* dan margin maksimal. algoritma SVC untuk mengklasifikasi pesan spamdengan menggunakan atribut data trainingberupa numerik dan teks [11]. Menurut website resmi *Scikit-learn*, algoritma SVC merupakan algoritma berbasis *library Support Vector Machine* (SVM). Adapun atribut numerik berisi panjang teks (*length*) dan jumlah tanda baca (*punct*), sedangkan atribut teks (*message*) adalah isi keseluruhan pesan. Regresi Logistik (Logit) merupakan suatu metode analisis statistika yang mendeskripsikan hubungan antara peubah respon (*dependent variable*) yang bersifat kualitatif memiliki dua kategori atau lebih dengan satu atau lebih peubah penjelas (*independent variable*) berskala kategori atau interval [12]. *Gradient Boosting* (GBR) adalah sebuah algoritma yang terdiri dari beberapa pohon keputusan untukmelakukan tugas klasifikasi dan regresi [13].

Penelitian ini bertujuan untuk menganalisis dan membandingkan visualisasi data dari empat model *machine learning (Random Forest, Logistic Regression*, SVC, dan *Gradient Boosting*) dalam prediksi penyakit diabetes. Hasil penelitian ini diharapkan dapat memberikan panduan dalam pemilihan teknik visualisasi yang optimal untuk mendukung interpretasi hasil prediksi diabetes berbasis *machine learning*, serta meningkatkan efektivitas komunikasi hasil prediksi kepada praktisi medis dan pasien. Penelitian ini menjadi penting mengingat kebutuhan akan sistem prediksi diabetes yang tidak hanya akurat tetapi juga dapat dipahami dengan baik oleh pengguna akhir. Dengan membandingkan berbagai teknik visualisasi dari model yang berbeda, penelitian ini diharapkan dapat memberikan kontribusi signifikan dalam pengembangan sistem prediksi diabetes yang lebih efektif dan *user-friendly*. Berdasarkan latar belakang yang sudah dijelaskan maka peneliti tertarik untuk membuat laporan penelitian tesis yang berjudul "Perbandingan Data Prediksi Penyakit Diabetes Menggunakan *Machine Learning*".

## 2. METODOLOGI PENELITIAN

### 2.1. Alur Penelitian

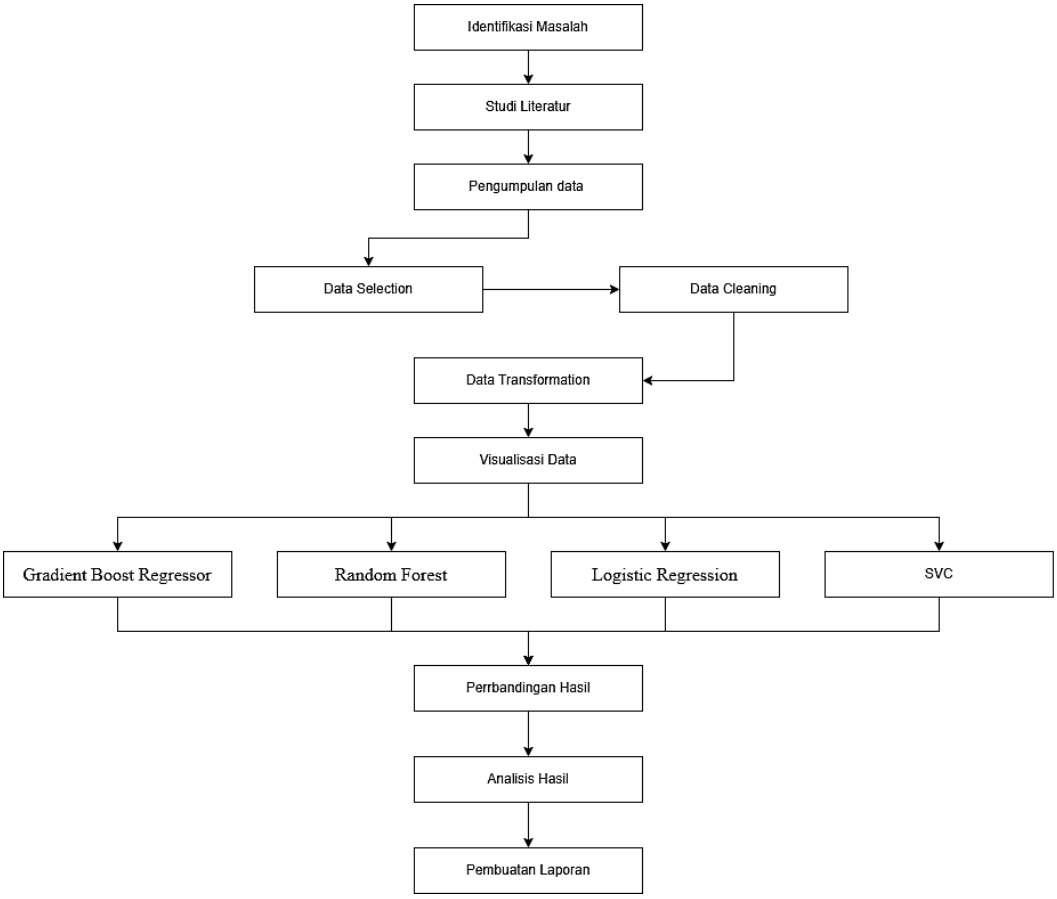
Adapun alur penelitian dapat dilihat pada Gambar 1 dan berikut penjelasannya:

1. Identifikasi Masalah
   Langkah ini dilakukan dengan berdiskusi bersama stakeholder dan mengamati kebutuhan yang ada untuk mengetahui permasalahan yang ingin diselesaikan melalui analisis data. Hasil dari tahap ini adalah perumusan masalah dan tujuan analisis yang jelas dan terukur.

2. Studi Literatur
   Melakukan pencarian dan telaah terhadap berbagai studi terdahulu melalui jurnal, artikel ilmiah, serta dokumentasi proyek sejenis.

3. Pengumpulan Data
   Data dikumpulkan melalui situs kaggle [14]. Hasilnya adalah *dataset* awal yang berisi informasi terkait masalah yang telah diidentifikasi sebelumnya.

4. *Data Selection*
   Data yang telah dikumpulkan kemudian diseleksi menggunakan teknik seperti *filtering*, analisis korelasi, dan pemilihan fitur yang relevan. Hasil dari tahap ini adalah *dataset* yang terfokus dan siap diproses lebih lanjut.




**Gambar 1.** Kerangka Kerja Penelitian

5. *Data Cleaning*
   Tahap ini dilakukan dengan membersihkan data dari nilai yang hilang (*missing values*), duplikasi, inkonsistensi, dan *outlier*. Hasilnya adalah *dataset* yang bersih, konsisten, dan dapat diandalkan untuk proses analisis berikutnya.

6. *Data Transformation*
   Data kemudian ditransformasi menggunakan teknik seperti normalisasi, *encoding* variabel kategorikal, dan rekayasa fitur untuk menyiapkannya dalam format yang optimal bagi algoritma *machine learning*.

