

**IMPLEMENTASI *EXTREME GRADIENT BOOSTING* (XGBoost) UNTUK  
KLASIFIKASI RISIKO DIABETES MELLITUS BERDASARKAN  
RIWAYAT KESEHATAN GUNA MENINGKATKAN DIAGNOSA DINI**

**LAPORAN TUGAS AKHIR**

Proposal ini Disusun untuk Memenuhi Salah Satu Syarat Memperoleh  
Gelar Sarjana Strata 1 (S1) Pada Program Studi Teknik Informatika  
Fakultas Teknologi Industri Universitas Islam Sultan Agung



**DISUSUN OLEH:**

**LAILATUL FITRIYAH**

**32602000107**

**PROGRAM STUDI TEKNIK INFORMATIKA  
FAKULTAS TEKNOLOGI INDUSTRI  
UNIVERSITAS ISLAM SULTAN AGUNG  
SEMARANG  
2024**

**FINAL PROJECT**

**IMPLEMENTASI *EXTREME GRADIENT BOOSTING* UNTUK  
KLASIFIKASI RISIKO DIABETES MELLITUS BERDASARKAN  
RIWAYAT KESEHATAN GUNA MENINGKATKAN DIAGNOSA DINI**

*Proposed to complete the requirement to obtain a bachelor's degree (S-1) at  
Informatics Engineering Departement of Industrial Technology  
Faculty Sultan Agung Islamic University*



*Arranged By :*

**LAILATUL FITRIYAH**

**NIM 32602000107**

**MAJORING OF INFORMATICS ENGINEERING**

**INDUSTRIAL TECHNOLOGY FACULTY**

**SULTAN AGUNG ISLAMIC UNIVERSITY**

**SEMARANG**

**2024**

## LEMBAR PENGESAHAN PEMBIMBING

Laporan Tugas Akhir dengan judul **“IMPLEMENTASI *EXTREME GRADIENT BOOSTING* (XGBoost) UNTUK KLASIFIKASI RISIKO DIABETES MELLITUS BERDASARKAN RIWAYAT KESEHATAN GUNA MENINGKATKAN DIAGNOSA DINI”** ini disusun oleh :

Nama : Lailatul Fitriyah

NIM : 32602000107

Program Studi : Teknik Informatika

Telah disahkan oleh dosen pembimbing pada :

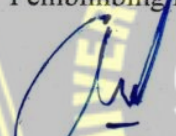
Hari : Jumat

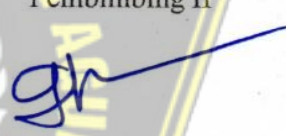
Tanggal : 30 Agustus 2024

Mengesahkan,

Pembimbing I

Pembimbing II

  
Andi Rihsyah, ST., M.Kom  
NIK. 210 616 053

  
Ghufron, ST., M.Kom  
NIK. 210 622 056

Mengetahui,

Ketua Program Studi Teknik Informatika  
Fakultas Teknologi Industri  
Universitas Islam Sultan Agung

  
Moch. Faufik, ST., MIT  
NIK. 210 604 034

## LEMBAR PENGESAHAN PENGUJI


Laporan tugas akhir dengan judul “Implementasi *Extreme Gradient Boosting* (XGBoost) Untuk Klasifikasi Risiko Diabetes Mellitus Berdasarkan Riwayat Kesehatan Guna Meningkatkan Diagnosa Dini” ini telah dipertahankan di depan dosen penguji Tugas Akhir pada :

Hari : *Senin*


Tanggal : *2 September 2024*

### TIM PENGUJI

Ketua Penguji

  
Badie'ah, ST., M.Kom  
NIK. 210615044

Anggota Penguji

  
Asih Widi Harini, S.Si., MT  
NIK. 110010398



## SURAT PERNYATAAN KEASLIAN TUGAS AKHIR

Yang bertanda tangan dibawah ini :

Nama : Lailatul Fitriyah

NIM : 32602000107

Judul Tugas Akhir : Implementasi *Extreme Gradient Boosting* (XGBoost)  
Untuk Klasifikasi Risiko Diabetes Mellitus Berdasarkan  
Riwayat Kesehatan Guna Meningkatkan Diagnosa Dini

Dengan bahwa ini saya menyatakan bahwa judul dan isi Tugas Akhir yang saya buat dalam rangka menyelesaikan Pendidikan Strata Satu (S1) Teknik Informatika tersebut adalah asli dan belum pernah diangkat, ditulis ataupun dipublikasikan oleh siapapun baik keseluruhan maupun sebagian, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam daftar pustaka, dan apabila di kemudian hari ternyata terbukti bahwa judul Tugas Akhir tersebut pernah diangkat, ditulis ataupun dipublikasikan, maka saya bersedia dikenakan sanksi akademis. Demikian surat pernyataan ini saya buat dengan sadar dan penuh tanggung jawab.

Semarang, 2 September 2024

Yang Menyatakan,



10000  
SEPULUH RIBU RUPIAH  
METERAI  
TEMPEL  
D04ALX358073589

Lailatul Fitriyah

## PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH

Saya yang bertanda tangan dibawah ini :

Nama : Lailatul Fitriyah

NIM : 32602000107

Program Studi : Teknik Informatika

Fakultas : Teknologi industri

Dengan ini menyatakan Karya Ilmiah berupa Tugas akhir dengan Judul : Implementasi *Extreme Gradient Boosting* (XGBoost) untuk Klasifikasi Risiko Diabetes Mellitus Berdasarkan Riwayat Kesehatan Guna Meningkatkan Diagnosa Dini.

Menyetujui menjadi hak milik Universitas Islam Sultan Agung serta memberikan Hak bebas Royalti Non-Eksklusif untuk disimpan, dialihmediakan, dikelola dan pangkalan data dan dipublikasikan diinternet dan media lain untuk kepentingan akademis selama tetap menyantumkan nama penulis sebagai pemilik hak cipta. Pernyataan ini saya buat dengan sungguh-sungguh. Apabila dikemudian hari terbukti ada pelanggaran Hak Cipta/Plagiarisme dalam karya ilmiah ini, maka segala bentuk tuntutan hukum yang timbul akan saya tanggung secara pribadi tanpa melibatkan Universitas Islam Sultan agung.

Semarang, 2 September 2024

Yang menyatakan,

  
Lailatul Fitriyah



## KATA PENGANTAR

Saya bersyukur kepada Allah SWT karena telah memberi saya kesempatan untuk menyelesaikan Tugas Akhir dengan judul “Implementasi *Extreme Gradient Boosting* (XGBoost) Untuk Klasifikasi Risiko Diabetes Mellitus Berdasarkan Riwayat Kesehatan Guna Meningkatkan Diagnosa Dini”. Tugas Akhir ini merupakan salah satu syarat untuk menyelesaikan studi dan memungkinkan saya untuk memperoleh gelas sarjana (S-1) di Program Studi Teknik Informatika di Fakultas Teknologi Industri Universitas Islam Sultan Agung Semarang.

Sebagai penulis, saya ingin mengucapkan terima kasih kepada berbagai pihak yang telah membantu dalam penyusunan dan pembuatan tugas akhir ini, antara lain :

1. Rektor UNISSULA, Bapak Prof. Dr. H. Gunarto, S.H., M.Hum;
2. Dekan Fakultas Teknologi Industri, Ibu Dr. Hj. Novi Marlyana ST., M.T., IPU, Asean Eng;
3. Kepala Prodi S1-Teknik Informatika UNISSULA Bapak Ir. Sri Mulyono. M.Eng;
4. Dosen Pembimbing I Bapak Andi Riansyah, S.T., M.Kom, dan Dosen Pembimbing II Bapak Ghufron, S.T., M.Kom, yang telah meluangkan waktu dan memberikan pengetahuan;
5. Orang tua dan keluarga penulis yang telah memberikan doa dan dukungan penuh untuk menyelesaikan penelitian ini sebagai syarat kelulusan Program Studi S1 Teknik Informatika;
6. Teman-teman kontrakan yang telah menemani, mendukung dan menjadi tempat berbagi cerita selama 4 tahun terakhir;
7. Dengan rendah hati, penulis menyadari bahwa laporan masih kekurangan kualitas, kuantitas, dan ilmu pengetahuan. Oleh karena itu, penulis mengharapkan kritikan dan saran yang membangun untuk membantu memperbaiki laporan ini;

Akhir kata, saya berharap Allah SWT berkenan membalas segala kebaikan semua pihak yang turut serta membantu sehingga laporan Tugas Akhir ini dapat

terselesaikan dengan baik dan semoga laporan ini dapat memberikan manfaat bagi pembaca.

Semarang, 25 Agustus 2024

Lailatul Fitriyah





## DAFTAR ISI

<b>HALAMAN JUDUL.....</b>	<b>1</b>
<b>FINAL PROJECT .....</b>	<b>ii</b>
<b>LEMBAR PENGESAHAN PEMBIMBING .....</b>	<b>iii</b>
<b>LEMBAR PENGESAHAN PENGUJI .....</b>	<b>iv</b>
<b>SURAT PERNYATAAN KEASLIAN TUGAS AKHIR .....</b>	<b>v</b>
<b>PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH.....</b>	<b>vi</b>
<b>KATA PENGANTAR.....</b>	<b>vii</b>
<b>DAFTAR ISL.....</b>	<b>ix</b>
<b>DAFTAR TABEL.....</b>	<b>xi</b>
<b>DAFTAR GAMBAR .....</b>	<b>xii</b>
<b>ABSTRAK.....</b>	<b>xiii</b>
<b>BAB I PENDAHULUAN.....</b>	<b>1</b>
1.1 Latar Belakang .....	1
1.2 Rumusan Masalah.....	2
1.3 Batasan Masalah.....	2
1.4 Tujuan.....	3
1.5 Manfaat.....	3
1.6 Sistematika Penulisan .....	3
<b>BAB II TINJAUAN PUSTAKA DAN DASAR TEORI.....</b>	<b>5</b>
2.1 Tinjauan Pustaka .....	5
2.2 Dasar Teori.....	8
2.2.1 Diabetes Mellitus .....	8
2.2.2 Klasifikasi .....	9
2.2.3 Data <i>Pre Processing</i> .....	9
2.2.4 <i>Decision Tree</i> (Pohon Keputusan) .....	10
2.2.5 <i>Extremen Gradient Boosting (XGBoost)</i> .....	12
2.2.6 Evaluasi Model.....	16
<b>BAB III METODE PENELITIAN .....</b>	<b>18</b>

3.1	Metode Penelitian .....	18
3.1.1	Studi Literatur .....	18
3.1.2	Pengumpulan Data .....	19
3.1.3	Pemodelan Sistem.....	22
3.1.4	Evaluasi Model.....	23
3.1.5	Implementasi Streamlit .....	24
<b>BAB IV HASIL DAN ANALISIS PENELITIAN.....</b>		<b>25</b>
4.1	Hasil Pemodelan .....	25
4.1.1	Memuat Data.....	25
4.1.2	<i>Preprocessing</i> .....	26
4.1.3	<i>Split</i> Data .....	28
4.1.4	Format Dmatrix .....	28
4.1.5	Menentukan Hyperparameter .....	30
4.1.6	Pelatihan Model.....	31
4.1.7	Evaluasi Model.....	32
4.1.8	Hasil Implementasi Streamlit .....	34
4.2	Analisis Penelitian .....	38
<b>BAB V KESIMPULAN DAN SARAN.....</b>		<b>44</b>
5.1	Kesimpulan .....	44
5.2	Saran.....	44
<b>DAFTAR PUSTAKA</b>		

## DAFTAR TABEL

Tabel 3. 1 Atribut dalam dataset.....	19
Tabel 4. 1 Hasil mengubah format ke dmatrix.....	29
Tabel 4. 2 Nilai confusion matrix .....	33
Tabel 4. 3 Hasil evaluasi kinerja model .....	33
Tabel 4. 4 Hasil evaluasi setiap kelas.....	33
Tabel 4. 5 Penjelasan tampilan utama website.....	35
Tabel 4. 6 Analisis Penelitian.....	38



## DAFTAR GAMBAR

Gambar 2. 1 Struktur pohon keputusan (Husaini <i>dkk.</i> , 2023) .....	11
Gambar 2. 2 Arsitektur <i>XGBoost</i> (Mehdary <i>dkk.</i> , 2024) .....	13
Gambar 2. 3 Cara kerja <i>XGBoost</i> (AWS, 2024) .....	14
Gambar 3. 1 Alur Penelitian.....	18
Gambar 3. 2 <i>Flowchart</i> pemodelan sistem .....	22
Gambar 4. 1 <i>Output</i> dari dataset.....	25
Gambar 4. 2 <i>Output</i> sebelum di encoding data .....	26
Gambar 4. 3 <i>Output</i> setelah di <i>encoding</i> data.....	27
Gambar 4. 4 <i>Output</i> dari memisahkan fitur dan target .....	27
Gambar 4. 5 <i>Output</i> <i>standarisasi</i> data.....	28
Gambar 4. 6 <i>Output</i> pembagian data.....	28
Gambar 4. 7 <i>Output</i> mengubah format ke Dmatrix.....	29
Gambar 4. 8 <i>Output</i> mendefinisikan <i>hyperparameter</i> .....	30
Gambar 4. 9 <i>Output</i> pelatihan model. ....	31
Gambar 4. 10 <i>Output</i> evaluasi model.....	32
Gambar 4. 11 Tampilan awal webs ite .....	34
Gambar 4. 12 Tampilan setelah memasukkan input dengan hasil positif .....	36
Gambar 4. 13 Tampilan setelah memasukkan input hasil dengan hasil negatif ...	36

## ABSTRAK

DM atau diabetes mellitus adalah penyakit kronis yang terus meningkat diseluruh dunia, membawa dampak signifikan terhadap kualitas hidup penderita. Prediksi dini dan manajemen yang tepat adalah kunci untuk meminimalkan komplikasi yang terkait dengan penyakit ini. Namun, kompleksitas dalam proses diagnosis sering kali menjadi hambatan, terutama di daerah dengan akses terbatas ke layanan kesehatan. Oleh karena itu, diperlukan solusi inovatif yang mampu mendukung upaya pencegahan dan perawatan diabetes. Penelitian ini berfokus pada pengembangan sistem berbasis web menggunakan algoritma *Extreme Gradient Boosting* (XGBoost) untuk mengklasifikasi risiko diabetes berdasarkan riwayat kesehatan. Hasil penelitian menunjukkan bahwa sistem ini dapat memberikan prediksi yang akurat dan konsisten, dengan metrik kinerja yang menunjukkan tingkat akurasi 0.9888, precision 1.0000, recall 0.9718 dan F1-score 0.9857 selama tahap pengujian sistem. Temuan ini menunjukkan bahwa sistem berbasis web menggunakan *Extreme Gradient Boosting* (XGBoost) memiliki potensi besar sebagai alat bantu yang efektif untuk prediksi dini, sehingga dapat meningkatkan dan membantu dalam pengelolaan penyakit diabetes secara lebih baik, terutama di daerah dengan akses terbatas ke layanan kesehatan.

**Kata Kunci :** Diabetes, Klasifikasi, Prediksi, XGBoost

### ABSTRACT

*DM or diabetes mellitus is a chronic disease that continues to rise globally, significantly impacting the quality of life of those affected. Early prediction and proper management are crucial to minimizing the complications associated with this disease. However, the complexity of the diagnostic process often poses a challenge, especially in areas with limited access to healthcare services. Therefore, innovative solutions are needed to support efforts in diabetes prevention and care. This research focuses on the development of a web-based system utilizing the *Extreme Gradient Boosting* (XGBoost) algorithm to classify diabetes risk based on medical history. The study results indicate that this system can provide accurate and consistent predictions, with performance metrics showing an accuracy of 0.9888, a precision of 1.0000, a recall of 0.9718, and an F1-score of 0.9857 during the system testing phase. These findings suggest that the web-based system using *Extreme Gradient Boosting* (XGBoost) holds great potential as an effective tool for early prediction, thereby improving and assisting in better diabetes management, particularly in areas with limited access to healthcare services.*

**"Keywords :** Diabetes, Classification, Prediction, XGBoost"

# BAB I

## PENDAHULUAN

### 1.1 Latar Belakang

Pada dasarnya, diabetes merujuk pada sekelompok kondisi yang mempengaruhi proses dalam tubuh, kondisi ini ditunjukkan dengan meningkatnya jumlah gula dalam darah karena kerusakan pada fungsi organ seperti ginjal, jantung, saraf dan pembuluh darah. Penyakit ini dapat mempengaruhi berbagai organ tubuh dan menimbulkan gejala yang bervariasi. Diabetes merupakan masalah kesehatan yang dapat dikenali dengan bertambahnya jumlah kadar gula dalam darah secara berkelanjutan (Aqsha *dkk.*, 2024).

