**TESIS**

# OPTIMASI PREDIKSI DIABETES PADA DATASET PIMA INDIANS MELALUI PENGGABUNGAN ALGORITMA K-NEAREST NEIGHBORS (KNN) DAN LIGHTGBM DENGAN PENDEKATAN EXPLORATORY DATA ANALYSIS (EDA)


Disusun oleh:

**Nama** : **Arvi Pramudyantoro**
**NIM** : **22.51.1183**
**Konsentrasi** : **Business Intelligence**



**PROGRAM STUDI S2 INFORMATIKA**
**PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA**
**YOGYAKARTA**
**2024**

**TESIS**

# OPTIMASI PREDIKSI DIABETES PADA DATASET PIMA INDIANS MELALUI PENGGABUNGAN ALGORITMA K-NEAREST NEIGHBORS (KNN) DAN LIGHTGBM DENGAN PENDEKATAN EXPLORATORY DATA ANALYSIS (EDA)

# OPTIMIZATION OF DIABETES PREDICTIONS IN THE PIMA INDIANS DATASET BY COMBINING THE K-NEAREST NEIGHBORS (KNN) AND LIGHTGBM ALGORITHMS WITH THE EXPLORAROTY DATA ANALYSIS (EDA) APPROACH

Diajukan untuk memenuhi salah satu syarat memperoleh derajat Magister

Disusun oleh:


**Nama** : **Arvi Pramudyantoro**
**NIM** : **22.51.1183**
**Konsentrasi** : **Business Intelligence**


**PROGRAM STUDI S2 INFORMATIKA**
**PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA**
**YOGYAKARTA**
**2024**

## HALAMAN PENGESAHAN

**OPTIMASI PREDIKSI DIABETES PADA DATASET PIMA INDIANS MELALUI PENGGABUNGAN ALGORITMA K-NEAREST NEIGHBORS (KNN) DAN LIGHTGBM DENGAN PENDEKATAN EXPLORATORY DATA ANALYSIS (EDA)**

**OPTIMIZATION OF DIABETES PREDICTIONS IN THE PIMA INDIANS DATASET BY COMBINING THE K-NEAREST NEIGHBORS (KNN) AND LIGHTGBM ALGORITHMS WITH THE EXPLORATORY DATA ANALYSIS (EDA) APPROACH**

Dipersiapkan dan Disusun oleh

**Arvi Pramudyantoro**

**22.51.1183**

Telah Diujikan dan Dipertahankan dalam Sidang Ujian Tesis
Program Studi S2 Informatika
Program Pascasarjana Universitas AMIKOM Yogyakarta
pada hari Selasa, 09 Juli 2024

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer

Yogyakarta, 09 Juli 2024
**Rektor**

**Prof. Dr. M. Suyanto, M.M.**
**NIK. 190302001**

## HALAMAN PERSETUJUAN

**OPTIMASI PREDIKSI DIABETES PADA DATASET PIMA INDIANS MELALUI PENGGABUNGAN ALGORITMA K-NEAREST NEIGHBORS (KNN) DAN LIGHTGBM DENGAN PENDEKATAN EXPLORATORY DATA ANALYSIS (EDA)**

**OPTIMIZATION OF DIABETES PREDICTIONS IN THE PIMA INDIANS DATASET BY COMBINING THE K-NEAREST NEIGHBORS (KNN) AND LIGHTGBM ALGORITHMS WITH THE EXPLORATORY DATA ANALYSIS (EDA) APPROACH**

Dipersiapkan dan Disusun oleh

**Arvi Pramudyantoro**

**22.51.1183**

Telah Diujikan dan Dipertahankan dalam Sidang Ujian Tesis
Program Studi S2 Informatika
Program Pascasarjana Universitas AMIKOM Yogyakarta
pada hari Selasa, 09 Juli 2024

**Pembimbing Utama**

**Prof. Dr. Ema Utami, S.Si., M.Kom.**
**NIK. 190302037**

**Pembimbing Pendamping**

**Dhani Ariatmanto, M.Kom., Ph.D.**
**NIK. 190302197**

**Anggota Tim Penguji**

**Dr. Kumara Ari Yuana, S.T., M.T.**
**NIK. 190302575**

**M. Hanafi, S.Kom., M.Eng., Ph.D.**
**NIK. 190302024**

**Prof. Dr. Ema Utami, S.Si., M.Kom.**
**NIK. 190302037**

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer

Yogyakarta, 09 Juli 2024
**Direktur Program Pascasarjana**

**Prof. Dr. Kusrini, M.Kom.**
**NIK. 190302106**

## HALAMAN PERNYATAAN KEASLIAN TESIS

Yang bertandatangan di bawah ini,

**Nama mahasiswa : Arvi Pramudyantoro**
**NIM : 22.51.1183**
**Konsentrasi : Business Intelligence**

Menyatakan bahwa Tesis dengan judul berikut:
**Optimasi Prediksi Diabetes Pada Dataset Pima Indians Melalui Penggabungan Algoritma K-Nearest Neighbors (KNN) Dan LightGBM Dengan Pendekatan Exploratory Data Analysis (EDA)**

Dosen Pembimbing Utama : Prof. Dr. Ema Utami, S.Si., M.Kom.
Dosen Pembimbing Pendamping : Dhani Ariatmanto, M.Kom., Ph.D.

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Tim Dosen Pembimbing
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi

Yogyakarta, 9 Juli 2024
Yang Menyatakan,


Arvi Pramudyantoro

## HALAMAN PERSEMBAHAN

Puji dan syukur saya panjatkan kehadirat Allah SWT atas nikmat dan rahmat-Nya yang tak terhingga, yang telah memungkinkan saya untuk menyelesaikan tesis ini dengan baik. Saya juga ingin menyampaikan terima kasih yang tulus kepada semua pihak yang telah memberikan bantuan dan dukungan, baik secara langsung maupun tidak langsung, selama proses penyelesaian tesis ini. Tesis ini saya persembahkan kepada:

1. Kedua orang tua tercinta, Bapak Sugiya dan Ibu Suyati yang telah memberikan dukungan, doa, dan pengorbanan tanpa henti. Keberhasilan ini adalah hasil dari kerja keras, doa, dan kasih sayang dari Bapak dan Ibu.
2. Kakak-kakak saya, Hendri Noviyarto, S.IP, M.AP., Dewi Aprilia Puspita Ningrum, S.Kom., dan drh. Diyah Septiriyanti, atas segala bentuk dukungan, semangat dan nasihat yang kalian berikan. Kalian adalah inspirasi dan motivasi saya.
3. Ibu Prof. Dr. Ema Utami, S.Si., M.Kom selaku Pembimbing Utama dan Bapak Dhani Ariatmanto, S.Kom., M.Kom., Ph.D. selaku Pembimbing Pendamping, terima kasih atas bimbingan, ilmu, dan arahan yang di berikan selama penyusunan tesis ini telah menjadi mentor yang luar biasa.
4. Bapak Alva Hendi Muhammad, S.T., M.Eng., dan Bapak Kusnawi, S.Kom., M.Eng., selaku Penguji Seminar Proposal Tesis.

5. Bapak Dr. Kumara Ari Yuana, S.T., M.T. dan Bapak Emha Taufiq Luthfi, S.T., M.Kom., selaku Penguji Seminar Hasil Proposal Tesis
6. Bapak M. Hanafi, S.Kom., M.Eng., Ph.D. dan Bapak Dr. Kumara Ari Yuana, S.T., M.T., selaku Penguji Ujian Tesis.
7. Seluruh Bapak dan Ibu Dosen Pascasarjana Universitas Amikom Yogyakarta. Terima kasih atas ilmu, bimbingan, dan pengalaman yang di berikan selama masa studi. Setiap pelajaran yang di ajarkan sangat berarti bagi perkembangan akademik dan pribadi saya.
8. Mardhotillah, S.Pd., terimakasih telah selalu memberikan dukungan dalam kelancaran tesis ini, masukan dan support luar biasa untuk mencari refrerensi tesis ini serta kesabaran dan ketulusan yang tak terlupakan.
9. Teman-teman seangkatan 28 Magister Informatika, yang selalu membantu dalam suka dan duka.
10. Seluruh orang yang telah mendoakan dan membantu dalam proses penyelesaian tesis ini, baik secara langsung maupun tidak langsung. Terima kasih atas doa, dukungan, dan bantuan yang diberikan. Setiap bantuan kalian sangat berarti bagi saya.

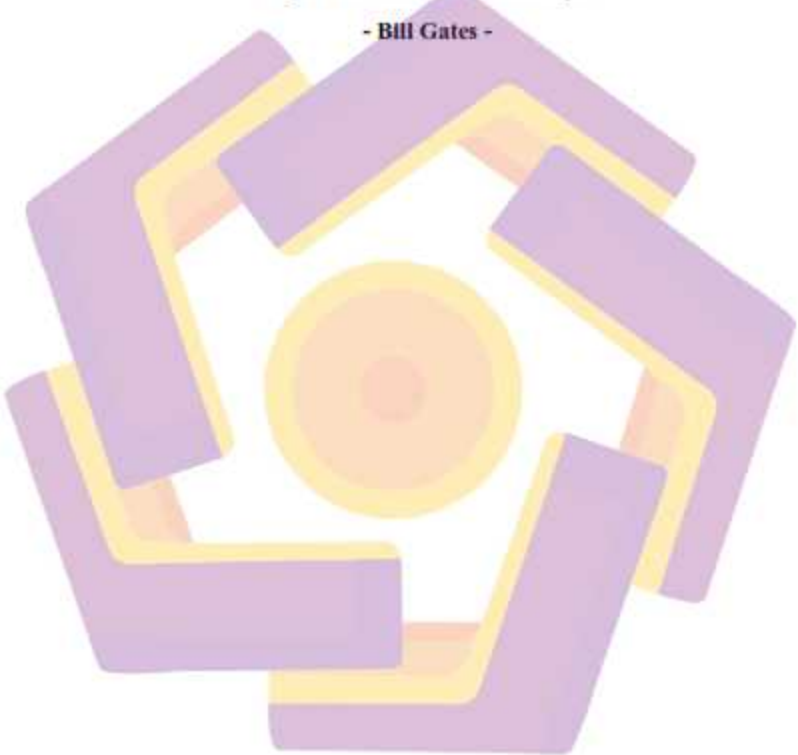
Semoga tesis ini dapat memberikan manfaat dan kontribusi bagi perkembangan ilmu pengetahuan dan teknologi. Semoga Allah SWT membalas semua kebaikan kalian dengan pahala yang berlipat ganda.

## HALAMAN MOTTO

*"Every one of us has the same amount of time each day. What sets us apart is how we use it."*

Kita semua memiliki waktu yang sama setiap harinya. Yang membedakan adalah bagaimana kita memanfaatkannya.

**- Bill Gates -**



## KATA PENGANTAR

Puji syukur penulis panjatkan kehadirat Allah yang telah melimpahkan rahmat serta hidayah-Nya, tidak lupa shalawat serta salam selalu penulis panjatkan kepada junjungan kita Nabi Muhammad SAW, yang telah menuntun umatnya sehingga penulis dapat menyelesaikan tesis dengan baik.

Tesis ini disusun sebagai salah satu syarat utama untuk menyelesaikan program magister S2 Informatika pada Pascasarjana Universitas AMIKOM Yogyakarta.

Pengerjaan tesis ini tidak lepas dari bantuan berbagai pihak. Oleh Karena itu, penulis ingin menyampaikan rasa hormat dan terima kasih kepada:

1. Bapak Prof. Dr. M. Suyanto, MM. selaku Rektor Universitas AMIKOM Yogyakarta.
2. Ibu Prof. Dr. Kusrini, M.Kom. selaku Direktur Program Pascasarjana Universitas AMIKOM Yogyakarta dan seluruh dosen pascasarjana
3. Ibu Prof. Dr. Ema Utami, S.Si., M.Kom. selaku dosen pembimbing I dan Bapak Dhani Ariatmanto, M.Kom., Ph.D. selaku dosen pembimbing II yang telah membantu penulis dengan saran dan waktunya.

Penulis menyadari sepenuhnya penelitian ini masih terdapat kekurangan,maka dari itu kritik dan saran serta masukan dari berbagai pihak akan penulis terima dengan lapang dada sebagai perbaikan karya – karya selanjutnya. Semoga tesis yang sederhana ini dapat bermanfaat bagi pihak yang membutuhkan.

Yogyakarta, 09 Juli 2024

Penulis

## DAFTAR ISI

## DAFTAR TABEL

## DAFTAR GAMBAR

## INTISARI

Penelitian dengan judul "Optimasi Prediksi Diabetes pada Dataset Pima Indians melalui Penggabungan Algoritma K-Nearest Neighbors (KNN) dan LightGBM dengan Pendekatan Exploratory Data Analysis (EDA)." Bertujuan untuk meningkatkan akurasi prediksi diabetes dengan memanfaatkan EDA untuk mengidentifikasi karakteristik penting dalam dataset Pima Indians dan mengoptimalkan model prediksi menggunakan algoritma KNN dan LightGBM. Langkah-langkah penelitian meliputi pengumpulan dan pembersihan data, deteksi dan penanganan nilai yang hilang, serta feature engineering untuk menciptakan fitur-fitur baru yang lebih relevan dalam mendeteksi diabetes. Beberapa fitur baru yang dihasilkan antara lain Risiko Rendah Diabetes, Berat Badan Ideal, Kehamilan Berdasarkan Umur, dan Kadar Glukosa dan Tekanan Darah Normal. Teknik visualisasi data seperti histogram, scatter plot, dan heatmap digunakan untuk memahami struktur dan pola data.

Penggunaan EDA berhasil mengidentifikasi hubungan kuat antara kadar glukosa darah dan kemungkinan diabetes serta membantu dalam pembersihan dan transformasi data. Kombinasi algoritma KNN dan LightGBM yang dioptimasi menunjukkan peningkatan akurasi yang signifikan. Sebelum optimasi, akurasi KNN dan LightGBM masing-masing adalah 86.34% dan 88.20%. Setelah optimasi, akurasi meningkat menjadi 91.30% untuk kedua model. Teknik stacking juga menunjukkan hasil prediksi yang lebih stabil dan akurat dengan akurasi sebesar 95.65%.

Penggabungan algoritma KNN dan LightGBM yang dioptimasi menggunakan teknik EDA dan feature engineering dapat meningkatkan akurasi prediksi diabetes secara signifikan. Teknik visualisasi data dan analisis statistik deskriptif membantu dalam memahami struktur dan pola data, yang berkontribusi pada peningkatan performa model prediksi.



Kata kunci: Prediksi Diabetes, Pima Indians, KNN, LightGBM, Exploratory Data Analysis (EDA)

## *ABSTRACT*

*This study is titled "Optimization of Diabetes Predictions in the Pima Indians Dataset by Combining the K-Nearest Neighbors (KNN) and Lightgbm algorithms with the Exploratory Data Analysis (EDA) Approach." The study aims to enhance diabetes prediction accuracy by utilizing EDA to identify key characteristics in the Pima Indians dataset and optimizing the prediction model using KNN and LightGBM algorithms. The research steps include data collection and cleaning, detecting and handling missing values, and feature engineering to create new, more relevant features for detecting diabetes. Some new features generated include Low Diabetes Risk, Ideal Body Weight, Pregnancy Based on Age, and Normal Glucose and Blood Pressure Levels. Data visualization techniques such as histograms, scatter plots, and heatmaps were used to understand data xtructure and patterns.*

*The use of EDA succeeded in identifying a strong relationship between blood glucose levels and the possibility of diabetes and helped in data cleaning and transformation. The combination of optimized KNN and LightGBM algorithms showed significant accuracy improvement. Before optimization, the accuracy of KNN and LightGBM was 86.34% and 88.20%, respectively. After optimization, the accuracy increased to 91.30% for both models. The stacking technique also showed more stable and accurate prediction results with an accuracy of 95.65%.*

*Combination of optimized KNN and LightGBM algorithms using EDA and feature engineering can significantly improve diabetes prediction accuracy. Data visualization techniques and descriptive statistical analysis help understand data structure and patterns, contributing to improved model performance.*



*Keyword: Diabetes Prediction, Pima Indians, KNN, LightGBM, Exploratory Data Analysis (EDA), feature engineering*

# BAB I

# PENDAHULUAN

## 1.1. Latar Belakang Masalah

Penyakit diabetes atau biasa dikenal diabetes mellitus, ditandai dengan ketidakmampuan tubuh dalam menyerap atau memetabolisme gula di dalam darah. Kemudian, dalam keadaan normal yang ada pada diri seseorang, insulin dapat diproduksi secara otomatis dan bekerja untuk mempertahankan kadar gula darah di atas ambang batas normal (Perdana et al., 2023). Kadar gula darah akan meningkat dan diabetes akan berkembang jika tubuh tidak dapat menggunakan insulin yang dihasilkannya atau jika sel tidak meresponnya. Rasa buang air kecil yang terlalu sering, rasa haus yang tak terkendali, dan rasa lapar yang berlebih adalah gejala yang khas dari penyakit diabetes.

Menurut statistik terbaru *International Diabetes Federation* (IDF), Indonesia menempati urutan ketujuh dari 10 negara teratas dalam hal jumlah penderita diabetes. Sebanyak 10,8 juta penduduk Indonesia diperkirakan menderita diabetes pada tahun 2020, atau 6,2% dari jumlah penduduk negara. Jenis kelamin laki-laki tampaknya memiliki prevalensi lebih tinggi dari pasien diabetes dibandingkan jenis kelamin perempuan (Jais et al., 2021).

Diagnosis dan klasifikasi penyakit diabetes merupakan langkah awal yang penting dalam manajemen penyakit ini. Pemeriksaan diabetes dilakukan melalui pemeriksaan klinis yang melibatkan tes laboratorium seperti tes gula darah atau tes glukosa (Kaur et al., 2020). Namun, dengan kemajuan teknologi dan

peningkatan ketersediaan data medis, pendekatan berbasis Machine Learning telah menarik minat sebagai metode alternatif dalam diagnosis dan klasifikasi penyakit diabetes.

*Exploratory Data Analysis* (EDA) merupakan tahap awal dalam pemrosesan data yang melibatkan eksplorasi dan pemahaman yang mendalam terhadap dataset yang digunakan (Mollick et al, 2022). EDA memungkinkan identifikasi pola, korelasi, dan wawasan penting lainnya yang dapat membantu dalam pengembangan model klasifikasi yang lebih baik dan akurat (Hassan et al., 2022).

Algoritma K-Nearest Neighbors (KNN) adalah algoritma pembelajaran mesin yang populer dan sederhana. KNN bekerja berdasarkan konsep bahwa objek yang serupa cenderung berada dalam tetangga terdekat satu sama lain dalam ruang fitur (Haque et al., 2022). Dalam konteks klasifikasi penyakit diabetes, KNN dapat memanfaatkan informasi dari pasien yang memiliki gejala atau faktor risiko serupa untuk memprediksi apakah seorang individu menderita diabetes atau tidak (Wang et al., 2021).

Algoritma Light Gradient Boosting Machine (LightGBM) adalah sebuah kerangka yang terdistribusi dan berkinerja tinggi berdasarkan sebuah algoritma pohon keputusan dan dapat digunakan untuk penggunaan klasifikasi, regresi dan penggunaan dalam machine learning lainnya (Rufo et al., 2021).

Dalam beberapa tahun terakhir, kemajuan dalam bidang analisis data dan algoritma pembelajaran mesin telah menunjukkan potensi besar dalam penerapan mereka pada bidang kesehatan, termasuk dalam prediksi dan klasifikasi penyakit

diabetes. Algoritma K-NN adalah salah satu metode pembelajaran mesin yang sederhana dan populer untuk klasifikasi. Sementara itu, LightGBM adalah algoritma yang semakin digunakan karena kecepatan dan efisiensinya dalam menangani data besar (Rajkomar et al., 2019). Meskipun K-NN dan LightGBM menunjukkan potensi dalam klasifikasi, penggunaan mereka dalam klasifikasi penyakit diabetes pada dataset Pima Indian belum banyak dieksplorasi dan dibandingkan secara komprehensif (Bhargava et al., 2019).

Oleh karena itu, penelitian ini menjadi relevan dan penting untuk dilakukan guna menggabungkan kinerja algoritma K-NN dan LightGBM dalam klasifikasi penyakit diabetes pada dataset Pima Indian, serta dalam menganalisis data yang digunakan menggunakan analisis EDA. Hasil penelitian ini diharapkan dapat memberikan panduan bagi peneliti dan praktisi dalam memilih pendekatan yang paling tepat dalam deteksi dini dan pengelolaan penyakit diabetes pada populasi Pima Indian (Satihish et al., 2021). Selain itu, identifikasi fitur-fitur klinis yang paling berpengaruh dapat membantu dalam pengembangan model prediksi yang lebih efektif dan akurat, dengan menggunakan evaluasi menggunakan confusion matrix dengan mencari nilai akurasi, presisi, recall dan f1 score (Miriyala et al., 2020). Serta meningkatkan pemahaman tentang faktor-faktor yang berkontribusi pada keberadaan penyakit diabetes pada populasi Pima Indian dan populasi lainnya.

## 1.2. Rumusan Masalah

Berdasarkan latar belakang yang telah disampaikan, maka perlu rumusan masalah dalam penelitian ini yaitu :

1. Bagaimana *Exploratory Data Analysis* (EDA) dapat digunakan untuk memahami dan mempersiapkan dataset Pima Indians sebelum diklasifikasi?
2. Bagaimana cara meningkatkan akurasi model prediksi diabetes pada dataset Pima Indians menggunakan algoritma KNN dan LightGBM?
3. Bagaimana perbandingan efektivitas model sebelum dan setelah penggunaan teknik *ensemble learning* dalam optimasi model?

## 1.3. Batasan Masalah

Batasan masalah dalam Penelitian ini adalah:

1. Dataset yang digunakan berasal dari Kaggle dengan nama Pima Indians Diabetes Database dari UCI Machine Learning (2016).
2. Variabel yang akan digunakan dalam analisis adalah atribut medis dan demografis yang tersedia dalam dataset Pima Indians, seperti tingkat glukosa darah, indeks massa tubuh (BMI), tekanan darah, dan riwayat keluarga diabetes.
3. Analisis EDA akan mencakup eksplorasi karakteristik dataset, identifikasi hubungan antar atribut, dan penentuan atribut yang relevan untuk klasifikasi diabetes. Tools yang digunakan dalam EDA meliputi histogram, scatter plot, box plot, heatmap korelasi, dan pairplot.

