SISTEM PAKAR DALAM ANALISIS PREDIKSI RISIKO DIABETES MENGGUNAKAN *MODEL* XGBOOST DAN ANALISIS FAKTOR RISIKO DENGAN EXPLAINABLE AI (SHAP)

SKRIPSI

DISUSUN OLEH

NADIA APRILIA SIREGAR

NPM. 2209010123

PROGRAM STUDI SISTEM INFORMASI

FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI

UNIVERSITAS MUHAMMADIYAH SUMATERA UTARA

MEDAN

2026

SISTEM PAKAR DALAM ANALISIS PREDIKSI RISIKO DIABETES MENGGUNAKAN *MODEL* XGBOOST DAN ANALISIS FAKTOR RISIKO DENGAN EXPLAINABLE AI (SHAP)

SKRIPSI

Diajukan sebagai salah satu syarat untuk memperoleh gelar Sarjana Komputer (S.Kom) dalam Program Studi Sistem Informasi pada Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Muhammadiyah Sumatera Utara

NADIA APRILIA SIREGAR
NPM. 2209010123

PROGRAM STUDI SISTEM INFORMASI
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI
UNIVERSITAS MUHAMMADIYAH SUMATERA UTARA
MEDAN
2026

LEMBAR PENGESAHAN

Judul Skripsi	:	SISTEM PAKAR DALAM ANALISIS PREDIKSI RISIKO DIABETES MENGGUNAKAN MODEL XGBOOST DAN ANALISIS FAKTOR RISIKO DENGAN EXPLAINABLE AI (SHAP)
Nama Mahasiswa	:	NADIA APRILIA SIREGAR
NPM	:	2209010123
Program Studi	:	SISTEM INFORMASI

Menyetujui
Komisi Pembimbing

(Dr. Firahmi Rizky, M.Kom)
NIDN. 0116079201

Ketua Program Studi

(Mahardika Abdi Prawira, S.Kom M.Kom.)
NIDN. 0117088902

Dekan

(Dr. Al-Khowarizmi, S.Kom., M.Kom.)
NIDN. 0127099201

PERNYATAAN ORISINALITAS

SISTEM PAKAR DALAM ANALISIS PREDIKSI RISIKO DIABETES MENGGUNAKAN *MODEL* XGBOOST DAN ANALISIS FAKTOR RISIKO DENGAN EXPLAINABLE AI (SHAP)

SKRIPSI

Saya menyatakan bahwa karya tulis ini adalah hasil karya sendiri, kecuali beberapa kutipan dan ringkasan yang masing-masing disebutkan sumbernya.

Medan, 13 Maret 2026

Yang membuat pernyataan

Nadia Aprilia Siregar

NPM. 2209010123

PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademika Universitas Muhammadiyah Sumatera Utara, saya bertanda tangan dibawah ini:

Nama	: Nadia Aprilia Siregar
NPM	: 2209010123
Program Studi	: Sistem Informasi
Karya Ilmiah	: Skripsi

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Muhammadiyah Sumatera Utara Hak Bedas Royalti Non-Eksekutif *(Non-Exclusive Royalty free Right)* atas penelitian skripsi saya yang berjudul:

SISTEM PAKAR DALAM ANALISIS PREDIKSI RISIKO DIABETES MENGGUNAKAN *MODEL* XGBOOST DAN ANALISIS FAKTOR RISIKO DENGAN EXPLAINABLE AI (SHAP)

Beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Non-Eksekutif ini, Universitas Muhammadiyah Sumatera Utara berhak menyimpan, mengalih media, memformat, mengelola dalam bentuk database, merawat dan mempublikasikan Skripsi saya ini tanpa meminta izin dari saya selama tetap mencantumkan nama saya sebagai penulis dan sebagai pemegang dan atau sebagai pemilik hak cipta.

Demikian pernyataan ini dibuat dengan sebenarnya.

Medan, 13 Maret 2026

Yang membuat pernyataan

Nadia Aprilia Siregar

NPM. 2209010123

RIWAYAT HIDUP

DATA PRIBADI

Nama Lengkap	: Nadia Aprilia Siregar
Tempat dan Tanggal Lahir	: Medan, 02 April 2004
Alamat Rumah	: Jl. Rakyat No.109 Medan Perjuangan
Telepon/Faks/HP	: 0895-3876-48484
E-mail	: nadiaaprilia637@gmail.com
Instansi Tempat Kerja	: -
Alamat Kantor	: -

DATA PENDIDIKAN

SD	: Negeri 060874 Kota Medan	TAMAT: 2016
SMP	: Negeri 12 Medan Kota Medan	TAMAT: 2019
SMA	: Negeri 8 Medan Kota Medan	TAMAT: 2022

KATA PENGANTAR

Puji dan syukur penulis panjatkan ke hadirat Allah SWT yang telah melimpahkan rahmat dan karunia-Nya sehingga penulis dapat menyelesaikan skripsi yang berjudul "SISTEM PAKAR DALAM ANALISIS PREDIKSI RISIKO DIABETES MENGGUNAKAN *MODEL* XGBOOST DAN ANALISIS FAKTOR RISIKO DENGAN EXPLAINABLE AI (SHAP)" dengan baik dan tepat waktu. Skripsi ini disusun sebagai salah satu syarat untuk memperoleh gelar Sarjana pada Program Studi Sistem Informasi. Penulis menyadari bahwa penyelesaian skripsi ini tidak terlepas dari bantuan, bimbingan, dan dukungan dari berbagai pihak. Oleh karena itu, penulis ingin menyampaikan ucapan terima kasih yang sebesar-besarnya kepada:

1. Bapak Prof. Dr. Akrim, M.Pd., Rektor Universitas Muhammadiyah Sumatera Utara (UMSU), Semoga Allah SWT senantiasa melimpahkan kesehatan, keberkahan, dan kebahagiaan kepada beliau dan keluarga, Aamiin.
2. Bapak Dr. Al-Khowarizmi, S.Kom., M.Kom. Dekan Fakultas Ilmu Komputer dan Teknologi Informasi (FIKTI) UMSU, Semoga Allah SWT senantiasa melimpahkan kesehatan, keberkahan, dan kebahagiaan kepada beliau dan keluarga, Aamiin.
3. Ibu Dr. Firahmi Rizky, S.Kom.,M.Kom. Wakil dekan 1 sekaligus Pembimbing, Penulis mengucapkan terima kasih yang sebesar-besarnya atas segala bimbingan, arahan, serta kesabaran yang telah diberikan selama proses penyusunan skripsi ini. Dukungan dan masukan yang diberikan sangat berarti bagi penulis dalam menyelesaikan penelitian ini dengan baik.

4. Bapak Mhd. Basri S.Si.,M.Kom. Wakil dekan 3, Semoga Allah SWT senantiasa melimpahkan kesehatan, keberkahan, dan kebahagiaan kepada beliau dan keluarga, Aamiin.
5. Bapak Mahardika Abdi Prawira Tanjung, S.Kom.,M.Kom, Ketua Program Studi Sistem Informasi, Semoga Allah SWT senantiasa melimpahkan kesehatan, keberkahan, dan kebahagiaan kepada beliau dan keluarga, Aamiin.
6. Bapak Mulkan Azhari, S.Kom.,M.Kom Sekretaris Program Studi Sistem Informasi, Semoga Allah SWT senantiasa melimpahkan kesehatan, keberkahan, dan kebahagiaan kepada beliau dan keluarga, Aamiin.
7. Terutama dan paling utama Kepada Kedua Orang Tua tercinta ayahanda Mhd.Isroq Siregar dan ibunda Fadma Dewi yang selalu memberikan ketulusan doanya dan juga memberikan dukungan baik secara moril maupun materil, juga motivasi sehingga penulis dapat menyelesaikan Skripsi ini. Semoga Allah SWT senantiasa melimpahkan kesehatan, keberkahan, dan kebahagiaan kepada kedua orang tua penulis, Amiin.
8. kepada kakak dan adik tercinta yang selalu memberikan dukungan serta doa yang tulus kepada penulis, memberikan semangat yang luarbiasa bahkan sampai sekarang.
9. Ucapan terimakasih yang sangat istimewa penulis berikan kepada Krisna Aditya, Partner yang senantiasa hadir dengan dukungan, perhatian, dan ketulusan, tidak hanya selama proses penyusunan skripsi ini, tetapi juga dalam perjalanan penulis selama menempuh perkuliahan. Kehadirannya yang selalu menemani, baik dalam suka maupun duka, menjadi penguat yang tidak tergantikan bagi penulis. Dalam berbagai situasi, saat penulis berada pada titik semangat maupun kelelahan, dukungan, kesabaran, dan pengertian yang diberikan telah menjadi sumber motivasi yang berarti untuk terus melangkah hingga penelitian ini dapat diselesaikan dengan baik. Penulis menyadari bahwa tanpa kehadiran dan peran tersebut, proses yang dijalani tidak akan terasa sekuat dan sebermakna ini. Oleh karena itu, penulis menyampaikan rasa terima kasih yang mendalam atas segala

dukungan dan kebersamaan yang telah diberikan, yang secara tidak langsung turut menjadi bagian dari terselesaikannya skripsi ini.

10. Kepada Firda Haliza yang menjadi satu-satunya teman yang selalu mendukung dan memberikan motivasi dari awal perkuliahan sampai saat ini, penulis mengucapkan terimakasih dan semangat terus untuk kita berdua.
11. Semua pihak yang terlibat langsung ataupun tidak langsung yang tidak dapat penulis ucapkan satu-persatu yang telah membantu penyelesaian skripsi ini semoga Allah membalas dengan kebaikan, Aamiin.

SISTEM PAKAR DALAM ANALISIS PREDIKSI RISIKO DIABETES MENGGUNAKAN *MODEL* XGBOOST DAN ANALISIS FAKTOR RISIKO DENGAN EXPLAINABLE AI (SHAP)

ABSTRAK

Perkembangan *Machine Learning* dalam bidang kesehatan memberikan peluang dalam deteksi dini penyakit, termasuk diabetes mellitus yang memiliki prevalensi tinggi dan sering terlambat terdeteksi. Metode pemeriksaan konvensional memiliki keterbatasan dalam aksesibilitas dan biaya, sementara *model* seperti XGBoost meskipun akurat masih memiliki kelemahan dalam interpretabilitas. Oleh karena itu, penelitian ini bertujuan mengembangkan sistem pakar berbasis web untuk prediksi risiko diabetes menggunakan XGBoost serta memberikan penjelasan faktor risiko melalui metode Explainable AI (SHAP). Rumusan masalah penelitian ini mencakup analisis kebutuhan sistem, perancangan sistem berbasis web, pengujian kinerja prediksi dan interpretasi, serta implementasi sistem yang dapat digunakan oleh masyarakat. Metode yang digunakan adalah *Data Mining* dengan teknik klasifikasi menggunakan algoritma XGBoost. Dataset yang digunakan berjumlah 520 data dari Kaggle dengan variabel seperti usia, jenis kelamin, dan gejala klinis diabetes. Tahapan penelitian meliputi data cleaning, label *encoding*, pelatihan *model*, serta integrasi SHAP untuk interpretasi hasil. Perancangan sistem dilakukan menggunakan UML. Hasil penelitian menunjukkan bahwa sistem mampu memprediksi risiko diabetes dengan baik serta memberikan penjelasan faktor risiko secara transparan. Implementasi SHAP memungkinkan pengguna memahami kontribusi setiap variabel terhadap hasil prediksi. Sistem yang dikembangkan berbasis web sehingga mudah diakses tanpa memerlukan pemeriksaan laboratorium. Dengan demikian, sistem pakar ini tidak hanya memberikan prediksi yang akurat tetapi juga interpretatif, sehingga dapat membantu deteksi dini diabetes dan meningkatkan kesadaran masyarakat terhadap faktor risiko penyakit.