Data yang diperoleh dari hasil Riset Kesehatan Dasar (Riskesdes) Indonesia antara tahun 2013 hingga 2018 menunjukkan bahwa prevalensi diabetes meningkat di hampir seluruh provinsi, kecuali Nusa Tenggara Timur (0,9%), serta Maluku dan Papua (1,1%). Prevalensi diabetes yang didiagnosis oleh dokter dalam Riskesdes 2018 memberikan gambaran rinci tentang distribusi penyakit diabetes di seluruh Indonesia, menekankan pentingnya pemahaman mendalam untuk merencanakan intervensi yang efektif dalam pencegahan dan pengelolaan diabetes di masyarakat. Diabetes telah berkembang menjadi masalah kesehatan global dengan angka yang semakin tinggi setiap tahunnya, yang menunjukkan bahwa diabetes bukan hanya masalah kesehatan di tingkat lokal tetapi telah menjadi tantangan global yang membutuhkan perhatian dan tindakan bersama dari berbagai pihak yang memiliki kesadaran. Dengan adanya kecerdasan buatan diharapkan dapat memprediksi penyakit khususnya diabetes mellitus supaya pasien dapat mendapatkan perawatan yang tepat dan tanggap (Suharno *dkk.*, 2024).

Akhir-akhir ini, penggunaan teknik pembelajaran mesin untuk klasifikasi penyakit sangat populer. Pendekatan ini memberi kesempatan bagi para peneliti untuk menganalisis dataset lebih cermat lagi sehingga berhasil menciptakan model yang mampu mengelompokkan dan memprediksi dengan akurat (Salsabil *dkk.*, 2024).

Algoritma yang biasanya menerapkan pohon keputusan sebagai *classifier* serta mengaplikasikan *Gradient Boosting* sebagai dasarnya adalah *Extreme Gradient Boosting* (XGBoost). Prosesnya tidak berjalan secara berurutan, melainkan secara *multi-threaded*. Hal ini memanfaatkan inti CPU secara optimal, sehingga meningkatkan kecepatan dan kinerja algoritma secara signifikan (Abdurrahman dkk., 2022). Algoritma ini menerapkan teknik *ensemble* dan fokus pada *gradient* dari pengoptimalan fungsi kerugian, sehingga model yang dihasilkan akan sangat kuat dan akurat.

Dengan *Extreme Gradient Boosting* (XGBoost), penelitian ini dimaksudkan untuk menghasilkan model yang dapat memprediksi risiko diabetes berdasarkan riwayat kesehatan yang dialami seseorang untuk meningkatkan diagnosa sedini mungkin, peneliti berharap dapat membantu dalam memberikan kontribusi dalam pengembangan alat prediksi yang lebih akurat dan dapat membantu dalam upaya pencegahan dan penanganan dini bagi seseorang yang berisiko, serta memberikan informasi tambahan mengenai faktor risiko yang berhubungan tentang dm.

## 1.2 Rumusan Masalah

Bagaimana tingkat keberhasilan algoritma *Extreme Gradient Boosting* (XGBoost) dalam menghasilkan prediksi risiko diabetes mellitus berdasarkan riwayat kesehatan yang dialami seseorang secara akurat.

## 1.3 Batasan Masalah

Batasan masalah yang ditetapkan dalam penulisan laporan ini adalah sebagai berikut :

1. Hanya mengidentifikasi 2 kategori, yaitu positif untuk seseorang yang berisiko memiliki diabetes (1) dan negatif untuk seseorang yang tidak berisiko diabetes (0)
2. Penelitian ini hanya membahas simulasi prediksi risiko diabetes mellitus berdasarkan riwayat kesehatan.

#### 1.4 Tujuan

Penelitian ini memiliki tujuan untuk menyusun sebuah sistem yang mampu menghasilkan prediksi risiko diabetes mellitus seseorang berdasarkan riwayat kesehatan yang dialami dengan menerapkan algoritma XGBoost, Diharapkan mampu membangun model yang mampu memberikan prediksi secara akurat.

#### 1.5 Manfaat

Sistem yang dikembangkan dapat membantu pengguna dalam mendignosis apakah seseorang memiliki risiko diabetes mellitus atau tidak, berdasarkan riwayat kesehatan yang dialami

#### 1.6 Sistematika Penulisan

Rancangan penulisan laporan tugas akhir ini disusun dalam setiap bab, dengan alur penelitian yang telah dibuat oleh penulis :

##### BAB I : PENDAHULUAN

Membahas latar belakang yang menjadikan topik ini layak dijadikan sebagai judul penelitian, menguraikan masalah yang akan diselesaikan dalam rumusan masalah, membatasi ruang lingkup penelitian agar penelitian ini tidak menjadi lebar dalam batasan masalah, menetapkan tujuan yang ingin dicapai, menjelaskan manfaat yang diharapkan dari pengembangan sistem ini, dan menjabarkan sistematika penulisan sebagai rancang yang terarah dalam penulisan laporan tugas akhir.

##### BAB II : TINJAUAN PUSTAKA DAN DASAR TEORI

Membahas penelitian terdahulu dan teori-teori yang relevan untuk mendorong analisis permasalahan, sebagai dasar dalam penyusunan laporan tugas akhir ini. Tinjauan pustaka mencakup ulasan penelitian-penelitian terdahulu yang terkait dengan topik, sementara dasar teori menyediakan konsep-konsep dan prinsip-prinsip ilmiah yang digunakan sebagai acuan dalam menganalisis dan memecahkan masalah yang



dibahas dalam tugas akhir.

**BAB III : METODE PENELITIAN**

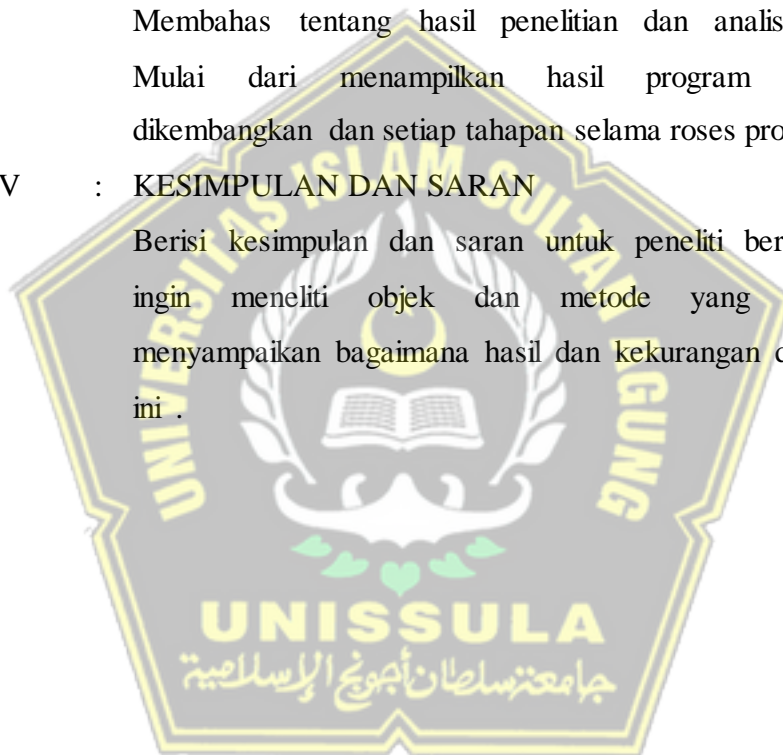
Membahas tentang rencana proses sistem dibangun sebagai prediksi risiko diabetes memakai algoritma XGBoost. Analisis ini meliputi perancangan sistem dan *interface design* yang mencakup tata letak visual, navigasi dan bagaimana pengguna dapat menggunakannya.

**BAB IV : HASIL DAN ANALISIS PENELITIAN**

Membahas tentang hasil penelitian dan analisis penelitian. Mulai dari menampilkan hasil program yang telah dikembangkan dan setiap tahapan selama proses program dibuat.

**BAB V : KESIMPULAN DAN SARAN**

Berisi kesimpulan dan saran untuk peneliti berikutnya yang ingin meneliti objek dan metode yang sama serta menyampaikan bagaimana hasil dan kekurangan dari penelitian ini .



## BAB II

### TINJAUAN PUSTAKA DAN DASAR TEORI

#### 2.1 Tinjauan Pustaka

Tinjauan pustaka ini mengulas berbagai penelitian yang menggunakan metode XGBoost. Seiring dengan pemahaman yang lebih mendalam tentang faktor-faktor risiko dan kemajuan teknologi dalam analisis data, berbagai pendekatan telah dikembangkan untuk meningkatkan akurasi dari model prediksi yang telah diterbitkan dalam literatur ilmiah.

Penggunaan metode *certainty factor* dalam diagnosis diabetes dengan sistem pakar hanya mampu menghasilkan *screening* awal kepada pengguna berdasarkan gejala yang terdapat dalam database. Hasil persentase bersifat subjektif karena penilaian ahli dapat bervariasi berdasarkan pengetahuan dan pengalaman mereka. Metode ini hanya menentukan kepastian dan ketidakpastian tanpa memberikan jaminan akurasi penuh, sehingga keakuratan analisis belum mencapai 100% (Hidayah, 2023).

Pada penelitian (Hana, 2020), klasifikasi diabetes menggunakan algoritma C4.5, dalam sistem mengidentifikasi apakah pasien termasuk penderita diabetes atau tidak. Dengan kumpulan data yang terdiri dari 16 atribut dan 520 entri yang dipisahkan menjadi 416 entri sebagai data latih dan 104 entri sebagai data uji. Hasil evaluasi menunjukkan bahwa kurva *ROC (Receiver Operating Characteristic)* memiliki nilai AUC sebesar 0.994, mengindikasikan klasifikasi yang sangat baik. dan menghasilkan akurasi sebesar 97,12%, *precision* sebesar 9,02% dan *recall* sebesar 100,00%. Temuan ini menegaskan bahwa penelitian ini mampu mengklasifikasikan penyakit diabetes dengan tingkat akurasi yang tinggi.

Dalam penelitiannya, tentang penerapan metode SVM, MLP dan XGBoost menunjukkan analisis yang menarik terkait dengan karakteristik diabetes 2. Ditemukan bahwa diabetes tipe 2 biasanya lebih umum terjadi pada seseorang umur 20 tahun ke atas, dengan proporsi yang menderita diabetes tipe 2 lebih dikuasai wanita dibanding laki-laki. Dari ke 3 metode, Penggunaan metode SVM berhasil mengkategorikan data dengan tingkat

akurasi mencapai 91.30% menunjukkan performa yang lebih unggul pada data ekspresi gen skeletal muscle normal (NGT), IGT dan diabetes 2 ketimbang menggunakan arsitektur multilayer perceptron menghasilkan 78.26% dan *Extreme Gradient Boosting* menghasilkan 73.91% (Rahman, 2020).

Penelitian lainnya dilakukan oleh (Abdurrahman *dkk.*, 2022), menunjukkan bahwa dalam melakukan klasifikasi data menggunakan algoritma XGBoost dalam meningkatkan nilai akurasi dapat dilakukan dengan menyesuaikan parameter-parameter tertentu yang tidak dipelajari oleh model itu sendiri, tetapi mempengaruhi kinerja dan perilaku model. Ketika *hyperparameter* tuning diterapkan dengan menggunakan metode *grid search* dan *random search*, nilai akurasi model meningkat secara signifikan menjadi 95%. Ini menandakan bahwa dengan menyesuaikan *hyperparameter*, model dapat menghasilkan prediksi yang lebih akurat. Walaupun keduanya (*Gridsearch* dan *Random Search*) menghasilkan nilai akurasi yang sama, namun proses *gridsearch* memerlukan waktu eksekusi yang lebih lama karena melakukan pencarian semua kemungkinan kombinasi *hyperparameter* yang telah ditentukan sebelumnya. Oleh karena itu, *Gridsearch* sering disebut algoritma *exhaustive search*. Meskipun memerlukan waktu eksekusi yang lebih lama, metode ini memberikan hasil yang lebih optimal dengan nilai akurasi yang lebih tinggi.

Penelitian ini mengevaluasi metode *boosting*, di mana SAMME pada algoritma AdaBoost menunjukkan akurasi tertinggi mencapai 91.14%, sedangkan algoritma AdaBoost dengan metode SAMME menghasilkan akurasi lebih rendah, yaitu menghasilkan *f1-score* 0.93 dan nilai AUC 0.9693. Sementara pada algoritma LightGBM, metode *boosting* GOSS menunjukkan tingkat akurasi tertinggi, mencapai 91.67% dibandingkan dengan metode *Samme.R* dengan akurasi 89.58%, dan metode DART yang mencapai akurasi 91.14% dan metode GBDT dengan akurasi 89.58%. LightGBM dengan metode *boosting* GOSS juga menghasilkan nilai *precision* 0.96 dengan *recall* 0.92, menghasilkan *f1-score* 0.94 dan AUC 0.9704. Dari hasil penelitian ini, terlihat bahwa tingkat akurasi LightGBM dengan metode *Boosting* GOSS lebih tinggi dari pada AdaBoost dengan metode *Boosting Samme* (Ahsana *dkk.*, 2021).

Penelitian tentang implementasi data mining menggunakan *Random Forest* Dan *XGBoost*, hasilnya menunjukkan bahwa kedua metode tersebut memberikan prediksi yang tepat dan stabil. Penelitian ini menggunakan analisis data yang terdiri dari 768 data dan 9 tanda yang didapatkan dari situs *kaggle*. Data tersebut kemudian diproses menggunakan langkah preprocessing, mencakup penanganan nilai yang tidak lengkap, penanganan outlier dan normalisasi, sehingga menghasilkan data sebanyak 688 data yang siap untuk dianalisis. Setelah tahap *preprocessing*, data dipelajari dan diuji menggunakan metode *Cross Validation* untuk menemukan parameter terbaik dalam model, baik *Random Forest* maupun *XGBoost*, keduanya memberikan performa yang baik. Akurasi keseluruhan yang diperoleh dari penggunaan *Random forest* adalah 74%, sedangkan penggunaan *XGBoost* adalah 76%. Temuan ini menekankan pentingnya teknik Data Mining untuk mengatasi permasalahan kesehatan global, khususnya dalam deteksi dini dan manajemen penyakit seperti diabetes (Salsabil dkk., 2024).

Penelitian selanjutnya dilakukan oleh (Syahrani, 2019) tentang Perbandingan antara Boosting (*XGBoost*) dan Bagging (*Random forest*) untuk pengelompokan Sekuens DNA Sambungan Splicing. Penelitian ini bertujuan untuk melihat bagaimana perbandingan efektivitas kedua teknik tersebut, dalam mengklasifikasikan kategori dalam penelitian ini, menunjukkan tingkat akurasi 96.24% untuk *XGBoost* dan 95.11% untuk *Random Forest*.

Berdasarkan beberapa hasil penelitian di atas, dapat disimpulkan bahwa penerapan algoritma *XGBoost* mampu menghasilkan nilai akurasi yang tinggi sehingga dapat menghasilkan prediksi yang akurat. Ini berarti bahwa *XGBoost* adalah pilihan yang tepat untuk memproses dan menganalisis data terkait penyakit diabetes mellitus. Algoritma *XGBoost* telah terbukti efisien dalam memprediksi data dengan tepat, memberikan hasil prediksi yang akurat dengan akurasi yang tinggi. Dengan demikian, temuan ini menunjukkan bahwa penerapan *XGBoost* dalam konteks ini dapat memberikan kontribusi yang berharga dalam upaya identifikasi dan penanganan penyakit diabetes mellitus.

## 2.2 Dasar Teori

### 2.2.1 Diabetes Mellitus

Diabetes Mellitus merupakan penyakit yang sangat sering dijumpai. Istilah diabetes berasal dari bahasa Yunani, "*diabainen*" yang berarti "mengalir terus-menerus". Sedangkan, "mellitus" berasal dari bahasa Latin "*millitus*" yang bermakna "madu". Kata itu dipilih untuk menggambarkan penyakit yang ditandai dengan tingginya kadar gula dalam urine. Namun, dalam bahasa Indonesia, DM sering dikenal sebagai "penyakit kencing manis" (Kusuma *dkk.*, 2024).

Menurut (Ridwan, 2020), diabetes terjadi ketika pankreas tidak menghasilkan insulin secara aktif. Insulin sangat berarti dalam mengelola arah gula darah dari makanan ke dalam sel darah untuk dijadikan energi. Karbohidrat dari makanan dipecah menjadi kadar gula dalam darah yang ditransfer oleh insulin ke dalam sel. Ketika produksi insulin tidak mencukupi, kadar gula dalam darah akan meningkat, seiring waktu dapat menimbulkan kerusakan pada organ dan mempengaruhi cara kerja organ serta jaringan tubuh. Biasanya seseorang yang berisiko diabetes memiliki gejala umum, meliputi frekuensi buang air kecil yang tinggi, rasa haus yang berlebihan, peningkatan nafsu makan, penurunan berat badan yang mendadak, kelelahan, obesitas, luka yang lambat sembuh, gangguan penglihatan, gatal-gatal dan lain-lain.

