FORMAT: Jurnal Ilmiah Teknik Informatika
Akreditasi Sinta 5 RISTEKBRIN No. 164/E/KPT/2021; E-ISSN 2722-7162 | P-ISSN 2089-5615

Vol. 14, No. 1,
January 2025,
13-23. 2025

Pengembangan Model Prediksi Diabetes Melitus Menggunakan Metode *Stochastic Gradient Boosting*

Andrian Sah [1] ; Chaeroen Niesa [2] ; Amat Damuri [3] ; Nur Amalia Hasma [4]

[1,] *Program Studi Sistem Informasi, Fakultas Ilmu Komputer, Universitas Yopis Papua, Jl. Dr.Sam Ratulangi No.11, Trikora, Kec.Jayapura Utara, Kotaa Jayapura, Papua 99113.*
[2] *Program Studi Informatika, Fakultas Komputer dan Multimedia, Universitas Islam Kebangsaan Indonesia, Jl. Medan – Banda Aceh, Bladeh, Kec. Jeumpa, Kabupaten Bireuen, Aceh 24251*
[3] *Program Studi Manajemen Informatika, STMIK Al Muslim, Jl. Raya Setu, Kp. Bahagia, Kec. Tambun Selatan, Kabupaten Bekasi, Jawa Barat 17510*
[4] *Program Studi Sistem Informasi, Fakultas Komputer dan Multimedia, Universitas Islam Kebangsaan Indonesia, Jl. Medan – Banda Aceh, Bladeh, Kec. Jeumpa, Kabupaten Bireuen, Aceh 24251*

---

[1] cyberdefance23@gmail.com, [2] jeumalaniesa@gmail.com, [3] amat.damuri@almuslim.ac.id, [4] nuramaliahasma05@gmail.com

Kata kunci:
Data Mining, Diabetes Mellitus, Prediction Model, Machine Learning, Stochastic Gradient Boosting

Abstract

Diabetes mellitus is one of the global health issues with a continuously increasing prevalence. The high prevalence of diabetes mellitus has a significant impact on economic burdens and healthcare systems, as it often leads to severe complications such as cardiovascular disease and kidney failure. Therefore, early prediction and detection of diabetes mellitus are crucial aspects in mitigating its adverse effects. Data mining and machine learning technologies offer innovative solutions for processing complex medical data, providing deeper insights, and supporting data-driven decision-making. This study aims to develop a diabetes mellitus prediction model using the Stochastic Gradient Boosting (SGB) algorithm. The model utilizes a dataset comprising clinical variables such as glucose levels, blood pressure, body mass index (BMI), and genetic history to identify diabetes risks. The results indicate that the developed prediction model achieved a high accuracy of 92.75%, with an Area Under the Curve (AUC) value of 0.98, demonstrating the model's ability to differentiate between positive (diabetes) and negative (non-diabetes) classes. Consistent values for precision, recall, and F1-score across both classes indicate the model's reliability in capturing complex patterns in the dataset. The model's high accuracy is attributed to the iterative boosting approach in the SGB algorithm, which adaptively corrects prediction errors at each iteration. Additionally, regulatory mechanisms such as learning rate and subsampling help prevent overfitting, making the algorithm effective for datasets with complex patterns.

Pendahuluan

Diabetes melitus merupakan salah satu masalah kesehatan global yang semakin meningkat prevalensinya. Menurut laporan terbaru dari *International Diabetes Federation* (IDF), pada tahun 2023 terdapat sekitar 537 juta orang dewasa (usia 20-79 tahun) di seluruh dunia yang hidup dengan diabetes [1]. Di Indonesia, diabetes adalah penyebab kematian tertinggi ketiga setelah stroke dan penyakit jantung [2]. Data Kementerian Kesehatan mencatat bahwa pada tahun 2022, prevalensi diabetes melitus pada penduduk usia di atas 15 tahun mencapai 11,7%, meningkat dari 10,9% pada tahun sebelumnya [3]. Peningkatan ini menunjukkan bahwa lebih dari 19 juta penduduk Indonesia menderita diabetes, atau sekitar 10,8% dari total populasi dewasa [3]. Kondisi ini diperparah dengan tingginya proporsi penderita diabetes yang tidak terdiagnosis, mencapai 73,7% pada kelompok usia 20-79 tahun [4]. Fenomena ini menambah beban ekonomi dan sistem kesehatan nasional, mengingat diabetes sering kali menyebabkan komplikasi serius seperti penyakit kardiovaskular, gagal ginjal, dan amputasi. Dalam menghadapi tantangan ini, prediksi dan deteksi dini menjadi aspek yang sangat penting. Teknologi dapat menjadi solusi dalam menyelesaikan berbagai tantangan kompleks, seperti meningkatkan efisiensi, mempercepat proses, memberikan akses informasi yang lebih luas, serta mendukung pengambilan keputusan yang lebih akurat di berbagai bidang [5], [6]. Teknologi *data mining* dan *machine learning* telah menjadi solusi inovatif untuk memproses data medis yang kompleks dan menyediakan wawasan yang lebih mendalam terhadap pola-pola risiko diabetes. *Data mining* memungkinkan penggalian pengetahuan dari dataset besar [7], sementara *machine learning* mampu belajar dari data historis untuk membuat prediksi yang akurat dan *real-time* [8].

Penelitian sebelumnya terkait prediksi atau klasifikasi diabetes melitus telah banyak dilakukan dengan berbagai pendekatan. Salah satu penelitian menggunakan algoritma *K-Nearest Neighbors* (K-NN) dengan dan tanpa seleksi fitur *Information Gain* [9]. Hasilnya menunjukkan bahwa model K-NN tanpa *Information Gain* memiliki akurasi sebesar 69,11%, sedangkan model K-NN

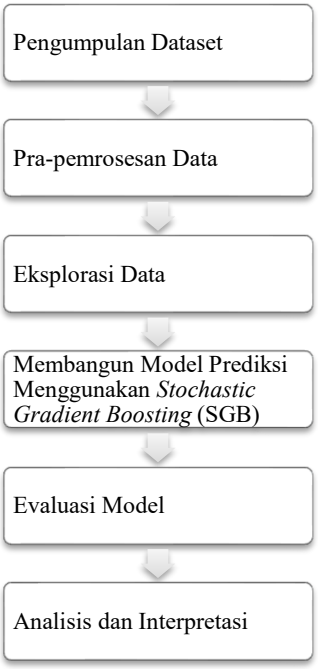
dengan *Information Gain* mencapai akurasi tertinggi sebesar 72,93% pada nilai K=17. Penelitian lain menggunakan algoritma *Naïve Bayes* untuk mengklasifikasikan diabetes, dengan hasil akurasi sebesar 73,16% dan nilai Area *Under the Curve* (AUC) pada kurva ROC sebesar 0,818 [10]. Pendekatan lain yang digunakan adalah *Support Vector Machine* (SVM), yang dalam pengujiannya menghasilkan rata-rata akurasi 0,87, *precision* 0,82, *sensitivity* 0,78, dan *specificity* 0,92 [11]. Selain itu, ada penelitian yang membandingkan algoritma AdaBoost (*Adaptive Boosting*) dengan LightGBM (*Light Gradient Boosting Machine*) [12]. Hasilnya menunjukkan bahwa algoritma LightGBM memiliki akurasi lebih tinggi (91,67%) dan nilai AUC sebesar 0,9704 dibandingkan dengan algoritma AdaBoost yang memiliki akurasi 91,14% dan nilai AUC sebesar 0,9693.