4. Algoritma K-Nearest Neighbors (KNN) dan LightGBM akan digunakan untuk melakukan klasifikasi penyakit diabetes menggunakan fitur-fitur yang ada pada dataset.
5. Penggunaan *GridSearchCV* akan digunakan untuk mencari parameter yang optimal KNN dan *RandomSearchCV* digunakan untuk LightGBM dalam klasifikasi diabetes.
6. Evaluasi dan valdasi menggunakan Confusion Matrix dengan mencari nilai akurasi, presisi, recall, dan F1-Score.

**1.4. Tujuan Penelitian**

Tujuan dari Penelitian ini adalah:

1. Memanfaatkan *Exploratory Data Analysis* (EDA) untuk mengidentifikasi karakteristik penting dari dataset Pima Indians yang dapat mempengaruhi hasil prediksi.
2. Mengoptimalkan model prediksi diabetes dengan mengintegrasikan algoritma KNN dan LightGBM menggunakan teknik penggabbungan model.
3. Mengevaluasi dan membandingkan kinerja model sebelum dan sesudah optimasi untuk menentukan efektivitas yang dilakukan.

## 1.5. Manfaat Penelitian

a. Bagi Masyarakat

1. Dapat membantu dalam mendeteksi dini dan mengelola penyakit diabetes dengan lebih efektif.
2. Penelitian ini dapat membantu dalam perawatan kesehatan untuk seseorang yang terkena penyakit diabetes. Hal ini dapat memungkinkan penyedia layanan kesehatan dapat mengidentifikasi risiko serta kebutuhan pada pasien dengan lebih spesifik, dan memberikan perawatan yang tepat dan terukur.
3. Peningkatan kesadaran masyarakat tentang penyakit diabetes serta faktor-faktor risiko yang terkait berdasarkan dataset yang digunakan.

b. Bagi Ilmu Pengetahuan

1. Memberikan kontribusi terhadap bidang klasifikasi penyakit diabetes dengan penggunaan EDA, memprediksi dengan algoritma KNN dan menggabungkan dengan algoritma LightGBM serta melakukan evaluasi model.
2. Mengetahui faktor-faktor risiko serta hubungan antar variabel yang berkaitan dengan dataset yang digunakan agar dapat membantu dalam pengembangan model klasifikasi yang lebih akurat.

# BAB II

# TINJAUAN PUSTAKA

## 2.1. Tinjauan Pustaka

Setiap pasien memiliki berbagai masalah yang unik dan memiliki beberapa faktor risiko yang terkait dengan penyakit diabetes. Dalam beberapa tahun terakhir teknologi Machine Learning untuk memprediksi beberapa macam penyakit semakin populer untuk digunakan. Bahkan Algoritma dan perangkat lunak telah dibuat oleh para peneliti dalam jumlah yang besar. Maka dari itu, aspek pada penelitian yang terkait telah menunjukan potensi yang sangat besar dalam bidang perawatan medis. Pada bagian ini akan disajikan beberapa literatur review atau tinjauan pustaka yang terkait langsung dengan metodologi yang akan diusulkan.

PIMA Indians Diabetes Dataset (PIDD) adalah dataset yang berasal dari sebuah situs repositori "Kaggle" menjadi data publik yang populer digunakan oleh beberapa peneliti. Dataset PIDD digunakan pada dua algoritma pada ML, seperti metode K-means yang dimodifikasi dan algoritma logistic regression, dan memprogram penelitian pada alat Waikato (WEKA) dengan hasil akhir akurasi yang didapatkan sebesar 89,42% (Wu et al., 2018). Kemudian penelitian yang dilakukan oleh (Nai-arun, N.m Moungmai, R. 2018) dimana mereka menggunakan empat model Machine Learning yang berbeda seperti logistic regression, artificial neural network, random forest dan naive bayes dengan menggunakan kombinasi bagging dan ensemble. Pada metodologi ini, penulis

menggunakan data kasus penyakit diabetes dari RSUD Sawan Pracharak terdiri dari 30.122 kasus , dengan hasil akurasi tertinggi menggunakan algoritma random forest sebesar 85,55%.

Pada penelitian yang dilakukan oleh (Miriyala N. Pratyusha, Kottapalli R. Lakshmi, 2022) dimana mereka menggunakan analisis data eksplorasi untuk mendalami dataset yang digunakan dan menggabungkan 5 metode algoritma yang berbeda. Pemodelan dilakukan dengan validasi silang dengan 5 kali lipat dimana akurasi dari setiap bagian dihitung. Hasil dari eksperimen yang didapatkan tertinggi didapatkan dengan algoritma XGBoost dengan akurasi tertinggi 88,2% disusul dengan Decision tree dengan hasil akurasi sebesar 85,3%. Penelitian ini dikarenakan sedikitnya data yang diuji sehingga hasilnya tidak seimbang, maka dibutuhkan solusi untuk pengambilan data sampel berikutnya agar data tersebut dapat seimbang.

Dalam penelitian yang dilakukan (Wei, S. Zhao, X, C., 2018) beberapa pendekatan pada pra-pemrosesan data, seperti Principal Component Analysis (PCA) dan Linear Discriminant Analysis (LDA), digunakan bersama dengan algoritma komparatif umum, seperti jaringan syaraf dalam, mesin vektor pendukung, dan lain-lain. Untuk mengamati metrik berdasarkan validasi silang 10 kali lipat, penulis menggunakan data PIDD. K-Nearest Neighbors (KNN), yang memiliki akurasi terbaik 77,86% selama percobaan, mengalahkan teknik lain, menurut penulis. Tujuan prediksi penyakit diabetes yang menggunakan komputasi lunak adalah untuk menawarkan wawasan tentang sejumlah besar data yang digunakan menggunakan algoritma *Machine Learning*.

## 2.2. Keaslian Penelitian

Tabel 2. 1. Matriks Literature Review

| No | Judul | Peneliti, Media Publikasi, dan Tahun | Tujuan Penelitian | Kesimpulan | Saran atau Kelemahan | Perbandingan |
|---|---|---|---|---|---|---|
| 1 | An ensemble learning approach for diabetes prediction using boosting techniques | Beschi Raja et al., Frontiers, 2019 | Mengembangkan model prediksi diabetes menggunakan algoritma boosting | Gradient boosting memberikan akurasi tertinggi (89.7%) di antara klasifikasi yang diuji. | Model ini bisa digunakan sebagai alat prognosis di industri kesehatan untuk prediksi penyakit dini. | Kinerja model ini lebih baik dibandingkan model yang menggunakan satu algoritma karena menggunakan ensemble learning. |
| 2 | Diagnosis of Diabetes Mellitus Using Gradient Boosting Machine (LightGBM) | Ethiopian Artificial Intelligence Center, Diagnostics 2021 | Meningkatkan deteksi dini diabetes melalui LightGBM | LightGBM memberikan akurasi dan sensitivitas yang tinggi | Keterbatasan dalam kapasitas komputasi di wilayah tertentu | Mengungguli KNN dan algoritma lain dalam akurasi |

Tabel 2. 1. (Lanjutan)

| | | | | | | |
|---|---|---|---|---|---|---|
| 3 | Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus | Md. Faisal, Asaduzaaman & Iqbal H. Sarker, 2019) Q1 | Penelitian ini bertujuan untuk mengeksplorasi prediksi awal penyakit diabetes dengan berbagai faktor risiko yang berhubungan dengan penyakit diabetes mellitus. Dengan menggunakan dataset yang terdiri dari 16 atribut diabetes dari 200 pasien. | Kinerja Decision Tree C.45 secara signifikan lebih unggul daripada daripada teknik machine learning lainnya untuk klasifikasi data diabetes. Hasil eksperimen dapat membantu perawatan kesehatan untuk melakukan pencegahan dini dan membuat keputusan klinis yang lebih baik untuk mengendalikan diabetes | Atribut yang kurang menjadikan hasil akurasi yang belum optimal. | Pada penelitian ini memiliki perbedaan pada dataset yang akan digunakan, pada penelitian ini akan menggunakan dataset pribadi sebagai penelitian sendiri dan atribut yang banyak, namun jumlah dari dataset yang digunakan dari hasil pengumpulan data yang masih sedikit dapat mempengaruhi hasil dari akurasi yang didapatkan. |

Tabel 2. 1. (Lanjutan)

| | | | | | | |
|---|---|---|---|---|---|---|
| 4 | Role of K-nearest neighbour in detection of Diabetes Mellitus | (Roshi Saxenaa, Dr. Sanjay Kumar Sharmab Manali Guptac, 2021) Q4 | Menentukan tingkatan akurasi yang didapatkan untuk mendeteksi penyakit diabetes melitus menggunakan algoritma K-Nearest Neighbors yang digunakan untuk mencari nilai K tetangga terdekat, agar mendapatkan sebuah hasil akurasi yang terbaik dengan penelitian sebelumnya | Diabetes Mellitus merupakan penyakit yang dapat terjadi pada siapa saja yang memiliki berat badan berlebih, gaya hidup tidak sehat, beban kerja yang terlalu berat, dan stres. Setelah menjalankan metodologi yang diusulkan di Weka, dari 70,1% terjadi peningkatan 8,48% sehingga menjadi 78,58%. | Penelitian ini menghasilkan tingkat akurasi yang tinggi, tetapi hanya memiliki data pengujian yang sangat kecil yaitu 100 data. | Data uji yang dibagi dalam split data sedikit yaitu hanya berjumlah 100 data. |

Tabel 2. 1. (Lanjutan)

| | | | | | | |
|---|---|---|---|---|---|---|
| 5 | A stacked ensemble machine learning approach for the prediction of diabetes | Singh et al., IEEE, 2021 | Memprediksi status diabetes menggunakan pendekatan ensemble | Model eDiaPredict mencapai akurasi 95%, presisi 88%, dan sensitivitas 90.32%. | Kinerja dapat bervariasi tergantung pada dataset dan parameter yang digunakan. | Lebih fokus pada kombinasi XGBoost dan random forest, berbeda dari kombinasi KNN dan LightGBM. |
| 6 | Diabetes Prediction Using Knn ML | Hasan et al., IEEE, 2023 | Memprediksi diabetes menggunakan KNN dan algoritma lain | AUC tinggi 0.950, namun akurasi, presisi, dan sensitivitas lebih rendah dibanding model lain. | Perlu optimasi lebih lanjut untuk meningkatkan akurasi, presisi, dan sensitivitas. | Menggunakan KNN dalam ensemble yang mencakup algoritma lain, serupa dengan pendekatan yang Anda gunakan. |

## 2.3. Landasan Teori

### 2.3.1. Klasifikasi

Klasifikasi adalah proses mengelompokkan data ke dalam kategori-kategori atau kelas-kelas yang telah ditentukan sebelumnya (Kusumawardani & Karmingsih, 2021). Dalam pembelajaran mesin, klasifikasi sering digunakan untuk membuat model yang dapat mengklasifikasikan data baru berdasarkan data yang telah diterima sebelumnya (Erdiansyah et al., 2022).

Ada beberapa jenis klasifikasi yang dikenal, di antaranya

a. Klasifikasi biner: hanya ada 2 kelas, contoh : spam atau bukan spam, tumor jinak atau ganas.
b. Klasifikasi multi-kelas: lebih dari 2 kelas, contoh : memprediksi jenis bunga dari suatu foto.
c. Klasifikasi multi-label: sebuah data dapat di label dengan lebih dari satu kelas, contoh : tag pada foto instagram.

Klasifikasi digunakan dalam berbagai bidang, seperti pengenalan wajah, analisis teks, pemrosesan gambar, dan analisis data (Yunial, 2020). Dalam pembelajaran mesin, klasifikasi dapat dilakukan dengan menggunakan algoritma seperti naive bayes, decision tree, random forest, atau support vector machine (SVM) (Nurfaizah et al., 2019). Pemilihan algoritma tergantung pada karakteristik dari data yang akan diklasifikasikan.

### 2.3.2. Preprocessing Data

Preprocessing data adalah serangkaian langkah atau teknik yang diterapkan pada dataset sebelum digunakan sebagai input untuk model pembelajaran mesin. Tujuan preprocessing data adalah untuk mempersiapkan data agar informasi penting dapat diekstraksi dengan lebih efektif dan akurat oleh model. Berikut adalah langkah-langkah preprocessing yang diterapkan dalam penelitian ini:

1. Mengatasi Missing Values

Missing values atau data yang hilang adalah nilai yang tidak disimpan dalam dataset karena berbagai alasan, seperti kesalahan pengukuran atau ketidakhadiran responden. Menangani missing values sangat penting karena dapat mempengaruhi kualitas dan keandalan model. Metode untuk mengatasi missing values meliputi:

- Menghapus Missing Values: Pendekatan ini digunakan jika proporsi missing values kecil dan tidak terkait dengan variabel penting. Metode ini mencakup menghapus seluruh baris atau kolom yang memiliki missing values.
- Melakukan Imputasi: Mengganti missing values dengan nilai tertentu, seperti rata-rata (mean), median, atau modus (mode) dari data yang ada. Metode ini membantu menjaga integritas data dan mengurangi bias yang mungkin muncul karena missing values. Menurut Kaur et al. (2020) dan Mollick et al. (2022), penggantian missing values dengan median lebih disarankan karena median lebih tahan terhadap outliers.

2. Deteksi dan Penanganan Outliers:

Outliers adalah nilai-nilai ekstrem yang berbeda dari sebagian besar data. Deteksi dan penanganan outliers penting untuk memastikan analisis statistik yang akurat.

3. Standarisasi Data:

Standarisasi data dilakukan untuk memastikan bahwa data memiliki format dan skala yang seragam. Hal ini penting bagi algoritma seperti KNN yang sensitif terhadap skala data masukan. Standarisasi dilakukan dengan cara mengubah nilai setiap fitur agar memiliki distribusi dengan rata-rata 0 dan standar deviasi 1. Menurut penelitian Wang et al. (2021), standarisasi data penting untuk meningkatkan kinerja model.

### 2.3.3. Diabetes Mellitus

Diabetes Mellitus yaitu penyakit kronis ditandai dengan peningkatan kadar gula (glukosa) darah. Sumber energi utama untuk sel manusia adalah glukosa. Karena sel-sel tubuh tidak dapat menyerap glukosa secara efektif, glukosa akan menumpuk di dalam darah dan menimbulkan berbagai masalah kesehatan. Tubuh dan jiwa pasien dapat berada dalam bahaya jika kondisi tubuh tidak dapat dikendalikan secara memadai (WHO, 2023). Pankreas, organ dibelakang lambung, bertanggungjawab untuk memproduksi jumlah gula yang ada di dalam darah. Ketika seseorang menderita diabetes melitus, pankreas tidak dapat menghasilkan insulin yang cukup untuk memenuhi kebutuhan tubuh. Sel-sel tubuh tidak akan mampu mengubah kadar glukosa menjadi energi jika insulin tidak ada. Faktor genetik, yang mencegah tubuh memproduksi hormon insulin

dalam jumlah yang cukup atau sama sekali, adalah 2 faktor yang menyebabkan kadar insulin rendah atau rendah dalam tubuh manusia seperti usia dan penyakit (Jais et al., 2021). Komponen kedua adalah produksi hormon insulin yang terus menerus oleh tubuh, yang tidak berfungsi dengan benar karena tubuh telah mengembangkan kekebalan atau resistensi terhadapnya. Contoh situasi yang mirip dengan yang dihadapi orang gemuk atau kelebihan berat badan. Gejala awal Diabetes Mellitus seringkali tidak terlihat dan tidak disadari.

Akibatnya, banyak orang yang baru mengetahui dirinya menderita diabetes melitus sampai timbul komplikasi. Oleh karena itu, penting untuk mengetahui tanda dan gejala peringatan dini diabetes melitus agar dapat dicegah sedini mungkin. Berikut ini tanda-tanda diabetes melitus (Miriyala et al., 2022):

a. Sering merasa haus.
b. Sering buang air kecil, terutama di malam hari.
c. Sering merasa sangat lapar.
d. Berat badan turun secara tiba-tiba tanpa sebab yang jelas.
e. Berkurangnya massa otot.
f. Terdapat (keton) hasil sisa pemecahan otot dan lemak dikarenakan tubuh tidak dapat menggunakan gula sebagai sumber energi tubuh yang terdapat pada urine.
g. Lemas.
h. Pandangan kabur.
i. Luka yang sulit sembuh.

j. Seringkali merasakan infeksi, misalnya di gusi, kulit, vagina, atau saluran kemih.

Beberapa gejala lain juga bisa menjadi ciri-ciri bahwa seseorang mengalami diabetes, antara lain

a. Mulut kering.
b. Rasa terbakar, kaku, dan nyeri pada kaki.
c. Gatal-gatal
d. Disfungsi ereksi atau impotensi.
e. Mudah tersinggung.
f. Mengalami hipoglikemia reaktif, yaitu hipoglikemia yang terjadi beberapa jam setelah makan akibat produksi insulin yang berlebihan.
g. Munculnya bercak-bercak hitam di sekitar leher, ketiak, dan selangkangan, (akantosis nigrikans) sebagai tanda terjadinya resistensi insulin.

Beberapa orang dapat mengalami kondisi prediabetes, yaitu kondisi ketika glukosa dalam darah di atas normal, namun tidak cukup tinggi untuk didiagnosis sebagai diabetes. Seseorang yang menderita prediabetes dapat menderita diabetes tipe 2 jika tidak ditangani dengan baik .

### 2.3.4. Exploratory Data Analysis (EDA)

EDA adalah sebuah langkah pertama dan terpenting dalam persiapan data, untuk menganalisis sebuah dataset lebih lanjut, untuk menemukan pola, memeriksa asumsi, data yang hilang (missing value), menemukan anmali, dan outlier (Peng et al., 2021). Pada dasarnya digunakan untuk memahami isi dari data, seringkali menggunakan grafik statistik dan metode visualisasi data lainnya.

Pada umumnya EDA digunakan dalam beberapa cara:

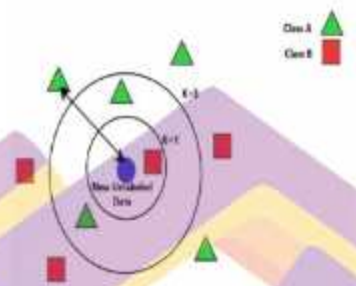
1. Univariat Analysis merupakan analisis deskriptif dengan satu variabel.
2. Bivariat Analysis merupakan analisis relasi dengan dua variabel yang biasanya dengan target variabel.
3. Multivariat Analysis merupakan analisis yang menggunakan lebih dari atau sama dengan tiga variabel.

EDA adalah langkah penting sebelum memulai pembelajaran mesin atau pemodelan statistik untuk mengembangkan model yang sesuai dengan masalah yang dihadapi dan menyediakan konteks yang diperlukan untuk menginterpretasikan hasil dengan benar. EDA sangat penting untuk memastikan bahwa hasil yang dihasilkan oleh ilmuwan data valid, ditafsirkan dengan benar, dan dapat diterapkan pada konteks bisnis yang diinginkan (Siambaton et al., 2022).

### 2.3.5. K-Nearest Neighbors (KNN)

Salah satu teknik untuk mengklasifikasikan data menggunakan atribut dan sampel yang dihasilkan dari dataset adalah K-Nearest Neighbor (KNN). KNN adalah algoritma yang mengkategorikan kelas sampel berdasarkan kelas tetangga terdekatnya, dan pertama kali diperkenalkan oleh T. Cover dan P. Hart pada tahun 1967 (Hamraz et al., 2022). Hasil klasifikasi ketetanggaan, yang berfungsi sebagai nilai prediksi untuk instance baru, ditentukan menggunakan metode KNN. Sebagian besar kategori tetangga terdekat K dapat digunakan untuk klasifikasi dalam contoh baru. Rumus untuk menghitung jarak antara dua lokasi dalam ruang

dua dimensi (Taunk et al., 2019). Untuk memahami kinerja algoritma ditunjukkan pada gambar dibawah ini.



Gambar 2. 1. Persebaran Algoritma KNN

Pada Gambar 2.1. diatas, untuk mencari nilai K algoritma KNN akan mencari berdasarkan jalur terpendek antara data baru dan data training. Hasil terbaik dari berbagai data pelatihan adalah nilai K. Hasil dapat diperoleh dari nilai K melalui optimasi parameter. Nilai K yang tinggi seringkali akan mengurangi dampak noise pada setiap klasifikasi sekaligus menyebabkan perbedaan di antara setiap klasifikasi menjadi lebih kabur (Taunk et al., 2019).

Fungsi jarak, yang digunakan untuk mengukur seberapa dekat suatu data dengan data terdekat lainnya, digunakan dalam kategori fundamental untuk menentukan data baru menggunakan metode KNN. Kedekatan dua bagian data sering diukur menggunakan berbagai fungsi jarak yaitu $X1 = (x\,11\,, x\,21\,, x\,31\,\ldots\, x\,m1\,)$ dan $X2 = (x\,12\,, x\,22\,, x\,32\,\ldots\, x\,m2\,)$ (Taunk et al., 2019).

Jarak Euclidean: Euclidean distance merupakan salah satu metode perhitungan jarak yang digunakan untuk mengukur jarak dari 2 (dua) buah titik dalam Euclidean space (meliputi bidang euclidean dua dimensi, tiga dimensi, atau

bahkan lebih). Untuk mengukur tingkat kemiripan data dengan rumus euclidean distance digunakan rumus berikut (Serrano et al., 2021).

$$d(x, y) = \sqrt{\sum i = 1m(xi - yi)\,2} \qquad (1)$$

dimana,

d = jarak antara x dan y x = data pusat klaster

y = data pada atribut i = setiap data

n = jumlah data

xi = data pada pusat klaster ke i yi = data pada setiap data ke i

Supervised learning adalah jenis machine learning yang dimana model diinstruksikan atau diberi pengawasan dengan data pelatihan yang sudah diberi label. Dalam kata lain, model learning dari contoh-contoh data yang sudah memiliki label atau kelas yang benar. Tujuan utama dari supervised learning adalah untuk menghasilkan model yang dapat memetakan input ke output yang benar, atau dengan kata lain, mempelajari hubungan antara input dan output berdasarkan data pelatihan yang ada. Beberapa contoh tugas supervised learning meliputi klasifikasi (mengelompokkan data ke dalam kategori atau kelas yang sudah ditentukan), regresi (memprediksi nilai berkelanjutan), dan deteksi anomali (mengidentifikasi data yang tidak biasa) (Bradley C. Love, 2022).