7. Visualisasi Data
   Visualisasi data dilakukan dengan membuat grafik, diagram, dan menggunakan *tools* seperti Matplotlib atau Seaborn. Hasilnya berupa wawasan awal dari data yang mempermudah pengambilan keputusan dalam tahap *modeling*.

8. Pemodelan (*Gradient Boost, Random Forest, Logistic Regression*, SVC)
   Hasil dari tahap ini adalah beberapa model terlatih dengan performa awal yang bisa dibandingkan.
   a. *Gradient Boost*: Untuk masalah regresi dengan fokus pada peningkatan akurasi melalui *boosting* [15].
   b. *Random Forest*: Model *ensemble* berbasis pohon keputusan untuk regresi atau klasifikasi [16].

c. *Logistic Regression*: Digunakan untuk klasifikasi biner (misalnya, ya/tidak) [12].
d. SVC: Algoritma berbasis *hyperplane* untuk klasifikasi.

9. Perbandingan Hasil
Model-model yang telah dibangun kemudian dibandingkan menggunakan metrik evaluasi seperti akurasi, *precision*, *recall*, RMSE, dan lain-lain. Tujuannya adalah untuk memilih model terbaik berdasarkan performa kuantitatif.

10. Analisis Hasil
Model terbaik dianalisis lebih lanjut untuk memahami kontribusi fitur, kelebihan model, serta potensi penggunaannya dalam konteks nyata. Hasilnya adalah interpretasi mendalam dari model dan *insight* yang relevan dengan tujuan awal proyek.

11. Pembuatan Laporan
Langkah terakhir adalah menyusun laporan yang mencakup semua proses, temuan, dan rekomendasi yang diperoleh selama analisis. Laporan ini disusun secara sistematis dan disajikan kepada *stakeholder*. Hasil akhirnya adalah laporan akhir proyek analisis data yang lengkap dan informatif.

### 2.2. Bahan dan Data Metode Penelitian

Bahan penelitian mencakup referensi yang relevan, data yang digunakan, serta sumber data yang telah diperoleh sebelumnya [17]. Adapun bahan yang dibutuhkan dalam penelitian ini adalah data yang bersumber dari website Kaggle "https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset/data". Kemudian terdapat 9 atribut yang digunakan yaitu *Gender, Age, Hypertension, Heart Disease, Smoking History, BMI, HbA1c Level, Blood Glucose, Level Diabetes*.

### 2.3. Alat Bantu Penelitian

1. *Hardware* (perangkat keras) :
   a. Laptop Lenovo dengan prosesor AMD Ryzen 5 5600H 3.30GHz RAM 8 GB SSD 512 GB
   b. Mouse 3200DPI
   c. Mesin Cetak (printer)
   d. Dan beberapa perangkat keras lainnya

2. Perangkat Lunak (*software*) yang digunakan pada penelitian ini yaitu :
   a. Sistem Operasi (OS) windows 11
   b. Microsoft Office
   c. Google Chrome
   d. Google Colab
   e. Kaggle
   f. Mendeley

### 2.4. *Literature Review*

Penelitian tentang prediksi diabetes menggunakan *machine learning* telah banyak dilakukan dalam beberapa tahun terakhir dengan berbagai pendekatan dan metode. Tasin et al. mengembangkan sistem prediksi diabetes otomatis menggunakan *dataset* pribadi dari pasien perempuan Bangladesh yang dikombinasikan dengan *dataset* Pima Indian. Penelitian mereka menerapkan mutual information untuk seleksi fitur dan menggunakan model semi-*supervised* dengan *extreme gradient boosting* untuk memprediksi fitur insulin. Hasil penelitian menunjukkan bahwa XGBoost classifier dengan pendekatan ADASYN mencapai akurasi 81% dan AUC sebesar 0.84. Meskipun demikian, penelitian ini masih terbatas pada populasi tertentu dan belum mengeksplorasi perbandingan komprehensif dengan algoritma lain seperti *Random Forest* dan *Gradient Boosting* dalam konteks yang sama.

El-Sofany et al. mengusulkan teknik berbasis *machine learning* untuk prediksi diabetes melalui aplikasi mobile, dengan mengevaluasi sepuluh teknik klasifikasi termasuk *logistic regression*, *random forest*, *k-Nearest Neighbors* (K-NN), *decision tree*, *bagging*, *AdaBoost*, *XGBoost*, *voting*, SVM, dan *Naive Bayes*. Penelitian ini menggunakan *dataset* pribadi yang terdiri dari 300 sampel data sukarelawan dari rumah sakit khusus di Arab Saudi dan Mesir. Namun, penelitian ini belum mengintegrasikan teknik visualisasi data yang komprehensif untuk membantu interpretasi hasil prediksi oleh praktisi medis, yang menjadi fokus penting dalam aplikasi klinis [8].

Modak dan Jha mengembangkan model prediksi diabetes menggunakan berbagai algoritma *machine learning* termasuk *Logistic Regression*, SVM, *Naïve Bayes*, *Random Forest*, serta metode *ensemble* seperti *XGBoost*, *LightGBM*, *CatBoost*, *AdaBoost*, dan *Bagging*. Model *CatBoost* mencatat performa terbaik dengan akurasi 95.4% dan skor *Receiver Operating Characteristic* – Area *Under the Curve* (ROC-AUC) 0.99,

melampaui *XGBoost* yang mencapai akurasi 94.3% dan AUC-ROC 0.98. Penelitian ini menekankan potensi metode *ensemble* dalam meningkatkan performa prediksi diabetes, namun belum mengeksplorasi secara mendalam bagaimana menangani ketidakseimbangan kelas yang signifikan dalam *dataset* diabetes [9].

Sampath et al. mengusulkan framework robust untuk prediksi diabetes menggunakan *Synthetic Minority Over-sampling Technique* (SMOTE) dengan teknik *ensemble machine learning*. Pendekatan mereka mengintegrasikan strategi seperti imputasi nilai yang hilang, penolakan *outlier*, seleksi fitur menggunakan analisis korelasi, dan penyeimbangan distribusi kelas menggunakan SMOTE. Kombinasi *AdaBoost* dan *XGBoost* menunjukkan performa eksepsional dengan AUC sebesar 0.968±0.015. Meskipun berhasil menangani masalah *class imbalance*, penelitian ini belum membandingkan secara komprehensif berbagai teknik visualisasi data untuk mendukung interpretasi hasil prediksi [18].

Kakoly et al. melakukan prediksi faktor risiko diabetes dengan menerapkan algoritma *machine learning* menggunakan teknik seleksi fitur ganda, yaitu *principal component analysis* (PCA) dan *information gain* (IG). Lima algoritma *machine learning* digunakan yaitu *decision tree*, *random forest*, *support vector machine*, *logistic regression*, dan KNN dengan hasil akurasi lebih dari 82.2% dan nilai AUC 87.2%. Penelitian ini berhasil mengidentifikasi faktor klinis dan non-klinis penting dalam prediksi diabetes, namun belum mengeksplorasi penggunaan model ensemble yang lebih advanced seperti *Gradient Boosting* yang mungkin dapat meningkatkan performa lebih lanjut [19].