Sedangkan menurut (Abdurrahman *dkk.*, 2022), *American Diabetes Association* (ADA) merupakan kondisi medis yang memengaruhi bagaimana tubuh memproses glukosa, yang pada akhirnya dapat menyebabkan kadar gula darah menjadi tidak normal. Setiap 21 detik ada seorang pasien diabetes baru, menunjukkan tingkat kejadian yang signifikan. Penyebab kematian prematur di seluruh dunia salah satunya disebabkan oleh diabetes, juga menjadi pemicu terbesar dalam masalah kesehatan seperti kebutaan, penyakit jantung, dan gagal ginjal.

Penyakit ini adalah salah satu kondisi kesehatan yang berpotensi mengancam nyawa, namun sebagian besar dari mereka tidak menyadari adanya penyakit ini. Seseorang yang terkena penyakit ini sering tidak melakukan pemeriksaan atau pengobatan untuk menangani DM. Banyak faktor yang

menyebabkan siapa saja yang berisiko terkena penyakit ini. Banyak orang beranggapan bahwa diabetes mellitus hanya dialami oleh lansia, padahal kenyataannya, semua orang dari golongan muda sampai tua pun sangat bisa terkena penyakit ini. DM dapat berakibat fatal dengan merusak pembuluh darah dalam darah. Umumnya, kerusakan tersebut terjadi sebelum seseorang menyadari atau memeriksa kondisinya. Penyakit ini kerap berkaitan dengan kondisi “*silent killer*”, seperti penyakit jantung, hipertensi dan kadar kolesterol yang tinggi, yang mana semua ini bisa dipengaruhi oleh gaya hidup seseorang (Murtiningsih *dkk.*, 2021).

### 2.2.2 Klasifikasi

Klasifikasi termasuk algoritma pembelajaran mesin dalam kategori *supervised learning* (pembelajaran yang diawasi), di mana algoritma ini menggunakan data pelatihan yang telah diberi hasil dengan jawaban yang benar (Rahman, 2020). Biasanya setiap objek di kelompokkan pada setiap kategori yang sudah ditentukan sebelumnya berdasarkan atribut yang dimiliki (Maulana *dkk.*, 2024). Prosesnya dimulai ketika data latih dan data uji dibuat sebagai aturan klasifikasi (Herni Yulianti *dkk.*, 2022).

Klasifikasi juga mengarah pada proses pengelompokan fitur ke dalam kelas yang tepat. Vektor fitur pelatihan dengan kelas yang telah diketahui kemudian digunakan untuk membangun pembatas, proses ini juga dikenal sebagai pembelajaran terarah. Banyak teknik klasifikasi telah dicoba oleh peneliti, dan secara umum metode klasifikasi dapat dikelompokkan menjadi beberapa kategori berbeda yang dikenal dengan pengklasifikasian berbasis keputusan (Rabbani *dkk.*, 2023).

### 2.2.3 Data Pre Processing

Adalah proses mempelajari data lebih lanjut mengenai karakteristik data yang digunakan serta menyelesaikan dan membersihkan masalah pada data, sehingga data siap digunakan untuk pemodelan menggunakan *extreme gradient*

*boosting* (Setio *dkk.*, 2020). Berikut merupakan tahapan *preprocessing* yang digunakan dalam penelitian ini untuk membangun model XGBoost :

#### 1. Label *Encoding*

Dalam dataset yang digunakan, terdapat fitur dengan data *object*. *Encoding* berguna untuk mengonversi fitur data *object* menjadi data numerik/biner (pelabelan). Seperti pada fitur dalam dataset yang digunakan memiliki dua kategori, yaitu *yes* dan *no* diubah menjadi 1 dan 0. begitupun pada kelas target "*class*" yaitu positif dan negatif diubah menjadi 1 dan 0 (Ubaidillah *dkk.*, 2022).

#### 2. Memisahkan Fitur dan Target

Pemisahan atribut merupakan langkah krusial dalam *preprocessing* data. Fitur adalah variabel independen yang digunakan untuk memprediksi hasil, sedangkan target adalah variabel dependen yang ingin diprediksi oleh model (Fajar *dkk.*, 2024).

#### 3. Standarisasi Data

Standarisasi merupakan proses mengatur nilai dalam dataset yang digunakan dalam penelitian agar berada dalam skala yang sama, dengan mengubah data menjadi nilai dengan mean = 0 dan standar deviasi = 1. Tujuannya untuk memastikan bahwa semua fitur atau variabel dalam dataset memberikan kontribusi yang seimbang dalam model machine learning, terutama ketika variabel memiliki skala yang berbeda-beda (Maulana *dkk.*, 2023).

#### 4. *Split* Data

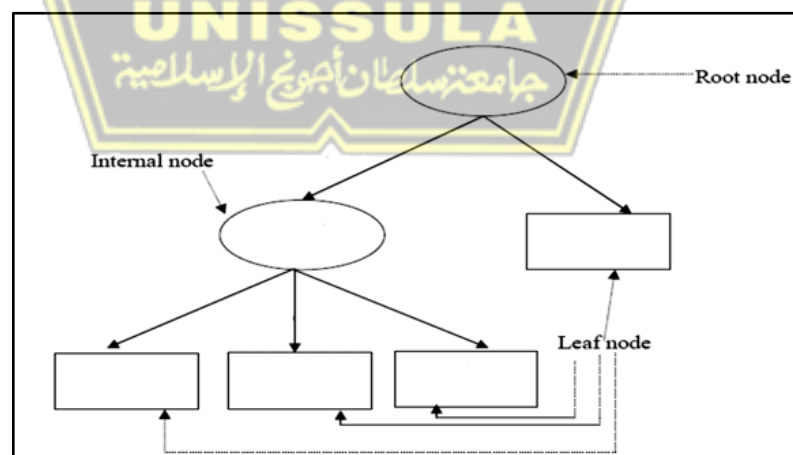
Dataset yang digunakan akan dibagi menjadi dua, yaitu data pelatihan untuk melatih model dan data pengujian untuk mengukur kinerja model. Pembagian ini dilakukan dengan proporsi 80% (*training*) banding 20% (*testing*), (Ubaidillah *dkk.*, 2022).

### 2.2.4 *Decision Tree* (Pohon Keputusan)

Secara garis besar, *Decision Tree* ialah teknik pengambilan keputusan yang mengorganisir setiap pilihan ke dalam struktur bercabang. Menurut (Blockeel *dkk.*,

2023), Pohon Keputusan merupakan metode populer dan efektif dalam bidang klasifikasi dan prediksi yang mengubah data menjadi pohon aturan keputusan yang mudah dipahami. Aturan ini bisa diekspresikan dengan mencari atau memfilter data sesuai kriteria yang dihasilkan dari model klasifikasi. Pohon keputusan membuat kumpulan data menjadi lebih kecil dari kumpulan data besar. Dimana masing-masing simpul daun mewakili label kelas. Simpul akar dan internal berisi kondisi tes atribut, biasanya berbentuk oval, sedangkan simpul daun berbentuk persegi.

Pohon keputusan mengklasifikasi data dengan mengajukan pertanyaan yang berhubungan dengan karakteristik setiap fitur dalam data. Setiap pertanyaan diwakili oleh simpul internal yang terhubung ke anak simpul yang mewakili jawaban dari pertanyaan. Dengan cara ini, pertanyaan-pertanyaan tersebut membentuk struktur hierarki dalam bentuk pohon. Pohon keputusan dibangun dengan menambahkan simpul pertanyaan secara bertahap, dipandu oleh data pelatihan berlabel. Pohon keputusan secara alami menyelesaikan masalah dengan memecahnya menjadi bagian yang lebih kecil, digunakan untuk membagi data menjadi subnet yang lebih spesifik dengan terus-menerus mengajukan pertanyaan hingga mencapai kondisi di mana pembagian lebih lanjut tidak diperlukan lagi (Charbuty *dkk.*, 2021).



Gambar 2. 1 Struktur pohon keputusan (Husaini *dkk.*, 2023)

Gambar 2. 1 merupakan rangkaian pohon keputusan, di mana setiap node internal menunjukkan evaluasi terhadap suatu atribut, setiap bagian

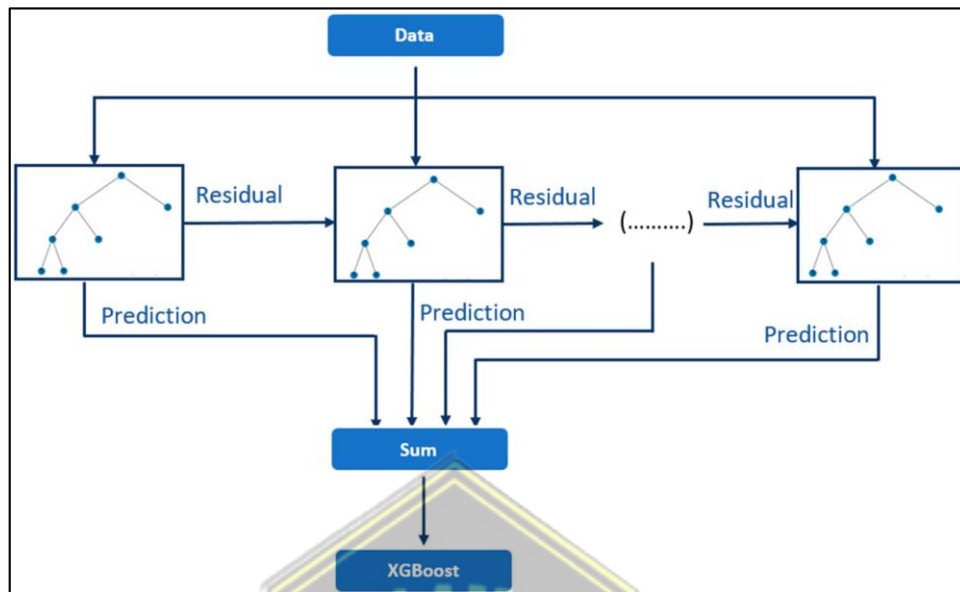


mengindikasikan hasil evaluasi tersebut dan leaf node menunjukkan distribusi kelas. Root node dikenal sebagai node paling atas atau node akar, node ini memiliki beberapa cabang keluar tanpa cabang masuk. Sedangkan internal node memiliki satu alur masuk dan beberapa alur keluar. Sebaliknya, leaf node hanya memiliki satu alur masuk dan tanpa alur keluar. Lalu pohon ini digunakan untuk mengelompokkan data yang belum memiliki kategori ke dalam kelas yang sudah ada.

### 2.2.5 *Extreme Gradient Boosting (XGBoost)*

XGBoost atau *Extreme Gradient Boosting*, diperkenalkan pada tahun 2011 oleh Tianqi Chen dan Carlos Guestrin (Purbolingga *dkk.*, 2023). Sejak itu, banyak peneliti mulai mengembangkan model ini dengan memodifikasi dan kemudian disempurnakan. Algoritma ini adalah hasil pengembangan dari *Gradient Boosting Decision Tree* yang dikenalkan oleh Friedman (Pinata *dkk.*, 2020). Algoritma ini sangat unggul dalam membuat model pohon keputusan untuk menyelesaikan masalah regresi dan klasifikasi dengan mengandalkan *Gradient Boosting Decision Tree* (GBDT). Dalam prosesnya, pohon yang dibangun akan diperbaiki oleh pohon sebelumnya dan berikutnya sehingga menciptakan ketergantungan antar pohon, XGBoost juga mengatur bobot pada setiap pohon yang di bangun hingga membentuk pohon yang tangguh dan dapat diandalkan (Maulana *dkk.*, 2024) .

XGBoost adalah algoritma *ensemble* yang terdiri dari beberapa beberapa pohon, berikut adalah gambaran dari arsitektur *XGBoost* :



Gambar 2. 2 Arsitektur *XGBoost* (Mehdary dkk., 2024)

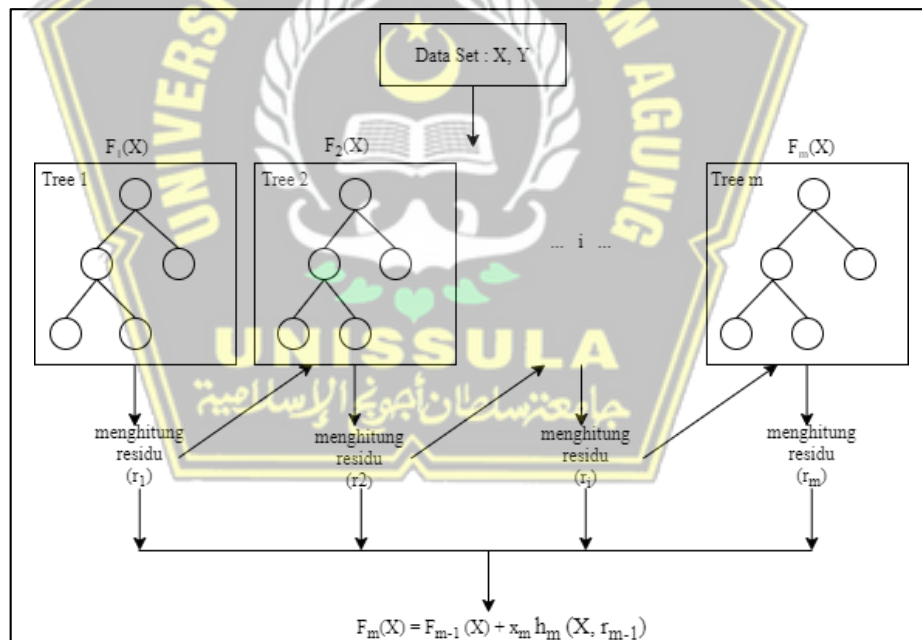
Pada gambar 2. 2 di atas merupakan arsitektur *XGBoost* yang menjelaskan tentang penerapan teknik *ensemble* dari beberapa pohon keputusan dalam algoritma *XGBoost*. Teknik *ensemble* adalah menggabungkan beberapa model pohon keputusan untuk membuat prediksi yang lebih kuat dan lebih akurat dibandingkan dengan menggunakan model individu. Pendekatan ini meningkatkan performa model dengan mengurangi bias dan varians.

Kemudian teknik *boosting* melibatkan pelatihan model secara berurutan. Dalam *XGBoost*, setiap model baru yang dilatih untuk memperbaiki kesalahan yang dibuat oleh model sebelumnya. Model-model ini bekerja bersama-sama, dengan setiap model bertujuan untuk mengurangi kesalahan yang masih ada setelah model sebelumnya diterapkan. pendekatan ini secara iterative meningkatkan akurasi prediksi dengan memfokuskan pelatihan pada contoh-contoh yang sulit diprediksi oleh model sebelumnya.

Lalu teknik *regularization* digunakan untuk mengurangi *overfitting*, yaitu ketika model menjadi sangat rumit dan terlalu sesuai dengan data pelatihan, performanya cenderung menurun saat diterapkan pada data yang belum pernah dilihat sebelumnya. *XGBoost* menambah regularisasi dalam proses pelatihannya, yang membantu menjaga keseimbangan antara kompleksitas model dan kemampuannya untuk regeralisasi (Mohiuddin dkk., 2023). pendekatan ini

membantu memastikan bahwa model tidak sekedar terampil pada data pelatihan tapi juga kompeten dalam menghasilkan prediksi yang tepat walaupun belum pernah dilihat sama sekali, sehingga meningkatkan performa model secara keseluruhan.

XGBoost bekerja dengan setiap pohon regresi yang memetakan titik data masukan ke salah satu daunnya yang memiliki skor berkelanjutan. XGBoost meminimalkan fungsi tujuan regularisasi (L1 dan L2) yang menggabungkan fungsi kerugian cembung (berdasarkan perbedaan antara keluaran prediksi dan target). Pelatihan terjadi secara iteratif dengan menambahkan pohon baru yang memprediksi kesalahan dari pohon sebelumnya yang kemudian digabungkan dengan pohon sebelumnya untuk menghasilkan prediksi akhir. Disebut peningkatan *gradient* karena menggunakan algoritma penurunan *gradient* untuk meminimalkan kerugian saat menambahkan model baru.