Kata Kunci: Sistem Pakar; Diabetes Mellitus; XGBoost; SHAP; *Machine Learning*

EXPERT SYSTEM IN DIABETES RISK *PREDICT*ION ANALYSIS USING THE XGBOOST MODEL AND RISK FACTOR ANALYSIS WITH EXPLAINABLE AI (SHAP)

ABSTRACT

Advances in Machine Learning in the field of healthcare offer opportunities for the early detection of diseases, including diabetes mellitus, which has a high prevalence and is often detected too late. Conventional screening methods have limitations in terms of accessibility and cost, while models such as XGBoost, although accurate, still have weaknesses in interpretability. Therefore, this study aims to develop a web-based expert system for Predicting diabetes risk using XGBoost and providing explanations of risk factors through the Explainable AI (SHAP) method. The research problem formulation includes analysis of system requirements, web-based system design, testing of Predictive performance and interpretation, as well as system implementation that can be used by the public. The method used is Data Mining with classification techniques using the XGBoost algorithm. The dataset consists of 520 data points from Kaggle with variables such as age, gender, and clinical symptoms of diabetes. The research stages include data cleaning, label encoding, model training, and SHAP integration for result interpretation. System design was performed using UML. The results show that the system is capable of Predicting diabetes risk effectively and provides a transparent explanation of risk factors. The implementation of SHAP allows users to understand the contribution of each variable to the Prediction results. The system is web-based, making it easily accessible without the need for laboratory testing. As a result, this expert system not only provides accurate Predictions but also interpretive insights, thereby aiding in the early detection of diabetes and raising public awareness of the disease's risk factors.

Keywords: *Expert System; Diabetes Mellitus; XGboost; SHAP; Machine Learning*

DAFTAR ISI

LEMBAR PENGESAHAN i
PERNYATAAN ORISINALITAS ii
RIWAYAT HIDUP iv
KATA PENGANTAR v
ABSTRAK viii
ABSTRACT ix
DAFTAR ISI x
DAFTAR TABEL xii
DAFTAR GAMBAR xiii
BAB I PENDAHULUAN 1
1.1. Latar Belakang Masalah 1
1.2. Rumusan Masalah 6
1.3. Batasan Masalah 7
1.4. Tujuan Penelitian 8
1.5. Manfaat Penelitian 8
BAB II LANDASAN TEORI 10
2.1. *Data Mining* 10
2.2. Sistem Pakar 11
2.3. *Machine Learning* Bidang Kesehatan 15
2.4. Extreme *Gradient boosting* (XGBoost) 19
2.5. Explainable AI (SHAP) 23
2.6. Resiko Diabetes 27
2.7. Unified *Model*ing Language (UML) 29
2.8. Penelitian Terdahulu 32
BAB III ANALISA DAN PERANCANGAN SISTEM 41
3.1. Analisis Permasalahan 41
3.2. Analisis Kebutuhan Sistem 43
3.2.1. Kebutuhan Fungsional Sistem 43
3.2.2. Kebutuhan Non-Fungsional 43
3.3. Perancangan Sistem 44
3.3.1. Usecase Diagram 44
3.3.2. Activity Diagram 45
3.3.3. Sequence Diagram 47
3.3.4. Class Diagram 48
3.4. Tahapan Pengolahan Data 49

3.4.1. Data Cleaning 49
3.4.2. Label *Encoding* 50
3.5. Populasi dan Sampel Penelitian 50
3.6. Validasi dan Analisis Algoritma 51
3.6.1. Extreme *Gradient boosting* (XGBoost) 51
3.6.2. Shapley Additive Explanations (SHAP) 61
BAB IV IMPLEMENTASI DAN PENGUJIAN SISTEM 83
4.1. Kebutuhan Sistem 83
4.2. Persiapan dan Pre-processing Dataset 85
4.3. Implementasi Sistem 87
4.4. Ringkasan Pengujian dan Evaluasi 100
BAB V PENUTUP 106
5.1. Kesimpulan 106
5.2. Saran 107
DAFTAR PUSTAKA 112
LAMPIRAN 115

DAFTAR TABEL

Tabel 2. 1 Simbol Usecase Diagram 30
Tabel 2. 2 Simbol Activity Diagram 30
Tabel 2. 3 Simbol Sequence Diagram 31
Tabel 2. 4 Simbol Class Diagram 31
Tabel 2. 5 Penelitian Terdahulu 36
Tabel 3. 1 Subset Training Xgboost 54
Tabel 3. 2 Probabilitas Awal Prediksi 54
Tabel 3. 3 Subset Fitur Resiko Tinggi 66
Tabel 3. 4 Subset Pasien Resiko Rendah 77
Tabel 3. 5 Subset Fitur Pasien Resiko Sedang 78
Tabel 3. 6 Verifikasi SHAP Values 79
Tabel 4. 1 *Endpoint* API 90