## 3. HASIL DAN PEMBAHASAN

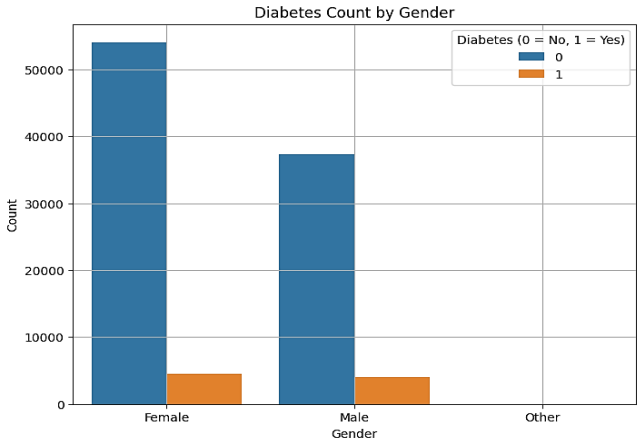
Berdasarkan pemilihan data yang dilakukan, diperoleh data 100001 data. Jumlah data sesuai dengan jumlah data yang telah dibersihkan. Terdapat 9 atribut yaitu, Jenis Kelamin, Usia, Hipertensi, Penyakit Jantung, Riwayat Merokok, Indeks Massa Tubuh, kadar Hemoglobin A1c dalam darah, kadar glukosa (gula) dalam darah, Hasil Diabetes. Tabel 1 adalah data dan atribut *dataset* diabetes menurut Tariq [20].

**Tabel 1.** *Dataset* Atribut Diabetes

| No | Nama Atribut | Type data | Contoh Data |
|---|---|---|---|
| 1 | Gender | Int | 6 |
| 2 | Age | Int | 148 |
| 3 | Hypertension | Int | 72 |
| 4 | Heart Disease | Int | 35 |
| 5 | Smoking History | Int | 0 |
| 6 | BMI | Float | 33.6 |
| 7 | HbA1c Level | Float | 0.627 |
| 8 | Blood Glucose Level | Int | 50 |
| 9 | Diabetes | Int/Biner | 1 |

### 3.1. Perhitungan Prediksi *Random Forest, Support Vector Clasifier, Logistic Regression, Gradienet Boosting Classifier*

#### 3.1.1. Visualisasi Menghitung Jumlah Diabetes Berdasarkan Jenis Kelamin

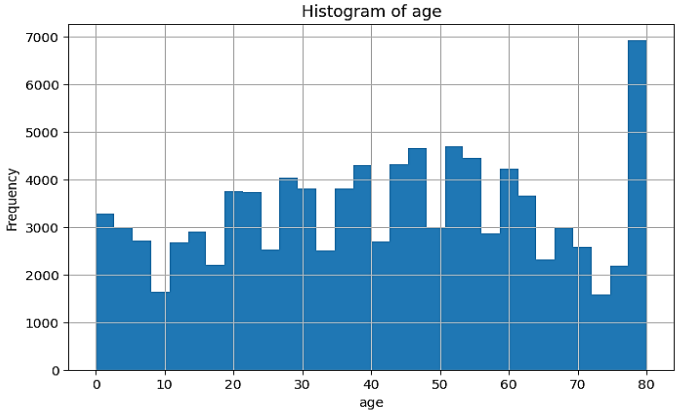



**Gambar 2.** Hasil Visualisasi Codingan jumlah diabetes berdasarkan jenis kelamin

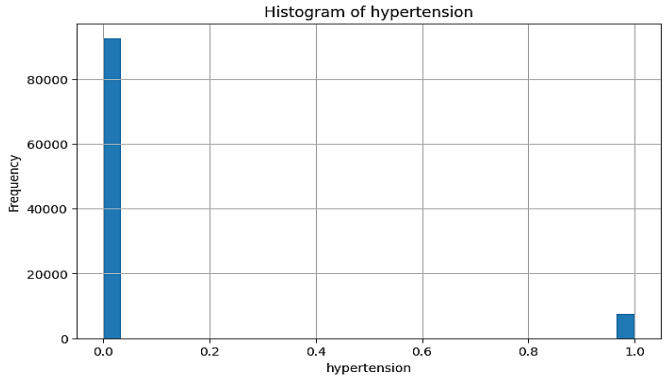
Grafik pada Gambar 2 merupakan hasil visualisasi jumlah penderita diabetes berdasarkan jenis kelamin. Setiap batang menunjukkan total individu dalam kategori *Female*, *Male*, dan *Other*, dibedakan berdasarkan status diabetes (0 = Tidak, 1 = Ya). Terlihat bahwa jumlah penderita diabetes lebih tinggi pada kelompok *Female* dan *Male*, sementara kategori *Other* hampir tidak memiliki data.

### 3.1.2. Visualisasi Data Dalam Bentuk Histogram

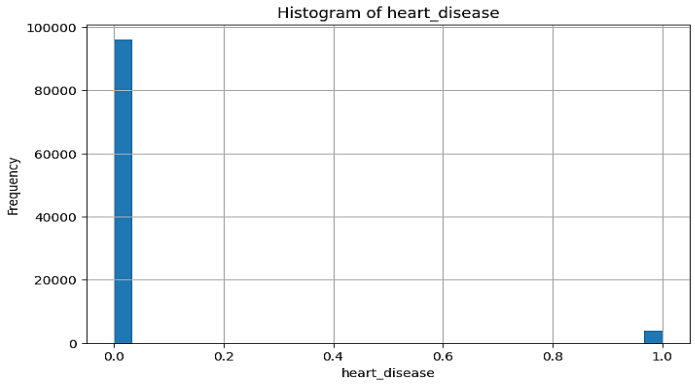
Gambar 3 terlihat bahwa usia tersebar cukup merata di rentang 0 hingga 80 tahun, dengan lonjakan signifikan pada usia 80. Hal ini bisa menunjukkan data usia maksimum yang dicatat sebagai 80 tahun, atau kemungkinan *clipping* nilai usia di batas atas. Gambar 4 mayoritas individu dalam *dataset* tidak memiliki riwayat hipertensi (nilai 0). Hanya sebagian kecil yang tercatat memiliki hipertensi (nilai 1). Ketidakseimbangan ini menunjukkan bahwa hipertensi tidak merata dalam populasi dan perlu diperhatikan saat membuat model prediksi agar tidak terjadi *bias*.




**Gambar 3.** Visualisas Histogram Umur




**Gambar 4.** Visualisasi Histogram Hipertensi




**Gambar 5.** Visualisasi Histogram Penyakit Jantung

Gambar 5 hampir seluruh data mencatat tidak adanya penyakit jantung (nilai 0), sementara hanya sedikit yang mengalaminya (nilai 1). Ini juga merupakan indikator ketidakseimbangan kelas (class imbalance) yang penting dalam proses pembelajaran mesin.