Gambar 2. 3 Cara kerja XGBoost (AWS, 2024)

Gambar 2.3 di atas merupakan gambaran cara kerja xgboost, dimana  $\alpha_i$  adalah parameter regularisasi dan  $r_i$  adalah residu dihitung dengan pohon ke-I, dan  $h_i$  adalah fungsi yang dilatih untuk memprediksi residu,  $r_i$  menggunakan  $X$  untuk pohon ke-i. Untuk menghitung  $\alpha_i$  menggunakan residu yang dihitung  $r_i$  dan hitung dalam rumus berikut :

$$\arg \min_{\alpha} = \sum_{i=1}^m L(Y_i, F_{i-1}(X_i) + \alpha h_i(X_i, r_{i-1})) \quad (1)$$

$L(Y, F(X))$  adalah fungsi kerugian terdiferensiasi

Keterangan :

- Kumpulan data  $(X, Y)$  : proses dimulai dengan kumpulan data yang terdiri dari fitur  $X$  dan target  $Y$ .
- Pohon 1 ( $F_1(X)$ ) : model pertama dilatih menggunakan kumpulan data. setelah pohon dibuat, menghitung  $r_1$ , yaitu selisih antara prediksi pohon 1 dengan nilai sebenarnya  $Y$ .
- Hitung  $\alpha_1$  : nilai  $\alpha_1$  dihitung, merupakan parameter regularisasi yang membantu mengurangi overfitting.
- pohon 2 ( $F_2(X)$ ) : Model kedua, dilatih untuk memprediksi residu  $r_1$ . Setelah pohon 2 dilatih, lalu menghitung sisa  $r_2$ , yaitu selisih antara prediksi pohon 2 dengan sisa  $r_1$ .
- Hitung  $\alpha_2$  : nilai  $\alpha_2$  dihitung untuk pohon 2.
- Iterasi proses : proses ini diulangi untuk banyak pohon (dari pohon 3 hingga pohon  $m$ ). Setiap pohon baru dilatih untuk memprediksi residu dari pohon sebelumnya, dan residu baru dihitung setiap saat. Nilai  $\alpha_i$  juga dihitung setiap kali untuk setiap pohon  $i$ .
- Model Akhir  $F_m(X)$  : kombinasi dari setiap pohon yang telah dilatih. Kombinasi ini ditulis sebagai :

$$F_m(X) = F_{m-1}(X) + \alpha_m h_m(X, r_{m-1}) \quad (2)$$

Dimana  $h_i$  merupakan fungsi yang dilatih untuk memprediksi residu  $r_i$  pada pohon ke- $i$ .

- Mengoptimalkan  $\alpha$  : untuk menghitung nilai  $\alpha$ , kemudian meminimalkan fungsi kerugian terdiferensiasi  $L$ , menggunakan rumus berikut :

$$\arg \min_{\alpha} \sum_{i=1}^m L(Y_i, F_{i-1}(X_i) + \alpha h_i(X_i, r_{i-1})) \quad (3)$$

Keterangan :

$L$  = Fungsi kerugian

$Y_i$  = Nilai target sebenarnya (0 atau 1)

$F_{i-1}(X_i)$  = Prediksi model pada iterasi sebelumnya

$\alpha$  = Koefisien yang menentukan seberapa besar kontribusi model baru  $h_i$

$h_i(X_i, r_{i-1})$  = Model baru yang ditambahkan pada iterasi ke- $i$

$r_{i-1}$  = Residual atau kesalahan pada iterasi sebelumnya

Penggunaan teknologi seperti algoritma pembelajaran mesin dalam memprediksi penyakit diabetes juga berkembang pesat sehingga penting untuk menyadari apakah seseorang termasuk dalam kategori pasien yang memiliki penyakit diabetes sebelum penderita Diabetes Mellitus mengalami komplikasi serius, sehingga mereka dapat menjalani pengobatan yang sesuai dan tepat waktu. Algoritma XGBoost telah populer digunakan untuk prediksi sebuah penyakit. Maka dari itu, dalam penulisan laporan ini akan memanfaatkan algoritma XGBoost sebagai salah satu implementasi untuk menghasilkan prediksi risiko penyakit Diabetes Mellitus. Sistem ini diharapkan dapat bermanfaat bagi petugas kesehatan serta membantu meningkatkan layanan rumah sakit kepada pasien, terutama dalam deteksi dini khususnya penyakit diabetes mellitus dan personalisasi perawatannya.

### 2.2.6 Evaluasi Model

Model yang akan dievaluasi dalam penelitian ini adalah menggunakan *matrix confusion*, merupakan salah satu cara untuk mengukur suatu kinerja model dalam masalah klasifikasi (Rombe, 2021), dimana hasil klasifikasi bisa berada dalam dua kelas atau lebih. Evaluasi menggunakan *Confusion matrix* membantu dalam mengidentifikasi bagaimana model memprediksi kondisi diabetes (positif dan negatif) secara akurat (Yusufi dkk., 2022).

Nilai evaluasi yang akan dipertimbangkan meliputi nilai akurasi, presisi, recall, dan f1-score, Berikut merupakan penjelasan dan rumus yang digunakan untuk mengetahui ke-4 nilai tersebut (Paramitha dkk., 2023) :

Nilai akurasi adalah perbandingan antara jumlah data yang berhasil diidentifikasi dengan benar terhadap total data yang diuji. Model melakukan evaluasi berdasarkan jumlah data yang diprediksi dengan benar. nilai ini dapat dihitung menggunakan rumus 4.

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)} \quad (4)$$

Nilai presisi merupakan perbandingan antara prediksi benar dan total jumlah prediksi yang dibuat. Nilai presisi dapat dihitung menggunakan rumus 5.

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

Nilai *recall* merupakan perbandingan seberapa banyak prediksi yang benar dibandingkan dengan seluruh data yang sebenarnya, nilai ini dapat dihitung menggunakan rumus 6.

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

Nilai *F1-score* adalah menggabungkan nilai *presisi* dan *recall* dalam satu metric, ini sangat berguna untuk menghadapi data yang tidak seimbang, seperti jumlah target pada kategori positif dan negatif yang sangat berbeda. Nilai ini dapat dihitung menggunakan rumus 7.

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \quad (7)$$

Keterangan :

*True Positive* (TP) : prediksi positif yang benar

*True Negative* (TN) : prediksi negatif yang benar

*False Positive* (FP) : prediksi positif tapi negatif

*False Negative* (FN) : prediksi negatif tapi positif

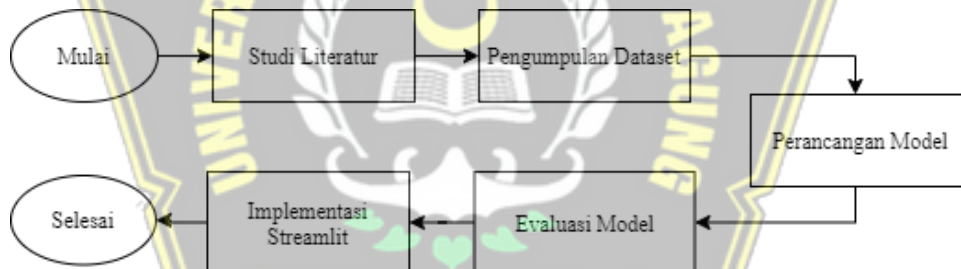
## BAB III

### METODE PENELITIAN

Penelitian ini bermaksud untuk menganalisis teori bahwa sistem yang dirancang mampu memprediksi risiko penyakit diabetes mellitus berdasarkan riwayat kesehatan yang dialami dengan tingkat akurasi yang bagus. Analisis ini menggunakan data hasil dari pasien yang memiliki risiko penyakit diabetes dan yang tidak, yang kemudian dianalisis menggunakan *Extreme Gradient Bossting* (XGBoost).

#### 3.1 Metode Penelitian

XGBoost (*Extreme Gradient Boosting*) dipakai sebagai algoritma dalam Penelitian ini, yang dikembangkan untuk membuat model prediksi sehingga bisa mendiagnosis risiko diabetes mellitus secara akurat.



Gambar 3. 1 Alur Penelitian

Gambar 3.1 adalah rancangan yang akan diimplementasikan dalam penelitian ini. Mulai dari mencari informasi seputar penelitian yang diteliti sampai menghasilkan sistem prediksi risiko diabetes.

##### 3.1.1 Studi Literatur

Peneliti melakukan analisis sumber secara mendalam. Termasuk menelaah berbagai sumber seperti makalah, jurnal, artikel, tesis, skripsi, serta melihat dan membaca banyak situs web. Melalui proses ini, peneliti memahami perkembangan terkini dalam bidang yang sedang diteliti, teknik yang digunakan, dan metodologi yang umum diterapkan oleh peneliti lain. analisis sumber ini juga berguna untuk mendeteksi kekurangan dalam penelitian sebelumnya yang dapat diperbaiki dalam

penelitian ini. Tujuannya untuk memahami teori yang mendasari bagaimana model mengklasifikasikan risiko diabetes mellitus berdasarkan riwayat kesehatan menggunakan XGBoost.

### 3.1.2 Pengumpulan Data

Dataset ini bernama *Early Stage Diabetes Risk Prediction Diabetes* dan diunduh dari kaggle<sup>1</sup>, data yang digunakan dalam penelitian ini adalah data sekunder yang diperoleh dari kuesioner yang diisi oleh pasien di rumah sakit *Sylhet Diabetic Hospital* di *Sylhet, Bangladesh* dan telah mendapat persetujuan dari seorang dokter. Penelitian ini melibatkan sebanyak 520 data, terdiri dari 320 data untuk kategori positif diabetes mellitus dan 200 untuk kategori non-diabetes mellitus. Dataset ini mencakup berbagai fitur yang terkait dengan gejala medis yang berkaitan dengan diabetes, di antaranya :

Tabel3. 1 Atribut dalam dataset

No	Nama Atribut	Kategori	Keterangan
1	<i>Age</i>	Tahun	Umur
2	<i>Gender</i>	<i>Male/Female</i>	Jenis kelamin seseorang
3	<i>Polyuria</i>	<i>Yes/No</i>	Keadaan di mana seseorang mengalami frekuensi buang air kecil yang melampaui batas normal
4	<i>Polydipsia</i>	<i>Yes/No</i>	Keadaan di mana seseorang mengalami rasa haus yang berlebihan atau tidak
5	<i>Sudden Weigh Loss</i>	<i>Yes/No</i>	Keadaan di mana seseorang berat badannya turun secara mendadak
6	<i>Weakness</i>	<i>Yes/No</i>	Keadaan di mana seseorang sering kelelahan yang tidak wajar
7	<i>Polyphagia</i>	<i>Yes/No</i>	Keadaan di mana seseorang memiliki nafsu makan diluar batas normal

<sup>1</sup> <https://www.kaggle.com/datasets/ishandutta/early-stage-diabetes-risk-prediction-dataset>



No	Nama Atribut	Kategori	Keterangan
8	<i>Genital Trush</i>	<i>Yes/No</i>	Keadaan di mana seseorang terjangkit infeksi jamur di area genital (kemaluan)
9	<i>Visual Blurring</i>	<i>Yes/No</i>	Keadaan di mana seseorang mengalami gangguan penglihatan yang tampak buram
10	<i>Itching</i>	<i>Yes/No</i>	Keadaan di mana seseorang merasakan gatal pada kulit
11	<i>Irritability</i>	<i>Yes/No</i>	Keadaan di mana seseorang menjadi mudah tersinggung atau emosional
12	<i>Delayed Healing</i>	<i>Yes/No</i>	Keadaan di mana seseorang mengalami proses penyembuhan luka yang lambat
13	<i>Partial Paresis</i>	<i>Yes/No</i>	Keadaan di mana seseorang mengalami kehilangan kekuatan pada sebagian otot
14	<i>Muscle Stiffness</i>	<i>Yes/No</i>	Keadaan di mana seseorang merasakan kekakuan pada otot-otot
15	<i>Alopecia</i>	<i>Yes/No</i>	Keadaan di mana seseorang mengalami kerontokan rambut yang berlebihan
16	<i>Obesity</i>	<i>Yes/No</i>	Keadaan di mana seseorang memiliki berat badan melebihi normal (kegemukan/obesitas)
17	<i>Class (Hasil)</i>	<i>Positif/Negatif</i>	Menunjukkan hasil positif apabila seseorang berisiko diabetes dan negatif untuk seseorang yang tidak berisiko diabetes

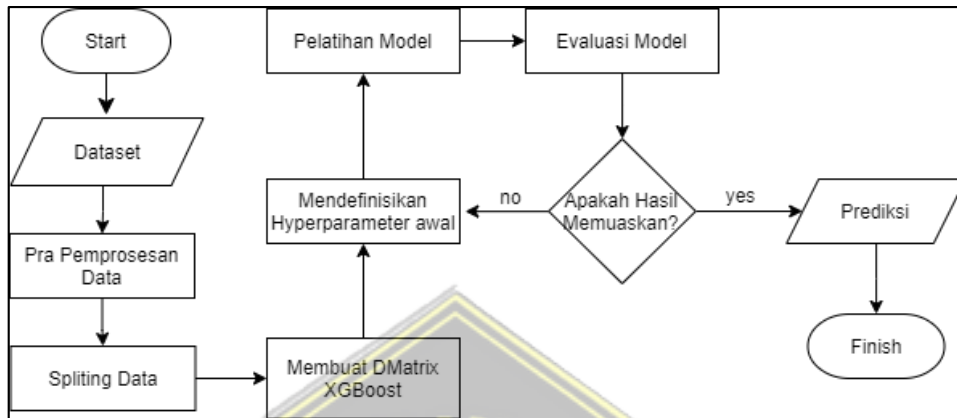
Tabel 3.1 merupakan jenis fitur yang ada dalam dataset “*Early Stage Diabetes Risk Prediction Diabetes*” yaitu ada enam belas fitur dengan satu target (*class*). Tabel di atas merupakan penjelasan dari setiap fitur dalam dataset.

Untuk mengetahui nilai *mean* atau nilai rata-rata suatu data dapat diperoleh dengan menjumlahkan seluruh nilai dan membaginya dengan jumlah total data. Sebagai contoh, *mean* untuk fitur umur yaitu 48.03. *Median* atau nilai tengah dalam data yang telah diurutkan, adalah nilai yang berada di posisi tengah jika jumlah data ganjil. Jika jumlah data genap, *median* adalah rata-rata dari dua nilai tengah. Seperti *median* pada fitur umur yaitu 47.5. Sedangkan standar deviasi untuk mengukur seberapa tersebar nilai-nilai dalam setiap fitur dataset. Seperti standar deviasi pada fitur umur yaitu 12.15. Dengan distribusi dari setiap kategori positif sebanyak 320 dan negatif sebanyak 200.



### 3.1.3 Pemodelan Sistem

Desain pemodelan sistem dapat dilihat melalui bentuk *flowchart* pada gambar 3. 2 di bawah :



Gambar 3. 2 *Flowchart* pemodelan sistem

Sistem pemodelan ini diproses dengan membagi menjadi 3 tahap yaitu :

- 1) Dataset dengan format file CSV di eksplor untuk dimuat. Dataset ini ada 17 fitur yang akan dipakai dalam pemodelan. Dataset ini diimpor menggunakan pustaka seperti *pandas*.
- 2) Data *Preprocessing*, proses ini mempersiapkan data untuk diolah dalam pemodelan dengan algoritma XGBoost. Langkah awal melakukan label encoding, proses mengubah fitur data yang aslinya teks menjadi numerik, agar bisa dikelola oleh algoritma karena membutuhkan input data numerik untuk melakukan kalkulasi. Kemudian, fitur dan target dipisahkan untuk memastikan bahwa pelatihan model hanya melibatkan fitur-fitur yang saling berkaitan dan menggunakan target untuk membuat prediksi. Selanjutnya, langkah terakhir dalam tahap *preprocessing* adalah standarisasi fitur untuk melihat semua fitur memiliki skala yang beragam, karena algoritma XGBoost sensitif. Data dibagi menjadi dua bagian, data *train* untuk melatih model dan data *test* untuk menjamin evaluasi model dilakukan dengan optimal. Lalu, setelah data dibagi menjadi 2, format data diubah menjadi format Dmatrix yang disempurnakan untuk menaikkan performa. Langkah-langkah tersebut kan menghasilkan dataset yang bersih, terstruktur dan teroptimalkan,

sehingga siap digunakan ke tahap berikutnya, yaitu pemodelan XGBoost yang efektif dan akurat.

- 3) Proses pemodelan, dilakukan setelah memastikan data dalam kondisi siap untuk pelatihan model XGBoost. Pertama, melakukan pencarian *hyperparameter* menggunakan *grid search*, terdapat 3 *hyperparameter* yang digunakan untuk melatih model XGBoost, yaitu *max\_depth* (kedalaman maksimum pohon), *learning\_rate* (tingkat pembelajaran), *n\_estimators* (jumlah pohon keputusan/ model yang dibangun). Proses ini dilakukan dengan *5-fold cross-validation*, di mana model dievaluasi untuk setiap kombinasi *hyperparameter* untuk mendapatkan kombinasi parameter terbaik yang menghasilkan performa model paling optimal. Lalu, untuk memaksimalkan performa model, model dilatih dengan data pelatihan dan di evaluasi menggunakan data pengujian. Proses ini melibatkan iterasi berulang selama 5 kali untuk meminimalkan kesalahan prediksi dan *overfitting* dapat diminimalkan sehingga model lebih andal dalam membuat prediksi pada data yang tidak terlihat sama sekali selama proses pelatihan model. Setelah proses pelatihan dan optimasi, model disimpan untuk digunakan di masa mendatang dan dapat diterapkan dalam aplikasi nyata untuk prediksi risiko diabetes secara otomatis. Dengan pendekatan ini, dapat dibangun model prediksi risiko diabetes yang akurat dan handal menggunakan algoritma XGBoost.