Berdasarkan berbagai hasil penelitian tersebut, pendekatan dengan teknik *boosting* terbukti mampu meningkatkan akurasi prediksi dibandingkan dengan metode tradisional. Teknik *boosting* unggul karena kemampuannya untuk memperbaiki kelemahan model sederhana (*weak learners*) melalui proses iteratif yang memperkuat data sulit [13]. Namun, teknik *boosting* seperti AdaBoost dan LightGBM memiliki kelemahan, antara lain kecenderungan untuk *overfitting* pada dataset kecil atau terlalu kompleks, serta tingginya kebutuhan komputasi terutama pada dataset besar [14]. Salah satu pendekatan yang populer dalam teknik boosting adalah *Stochastic Gradient Boosting* (SGB). Metode ini dipilih dalam penelitian ini karena kemampuannya untuk menangani dataset besar dengan fitur yang kompleks, mengatasi *overfitting* melalui regulasi (seperti *shrinkage* dan *subsampling*), serta menghasilkan prediksi yang lebih akurat [15]. Keunggulan SGB terletak pada pendekatan iteratifnya, di mana model secara bertahap mempelajari kesalahan dari model sebelumnya dan memprioritaskan perbaikan pada data yang sulit diprediksi [16].

Penelitian ini bertujuan untuk membangun model prediksi diabetes melitus menggunakan metode *Stochastic Gradient Boosting* (SGB). Model ini memanfaatkan dataset berisi variabel klinis seperti kadar glukosa, tekanan darah, indeks massa tubuh (BMI), dan riwayat genetik untuk mengidentifikasi risiko diabetes. Dengan adanya model prediksi ini, dapat berkontribusi dalam memanfaatkan teknologi untuk meningkatkan deteksi dini diabetes secara lebih akurat dan efisien, sehingga membantu tenaga medis dan pihak terkait dalam pengambilan keputusan.

Metode penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan memanfaatkan algoritma *Stochastic Gradient Boosting* (SGB) dalam membangun model prediksi diabetes melitus. Penelitian ini menggunakan metode eksperimen untuk menguji performa model prediksi dengan langkah-langkah sistematis, mulai dari pengumpulan data, pra-pemrosesan, hingga evaluasi model. Tahapan-tahapan yang dilakukan divisualisasikan pada Gambar 1.



Gambar 1. Tahapan penelitian yang digunakan

A. Pengumpulan Dataset

Tahapan awal penelitian ini adalah pengumpulan dataset yang relevan untuk prediksi diabetes melitus. Dataset yang digunakan diambil dari platform Kaggle, dengan nama dataset "Diabetes" dan file bernama "diabetes.csv" (https://www.kaggle.com/datasets/johndasilva/diabetes) [17]. Dataset ini menyediakan data rekam medis dengan variabel penting seperti jumlah kehamilan (*Pregnancies*), kadar glukosa (*Glucose*), tekanan darah (*BloodPressure*), ketebalan kulit (*SkinThickness*), kadar insulin (*Insulin*), indeks massa tubuh (BMI), skor riwayat genetik diabetes (*DiabetesPedigreeFunction*), usia (*Age*), dan *outcome* (*Outcome*, di mana 1 menunjukkan diabetes dan 0 menunjukkan non-diabetes). Dataset ini mencakup 2000 entri dengan 9 kolom, dan dipilih karena kualitasnya yang terpercaya serta relevansi fitur-fitur yang tersedia dalam mendukung analisis dan prediksi diabetes.

B. Pra-pemrosesan Data

Setelah dataset terkumpul, langkah selanjutnya adalah melakukan pra-pemrosesan data. Tahap pra-pemrosesan bertujuan untuk memastikan bahwa data berada dalam kondisi optimal untuk analisis dan pemodelan [18]. Proses ini mencakup beberapa langkah utama, antara lain:

1) Penanganan data tidak lengkap: Nilai nol atau kosong pada fitur-fitur kritis seperti kadar glukosa dan tekanan darah diidentifikasi. Baris dengan data tidak lengkap dihapus, atau nilai-nilai tersebut diisi dengan nilai median dari masing-masing fitur.
2) Normalisasi data: Fitur numerik dinormalisasi untuk menyamakan skala data sehingga menghindari bias dalam model prediksi.
3) Pemisahan data: Dataset dibagi menjadi data pelatihan (80%) dan data pengujian (20%) secara stratifikasi untuk memastikan distribusi kelas target tetap konsisten di kedua subset.

C. Eksplorasi Data

Pada tahap eksplorasi data, dilakukan analisis awal untuk memahami karakteristik dataset. Analisis ini mencakup distribusi data pada setiap fitur melalui visualisasi data seperti distribusi variabel numerik, perbandingan distribusi fitur numerik, dan hubungan antar fitur utama. Analisis ini bertujuan untuk mengidentifikasi pola awal yang dapat memengaruhi prediksi. Selain itu, ketidakseimbangan kelas juga diperiksa untuk menentukan apakah diperlukan penyesuaian seperti *oversampling* atau *undersampling*. Hasil eksplorasi ini memberikan wawasan yang mendalam tentang dataset dan membantu dalam proses pemilihan fitur yang relevan.

D. **Membangun Model Prediksi Menggunakan *Stochastic Gradient Boosting* (SGB)**

Stochastic Gradient Boosting (SGB) adalah algoritma *machine learning* berbasis ensemble yang dirancang untuk meningkatkan akurasi prediksi dengan cara menggabungkan sejumlah model sederhana (*weak learners*), seperti pohon keputusan, menjadi satu model yang kuat [19]. Algoritma ini bekerja secara iteratif dengan membangun model baru yang fokus pada kesalahan model sebelumnya, sehingga prediksi secara keseluruhan menjadi lebih akurat [20]. Proses ini didasarkan pada pendekatan *boosting*, di mana setiap model baru yang ditambahkan mencoba meminimalkan fungsi loss dengan menggunakan metode *gradient descent*.