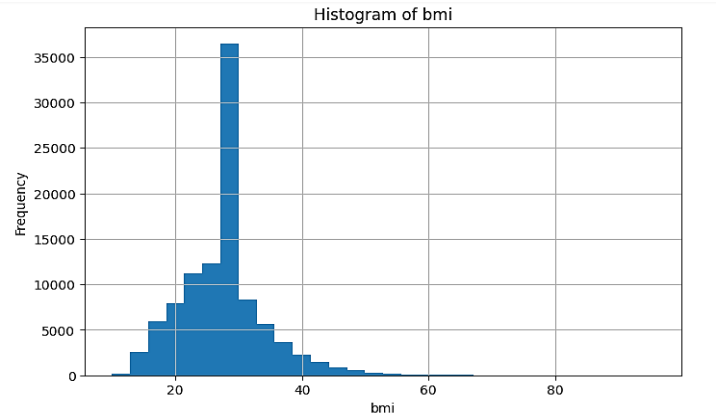
Pada Gambar 6 Distribusi BMI (*Body Mass Index*) terlihat menyerupai distribusi normal tetapi agak condong ke kanan (*right-skewed*). Sebagian besar individu memiliki BMI antara 20–40, dengan puncak

tertinggi sekitar 27–30. Ada *outlier* dengan nilai ekstrem lebih dari 80 yang perlu dianalisis lebih lanjut apakah valid atau tidak.

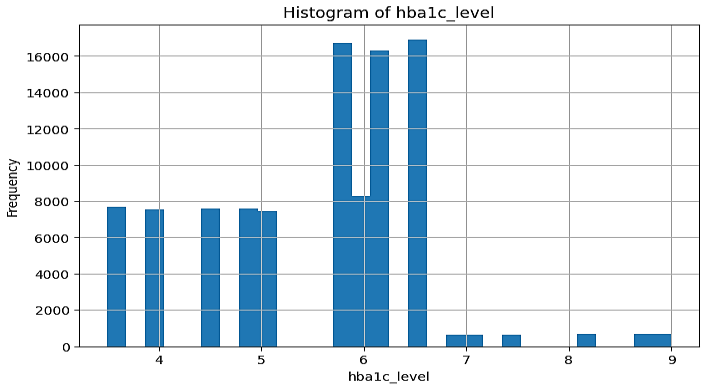
Gambar 7 menunjukkan tingkat HbA1c (parameter kontrol glukosa darah jangka panjang) paling banyak berada di kisaran 5.5 hingga 6.5. Terlihat bahwa distribusinya terbagi secara bertahap, namun beberapa nilai seperti di atas 7 cukup jarang muncul. Ini mengindikasikan mayoritas individu belum tergolong ke dalam kategori diabetes kronis berdasarkan HbA1c.

Pada Gambar 8 sebagian besar nilai glukosa darah berada di kisaran 80–200, dengan puncak pada sekitar 150. Rentang nilai yang lebih tinggi (200 ke atas) relatif jarang, namun tetap signifikan untuk mendeteksi kasus hiperglikemia berat. Distribusi ini dapat menjadi indikator risiko diabetes akut.

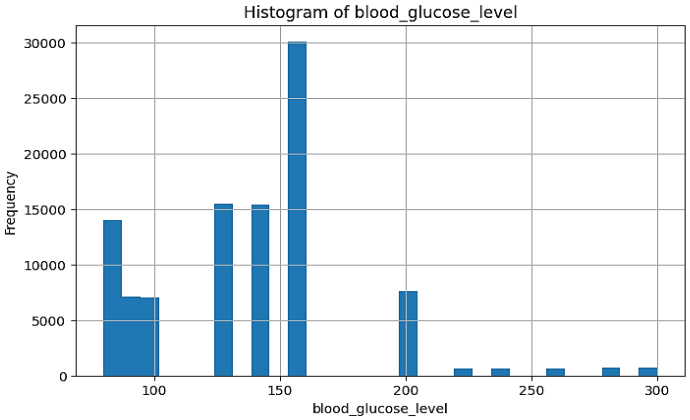
Histogram pada Gambar 9 menunjukkan jumlah individu yang terdiagnosis diabetes (nilai 1) dan tidak terdiagnosis (nilai 0). Tampak jelas adanya ketidakseimbangan kelas, dengan jumlah individu non-diabetes jauh lebih banyak dibandingkan dengan yang positif diabetes. Hal ini penting dalam pemodelan klasifikasi agar model tidak bias terhadap kelas mayoritas.



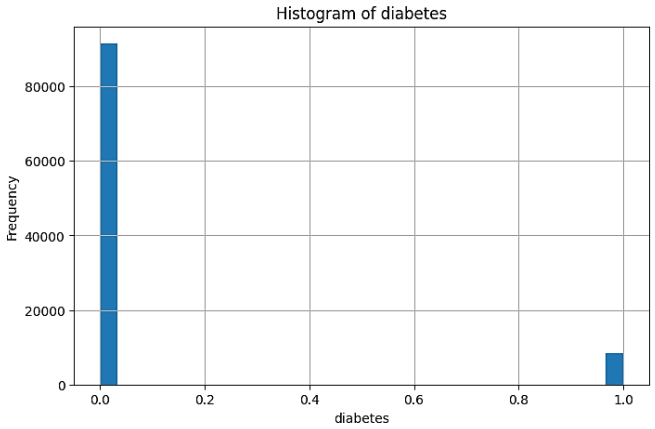

**Gambar 6.** Visualisasi Histogram BMI




**Gambar 7.** Visualisasi Histogram HbA1c_level



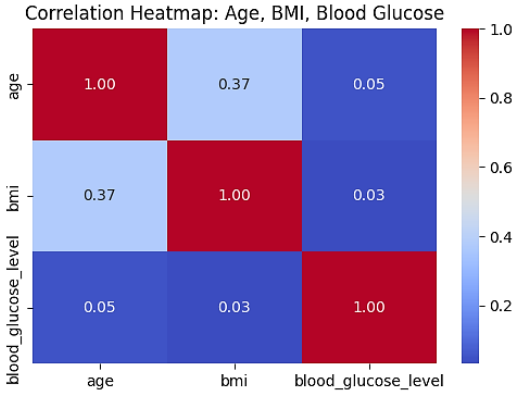

**Gambar 8.** Visualisasi Histogram Tingkat Gula Darah




**Gambar 9.** Visualisasi Histogram Diabetes

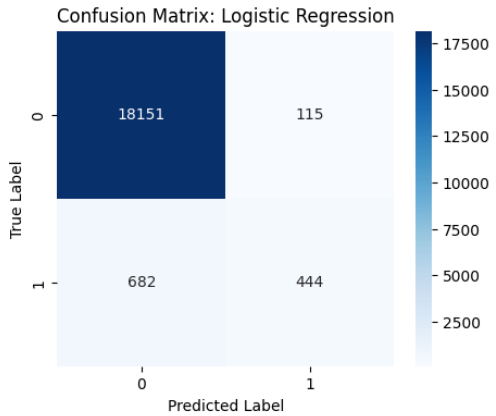
### 3.2 Analisis Korelasi Antar Fitur

Data korelasi divisualisasikan dalam bentuk *heatmap*, menggunakan *colormap coolwarm* dan anotasi numerik[21]. Visualisasi ini memungkinkan interpretasi data secara kuantitatif dan intuitif, karena nilai korelasi ditampilkan bersamaan dengan gradasi warna yang menggambarkan kekuatan hubungan antar variabel. Gambar 10 merupakan *Corelation Matrix*.