#### 3.1.4 Evaluasi Model

Langkah ini, mengukur akurasi model menggunakan confusion matrix. Teknik ini dapat mengetahui nilai prediksi benar dan salah dalam kategori ‘*True Positive*’, ‘*True Negative*’, ‘*False Positive*’, dan ‘*False Negative*’. Hal ini memungkinkan diukur melalui metrik akurasi, presisi, recall dan F1-score. Nilai-nilai inilah yang nantinya akan menghasilkan jawaban mendalam mengenai cara model mengatasi berbagai masalah dan membantu mencari akar permasalahan yang perlu diperbaiki sampai memastikan kinerja model berfungsi dengan optimal.

### 3.1.5 Implementasi Streamlit

Antarmuka web dikembangkan melalui *streamlit*, merupakan library python yang menyediakan pembuatan antarmuka pengguna dengan mudah dan sederhana dalam penerapannya. Sedangkan untuk pengujian sistem hanya memprioritaskan *input* dan *output* pada sistem untuk memastikan sistem yang dikembangkan berfungsi sesuai dengan spesifikasi yang telah ditentukan. Jadi penelitian ini hanya diuji secara fungsionalitas saja tanpa meninjau struktur dalam kode program.



## BAB IV

### HASIL DAN ANALISIS PENELITIAN

#### 4.1 Hasil Pemodelan

Pemodelan sistem menggunakan algoritma XGBoost telah berhasil diterapkan dalam penelitian ini untuk memprediksi risiko diabetes berdasarkan riwayat kesehatan. Model ini dipilih karena keunggulannya dalam menangani data yang kompleks dan distribusi data yang tidak seimbang. Dari hasil pemodelan menunjukkan performa yang baik. Penjelasan lebih detail mengenai kinerja model serta interpretasi akan dibahas dan ditunjukkan melalui tangkap layar dibawah :

##### 4.1.1 Memuat Data

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 520 entries, 0 to 519
Data columns (total 17 columns):
#   Column                                Non-Null Count  Dtype
---  ---                                -
0   Age                                    520 non-null    int64
1   Gender                                520 non-null    object
2   Polyuria                               520 non-null    object
3   Polydipsia                             520 non-null    object
4   sudden weight loss                     520 non-null    object
5   weakness                               520 non-null    object
6   Polyphagia                             520 non-null    object
7   Genital thrush                         520 non-null    object
8   visual blurring                        520 non-null    object
9   Itching                                520 non-null    object
10  Irritability                           520 non-null    object
11  delayed healing                        520 non-null    object
12  partial paresis                        520 non-null    object
13  muscle stiffness                       520 non-null    object
14  Alopecia                               520 non-null    object
15  Obesity                                520 non-null    object
16  class                                  520 non-null    object
dtypes: int64(1), object(16)
memory usage: 69.2+ KB

class
Positive    320
Negative    200
Name: count, dtype: int64
```

Gambar 4. 1 *Output* dari dataset

Gambar 4.1 merupakan hasil dari memuat dataset berformat CSV yang digunakan dalam pemodelan menggunakan algoritma XGBoost. gambar di atas merupakan hasil dari perintah untuk menampilkan jumlah data yaitu sebanyak 520 data, jumlah atribut/fitur sebanyak 16 fitur dengan 1 target “class”, menampilkan atribut/fitur yang ada dalam dataset, serta distribusi kelas target “class” (*positive/negative*) yang mana pada kategori positif ada 320 data dan negative 200 data.

#### 4.1.2 Preprocessing

Proses ini mempersiapkan data untuk diolah dalam pemodelan dengan algoritma XGBoost. Langkah awal melakukan pelabelan yaitu proses mengubah fitur data yang aslinya teks menjadi numerik. Kedua, fitur dan target dipisah untuk memastikan bahwa model dilatih hanya pada fitur yang relevan dan menggunakan target untuk membuat prediksi. Selanjutnya, langkah terakhir dalam *preprocessing* adalah standarisasi data untuk menyesuaikan data agar semua fitur berada dalam rentang yang seragam. Berikut langkah lebih jelasnya dari tahapan preprocessing yang dibutuhkan dalam penelitian ini:

##### 1. Label Encoding

Langkah ini mencakup konversi semua data yang teks menjadi data numerik menggunakan teknik pelabelan, agar algoritma pembelajaran mesin bisa memproses data untuk melakukan perhitungan.

Data Sebelum Encoding:

	Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing	partial paresis	muscle stiffness	Alopecia	Obesity	class
0	40	Male	No	Yes	No	Yes	No	No	No	Yes	No	Yes	No	Yes	Yes	Yes	Positive
1	58	Male	No	No	No	Yes	No	No	Yes	No	No	No	Yes	No	Yes	No	Positive
2	41	Male	Yes	No	No	Yes	Yes	No	No	Yes	No	Yes	No	Yes	Yes	No	Positive
3	45	Male	No	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes	No	No	No	No	Positive
4	60	Male	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Positive

Gambar 4. 2 Output sebelum di encoding data

Gambar 4.2 adalah hasil dari menampilkan baris pertama dataset sebelum dilakukan label encoding, gambar di atas adalah bentuk dataset asli yang digunakan untuk penelitian ini.

Data Setelah Encoding:

	Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing	partial paresis	muscle stiffness	Alopecia	Obesity	class
0	40	1	0	1	0	1	0	0	0	1	0	1	0	1	1	1	1
1	58	1	0	0	0	1	0	0	1	0	0	0	1	0	1	0	1
2	41	1	1	0	0	1	1	0	0	1	0	1	0	1	1	0	1
3	45	1	0	0	1	1	1	1	0	1	0	1	0	0	0	0	1
4	60	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1

Gambar 4. 3 Output setelah di encoding data

Sedangkan pada gambar 4.3 merupakan hasil dari menampilkan pengolahan dataset dengan fokus pada beberapa baris pertama dari dataset yang digunakan untuk melihat bagaimana nilai-nilai kategori diubah menjadi numerik.

## 2. Memisahkan fitur dan target

Menyiapkan fitur secara terpisah dari target guna memastikan bahwa pelatihan model hanya melibatkan fitur yang relevan, sedangkan target digunakan menghasilkan prediksi.

```

Fitur (x):
  Age  Gender  Polyuria  Polydipsia  sudden weight loss  weakness  \
0  40     1         0         1         0         1
1  58     1         0         0         0         1
2  41     1         1         0         0         1
3  45     1         0         0         1         1
4  60     1         1         1         1         1

  Polyphagia  Genital thrush  visual blurring  Itching  Irritability  \
0           0           0           0         1           0
1           0           0           1         0           0
2           1           0           0         1           0
3           1           1           0         1           0
4           1           0           1         1           1

  delayed healing  partial paresis  muscle stiffness  Alopecia  Obesity
0                1                0                1        1        1
1                0                1                0        1        0
2                1                0                1        1        0
3                1                0                0        0        0
4                1                1                1        1        1

Target (y):
0  1
1  1
2  1
3  1
4  1
Name: class, dtype: int64

```

Gambar 4. 4 Output dari memisahkan fitur dan target

Gambar 4. 4 adalah hasil pemisahan fitur dan target pada dataset. Fitur adalah variabel input yang digunakan oleh model untuk membuat prediksi, sedangkan target adalah variabel hasil yang ingin di prediksi. Pembagian



fitur dalam dataset yang dipisah mulai dari fitur *Age* (umur) sampai *Obesity* (obesitas) termasuk atribut sedangkan “*class*” merupakan target.

### 3. Standarisasi data

Data Setelah Standarisasi:																
	Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing	partial paresis	muscle stiffness	Alopecia	Obesity
0	-0.661367	0.765092	-0.992337	1.109847	-0.846269	0.839594	-0.915126	-0.535844	-0.901025	1.027295	-0.565506	1.084312	-0.869918	1.290994	1.380227	2.215647
1	0.821362	0.765092	-0.992337	-0.901025	-0.846269	0.839594	-0.915126	-0.535844	1.109847	-0.973430	-0.565506	-0.922244	1.149534	-0.774597	1.380227	-0.451335
2	-0.578993	0.765092	1.007722	-0.901025	-0.846269	0.839594	1.092746	-0.535844	-0.901025	1.027295	-0.565506	1.084312	-0.869918	1.290994	1.380227	-0.451335
3	-0.249498	0.765092	-0.992337	-0.901025	1.181657	0.839594	1.092746	1.866215	-0.901025	1.027295	-0.565506	1.084312	-0.869918	-0.774597	-0.734518	-0.451335
4	0.986110	0.765092	1.007722	1.109847	1.181657	0.839594	1.092746	-0.535844	1.109847	1.027295	1.768328	1.084312	1.149534	1.290994	1.380227	2.215647

Gambar 4. 5 Output standarisasi data

Gambar 4.5 merupakan hasil dari standarisasi data menggunakan *scaler*, yaitu proses mengubah fitur-fitur sehingga memiliki rata-rata 0 dan deviasi standar 1. Standarisasi dilakukan agar fitur menjadi lebih konsisten dalam hal skala, sehingga memudahkan model dalam memproses data yang memiliki skala fitur beragam, hal tersebut membantu meningkatkan performa dan stabilitas model.

#### 4.1.3 Split Data

Proses pembagian data sangat penting dilakukan untuk menjamin model dilatih menggunakan data tertentu lalu diuji performanya menggunakan data asing yang sebelumnya belum pernah dilihat selama pelatihan. Data dibagi menjadi 2 untuk memastikan bahwa model yang dilatih dapat dievaluasi secara optimal.

Training data: (416, 16)  
Testing data: (104, 16)

Gambar 4. 6 Output pembagian data

Gambar 4. 6 di atas adalah hasil pembagian data dengan proporsi pembagian 80:20 yaitu data yang digunakan untuk melatih model adalah data pelatihan, sedangkan data pengujian digunakan untuk menilai performa model. Dari proses pembagian tersebut menghasilkan sebanyak 416 untuk data pelatihan dan data pengujian sebanyak 104.

#### 4.1.4 Format Dmatrix

Setelah proses pembagian data, formatnya dikonversi menjadi Dmatrix, merupakan format khusus untuk XGBoost. Format Dmatrix adalah struktur yang

dirancang secara khusus oleh XGBoost untuk mengarsip data. Format ini menawarkan efisiensi penyimpanan yang lebih tinggi dan memberikan kecepatan yang signifikan dalam memproses pelatihan dan prediksi saat menggunakan XGBoost.

```
DTrain Features:
['Age', 'Gender', 'Polyuria', 'Polydipsia', 'sudden weight loss', 'weakness', 'Polyphagia', 'Genital thrush', 'visual blurring', 'Itching', 'Irritability', 'delayed healing', 'partial paresis', 'muscle stiffness', 'Alopecia', 'Obesity']
DTrain Labels:
[[1. 1. 0. 0. 0. 0. 0. 1. 0. 0. 0. 1. 0. 0. 1. 0. 0. 1. 0. 1. 1. 1. 1. 1. 1. 1. 0.
 1. 0. 1. 1. 1. 1. 0. 1. 1. 1. 0. 1. 0. 0. 1. 1. 1. 1. 1. 0. 1. 1. 0. 1. 1. 0. 1.
 1. 0. 1. 0. 0. 1. 0. 0. 1. 0. 1. 0. 1. 1. 0. 1. 1. 0. 1. 0. 1. 0. 1. 1. 1.
 0. 1. 0. 0. 1. 1. 0. 1. 1. 0. 1. 1. 1. 0. 1. 1. 1. 0. 0. 1. 1. 1. 0. 0.
 1. 1. 1. 1. 1. 0. 0. 1. 1. 0. 1. 1. 1. 0. 1. 1. 1. 0. 0. 1. 1. 1. 0. 0.
 1. 1. 1. 1. 0. 0. 1. 0. 1. 0. 1. 0. 1. 0. 1. 0. 1. 0. 1. 0. 1. 1. 1.
 1. 0. 0. 1. 0. 1. 1. 0. 0. 0. 1. 0. 0. 0. 1. 0. 0. 0. 1. 1. 1. 1. 1.
 1. 0. 1. 0. 1. 0. 1. 1. 0. 1. 1. 0. 1. 0. 1. 0. 1. 0. 1. 0. 1. 1. 1.
 1. 1. 1. 0. 1. 1. 0. 0. 1. 1. 1. 1. 0. 1. 0. 1. 0. 1. 0. 1. 1. 1. 0.
 0. 1. 0. 1. 0. 1. 0. 1. 0. 1. 0. 0. 0. 1. 0. 0. 1. 1. 1. 0. 1. 0. 0.
 1. 1. 1. 0. 1. 0. 1. 0. 1. 1. 1. 0. 1. 0. 1. 0. 1. 0. 1. 0. 1. 1. 1.
 1. 1. 1. 1. 0. 0. 1. 1. 0. 0. 0. 1. 1. 1. 0. 0. 0. 1. 1. 1. 1. 1. 0.
 1. 0. 1. 0. 1. 1. 1. 0. 0. 1. 1. 1. 1. 0. 0. 0. 1. 0. 0. 1. 1. 0. 1. 0.
 1. 1. 1. 1. 1. 0. 1. 0. 1. 1. 1. 1. 1. 1. 1. 0. 1. 0. 0. 0. 1. 1. 1. 0.
 0. 0. 1. 0. 1. 0. 0. 1. 0. 1. 0. 1. 1. 0. 0. 1. 0. 1. 1. 1. 0. 1. 0. 0.
 1. 1. 1. 0. 1. 1. 1. 1. 0. 1. 1. 1. 0. 1. 1. 1. 0. 1. 0. 1. 0. 0. 1. 0.
 1. 1. 0. 0. 0. 0. 1. 1. 1. 0. 1. 1. 1. 1. 0. 1. 0. 1. 1. 0. 1. 1. 0. 1.
 1. 0. 1. 1. 0. 1. 1. 1. 0. 0. 0. 1. 1. 0. 1. 1. 1. 1. 1. 0. 1. 1. 0. 0.
 0. 1. 1. 1. 1. 1. 1. 1.]
```

Gambar 4. 7 Output mengubah format ke Dmatrix

Pada gambar 4. 7 adalah hasil dari proses modifikasi format menjadi dmatrix setelah proses *split* data. dibawah merupakan tabel dari hasil perubahan format ke dmatrix agar lebih jelas.

Tabel 4. 1 Hasil mengubah format ke dmatrix

<p>DTrain Features:                  ['Age', 'Gender', 'Polyuria', 'Polydipsia', 'sudden weight loss', 'weakness', 'Polyphagia', 'Genital thrush', 'visual blurring', 'Itching', 'Irritability', 'delayed healing', 'partial paresis', 'muscle stiffness', 'Alopecia', 'Obesity']</p>
<p>DTrain Labels:                  [[1. 1. 0. 0. 0. 0. 0. 1. 0. 0. 0. 1. 0. 0. 1. 0. 0. 1. 0. 1. 1. 1. 1. 1. 1. 1. 1. 0.                 1. 0. 1. 1. 1. 1. 0. 1. 1. 1. 0. 1. 0. 0. 1. 1. 1. 1. 1. 0. 1. 1. 0. 1.                 1. 0. 1. 0. 0. 1. 0. 0. 1. 0. 0. 1. 0. 1. 1. 1. 0. 1. 1. 0. 1. 0. 1. 1. 1.                 0. 0. 1. 0. 0. 1. 1. 1. 0. 1. 1. 1. 0. 1. 1. 1. 1. 0. 0. 1. 1. 1. 0. 0.                 1. 0. 0. 1. 0. 1. 0. 1. 1. 0. 0. 0. 1. 1. 1. 1. 1. 1. 1. 0. 0. 1. 1.                 1. 1. 1. 1. 0. 0. 1. 1. 0. 0. 0. 1. 0. 0. 0. 1. 0. 0. 0. 1. 1. 1. 1.                 1. 0. 0. 1. 0. 1. 1. 0. 0. 0. 0. 1. 0. 0. 0. 1. 0. 0. 0. 1. 0. 0. 0. 1. 1. 1.                 1. 0. 0. 1. 1. 0. 1. 0. 1. 1. 0. 1. 1. 0. 0. 0. 0. 1. 1. 1. 1. 0. 1.                 1. 1. 0. 1. 0. 1. 0. 1. 1. 1. 0. 1. 0. 1. 0. 1. 0. 1. 0. 1. 0. 1. 1. 1.                 1. 1. 1. 1. 0. 0. 1. 1. 0. 0. 0. 0. 1. 1. 1. 0. 0. 0. 1. 1. 1. 1. 1. 0.                 1. 0. 1. 0. 1. 1. 1. 0. 0. 1. 1. 1. 1. 0. 0. 0. 1. 0. 0. 1. 1. 0. 1. 0.                 1. 1. 1. 1. 1. 0. 1. 0. 1. 1. 1. 1. 1. 1. 1. 0. 1. 0. 0. 0. 1. 1. 1. 0.                 0. 0. 1. 0. 1. 0. 0. 1. 0. 1. 0. 1. 1. 0. 0. 1. 0. 1. 1. 1. 0. 1. 0. 0.                 1. 1. 1. 0. 1. 1. 1. 1. 0. 1. 1. 1. 0. 1. 1. 1. 0. 1. 0. 1. 1. 0. 1.                 1. 0. 1. 1. 0. 1. 1. 1. 0. 0. 0. 1. 1. 0. 1. 1. 1. 1. 1. 0. 1. 1. 0. 0.                 0. 1. 1. 1. 1. 1. 1. 1.]</p>

Tabel 4. 1 diatas adalah hasil dari modifikasi format menjadi *dmatrix* setelah proses *split* data, untuk memperjelas hasil dari gambar 4.10.

#### 4.1.5 Menentukan Hyperparameter

Proses selanjutnya adalah menentukan *hyperparameter* menggunakan *Grid search* dengan *5-fold Cross-Validation*. *Grid Search* adalah teknik biasa yang dipakai untuk menemukan kombinasi *hyperparameter* terbaik agar mendapatkan model yang memiliki performa paling baik. Di sisi lain, untuk mengevaluasi kinerja model dengan *5-fold cross validation* dengan membagi data menjadi 5 fold yang berbeda sehingga menghasilkan gambaran yang lebih valid tentang bagaimana model akan bekerja pada data yang tidak terlihat. Jadi proses ini tentang bagaimana kombinasi keduanya bekerja dan manfaat penggabungannya dalam proses pemodelan.