DAFTAR GAMBAR

Gambar 3. 1 Usecase Diagram 45
Gambar 3. 2 Activity Diagram 46
Gambar 3. 3 Sequence Diagram 47
Gambar 3. 4 Class Diagram 48
Gambar 3. 5 Node Pohon XGboost 55
Gambar 4. 1 Struktur Proyek 89
Gambar 4. 2 Arsitektur *Frontend* 93
Gambar 4. 3 Konponen Reusable 94
Gambar 4. 4 Halaman Home 95
Gambar 4. 5 Halaman Prediksi 96
Gambar 4. 6 Tampilan Resiko 97
Gambar 4. 7 SHAP Values 98
Gambar 4. 8 Tampilan Rekomendasi 98
Gambar 4. 9 Tampilan Informasi *Model* 99
Gambar 4. 10 Tampilan Indikator 100
Gambar 4. 11 Dokumentasi API 101
Gambar 4. 12 Testing Prediksi 102
Gambar 4. 13 Hasil Prediksi 103
Gambar 4. 14 Status Request 103
Gambar 4. 15 Status Response 200 104

BAB I

PENDAHULUAN

1.1. Latar Belakang Masalah

Perkembangan teknologi kecerdasan buatan dan *Machine Learning* dalam beberapa tahun terakhir telah membuka peluang besar dalam berbagai bidang, termasuk sektor kesehatan. Sistem pakar berbasis *Machine Learning* kini mampu mengadopsi pengetahuan dan kemampuan penalaran layaknya seorang ahli untuk membantu dalam proses diagnostik dan pengambilan keputusan medis. Salah satu algoritma *Machine Learning* yang menunjukkan performa superior adalah XGBoost (Extreme *Gradient boosting*), yang telah terbukti unggul dalam berbagai studi prediksi medis. Sebagai contoh, algoritma XGBoost telah menunjukkan akurasi yang lebih baik dalam hal performa prediksi dibandingkan dengan metode konvensional seperti regresi logistik atau *decision tree* sederhana (Li Wenguang AND Peng, 2024). Hal ini diperkuat dengan studi optimisasi diagnosis anemia menggunakan XGBoost dan beberapa algoritma lain dengan hasil akhir menunjukkan bahwa XGBoost unggul dalam segi performa, presisi, dan stabilitas *model* (Saputra et al., 2025). Keunggulan XGBoost terletak pada kemampuannya menangani data yang tidak seimbang (imbalanced data), mengatasi masalah *overfitting* melalui mekanisme regularisasi yang ketat, menangani *missing value*s secara otomatis, serta memberikan hasil prediksi dengan tingkat akurasi dan generalisasi yang tinggi bahkan pada dataset dengan karakteristik kompleks.

Namun demikian, penerapan *Machine Learning* dalam sistem pakar kesehatan menghadapi tantangan fundamental terkait interpretabilitas *model*.

Banyak *model Machine Learning*, termasuk XGBoost, bersifat "black-box" yang berarti proses pengambilan keputusan di dalam *model* sangat kompleks dan sulit untuk dijelaskan serta dipahami oleh pengguna non-teknis seperti tenaga medis, dokter, perawat, atau bahkan pasien itu sendiri. Dalam praktik medis, setiap keputusan harus dapat dijelaskan baik kepada pasien maupun dalam konteks dokumentasi medis dan pertanggungjawaban profesional. Sebuah sistem pakar yang efektif tidak hanya memberikan rekomendasi, tetapi juga harus mampu menjelaskan alasan di balik rekomendasi tersebut layaknya seorang ahli yang berkonsultasi dengan pasien. Di sinilah pentingnya penerapan Explainable AI (XAI) sebagai komponen integral dalam sistem pakar. Salah satu metode XAI yang paling populer, *robust*, dan efektif adalah SHapley Additive Explanations (SHAP), yang memungkinkan interpretasi kontribusi setiap fitur atau variabel terhadap hasil prediksi secara kuantitatif, konsisten, dan dapat divisualisasikan dengan jelas (Salih et al., 2025). SHAP bekerja berdasarkan konsep nilai Shapley dari teori permainan kooperatif yang mengalokasikan kontribusi setiap fitur secara adil dan proporsional terhadap prediksi akhir dengan mempertimbangkan semua kemungkinan kombinasi fitur. Dengan mengintegrasikan SHAP ke dalam sistem pakar, sistem tidak hanya memberikan output berupa hasil prediksi tetapi juga dapat memberikan penjelasan mendalam mengenai faktor-faktor spesifik mana yang paling berpengaruh dalam prediksi tersebut untuk setiap individu (Ponce-Bobadilla et al., 2024).

Salah satu bidang kesehatan yang sangat memerlukan sistem prediksi dini berbasis teknologi adalah penyakit Diabetes Mellitus. Diabetes merupakan salah satu masalah kesehatan global yang signifikan dengan tingkat prevalensi yang

terus meningkat dari tahun ke tahun. Menurut laporan *International Diabetes Federation* (IDF), diperkirakan terdapat 537 juta orang dewasa dengan diagnosa diabetes dalam usia rentang 20 sampai 70 tahun di seluruh dunia dan angka ini diprediksi akan terus meningkat sampai dengan 46% sampai dengan tahun 2045 (Sun et al., 2022). Lonjakan jumlah penderita ini tidak hanya terjadi di negara-negara maju, namun juga menyebar ke berbagai negara berkembang di Asia, Afrika, dan Amerika Latin. Negara-negara berpendapatan rendah hingga menengah mungkin menghadapi masalah sumber daya kesehatan yang tidak memadai untuk menangani peningkatan kasus diabetes yang cepat (Hossain et al., 2024). Angka tersebut menunjukkan bahwa diabetes telah menjadi ancaman serius bagi sistem kesehatan global dan memerlukan perhatian khusus dalam upaya pencegahan dan penanganan yang lebih efektif.