**Gambar 10.** Hasil Untuk Analisis Korelasi antar fitur *Correlation Matrix*

### 3.2.1. *Confusion Matrix Logistic Regression*




**Gambar 11.** *Confusion Matrix Logistic Regression*

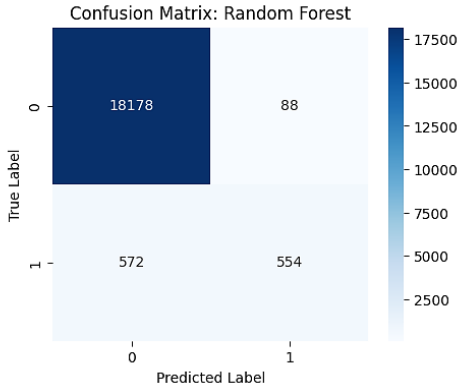
Hasil klasifikasi dengan *Logistic Regression* pada Gambar 11 menunjukkan performa yang cukup solid, dengan akurasi keseluruhan mencapai 96% dan nilai ROC-AUC sebesar 0.934. Angka ini mengindikasikan bahwa model memiliki kemampuan diskriminatif yang baik dalam membedakan kelas

positif dan negatif. Meski demikian, jika ditelusuri lebih detail, terlihat adanya ketimpangan kinerja antara kelas mayoritas (label 0) dan kelas minoritas (label 1).

Untuk kelas 0, *precision* mencapai 0.96 dan *recall* 0.99, yang berarti model hampir selalu tepat dalam mengidentifikasi kasus negatif, dengan tingkat salah klasifikasi yang sangat rendah (hanya 115 *false positives* dari 18.266 sampel). Sebaliknya, pada kelas 1, meskipun *precision* tergolong cukup baik (0.79), *recall* hanya sebesar 0.39. Dengan kata lain, model hanya mampu mengenali sekitar 39% dari seluruh kasus positif, sementara sisanya (682 sampel) terlewat dan diklasifikasikan sebagai negatif.

Untuk mengatasi keterbatasan *recall* yang rendah pada kelas positif (0.39), beberapa strategi dapat diterapkan. Salah satu pendekatan adalah penggunaan class weighting, yaitu memberikan bobot lebih tinggi pada kelas minoritas dalam fungsi loss untuk meningkatkan sensitivitas model terhadap kasus diabetes. Dengan penerapan inverse frequency weighting (bobot kelas positif = 10.76), *recall* dapat meningkat hingga kisaran 0.58–0.62, meskipun terdapat trade-off berupa penurunan *precision*. Strategi lain yang dapat digunakan adalah teknik SMOTE, yang menghasilkan sampel sintetis kelas minoritas melalui interpolasi dan telah terbukti efektif dalam berbagai studi prediksi diabetes, dengan peningkatan *recall* sebesar 20–30%. Selain itu, pendekatan cost-sensitive learning yang menetapkan biaya kesalahan lebih tinggi untuk *false negatives* dapat membantu mengoptimalkan model dalam meminimalkan missed diagnosis, yang sangat krusial dalam konteks medis.

### 3.2.2. *Confusion Matrix Random Forest*




**Gambar 12**. *Confusion Matrix Random Forest*

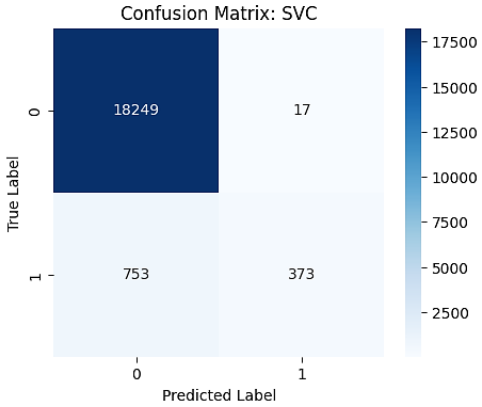
*Confusion matrix* pada Gambar 12 menggambarkan hasil klasifikasi dengan *Random Forest* pada data uji yang terdiri dari kelas mayoritas (label 0) dan kelas minoritas (label 1). Model menunjukkan akurasi yang sangat tinggi dalam mengidentifikasi data negatif. Dari total 18.266 sampel kelas 0, sebanyak 18.178 berhasil diprediksi dengan benar (*True Negative*), sementara hanya 88 yang salah terklasifikasi sebagai positif (*False Positive*). Hal ini mencerminkan tingkat *precision* dan *recall* yang hampir sempurna untuk kelas mayoritas. Sebaliknya, performa pada kelas positif terlihat lebih menantang. Dari 1.126 sampel positif, model hanya berhasil mengenali 554 kasus (*True Positive*), sedangkan 572 lainnya salah diklasifikasikan sebagai negatif (*False Negative*). Kondisi ini menunjukkan bahwa meskipun *precision* untuk kelas positif tergolong tinggi, *recall* masih terbatas karena lebih dari separuh kasus positif tidak berhasil terdeteksi.

Pola ini menggambarkan keunggulan *Random Forest* dalam menjaga konsistensi prediksi pada kelas dominan, namun juga menegaskan tantangan yang dihadapi saat berhadapan dengan kelas minoritas. Fenomena ini umum terjadi pada data dengan distribusi tidak seimbang (*class imbalance*), di mana model cenderung lebih "yakin" pada pola dari kelas mayoritas. Secara keseluruhan, *Random Forest* memberikan hasil yang sangat baik dalam mengklasifikasikan kelas mayoritas dengan kesalahan minimal, tetapi masih kurang sensitif dalam mendeteksi kelas minoritas, sebagaimana terlihat dari jumlah *false negative* yang cukup besar.