```
Fitting 5 folds for each of 27 candidates, totalling 135 fits
Best Parameters: {'eval_metric': 'logloss', 'learning_rate': 0.1, 'max_depth': 4, 'n_estimators': 100, 'objective': 'binary:logistic'}
Best Estimator: XGBClassifier(base_score=None, booster=None, callbacks=None,
                               colsample_bylevel=None, colsample_bynode=None,
                               colsample_bytree=None, device=None, early_stopping_rounds=None,
                               enable_categorical=False, eval_metric='logloss',
                               feature_types=None, gamma=None, grow_policy=None,
                               importance_type=None, interaction_constraints=None,
                               learning_rate=0.1, max_bin=None, max_cat_threshold=None,
                               max_cat_to_onehot=None, max_delta_step=None, max_depth=4,
                               max_leaves=None, min_child_weight=None, missing=nan,
                               monotone_constraints=None, multi_strategy=None, n_estimators=100,
                               num_jobs=None, num_parallel_tree=None, random_state=None, ...)
```

Gambar 4. 8 *Output* mendefinisikan *hyperparameter*

Gambar 4.8 adalah hasil dari tangkapan layar selama proses pencarian kombinasi *hyperparameter* menggunakan *grid search* dengan *5-fold Cross Validation*. Proses penentuan *Hyperparameter* hanya menggunakan 3 parameter saja, berikut parameter yang peneliti pakai dalam penelitian ini :

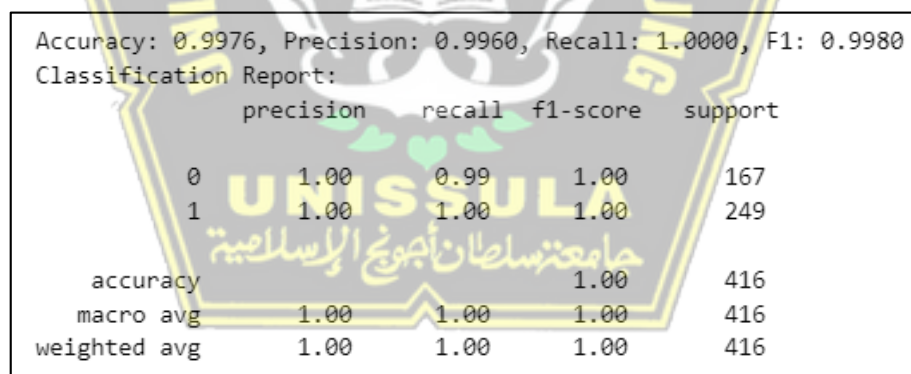
- *eval\_metric* : *logloss*
- *objective* : *binary:logistic*
- *max\_depth* : [3, 4, 5] = 3 parameter
- *learning\_rate* : [0.01, 0.1, 0.2] = 3 parameter
- *n\_estimators* : [50, 100, 200] = 3 parameter

Setelah dijumlah secara manual, kombinasi *hyperparameter* adalah  $3 \times 3 \times 3 = 27$ . Dengan menggunakan *5-fold Cross-Validation* (iterasi sebanyak 5 kali), yang artinya setiap kombinasi akan diuji sebanyak lima kali, sehingga

menghasilkan  $27 \times 5 = 135$  fits. Lalu hasil *grid search* menunjukkan parameter terbaik yang menghasilkan {'eval\_metric': 'logloss', 'learning\_rate': 0.1, 'max\_depth': 4, 'n\_estimators': 100, 'objective': 'binary:logistic'}. Serta menampilkan model terbaik yang telah dilatih dengan *hyperparameter* terbaik yang ditemukan, ini merupakan model akhir yang akan dipakai untuk membuat prediksi pada data baru.

#### 4.1.6 Pelatihan Model

Algoritma *xgboost* dimanfaatkan untuk menganalisis bentuk data pelatihan agar dapat menghasilkan prediksi risiko diabetes pada data baru. Dalam konteks pelatihan model, *5-fold Cross-Validation* diterapkan untuk menilai dan menjamin model dapat menggeneralisasi dengan baik serta mengurangi kemungkinan terjadinya *overfitting*. Prosesnya diiterasi selama lima kali, setiap fold berperan sebagai data validasi satu kali. Cara ini, memberikan gambaran proses kerja model pada data asing.