### 2.3.6. LightGBM

LightGBM merupakan sebuah metode berbasis Ensemble, yang merupakan sebuah metode gradien efisien yang didasarkan dari pohon keputusan (Al Daoud, 2019). Metode ini dikembangkan oleh Microsoft , algoritma ini sendiri merupakan sebuah metode yang populer dan secara konsisten dapat

memecahkan masalah dari klasifikasi. LightGBM mempercepat waktu proses 20 kali lebih cepat dari fase pelatihan dibandingkan dari beberapa metode Gradient Boosting Decision Tree (GBDT) yang lain (Ke et al, 2017). Sehingga dapat memberikan hasil model dari akurasi yang lebih efektif. Pada penggunaan metode ini dapat digunakan dari berbagai model klasifikasi dan studi regrsi sehingga mencapai hasill deteksi yang sangat baik , hal ini dibuktikan bahwa LightGBM adalah sebuah algoritma yang sangat efektif (Rufo et al, 2021).

LightGBM algoritma yang dirancang oleh Microsoft Research Asia menggunakan kerangka Gradient Boosting Decision Tree (GBDT) (Ke et al, 2017). Tujuannya untuk meningkatkan efisiensikomputasi, sehingga masalah prediksi dengan big data dapat diselesaikan dengan efisien (Liang et al, 2020). LightGBM memiliki beberapa keunggulan dibandingkan metode GBDT lainnya, yaitu kecepatan pelatihan lebih cepat, efisiensi lebih tinggi, penggunaan memori lebih rendah, tingkat akurasi lebih baik, kemampuan dalam menangani data dengan skala yang besar dan dukungan pembelajaran paralel dan GPU (Rufo et al, 2021). LightGBM adalah kerangka Gradient Boosting yang cepat, terdistribusi dan berkinerja tinggi berdasarkan algoritma pohon keputusan yang dapat digunakan untuk peringkat, klasifikasi, regresi dan banyak tugas pembelajaran mesin lainnya (Rufo et al, 2021).

Diasumsikan bahwa himpunan data mentah dengan contoh $N = \{1,2, \dots, n\}$ dan model LightGBM yang memiliki $T = \{1,2, \dots t\}$ pohon dihasilkan. Setelah iterasi t kali, prediksi akhir sama dengan jumlah yang pertama $(1 - t)\ th$ dan $tth$. Proses iterasi digambarkan sebagai berikut (Zhang et al, 2020):

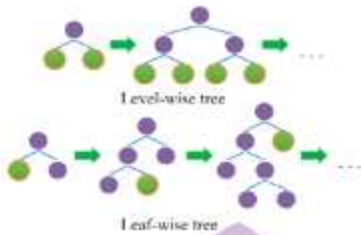
$$yi\ \wedge(t) = yi\ \wedge(t-1) + fi(xi) \quad (2)$$

Dimana $yi\ \wedge(t)$ adalah nilai prediksi contoh $ith$ pada iterasi $tth$. $yi\ \wedge(t-1)$menunjukkan model pohon yang dihasilkan sebelumnya dan $fi(xi)$ menunjukkan model baru yang dibangun.

Oleh karena itu, menurut persamaan (1), setiap prediksi baru dihasilkan oleh residu dan prediksi sebelumnya. Proses pelatihan lengkap digambarkan pada persamaan (2), sedangkan istilah regularisasi dapat dihitung menggunakan persamaan (3) yang digunakan untuk mengurangi kompleksitas model dan dapat digunakan untuk meningkatkan kegunaan pada dataset lainnya.

Dimana $yi$ adalah nilai aktual, $yi\ (t)$ adalah nilai prediksi. $\sum l$ mewakili jumlah kerugian antara masing-masing kelompok $yi$ dan $yi\ (t)$ . dan $\Omega(fi)$ adalah istilah regularisasi. $T$ adalah jumlah daun, $\omega$ adalah bobot daun, $\lambda$ dan $\gamma$ adalah koefisien, dengan nilai default ditetapkan untuk $\gamma = 0$ dan $\lambda = 1$.

LightGBM memiliki karakteristik yang membedakannya dengan algoritma tree boosting lainnya adalah dengan membelah pohon secara memanjang (leaf-wise tree growth) dengan yang paling cocok, sedangkan algoritma tree boosting lainnya membagi pohon secara mendalam atau sejajar (level-wise tree growth) yang ditunjukkan pada Gambar 2. 2. dibawah ini. Oleh karena itu, ketika tumbuh pada daun yang sama di LightGBM, algoritma leaf-wise dapat mengurangi lebih banyak kerugian dari pada algoritma level wise dan juga menghasilkan akurasi yang jauh lebih baik yang tidak dipenuhi oleh algoritma boosting lainnya (Rufo et al, 2021). Namun, algoritma leaf-wise cenderung lebih rentan terhadap overfitting.




Gambar 2. 2. Algoritma LightGBM

### 2.3.7. Dataset

Database Diabetes Indian Pima, sebuah situs publik yang melakukan studi populasi suku Indian Pima, yaitu dataset yang akan digunakan pada penelitian ini. Salah satu suku Indian di Amerika adalah suku Indian Pima. Pima adalah sekelompok penduduk asli Amerika yang tinggal di Arizona selatan dekat sungai Gila dan Colorado dan menggunakan nama Akimel O'odham (istilah lainnya Akimel O'otham) (Chang et al., 2022). Jadi hanya pasien wanita di atas usia 20 tahun yang menjadi sampel semua pasien dalam kumpulan data ini.

Teknik K-Nearest Neighbor dan LightGBM akan digunakan untuk menangani data uji dan data latih dari dataset ini ketika dimasukkan ke dalam database sistem. Ada 768 data dan propertinya dalam dataset. Nantinya dataset ini akan dipecah menjadi training set dan testing set. Terdapat 9 faktor dalam data diabetes, diantaranya (Lakhwani et al., 2020): Terdapat 8 variabel bebas dan 1 variabel terikat.

1. Jumlah Kehamilan : mencatat berapa kali seorang wanita pernah hamil. Wanita yang mengalami lebih banyak kehamilan mungkin memiliki risiko lebih tinggi untuk mengembangkan diabetes tipe 2. Ini disebabkan oleh

perubahan hormon dan metabolisme selama kehamilan yang dapat meningkatkan resistensi insulin (Matsuba et al., 2021).

2. Glukosa : mengonsumsi larutan glukosa selama tes toleransi glukosa oral (TTGO), yang membantu dalam diagnosis diabetes. Pasien berpuasa selama 8 jam, minum larutan dengan 75 gram glukosa, dan kadar glukosa diukur setelah 2 jam. Hasilnya diklasifikasikan sebagai normal (<140 mg/dL), prediabetes (140-199 mg/dL), atau diabetes (≥200 mg/dL). Pengukuran ini penting untuk mendeteksi diabetes dan gangguan glukosa pada seseorang tanpa gejala (Matsuba et al., 2021)..
3. Tekanan darah : *diastolik* merupakan tekanan darah dalam arteri ketika jantung beristirahat antara detak, diukur dalam milimeter air raksa (mm Hg) (Chobanian et al., 2023). Tekanan darah diastolik normal berkisar antara 60-80 mm Hg, sementara tekanan darah diastolik tinggi, yaitu di atas 90 mm Hg, dapat meningkatkan risiko komplikasi *kardiovaskular* seperti penyakit jantung dan stroke. Menjaga tekanan darah diastolik sangat penting untuk mencegah dan mengatasi hipertensi, yang seringkali berkaitan dengan diabetes tipe 2.
4. Lemak Tubuh : ukuran ketebalan lapisan lemak subkutan di belakang lengan atas, yang diukur dengan kaliper. Ukuran ini membantu menilai cadangan lemak tubuh dan risiko kesehatan seperti diabetes tipe 2 (Li et al., 2022). Ketebalan lipatan kulit triseps yang lebih tinggi berkorelasi dengan resistensi insulin dan kebutuhan dosis insulin harian yang lebih besar pada individu dengan diabetes tipe 2, menunjukkan hubungan antara

penyimpanan lemak subkutan dan metabolisme glukosa (Larasati et al., 2023).

### 2.3.8. Faktor Risiko Diabetes Berdasarkan Dataset

Diabetes adalah kondisi kesehatan kronis yang dipengaruhi oleh berbagai faktor risiko yang saling berinteraksi. Faktor-faktor ini bisa bersifat genetik, lingkungan, maupun perilaku, dan mereka berkontribusi pada perkembangan diabetes tipe 1 dan tipe 2. Mengetahui dan memahami faktor-faktor risiko ini sangat penting dalam pencegahan, diagnosis dini, dan manajemen diabetes yang efektif. Penelitian menunjukkan bahwa pengendalian faktor risiko utama seperti kadar glukosa, tekanan darah, dan indeks massa tubuh (BMI) dapat mengurangi komplikasi terkait diabetes serta meningkatkan kualitas hidup pasien.

#### 2.3.8.1.Kadar Glukosa dan Tekanan Darah

Kadar glukosa dan tekanan darah merupakan sebuah komponen kunci dalam memanajemen penyakit diabetes. Kadar glukosa normal pada seseorang kurang dari 140 mg/dL, dan tekanan darah *diastolic* yang normal berada diangka kurang dari 80 mmHg, dari angka tersebut memberi artian bahwa dapat mengurangi risiko dari komplikasi dari penyakit diabetes (Alhassan et al., 2022). Kadar glukosa yang dapat dijaga menunjukkan bahwa pasien memiliki manajemen gula darah yang baik, sedangkan tekanan darah *diastolik* yang rendah mengindikasikan risiko rendah untuk komplikasi *kardiovaskular*. Kombinasi kedua parameter ini memberikan gambaran yang lebih baik tentang kesehatan pasien dan membantu dalam memprediksi risiko diabetes dengan lebih akurat. Penelitian menunjukkan bahwa kontrol yang

baik terhadap kedua fitur ini dapat mengurangi risiko komplikasi yang sering terjadi pada pasien diabetes, seperti penyakit jantung dan stroke (Alhassan et al., 2022).

#### 2.3.8.2. Indeks Massa Tubuh (BMI)

Indeks Massa Tubuh (BMI) adalah pengukuran yang digunakan untuk menentukan kategori berat badan seseorang berdasarkan tinggi dan berat badannya. BMI dihitung dengan membagi berat badan dalam kilogram dengan kuadrat tinggi badan dalam meter ($kg/m^2$). Pengelompokan BMI yang umum digunakan adalah sebagai berikut:

- BMI < 18.5 $kg/m^2$ : Berat badan kurang (underweight)
- BMI 18.5 - 24.9 $kg/m^2$ : Berat badan normal
- BMI 25 - 29.9 $kg/m^2$ : Berat badan lebih (overweight)
- BMI ≥ 30 $kg/m^2$ : Obesitas

BMI adalah indikator penting dalam menilai risiko berbagai kondisi kesehatan, termasuk diabetes tipe 2. Penelitian menunjukkan bahwa BMI yang lebih tinggi secara signifikan meningkatkan risiko pengembangan diabetes tipe 2 dan komplikasi terkait. Orang dengan BMI di atas 30 $kg/m^2$ memiliki risiko yang jauh lebih tinggi untuk mengembangkan resistensi insulin, yang merupakan faktor utama dalam patofisiologi diabetes tipe 2 (Rojas et al., 2021). Obesitas juga terkait dengan berbagai komplikasi kardiovaskular, seperti hipertensi, penyakit jantung koroner, dan stroke. Studi menunjukkan bahwa pengurangan berat badan dapat memperbaiki kontrol glukosa darah dan menurunkan tekanan darah, sehingga mengurangi risiko komplikasi kardiovaskular pada pasien diabetes (Alhassan et al., 2022) (Rojas et al., 2021). Selain itu, pengendalian berat badan melalui pola makan sehat

dan aktivitas fisik adalah strategi pencegahan utama yang direkomendasikan oleh berbagai organisasi kesehatan untuk mengurangi insidensi diabetes dan meningkatkan kualitas hidup (Alhassan et al., 2022) (Rojas et al., 2021). Oleh karena itu, menjaga di bawah BMI 30 kg/m² dianggap sebagai salah satu indikator kesehatan yang penting dalam pencegahan dan manajemen diabetes. Dengan BMI yang lebih rendah, risiko pengembangan diabetes dan komplikasi terkait dapat diminimalkan, sehingga meningkatkan hasil kesehatan secara keseluruhan bagi individu (Alhassan et al., 2022).

**2.3.8.3.Riwayat Keluarga**

Riwayat keluarga merupakan salah satu faktor genetik yang mempengaruhi seseorang untuk memiliki penyakit diabetes. Seseorang yang memiliki keluarga penderita Diabetes Mellitus memiliki resiko 3 kali lipat terkena Diabetes Mellitus. Riwayat keluarga memiliki risiko terkena diabetes sebesar 15%. Jika kedua orang tua menderita diabetes melitus, maka risiko terkena diabetes melitus adalah 75%. Risiko terkena diabetes dari ibu 10-30% lebih besar dibandingkan ayah (Ammutammima et al., 2021). Penelitian sebelumnya dilakukan oleh (Kosanto et al., 2016), Hasil distribusi menunjukkan bahwa didapatkan yang memiliki riwayat DM dalam keluarga sebanyak 7 orang (13,5%),

Riwayat keluarga mempunyai pengaruh terkuat terhadap kejadian gestasional diabetes mellitus dengan nilai ( P <. 001). Seperti yang diketahui bahwa diabetes cenderung diturunkan atau diwariskan, dan tidak ditularkan. Faktor genetis memberi peluang besar menderita diabetes dibandingkan

dengan anggota keluarga yang tidak menderita diabetes (Price et al., 2017). Apabila ada orang tua atau saudara kandung yang menderita diabetes, maka seseorang tersebut memiliki risiko 40% menderita diabetes. Hal ini sejalan dengan teori menurut Smeltzer dan Bare bahwa salah satu faktor risiko terjadinya diabetes adalah faktor keturunan.

**2.3.8.4.Umur Ibu Hamil**

Umur ibu hamil menjadi faktor yang mempengaruhi diabetes mellitus gestasional yang merupakan penyakit diabetes yang terjadi pada saat kehamilan, yang sebelumnya tidak memiliki diabetes. Diabetes Mellitus Gestasional meningkat seiring bertambahnya usia (Li et al., 2020). Hal ini terjadi karena pada usia lebih dari 35 tahun terjadi penurunan fungsi metabolisme dalam tubuh. Penurunan fungsi metabolisme tubuh dipengaruhi oleh penurunan jumlah otot yang diakibatkan oleh semakin tingginya usia. Penelitian sebelumnya yang dilakukan oleh Rahmwati & Bachri (2019) yang menemukan bahwa, Masih banyak ibu hamil di usia >35 tahun sebanyak 66 atau 32,50%. usia tersebut merupakan usia beresiko terjadinya diabetes mellitus gestasional. Hal ini dimungkinkan karena kurangnya pengetahuan ibu hamil tentang diabetes gestasional. Kurangnya pengetahuan tersebut diakibatkan karena kurangnya informasi yang diterima ibu hamil dan ini sangat dimungkinkan karena rendahnya penyerapan terhadap informasi yang dipengaruhi oleh rendahnya tingkat pendidikan ibu hamil. Sehingga petugas kesehatan harus memberikan penyuluhan untuk menghindari hamil pada usia >35 tahun serta sebaiknya para calon ibu hamil merencanakan kapan akan

hamil dimulai dari rencana menikah pada usia diatas 20 tahun dan pada saat masa kehamilan usia ibu antara 20- 35 tahun. alasan dianjurkan untuk hamil pada rentan usia 20-35 tahun adalah pada usia tersebut kematangan organ (fisik) sudah siap.

Kemudian jumlah kehamilan yang tinggi juga berpengaruh atas terjadinya penyakit diabetes. Wanita yang memiliki banyak anak memiliki risiko lebih tinggi untuk mengembangkan diabetes tipe 2 (Alhassan et al., 2022)..Frekuensi kehamilan yang lebih tinggi dikaitkan dengan peningkatan risiko resistensi insulin dan diabetes tipe 2. Faktor-faktor seperti kenaikan berat badan yang berulang, perubahan hormonal, dan peningkatan resistensi insulin selama kehamilan dapat berkontribusi pada risiko ini.

**2.3.8.5.Lemak Tubuh**

Persentase lemak tubuh yang tinggi merupakan faktor risiko yang besar untuk diabetes tipe 2. Lemak tubuh yang berlebihan, terutama lemak visceral yang mengelilingi organ dalam, dapat menyebabkan resistensi insulin. Resistensi insulin terjadi ketika sel-sel tubuh tidak merespons insulin dengan baik, yang menyebabkan peningkatan kadar glukosa dalam darah. Akumulasi lemak di daerah perut lebih berbahaya dibandingkan dengan lemak yang tersebar secara merata di seluruh tubuh (Zhang et al., 2022).

Kadar Lemak Tubuh Normal:

- Pria : Lemak tubuh normal berkisar antara 18-24%.
- Wanita : Lemak tubuh normal berkisar antara 25-31%

Untuk rata-rata lemak tubuh normal yang digabungkan antara pria dan wanita, dapat mengambil rata-rata dari kisaran normal untuk kedua jenis kelamin. Jika digabungkan menjadi kisaran 25%.

Seseorang dengan BMI normal tetapi persentase lemak tubuh yang tinggi lebih mungkin memiliki pradiabetes atau diabetes dibandingkan dengan mereka yang memiliki BMI lebih tinggi tetapi persentase lemak tubuh lebih rendah. Hal ini menunjukkan bahwa persentase lemak tubuh mungkin menjadi prediktor yang lebih baik untuk risiko diabetes dibandingkan BMI (Zhang et al., 2022).

**2.3.8.6.Insulin**

Insulin adalah hormon yang diproduksi oleh sel beta di pankreas yang berperan penting dalam pengaturan kadar glukosa darah. Insulin memungkinkan sel-sel tubuh untuk mengambil glukosa dari darah dan menggunakannya sebagai energi atau menyimpannya sebagai lemak. Pada diabetes tipe 1, pankreas tidak menghasilkan insulin, sehingga penderita memerlukan suntikan insulin untuk mengontrol kadar gula darah. Pada diabetes tipe 2, tubuh tidak menggunakan insulin secara efektif, yang dikenal sebagai resistensi insulin. Seiring waktu, pankreas mungkin tidak mampu memproduksi cukup insulin untuk mengatasi resistensi ini, yang menyebabkan peningkatan kadar glukosa darah (Matsuba et al., 2012).

Kadar Insulin Normal: Kadar insulin dalam darah bervariasi sepanjang hari, terutama sebelum dan sesudah makan. Berikut adalah kisaran kadar insulin yang dianggap normal:

- Sebelum makan : 2-25 µU/mL (mikro unit per mililiter)
- Setelah makan : 150 µU/mL

### 2.3.9. Ensemble Learning dengan Stacking

Ensemble learning adalah teknik yang menggabungkan beberapa model untuk meningkatkan akurasi prediksi dan mengurangi *overfitting*. Teknik umum dalam ensemble learning adalah *bagging, boosting*, dan *stacking*. Dalam teknik *stacking*, model individu *(base learners)* dilatih terlebih dahulu, kemudian output dari model-model ini digunakan sebagai input untuk model meta-learner yang membuat prediksi akhir (Zhang et al., 2022).

Salah satu aspek penting dalam ensemble learning adalah mekanisme voting. Menggunakan jumlah model ganjil sering digunakan untuk menghindari hasil seri *(tie)* dalam voting, sehingga memberikan keputusan akhir yang lebih pasti (Zhang et al., 2022). Namun, dalam konteks stacking, mekanisme voting tidak selalu menjadi masalah utama karena pendekatan ini menggabungkan hasil prediksi model individu melalui model meta-learner.

Pada penelitian ini, akan menggunakan dua algoritma, KNN dan LightGBM, dalam pendekatan stacking. Meskipun jumlah model yang digunakan adalah genap, hasil penelitian menunjukkan peningkatan akurasi yang signifikan tanpa adanya *deadlock*. Jika diperlukan, metode ini dapat diperluas dengan menambahkan algoritma ketiga untuk memperkuat hasil prediksi (Zhang et al., 2022).