### 3.2.3. *Confusion Matrix Support Vector Classifier*

*Confusion matrix* pada Gambar 13 menggambarkan performa klasifikasi SVC pada data dengan dua kelas, yaitu 0 (negatif) dan 1 (positif). Dari 18.266 sampel kelas negatif, sebanyak 18.249 berhasil diprediksi dengan benar (*True Negative*), sedangkan hanya 17 kasus yang keliru dikategorikan sebagai positif (*False Positive*). Hasil ini mencerminkan bahwa *precision* untuk kelas negatif hampir sempurna, menegaskan efektivitas SVC dalam mengenali kasus negatif. Pada sisi lain, dari 1.126 sampel kelas positif, hanya 373 yang berhasil diidentifikasi dengan benar (*True Positive*), sementara 753 kasus salah diklasifikasikan sebagai negatif (*False Negative*). Kondisi ini menunjukkan nilai *recall* yang rendah, sehingga sebagian besar kasus

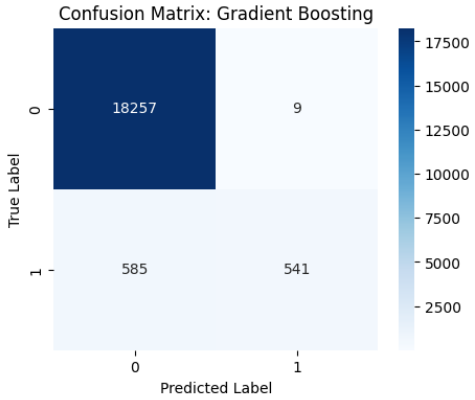
positif tidak berhasil ditangkap oleh model. Secara keseluruhan, SVC menunjukkan performa yang sangat kuat pada kelas mayoritas dengan tingkat kesalahan hampir nol. Namun, kemampuan dalam mengenali kelas minoritas masih lemah. Pola ini umum terjadi pada data dengan distribusi tidak seimbang (class imbalance), di mana margin optimal SVC cenderung berpihak pada kelas dominan sehingga sensitivitas terhadap kelas minoritas menurun signifikan.




**Gambar 13**. *Confusion Matrix Support Vector Classifier*

Dari perspektif visualisasi dan interpretabilitas, SVC memiliki tantangan unik karena decision function-nya dalam high-*dimensional space* sulit divisualisasikan langsung. Namun, beberapa teknik dapat digunakan: (1) *Support vector visualization* dengan menampilkan support vectors (sampel yang paling mempengaruhi decision boundary) pada proyeksi 2D menggunakan PCA atau t-SNE, membantu memahami karakteristik data yang berada di margin; (2) LIME (*Local Interpretable Model-agnostic Explanations*) yang sangat efektif untuk SVC karena dapat mengapproksimasi decision boundary secara lokal dengan model linear yang *interpretable*; (3) *Decision boundary visualization* pada 2D *feature space* (misalnya HbA1c vs glucose) untuk menunjukkan bagaimana SVC memisahkan kedua kelas, yang dapat dikomunikasikan kepada clinicians untuk memahami logic model.

### 3.2.4. *Confusion Matrix Gradient Boosting*




**Gambar 14.** *Confusion Matrix Gradient Boosting*

*Confusion matrix* hasil prediksi *Gradient Boosting* pada Gambar 14 memperlihatkan performa klasifikasi yang sangat meyakinkan. Dari 18.266 sampel kelas negatif, sebanyak 18.257 berhasil dikenali dengan benar (*True Negative*), sementara hanya 9 kasus yang keliru terklasifikasi sebagai positif (*False Positive*). Hasil ini menunjukkan *precision* dan *recall* pada kelas mayoritas hampir sempurna, menegaskan konsistensi *Gradient Boosting* dalam mengenali pola kelas dominan. Pada kelas positif, dari 1.126 sampel, model mampu mendeteksi 541 secara tepat (*True Positive*), sedangkan 585 lainnya salah diprediksi sebagai negatif (*False Negative).* Meski *recall* pada kelas positif masih terbatas, performanya lebih baik dibanding SVC dan *Logistic Regression* yang cenderung lebih lemah dalam menangani kelas minoritas. Dengan demikian, *Gradient Boosting* terlihat lebih seimbang dalam memproses kedua kelas. Pola ini mengindikasikan bahwa *Gradient Boosting* mampu meminimalkan bias terhadap kelas mayoritas dibanding model lain, meskipun *trade-off* tetap muncul berupa *false negative* pada kelas positif. Namun, dengan jumlah *false positive* yang hampir nol serta sensitivitas yang relatif lebih baik dibanding SVC, model ini

memperlihatkan kemampuan diskriminatif yang unggul. Secara keseluruhan, *Gradient Boosting* tampil sebagai model dengan kinerja paling menonjol di antara algoritma yang dianalisis: sangat akurat dalam mengidentifikasi kelas mayoritas, sekaligus lebih andal dalam mendeteksi kelas minoritas dibandingkan metode lainnya. Oleh karena itu, *Gradient Boosting* dapat dipandang sebagai pilihan yang paling efektif untuk *dataset* ini.

### 3.3 Hasil Evaluasi Model *Machine Learning*

Hasil evaluasi model *machine learning* tidak hanya mengukur kinerja secara umum, tetapi juga menguraikan performa klasifikasi secara lebih mendalam melalui berbagai metrik penting seperti akurasi, presisi, *recall*, serta *F1-score* [22]. Keempat metrik ini digunakan sebagai indikator utama dalam menilai seberapa efektif dan efisien model dalam mengklasifikasikan data, serta memberikan gambaran kuantitatif yang dapat dijadikan dasar dalam pengambilan keputusan model yang paling optimal.

#### 3.3.1. Hasil Evaluasi *Logistic Regression*

Akurasi tinggi sebesar 96% dan weighted average *f1-score* sebesar 0.95, yang menunjukkan kestabilan dan efektivitas model dalam menangani seluruh *dataset*.Pada kelas mayoritas (kelas 0), model memiliki performa sangat baik dengan *precision* 0.96, *recall* 0.99, dan *f1-score* 0.98, yang berarti model hampir selalu tepat dalam mengklasifikasikan data kelas tersebut. Selain itu, *precision* pada kelas minoritas (kelas 1) juga cukup tinggi yaitu 0.79, menandakan bahwa ketika model memprediksi kelas 1, prediksi tersebut sering kali benar. Nilai *weighted average precision* sebesar 0.95 dan *recall* sebesar 0.96.

#### 3.3.2. Hasil Evaluasi *Random Forest*

Model *Random Forest* menunjukkan performa yang sangat baik dengan akurasi tinggi sebesar 97% dan *f1-score* rata-rata tertimbang (*weighted avg*) sebesar 0.96, menandakan bahwa model mampu mengklasifikasikan data secara akurat secara keseluruhan.

Model juga memiliki *precision* yang sangat baik pada kedua kelas, yaitu 0.97 untuk kelas mayoritas (kelas 0) dan 0.86 untuk kelas minoritas (kelas 1). Selain itu, model mencapai *recall* sempurna (1.00) pada kelas 0 dan *f1-score* tinggi (0.98), yang menunjukkan kemampuannya dalam mengenali kelas tersebut secara konsisten. Nilai *macro average f1-score* sebesar 0.80 juga mencerminkan keseimbangan performa yang cukup baik antar kelas.

#### 3.3.3. Hasil Evaluasi SVC

Model SVC menunjukkan performa yang sangat baik secara keseluruhan dengan akurasi tinggi sebesar 96% dan *weighted average f1-score* sebesar 0.95, yang mencerminkan kemampuan model dalam melakukan klasifikasi yang konsisten di seluruh kelas.