Keunggulan utama dari SGB adalah penggunaan teknik *gradient descent* untuk meminimalkan fungsi kehilangan (*loss function*) secara bertahap [21]. Pada setiap iterasi, SGB menghitung gradien dari fungsi kehilangan terhadap prediksi model sebelumnya dan membangun model baru yang berfokus pada meminimalkan gradien tersebut. Selain itu, SGB menggunakan teknik *stochastic sampling*, di mana subset data secara acak dipilih pada setiap iterasi [22]. Hal ini membantu mengurangi *overfitting* dan meningkatkan generalisasi model.

Untuk menerapkan algoritma SGB melalui beberapa proses, diantaranya:

1) Fungsi Model Awal

Fungsi model awal adalah model pertama yang digunakan sebelum proses iteratif boosting dimulai. Model ini merupakan prediksi awal yang sederhana, biasanya berupa nilai konstanta, yang meminimalkan fungsi kehilangan (L) pada seluruh dataset. Proses ini menggunakan persamaan (1).

$$F_0(x) = \arg\min_{\gamma} L(y_i, \gamma) \qquad (1)$$

di mana $L(y_i, \gamma)$ merupakan fungsi kehilangan (misalnya *Mean Squared Error* untuk regresi atau *Log Loss* untuk klasifikasi), y_i menunjukkan nilai aktual dari data target, dan γ adalah nilai konstanta yang dipilih untuk meminimalkan fungsi kehilangan.

2) Pembaruan Gradien

Setelah model awal dibuat, SGB menghitung gradien dari fungsi kehilangan untuk semua sampel. Gradien ini menunjukkan arah perbaikan atau kesalahan yang perlu dikoreksi oleh model baru. Untuk mendapatkan pembaruan gradien untuk setiap sampel i menggunakan persamaan (2).

$$g_{im} = -\frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)} \quad (2)$$

di mana g_{im} menunjukkan gradien pada iterasi mmm untuk sampel ke- i, $F_{m-1}(x_i)$ merupakan prediksi model pada iterasi sebelumnya, dan L menunjukkan Fungsi kehilangan.

3) Pembaruan Model

Model baru (hmh_mhm) dibangun untuk meminimalkan gradien yang dihitung pada langkah sebelumnya. Pada umumnya, model ini berupa pohon keputusan yang dilatih untuk memprediksi nilai gradien (gimg_{im}gim). Proses ini menggunakan persamaan (3).

$$h_m(x) = \arg\min_h \sum_{i=1}^{n} \left(g_{im} - h(x_i)\right)^2 \quad (3)$$

di mana $h_m(x)$ merujuk pada pohon keputusan atau model baru pada iterasi m, g_{im} merupakan Gradien yang menjadi target pelatihan model, dan $h(x_i)$ merupakan prediksi model baru untuk sampel i.

4) Fungsi Model yang Diperbarui

Setelah model baru (hmh_mhm) dibangun, model utama diperbarui dengan menambahkan hasil model baru yang dikalikan dengan faktor pembelajaran (η\etaη). Pada tahapan ini digunakan persamaan (4).

$$F_m(x) = F_{m-1}(x) + \eta \cdot h_m(x) \quad (4)$$

di mana $F_m(x)$ adalah model pada iterasi mmm yang diperbarui, $F_{m-1}(x)$ menunjukkn model pada iterasi sebelumnya, η merupakan *Learning rate*, yang mengontrol kontribusi dari model baru terhadap model utama, dan $h_m(x)$ merupakan model baru yang dibangun untuk memperbaiki kesalahan.

E. Evaluasi Model

Tahapan evaluasi model dilakukan untuk mengukur kinerja model prediksi yang dibangun, dalam hal ini menggunakan algoritma *Stochastic Gradient Boosting* (SGB). Evaluasi dilakukan dengan membandingkan prediksi model terhadap nilai aktual pada data pengujian yang telah dipisahkan sebelumnya [23]. Evaluasi dimulai dengan *confusion matrix* yang digunakan untuk melihat distribusi prediksi benar dan salah pada masing-masing kelas, yaitu *True Positive* (TP), *False Positive* (FP), *True Negative* (TN), dan *False Negative* (FN) [24]. Dari *confusion matrix*, diperoleh beberapa metrik untuk mengukur kinerja model, antara lain *precision*, *recall*, dan *F1-score*. Akurasi menghitung persentase prediksi yang tepat dari seluruh data, sedangkan *precision* dan *recall* masing-masing memberikan gambaran tentang kemampuan model dalam mengidentifikasi kelas positif dengan benar serta mendeteksi semua kasus positif yang ada. *F1-score*, sebagai rata-rata harmonis dari *precision* dan *recall*, digunakan untuk memberikan gambaran yang seimbang terutama jika terdapat ketidakseimbangan kelas. Evaluasi juga mencakup visualisasi ROC (*Receiver Operating Characteristic*) *Curve* untuk menganalisis *trade-off* antara tingkat deteksi positif (TPR) dan tingkat kesalahan positif (FPR), dengan *Area Under the Curve* (AUC) menunjukkan sejauh mana model mampu membedakan antara kelas positif dan negatif.