Namun, dengan besarnya prevalensi tersebut, sangat disayangkan karena dari segi deteksi dini dan pencegahan, situasinya masih jauh dari ideal. Tingginya *prevalensi* diabetes membawa dampak besar baik dari sisi kesehatan individu maupun beban ekonomi sistem kesehatan secara keseluruhan. Dari perspektif kesehatan individu, diabetes dapat memicu berbagai komplikasi serius seperti kerusakan saraf (neuropati diabetik), gangguan fungsi ginjal (nefropati diabetik), kerusakan retina mata (retinopati diabetik), hingga penyakit kardiovaskular yang dapat berujung pada serangan jantung, stroke, bahkan kematian dini (Upadhaya et al., 2025). Kondisi ini tidak hanya menurunkan kualitas hidup penderita, tetapi juga membebani keluarga dan sistem perawatan kesehatan. Sementara itu, dari segi ekonomi, biaya perawatan jangka panjang untuk penderita diabetes dan penanganan berbagai komplikasinya memberikan beban finansial yang sangat

besar bagi sistem kesehatan nasional maupun keluarga penderita. Biaya tersebut mencakup pemeriksaan rutin, pengobatan, rawat inap, hingga perawatan komplikasi yang memerlukan teknologi medis canggih.

Mengingat karakteristik penyakit diabetes yang bersifat progresif namun dapat dicegah apabila terdeteksi sejak dini, maka pengembangan sistem pakar untuk prediksi dini menjadi sangat krusial dan strategis. Deteksi risiko secara lebih awal melalui sistem pakar memungkinkan dilakukannya intervensi preventif seperti perubahan gaya hidup, pengaturan pola makan yang lebih sehat, peningkatan aktivitas fisik, atau pemberian terapi farmakologis sebelum kondisi berkembang menjadi lebih parah dan menimbulkan komplikasi yang sulit diobati (Larassati et al., 2022). Intervensi dini terbukti dapat menunda atau bahkan mencegah onset diabetes tipe 2 dan 1 pada individu yang berisiko tinggi. Namun demikian, metode skrining tradisional yang mengandalkan pemeriksaan laboratorium seperti tes gula darah puasa, tes toleransi glukosa oral, atau pemeriksaan HbA1c memiliki sejumlah keterbatasan signifikan dalam hal aksesibilitas, biaya pemeriksaan yang relatif mahal, waktu tunggu hasil yang cukup lama, serta kebutuhan akan fasilitas laboratorium yang memadai. Keterbatasan ini menjadi hambatan serius terutama di daerah-daerah dengan fasilitas kesehatan terbatas, wilayah pedesaan, atau masyarakat dengan kondisi ekonomi menengah ke bawah yang tidak memiliki akses mudah ke layanan kesehatan komprehensif. Oleh karena itu, diperlukan penggunaan sistem pakar berbasis kecerdasan buatan sebagai pendekatan alternatif yang lebih mudah diakses untuk mengenali pola risiko diabetes dan membangun *model* prediksi

yang akurat tanpa ketergantungan pada pemeriksaan laboratorium yang mahal dan invasif.

Integrasi antara sistem pakar, *model* prediktif XGBoost yang akurat, dan kemampuan interpretasi SHAP yang transparan menjadi kombinasi yang ideal dalam pengembangan sistem informasi kesehatan yang tidak hanya canggih secara teknologi tetapi juga dapat diterima, dipercaya, dan digunakan secara luas oleh praktisi kesehatan, pengambil kebijakan kesehatan, maupun masyarakat umum. Sistem pakar yang mampu memberikan penjelasan yang mudah dipahami, visual, dan relevan akan meningkatkan kepercayaan pengguna terhadap teknologi, mendorong adopsi yang lebih luas di berbagai level layanan kesehatan, meningkatkan literasi kesehatan masyarakat, dan pada akhirnya berkontribusi signifikan pada upaya pencegahan diabetes secara lebih efektif, terukur, dan berkelanjutan. Selain itu, sistem pakar dengan kemampuan eksplanasi yang baik juga dapat berfungsi sebagai alat edukasi yang membantu masyarakat memahami faktor-faktor risiko diabetes dan mendorong perubahan perilaku kesehatan yang lebih baik.

Dengan demikian, penelitian ini hadir dengan urgensi ganda yang sangat penting pertama, mengembangkan sistem pakar analisis prediksi risiko diabetes yang akurat dan andal menggunakan algoritma XGBoost dengan memanfaatkan data demografis dan klinis yang dapat dikumpulkan tanpa pemeriksaan laboratorium invasif, sehingga sistem dapat diakses oleh masyarakat luas dengan biaya minimal; dan kedua, memastikan bahwa sistem pakar tersebut dapat memberikan penjelasan yang sistematis, transparan, dan komprehensif melalui

implementasi metode SHAP sehingga layak untuk diintegrasikan dalam sistem informasi kesehatan dan dapat digunakan sebagai alat bantu pengambilan keputusan yang tidak hanya akurat tetapi juga dapat dipertanggungjawabkan secara ilmiah dan etis. Melalui penelitian ini, diharapkan dapat dihasilkan sebuah prototipe sistem pakar yang tidak hanya memberikan prediksi risiko diabetes secara otomatis tetapi juga memberikan analisis faktor risiko yang mendalam dan edukasi yang bermakna kepada pengguna mengenai faktor-faktor risiko spesifik yang mempengaruhi kondisi kesehatan mereka, sehingga mendorong peningkatan kesadaran kesehatan dan perubahan perilaku menuju gaya hidup yang lebih sehat, pola makan yang lebih baik, dan aktivitas fisik yang lebih teratur.