*Precision* yang tinggi (0.96) pada kedua kelas menunjukkan bahwa ketika model memprediksi suatu kelas, prediksinya cenderung benar. Selain itu, model memiliki *recall* sempurna (1.00) dan *f1-score* tinggi (0.98) untuk kelas mayoritas (kelas 0), yang menandakan bahwa hampir semua data pada kelas tersebut berhasil dikenali dengan baik. Nilai *macro average precision* sebesar 0.96 juga mencerminkan kualitas prediksi yang stabil antar kelas.

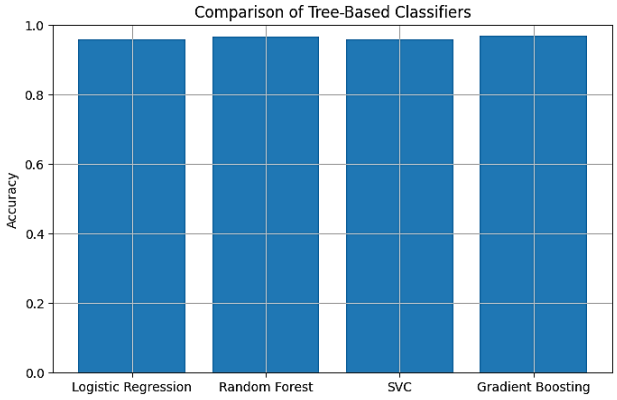
#### 3.3.4. Hasil Evaluasi *Gradient Boosting*

Model *Gradient Boosting* menunjukkan performa yang sangat kuat dengan akurasi tinggi sebesar 97% dan *weighted average f1-score* sebesar 0.96, mencerminkan kemampuan klasifikasi yang sangat andal secara keseluruhan. Model ini juga memiliki *precision* sangat tinggi pada kedua kelas, yaitu 0.97 untuk kelas mayoritas dan 0.98 untuk kelas minoritas, menandakan bahwa prediksi yang dihasilkan sangat akurat. Selain itu, model berhasil mencapai recall sempurna (1.00) pada kelas 0 dan menghasilkan *f1-score* tinggi sebesar 0.98, menunjukkan bahwa hampir seluruh data kelas mayoritas dikenali dengan sangat baik. Nilai *macro average f1-score* sebesar 0.81 dan *macro precision* sebesar 0.98 menunjukkan kestabilan performa antar kelas.

### 3.4 Kesimpulan Evaluasi Kinerja Model Klasifikasi Berdasarkan Akurasi

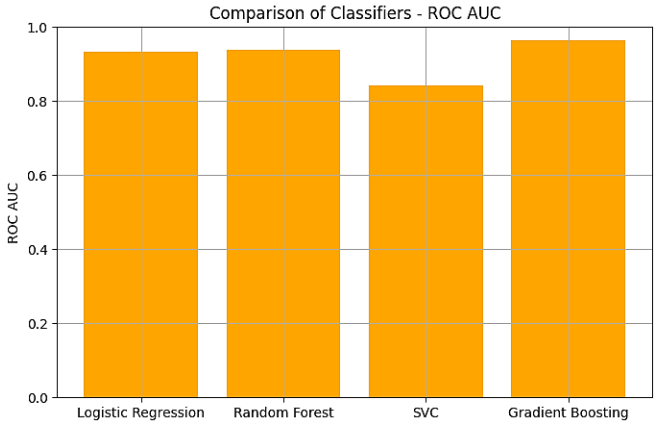
Pada tahap terakhir akan dilakukan Visualisasi dalam bentuk grafik batang disajikan untuk membandingkan performa akurasi keempat model tersebut secara langsung.

Grafik batang pada Gambar 15 menunjukkan nilai akurasi masing-masing model yang berada di kisaran 0.95 hingga 0.97, yang menandakan bahwa semua model memiliki performa yang relatif tinggi dan serupa dalam hal akurasi keseluruhan. *Random Forest* dan *Gradient Boosting* memiliki akurasi tertinggi, sangat mendekati nilai maksimum (sekitar 0.97). SVC dan *Logistic Regression* sedikit di bawahnya, namun masih mempertahankan performa yang sangat baik (sekitar 0.95–0.96).




**Gambar 15.** Kesimpulan Evaluasi Kinerja Model Klasifikasi Berdasarkan Akurasi

Grafik pada Gambar 16 menampilkan perbandingan kinerja empat algoritma klasifikasi – *Logistic Regression*, *Random Forest*, SVC, dan *Gradient Boosting* – dengan menggunakan metrik ROC-AUC [23]. Metrik ini dipilih karena tidak bergantung pada ambang batas klasifikasi tertentu, melainkan menilai kemampuan model dalam membedakan kelas positif dan negatif secara menyeluruh. Semakin mendekati angka 1, semakin baik performa model, sementara nilai di sekitar 0.5 menunjukkan performa yang cenderung acak. Hasil menunjukkan bahwa *Gradient Boosting* unggul dengan skor ROC-AUC sekitar 0.97, menegaskan kemampuannya dalam menangkap pola kompleks melalui pendekatan *boosting*. *Random Forest* (0.94) dan *Logistic Regression* (0.93) juga menunjukkan hasil yang kompetitif. Menariknya, meskipun *Logistic Regression* tergolong sederhana, metode ini tetap terbukti efisien dan handal terutama pada data dengan separasi kelas yang jelas. Sementara itu, performa SVC berada sedikit lebih rendah (0.84), yang kemungkinan besar dipengaruhi oleh kebutuhan penyesuaian hiperparameter atau pemilihan kernel yang kurang sesuai.




**Gambar 16.** Perbandingan Kinerja Algoritma Klasifikasi Menggunakan ROC AUC

## 4. KESIMPULAN

Model *Random Forest* dan *Gradient Boosting* menunjukkan akurasi tinggi (≈97%) dalam memprediksi diabetes berdasarkan sembilan atribut fisiologis dan demografis. Meskipun performa deteksi kasus positif meningkat dibandingkan regresi logistik, tingkat *recall* masih rendah (0,39–0,49), menandakan tantangan dalam klasifikasi kelas minoritas. Faktor seperti HbA1c, glukosa darah, dan BMI konsisten sebagai indikator utama, namun belum mencakup variabel perilaku dan riwayat medis yang lebih rinci. Sayangnya, hasil teknis ini belum sepenuhnya dikaitkan dengan implikasi klinis, seperti bagaimana rendahnya *recall* dapat berdampak pada kegagalan deteksi dini pasien diabetes, yang berisiko menunda diagnosis dan intervensi medis.