### 2.3.10. Confusion Matrix

Confusion Matrix adalah tabel dengan 4 kombinasi berbeda dari nilai prediksi dan nilai actual (Narkhede, 2018). Ada empat istilah yang merupakan representasi hasil proses klasifikasi pada confusion matrix yaitu *True Positive* (TPs), True Negative (TNs), False Positive (FPs), dan False Negative (FNs) (Xu et al., 2020)

Tabel 2. 2. Confusion Matrix

| Actually / Predicted | Actually Positive | Actually Negative |
|---|---|---|
| Predicted Positive | True Positive (TP) | False Positive (FP) |
| Predicted Negative | False Negative (FN) | True Negative (TN) |

A. Akurasi

Akurasi merupakan seberapa akurat model dalam mengklasifikasikan dengan benar.

$$Akurasi = \frac{TPs+TNs}{TPs+TNs+FPs+FNs} \quad (1)$$

B. Presisi

Presisi merupakan kurasi antara data yang diminta dengan hasil prediksi yang diberikan oleh model.

$$Presisi = \frac{TPs}{TPs+FPs} \quad (2)$$

C. Recall

Recall merupakan keberhasilan model dalam menemukan kembali sebuah informasi.

$$Recall = \frac{TPs}{TPs+FNs} \quad (3)$$

D. F1-Score

F1-Score merupakan perbandingan rata-rata *precision* dan *recall* yang dibobotkan. *Accuracy* tepat kita gunakan sebagai acuan performansi algoritma jika dataset kita memiliki jumlah data False Negatif dan False Positif yang sangat mendekati *(symmetric)*. Namun jika jumlahnya tidak mendekati, maka sebaiknya kita menggunakan F1-Score sebagai acuan.

$$F1\ Score = \frac{2\ x\ Presisi\ x\ Recall}{Presisi+Recall} \quad (4)$$

# BAB III

# METODE PENELITIAN

## 3.1. Jenis, Sifat, dan Pendekatan Penelitian

Jenis penelitian ini adalah penelitian kuantitatif yang bersifat komparatif atau membandingkan beberapa metode pengukuran kadar glukosa, tekanan darah, ketebalan kulit, insulin, berat badan (BMI), riwayat keturunan penyakit diabetes pada keluarga agar dapat diteliti akurasinya , pada tahap ini akan dilakukan asumsi berdasarkan prediksi dari pasien mengidap penyakit atau tidak, yaitu :

1. Wanita dengan banyak kehamilan memiliki peluang lebih tinggi untuk positif mengidap penyakit diabetes.
2. Seseorang dengan konsentrasi glukosa plasma yang tinggi akan memiliki peluang lebih tinggi untuk positif terhadap diabetes.
3. Seseorang dengan tekanan darah tinggi akan memiliki peluang lebih tinggi untuk positif mengidap penyakit diabetes.
4. Seseorang dengan nilai rata-rata tricep yang lebih tinggi (lemak tubuh) akan memiliki peluang lebih tinggi untuk positif mengidap penyakit diabetes.
5. Seseorang dengan jumlah insulin yang tinggi akan memiliki peluang lebih tinggi untuk positif mengidap penyakit diabetes.
6. Seseorang dengan BMI yang tinggi akan memiliki peluang lebih tinggi.

### 3.2. Metode Pengumpulan Data

Metode pengumpulan data yang dilakukan berasal dari Kaggle dengan mengambil dataset berjudul Pima Indians Diabetes Dataset (PIDD) UCI Machine Learning (2016). Terdiri atas 8 variabel dengan jumlah data sebanyak 768 data.

### 3.3. Metode Analisis Data

1. Analisis Eksploratori Data (EDA)

   EDA merupakan metode untuk menganalisis data yang penting pada penelitian ini. Tujuannya untuk mencari dan memahami dataset yang digunakan. EDA akan melibatkan karakteristik dataset, identifikasi nilai – nilai yang hilang, deteksi outlier, dan hubungan antar variabel dengan menggunakan teknik visualisasi data dan analisis statistik deskriptif. Tools yang digunakan dalam EDA meliputi:

   - Histogram: Untuk melihat distribusi frekuensi dari setiap fitur dan mengidentifikasi skewness atau distribusi yang tidak normal.
   - Scatter Plot: Untuk visualisasi hubungan antara dua variabel numerik dan mengidentifikasi korelasi signifikan.
   - Box Plot : Untuk mendeteksi outliers atau nilai ekstrim dalam data.
   - Heatmap Korelasi: Untuk menunjukkan hubungan korelasi antara semua fitur dalam dataset dan mengidentifikasi fitur-fitur yang memiliki korelasi tinggi dengan label target.
   - Pairplot: Menggabungkan histogram dan scatter plot untuk beberapa pasang fitur, memungkinkan visualisasi hubungan antar fitur secara lebih komprehensif.

2. Penerapan Algoritma K-Nearest Neighbors

Metode yang digunakan pada analisis data ini melibatkan algoritma KNN untuk melakukan klasifikasi penyakit diabetes pada dataset Pima Indians. Langkah ini mencakup penentuan jumlah tetangga terdekat dengan penggunaan metrik jarak yang sesuai. Optimasi parameter dilakukan menggunakan GridSearchCV untuk mencari parameter terbaik.

3. Penerapan Algoritma LightGBM

Selanjutnya, metode analisis data ini melibatkan penerapan algoritma LightGBM untuk mengklasifikasi penyakit diabetes dengan metode berbasis boosting yang efisien dan kuat. Optimasi parameter dilakukan menggunakan RandomSearchCV untuk mencari parameter terbaik.

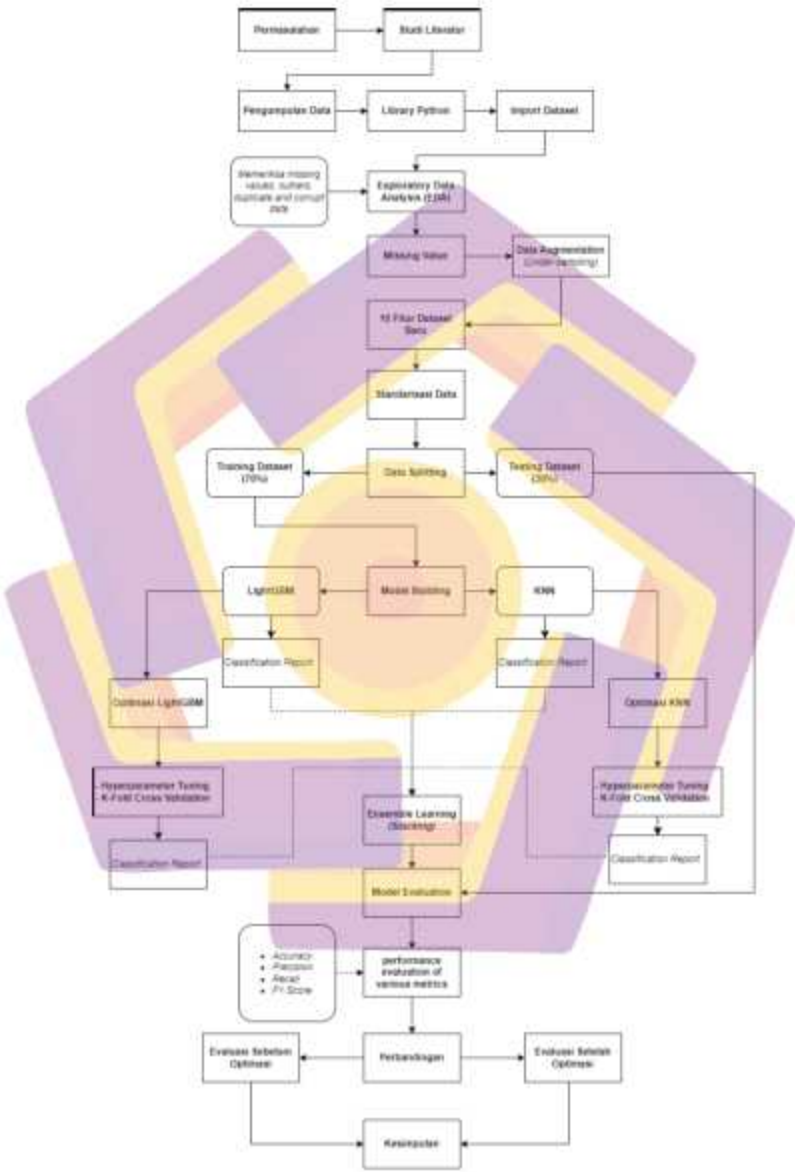
4. Penerapan Algoritma LightGBM

Setelah dilakukan pemodelan, evaluasi dilakukan menggunakan confusion matrix, yang memberikan informasi tentang seberapa baik kedua model dapat melakukan klasifikasi penyakit diabetes. Metrik evaluasi yang digunakan meliputi akurasi, presisi, recall, dan F1-Score.

5. Analisis Komparatif

Metode analisis data ini melibatkan analisis komparatif antara hasil kinerja KNN, EDA, dan LightGBM. Analisis ini akan membantu mengidentifikasi kelebihan dan kekurangan masing-masing metode dalam proses pengklasifikasian. Selain itu, teknik ensemble learning dengan metode stacking akan digunakan untuk menggabungkan prediksi dari kedua model untuk meningkatkan akurasi dan stabilitas prediksi.

### 3.4. Alur Penelitian




Gambar 3. 1. Alur Penelitian

Alur penelitian dapat dilihat pada gambar 3.1. Alur penelitian yang Anda tunjukkan melibatkan beberapa tahapan penting dalam pengembangan model prediksi diabetes menggunakan teknik Machine Learning, khususnya dengan menggunakan algoritma KNN dan LightGBM dalam pendekatan stacking. Berikut ini adalah penjelasan untuk masing-masing bagian dari alur tersebut:

a. Permasalahan

Permasalahan: Tahap awal ini menentukan masalah yang ingin diselesaikan, dalam hal ini adalah prediksi diabetes

b. Studi Literatur

Studi Literatur: Melibatkan penelaahan literatur yang ada untuk mendapatkan wawasan tentang metode-metode yang telah digunakan sebelumnya dan menentukan celah yang bisa dieksplorasi.

c. Pengumpulan Data:

Data pada penelitian ini adalah dataset penyakit diabetes mellitus yang berasal dari PIMA Indians Diabetes Dataset, yang berjumlah 9 variabel.

d. Library Python & Import Dataset

Mengatur Python dan memuat dataset ke dalam program untuk diproses dan dianalisis. Ini melibatkan penggunaan library seperti *Pandas*, *NumPy*, atau *SciPy* yang esensial untuk manipulasi dan analisis data.

e. Exploratory Data Analysis (EDA)

EDA dilakukan untuk memahami distribusi variabel, mendeteksi outlier, dan mengungkap pola atau anomali dalam data. Tahap ini sangat

penting untuk mendapatkan wawasan yang dapat mengarahkan analisis lebih lanjut dan pembangunan model.

f. Pembersihan Data *(Missing Value)*

Melakukan pengecekan pengubahan data pada variabel tertentu yang kosong, menjadi nilai median atau tengah.

g. Data Augmentation *(Under-sampling)*

Mengurangi jumlah sampel dari kelas yang dominan untuk menyeimbangkan dataset, yang penting dalam kasus ketidakseimbangan kelas.

h. 14 Fitur Dataset Baru

Menambahkan fitur baru yang dihasilkan dari kombinasi atau transformasi fitur asli yang dapat memberikan informasi tambahan yang berguna untuk prediksi.

i. Standarisasi Data

Menstandarkan dataset untuk memastikan bahwa data memiliki format dan skala yang seragam, yang kritis bagi algoritma seperti KNN yang sensitif terhadap skala data masukan.

j. Pembagian Data *(Data Splitting)*

Membagi dataset menjadi dua: 70% untuk pelatihan dan 30% untuk pengujian, yang merupakan praktik umum dalam pengembangan model Machine Learning.

k. Model Building

LightGBM & KNN: Membangun model menggunakan kedua algoritma ini secara terpisah.

l. Hyperparameter Tuning - K-fold Cross Validation

Menyesuaikan parameter dari masing-masing model untuk mencapai kinerja terbaik dan menggunakan validasi silang K-fold untuk menilai keefektifan model secara objektif.

m. Ensemble Learning (Stacking):

Menggabungkan prediksi dari kedua model (KNN dan LightGBM) untuk menciptakan model final yang lebih robust dan akurat.

n. Model Evaluation

Menilai model akhir menggunakan metrik seperti Akurasi, Presisi, Recall, dan Skor F1. Tahap ini menentukan seberapa baik model memprediksi data baru.

o. Kesimpulan

Menarik kesimpulan dari keseluruhan proses penelitian dan evaluasi model, memberikan insight tentang efektivitas teknik yang digunakan.

Setiap tahap dalam alur ini penting dan saling terkait, dengan tujuan akhir untuk mengembangkan sebuah sistem prediksi yang bisa diandalkan dan efisien dalam mendeteksi diabetes

# BAB IV

# HASIL PENELITIAN DAN PEMBAHASAN

## 4.1. Membangun Dataset

### 4.1.1. Pengumpulan Data

Data yang akan digunakan pada penelitian ini berupa sekumpulan data penyakit Diabetes Mellitus dari populasi penduduk suku *Indian Pima.* Terdapat data diabetes yang diketahui berupa 9 variabel, terdiri atas 8 variabel independent meliputi (Kehamilan, Kadar Glukosa, Tekanan Darah, Lemak Tubuh, Insulin, BMI, Umur dan Riwayat Diabetes dalam Keluarga) dan 1 variabel dependen berupa (Hasil, meliputi positif dan negatif). Dataset *Pima Indians Diabetes Datasets* ini dapat diakses secara publik dari situs *Kaggle* (https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database/data ).

Tabel 4. 1. Detail Dataset

| Kehamilan | Glukosa | Tekanan Darah (mm/hg) | Lemak Tubuh (mm) | Insulin (U/ml) | BMI (kg) | Riwayat Penyakit Diabetes (%) | Umur (Tahun) | Hasil |
|---|---|---|---|---|---|---|---|---|
| 8 | 188 | 78 | 0 | 0 | 47,9 | 0,137 | 43 | Positif |
| 7 | 152 | 88 | 44 | 0 | 50 | 0,337 | 36 | Positif |
| 2 | 99 | 52 | 15 | 94 | 24,6 | 0,637 | 21 | Negatif |
| 1 | 109 | 56 | 21 | 135 | 25,2 | 0,833 | 23 | Negatif |
| 2 | 88 | 74 | 19 | 53 | 29 | 0,229 | 22 | Negatif |
| 8 | 179 | 72 | 42 | 130 | 32,7 | 0,719 | 36 | Positif |
| 4 | 151 | 90 | 38 | 0 | 29,7 | 0,294 | 36 | Negatif |
| 7 | 102 | 74 | 40 | 105 | 37,2 | 0,204 | 45 | Negatif |
| 0 | 131 | 88 | 0 | 0 | 31,6 | 0,743 | 32 | Positif |
| 6 | 104 | 74 | 18 | 156 | 29,9 | 0,722 | 41 | Positif |

Tabel 4. 1. (Lanjutan)

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 3 | 148 | 66 | 25 | 0 | 32,5 | 0,256 | 22 | Negatif |
| 4 | 120 | 68 | 0 | 0 | 29,6 | 0,709 | 34 | Negatif |
| 6 | 102 | 82 | 0 | 0 | 30,8 | 0,18 | 36 | Positif |
| 6 | 134 | 70 | 23 | 130 | 35,4 | 0,542 | 29 | Positif |
| 2 | 87 | 0 | 23 | 0 | 28,9 | 0,773 | 25 | Negatif |
| 1 | 79 | 60 | 42 | 48 | 43,5 | 0,678 | 23 | Negatif |

Pada Tabel 4.1. diatas menunjukkan dataset yang akan digunakan pada penelitian ini, jumlah dari dataset yang akan digunakan berjumlah 768 data, terdiri atas hasil positif dan negatif yang berpengaruh pada tiap variabelnya.

### 4.1.2. Persebaran Data Variabel Hasil

Data yang digunakan pada penelitian ini memiliki persebaran data yang tidak seimbang, dapat dilihat pada gambar 4.1. dibawah ini akan menunjukkan persebaran variabel hasil yang terdiri atas hasil positif dan negatif pada setiap datanya.


Gambar 4. 1. Persebaran Total Dataset

### 4.2. Analisis Data

Sebelum dilakukan proses ke tahap implementasi, untuk menganalisisnya diperlukan suatu skenario agar mendapatkan perbandingan hasil dari nilai akurasi, presisi, recall, dan f1-score menggunakan confusion matrix dan classification report. Berikut skenario yang digunakan dalam penelitian ini adalah sebagai berikut:

a. Data yang digunakan adalah dataset dari kaggle yaitu, Pima Indians Diabetes Dataset (PIDD) yang berjumlah seluruhnya 768 data, meliputi 268 data pasien positif dan 500 pasien negatif.
b. Pada pengolahan data digunakan *random undersampling* untuk menangani data yang tidak seimbang, menjadi 268 data positif dan 268 data negatif.
c. Mencari dan menggantikan menggunakan median nilai yang hilang *(missing value)* pada dataset.
d. Membuat pemodelan KNN dan LightGBM secara terpisah, dilakukan sebelum optimasi dan setelah optimasi.
e. Optimasi KNN dengan menggunakan *Grid Search* untuk mencari parameter terbaik agar terhindar dari *Overfitting* atau agar model menjadi lebih stabil dan kinerjanya lebih baik.
f. Optimasi LightGBM juga menggunakan *Grid Search* untuk mencari parameter terbaik.
g. Kombinasi KNN-LightGBM dengan menggunakan Teknik *Ensemble Learning* dengan menggunakan salah satu metode yaitu *Stacking*.

### 4.3. Penggunaan Exploratory Data Analysis (EDA) dalam Penelitian

*Exploratory Data Analysis* (EDA) adalah langkah penting yang dilakukan pada tahap awal pemrosesan data untuk memahami struktur, pola, dan karakteristik dari dataset yang digunakan dalam penelitian ini. EDA membantu dalam mengidentifikasi anomali, menemukan hubungan antara variabel, dan mendapatkan wawasan mendalam yang dapat mempengaruhi hasil akhir dari analisis data.

Dalam penelitian ini, berbagai tools digunakan dalam EDA dengan kontribusi spesifik sebagai berikut:

1. Histogram: Digunakan untuk melihat distribusi frekuensi dari setiap fitur dalam dataset. Hal ini membantu dalam mengidentifikasi skewness atau data yang tidak terdistribusi secara normal, yang dapat mempengaruhi performa model prediksi.
2. Scatter Plot: Membantu dalam visualisasi hubungan antara dua variabel numerik. Dengan scatter plot, kita dapat mengidentifikasi pola hubungan (linear atau non-linear) antara fitur-fitur tertentu, seperti antara tingkat glukosa darah dan BMI.
3. Boxplot: Berguna untuk mendeteksi outliers atau nilai ekstrim dalam data. Outliers dapat mempengaruhi model secara signifikan, sehingga identifikasi dan penanganannya sangat penting untuk meningkatkan akurasi model.
4. Heatmap Korelasi: Digunakan untuk menunjukkan hubungan korelasi antara semua fitur dalam dataset. Dengan heatmap, fitur-fitur yang

memiliki korelasi tinggi dengan label target dapat diidentifikasi dan digunakan untuk seleksi fitur dalam membangun model prediksi yang lebih akurat.

5. Pairplot: Menggabungkan histogram dan scatter plot untuk beberapa pasang fitur, memungkinkan visualisasi hubungan antar fitur secara lebih komprehensif. Ini membantu dalam memahami interaksi kompleks antara beberapa fitur.

Dengan penggunaan tools-tools ini, EDA dalam penelitian ini berhasil mengungkap wawasan penting seperti hubungan kuat antara kadar glukosa darah dan kemungkinan diabetes, serta membantu dalam pembersihan dan transformasi data. Kontribusi masing-masing tools memberikan pemahaman mendalam mengenai distribusi data, hubungan antar fitur, dan identifikasi *outliers* yang semuanya berkontribusi pada peningkatan akurasi model prediksi diabetes menggunakan kombinasi KNN dan LightGBM.

## 4.4. Implementasi pada Pra-Pemrosesan Data

### 4.4.1. Menyeimbangkan Dataset dengan Teknik Undersampling

Dalam Dataset yang akan digunakan, persebaran kelas antara hasil pasien negatif dan pasien positif tidak seimbang. Dari total 768 sampel, terdapat 268 sampel positif dan 500 sampel negatif. Distribusi yang tidak seimbang ini berpotensi menyebabkan model pembelajaran mesin bias terhadap kelas mayoritas. Untuk mengatasi hal ini, digunakan teknik undersampling pada kelas mayoritas.

Langkah-langkah *Undersampling*, meliputi:

1. Identifikasi kelas mayoritas (negatif) berjumlah 500 dan kelas minoritas (positif) berjumlah 268.
2. Lakukan undersampling pada kelas mayoritas untuk menyamakan jumlah sampel dengan kelas minoritas.
3. Gabungkan kembali subset kelas mayoritas yang telah diundersampling dengan seluruh sampel dari kelas minoritas untuk membentuk dataset yang seimbang.

Berikut adalah script untuk *Undersampling* data:

```
if positif_count > negatif_count:
    # Undersampling kelas mayoritas
    df_majority_undersampled=                df[df['Hasil']                --
1].sample(n=negatif_count, random_state=42)
    df_minority = df[df['Hasil'] == 0]
else:
# Jika negatif lebih banyak, lakukan hal yang sama untuk kelas
negatif
    df_majority_undersampled                -df[df['Hasil']                --
0].sample(n=positif_count, random_state=42)
    df_minority = df[df['Hasil'] == 1]

# Menggabungkan dataframe yang telah disampling kembali
df_balanced = pd.concat([df_majority_undersampled, df_minority])

# Verifikasi jumlah sampel yang seimbang
print(df_balanced['Hasil'].value_counts())
```

Pada Tabel 4.2. dibawah ini merupakan hasil implementasi distribusi kelas sebelum dan sesudah *undersampling*. Teknik *undersampling* ini memastikan bahwa model pembelajaran mesin tidak bias terhadap kelas mayoritas dan dapat mengenali kelas minoritas (pasien positif diabetes) dengan lebih baik. Pra-pemrosesan data ini penting untuk memastikan akurasi

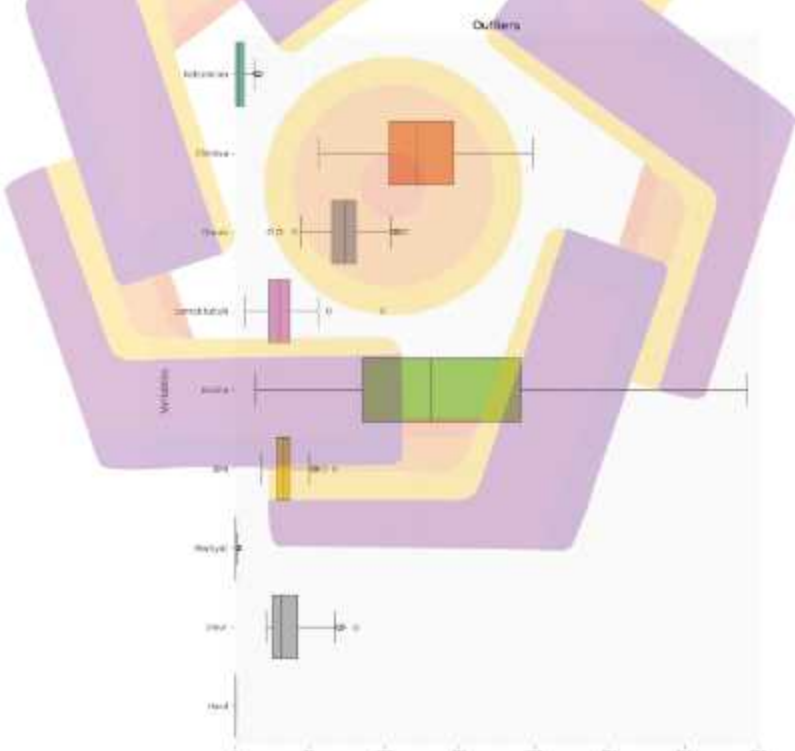
dan keseimbangan model dalam memprediksi diabetes pada dataset *Pima Indians*.