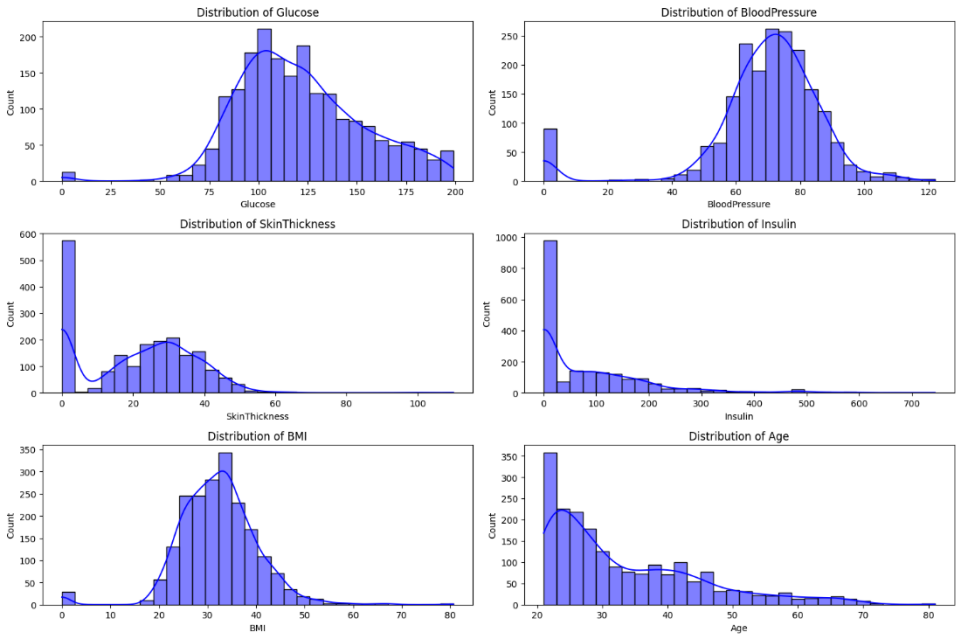
F. Analisis dan Interpretasi

Tahap analisis dan interpretasi dilakukan untuk memahami lebih mendalam kinerja model prediksi yang telah dibangun dan mengidentifikasi wawasan penting yang dihasilkan dari proses pemodelan. Pada tahap ini, hasil evaluasi model, seperti metrik akurasi, *precision*, *recall*, *F1-score*, dan ROC *Curve*, dianalisis untuk menilai kemampuan model dalam memprediksi diabetes melitus. Selain itu, dilakukan analisis terhadap distribusi kesalahan dan bagaimana model menangani data dengan karakteristik yang berbeda.

Hasil dan diskusi

Untuk membangun model prediksi model prediksi diabetes melitus menggunakan algoritma *Stochastic Gradient Boosting* (SGB) maka langkah pertama yaitu menyiapkan dataset yang digunakan untuk pelatihan dan pengujian. Dataset yang digunakan dalam penelitian ini diambil dari platform Kaggle dengan nama dataset "Diabetes" (https://www.kaggle.com/datasets/johndasilva/diabetes) [17]. Dataset ini berisi 2000 entri dan 9 kolom yang mencakup variabel-variabel penting seperti jumlah kehamilan (*Pregnancies*), kadar glukosa (*Glucose*), tekanan darah (*BloodPressure*), ketebalan kulit (*SkinThickness*), kadar insulin (*Insulin*), indeks massa tubuh (BMI), skor riwayat genetik diabetes (*DiabetesPedigreeFunction*), usia (*Age*), dan hasil akhir diagnosis diabetes (*Outcome*, di mana 1 menunjukkan diabetes dan 0 menunjukkan non-diabetes). Namun, analisis awal menunjukkan adanya nilai nol dalam kolom-kolom penting seperti *Glucose*, *BloodPressure*, *SkinThickness*, dan *Insulin*, yang dapat mengindikasikan data tidak lengkap dan memerlukan penanganan lebih lanjut. Maka Langkah selanjutnya yakni pra-pemrosesan untuk memastikan dataset siap digunakan dalam proses pemodelan. Pertama, nilai-nilai nol pada kolom penting diidentifikasi dan ditangani dengan mengganti nilai nol dengan nilai median masing-masing kolom, karena median lebih tahan terhadap outlier dibandingkan rata-rata. Kedua, dataset dibagi menjadi data pelatihan (80%) dan data pengujian (20%) secara stratifikasi untuk menjaga proporsi kelas target (*Outcome*) tetap konsisten. Selain itu, dilakukan normalisasi pada fitur numerik seperti *Glucose*, BMI, dan *Age* untuk memastikan semua fitur berada dalam skala yang sebanding, sehingga model dapat memproses data secara optimal tanpa bias terhadap fitur dengan nilai besar. Langkah ini bertujuan untuk mengatasi permasalahan kualitas data dan memastikan bahwa dataset yang digunakan memiliki konsistensi dan struktur yang baik untuk analisis lanjutan.

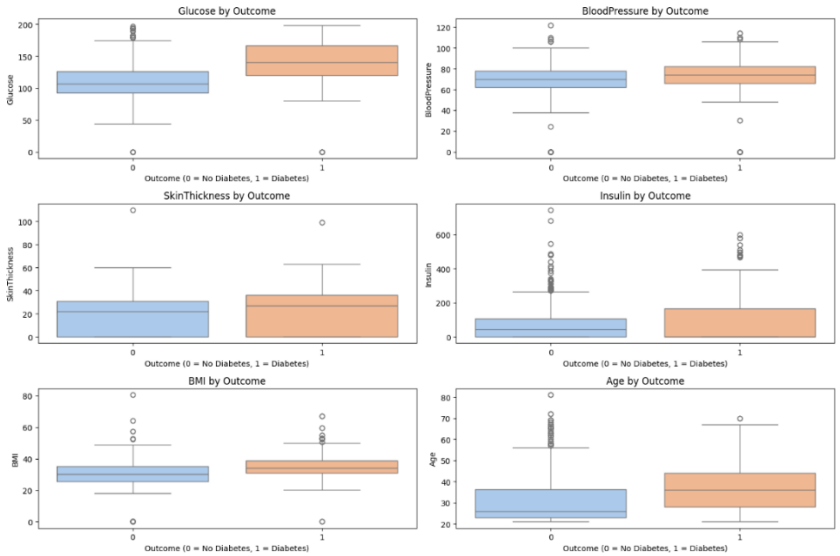
Proses selanjutnya yaitu eksplorasi data yang digunakan untuk menggali pemahaman mendalam mengenai karakteristik dataset. Proses ini melibatkan pemeriksaan distribusi data pada setiap fitur melalui visualisasi data seperti distribusi variabel numerik, perbandingan distribusi fitur numerik, dan hubungan antar fitur utama. Eksplorasi data yang pertama yaitu visualisasi distribusi variabel numerik dalam dataset yang digunakan. Visualisasi ini penting karena membantu mengidentifikasi karakteristik dari sebaran data numerik yang digunakan dan pola datanya. Visualisasi data distribusi variabel numerik dari dataset yang digunakan divisualisasikan pada Gambar 2.



Gambar 2. Visualisasi distribusi variabel numerik

Visualisasi data pada Gambar 2 menunjukkan distribusi variabel numerik dalam dataset diabetes, seperti *Glucose*, *BloodPressure*, *SkinThickness*, *Insulin*, BMI, dan *Age*. Sebagian besar variabel memiliki pola distribusi yang mendekati normal, seperti *Glucose*, *BloodPressure*, dan BMI, dengan puncak pada rentang nilai yang sesuai dengan karakteristik medis. Visualisasi data selanjutnya yaitu boxplot untuk untuk membandingkan distribusi fitur numerik utama (*Glucose*, *BloodPressure*, *SkinThickness*, *Insulin*, BMI, dan *Age*).