```

Accuracy: 0.9976, Precision: 0.9960, Recall: 1.0000, F1: 0.9980
Classification Report:

```

	precision	recall	f1-score	support
0	1.00	0.99	1.00	167
1	1.00	1.00	1.00	249
accuracy			1.00	416
macro avg	1.00	1.00	1.00	416
weighted avg	1.00	1.00	1.00	416

Gambar 4. 9 *Output* pelatihan model.

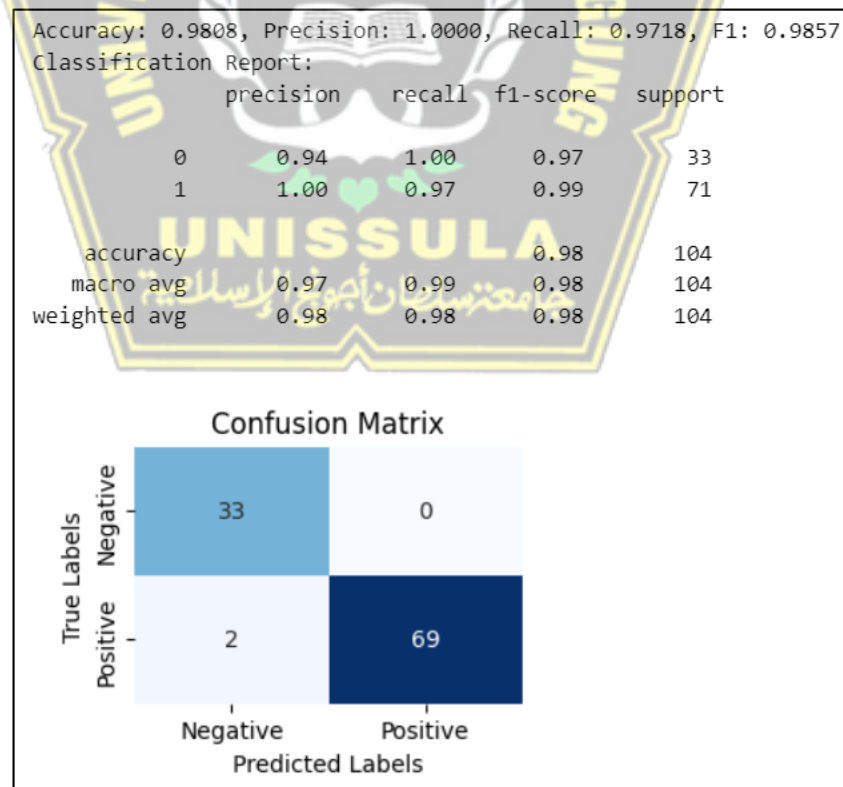
Gambar 4.9 merupakan hasil dari menyimpan model yang telah melalui tahap *grid search*, model dengan kombinasi *hyperparameter* terbaik berdasarkan kinerja selama perulangan 5 kali (*cross validation*). Model ini lah yang akan digunakan untuk membuat prediksi pada data *testing*. Hasil metrik yang dihasilkan Selama pelatihan model dalam format yang lebih mudah dibaca dengan menampilkan empat angka desimal. Dibuktikan dengan menampilkan/menghitung

metrik evaluasi pada data *train*. *Confusion matrix* untuk melihat seberapa banyak prediksi benar dan salah yang dilakukan oleh model pada data *training*.

*Accuracy* yang dihasilkan dari gambar di atas adalah untuk menghitung akurasi model dalam memprediksi yang benar, dengan nilai *accuracy* 0.9979, *precision* adalah rasio prediksi positif yang benar dari semua prediksi positif dengan hasil 0.9960, *recall* adalah rasio kejadian positif yang benar-benar diidentifikasi oleh model dengan hasil 1.000, dan f1-score ialah harmonisasi antara *precision* dan *recall* yang menghasilkan 0.9980.

#### 4.1.7 Evaluasi Model

Evaluasi model merupakan proses performa model yang sudah dilatih dan diuji menggunakan data asing. Tujuannya untuk mengukur seberapa normal model jika diterapkan pada data asing. Model juga akan menghasilkan probabilitas prediksi risiko diabetes untuk setiap kategori untuk mengukur tingkat keyakinan model dalam menghasilkan prediksi risiko diabetes.



Gambar 4. 10 Output evaluasi model

Gambar 4.10 adalah hasil tangkapan layar dari evaluasi model menggunakan *confusion matrix*. Hasil evaluasi lebih jelas akan ditampilkan melalui tabel di bawah ini :

Tabel 4.2 Nilai confusion matrix

<b>True Positif (TP)</b>	<b>False Positif (FP)</b>	<b>False Negatif (FN)</b>	<b>True Negatif (TN)</b>
33	0	2	69

Tabel 4.2 diatas merupakan hasil dari *confusion matrix* dengan nilai *true positif* sebanyak 33, *false positif* 0, *false negative* 2 dan *true negatif* 69. Nilai-nilai ini dapat dihitung menggunakan rumus persamaan akurasi (5), precision (6), recall (7) dan f1-score (9) pada bab 2 sebelumnya.

Tabel 4.3 Hasil evaluasi kinerja model

	<b>Accuracy</b>	<b>Precision</b>	<b>Recall</b>	<b>F1-score</b>
<b>Pelatihan</b>	99%	99%	100%	99%
<b>Pengujian</b>	98%	100%	97%	98%

Pada tabel 4.3 menampilkan hasil evaluasi kinerja model pada data pelatihan dan data pengujian yang menunjukkan bahwa model memiliki performa yang paling bagus pada kedua pembagian data tersebut. Dengan akurasi yang tinggi model mampu memprediksi label dengan sangat baik. nilai presisi yang tinggi menunjukkan kemampuan model yang baik dalam mengidentifikasi data positif secara akurat. Selain itu, F1-score yang tinggi mengindikasikan performa model yang solid secara keseluruhan.

Tabel 4.4 Hasil evaluasi setiap kelas

	<b>Precision</b>	<b>Recall</b>	<b>F1-score</b>
<b>Non Diabetes (0)</b>	94%	100%	97%
<b>Diabetes (1)</b>	100%	97%	99%

Pada tabel 4.4 diatas menunjukkan hasil skor dari setiap yang diperoleh dari model prediksi diabetes. Dalam tabel ini, terdapat 2 kategori yang dievaluasi, yaitu kelas 0 untuk non-diabetes yang mengindikasikan bahwa seseorang dalam kategori ini tidak menderita diabetes, model prediksi mengklasifikasikan data

yang masuk ke dalam kelas ini berdasarkan karakteristik yang menunjukkan ketidakadaan penyakit diabetes. Sebaliknya, kelas 1 untuk diabetes yang mengindikasikan bahwa seseorang dalam kategori ini menderita diabetes, model prediksi mengklasifikasikan data yang masuk ke dalam kelas ini berdasarkan karakteristik yang menunjukkan adanya risiko penyakit diabetes

Setiap nilai skor yang ditampilkan dalam tabel adalah hasil evaluasi kinerja model terhadap data yang digunakan. Untuk mengukur model dalam membuat prediksi yang benar (Akurasi), untuk mengukur prediksi positif yang benar dibandingkan prediksi positif yang dibuat model untuk setiap kategori (precision), lalu untuk mengukur seberapa banyak kasus positif yang benar-benar terdeteksi oleh model dari total kasus positif yang ada (recall) dan untuk menghitung rata-rata harmonis dari presisi dan recall yang memberikan gambaran umum tentang keseimbangan antara keduanya (*F1-score*).

#### 4.1.8 Hasil Implementasi Streamlit

Laporan penelitian ini menghasilkan model yang telah diimplementasikan ke dalam bentuk website. antarmuka yang ditampilkan sesuai dengan rencana yang telah dijelaskan sebelumnya pada tahap pemodelan. Di bawah ini terdapat hasil tangkapan layar dari demo program yang telah dikembangkan.

##### 1. Tampilan utama website



The screenshot shows a web browser window displaying a Streamlit application. The page title is "Masukkan Riwayat Kesehatan". The form contains several input fields and dropdown menus for health history, including Age, Gender, Polyuria, Polydipsia, Sudden Weight Loss, Weakness, Polyphagia, and Genital Thrush. On the right side, there is a prominent heading: "KLASIFIKASI RISIKO DIABETES MELLITUS BERDASARKAN RIWAYAT KESEHATAN MENGGUNAKAN XGBOOST". Below this heading, a note reads: "NB: Silahkan isi data terlebih dahulu "Masukkan Riwayat Kesehatan" di samping dan klik tombol "Prediksi Risiko Diabetes" pada bagian bawah untuk melihat hasilnya."

Gambar 4. 11 Tampilan awal website

Gambar 4.11 merupakan tampilan utama dari website yang dikembangkan menggunakan streamlit. Di halaman ini, pengguna dapat menemukan panduan tentang cara menggunakan website untuk memprediksi risiko diabetes. Penjelasan lebih detailnya dijabarkan dalam bentuk tabel dibawah :

Tabel 4. 5 Penjelasan tampilan utama website

Tampilan website	Penjelasan
"NB: Silakan isi data terlebih dahulu di bagian 'Masukkan Riwayat Kesehatan' di samping, dan klik tombol 'Prediksi Risiko Diabetes' di bagian bawah untuk melihat hasil prediksi."	Petunjuk ini memberi tahu pengguna bahwa mereka harus mengisi informasi riwayat kesehatan mereka terlebih dahulu sebelum dapat melihat hasil prediksi risiko diabetes.
"Masukkan Riwayat Kesehatan"	Sidebar berisi formulir di mana pengguna dapat memasukkan data riwayat kesehatan dari gejala yang mereka alami. Formulir ini harus diisi dengan lengkap agar model dapat melakukan prediksi dengan akurat.
"Prediksi Risiko Diabetes"	Setelah mengisi semua data yang diperlukan, pengguna harus menekan tombol ini untuk memulai proses prediksi. Hasil prediksi kemudian akan ditampilkan di halaman utama (Risiko dan probabilitas diabetes).



## 2. Tampilan hasil positif

**Masukkan Riwayat Kesehatan**

Age/Umur: 34  
Visual Blurring (penglihatan kabur): Yes  
Gender/Jenis Kelamin: Male  
Polyuria ( buang air kecil yang berlebihan): Yes  
Polydipsia (haus berlebihan): Yes  
Sudden Weight Loss (Turun Berat badan secara tiba-tiba): Yes  
Weakness (lemah): Yes  
Polyphagia (peringkatan nafsu makan): Yes  
Genital Thrush (infeksi jamur pada area genital): No

Visual Blurring (penglihatan kabur): No  
Itching (gatal-gatal pada kulit): No  
Instability (muntah-muntah): No  
Delayed Healing (proses penyembuhan luka yang lambat): Yes  
Partial Paresis (lemah sebagian otot): Yes  
Muscle Stiffness (kekakuan pada otot): Yes  
Alopecia (rambut rontok): No  
Obesity (obesitas): Yes

**KLASIFIKASI RISIKO DIABETES MELLITUS BERDASARKAN RIWAYAT KESEHATAN MENGGUNAKAN XGBOOST**

Data Pasien Baru Setelah Preprocessing:

Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital Thrush
0	0.8214	-0.7651	1.0077	1.1068	1.1817	-0.8396	1.0027

**HASIL PREDIKSI**  
Risiko Diabetes: Positive  
Probabilitas Risiko Diabetes: 1.00

Gambar 4. 12 Tampilan setelah memasukkan input dengan hasil positif

Pada gambar 4.12 merupakan tahapan yang terjadi setelah pengguna memasukkan data riwayat kesehatan yang dialami pengguna/pasien. Dari gambar diatas menunjukkan bahwa pengguna/pasien di prediksi risiko diabetes positif dengan hasil probabilitas 1.00 yang artinya model sangat yakin bahwa pasien tersebut memiliki penyakit diabetes.

## 3. Tampilan hasil negatif

**Masukkan Riwayat Kesehatan**

Age/Umur: 44  
Visual Blurring (penglihatan kabur): No  
Gender/Jenis Kelamin: Male  
Polyuria ( buang air kecil yang berlebihan): No  
Polydipsia (haus berlebihan): No  
Sudden Weight Loss (Turun Berat badan secara tiba-tiba): No  
Weakness (lemah): No  
Polyphagia (peringkatan nafsu makan): No  
Genital Thrush (infeksi jamur pada area genital): Yes

Visual Blurring (penglihatan kabur): No  
Itching (gatal-gatal pada kulit): No  
Instability (muntah-muntah): No  
Delayed Healing (proses penyembuhan luka yang lambat): Yes  
Partial Paresis (lemah sebagian otot): No  
Muscle Stiffness (kekakuan pada otot): No  
Alopecia (rambut rontok): Yes  
Obesity (obesitas): Yes

**KLASIFIKASI RISIKO DIABETES MELLITUS BERDASARKAN RIWAYAT KESEHATAN MENGGUNAKAN XGBOOST**

Data Pasien Baru Setelah Preprocessing:

Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital Thrush
0	-0.3219	-0.7651	1.0077	-0.961	1.1817	-0.8396	1.8912

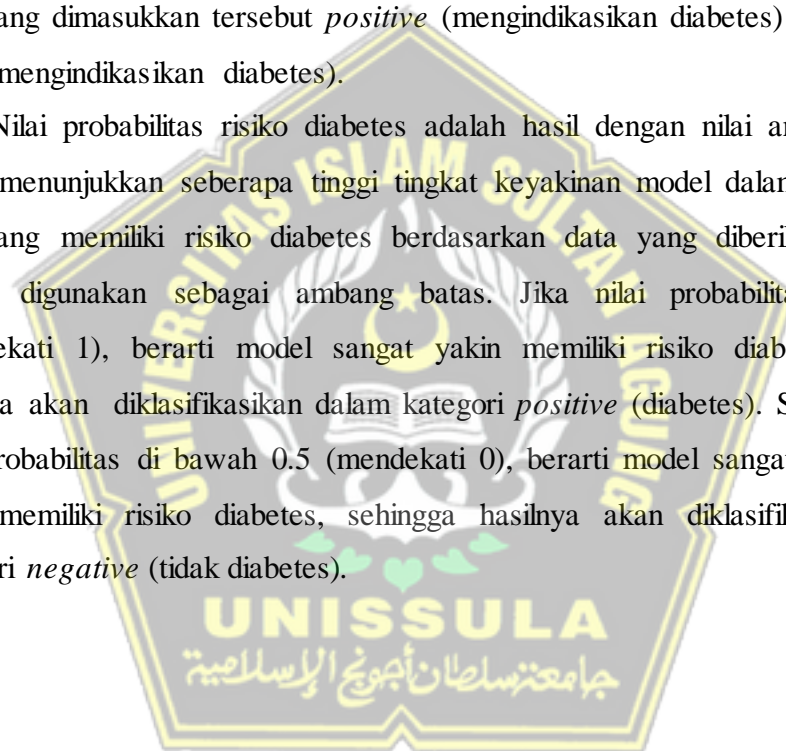
**HASIL PREDIKSI**  
Risiko Diabetes: Negative  
Probabilitas Risiko Diabetes: 0.00

Gambar 4. 13 Tampilan setelah memasukkan input hasil dengan hasil negatif

Pada gambar 4.13 merupakan tahapan yang terjadi setelah pengguna memasukkan data riwayat kesehatan yang pasien alami. Dari gambar diatas menunjukkan bahwa pengguna di prediksi risiko diabetes negative dengan nilai probabilitas 0.00 yang artinya model sangat yakin bahwa pasien tidak terkena diabetes.

Gambar 4.12 dan 4.13 ialah tampilan hasil dari serangkaian proses yang terjadi setelah pengguna memasukkan data riwayat kesehatan. Pertama, melibatkan proses standarisasi terhadap data yang telah dimasukkan, agar nilai-nilai input yang baru dimasukan berada dalam skala yang beragam berdasarkan parameter yang telah diolah oleh model sebelumnya yang sudah dilatih dan hasilnya akan ditampilkan di halaman utama. Kemudian model ini akan menghasilkan prediksi risiko diabetes dan probabilitas risiko diabetes berdasarkan data input yang dimasukkan. Model akan menentukan apakah risiko diabetes dari data yang dimasukkan tersebut *positive* (mengindikasikan diabetes) atau *negative* (tidak mengindikasikan diabetes).

Nilai probabilitas risiko diabetes adalah hasil dengan nilai antara 0 dan 1 untuk menunjukkan seberapa tinggi tingkat keyakinan model dalam memprediksi seseorang memiliki risiko diabetes berdasarkan data yang diberikan. Nilai 0.5 (50%) digunakan sebagai ambang batas. Jika nilai probabilitas diatas 0.5 (mendekati 1), berarti model sangat yakin memiliki risiko diabetes, sehingga hasilnya akan diklasifikasikan dalam kategori *positive* (diabetes). Sebaliknya, jika nilai probabilitas di bawah 0.5 (mendekati 0), berarti model sangat yakin bahwa tidak memiliki risiko diabetes, sehingga hasilnya akan diklasifikasikan dalam kategori *negative* (tidak diabetes).



## 4.2 Analisis Penelitian

Tahap ini, menganalisis sistem yang telah dikembangkan menggunakan streamlit dengan menguji sistem melalui *input* dan *output* untuk memastikan sistem bekerja sesuai spesifikasi dan tujuan yang telah ditetapkan pada sistem, berikut merupakan hasil dari beberapa prediksi yang telah dicoba :

Tabel 4. 6 Analisis Penelitian

<i>Input</i>	<b>Prediksi yang diharapkan</b>	<i>Output</i>	<b>Kesimpulan</b>
 <p>Masukkan Riwayat Kesehatan</p> <p>Age/Umur: 72</p> <p>Visual Blurring (penglihatan kabur): Yes</p> <p>Gender/Jenis Kelamin: Male</p> <p>Itching (gatal-gatal pada kulit): Yes</p> <p>Polyuria (Buang air kecil yang berlebihan): Yes</p> <p>Irritability (emosional): No</p> <p>Polydipsia (Haus berlebihan): No</p> <p>Delayed Healing (proses penyembuhan luka yang lambat): Yes</p> <p>Sudden Weight Loss (Turun berat badan secara tiba-tiba): No</p> <p>Partial Paresis (lemah sebagian otot): Yes</p> <p>Weakness (lemah): No</p> <p>Muscle Stiffness (kekakuan pada otot): Yes</p> <p>Polyphagia (peningkatan nafsu makan): Yes</p> <p>Alopecia (rambut rontok): Yes</p> <p>Genital Thrush (infeksi jamur pada area genital): No</p> <p>Obesity (obesitas): No</p> <p>Prediksi Risiko Diabetes</p>	negatif	positif	Sistem tidak akurat
 <p>Masukkan Riwayat Kesehatan</p> <p>Age/Umur: 40</p> <p>Visual Blurring (penglihatan kabur): No</p> <p>Gender/Jenis Kelamin: Female</p> <p>Itching (gatal-gatal pada kulit): Yes</p> <p>Polyuria (Buang air kecil yang berlebihan): Yes</p> <p>Irritability (emosional): No</p> <p>Polydipsia (Haus berlebihan): Yes</p> <p>Delayed Healing (proses penyembuhan luka yang lambat): No</p> <p>Sudden Weight Loss (Turun berat badan secara tiba-tiba): No</p> <p>Partial Paresis (lemah sebagian otot): Yes</p> <p>Weakness (lemah): Yes</p> <p>Muscle Stiffness (kekakuan pada otot): No</p> <p>Polyphagia (peningkatan nafsu makan): Yes</p> <p>Alopecia (rambut rontok): No</p> <p>Genital Thrush (infeksi jamur pada area genital): No</p> <p>Obesity (obesitas): No</p> <p>Prediksi Risiko Diabetes</p>	positif	positif	Sistem memprediksi dengan akurat

<i>Input</i>	<b>Prediksi yang diharapkan</b>	<i>Output</i>	<b>Kesimpulan</b>
<p>Masukkan Riwayat Kesehatan</p> <p>Age/Umur <input type="text" value="57"/> Visual Blurring (penglihatan kabur) <input type="text" value="No"/></p> <p>1 180</p> <p>Gender/Jenis Kelamin <input type="text" value="Male"/> Itching (gatal-gatal pada kulit) <input type="text" value="No"/></p> <p>Polyuria (Buang air kecil yang berlebihan) <input type="text" value="Yes"/> Irritability (emosional) <input type="text" value="No"/></p> <p>Polydipsia (Haus berlebihan) <input type="text" value="Yes"/> Delayed Healing (proses penyembuhan luka yang lambat) <input type="text" value="Yes"/></p> <p>Sudden Weight Loss (Turun berat badan secara tiba-tiba) <input type="text" value="No"/> Partial Paresis (lemah sebagian otot) <input type="text" value="Yes"/></p> <p>Weakness (lemah) <input type="text" value="Yes"/> Muscle Stiffness (kekakuan pada otot) <input type="text" value="No"/></p> <p>Polyphagia (peningkatan nafsu makan) <input type="text" value="Yes"/> Alopecia (rambut rontok) <input type="text" value="No"/></p> <p>Genital Thrush (infeksi jamur pada area genital) <input type="text" value="Yes"/> Obesity (obesitas) <input type="text" value="No"/></p> <p><a href="#">Prediksi Risiko Diabetes</a></p>	positif	positif	Sistem memprediksi dengan akurat
<p>Masukkan Riwayat Kesehatan</p> <p>Age/Umur <input type="text" value="41"/> Visual Blurring (penglihatan kabur) <input type="text" value="Yes"/></p> <p>1 180</p> <p>Gender/Jenis Kelamin <input type="text" value="Male"/> Itching (gatal-gatal pada kulit) <input type="text" value="Yes"/></p> <p>Polyuria (Buang air kecil yang berlebihan) <input type="text" value="Yes"/> Irritability (emosional) <input type="text" value="Yes"/></p> <p>Polydipsia (Haus berlebihan) <input type="text" value="Yes"/> Delayed Healing (proses penyembuhan luka yang lambat) <input type="text" value="No"/></p> <p>Sudden Weight Loss (Turun berat badan secara tiba-tiba) <input type="text" value="Yes"/> Partial Paresis (lemah sebagian otot) <input type="text" value="No"/></p> <p>Weakness (lemah) <input type="text" value="Yes"/> Muscle Stiffness (kekakuan pada otot) <input type="text" value="No"/></p> <p>Polyphagia (peningkatan nafsu makan) <input type="text" value="Yes"/> Alopecia (rambut rontok) <input type="text" value="No"/></p> <p>Genital Thrush (infeksi jamur pada area genital) <input type="text" value="Yes"/> Obesity (obesitas) <input type="text" value="Yes"/></p> <p><a href="#">Prediksi Risiko Diabetes</a></p>	positif	positif	Sistem memprediksi dengan akurat

<i>Input</i>	<b>Prediksi yang diharapkan</b>	<i>Output</i>	<b>Kesimpulan</b>
<p>Masukkan Riwayat Kesehatan</p> <p>Age/Umur <input type="text" value="45"/> Visual Blurring (penglihatan kabur) <input type="checkbox"/> Yes <input checked="" type="checkbox"/> No</p> <p>Gender/Jenis Kelamin <input type="text" value="Female"/> Itching (gatal-gatal pada kulit) <input type="checkbox"/> Yes <input checked="" type="checkbox"/> No</p> <p>Polyuria (Buang air kecil yang berlebihan) <input type="checkbox"/> Yes <input checked="" type="checkbox"/> No Irritability (emosional) <input type="checkbox"/> Yes <input checked="" type="checkbox"/> No</p> <p>Polydipsia (Haus berlebihan) <input type="checkbox"/> Yes <input checked="" type="checkbox"/> No Delayed Healing (proses penyembuhan luka yang lambat) <input type="checkbox"/> Yes <input checked="" type="checkbox"/> No</p> <p>Sudden Weight Loss (Turun berat badan secara tiba-tiba) <input type="checkbox"/> Yes <input checked="" type="checkbox"/> No Partial Paresis (lemah sebagian otot) <input type="checkbox"/> Yes <input checked="" type="checkbox"/> No</p> <p>Weakness (lemah) <input type="checkbox"/> Yes <input checked="" type="checkbox"/> No Muscle Stiffness (kekakuan pada otot) <input type="checkbox"/> Yes <input checked="" type="checkbox"/> No</p> <p>Polyphagia (peningkatan nafsu makan) <input type="checkbox"/> Yes <input checked="" type="checkbox"/> No Alopecia (rambut rontok) <input type="checkbox"/> Yes <input checked="" type="checkbox"/> No</p> <p>Genital Thrush (infeksi jamur pada area genital) <input type="checkbox"/> Yes <input checked="" type="checkbox"/> No Obesity (obesitas) <input type="checkbox"/> Yes <input checked="" type="checkbox"/> No</p> <p><input type="button" value="Prediksi Risiko Diabetes"/></p>	positif	positif	Sistem memprediksi dengan akurat