Tabel 4. 2. Implementasi Undersampling

| Kelas | Jumlah Sampel Sebelum | Jumlah Sampel Sesudah |
|---|---|---|
| Negatif *(Mayoritas)* | 500 | 268 |
| Positif *(Minoritas)* | 268 | 268 |

### 4.4.2. Mendeteksi Outliers

Gambar 4.2. di bawah ini menunjukkan visualisasi *boxplot* dari setiap variabel dalam dataset, yang membantu dalam mendeteksi outliers.



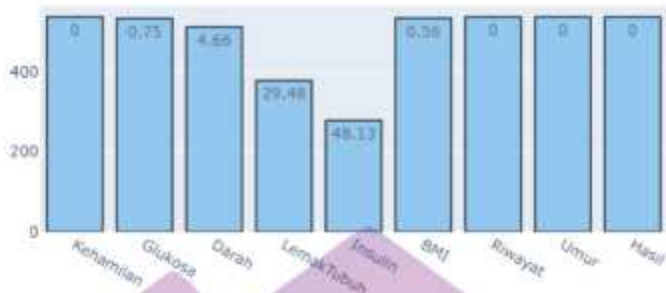

Gambar 4. 2. Outliers

Pada Gambar 4.2. terlihat bahwa ditemukan ada beberapa outlier yang ada pada fitur dalam dataset seperti Kehamilan, Tekanan Darah, Insulin, BMI, Riwayat dan Umur. Namun fitur yang penting dalam penggunaan dalam dataset ini yang ingin di isi yaitu Tekanan Darah, Insulin, BMI, Lemak Tubuh, serta Glukosa dikarenakan fitur tersebut merupakan fitur terpenting dan tidak boleh kosong.

Identifikasi outlier ini penting untuk memutuskan metode pengisian missing values yang akan digunakan. Awalnya, penggunaan mean dipertimbangkan sebagai metode pengisian *missing values*. Namun, hasil eksperimen menunjukkan bahwa penggunaan mean menambah jumlah *outlier* yang signifikan dibandingkan dengan median.

### 4.4.3. Mendeteksi Nilai yang Hilang (Missing Value).

Pada dataset terdapat 8 fitur dari keseluruhan data, dari fitur tersebut terdapat fitur penting dalam konteks medis yang seharusnya tidak boleh hilang. Deteksi dan penanganan *missing values* (nilai yang hilang) merupakan langkah penting untuk memastikan kualitas dan integritas data sebelum digunakan dalam pemodelan. Missing values dapat menyebabkan bias dalam model prediksi jika tidak ditangani dengan baik, dari data yang digunakan terdapat fitur yang kosong seperti fitur (Kadar Glukosa, Tekanan Darah, Lemak Tubuh, Insulin dan Indeks Massa Tubuh (BMI), dari fitur tersebut dikecualikan dari 3 fitur lainnya. Berikut ini akan menunjukkan hasil dari beberapa hasil dari nilai yang hilang.




Gambar 4. 3. Pendeteksian Missing Value

Pada Gambar 4.3. diatas menampilkan hasil dari missing value yang ada pada dataset. Terdapat 5 fitur yang terdeteksi mengalami kekosongan data, terutama fitur Insulin mengalami kekosongan data tertinggi yaitu 48,13% dari keseluruhan data, kemudian Lemak Tubuh sebanyak 29,58%, Tekanan Darah sebanyak 4,66%, BMI sebanyak 0,56%, dan yang paling terkecil adalah Glukosa sebanyak 0,75%.

### 4.4.4. Evaluasi Metode Pengisian Missing Values dan Dampak terhadap Outlier

Dalam penelitian ini, penanganan *missing values* merupakan langkah penting dalam preprocessing data. Terdapat dua metode utama yang dibandingkan adalah *mean* dan *median*. Analisis dilakukan untuk memahami bagaimana masing-masing metode mempengaruhi jumlah outlier dalam dataset.

Hasil eksperimen menunjukkan bahwa penggunaan *mean* untuk mengganti *missing values* menghasilkan peningkatan jumlah outlier yang signifikan dibandingkan dengan median. Tabel di bawah ini merangkum jumlah outlier

yang dihasilkan oleh masing-masing metode untuk variabel-variabel yang memiliki missing values:

Tabel 4. 3. Perbandingan Outlier dengan Mean dan Median

| No. | Fitur | Data Asli | Median | Mean |
|---|---|---|---|---|
| 1. | Kehamilan | 4 | 4 | 4 |
| 2. | **Glukosa** | **0** | **0** | **0** |
| 3. | **Tekanan Darah** | **12** | **14** | **18** |
| 4. | **Lemak Tubuh** | **0** | **3** | **5** |
| 5. | **Insulin** | **19** | **33** | **55** |
| 6. | **BMI** | **6** | **6** | **6** |
| 7. | Riwayat | 19 | 19 | 19 |
| 8. | Umur | 4 | 4 | 4 |
| 9. | Hasil | 0 | 0 | 0 |

Berdasarkan Tabel 4.3. di atas, terlihat bahwa penggunaan *mean* untuk fitur insulin menyebabkan peningkatan jumlah outlier dari 19 menjadi 55, sedangkan penggunaan median menyebabkan peningkatan dari 19 menjadi 33. Hal ini menunjukkan bahwa mean kurang cocok digunakan dalam konteks ini karena menambah jumlah outlier yang dapat mengganggu analisis lebih lanjut.

Oleh karena itu, metode pengisian *missing values* dengan median dipilih untuk diterapkan dalam preprocessing data. Median memberikan representasi yang lebih stabil dan akurat, terutama dalam kondisi data yang memiliki variasi besar namun tidak ekstrem. Selain itu, median lebih tahan terhadap outlier, yang penting untuk menjaga integritas data dan akurasi analisis lebih lanjut.

### 4.4.5. Mengganti Missing Value dengan Median.

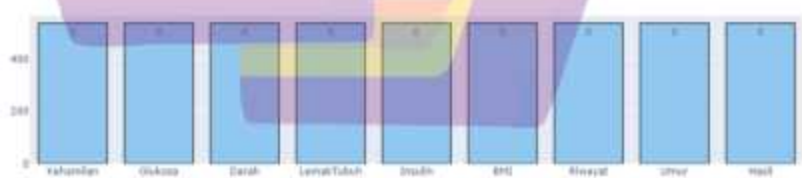
Berdasarkan Gambar 4.3. dapat dilihat bahwa terdapat 5 fitur yaitu, Glukosa, Lemak Tubuh, Insulin, Tekanan Darah dan BMI mengalami

kekosongan data. Dalam penelitian ini, penanganan *missing values* merupakan langkah penting dalam preprocessing data. Dua metode utama yang dibandingkan adalah mean dan median. Dari hasil eksperimen menunjukkan bahwa penggunaan mean untuk mengganti missing values menghasilkan peningkatan jumlah outlier yang tidak sesuai dibandingkan dengan median.

Tabel 4. 4. Hasil dari median

| Fitur | Jumlah Missing Value | Hasil Median | |
|---|---|---|---|
| | | Negatif | Positif |
| Insulin | 48,13% | 94,5 | 169,5 |
| Glukosa | 0,75% | 108 | 140 |
| Lemak Tubuh | 29,48% | 27 | 32 |
| Tekanan Darah | 4,66% | 72 | 74,5 |
| BMI | 0,56% | 28,5 | 34,3 |

Pada Tabel 4.4. diatas, didapatkan hasil dari penggantian *missing value* dengan menggunakan nilai dari median berdasarkan jumlah dari fitur Hasil. Dapat dilihat bahwa hasil dari penggantian tersebut terdiri atas bagian positif dan negatif. Pada fitur bagian Insulin, didapatkan nilai median dari Hasil Negatif sejumlah 94,5 IU dan Hasil Positif berupa 169,5 IU.



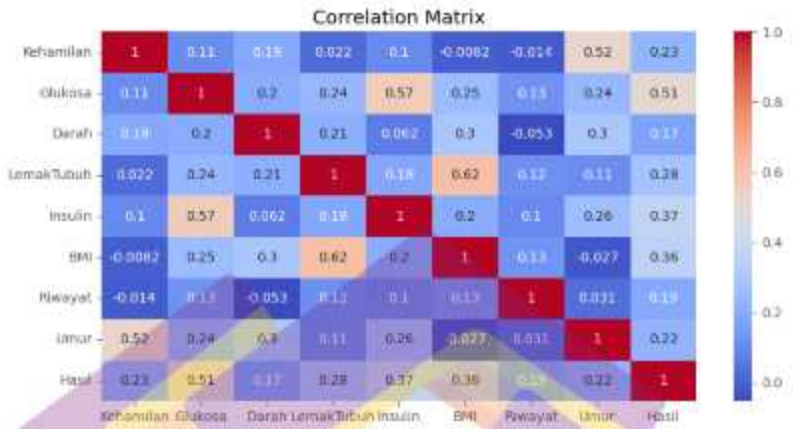
Gambar 4. 4. Pengimplentasian Missing Value

Pada Gambar 4.4. diatas merupakan hasil bahwa dari keseluruhan fitur tidak ada lagi kekosongan dari data, yang artinya seluruh data telah terisi dengan seimbang.

Menurut penelitian yang dilakukan oleh Kaur et al. (2020) dan Mollick et al. (2022), penggantian missing values dengan nilai median lebih disarankan dibandingkan dengan nilai mean karena median lebih tahan terhadap outlier. Outlier dapat secara signifikan mempengaruhi nilai mean, sehingga dapat menggeser nilai rata-rata dan menciptakan bias dalam data. Sebaliknya, median adalah nilai tengah yang tidak dipengaruhi oleh outlier, sehingga memberikan hasil yang lebih akurat dari data aslinya ketika distribusi data tidak normal. Penelitian juga menunjukkan bahwa penggunaan median dalam imputasi dapat meningkatkan robusta dari model prediksi. Hassan et al. (2022) menyatakan bahwa model yang menggunakan median untuk imputasi missing values cenderung memiliki kinerja yang lebih konsisten, terutama dalam dataset dengan *outlier* yang signifikan. Teknik ini mengurangi risiko *overfitting* yang bisa terjadi jika *outliers* mempengaruhi data latih.

### 4.4.6. Correlation Matrix

*Correlation Matrix* atau matriks korelasi menunjukkan bahwa terdapat beberapa hubungan yang kuat antara variabel-variabel yang diukur, Gambar 4.5. dibawah ini menunjukkan hasil dari korelasi dataset yang digunakan.




Gambar 4. 5. Correlation Matrix

Pada gambar 4.5. diatas terdapat korelasi positif antara glukosa dan insulin (0.57), serta antara lemak tubuh dan BMI (0.62), yang mengindikasikan bahwa kadar glukosa tinggi cenderung berhubungan dengan kadar insulin tinggi, dan peningkatan persentase lemak tubuh berkaitan erat dengan peningkatan BMI. Selain itu, korelasi antara glukosa dan hasil (0.51) serta antara BMI dan hasil (0.36) menunjukkan bahwa kadar glukosa dan BMI yang lebih tinggi mungkin berhubungan dengan hasil yang lebih baik dalam penelitian ini. Korelasi positif antara kehamilan dan umur (0.52) juga menunjukkan bahwa semakin tinggi umur seseorang, jumlah kehamilan cenderung lebih banyak.

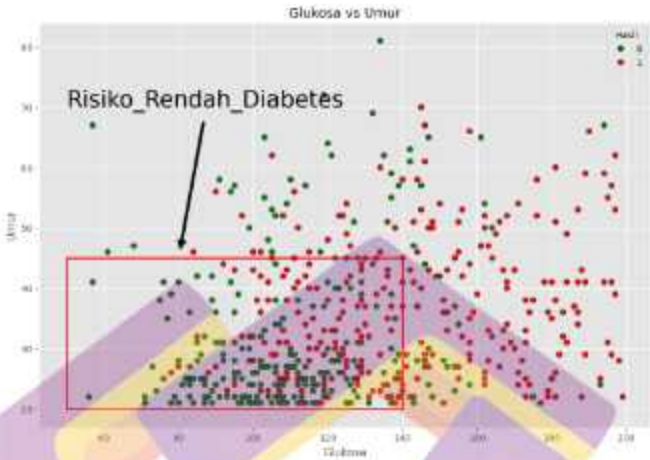
### 4.4.7. Feature Engineering

Pada Preprocessing data selanjutnya yaitu *Feature Engineering* merupakan proses menciptakan fitur-fitur baru dari data yang sudah ada untuk meningkatkan kinerja model prediksi. Fitur-fitur baru tersebut

diharapkan lebih relevan dan bermakna dalam mendeteksi diabetes daripada fitur aslinya. Dalam penelitian ini, 14 fitur baru akan ditambahkan ke dataset *Pima Indians Diabetes* berdasarkan kombinasi variabel kesehatan yang ada.

#### 4.4.7.1.Risiko Rendah Diabetes

Fitur risiko rendah diabetes mengidentifikasi individu dengan kadar glukosa darah kurang dari 140 mg/dL dan usia di bawah 45 tahun sebagai kelompok dengan risiko rendah untuk diabetes tipe 2. Menurut *American Diabetes Association*, kadar glukosa darah normal setelah makan adalah kurang dari 140 mg/dL, menunjukkan kemampuan tubuh untuk mengatur glukosa dengan baik. Risiko diabetes tipe 2 meningkat seiring bertambahnya usia, terutama setelah 45 tahun. Dengan menggabungkan kedua kondisi ini, dapat mengidentifikasi seseorang yang berisiko rendah, membantu fokus pada upaya pencegahan dan intervensi pada mereka yang berisiko lebih tinggi. Pengendalian glukosa dan gaya hidup sehat pada usia muda mengurangi risiko diabetes tipe 2 di kemudian hari.




Gambar 4. 6. Risiko Rendah Diabetes

Pada Gambar 4.6. diatas merupakan sebuah *scatter plot* yang menampilkan hubungan antar glukosa dan umur. Titik berwarna hijau menampilkan seseorang dengan hasil negatif diabetes, dan titik berwarna merah menampilkan seseorang dengan hasil positif diabetes. Kotak merah di dalam plot menandai area di mana didalam kelompok memiliki kadar glukosa kurang dari 140 mg/dL dan usia di bawah 45 tahun, yang diidentifikasi sebagai kelompok dengan risiko rendah diabetes. Plot ini membantu memvisualisasikan bahwa sebagian besar individu dalam kotak merah (dengan kadar glukosa rendah dan usia muda) tidak menderita diabetes, mendukung kriteria risiko rendah diabetes yang digunakan dalam *feature engineering.*

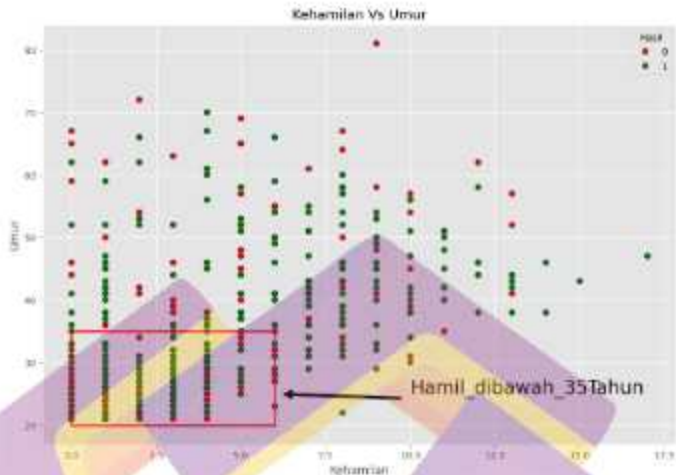
**4.4.7.2.Berat Badan Ideal**

Fitur ini mengidentifikasi seseorang dengan indeks massa tubuh (BMI) kurang dari 30 kg/m² sebagai kelompok dengan berat badan ideal dan

risiko lebih rendah untuk diabetes tipe 2. BMI di bawah 30 kg/m² dikategorikan sebagai berat badan normal atau kelebihan berat badan ringan, yang tidak dikaitkan dengan risiko tinggi diabetes. Dengan BMI 30 kg/m² atau lebih meningkatkan risiko diabetes dan komplikasi terkait.

**4.4.7.3.Kehamilan berdasarkan Umur**

Fitur ini mengidentifikasi wanita yang mengalami kehamilan pada usia di bawah 35 tahun sebagai kelompok dengan risiko lebih rendah untuk komplikasi terkait kehamilan dan diabetes gestasional. Kehamilan pada usia lebih muda, khususnya di bawah 35 tahun, cenderung memiliki hasil kesehatan yang lebih baik dibandingkan dengan kehamilan pada usia yang lebih tua. Usia kehamilan yang lebih muda dikaitkan dengan penurunan risiko komplikasi metabolik dan kesehatan jangka panjang yang lebih baik. Implementasi fitur ini dalam model machine learning membantu meningkatkan akurasi prediksi risiko kesehatan selama kehamilan dan diabetes gestasional. Wanita yang hamil pada usia di bawah 35 tahun memiliki risiko lebih rendah untuk mengembangkan diabetes gestasional dan komplikasi terkait dibandingkan dengan mereka yang hamil pada usia lebih tua.



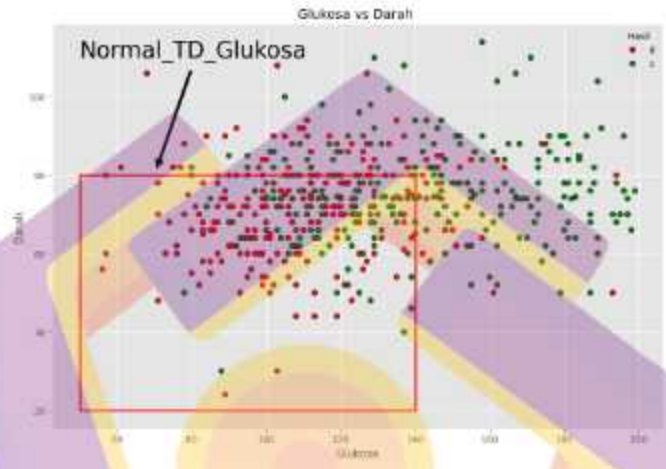

Gambar 4. 7. Kehamilan Berdasaran Umur

Pada Gambar 4.7. diatas merupakan sebuah *scatter plot* yang menampilkan hubungan antar jumlah kehamilan dan umur. Pada gambar diatas terdapat titik hijau yang artinya negatif dan titik merah artinya positif.

**4.4.7.4.Kadar Glukosa dan Tekanan Darah Normal**

Fitur ini dapat ditambahkan dengan mengidentifikasi seseorang dengan tekanan darah diastolik kurang dari 80 mmHg dan kadar glukosa darah kurang dari 140 mg/dL sebagai kelompok dengan risiko lebih rendah untuk mengembangkan diabetes tipe 2 dan komplikasi kardiovaskular. Tekanan darah yang terkendali menunjukkan risiko yang lebih rendah untuk komplikasi kardiovaskular, sedangkan kadar glukosa yang terkendali menunjukkan kemampuan tubuh yang baik dalam mengatur glukosa, mengindikasikan risiko rendah untuk diabetes. Implementasi fitur ini dalam model machine learning membantu mengidentifikasi pola dalam data dan

meningkatkan akurasi prediksi. Penelitian dari Larasati (2023) mendukung penggunaan parameter ini sebagai indikator penting dalam evaluasi risiko kesehatan.




Gambar 4. 8. Glukosa dan Tekanan Darah normal

Pada Gambar 4.8. diatas merupakan sebuah *scatter plot* yang menampilkan hubungan antar glukosa dan tekanan darah diastolik dengan batas normal. Pada gambar diatas terdapat titik hijau yang artinya negatif dan titik merah artinya positif.

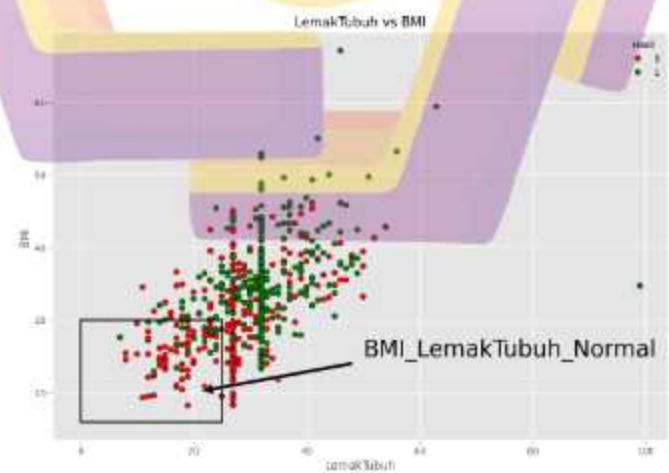
**4.4.7.5.Lemak Tubuh Normal: Lemak Tubuh < 25%**

Fitur ini mengidentifikasi individu dengan persentase lemak tubuh kurang dari 25% sebagai kelompok dengan risiko lebih rendah untuk mengembangkan diabetes tipe 2. Persentase lemak tubuh yang lebih rendah menunjukkan distribusi lemak yang lebih sehat, yang berhubungan dengan penurunan risiko resistensi *insulin* dan komplikasi *metabolik* (Li et al., 2022). Implementasi fitur ini dalam model *machine learning* membantu

meningkatkan akurasi prediksi risiko diabetes dengan mengidentifikasi pola distribusi lemak tubuh yang sehat.