Berdasarkan latar belakang di atas, maka penulis mengangkat sebuah topik yang menghasilkan judul "Sistem Pakar dalam Analisis Prediksi Risiko Diabetes Menggunakan *Model* XGBoost dan Analisis Faktor Risiko dengan Explainable AI (SHAP)".

1.2. Rumusan Masalah

Berdasarkan latar belakang di atas, maka rumusan masalah penelitian ini yaitu:

1. Bagaimana menganalisis kebutuhan sistem pakar yang diperlukan untuk melakukan prediksi risiko diabetes dan analisis faktor risiko menggunakan *model* XGBoost dan SHAP?
2. Bagaimana merancang sistem pakar berbasis web yang mampu melakukan prediksi risiko diabetes menggunakan algoritma XGBoost dan memberikan analisis faktor risiko melalui metode Explainable AI (SHAP)?

3. Bagaimana menguji sistem dalam memprediksi risiko diabetes dan akurasi interpretasi SHAP dalam menganalisis faktor-faktor risiko yang mempengaruhi prediksi?
4. Bagaimana mengimplementasikan sistem pakar berbasis web yang terintegrasi dengan *model* XGBoost dan SHAP sehingga dapat digunakan oleh pengguna untuk prediksi risiko diabetes dengan penjelasan faktor risiko yang mudah dipahami?

1.3. Batasan Masalah

Agar penelitian tetap fokus dan terukur, maka ditetapkan beberapa batasan sebagai berikut:

1. Dataset yang digunakan adalah dataset diabetes yang bersumber dari platform kaggle dengan variabel faktor risiko sebagai berikut Usia (Age), Jenis kelamin (Gender), *Poliuria* (sering buang air kecil), *Polydipsia* (sering merasa haus), Penurunan berat badan secara tiba-tiba (*Sudden Weight Loss*), Kelemahan tubuh (*Weakness*), *Polyphagia* (sering merasa lapar), Gangguan penglihatan (*Visual Blurring*), Penyembuhan luka yang lambat (*Delayed Healing*), Obesitas (*Obesity*)
2. Penelitian ini tidak mengeksplorasi banyak algoritma secara mendalam hanya fokus pada algoritma XGBoost sebagai algoritma utama.
3. Proses tuning *model* dilakukan secara manual tanpa optimasi *hyperparameter* yang sangat kompleks.
4. Interpretasi *model* difokuskan pada SHAP untuk menjelaskan faktor-faktor penting, tanpa melakukan penelitian komparatif mendalam terhadap berbagai metode XAI lainnya.

5. Pengembangan sistem informasi bersifat prototipe sederhana yaitu berupa *web app* untuk demonstrasi, bukan sistem produksi penuh.

1.4. Tujuan Penelitian

Penelitian ini memiliki rumusan penelitian sebagai berikut:

1. Menganalisis kebutuhan sistem pakar untuk prediksi risiko diabetes dan mengidentifikasi variabel-variabel yang diperlukan dalam *model* XGBoost serta metode analisis faktor risiko menggunakan SHAP.
2. Merancang sistem pakar berbasis web yang mengintegrasikan algoritma XGBoost untuk prediksi risiko diabetes dengan metode SHAP untuk memberikan analisis dan penjelasan faktor-faktor risiko secara interaktif.
3. Menguji dan mengevaluasi sistem serta menguji kemampuan SHAP dalam memberikan interpretasi faktor risiko yang akurat dan mudah dipahami.
4. Mengimplementasikan sistem pakar berbasis *web* yang terintegrasi dengan *model* XGBoost dan visualisasi SHAP sehingga dapat diakses dan digunakan oleh masyarakat untuk melakukan prediksi risiko diabetes beserta penjelasan faktor-faktor risiko yang mempengaruhinya.

1.5. Manfaat Penelitian

Adapula manfaat penelitian sebagai berikut:

1. Memberikan gambaran yang jelas mengenai kebutuhan sistem pakar dalam melakukan prediksi risiko diabetes, termasuk identifikasi variabel-variabel yang relevan untuk membangun *model* prediksi menggunakan algoritma XGBoost serta analisis faktor risiko dengan metode SHAP.

2. Menyediakan rancangan sistem pakar berbasis web yang mengintegrasikan algoritma XGBoost dan metode SHAP, sehingga mampu menghasilkan prediksi risiko diabetes sekaligus memberikan penjelasan faktor-faktor risiko secara interaktif dan mudah dipahami oleh pengguna.
3. Menghasilkan sistem yang telah diuji dan dievaluasi kinerjanya, serta menunjukkan kemampuan metode SHAP dalam memberikan interpretasi faktor risiko yang akurat, transparan, dan dapat mendukung pemahaman pengguna terhadap hasil prediksi *model*.
4. Menghasilkan implementasi sistem pakar berbasis web yang dapat diakses dan digunakan oleh masyarakat sebagai sarana prediksi risiko diabetes, dilengkapi dengan visualisasi faktor risiko yang memengaruhi hasil prediksi sebagai upaya meningkatkan kesadaran dan pemahaman terhadap kondisi kesehatan.