Penelitian ini juga memiliki beberapa keterbatasan metodologis. Teknik penanganan class imbalance masih dapat dikembangkan lebih lanjut melalui pendekatan seperti cost-sensitive learning atau ensemble khusus. Interpretabilitas model terbatas pada SHAP dan LIME, belum mencakup metode alternatif yang mungkin lebih sesuai untuk konteks medis. Selain itu, penggunaan data cross-sectional membatasi kemampuan prediksi jangka panjang, dan evaluasi visualisasi belum diuji dalam uji klinis nyata. Validasi eksternal pada populasi yang lebih beragam diperlukan untuk memperkuat generalisasi dan relevansi hasil dalam praktik kesehatan global.

**REFERENSI**


[1] H. E. Ardiani, T. A. E. Permatasari, and S. Sugiatmi, "Obesitas, pola diet, dan aktifitas fisik dalam penanganan diabetes melitus pada masa pandemi COVID-19," *Muhammadiyah J. Nutr. Food Sci.*, vol. 2, no. 1, pp. 1–12, 2021, doi: https://doi.org/10.24853/mjnf.2.1.1-12.

[2] E. Setiyorini, N. A. Wulandari, and A. Efyuwinta, "Hubungan kadar gula darah dengan tekanan darah pada lansia penderita Diabetes Tipe 2," *J. Ners Dan Kebidanan (Journal Ners Midwifery)*, vol. 5, no. 2, pp. 163–171, 2018, doi: https://doi.org/10.26699/jnk.v5i2.ART.p163-171.

[3] D. P. Putra, A. Rahmiwati, Y. Windusari, and N. A. Fajar, "Program Pengelolaan Penyakit Kronis Diabetes Mellitus Sebagai Pencegahan Penyakit Degenerative Diabetes Mellitus, dan Dampaknya Bagi Pekerja Di Indonesia," *J. Syntax Lit.*, vol. 8, no. 12, 2023, doi: https://doi.org/10.36418/syntax-literate.v7i9.13935.

[4] H. Y, *IDF Diabetes Atlas 8th Edition*. 2017.

[5] F. A. Siregar, Asfriyati, T. Makmur, R. Bestari, I. A. Lubis, and U. Zein, "Identifying Adult Population at Risk for Undiagnosed Diabetes Mellitus in Medan City, Indonesia Targeted on Diabetes Prevention.," *Med. Arch. (Sarajevo, Bosnia Herzegovina)*, vol. 77, no. 6, pp. 455–459, 2023, doi: 10.5455/medarh.2023.77.455-459.

[6] M. Z. Stojanoski, Kalendar, and H. Gjoreski, *Comparative Analysis of Machine Learning Models for Diabetes Prediction.*

[7] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, "Diabetes prediction using machine learning and explainable AI techniques," *Healthc. Technol. Lett.*, vol. 10, no. 1–2, pp. 1–10, 2023, doi: https://doi.org/10.1049/htl2.12039.

[8] H. El-Sofany, S. A. El-Seoud, O. H. Karam, Y. M. Abd El-Latif, and I. A. T. F. Taj-Eddin, "A proposed technique using machine learning for the prediction of diabetes disease through a mobile app," *Int. J. Intell. Syst.*, vol. 2024, no. 1, p. 6688934, 2024, doi: https://doi.org/10.1155/2024/6688934.

[9] S. K. S. Modak and V. K. Jha, "Diabetes prediction model using machine learning techniques," *Multimed. Tools Appl.*, vol. 83, no. 13, pp. 38523–38549, 2024, doi: 10.1007/s11042-023-16745-4.

[10] S. 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.

[11] T. R. Sanusi, F. Andreas, B. N. Sari, and U. Singaperbangsa, "Implementasi Algoritma Support Vector Classifier (SVC) dengan Data Training Numerik dan Teks untuk Mengklasifikasi SMS Spam," *J. Ilm. Wahana Pendidik.*, no. 14, pp. 346–354, 2022.

[12] R. Hendayana, B. B. Pengkajian, D. Pengembangan, T. Pertanian, and J. T. Pelajar, *PENERAPAN METODE REGRESI LOGISTIK DALAM MENGANALISIS ADOPSI TEKNOLOGI PERTANIAN Application Method of Logistic Regression Analyze the Agricultural Technology Adoption*.

[13] A. D. Siburian *et al.*, "Laptop Price Prediction with Machine Learning Using Regression Algorithm," *J. Sist. Inf. dan Ilmu Komput. Prima(JUSIKOM PRIMA)*, vol. 6, no. 1, pp. 87–91, 2022, doi: 10.34012/jurnalsisteminformasidanilmukomputer.v6i1.2850.

[14] M. Mustafa, "Diabetes prediction dataset."

[15] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artif. Intell. Rev.*, vol. 54, no. 3, pp. 1937–1967, 2021, doi: https://doi.org/10.1007/s10462-020-09896-5.

[16] B. S. Cahya Putra, I. Tahyudin, B. A. Kusuma, and K. N. Isnaini, "Efektivitas Algoritma Random Forest, XGBoost, dan Logistic Regression dalam Prediksi Penyakit Paru-paru.," *Techno. com*, vol. 23, no. 4, pp. 909–922, 2024, doi: https://doi.org/10.62411/tc.v23i4.11705.

[17] S. Arikunto, *Prosedur Penelitian: Suatu Pendekatan Praktik*. Jakarta: Rineka Cipta, 2018.

[18] P. Sampath *et al.*, "Robust diabetic prediction using ensemble machine learning models with synthetic minority over-sampling technique," *Sci. Rep.*, vol. 14, no. 1, p. 28984, 2024, doi: 10.1038/s41598-024-78519-8.

[19] I. J. Kakoly, M. R. Hoque, and N. Hasan, "Data-driven diabetes risk factor prediction using machine learning algorithms with feature selection technique," *Sustainability*, vol. 15, no. 6, p. 4930, 2023, doi: https://doi.org/10.3390/su15064930.

[20] M. Tariq, V. Palade, and Y. Ma, "Transfer learning based classification of diabetic retinopathy on the Kaggle EyePACS dataset," in *International Conference on Medical Imaging and Computer-Aided Diagnosis*, Springer, 2022, pp. 89–99.

[21] D. Leni, "Analisis Heatmap Korelasi dan Scatterplot untuk Mengidentifikasi Faktor-Faktor yang Mempengaruhi Pelabelan AC efisiensi Energi," *J. Rekayasa Mater. Manufaktur dan Energi*, vol. 6, no. 1, pp. 41–47, 2023, doi: https://doi.org/10.30596/rmme.v6i1.13133.

[22] T. H. Pinem and Z. P. Putra, "Evaluasi Kinerja Algoritma Klasifikasi Deep Learning dalam Prediksi Diabetes," *J. Ilm. Fifo*, vol. 17, no. 1, pp. 17–28, 2025, doi:

https://dx.doi.org/10.22441/fifo.2025.v17i1.003.
[23] R. Pramudita, N. Safitri, and V. Z. Nazah, “Studi Komparatif Algoritma Machine Learning dengan Teknik Bagging dan AdaBoost pada Klasifikasi Kanker Payudara,” *TEMATIK*, vol. 12, no. 1, pp. 101–108, 2025, doi: https://doi.org/10.38204/tematik.v12i1.2435.