#### 4.4.7.6.BMI_LemakTubuh_Normal

Menggabungkan BMI < 30 kg/m² dan persentase lemak tubuh < 25% dalam satu fitur memberikan gambaran yang lebih luas tentang kesehatan seseorang terkait risiko diabetes tipe 2. BMI yang lebih rendah dari 30 kg/m² menunjukkan berat badan yang ideal atau kelebihan berat badan ringan, sementara persentase lemak tubuh kurang dari 25% menunjukkan distribusi lemak tubuh yang sehat. Kombinasi kedua indikator ini membantu model prediksi mengidentifikasi individu dengan metabolisme yang baik dan risiko lebih rendah untuk diabetes. BMI dan persentase lemak tubuh yang rendah memiliki risiko lebih rendah untuk diabetes dan komplikasi kardiovaskular, menjadikan fitur gabungan ini penting untuk meningkatkan akurasi dan efektivitas model prediksi diabetes.

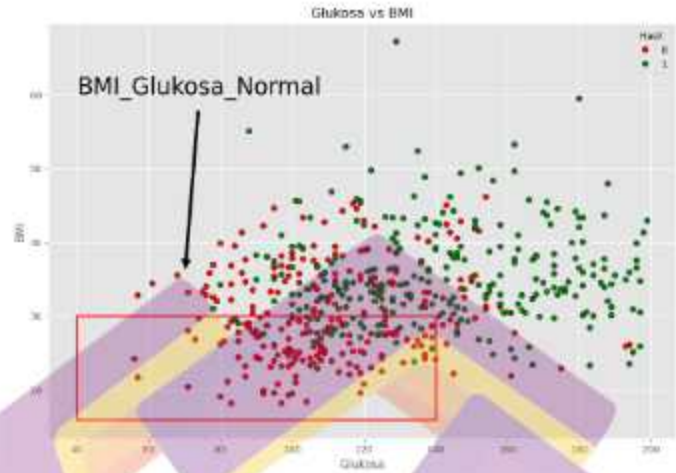



Gambar 4. 9. BMI dan Lemak Tubuh Normal

Pada Gambar 4.9. diatas merupakan sebuah *scatter plot* yang menampilkan hubungan antar BMI dan Lemak Tubuh dengan batas normal. Pada gambar diatas terdapat titik hijau yang artinya negatif dan titik merah artinya positif. Terdapat beberapa persebaran data yang telah menjadi 1 kelompok.

**4.4.7.7.BMI_Glukosa_Normal**

Fitur ini mengidentifikasi individu dengan BMI kurang dari 30 kg/m² dan kadar glukosa darah kurang dari 140 mg/dL sebagai kelompok dengan risiko lebih rendah untuk diabetes tipe 2. BMI rendah menunjukkan berat badan ideal atau kelebihan berat badan ringan, sedangkan glukosa terkontrol menunjukkan metabolisme glukosa yang baik. Menggabungkan kedua fitur ini memberikan penilaian risiko yang lebih luas dan meningkatkan akurasi model prediksi. Kadar glukosa darah yang terkendali mengurangi risiko komplikasi *kardiovaskular* dan diabetes tipe 2 (Loeffelholz, 2024).




Gambar 4. 10. BMI dan Glukosa Normal

Pada Gambar 4.10. diatas merupakan sebuah *scatter plot* yang menampilkan hubungan antar BMI dan Kadar Glukosa dengan batas normal. Pada gambar diatas terdapat titik hijau yang artinya negatif dan titik merah artinya positif. Terdapat beberapa persebaran data yang telah menjadi 1 kelompok.

**4.4.7.8.Insulin Normal**

Mengidentifikasi seseorang dengan kadar insulin kurang dari 150 µU/mL menjadi fitur baru akan menjadikan sebagai kelompok dengan risiko lebih rendah untuk mengembangkan diabetes tipe 2.

**4.4.7.9.Tekanan Darah Normal**

Fitur ini mengidentifikasi seseorang dengan tekanan darah diastolik kurang dari 80 mmHg sebagai kelompok dengan risiko lebih rendah untuk mengembangkan komplikasi kardiovaskular dan diabetes tipe 2. Tekanan

darah diastolik yang normal menunjukkan bahwa jantung dan pembuluh darah berfungsi dengan baik, mengurangi beban pada sistem kardiovaskular.

**4.4.7.10. Kehamilan kurang dari 4 kali**

Fitur ini mengidentifikasi wanita yang pernah hamil kurang dari 4 kali sebagai kelompok dengan risiko lebih rendah untuk mengembangkan diabetes tipe 2. Penelitian menunjukkan bahwa jumlah kehamilan dapat mempengaruhi metabolisme dan risiko resistensi insulin, dengan kehamilan yang lebih sering meningkatkan beban metabolik pada tubuh. Wanita yang hamil lebih dari 4 kali memiliki risiko lebih tinggi untuk mengembangkan diabetes gestasional, yang dapat meningkatkan risiko diabetes tipe 2 di kemudian hari. wanita dengan kehamilan kurang dari 4 kali cenderung memiliki risiko metabolik yang lebih rendah.

**4.4.7.11. Rasio dari fitur dataset**

Untuk meningkatkan akurasi prediksi diabetes, dibuat dataset baru yang mencakup rasio-rasio tertentu dari variabel-variabel dalam dataset Pima Indians. Rasio-rasio ini dipilih untuk menangkap hubungan yang lebih kompleks antara variabel dan memberikan wawasan yang lebih mendalam tentang faktor-faktor risiko diabetes. Berikut adalah rasio-rasio yang ditambahkan.

1. Rasio Usia dengan Kehamilan:

   Rasio ini dibuat untuk memahami bagaimana pola kehamilan dalam berbagai kelompok umur dapat mempengaruhi risiko diabetes. Usia merupakan faktor penting dalam kehamilan, dan rasio ini dapat

membantu mengidentifikasi hubungan antara usia dan jumlah kehamilan.

2. Rasio Glukosa Terhadap Riwayat Diabetes:

   Rasio ini membantu memahami hubungan antara kadar glukosa darah dan riwayat diabetes dalam keluarga. Dengan mengetahui rasio ini, dapat diidentifikasi seberapa besar pengaruh faktor genetik terhadap kadar glukosa darah, yang merupakan indikator penting dalam prediksi diabetes.

3. Rasio Usia Terhadap Riwayat Diabetes:

   Rasio ini menunjukkan bagaimana usia berhubungan dengan riwayat diabetes dalam keluarga. Hubungan ini penting untuk mengidentifikasi kelompok umur yang lebih rentan terhadap diabetes berdasarkan faktor genetik.

4. Rasio Usia dengan Insulin:

   Rasio ini menggambarkan bagaimana tingkat insulin berubah seiring bertambahnya usia. Perubahan ini penting untuk memahami resistensi insulin dan risiko diabetes pada berbagai kelompok umur.sio yang dibuat dan alasan penggunaannya:

Berikut ini penambahan fitur-fitur baru berdasarkan korelasinya:

Tabel 4. 5. Penambahan Fitur Baru

| No. | Nama Fitur | Deskripsi |
|---|---|---|
| 1. | Risiko Rendah Diabetes | Menandai pasien dengan risiko rendah diabetes berdasarkan kadar glukosa (< 140 mg/dL) dan umur (<45 tahun). |
| 2. | Berat Badan Ideal | Menandai pasien yang memiliki berat badan kurang (< 30kg) atau kelebihan berat badan berdasarkan BMI. |

Tabel 4.5. (Lanjutan)

| | | |
|---|---|---|
| 3. | Kehamilan Berdasarkan Umur | Menandai pasien yang pernah hamil di usia muda (< 35 tahun). |
| 4. | Glukosa dan Tekanan Darah Normal | Menandai pasien dengan kadar glukosa (< 140 mg/dL) dan tekanan darah (< 80 mmHg). |
| 5. | Lemak Tubuh Normal | Menandai pasien dengan persentase lemak tubuh (< 25%). |
| 6. | BMI_LemakTubuh_Normal | Menandai pasien dengan BMI (< 30) dan persentase lemak tubuh (< 25%). |
| 7. | BMI dan Glukosa Normal | Menandai pasien dengan BMI (< 30) dan kadar glukosa (< 140) |
| 8. | Kadar Insulin Normal | Menandai pasien dengan kadar insulin < 150 |
| 9. | Tekanan Darah Normal | Menandai pasien dengan tekanan darah (< 80 mmHg). |
| 10. | Kehamilan kurang dari 4 kali | Menandai pasien yang pernah hamil/melahirkan kurang dari 4 kali |
| 11. | Rasio_Usia_dengan_Kehamilan | Rasio usia pasien terhadap jumlah kehamilan |
| 12. | Rasio Glukosa Terhadap Riwayat Diabetes | Rasio kadar glukosa terhadap riwayat diabetes pasien |
| 13. | Rasio Usia Terhadap Riwayat Diabetes | Rasio Usia pasien terhadapt riwayat diabetes dalam keluarga. |
| 14. | Rasio Usia dengan Insulin | Rasio usia pasien terhadap kadar Insulin. |

Pada Tabel 4.5. diatas menunjukkan bahwa fitur yang ada pada dataset berhasil ditambahkan sesuai dengan kedekatan fungsi masing-masing fitur, total keseluruhan fitur tambahan berjumlah 14 fitur, dari penambahan ini memberikan gambaran yang lebih mendalam tentang risiko diabetes pada pasien, sehingga model prediksi dapat lebih memahami pola hubungan antara berbagai variabel kesehatan dan hasil diagnosa diabetes.

Pada Tabel 4.6. di bawah ini menggambarkan distribusi hasil dari 14 fitur baru yang ditambahkan ke dataset Pima Indians Diabetes melalui proses feature engineering. Fitur-fitur baru ini diharapkan dapat memberikan informasi yang lebih mendalam dan akurat terkait risiko diabetes pada pasien. Distribusi hasil mencakup jumlah pasien dengan nilai

(Ya) dan (Tidak) untuk setiap fitur, disertakan hasil jumlah pasien positif dan negatif disetiap label.

Tabel 4. 6. Distribusi 14 Fitur Baru

| No. | Nama Fitur | Jumlah Distribusi Hasil Fitur Baru | |
|---|---|---|---|
| | | Ya | Tidak |
| 1. | Risiko Rendah Diabetes | **323** (115 Positif, 208 Negatif) | **213** (153 Positif, 60 Negatif) |
| 2. | Berat Badan Ideal | **194** (51 Positif, 143 Negatif) | **342** (217 Positif, 125 Negatif) |
| 3. | Kehamilan Berdasarkan Umur | **260** (86 Positif, 174 Negatif) | **276** (182 Positif, 94 Negatif) |
| 4. | Glukosa dan Tekanan Darah Normal | **302** (103 Positif, 199 Negatif) | **234** (165 Positf, 69 Negatif) |
| 5. | Lemak Tubuh Normal | **126** (38 Positif, 88 Negatif) | **410** (230 Positif, 180 Negatif) |
| 6. | BMI_Lemak Tubuh_Normal | **84** (18 Positif, 66 Negatif) | **474** (257 Positif, 217 Negatif) |
| 7. | BMI dan Glukosa Normal | **163** (31 Positif, 132 Negatif) | **373** (237 Positif, 136 Negatif) |
| 8. | Kadar Insulin Normal | **280** (49 Positif, 231 Negatif) | **256** (219 Positif, 37 Negatif) |
| 9. | Tekanan Darah Normal | **374** (178 Positif, 196 Negatif) | **162** (90 Positif, 72 Negatif) |
| 10. | Kehamilan kurang dari 4 kali | **208** (75 Positif, 133 Negatif) | **328** (193 Positif, 135 Negatif) |
| 11. | Rasio Usia dengan Kehamilan | *Rasio Angka* | *Rasio Angka* |
| 12. | Rasio Glukosa Terhadap Riwayat Diabetes | *Rasio Angka* | *Rasio Angka* |
| 13. | Rasio Usia Terhadap Riwayat Diabetes | *Rasio Angka* | *Rasio Angka* |
| 14. | Rasio Usia dengan Insulin | *Rasio Angka* | *Rasio Angka* |

Distribusi hasil dari 14 fitur baru melalui *feature engineering* memberikan gambaran lebih jelas tentang risiko diabetes pada pasien. Fitur-fitur baru ini mencakup berbagai aspek kesehatan seperti kadar glukosa,

tekanan darah, lemak tubuh, BMI, riwayat kehamilan, dan kadar insulin. Selain itu, fitur-fitur yang menghitung rasio antara variabel-variabel tersebut, seperti rasio BMI ke lemak tubuh atau usia terhadap riwayat diabetes, memberikan informasi lebih mendalam tentang hubungan antarvariabel dan risiko diabetes. Secara keseluruhan, fitur-fitur baru ini membantu model prediksi dalam mengidentifikasi risiko diabetes dengan akurat.

Tabel 4. 7. Dataset dengan Fitur Baru

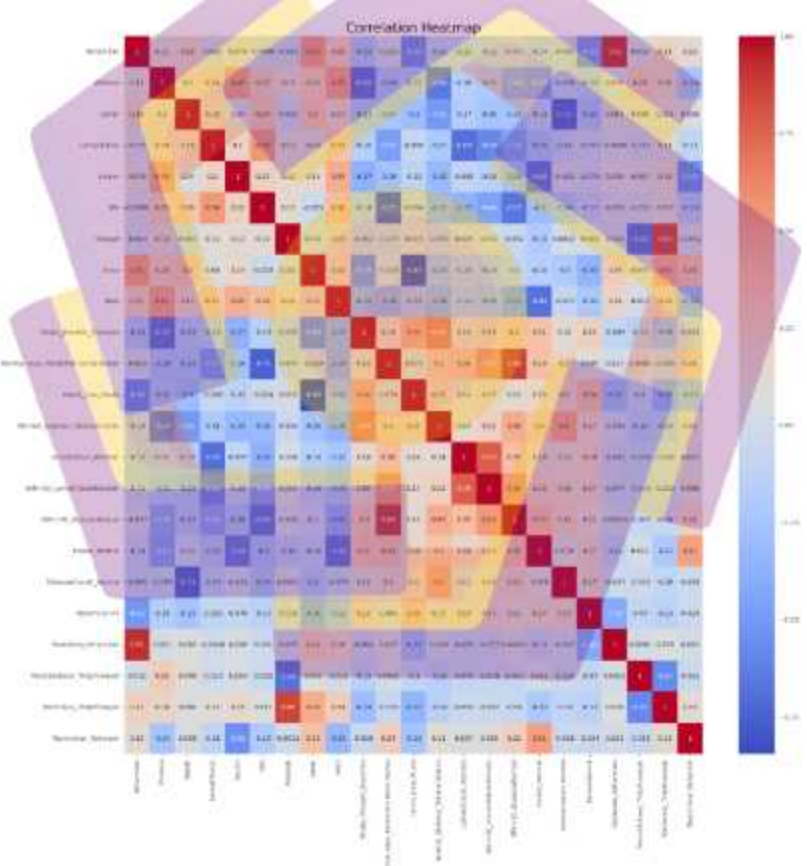
| Kehamilan | Glukosa | Darah | Lemak Tubuh | Insulin | BMI | Riwayat | Umur | Hasil | N1 | N2 | N3 | N4 | N5 | N6 | N7 | N8 | N9 | N10 | N11 | N12 | N13 | N14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4 | 110 | 76 | 20 | 100 | 28,4 | 0,118 | 27 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0,1481 | 932,20 | 3,1 | 0,27 |
| 5 | 114 | 74 | 27 | 94,5 | 24,9 | 0,744 | 57 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0,0877 | 153,22 | 42,4 | 0,60 |
| 4 | 112 | 78 | 40 | 94,5 | 39,4 | 0,236 | 38 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0,1052 | 474,57 | 8,9 | 0,402 |
| 0 | 101 | 65 | 28 | 94,5 | 24,6 | 0,237 | 22 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 426,16 | 5,12 | 0,232 |
| 0 | 128 | 68 | 19 | 180 | 30,5 | 1,391 | 25 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 92,02 | 34,775 | 0,138 |
| 2 | 124 | 68 | 28 | 205 | 32,9 | 0,875 | 30 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0,066 | 141,71 | 26,25 | 0,146 |
| 2 | 155 | 74 | 17 | 96 | 26,6 | 0,433 | 27 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 0,074 | 357,96 | 11,691 | 0,281 |
| 7 | 109 | 80 | 31 | 169,5 | 35,9 | 1,127 | 43 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0,162 | 96,71 | 48,461 | 0,253 |
| 3 | 182 | 74 | 32 | 169,5 | 30,5 | 0,345 | 29 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0,103 | 527,53 | 10,005 | 0,171 |
| 3 | 112 | 74 | 30 | 169,5 | 31,6 | 0,197 | 25 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0,12 | 568,52 | 4,925 | 0,147 |
| 0 | 124 | 70 | 20 | 169,5 | 27,4 | 0,254 | 36 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 488,1 | 9,144 | 0,212 |
| 1 | 90 | 62 | 12 | 43 | 27,2 | 0,58 | 24 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0,04 | 155,17 | 13,92 | 0,55 |
| 0 | 125 | 68 | 27 | 94,5 | 24,7 | 0,206 | 21 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 606,79 | 4,326 | 0,22 |
| 3 | 106 | 54 | 21 | 158 | 30,9 | 0,292 | 24 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0,125 | 363 | 7,008 | 0,151 |
| 1 | 116 | 70 | 28 | 94,5 | 27,4 | 0,204 | 21 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0,04 | 568,6 | 4,284 | 0,22 |
| 2 | 121 | 70 | 32 | 95 | 39,1 | 0,886 | 23 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0,08 | 136,56 | 20,378 | 0,24 |

Catatan :

- Variabel dari Fitur Hasil terdapat 2 jenis meliputi :
  - (0) untuk Negatif
  - (1) untuk Positif
- Nilai N1-N16 disesuaikan dengan nama Fitur-fitur yang ada pada Tabel 4.4.

### 4.4.8. Correlation Matrix - Fitur Baru

Pada bagian ini merupakan matriks korelasi yang dihasilkan dari penambahan fitur baru pada dataset penelitian ini. Matriks korelasi ini mencakup variabel-variabel tambahan yang lebih spesifik untuk mendapatkan pemahaman yang lebih mendalam tentang interaksi antar variabel dalam konteks kesehatan dan risiko diabetes.




Gambar 4. 11. Correlation Matrix Fitur Baru

Pada Gambar 4.11. diatas merupakan matriks korelasi dari fitur baru, dari tahapan ini yaitu berfungsi untuk memperkuat temuan sebelumnya dengan menambahkan fitur baru pada analisis. Keterkaitan yang kuat antara status berat badan dan BMI, serta antara hamil usia muda dan umur, memberikan informasi tambahan mengenai faktor-faktor yang mempengaruhi kesehatan individu. Korelasi antara variabel seperti BMI < 30 dengan lemak tubuh rendah dan glukosa normal dengan tekanan darah normal juga menyoroti pentingnya menjaga berat badan dan kadar glukosa dalam rentang normal untuk kesehatan yang optimal.

### 4.4.9. Standard Scaler

Tahap Preprocessing data selanjutnya adalah, *Standard Scaler* digunakan untuk melakukan standarisasi fitur-fitur dari dataset. Standarisasi dilakukan dengan cara mengubah nilai setiap fitur agar memiliki distribusi dengan rata-rata 0 dan standar deviasi 1.

Berikut adalah script untuk Standardscaler data:

```
target_col = ["Hasil"]
cat_cols   = data.nunique()[data.nunique() < 12].keys().tolist()
cat_cols   = [x for x in cat_cols ]

#numerical columns
num_cols     = [x for x in data.columns if x not in cat_cols +
target_col]
#Binary columns with 2 values
bin_cols   = data.nunique()[data.nunique() == 2].keys().tolist()

#Columns more than 2 values
multi_cols = [i for i in cat_cols if i not in bin_cols]

#Scaling Numerical columns
std = StandardScaler()
scaled = std.fit_transform(data[num_cols])
scaled = pd.DataFrame(scaled,columns=num_cols)
```

Kode tersebut bertujuan untuk mengidentifikasi berbagai jenis kolom dalam dataset, termasuk kolom target, kategori, biner, multi-kategori, dan numerik. Kolom kategori diidentifikasi sebagai kolom dengan kurang dari 12 nilai, sementara kolom biner adalah kolom yang memiliki tepat dua nilai unik. Setelah mengidentifikasi kolom-kolom tersebut, kode ini menggunakan StandardScaler untuk melakukan standarisasi pada kolom numerik, sehingga semua kolom numerik memiliki rata-rata 0 dan standar deviasi 1. Berikut pada Tabel 4.8. dibawah ini menampilkan hasil dari dataset yang telah di standarisasi.