BAB II

LANDASAN TEORI

2.1. *Data Mining*

Data Mining merupakan suatu proses untuk mengekstraksi informasi, pola, atau pengetahuan yang bernilai dari kumpulan data berukuran besar dengan menggunakan teknik statistik, matematika, dan kecerdasan buatan. Tujuan utama dari *Data Mining* adalah menemukan pola tersembunyi yang sebelumnya tidak diketahui dan dapat dimanfaatkan untuk mendukung pengambilan keputusan. *Data Mining* merupakan bagian dari proses yang lebih luas yang dikenal sebagai *Knowledge Discovery in Databases* (KDD) (Rizky et al., 2019). Proses KDD mencakup beberapa tahapan, yaitu seleksi data, pembersihan data, transformasi data, proses *Data Mining*, serta evaluasi dan interpretasi hasil. Melalui tahapan tersebut, data mentah diolah menjadi informasi yang bermakna. Dalam penerapannya, *Data Mining* memiliki berbagai fungsi, antara lain klasifikasi, klasterisasi, asosiasi, dan prediksi**.** Pada penelitian ini, fungsi *Data Mining* yang digunakan adalah klasifikasi, yaitu proses untuk mengelompokkan data ke dalam kelas tertentu berdasarkan karakteristik yang dimilikinya (Rizky et al., 2025). Klasifikasi digunakan untuk menentukan apakah seseorang termasuk ke dalam kelompok berisiko diabetes atau tidak berdasarkan faktor-faktor yang telah ditentukan. Pemanfaatan *Data Mining* dalam bidang kesehatan sangat penting, khususnya untuk mendukung deteksi dini suatu penyakit. Dengan memanfaatkan data gejala dan kondisi fisik pasien, *Data Mining* dapat membantu mengidentifikasi potensi risiko penyakit secara lebih cepat dan efisien. Oleh

karena itu, *Data Mining* dalam penelitian ini digunakan sebagai dasar dalam pembangunan sistem pakar deteksi dini risiko diabetes yang berbasis web dan ditujukan bagi masyarakat umum.

2.2. Sistem Pakar

Sistem pakar merupakan salah satu aplikasi kecerdasan buatan yang dirancang untuk meniru kemampuan pengambilan keputusan seorang ahli manusia dalam domain spesifik (Widyanti et al., 2020) ,(Budholiya et al., 2022). Sistem pakar mengkombinasikan *Knowledge Base* yang berisi pengetahuan faktual dan heuristik dari domain tertentu dengan *inference engine* yang dapat memanipulasi pengetahuan tersebut untuk menarik kesimpulan atau memberikan rekomendasi. Konsep sistem pakar pertama kali dikembangkan pada tahun 1960-an dan 1970-an dengan sistem seperti DENDRAL untuk analisis kimia dan MYCIN untuk diagnosis penyakit infeksi, yang menunjukkan bahwa komputer dapat meniru proses penalaran ahli manusia dalam domain terbatas.

Komponen utama sistem pakar meliputi *Knowledge Base*, *inference engine*, dan *user interface*. *Knowledge Base* menyimpan pengetahuan domain dalam berbagai bentuk seperti *rules*, *frames*, atau *semantic networks* yang merepresentasikan fakta, hubungan, dan heuristik yang digunakan oleh ahli dalam pengambilan keputusan. *Inference engine* bertugas untuk memanipulasi pengetahuan dalam *Knowledge Base* untuk menarik kesimpulan baru atau memberikan jawaban terhadap *query* pengguna, menggunakan strategi penalaran seperti *forward chaining* yang dimulai dari fakta untuk mencapai kesimpulan atau *backward chaining* yang dimulai dari goal untuk mencari fakta pendukung. *User*

interface memfasilitasi interaksi antara pengguna dan sistem, memungkinkan pengguna untuk memasukkan informasi, mengajukan pertanyaan, serta menerima penjelasan dan rekomendasi dari sistem dalam format yang mudah dipahami.

Dalam perkembangannya, sistem pakar modern telah berevolusi dengan mengintegrasikan teknik-teknik *Machine Learning* untuk mengatasi keterbatasan sistem pakar tradisional. Sistem pakar tradisional yang berbasis *rules* memiliki beberapa kelemahan *fundamental* termasuk kesulitan dalam *knowledge acquisition* dimana proses mengekstrak dan mengkodifikasi pengetahuan ahli memerlukan waktu lama dan mahal, *brittleness* dimana sistem tidak dapat menangani situasi di luar *scope rules* yang telah didefinisikan, serta kesulitan dalam *maintenance* ketika domain *knowledge* berkembang atau berubah. Integrasi *Machine Learning* memungkinkan sistem pakar untuk belajar dari data secara otomatis tanpa memerlukan *encoding manual* dari semua *rules*, dapat menangani *uncertainty* dan *incomplete information* dengan lebih baik, serta dapat beradaptasi ketika menerima data baru.

Sistem pakar berbasis *Machine Learning* dalam konteks kesehatan memanfaatkan algoritma pembelajaran untuk membangun *model* prediktif dari data pasien historis, sambil mengintegrasikan kemampuan *explanation* yang membuat keputusan *model* dapat dipahami dan diverifikasi oleh profesional medis. Pendekatan *hybrid* ini menggabungkan kekuatan pembelajaran otomatis dari data dengan transparansi dan interpretabilitas yang diperlukan dalam aplikasi medis. *Model Machine Learning* seperti XGBoost berfungsi sebagai mesin inferensi yang *powerful* , sementara metode *Explainable AI* seperti SHAP

menyediakan *layer explanation* yang memungkinkan sistem untuk menjelaskan alasan di balik prediksi atau rekomendasi yang diberikan.