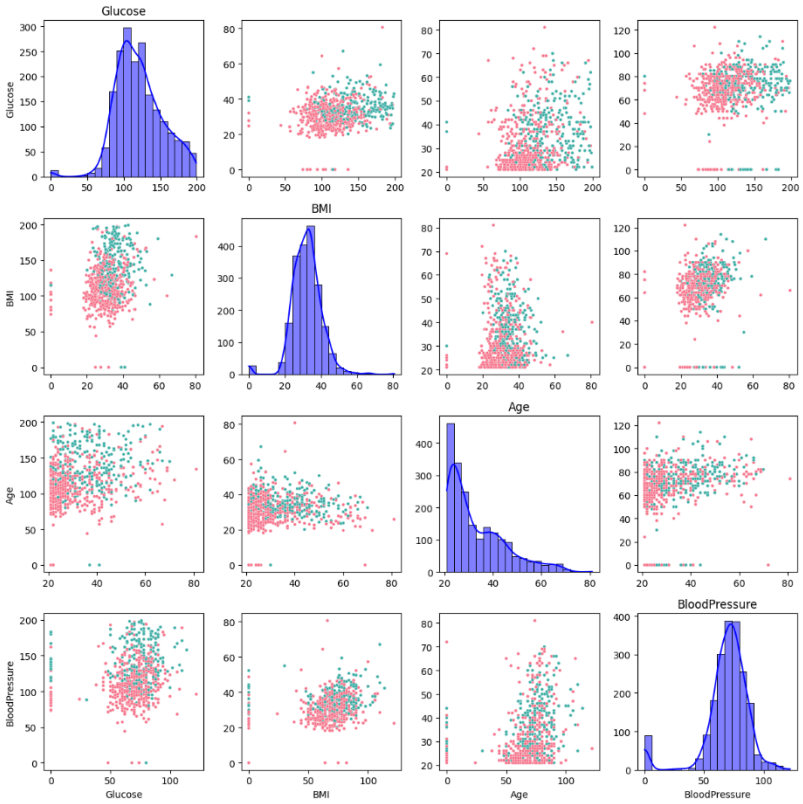
Visualisasi ini penting untuk memahami perbedaan karakteristik fitur antara dua kategori *Outcome*. Hal ini membantu dalam mengidentifikasi fitur-fitur yang berpengaruh secara signifikan terhadap prediksi diabetes. Visualisasi melalui *boxplot* untuk perbandingan distribusi fitur numerik utama pada dataset disajikan pada Gambar 3.



Gambar 3. *Boxplot* perbandingan distribusi fitur numerik

Visualisasi pada Gambar 3 menampilkan boxplot untuk setiap fitur numerik dalam dataset, yaitu *Glucose*, *BloodPressure*, *SkinThickness*, *Insulin*, BMI, dan *Age*, yang dibagi berdasarkan kategori *Outcome* (0 = Tidak Diabetes, 1 = Diabetes). Boxplot ini menggambarkan persebaran data, nilai tengah (*median*), rentang interkuartil (IQR), serta keberadaan *outlier* untuk setiap fitur. Dari visualisasi ini, terlihat bahwa individu dengan diabetes cenderung memiliki nilai yang lebih tinggi pada fitur seperti *Glucose*, BMI, dan *Age*, dibandingkan dengan individu yang tidak menderita diabetes. Sementara itu, fitur seperti *Insulin* dan *SkinThickness* memiliki persebaran data yang luas dengan banyak *outlier*.

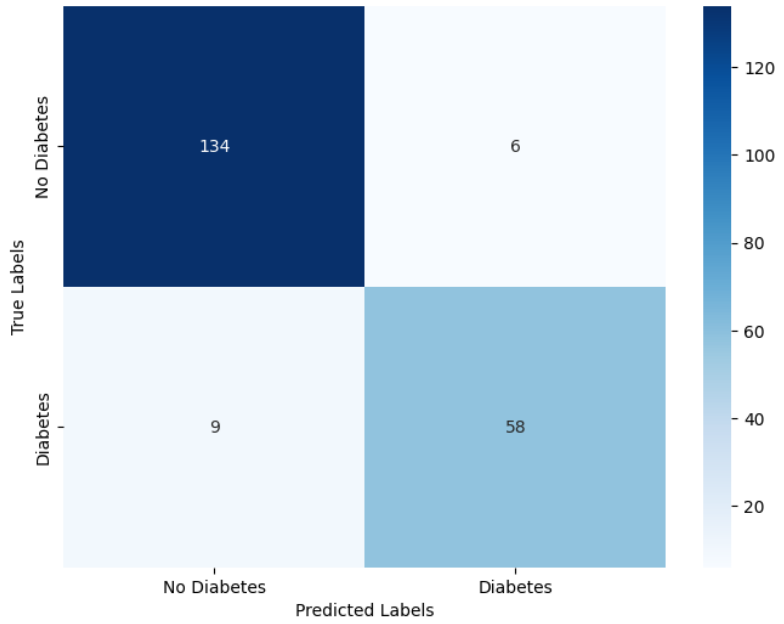
Visualisasi data yang diperlukan selanjutnya adalah visualisasi hubungan antar fitur utama dalam dataset menggunakan *pairplot*. Visualisasi ini bertujuan untuk memahami korelasi antar fitur serta bagaimana fitur-fitur tersebut memengaruhi *outcome*. Selain itu, *pairplot* juga membantu dalam mengidentifikasi pola klaster antar kelas, mendeteksi *outlier*, dan memahami hubungan signifikan antara variabel. *Pairplot* hubungan antar fitur pada dataset dipresentasikan pada Gambar 4.