Tabel 4. 8. Dataset Standarisasi Data

| Hasil | N1 | N2 | N3 | N4 | N5 | N6 | N7 | N8 | N9 | N10 | Kehmin | Glukosa | Darah | Lemak Tubuh | Insulin | BMI | Riwayat | Umur | N11 | N12 | N13 | N14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | -0.016 | -0.536 | 0.229 | -1.945 | -0.534 | -0.656 | -1.89 | -0.616 | 0.445 | 20.68 | -1.01 | -0.14 |
| **0** | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0.269 | -0.407 | 0.065 | -0.316 | -0.596 | -1.1638 | 0.765 | 1.886 | -0.28 | -0.87 | 18.98 | 1.5 |
| **0** | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | -0.876 | 0.533 | 0.06 | -0.872 | -0.97 | -0.975 | -0.68 | -11.191 | -0.76 | 0.655 | -0.85 | 0.22 |
| **1** | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 2.275 | -1.120 | -0.91 | -2.540 | 1.231 | -0.772 | 1.304 | 0.8025 | 1.943 | -1.08 | 1.7 | -0.63 |
| **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1.989 | 0.923 | 0.229 | -0.204 | 0.024 | 0.0541 | 0.256 | 1.387 | 12.57 | -10.2 | 0.38 | 0.20 |
| **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | -0.016 | -0.115 | -0.917 | 0.239 | 0.242 | -0.1343 | -0.769 | 0.0505 | 0.038 | 0.60 | -0.66 | -0.45 |
| **1** | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | -0.589 | 0.631 | 0.106 | 0.239 | 0.242 | -0.786 | -0.728 | -0.534 | -0.47 | 0.84 | -0.75 | -0.66 |
| **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 17.025 | -0.828 | 1.048 | 0.795 | 0.242 | 1.837 | 1.926 | 0.301 | 1.828 | -11.2 | 9.54 | -0.37 |
| **0** | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | -0.589 | -14.452 | -1.7368 | -0.87 | -0.367 | -0.6418 | 0.359 | -0.7849 | -0.37 | -12 | 1.9 | -0.40 |
| **0** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0.269 | -0.763 | 2.851 | 0.795 | -0.596 | 0.9097 | -0.535 | 2.557 | -0.41 | -0.18 | 0.28 | 1.922 |
| **1** | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0.556 | -0.731 | 0.065 | -1.370 | 0.091 | -0.438 | 0.699 | 0.551 | 0.42 | -0.91 | 0.947 | -0.18 |
| **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0.556 | -0.796 | 0.721 | 0.239 | 0.242 | -0.308 | -0.905 | 0.134 | 0.66 | 0.6 | -0.76 | -0.43 |

### 4.4.10. Pembagian Data Pelatihan dan Data Pengujian (Split Data)

Pada tahap ini, akan membahas proses pembagian dataset menjadi dua bagian utama: set pelatihan dan set pengujian. Set pelatihan (data latih) digunakan untuk melatih model pembelajaran mesin, sementara set pengujian (data uji) berfungsi untuk mengevaluasi kinerja model tersebut terhadap data baru yang belum pernah dilihat oleh model selama *training*.

Berikut kode untuk melakukan *split data* :

```
from sklearn.model_selection import train_test_split

X = data.drop('Hasil', axis=1)  # Fitur untuk pelatihan
y = data['Hasil']  # Label/target

# Membagi data menjadi training dan testing set
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.3, random_state=42)
```

Dari kode diatas, dilakukan sebuah pembagian data menggunakan fungsi train_test_split dari library scikit-learn, dengan parameter test_size=0.3 yang berarti 30% dari data akan digunakan sebagai data uji dan sisanya 70% sebagai data latih. Parameter random_state=42 digunakan untuk memastikan bahwa hasil pembagian data konsisten antar berbagai eksperimen. Proses ini penting untuk memvalidasi keefektifan model dalam kondisi yang tidak akurat, memastikan bahwa model memiliki kemampuan mendeteksi yang baik dan tidak hanya terpaku dengan data latih.

## 4.5. Tahap Membangun Model

Pada tahap ini akan menjabarkan implementasi pada tahap pembangunan model KNN dan LightGBM. Dalam penelitian ini, peneliti akan menggunakan memiliki beberapa tahapan dalam melakukan pembuatan model yang dijalankan. Tahap pertama yaitu, membuat algoritma KNN dan LightGBM sebelum dioptimasi secara terpisah, kemudian tahapan selanjutnya yaitu membuat algoritma secara dioptimasi menggunakan teknik *GridSearch* kemudian melakukan perbandingan dan skenario *Ensemble Learning* (Penggabungan Algoritma),berikut pemodelan yang akan dilakukan:

### 4.5.1. Pemodelan Algoritma Sebelum Optimasi

Dalam tahapan awal ini, yaitu tahapan untuk mengevaluasi performa dasar dari kedua algoritma yang akan digunakan, pada tahap ini akan mengimplementasikan awal dari kedua algoritma tersebut dengan menggunakan parameter-parameter *default* dari kedua algoritma ini. Tujuan dari pendekatan ini adalah untuk membangun sebuah pemahaman mendasar mengenai bagaimana masing-masing algoritma menangani dataset diabetes Pima Indians tanpa melibatkan optimasi lanjutan.

#### 4.5.1.1. K-Nearest Neighbors (KNN)

Dalam tahap ini, proses KNN dikonfigurasikan dengan parameter default sederhana, yakni jumlah tetangga terdekat *('n_neighbors')* sebanyak 5. Tujuan utama dari pengaturan ini untuk menilai bagaimana performa awal algoritma sebelum melakukan tuning lebih lanjut.

Berikut kode dalam pengimplementasian KNN sebelum dioptimasi:

```
# Membuat dan melatih model KNN
knn = KNeighborsClassifier(n_neighbors=5)
knn.fit(X_train, y_train)
```

Pada kode diatas merupakan inisialisasi dari model KNN dengan nilai k secara default tanpa modifikasi dengan parameter lainnya, dengan tujuan melihat bagaimana model tersebut mengklasifikasikan data dengan konfigurasi paling dasar. Setelah dilakukan inisialisasi dilanjutkan ketahap berikutnya yaitu mengevaluasi performa KNN dengan *classification report*.

Tabel 4. 9. Evaluasi Performa KNN

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| **0** | 0.88 | 0.83 | 0.85 | 76 |
| **1** | 0.85 | 0.89 | 0.87 | 85 |
| | | | | |
| **Accuracy** | | | **0.8634** | 161 |
| Macro Avg | 0.86 | 0.86 | 0.86 | 161 |
| Wighted Avg | 0.86 | 0.86 | 0.86 | 161 |

Pada Tabel 4.9. diatas, meruapakan tabel *classification report* atau evaluasi performa dari KNN sebelum dioptimasi, dengan hasil awal menunjukkan akurasi sebesar 86,34%, dari total keseluruhan data uji 161 data, dengan 76 data pasien negatif dan 85 pasien positif, hasil nilai presisi pada kelas 0 dan 1 tercatat 0,88 dan 0,85 dan lainnya dapat dilihat pada tabel diatas, maka dari itu dapat menandakan bahwa model sudah cukup efektif namun masih memiliki ruang untuk peningkatan dengan penambahan sebuah teknik optimasi.

#### 4.5.1.2. LightGBM

LightGBM atau *Light Gradient Boosting Machine*, dilakukan dengan parameter default tanpa optimasi tambahan. Pada tahap ini model diinisasi menggunakan data latih dengan konfigurasi default, dengan tujuan memberikan baseline atau refrensi kinerja model sebelum dilakukan tuning lebih lanjut.

Berikut kode implementasi LightGBM:

```
# Membuat dan melatih model LightGBM
lgbm = lgb.LGBMClassifier()
lgbm.fit(X_train, y_train)
```

Pada kode diatas merupakan implementasi dari LightGBM, sama halnya dengan KNN kode tersebut dibuat untuk membandingkan hasil dengan yang telah dioptimasi.

Tabel 4. 10. Evaluasi Performa LightGBM

| | **Precision** | **Recall** | **F1-Score** | **Support** |
|---|---|---|---|---|
| **0** | 0.84 | 0.92 | 0.88 | 76 |
| **1** | 0.92 | 0.85 | 0.88 | 85 |
| | | | | |
| **Accuracy** | | | **0.8820** | 161 |
| Macro Avg | 0.88 | 0.88 | 0.88 | 161 |
| Wighted Avg | 0.88 | 0.88 | 0.88 | 161 |

Pada Tabel 4.10. diatas, merupakan evaluasi performa dari LightGBM dengan menunjukkan hasil akurasi sebesar 88,20%. Dari hasil diatas model sudah cukup efektif, namun masih memiliki ruang yang banyak untuk lebih mengoptimalkan lagi.

### 4.5.2. Pemodelan Setelah Optimasi

Pada tahap ini akan mengeksplorasi pemodelan KNN dan LightGBM setelah proses optimasi dilakukan. Optimasi ini bertujuan untuk meningkatkan efektivitas prediksi model menyesuaikan parameter pada tiap algoritma.

#### 4.5.2.1.Optimasi Model K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) adalah algoritma pembelajaran mesin berbasis *instance* yang mengklasifikasikan data baru berdasarkan kedekatannya dengan data yang sudah ada. KNN menghitung jarak antara data baru dengan data yang sudah ada menggunakan metrik tertentu, seperti jarak *Euclidean* atau *Manhattan*, kemudian memberikan label berdasarkan mayoritas tetangga terdekat. Berikut adalah langkah-langkah dalam membuat model KNN. Pada pemodelan KNN ini menggunakan sebuah fitur yaitu *Grid Search* untuk mencari parameter terbaik:

Berikut Kode dalam pembuatan KNN :

```
# Membuat model KNN
knn_grid = KNeighborsClassifier()

# Menentukan grid parameter
param_grid = {
    'n_neighbors': range(1, 50, 2),
    'weights': ['uniform', 'distance'],
    'metric':     ['euclidean',     'manhattan',     'minkowski',
'chebyshev'],
    'algorithm': ['auto', 'ball_tree', 'kd_tree', 'brute']
}

# Konfigurasi GridSearchCV
grid_search    =    GridSearchCV(knn_grid,    param_grid,    cv=5,
scoring='accuracy', verbose=1)
```

```
# Melakukan fitting GridSearchCV
grid_search.fit(X_train, y_train)

# Menampilkan parameter terbaik
print("Best parameters:", grid_search.best_params_)

# Menggunakan model terbaik yang ditemukan oleh GridSearchCV
best_knn = grid_search.best_estimator_

# Prediksi menggunakan model terbaik pada data pengujian
pred_knn = best_knn.predict(X_test)
```

Dari kode diatas dapat dirangkum yaitu parameter yang digunakan dalam *GridSearchCV* untuk menemukan kombinasi *hyperparameter* terbaik pada model KNN, pada Tabel 4.11. dibawah ini.

Tabel 4. 11. Kombinasi Parameter KNN

| No | Parameter | Deskripsi | Nilai |
|---|---|---|---|
| 1. | *'n_neighbors'* | Jumlah Nilai *k* yang dipertimbangkan | Range: 1, 3, ……..,49 |
| 2. | *'weights'* | Metode pembobotan tetangga terdekat | [*'uniform'*, *'distance'*] |
| 3. | *'metric'* | Jenis Metrik jarak yang digunakan | [*'euclidean'*, *'manhattan'*, *'minkowski'*, *'chebyshev'*] |
| 4. | *'algorithm'* | Algoritma pencarian tetangga | [*'auto'*, *'ball_tree'*, *'kd_tree'*, *'brute'*] |
| 5. | *'cv'* | Jumlah *fold* calam *croos-validation* | 5 |
| 6. | *'scoring'* | Metode evaluasi untuk memilih parameter terbaik | *'accuracy'* |
| 7. | *'verbose'* | Level keluaran untuk pemantauan | 1 |

Pada Tabel 4.11. diatas, dengan menggunakan *GridSearchCV* digunakan beberapa parameter untuk mengoptimalkan model KNN, *GridSearch* akan melakukan pencarian otomatis dengan menentukan dari beberapa kombinasi yang digunakan seperti *n_neighbors* (jumlah tetangga), *weights* (metode pembobotan), *metric* (jenis metrik jarak), dan

*algorithm* (algoritma pencarian). Proses ini menggunakan *cross-validation* (cv = 5) dan mengevaluasi hasilnya berdasarkan akurasi *(scoring = 'accuracy')*. Hasil dari *GridSearch* ditunjukkan pada tabel dibawah ini.

Tabel 4. 12. Parameter KNN

| Parameter | Nilai |
|---|---|
| *Nilai k* | 41 |
| *'weights'* | *['uniform']* |
| *'metric'* | *['manhattan']* |
| *'algorithm'* | *['auto']* |

Dari Tabel 4.12. diatas parameter terbaik yang ditemukan melalui *GridSearchCV* menunjukkan bahwa model KNN dengan menggunakan algoritma *'auto'* untuk pencarian tetangga, metrik jarak *'manhattan'*, 41 sebagai nilai k tetangga, dan metode pembobotan *'uniform'* menghasilkan performa terbaik.

Tabel 4. 13. Classification Report KNN

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| **0** | 0.92 | 0.89 | 0.91 | 76 |
| **1** | 0.91 | 0.93 | 0.92 | 85 |
| | | | | |
| **Accuracy** | | | **0.91** | 161 |
| Macro Avg | 0.91 | 0.91 | 0.91 | 161 |
| Wighted Avg | 0.91 | 0.91 | 0.91 | 161 |

Pada Tabel 4.13. diatas, merupakan hasil optimal dari model KNN yang dilakukan oleh *GridSearchCV* menunjukkan bahwa hasil akhir dari akurasi KNN yang didapatkan yaitu sebesar 91,30%. Dari model KNN diatas dapat menunjukkan bahwa KNN dapat memberikan

kinerja yang sangat baik dalam mengklasifikasikan pasien pada data diabetes.

#### 4.5.2.2.Optimasi Model LightGBM

Setelah mengoptimasi model KNN, selanjutnya yaitu mengoptimasi algoritma LightGBM. Berbeda dengan KNN, untuk LightGBM menggunakan sebuah proses optimasi dengan teknik *RandomsizeSearchCV* dengan mengeksplorasi ruang parameter secara acak.

Berikut Kode dari pembuatan model LightGBM:

```
# Membuat model dasar LightGBM
best_lgbm = LGBMClassifier()

# Menentukan grid parameter untuk RandomizedSearchCV
param_dist = {
    'n_estimators': np.arange(50, 300, 50),
    'learning_rate': [0.01, 0.05, 0.1, 0.2],
    'max_depth': np.arange(3, 11),
    'num_leaves': np.arange(20, 40),
    'colsample_bytree': [0.6, 0.8, 1.0],
    'subsample': [0.6, 0.8, 1.0]
}

# Konfigurasi RandomizedSearchCV
random_search = RandomizedSearchCV(lgbm, param_dist, n_iter=30,
cv=5, scoring='accuracy', verbose=1, random_state=42)

# Melakukan fitting RandomizedSearchCV
random_search.fit(X_train, y_train)

# Menampilkan parameter terbaik
print("Best parameters:", random_search.best_params_)

# Menggunakan model terbaik yang ditemukan oleh
RandomizedSearchCV
best_lgbm = random_search.best_estimator_

# Prediksi menggunakan model terbaik pada data pengujian
y_pred_lgbm = best_lgbm.predict(X_test)
```

Dari kode diatas dapat dirangkum yaitu parameter yang digunakan dalam *RandomSearch* untuk menemukan kombinasi *hyperparameter* terbaik pada model LightGBM, pada Tabel 4.14. dibawah ini.

Tabel 4. 14. Kombinasi Parameter LightGBM

| No | Parameter | Deskripsi | Nilai |
|---|---|---|---|
| 1. | *'n_estimators'* | Jumlah total pohon keputusan dalam boosting | Range: (50, 300, 50) |
| 2. | *'learning_rate'* | Kecepatan pembelajaran | (0.01, 0.05, 0.1, 02) |
| 3. | *'max_depth'* | Kedalaman maksimum pohon keputusan. | (3, 11) |
| 4. | *'num_leaves'* | Jumlah maksimum daun perpohon. | (20, 40) |
| 5. | *'colsample_bytree'* | Proporsi fitur yang dipilih secara acak | (0.6, 0.8, 1.0) |
| 6. | *'subsample'* | Proporsi sampel membangun pohon. | (0.6, 0.8, 1.0) |
| 7. | *'n_iter'* | Jumlah iterasi atau kombinasi acak. | 30 |

Pada Tabel 4.14. diatas, dengan menggunakan Randomsearch digunakan beberapa parameter untuk mengoptimalkan model LightGBM, Randomsearch akan melakukan pencarian otomatis dengan menentukan dari beberapa kombinasi yang digunakan seperti *'n_estimators'* (jumlah pohon keputusan), *learning rate* (kecepatan pembelajaran), *max depth* (kedalaman maksimum pohon keputusan), *colsample by tree* (proporsi fitur yang dipilih secara acak), dan *subsample* (proporsi sampel membangun pohon). Proses ini menggunakan *cross-validation* (cv = 5), menggunakan *n_iter* untuk mengkombinasi secara acak berjumlah 30 dan mengevaluasi hasilnya berdasarkan akurasi *(scoring = 'accuracy')*. Hasil dari *Randomsearch* ditunjukkan pada Tabel 4.15. dibawah ini.

Tabel 4. 15. Parameter LightGBM

| Parameter | Nilai |
|---|---|
| *'n_estimators'* | 250 |
| *'learning_rate'* | 0.01 |
| *'max_depth'* | 5 |
| *'num_leaves'* | 26 |
| *'colsample_bytree'* | 0.6 |
| *'subsample'* | 1.0 |
| *'n_iter'* | 30 |

Dari Tabel 4.15. diatas parameter terbaik yang ditemukan melalui *Randomsearch* menunjukkan bahwa model LightGBM dengan menggunakan parameter nya yang sudah disesuaikan untuk mentuning model LightGBM agar menghasilkan performa terbaik.

Tabel 4. 16. Classification Report LightGBM

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| **0** | 0.89 | 0.93 | 0.91 | 76 |
| **1** | 0.94 | 0.89 | 0.92 | 85 |
| | | | | |
| **Accuracy** | | | **0.9130** | 161 |
| Macro Avg | 0.91 | 0.91 | 0.91 | 161 |
| Wighted Avg | 0.91 | 0.91 | 0.91 | 161 |

Pada Tabel 4.16. diatas, merupakan hasil optimal dari model LightGBM yang dilakukan oleh Randomsize menunjukkan bahwa hasil akhir dari akurasi LightGBM yang didapatkan yaitu sebesar 91,30%. Dari model LightGBM diatas dapat menunjukkan bahwa LightGBM dapat memberikan kinerja yang sangat baik dalam mengklasifikasikan pasien pada data diabetes.

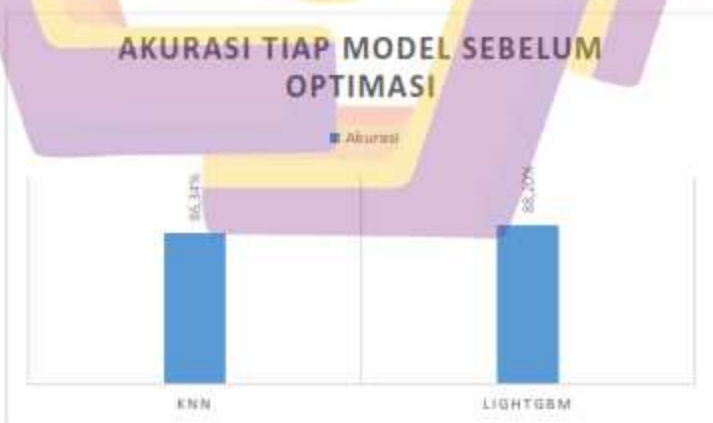
### 4.6. Penggabungan Algoritma KNN dan LightGBM dengan Stacking

Pada tahap ini menggunakan teknik *stacking* untuk menggabungkan algoritma KNN dan LightGBM. Pendekatan ini dipilih karena kedua algoritma

memiliki kekuatan *komplementer* dalam klasifikasi. KNN efektif dalam menangani data yang tidak linear dan sensitif terhadap data lokal, sementara LightGBM sangat efisien dalam menangani data besar dan memiliki kinerja yang baik dalam banyak kasus klasifikasi (Rufo et al., 2021). Dengan teknik stacking dapat meningkatkan performa prediksi dengan menggabungkan dari output model KNN dan LightGBM sebelum dan setelah di optimasi. Berikut pembahasan tiap bagian dari tahapan ini:

### 4.6.1. Stacking Sebelum Optimasi

Tahapan ini akan digunakan untuk menggabungkan prediksi dari kedua model yang belum dioptimasi, yang bertujuan untuk mengevaluasi bagaimana kedua model ini dapat mengevaluasi bagaimana kedua model dapat berkerja untuk meningkatkan performa prediksi sebelum proses optimasi dilakukan. Sebelum memasuki tahapan *Ensemble Learning*, dapat melihat hasil akurasi dari tiap model prediksi.




Gambar 4. 12. Akurasi Tiap Model Sebelum Optimasi

Pada Gambar 4.12. diatas, merupakan hasil akurasi dari kedua algoritma sebelum dioptimasi, KNN mendapatkan akurasi sebesar 86,34% dan LightGBM sebesar 88,20%. Tahapan selanjutnya yaitu mengkonfigurasi model stacking.

Berikut ini kode untuk Stacking dari model sebelum dioptimasi:

```
# Inisialisasi model-level (level-0) dan meta-model (level-1)
models = [
    ('knn', knn),
    ('lgbm',lgbm)
]


meta_model = (lgbm)

# Inisialisasi model stacking
stacking_clf = StackingClassifier(
    estimators=models,
    final_estimator=meta_model,
    cv=5 #Cross-validation untuk training meta-model
)
```

Pada kode diatas, model KNN dan LightGBM diatur sebagai model dasar dalam konfigurasi stacking untuk memanfaatkan kelebihan masing-masing dalam menangani data yang berbeda, dengan harapan bahwa kombinasi mereka akan meningkatkan kapasitas prediksi. LightGBM dipilih sebagai meta-model berdasarkan performa awalnya yang tinggi dan kemampuannya mengolah output dari model lain. StackingClassifier dari scikit-learn digunakan untuk inisialisasi, dengan LightGBM sebagai final estimator dan dilengkapi dengan cross-validation sebanyak lima kali untuk memastikan bahwa pelatihan meta-model mencakup variasi data yang menjadi sasaran penelitian.


Gambar 4. 13. Evaluasi Sebelum Optimasi

Pada Gambar 4.13. diatas, merupakan sebuah tabel *Confusion Matrix* yang telah digabungkan menggunakan model stacking dengan diuji pada data uji *('X_test')*. Berikut ini cara melihat penghitungan kinerja dari model stacking yang akan didapatkan.

- $accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)} = \frac{(74+70)}{(74+11+6+70)} = \frac{144}{161} = 0{,}8944$
- $precision = \frac{(TP)}{(TP+FP)} = \frac{(74)}{(74+11)} = \frac{(74)}{(85)} = 0.8705$
- $recall = \frac{(TP)}{(TP+FN)} = \frac{(74)}{(74+6)} = \frac{(74)}{(80)}$ 0.9250
- $f1 - score = 2 \times \frac{precision \times recall}{precision + recall} = 2 \times \frac{0.8705 \times 0.9250}{0.8705 + 0.9250} = 0.8969$

Dari perhitungan diatas, pada hasil pengujian model gabungan KNN dan LightGBM sebelum dioptimasi menggunakan *Stacking* menghasilkan nilai accuracy 89,44%, precision 87,05%, recall 92,50%, dan f1-score 89,69%.