<p>Masukkan Riwayat Kesehatan</p> <p>Age/Umur <input type="text" value="62"/> Visual Blurring (penglihatan kabur) <input type="checkbox"/> Yes <input checked="" type="checkbox"/> No</p> <p>Gender/Jenis Kelamin <input type="text" value="Female"/> Itching (gatal-gatal pada kulit) <input type="checkbox"/> Yes <input checked="" type="checkbox"/> No</p> <p>Polyuria (Buang air kecil yang berlebihan) <input type="checkbox"/> Yes <input checked="" type="checkbox"/> No Irritability (emosional) <input type="checkbox"/> Yes <input checked="" type="checkbox"/> No</p> <p>Polydipsia (Haus berlebihan) <input type="checkbox"/> Yes <input checked="" type="checkbox"/> No Delayed Healing (proses penyembuhan luka yang lambat) <input type="checkbox"/> Yes <input checked="" type="checkbox"/> No</p> <p>Sudden Weight Loss (Turun berat badan secara tiba-tiba) <input type="checkbox"/> Yes <input checked="" type="checkbox"/> No Partial Paresis (lemah sebagian otot) <input type="checkbox"/> Yes <input checked="" type="checkbox"/> No</p> <p>Weakness (lemah) <input type="checkbox"/> Yes <input checked="" type="checkbox"/> No Muscle Stiffness (kekakuan pada otot) <input type="checkbox"/> Yes <input checked="" type="checkbox"/> No</p> <p>Polyphagia (peningkatan nafsu makan) <input type="checkbox"/> Yes <input checked="" type="checkbox"/> No Alopecia (rambut rontok) <input type="checkbox"/> Yes <input checked="" type="checkbox"/> No</p> <p>Genital Thrush (infeksi jamur pada area genital) <input type="checkbox"/> Yes <input checked="" type="checkbox"/> No Obesity (obesitas) <input type="checkbox"/> Yes <input checked="" type="checkbox"/> No</p> <p><input type="button" value="Prediksi Risiko Diabetes"/></p>	positif	positif	Sistem memprediksi dengan akurat

<i>Input</i>	<b>Prediksi yang diharapkan</b>	<i>Output</i>	<b>Kesimpulan</b>
<p>Masukkan Riwayat Kesehatan</p> <p>Age/Umur <input type="text" value="25"/> Visual Blurring (penglihatan kabur) <input type="checkbox"/></p> <p>1 160 Yes <input type="checkbox"/></p> <p>Gender/Jenis Kelamin Female <input type="checkbox"/> Itching (gatal-gatal pada kulit) <input type="checkbox"/></p> <p>Polyuria (Buang air kecil yang berlebihan) No <input type="checkbox"/> Irritability (emosional) <input type="checkbox"/></p> <p>Polydipsia (Haus berlebihan) No <input type="checkbox"/> Delayed Healing (proses penyembuhan luka yang lambat) <input type="checkbox"/></p> <p>Sudden Weight Loss (Turun berat badan secara tiba-tiba) No <input type="checkbox"/> Partial Paresis (lemah sebagian otot) <input type="checkbox"/></p> <p>Weakness (lemah) No <input type="checkbox"/> Muscle Stiffness (kekakuan pada otot) <input type="checkbox"/></p> <p>Polyphagia (peningkatan nafsu makan) No <input type="checkbox"/> Alopecia (rambut rontok) <input type="checkbox"/></p> <p>Genital Thrush (infeksi jamur pada area genital) No <input type="checkbox"/> Obesity (obesitas) <input type="checkbox"/></p> <p><input type="button" value="Prediksi Risiko Diabetes"/></p>	positif	positif	Sistem memprediksi dengan akurat
<p>Masukkan Riwayat Kesehatan</p> <p>Age/Umur <input type="text" value="45"/> Visual Blurring (penglihatan kabur) <input type="checkbox"/></p> <p>1 160 No <input type="checkbox"/></p> <p>Gender/Jenis Kelamin Male <input type="checkbox"/> Itching (gatal-gatal pada kulit) <input type="checkbox"/></p> <p>Polyuria (Buang air kecil yang berlebihan) No <input type="checkbox"/> Irritability (emosional) <input type="checkbox"/></p> <p>Polydipsia (Haus berlebihan) No <input type="checkbox"/> Delayed Healing (proses penyembuhan luka yang lambat) <input type="checkbox"/></p> <p>Sudden Weight Loss (Turun berat badan secara tiba-tiba) No <input type="checkbox"/> Partial Paresis (lemah sebagian otot) <input type="checkbox"/></p> <p>Weakness (lemah) Yes <input type="checkbox"/> Muscle Stiffness (kekakuan pada otot) <input type="checkbox"/></p> <p>Polyphagia (peningkatan nafsu makan) No <input type="checkbox"/> Alopecia (rambut rontok) <input type="checkbox"/></p> <p>Genital Thrush (infeksi jamur pada area genital) No <input type="checkbox"/> Obesity (obesitas) <input type="checkbox"/></p> <p><input type="button" value="Prediksi Risiko Diabetes"/></p>	negatif	positif	Sistem tidak akurat

<i>Input</i>	<b>Prediksi yang diharapkan</b>	<i>Output</i>	<b>Kesimpulan</b>
<p>Masukkan Riwayat Kesehatan</p> <p>Age/Umur <input type="text" value="64"/> Visual Blurring (penglihatan kabur) <input type="text" value="No"/></p> <p>1 <input type="text" value="168"/> Itching (gatal-gatal pada kulit) <input type="text" value="No"/></p> <p>Gender/Jenis Kelamin <input type="text" value="Male"/> Irritability (emosional) <input type="text" value="Yes"/></p> <p>Polyuria (Buang air kecil yang berlebihan) <input type="text" value="No"/> Delayed Healing (proses penyembuhan luka yang lambat) <input type="text" value="Yes"/></p> <p>Polydipsia (Haus berlebihan) <input type="text" value="Yes"/> Sudden Weight Loss (Turun berat badan secara tiba-tiba) <input type="text" value="No"/></p> <p>Weakness (lemah) <input type="text" value="No"/> Partial Paresis (lemah sebagian otot) <input type="text" value="No"/></p> <p>Polyphagia (peningkatan nafsu makan) <input type="text" value="No"/> Muscle Stiffness (kekakuan pada otot) <input type="text" value="No"/></p> <p>Genital Thrush (infeksi jamur pada area genital) <input type="text" value="No"/> Alopecia (rambut rontok) <input type="text" value="No"/></p> <p><input type="text" value="No"/> Obesity (obesitas) <input type="text" value="No"/></p> <p><input type="button" value="Prediksi Risiko Diabetes"/></p>	positif	positif	Sistem memprediksi dengan benar
<p>Masukkan Riwayat Kesehatan</p> <p>Age/Umur <input type="text" value="46"/> Visual Blurring (penglihatan kabur) <input type="text" value="No"/></p> <p>1 <input type="text" value="160"/> Itching (gatal-gatal pada kulit) <input type="text" value="Yes"/></p> <p>Gender/Jenis Kelamin <input type="text" value="Male"/> Irritability (emosional) <input type="text" value="No"/></p> <p>Polyuria (Buang air kecil yang berlebihan) <input type="text" value="No"/> Delayed Healing (proses penyembuhan luka yang lambat) <input type="text" value="Yes"/></p> <p>Polydipsia (Haus berlebihan) <input type="text" value="No"/> Sudden Weight Loss (Turun berat badan secara tiba-tiba) <input type="text" value="No"/></p> <p>Weakness (lemah) <input type="text" value="Yes"/> Partial Paresis (lemah sebagian otot) <input type="text" value="No"/></p> <p>Polyphagia (peningkatan nafsu makan) <input type="text" value="No"/> Muscle Stiffness (kekakuan pada otot) <input type="text" value="No"/></p> <p>Genital Thrush (infeksi jamur pada area genital) <input type="text" value="No"/> Alopecia (rambut rontok) <input type="text" value="Yes"/></p> <p><input type="text" value="No"/> Obesity (obesitas) <input type="text" value="No"/></p> <p><input type="button" value="Prediksi Risiko Diabetes"/></p>	negatif	negatif	Sistem memprediksi dengan akurat

<i>Input</i>	<b>Prediksi yang diharapkan</b>	<i>Output</i>	<b>Kesimpulan</b>
	negatif	negatif	Sistem memprediksi dengan akurat

Dapat dilihat pada tabel 4. 6 menampilkan hasil analisis penelitian dari beberapa data yang telah diuji. Sistem berhasil memprediksi dengan tepat sembilan data, sedangkan dua data lainnya diprediksi secara keliru. Temuan ini mengidentifikasi bahwa sistem performa yang cukup memuaskan dalam memprediksi dan memberikan diagnosis yang akurat terhadap data asing yang belum pernah dilihat selama proses pelatihan model.

Kesalahan dalam memprediksi data mungkin saja terjadi karena data yang tidak seimbang, data jauh lebih banyak kategori positif daripada negatif. Sehingga model cenderung memprediksi kelas mayoritas dan mengurangi akurasi pada kelas minoritas. Berdasarkan analisis menyeluruh dari penelitian ini, dapat diambil kesimpulan bahwas sistem ini mampu memberikan hasil prediksi risiko diabetes mellitus berdasarkan riwayat kesehatan yang dimasukkan secara akurat.



## BAB V

### KESIMPULAN DAN SARAN

#### 5.1 Kesimpulan

Laporan pada penelitian ini membuktikan bahwa Algoritma *Extreme Gradient Boosting* menunjukkan kinerja yang optimal dan akurat untuk mengklasifikasikan risiko diabetes mellitus berdasarkan riwayat kesehatan untuk meningkatkan diagnosa sedini mungkin, berikut merupakan hasil penelitian dari laporan ini :

1. Algoritma *Extreme Gradient Boosting* (XGBoost) menunjukkan kinerja yang unggul dengan menghasilkan *accuracy* 0.9808, *precision* 1.000, *recall* 0.9718 dan *f1-score* 0.9857.
2. Dengan tingkat akurasi yang tinggi, model berhasil memprediksi apakah seseorang berisiko memiliki diabetes atau tidak dengan sangat baik berdasarkan riwayat kesehatan yang dialami, agar dapat meningkatkan diagnosa sedini mungkin.

#### 5.2 Saran

Walaupun penelitian ini menunjukkan hasil yang memuaskan. Untuk pengembangan lebih lanjut, penulis menyarankan agar sistem dapat diterapkan ke dalam platform mobile seperti Android. Selain itu, dengan memperluas jumlah dan keragaman data untuk meningkatkan generalisasi model, Data yang lebih bervariasi akan membantu model mempelajari berbagai perbedaan yang lebih luas, sehingga mampu memperbaiki keakuratan serta kestabilan prediksi.

## DAFTAR PUSTAKA

- Abdurrahman, G. *dkk.* (2022) “Optimasi Algoritma XGBoost Classifier Menggunakan Hyperparameter Gridsearch dan Random Search Pada Klasifikasi Penyakit Diabetes,” *INFORMAL: Informatics Journal*, 7(3), hal. 193. Tersedia pada: <https://doi.org/10.19184/isj.v7i3.35441>.
- Ahsana, R. *dkk.* (2021) “Perbandingan Akurasi Algoritma Adaboost Dan Algoritma Lightgbm Untuk Klasifikasi Penyakit Diabetes,” *e-Proceeding of Engineering*, 8(5), hal. 9738–9748.
- Aqsha, M. *dkk.* (2024) “Performa Klasifikasi Data Tidak Seimbang Dengan Pendekatan Machine Learning (Studi Kasus: Diabetes Indian Pima),” *Jurnal Matematika UNAND*, 12(2), hal. 176–193.
- AWS (2024) *Amazon SageMaker: Developer Guide*.
- Blockeel, H. *dkk.* (2023) “Decision trees: from efficient prediction to responsible AI,” *Frontiers in Artificial Intelligence*, 6. Tersedia pada: <https://doi.org/10.3389/frai.2023.1124553>.
- Charbuty, B. *dkk.* (2021) “Classification Based on Decision Tree Algorithm for Machine Learning,” *Journal of Applied Science and Technology Trends*, 2(01), hal. 20–28. Tersedia pada: <https://doi.org/10.38094/jastt20165>.
- Fajar, R. *AI dkk.* (2024) “Analisis Tingkat Kemiskinan di Indonesia Menggunakan Model Vanilla Long Short-Term Memory dan Stacked Long Short-Term Memory,” 8, hal. 1637–1646. Tersedia pada: <https://doi.org/10.30865/mib.v8i3.7877>.
- Hana, F.M. (2020) “Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma Decision Tree C4.5,” *Jurnal SISKOM-KB (Sistem Komputer dan Kecerdasan Buatan)*, 4(1), hal. 32–39. Tersedia pada: <https://doi.org/10.47970/siskom-kb.v4i1.173>.
- Herni Yulianti, S.E. *dkk.* (2022) “Penerapan Metode Extreme Gradient Boosting (XGBOOST) pada Klasifikasi Nasabah Kartu Kredit,” *Journal of Mathematics: Theory and Applications*, 4(1), hal. 21–26. Tersedia pada: <https://doi.org/10.31605/jomta.v4i1.1792>.

- Hidayah, A.R. (2023) "Rancang Bangun Sistem Pakar Untuk Mendiagnosa Penyakit Diabetes Dengan Metode Certainty Factor," *Seminar Riset Mahasiswa-Computer & Electrical (SERIMA-CE)*, 1(1), hal. 209–214.
- Husaini, B.Q. dkk. (2023) "Penerapan Algoritma Decision Tree C45 untuk Klasifikasi Penjurusan Siswa," *Jurnal Teknologi Informatika dan Komputer*, 9(1), hal. 455–470. Tersedia pada: <https://doi.org/10.37012/jtik.v9i1.1512>.
- Kusuma, U. dkk. (2024) "PENERAPAN SENAM KAKI DIABETIC TERHADAP PENURUNAN KADAR GLUKOSA GULA DARAH PADA PASIEN DIABETES MELLITUS."
- Maulana, H. dkk. (2023) "Clustering RFM (Recency, Frequency, Monetary) Publisher Gim Menggunakan Algoritma K-Means," *Seminar Nasional Informatika Bela Negara (SANTIKA)*, 3, hal. 2747–0563.
- Maulana, M. dkk. (2024) "Algoritma Xgboost Untuk Klasifikasi Kualitas Air Minum," *JATI (Jurnal Mahasiswa Teknik Informatika)*, 7(5), hal. 3251–3256. Tersedia pada: <https://doi.org/10.36040/jati.v7i5.7308>.
- Mehdary, A. dkk. (2024) "Hyperparameter Optimization with Genetic Algorithms and XGBoost: A Step Forward in Smart Grid Fraud Detection," *Sensors*, 24(4). Tersedia pada: <https://doi.org/10.3390/s24041230>.
- Mohiuddin, G. dkk. (2023) "Intrusion Detection using hybridized Meta-heuristic techniques with Weighted XGBoost Classifier," *Expert Systems with Applications*, 232(April), hal. 120596. Tersedia pada: <https://doi.org/10.1016/j.eswa.2023.120596>.
- Murtiningsih, M.K. dkk. (2021) "Gaya Hidup sebagai Faktor Risiko Diabetes Melitus Tipe 2," *e-CliniC*, 9(2), hal. 328. Tersedia pada: <https://doi.org/10.35790/ec1.v9i2.32852>.
- Paramitha, Y.N. dkk. (2023) "Klasifikasi Penyakit Stroke Menggunakan Metode Naïve Bayes," *Jurnal Siger Matematika*, 04(01), hal. 11–16. Tersedia pada: <https://www.kaggle.com/datasets/zzetrkalkpakbal/full-filled->.
- Pinata, N.N. dkk. (2020) "Prediksi Kecelakaan Lalu Lintas di Bali dengan XGBoost pada Python," *Jurnal Ilmiah Merpati (Menara Penelitian Akademika Teknologi Informasi)*, 8(3), hal. 188. Tersedia pada:

<https://doi.org/10.24843/jim.2020.v08.i03.p04>.

- Purbolingga, Y. *dkk.* (2023) “Perbandingan Algoritma CatBoost dan XGBoost dalam Klasifikasi Penyakit Jantung,” *Jurnal APTEK Vol. 15 No 2 (2023)* 126-133, 15(2), hal. 126–133. Tersedia pada: <http://journal.upp.ac.id/index.php/aptek/article/download/1930/1163/4970>.
- Rabbani, S. *dkk.* (2023) “Perbandingan Evaluasi Kernel SVM untuk Klasifikasi Sentimen dalam Analisis Kenaikan Harga BBM,” *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, 3(2), hal. 153–160. Tersedia pada: <https://doi.org/10.57152/malcom.v3i2.897>.
- Rahman, I.F. (2020) “Implementasi Metode SVM, MLP dan Xgboost pada Data Ekspresi Gen,” hal. 1–79. Tersedia pada: <https://dspace.uui.ac.id/handle/123456789/23679>.
- Ridwan, A. (2020) “Penerapan Algoritma Naïve Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus,” *Jurnal SISKOM-KB (Sistem Komputer dan Kecerdasan Buatan)*, 4(1), hal. 15–21. Tersedia pada: <https://doi.org/10.47970/siskom-kb.v4i1.169>.
- Rombe, Y. (2021) “PENGUNAAN METODE XGBOOST UNTUK KLASIFIKASI OBESITAS DI INDONESIA,” *Journal of Chemical Information and Modeling*, 53(February), hal. 2021. Tersedia pada: <https://doi.org/10.1080/09638288.2019.1595750><https://doi.org/10.1080/17518423.2017.1368728><http://dx.doi.org/10.1080/17518423.2017.1368728><https://doi.org/10.1016/j.ridd.2020.103766><https://doi.org/10.1080/02640414.2019.1689076><https://doi.org/>.
- Salsabil, M. *dkk.* (2024) “Implementasi Data Mining Dalam Melakukan Prediksi Penyakit Diabetes Menggunakan Metode Random Forest Dan Xgboost,” *Jurnal Ilmiah Komputasi*, 23(1), hal. 51–58. Tersedia pada: <https://doi.org/10.32409/jikstik.23.1.3507>.
- Setio, P.B.N. *dkk.* (2020) “Klasifikasi Dengan Pohon Keputusan Berbasis Algoritme C4.5,” *PRISMA, Prosiding Seminar Nasional Matematika*, 3, hal. 64–71.
- Suharno, J.A. *dkk.* (2024) “Hubungan Indeks Massa Tubuh dan Lingkar Perut

dengan Diabetes Melitus pada Orang Dewasa di Indonesia : Hasil Analisis Data Riskesdas 2018,” 26(1), hal. 1–10. Tersedia pada: <https://doi.org/10.29238/jnutri.v26i1.382>.

Syahrani, I.M. (2019) “Comparison analysis of ensemble technique with boosting (Xgboost) and bagging (Randomforest) for classify splice junction DNA sequence category,” *Jurnal Penelitian Pos dan Informatika*, 9(1), hal. 27–36.

Ubaidillah, R. *dkk.* (2022) “Implementasi XGBoost Pada Keseimbangan Liver Patient Dataset dengan SMOTE dan Hyperparameter Tuning Bayesian Search,” *Jurnal Media Informatika Budidarma*, 6(3), hal. 1723. Tersedia pada: <https://doi.org/10.30865/mib.v6i3.4146>.

Yusufi, A.H. *dkk.* (2022) “Prediksi Resiko Kematian Pada Penderita Penyakit Kardiovaskular Menggunakan Metode Ensemble Learning,” *Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)*, hal. 531–542. Tersedia pada: <https://conference.upnvj.ac.id/index.php/senamika/article/view/2209>.