Gambar 4. *Pairplot* hubungan antar fitur utama

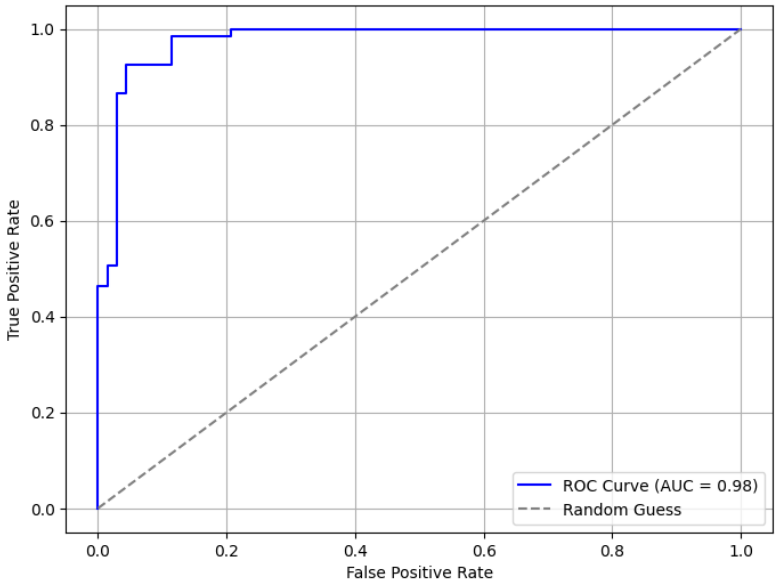
Visualisasi pada Gambar 4 adalah *pairplot* yang menggambarkan hubungan antar fitur utama dalam dataset diabetes, yaitu *Glucose*, BMI, *Age*, dan *BloodPressure*. Diagonal pada pairplot menampilkan histogram distribusi masing-masing fitur, sedangkan bagian non-diagonal menunjukkan *scatterplot* yang menggambarkan hubungan antar fitur, dengan warna yang membedakan dua kelas target: 1 (diabetes) dan 0 (non-diabetes). *Scatterplot* ini memperlihatkan pola keterkaitan antar fitur, seperti bagaimana *Glucose* memiliki kecenderungan nilai yang lebih tinggi pada kelas diabetes.

Proses selanjutnya adalah membangun model prediksi diabetes melitus menggunakan algoritma *Stochastic Gradient Boosting* (SGB). Setelah dataset dipersiapkan melalui tahap pra-pemrosesan, data didistribusikan kedalam dua bagian, yakni data latih (80%) dan data uji (20%), menggunakan metode stratifikasi untuk memastikan proporsi kelas target tetap konsisten pada kedua subset. Model SGB dibangun menggunakan *GradientBoostingClassifier* dengan parameter default, kemudian dilatih pada data pelatihan. Algoritma SGB bekerja secara iteratif dengan memperbaiki kesalahan prediksi dari iterasi sebelumnya, sehingga mampu mengenali pola kompleks dalam data secara efektif. Setelah model selesai dilatih, prediksi dilakukan terhadap data pengujian, dan model dievaluasi untuk mengukur kinerjanya. Evaluasi diawali dengan penggunaan *confusion matrix* untuk menganalisis distribusi prediksi yang tepat dan tidak tepat pada masing-masing kelas. *Confusion matrix* yang dihasilkan ditampilkan pada Gambar 5.



Gambar 4. Hasil evaluasi dengan *confusion matrix*

Evaluasi selanjutnya yaitu melalui ROC *Curve*, di mana grafik ini memperlihatkan hubungan antara *True Positive Rate* (TPR) dan *False Positive Rate* (FPR) pada berbagai *threshold* klasifikasi. Hasil ROC *Curve* dari model diabetes melitus yang dikembangkan divisualisasikan pada Gambar 5.



Gambar 5. ROC *Curve* untuk model prediksi diabetes melitus

Gambar 5 menunjukkan ROC *Curve* (*Receiver Operating Characteristic Curve*) untuk model prediksi diabetes melitus yang telah dibangun. Kurva biru menunjukkan performa model, sedangkan garis diagonal abu-abu menunjukkan baseline untuk tebakan acak (*random guess*). Pada gambar ini, nilai AUC (*Area Under the Curve*) yaitu 0.98, yang mengindikasikan bahwa model mampu dengan sangat baik dalam membedakan antara kelas positifnya (diabetes) serta kelas negatifnya (non-diabetes). Nilai AUC mendekati 1 mencerminkan performa prediksi yang hampir sempurna. Grafik ini menegaskan bahwa model mampu menghasilkan tingkat deteksi yang tinggi dengan tingkat kesalahan yang rendah, menjadikannya andal untuk mendukung deteksi dini diabetes melitus.

Proses evaluasi dilanjutkan dengan mencari metrik kinerja seperti *precision*, *recall*, *F1-score*, dan akurasi. Metrik-metrik tersebut diperoleh berdasarkan *confusion matrix* yang diperoleh sebelumnya. Hasil dari *precision*, *recall*, *F1-score*, dan akurasi dari model prediksi diabetes melitus yang dikembangkan disusun pada Tabel 1.

Tabel 1. Evaluasi kinerja model prediksi diabetes melitus

Kelas	***Precision***	***Recall***	***F1-Score***	***Accuracy***
No Diabetes	0.9371	0.9571	0.9470	0.9275
Diabetes	0.9062	0.8657	0.8855	

Hasil evaluasi model prediksi diabetes melitus pada Tabel 1 menunjukkan performa yang sangat baik untuk kedua kelas, yaitu No Diabetes (kelas 0) dan Diabetes (kelas 1). Untuk kelas No Diabetes, model mencapai *precision* sebesar 0.9371, yang berarti dari semua prediksi model untuk kelas ini, 93,71% di antaranya benar. *Recall* untuk kelas ini sebesar 0.9571, yang menunjukkan model berhasil mendeteksi 95,71% dari total kasus sebenarnya untuk kelas No Diabetes. Kombinasi *precision* dan *recall* menghasilkan *F1-score* sebesar 0.9470, yang mencerminkan keseimbangan performa model dalam prediksi kelas No Diabetes. Untuk kelas Diabetes, *precision* model adalah 0.9062, yang berarti dari semua prediksi model untuk kelas ini, 90,62% di antaranya benar. *Recall* untuk kelas ini sebesar 0.8657, yang menunjukkan model mampu mendeteksi 86,57% dari kasus diabetes yang sebenarnya. *F1-score* untuk kelas ini adalah 0.8855, yang menunjukkan keseimbangan antara *precision* dan *recall* dalam prediksi kelas Diabetes.

Secara keseluruhan, model memiliki tingkat akurasi sebesar 0.9275, yang berarti 92,75% dari seluruh prediksi model adalah benar. Hasil ini mencerminkan kemampuan model dalam membedakan kedua kelas dengan baik. Kemampuan model untuk mencapai nilai *precision*, *recall*, dan *F1-score* yang tinggi pada kedua kelas target menunjukkan bahwa SGB adalah algoritma yang efektif untuk menangani dataset ini. Keunggulan algoritma SGB terletak pada pendekatan boosting yang iteratif, di mana model secara bertahap memperbaiki kesalahan prediksi dari model sebelumnya (*weak learners*) dengan menggabungkan model-model sederhana, seperti pohon keputusan, menjadi satu model yang kuat.