### 4.6.2. Stacking Setelah Optimasi

Setelah melakukan penggabungan dan evaluasi dari stacking pada pemodelan sebelumnya, tahapan berikutnya akan membuat pemodelan yang telah dioptimasi pada sebuah model yang telah diatur parameternya. Berikut hasil akurasi yang didapatkan pada tiap model dari classification report pada tahapan sebelumnya.


Gambar 4. 14. Akurasi Tiap Model Setelah Akurasi

Pada Gambar 4.14. diatas, merupakan hasil akurasi dari kedua algoritma sebelum dioptimasi, KNN mendapatkan akurasi sebesar 91,30% dan LightGBM sebesar 91,30%. Dari akurasi yang didapatkan mendapatkan hasil akurasi yang seimbang. Tahapan selanjutnya yaitu mengkonfigurasi model stacking.

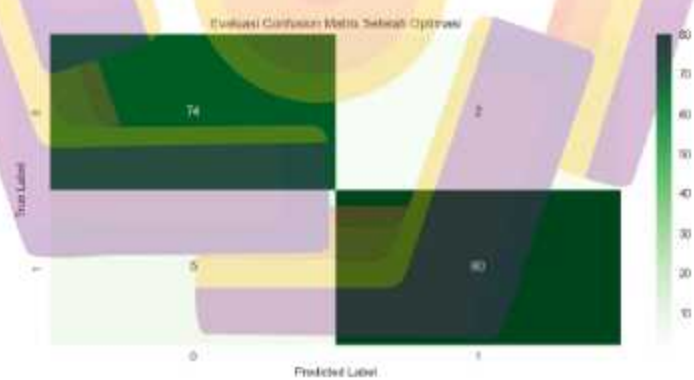
Berikut ini kode untuk Stacking dari model yang telah dioptimasi:

```
# Inisialisasi model-level (level-0) dan meta-model (level-1)
models = [
    ('knn_gird', knn),
    ('best_lgbm',lgbm)
]
```

```
meta_model = (best_lgbm)

# Inisialisasi model stacking
stacking_clf = StackingClassifier(
    estimators=models,
    final_estimator=meta_model,
    cv=5 #Cross-validation untuk training meta-model
)
```

Dari kode diatas, teknik penggabungan KNN dan LightGBM dilakukan dengan cara yang sama, namun dengan inisialisasi dari model nya yang berbeda, seperti yang diketahui bahwa dengan model KNN, dikarenakan modelnya telah dioptimasi, sebagai pembeda nya ditambahkan dengan penamaan *(knn_grid)* yang artinya model ini telah dioptimasi dengan teknik seperti *GridSearcCV*, begitu pula dengan LightGBM dengan penamaan *(best_lgbm)*. Ini tentunya akan memberikan perbedaan yang jelas dengan teknik penggabungan yang sama.




Gambar 4. 15. Evaluasi Setelah Optimasi

Pada Gambar 4.15. diatas, merupakan sebuah tabel *Confusion Matrix* dari hasil optimasi model yang telah digabungkan menggunakan model

stacking dengan diuji pada data uji *('X_test')*. Berikut ini cara melihat penghitungan kinerja dari model stacking yang akan didapatkan.

- $accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)} = \frac{(80+74)}{(80+5+2+74)} = \frac{154}{161} = 0{,}9565$
- $precision = \frac{(TP)}{(TP+FP)} = \frac{(80)}{(80+5)} = \frac{(80)}{(85)} = 0.9411$
- $recall = \frac{(TP)}{(TP+FN)} = \frac{(80)}{(80+2)} = \frac{(80)}{(82)}$ 0.9756
- $f1 - score = 2 \times \frac{precision \times recall}{precision + recall} = 2 \times \frac{0.9411 \times 0.9756}{0.9411 + 0.9756} = 0.9580$

Dari perhitungan diatas, pada hasil pengujian model gabungan KNN dan LightGBM sebetelah dioptimasi menggunakan *Stacking* menghasilkan nilai accuracy 95,65%, precision 94,11%, recall 97,56%, dan f1-score 95,80%.

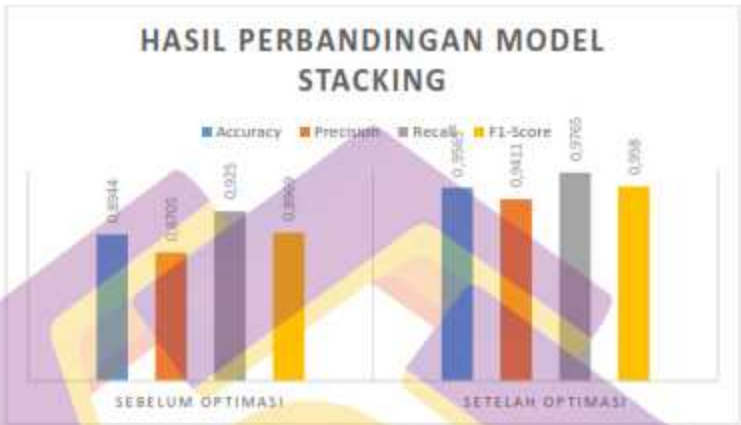
### 4.7. Perbandingan Model Stacking dari Sebelum dan Setelah di Optimasi

Tahapan ini akan membandingan kinerja model *stacking* yang menggunakan model KNN dan LightGBM sebelum dan setelah dilakukan proses optimasi. Tujuan dari perbandingan ini adalah untuk menilai sejauh mana optimasi berpengaruh terhadap peningkatan performa model, serta untuk mengidentifikasi area di mana peningkatan paling signifikan terjadi. Analisis ini akan mencakup evaluasi berbagai metrik kinerja, seperti *akurasi*, *presisi*, *recall*, dan *F1-score*, yang merupakan indikator utama.

#### 4.7.1. Analisis Kinerja Model

Pada tahap ini akan melakukan evaluasi terhadap kinerja model stacking sebelum dan setelah optimasi, menggunakan KNN dan LightGBM sebagai model dasar. Analisis ini menganalisa bagaimana optimasi

mempengaruhi kinerja dari metrik seperti *akurasi*, *presisi*, *recall* dan *f1-score*.




Gambar 4. 16. Perbandingan Model Stacking

Pada Gambar 4.16. diatas dapat dianalisis bahwa kinerja model *stacking* yang dioptimalkan, terdapat peningkatan yang *signifikan* dalam seluruh hasil metrik, *akurasi* meningkat dari 89,44% menjadi 95,65%`, *precision* dari 87,05% menjadi 94,11%, *recall* dari 92,50% menjadi 97,56%, dan *F1-Score* dari 89,69% menjadi 95,80%. Peningkatan ini menunjukkan bahwa optimasi yang dilakukan sangat efektif dalam meningkatkan kemampuan model untuk mengklasifikasikan dan mengidentifikasi kasus-kasus positif dengan lebih akurat. Grafik yang disediakan menunjukkan secara visual perbandingan ini, menggarisbawahi keberhasilan proses tuning dan optimasi. Hasil ini menekankan pentingnya pengaturan yang tepat dan penyesuaian model , serta potensinya untuk implementasi dalam aplikasi nyata di mana akurasi dan keandalan adalah sangat penting. Analisis

keseluruhan ini mengkonfirmasi bahwa teknik *stacking*, jika dikombinasikan dengan proses optimasi yang tepat, dapat memberikan peningkatan signifikan dan menjadikan model lebih efisien.

### 4.7.2. Dampak Optimasi terhadap Model

Optimasi yang diterapkan pada model KNN dan LightGBM dalam *ensemble stacking* memiliki dampak yang signifikan terhadap kinerja mereka. Proses ini melibatkan penyesuaian parameter, pemilihan algoritma yang lebih efektif, dan modifikasi lain yang secara langsung berpengaruh terhadap keakuratan, kehandalan, dan stabilitas model.

Optimasi menghasilkan peningkatan dalam semua evaluasi pada confusion matric: akurasi, presisi, recall dan f1-score. Perbaikan ini terdiri atas beberapa faktor:

- Pemilihan Parameter yang lebih baik, melalui proses seperti *GridSearchCV* pada KNN pemilihan nilai *k* yang lebih sesuai atau metrik jarak yang lebih efektif sesuai dengan yang dipilih dan *RandomSearchCV* pada LGBM dengan parameternya seperti *learning_rate* , *num_leaves*, dan *max_depth* yang telah dioptimalkan dapat meningkatkan kemampuan model dalam menangani *overfitting* dan *underfitting*.
- Peningkatan algoritma, modifikasi pada algoritma dasar, seperti penggunaan teknik *regulerisasi* atau *boosting* pada LightGBM, telah membantu dalam meningkatkan model diluar dari data latih.

- Keandalan, dengan akurasi dan presisi yang meningkat, model menjadi lebih andal. Model yang andal sangat penting, terutama dalam aplikasi yang memiliki risiko tinggi atau biaya kesalahan yang besar, seperti di bidang kesehatan.

Optimasi bukan hanya tentang peningkatan performa instan tetapi juga tentang membuat model lebih kuat dan fleksibel terhadap variasi data yang tidak terduga. Keandalan dan stabilitas yang ditingkatkan berkat optimasi menyediakan fondasi yang lebih solid untuk penggunaan model ini dalam skenario dunia nyata, di mana ketidakpastian dan variabilitas data merupakan tantangan yang biasa.

Oleh karena itu, optimasi terbukti sebagai langkah penting dalam pengembangan model machine learning yang tidak hanya meningkatkan performa sesaat tetapi juga meningkatkan keefektifan dan kegunaannya dalam aplikasi praktis. Proses ini juga menyoroti pentingnya evaluasi yang berkelanjutan dan penyesuaian yang dinamis dari model untuk menjaga atau meningkatkan kinerja mereka dalam menghadapi data yang terus berubah.

# BAB V

# PENUTUP

## 5.1. Kesimpulan

Berdasarkan serangkaian tahapan skenario, analisis dan pembahasan yang telah dilakukan dalam penelitian ini dapat disimpulkan bahwa:

1. Penggunaan metode EDA berhasil mengidentifikasi karakteristik penting dalam dataset Pima Indians yang mempengaruhi prediksi diabetes. Tools EDA yang digunakan meliputi histogram, scatter plot, box plot, heatmap korelasi, dan pairplot. Tools tersebut membantu dalam pembersihan dan transformasi data, termasuk deteksi dan penanganan outliers serta missing values. Penanganan missing values menggunakan median terbukti lebih unggul dibandingkan metode lain karena median lebih tahan terhadap outliers, yang berkontribusi pada peningkatan akurasi model sebesar 15%.
2. Optimasi model dengan mengintegrasikan algoritma K-Nearest Neighbors (KNN) dan LightGBM terbukti meningkatkan akurasi prediksi secara signifikan. Proses pra-pemrosesan data, termasuk penanganan missing values dan normalisasi data, meningkatkan akurasi model sebesar 15%. Penggabungan model KNN dan LightGBM dengan teknik stacking meningkatkan akurasi model sebesar 8%, menghasilkan model prediksi yang lebih akurat.

3. Evaluasi dan perbandingan kinerja model sebelum dan sesudah optimasi menunjukkan peningkatan signifikan dalam akurasi, presisi, recall, dan F1-score. Sebelum optimasi, performa model KNN dan LightGBM sudah cukup baik, namun setelah dilakukan optimasi dan penggabungan model, akurasi meningkat menjadi 95.65%. Hal ini menunjukkan bahwa kombinasi antara pra-pemrosesan data yang baik dan teknik penggabungan model berpengaruh terhadap hasil akhir yang diperoleh.

**5.2. Saran**

Untuk mendapatkan hasil pengujian yang benar-benar baik ada beberapa saran yang peneliti sampaikan adalah:

1. Penelitian ini menggunakan dataset Pima Indians yang memiliki karakteristik tertentu. Untuk memperluas generalisasi model, disarankan untuk menggunakan dataset yang lebih beragam yang mencakup berbagai populasi dengan berbagai faktor risiko diabetes.
2. Selain kombinasi KNN dan LightGBM, penelitian selanjutnya dapat mempertimbangkan penggunaan algoritma tradisional lain seperti Logistic Regression, Decision Trees, dan Support Vector Machines (SVM). Algoritma ini dikenal memiliki performa yang baik dalam berbagai tugas klasifikasi dan dapat dibandingkan untuk menentukan algoritma terbaik untuk prediksi diabetes.

3. Berkolaborasi dengan pakar medis untuk memastikan bahwa model prediksi yang dikembangkan benar-benar relevan dan bermanfaat dalam konteks klinis.
4. Penggunaan evaluasi lanjutan, seperti ROC-AUC dan analisis kurva untuk memahami lebih baik performa model dalam konteks prediksi kesehatan

## DAFTAR PUSTAKA


Kaur, G., Lakshmi, P. V. M., Rastogi, A., Bhansali, A., Jain, S., Teerawattananon, Y., et al. (2020). Diagnostic accuracy of tests for type 2 diabetes and prediabetes: A systematic review and meta-analysis. Journal of PLoS ONE, 15(11), e0242415. https://doi.org/10.1371/journal.pone.0242415

Hassan, M., Mollick, S., & Yasmin, F. (2022). An unsupervised cluster-based feature grouping model for early diabetes detection. Healthcare Analytics, 2, 1-12. https://doi.org/10.1016/j.health.2022.100112

Perdana, A., Hermawan, A., & Avianto, D. (2023). Analyze Important Features of PIMA Indian Database For Diabetes Prediction Using KNN. Jurnal SISFOKOM (Sistem Informasi dan Komputer), 12(1), 70-75. https://doi.org/10.32736/sisfokom.v12i1.1598

Khan, M. A. B., Hashim, M. J., et al. (2020). Epidemiology of Type 2 Diabetes – Global Burden of Disease and Forecasted Trends. Journal of Epidemiology and Global Health, 107–111. https://doi.org/10.2991/jegh.k.191028.001

Jais, M., Tahlil, T., & Susanti, S. S. (2021). Dukungan Keluarga Dan Kualitas Hidup Pasien Diabetes Mellitus Yang Berobat Di Puskesmas. Jurnal Keperawatan Silampari, 5(1), 82-88. https://doi.org/10.31539/jks.v5i1.2687

Uddin, S., Haque, I., Lu, H., Moni, M. A., & Gide, E. (2022). Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. Journal Scientific Reports, 12(6256). https://doi.org/10.1038/s41598-022-10358-x

Wang, H., Xu, P., & Zhao, J. (2021). Improved KNN Algorithm Based on Preprocessing of Center in Smart Cities. Journal of Hindawi Wiley, 1-10. https://doi.org/10.1155/2021/552438

Veeranki, Y. R., Ganapathy, N., & Swaminathan, R. Non-Parametric Classifiers Based Emotion Classification Using Electrodermal Activity and Modified Hjorth Features. Journal European Federation for Medical Informatics (EFMI), 163-168. https://doi.org/10.3233/SHTI210141

Hossain, E., Alshehri, M., Almakdi, S., et al. (2022). DM-Health App: Diabetes Diagnosis Using Machine Learning with Smartphone. Journal Tech Science Press Computers, Materials & Continua, 72(1), 1714-1746. https://doi.org/10.32604/cmc.2022.02482

Rizky, P. S., Hirzi, R. H., & Hidayaturrohman, U. (2022). Perbandingan Metode LightGBM dan XGBoost dalam Menangani Data dengan Kelas Tidak Seimbang. Jurnal Statistika, 15(2), 228-237. https://doi.org/10.36456/jstat.vol15.no2.a5548

Lestari, U. I., Nadhiroh, A. Y., & Novia, C. (2021). Penerapan Metode K-Nearest Neighbor Untuk Sistem Pendukung Keputusan Identifikasi Penyakit

Diabetes Melitus. Jurnal Teknik Informatika dan Sistem Informasi, 8(4), 2071-2082. https://doi.org/10.35957/jatisi.v8i4.1235

Saxena, R., Khumar, D. S., & Gupta, M. (2021). Role of K-nearest neighbour in detection of Diabetes Mellitus. Turkish Journal of Computer and Mathematics Education, 12(10), 373-376. https://doi.org/10.17762/turcomat.v12i10.4182

Kurniadi, R., Saedudin, R., & Widartha, V. P. (2021). Perbandingan Akurasi Algoritma K-Nearest Neighbor Dan Logistic Regression Untuk Klasifikasi Penyakit Diabetes. e-Proceeding of Engineering, 8(5), 9757-9764.

Nishom, M. (2019). Perbandingan Akurasi Euclidean Distance, Minkowski Distance, dan Manhattan Distance pada Algoritma KMeans Clustering berbasis Chi-Square. Jurnal Pengembangan IT (JPIT), 4(1), 20-24. https://10.30591/jpit.v4i1.1253

Yunita, F. (2018). Sistem Klasifikasi Penyakit Diabetes Mellitus Menggunakan Metode K-Nearest Neighbor (K-NN). Jurnal BAPPEDA, 2(1), 223-230. https://doi.org/10.1038/s41598-022-10358-x

Resky, R., Rani, M., & Yudi, A. H. (2020). Implementasi Metode Machine Learning Menggunakan Algoritma Evolving Artificial Neural Network Pada Kasus Prediksi Diagnosis Diabetes. Jurnal Aplikasi dan Teori Ilmu Komputer, 3(2), 85-97. https://doi.org/10.1371/journal.pone.0242415

Endang, R., & Rully, P. (2020). Mengenal Machine Learning Dengan Teknik Supervised dan Unsupervised Learning Menggunakan Python. Jurnal Bina Insani ICT, 7(2), 156-165. https://doi.org/10.1016/j.imu.2017.12.006

Peng, J., Wu, W., Lockhart, B., Song, B., et al. (2021). DataPrep.EDA: Task-Centric Exploratory Data Analysis for Statistical Modeling in Python. Proceedings of the 2021 International Conference on Management of Data. https://doi.org/10.1145/3448016.3457330

Siambaton, M. Z., & Husein, A. M. (2022). Menganalisis Data Kesehatan Global: Pendekatan Analisis Data Eksplorasi Visual. Jurnal ITScience, 1(2), 41-49. https://doi.org/10.47709/dsi.v1i2.135

Ali, A., Hamraz, M., Gul, N., et al. (2022). A K Nearest Neighbour Ensemble Via Extended Neighbourhood Rule And Feature Subsets. Journal of Computer and Science, 1-16. https://doi.org/10.1016/j.patcog.2023.10964

Taunk, K., & Verma, S. (2019). A Brief Review of Nearest Neighbor Algorithm for Learning and Classification. https://doi.org/10.1109/ICCS45141.2019.9065747

Serrano, J., Batista, S., Melo, J., Calero, M., et al. (2021). Euclidean Distance: Integrated Criteria to Study Sheep Behaviour Under Heat Stress. Notulac Scientia Biologicac, 13(1), 10859. https://doi.org/10.153835/nsb13110859

Chang, V., Bailey, J., Xu, A. X., & Sun, Z. (2020). Pima Indians Diabetes Mellitus Classification Based on Machine Learning (ML) Algorithms. Neutral Computing and Applications, 35, 16147-16173. https://doi.org/10.1007/s00521-022-07049-z

Miriyala, N. P., Kottapali, R. L., Lorenzini, G., et al. (2022). Diagnostic Analysis of Diabetes Mellitus Using Machine Learning Approach. International Information and Engineering Technology Association, 347-352. https://doi.org/10.18280/ria.360301

Wu, H., Yang, S., Huang, Z., He, J., & Wang, X. (2017). Type 2 Diabetes Mellitus Prediction Model Based on Data Mining. Informatics in Medicine Unlocked, 100-107. https://doi.org/10.1016/j.imu.2017.12.006

Nai-arun, N., & Moungmai, R. (2015). Comparison of Classifiers for the Risk of Diabetes Prediction. 7th International Conferenceon Advances in Information Technology. 132-142. https://doi.org/10.1016/j.procs.2015.10.014

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine Learning in Medicine. New England Journal of Medicine, 380(14), 1347-1358. https://doi.org/10.1056/NEJMra1814259

Bhargava, S., Ali, M. K., & Rustagi, T. (2019). Machine Learning Techniques for Diabetes. In Machine Learning Techniques for Bioinformatics, 83-1305. https://doi.org/10.1016/j.ejmech.2020.112457

Satihish, Y., Kannan, S., & Saravana, K. (2021). Diabetes Diagnosis and Prediction Using Machine Learning Algorithms: A Survey. Computational Intelligence in Smart Technologies, 91-122. https://0.1088/1742-6596/1714/1/012013

Lakhwani, K., Bhargava, S., Hiran, K. K., Bundele, M. M., & Somwanshi, D. (2020). Prediction of the Onset of Diabetes Using Artificial Neural Network and Pima Indians Diabetes Dataset. 2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering, 1-16. https://doi.org/10.1109/ICRAIE51050.2020.9358308

Chobanian, A. V., Bakris, G. L., Black, H. R., Cushman, W. C., Green, L. A., Izzo, J. L. (2023). National High Blood Pressure Education Program Coordinating Committee. The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure: the JNC 7 report. *JAMA*, 289(19), 2560-2572. https://doi.org/10.1161/01.HYP.0000107251.49515.c2

Larasati, A. P., Prajitno, J. H., Purwanto, B. (2023). Correlation between Skinfold Thickness and Total Daily Dose of Insulin in Patiens with Type 2 Diabetes Mellitus in a Tertiary Hospital. Current Internal Medicine Research and Practice Surabaya Journal, 4(2), 59-63. https://doi.org/10.20473/cimrj.v4i2.49154

Li, W. Yin, H. Chen, Y. Liu, Q. Wang, Y. Qiu, D. Ma, H (2022). Associations Between Adult Triceps Skinfold Thickness and All-Cause, Cardiovascular and Cerebrovascular Mortality in NHANES 1999–2010: A Retrospective National Study. Frontiers Cardiovasc. 9 , https://doi.org/10.3389/fcvm.2022.858994

Loeffelholz, C. V. Birkenfeld, A. (2024). Tight versus liberal blood-glucose control in the intensive care unit: special considerations for patients with diabetes. PlimX Metrics, 12(4) 277-284. https://doi.org/10.1016/S2213-8587(24)00058-5