Namun, terdapat sedikit penurunan pada *recall* kelas Diabetes, yang menunjukkan bahwa beberapa kasus diabetes mungkin tidak terdeteksi. Penurunan *recall* pada kelas Diabetes dapat disebabkan oleh beberapa faktor yang terkait dengan karakteristik dataset dan kinerja model. Salah satu penyebab utama adalah ketidakseimbangan kelas dalam dataset, di mana jumlah sampel pada kelas No Diabetes jauh lebih banyak dibandingkan kelas Diabetes. Ketidakseimbangan ini menyebabkan model cenderung lebih fokus pada kelas mayoritas, sehingga mengurangi sensitivitas terhadap kelas minoritas.

Kesimpulan

Penelitian ini berhasil mengembangkan model prediksi diabetes melitus menggunakan *metode Stochastic Gradient Boosting* (SGB) yang menunjukkan kinerja sangat baik. Model yang dibangun mencapai akurasi tinggi sebesar 92,75%, dengan nilai *precision*, *recall*, dan *F1-score* yang konsisten untuk kedua kelas, yaitu No Diabetes dan Diabetes. Selain itu, nilai AUC (*Area Under the Curve*) sebesar 0,98 mengindikasikan bahwa model memiliki kemampuan yang sangat baik dalam membedakan antara kelas positif (diabetes) dan kelas negatif (non-diabetes). Hasil ini menunjukkan efektivitas algoritma SGB dalam menangkap pola kompleks pada dataset dan menghasilkan prediksi yang andal. Kemampuan SGB untuk mencapai akurasi tinggi disebabkan oleh pendekatan boosting iteratif yang memanfaatkan gradien dari fungsi kehilangan untuk secara adaptif memperbaiki kesalahan prediksi pada setiap iterasi. Selain itu, mekanisme regulasi seperti *learning rate* dan *subsampling* turut berperan penting dalam mencegah *overfitting*, sehingga algoritma ini efektif untuk menangani dataset dengan pola yang rumit. Untuk penelitian selanjutnya, metode seperti penyeimbangan kelas melalui *oversampling* (misalnya SMOTE), *class weight balancing*, atau penyesuaian *threshold* klasifikasi dapat diterapkan untuk meningkatkan sensitivitas model terhadap kelas Diabetes. Dengan perbaikan ini, diharapkan model dapat lebih optimal dalam mendeteksi kasus diabetes, sehingga mendukung deteksi dini yang akurat dan andal dalam konteks medis.

Referensi

[1] T. Rawung, J. Posangi, and E. Nangoy, "Efektivitas Penggunaan Empagliflozin terhadap Nilai HbA1c pada Pasien Diabetes Melitus Tipe 2," *Med. Scope J.*, vol. 5, no. 2, pp. 232–239, 2023, doi: 10.35790/msj.v5i2.45424.

[2] H. Y. Resti and W. H. Cahyati, "Kejadian Diabetes Melitus Pada Usia Produktif Di Puskesmas Kecamatan Pasar Rebo," *Higeia J. Public Heal. Res. Dev.*, vol. 6, no. 3, pp. 350–361, 2022.

[3] K. W. D. Nugraha, T. Seviana, and F. Sibuea, *Profil Kesehatan Indonesia 2022*. Jakarta: Kementerian Kesehatan Republik Indonesia Jalan, 2023.

[4] A. Priyanto and E. D. H. Suprayetno, *Efektifitas Self Detection For Diabetic (SEDAB) Untuk Deteksi Dini Diabetes Militus*. Malang: Media Nusa Creative (MNC Publishing), 2022.

[5] A. Sah, J. Jusmawati, S. Nurhayati, M. Tonggiroh, and S. Bonay, "Sistem Informasi Manajemen Pada Puskesmas Kota Jayapura Berbasis Web," *JTIM J. Teknol. Inf. dan Multimed.*, vol. 4, no. 3, pp. 212–220, Nov. 2022.

[6] A. Sah, S. Suhardi, and S. Nurhayati, "Geographic Information System of Patient Development in Jayapura Hospital During Pandemic," *J. Teknol. Dan Open Source*, vol. 4, no. 2, pp. 149–154, 2021.

[7] R. I. Borman and M. Wati, "Penerapan Data Maining Dalam Klasifikasi Data Anggota Kopdit Sejahtera Bandarlampung Dengan Algoritma Naïve Bayes," *J. Ilm. Fak. Ilmu Komput.*, vol. 9, no. 1, pp. 25–34, 2020.

[8] R. I. Borman, R. Napianto, N. Nugroho, D. Pasha, Y. Rahmanto, and Y. E. P. Yudoutomo, "Implementation of PCA and KNN Algorithms in the Classification of Indonesian Medicinal Plants," in *International Conference on Computer Science, Information Technology and Electrical Engineering (ICOMITEE)*, IEEE, 2021, pp. 46–50.

[9] P. N. Sabrina and A. Komarudin, "Prediksi Penyakit Diabetes Dengan Metode K-Nearest Neighbor (KNN) dan Seleksi Fitur Information Gain," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 6, pp. 11320–11326, 2024.

[10] G. A. Putri, A. Trimaysella, and A. Khoiriah, "Penerapan Klasifikasi Data Mining pada Diabetes Menggunakan Metode Naive Bayes," *J. Ilmu Komput. Teknol. Terap.*, vol. 1, no. 14, pp. 1–9, 2024.

[11] A. W. Mucholladin, F. A. Bachtiar, and M. T. Furqon, "Klasifikasi Penyakit Diabetes menggunakan Metode Support Vector Machine," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 2, pp. 622–633, 2021.

[12] R. Ahsana, R. Rohmat Saedudin, and V. P. Widartha, "Perbandingan Akurasi Algoritma Adaboost dan Algoritma Lightgbm Untuk Klasifikasi Penyakit Diabetes," *e-Proceeding Eng.*, vol. 8, no. 5, pp. 9738–9748, 2021.

[13] A. F. L. Ptr, M. M. Siregar, and I. Daniel, "Analysis of Gradient Boosting, XGBoost, and CatBoost on Mobile Phone Classification," *J. Comput. Networks, Archit. High Perform. Comput.*, vol. 6, no. 2, pp. 661–670, 2024.

[14] I. K. Ananda, A. Z. Fanani, D. Setiawan, and D. F. Wicaksono, "Penerapan Random Oversampling dan Algoritma Boosting untuk Memprediksi Kualitas Buah Jeruk," *Edumatic J. Pendidik. Inform.*, vol. 8, no. 1, pp. 282–289, 2024.

[15] U. Schroeders, C. Schmidt, and T. Gnambs, "Detecting Careless Responding in Survey Data Using Stochastic Gradient Boosting," *Educ. Psychol. Meas.*, vol. 82, no. 1, pp. 29–56, Apr. 2021, doi: 10.1177/00131644211004708.

[16] A. Subasi, M. F. El-Amin, T. Darwich, and M. Dossary, "Permeability prediction of petroleum reservoirs using stochastic gradient boosting regression," *J. Ambient Intell. Humaniz. Comput.*, vol. 13, no. 7, pp. 3555–3564, 2022, doi: 10.1007/s12652-020-01986-0.

[17] J. Dasilva, "Diabetes Dataset," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/johndasilva/diabetes

[18] R. I. Borman, Y. Fernando, and Y. E. P. Yudoutomo, "Identification of Vehicle Types Using Learning Vector Quantization Algorithm with Morphological Features," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 2, pp. 339–345, 2022.

[19] M. F. El-Amin, "Turbulent Reynolds Stresses Prediction using Stochastic Gradient Boosting Regression," in *21st Learning and Technology Conference (L&T)*, 2024, pp. 139–143. doi: 10.1109/LT60077.2024.10468734.

[20] R. Setiawan, A. T. Wibowo, and M. Ridwan, "Pengembangan Sistem Rekomendasi Atlet Esports Berdasarkan Prediksi Elo Rating Menggunakan Model Stochastic Gradient Boosting," *J. Format*, vol. 11, no. 2, pp. 145–152, 2022.

[21] M. Kivrak, "Breast Cancer Risk Prediction with Stochastic Gradient Boosting," *Clin. Cancer Investig. J.*, vol. 11, no. 2, pp. 26–31, 2022, doi: 10.51847/21qrrkLo4Y.

[22] E. E. Başakın, Ö. Ekmekcioğlu, P. C. Stoy, and M. Özger, "Estimation of daily reference evapotranspiration by hybrid singular spectrum analysis-based stochastic gradient boosting," *MethodsX*, vol. 10, p. 102163, 2023, doi: https://doi.org/10.1016/j.mex.2023.102163.

[23] N. Hafiizah and R. A. Saputra, "Klasifikasi Kematangan Buah Jeruk Berdasarkan Fitur Warna Menggunakan Metode SVM," *FORMAT J. Ilm. Tek. Inform.*, vol. 13, no. 1, pp. 55–65, 2024.

[24] R. I. Borman, F. Rossi, Y. Jusman, A. A. A. Rahni, S. D. Putra, and A. Herdiansah, "Identification of Herbal Leaf Types Based on Their Image Using First Order Feature Extraction and Multiclass SVM Algorithm," in *International Conference on Electronic and Electrical Engineering and Intelligent System (ICE3IS)*, IEEE, 2021, pp. 12–17.

Ir. Andrian Sah, S.Kom., M.Kom

Penulis menyelesaikan Pendidikan S1 di prodi Teknik Informatika, STMIK AMIKOM Yogyakarta pada tahun 2012. Kemudian menyelesaikan pendidikan S2 di program Magister Informatika, Fakultas Teknologi Industri, Universitas Islam Indonesia (UII) pada tahun 2018. Saat ini penulis merupakan dosen di Prodi Sistem Informasi, Fakultas Ilmu Komputer, Universitas Yapis Papua. Selama menjadi dosen, penulis mengampu beberapa mata kuliah teori dan praktikum, diantaranya: Struktur Data, Pemograman Mobile, Jaringan Komputer, Teknik Riset Operasi, Pemograman Visual, Pemograman Berbasis Objek, Sistem Informasi Geografis, Sistem Informasi Manajemen, Keamanan Informasi dan mata kuliah lain yang terkait. Penulis memiliki bidang minat penelitian pada Sistem Informasi, Sistem Pakar, Keamanan Informasi, Multimedia Dan Digital Forensik.

Amat Damuri, S.Kom., M.Kom.

Saat ini sebagai dosen tetap kampus STMIK Al Muslim Bekasi prodi Latar belakang pendidikan penulis diantaranya yaitu menyelesaiksn Pendidikan S1 pada prodi Sistem Informasi di STMIK Kharisma Karawang pada tahun 2005 dan menyelesaikan Pendidikan S2 di STMIK Eresha Jakarta pada Prodi Teknik Informatika konsentrasi Sistem Informasi Managemen. Bidang penelitian penulis antara lain Rekayasa Perangkat Lunak, Sistem Informasi Manajemen, Sistem Pendukung Keputusan. Karya-karya penelitian dari penulis mencakup topik seperti klasifikasi data menggunakan metode algoritmik, e-learning, dan penerapan metode analisis dalam pengelolaan sumber daya berbasis teknologi. Karya penulis telah dipubilkasikan di berbagai jurnal lokal dan nasional baik pada jurnal penelitian maupun pengabdian kepada masyarakat.

Chaeroen Niesa, S.kom., M.Kom

Penulis adalah dosen di program studi Informatika, Fakultas Komputer dan Multimedia, Universitas Islam Kebangsaan Indonesia (UNIKI). Penulis menyelesaikan pendidikan Sarjana Komputer di Universitas Almuslim pada tahun 2013, kemudian saya melanjutkan studi S2 Magister Teknik Informatika di Universitas Sumatera Utara pada tahun 2016. Penulis berfokus pada bidang Sistem Informasi, Keamanan Komputer, Sistem Pendukung Keputusan, Database Dan Sistem Pakar. Beberapa publikasi saya ada di Jurnal Nasional dan sekarang lagi menunggu publikasi di Jurnal Internasional. Selain itu penulis aktif dalam membimbing mahasiswa serta kegiatan pengabdian kepada masyarakat (KKM). Penulis juga aktif di organisasi APTIKOM dan bergabung sejak 2021.

Nur Amalia Hasma, SST., M.T.

Penulis menyelesaikan pendidikan D4 vokasi di Politeknik Negeri Lhokseumawe, pada Jurusan Teknik Elektro, Program Studi Teknik Informatika. Kemudian melanjutkan S2 di Universitas Syiah Kuala, Banda Aceh, pada Program Studi Magister Teknik Elektro dengan bidang konsentrasi Teknologi Informasi. Saat ini penulis merupakan seorang dosen pada Fakultas Komputer dan Multimedia, Universitas Islam Kebangsaan Indonesia, Bireuen, Aceh. Fokus penelitian penulis adalah bidang Teknologi Informasi. Penelitian yang sudah pernah dilakukan salah satunya yaitu, “Pengenalan Gerakan Isyarat Bahasa Indonesia menggunakan algoritma SURF dan K-Nearest Neighbor” dan “Implementasi Machine Learning dalam menganalisis dan mendeteksi Berita Palsu pada portal berita Bahasa Inggris”.