Keuntungan sistem pakar dalam domain kesehatan sangat signifikan dan mencakup berbagai aspek. Pertama, sistem pakar dapat meningkatkan aksesibilitas layanan kesehatan dengan menyediakan kemampuan *diagnostic* atau *screening* di area dimana akses ke spesialis terbatas, terutama di daerah rural atau negara berkembang (Tamba et al., 2021). Kedua, sistem pakar dapat meningkatkan konsistensi dalam pengambilan keputusan medis dengan mengurangi variabilitas yang mungkin terjadi akibat perbedaan pengalaman, kelelahan, atau bias kognitif antar praktisi. Ketiga, sistem pakar dapat berfungsi sebagai *decision support tool* yang membantu dokter dengan menyediakan *second opinion* atau *flagging* kasus-kasus yang memerlukan perhatian khusus, sehingga mengurangi *diagnostic error* dan meningkatkan *patient safety*. Keempat, sistem pakar dapat memberikan edukasi kepada pasien dengan menjelaskan kondisi kesehatan mereka dan faktor-faktor yang mempengaruhinya dalam bahasa yang mudah dipahami, sehingga meningkatkan *health literacy* dan *patient engagement*.

Implementasi sistem pakar kesehatan berbasis *web* memberikan keuntungan tambahan dalam hal *deployment* dan aksesibilitas. *Platform web* memungkinkan sistem dapat diakses dari berbagai *device* tanpa memerlukan instalasi software khusus, memfasilitasi *update* dan *maintenance* yang terpusat tanpa perlu distribusi ke setiap pengguna, serta mendukung integrasi dengan *electronic health records* atau sistem informasi kesehatan lainnya. Arsitektur web-based juga memungkinkan *scalability* dimana sistem dapat melayani banyak pengguna secara

concurrent, serta memfasilitasi kolaborasi dimana *multiple stakeholders* dapat mengakses dan berkontribusi terhadap sistem.

Namun demikian, pengembangan dan *deployment* sistem pakar kesehatan juga menghadapi berbagai tantangan yang perlu dipertimbangkan dengan serius. Tantangan validasi klinis memerlukan bukti bahwa sistem memberikan rekomendasi yang *accurate* dan *safe* melalui *clinical trials* yang *rigorous* sebelum dapat digunakan dalam praktik klinis. Tantangan regulasi mencakup kepatuhan terhadap *regulasi medical devices* dan perlindungan data kesehatan seperti HIPAA di Amerika atau GDPR di Eropa yang mensyaratkan *safeguards* ketat untuk privasi dan keamanan data pasien. Tantangan integrasi workflow memerlukan desain sistem yang *seamlessly* terintegrasi dengan *workflow* klinis *existing* tanpa menambah beban kerja atau *disrupting care delivery*. Tantangan *acceptance* memerlukan strategi *change management* untuk memastikan bahwa *clinicians* dan *patients* menerima dan mempercayai sistem, termasuk training yang *adequate* dan *communication* yang efektif mengenai *capabilities* dan *limitations* sistem.

Aspek etis dalam pengembangan sistem pakar kesehatan juga merupakan konsiderasi penting yang tidak boleh diabaikan. *Fairness* dan *equity* harus dipastikan dimana sistem tidak menghasilkan *disparities* dalam kualitas prediksi atau rekomendasi across berbagai grup demografis atau sosial ekonomi, yang dapat terjadi jika *training* data tidak representatif atau mengandung *historical biases*. *Transparency* mengenai *limitations* sistem perlu dikomunikasikan dengan jelas kepada pengguna sehingga mereka memahami bahwa sistem adalah *tool*

yang membantu bukan menggantikan *clinical judgment*. *Accountability* perlu ditetapkan dengan jelas mengenai siapa yang bertanggung jawab ketika terjadi *error* atau *adverse outcomes* yang terkait dengan penggunaan sistem. *Informed consent* perlu diperoleh dari pasien mengenai penggunaan data mereka dan aplikasi AI dalam *care* mereka.

2.3. *Machine Learning* Bidang Kesehatan

Machine Learning merupakan cabang dari kecerdasan buatan yang memungkinkan sistem komputer untuk belajar dan meningkatkan performanya dari pengalaman tanpa diprogram secara eksplisit untuk setiap tugas spesifik (Sadr et al., 2025). Dalam konteks kesehatan, *Machine Learning* telah menunjukkan potensi yang sangat besar dalam berbagai aplikasi mulai dari diagnostik penyakit, prediksi risiko kesehatan, penemuan obat, personalisasi terapi, hingga analisis citra medis. Kemampuan *Machine Learning* untuk mengidentifikasi pola kompleks dalam data medis yang besar dan *multidimensional* menjadikannya alat yang sangat berharga dalam era digitalisasi kesehatan modern.

Penerapan *Machine Learning* dalam bidang kesehatan memberikan berbagai keuntungan signifikan (Shailaja et al., 2018). Pertama, *Machine Learning* mampu memproses dan menganalisis volume data yang sangat besar dalam waktu singkat, sesuatu yang tidak mungkin dilakukan secara manual oleh tenaga medis. Kedua, algoritma *Machine Learning* dapat mengidentifikasi pola dan hubungan kompleks antar variabel yang mungkin tidak terdeteksi oleh analisis statistik konvensional atau bahkan oleh intuisi klinis manusia. Ketiga, sistem berbasis *Machine Learning*

dapat memberikan prediksi atau rekomendasi yang konsisten dan objektif tanpa dipengaruhi oleh faktor subjektif seperti kelelahan atau *bias* kognitif yang dapat mempengaruhi pengambilan keputusan manusia. Keempat, *Machine Learning* memungkinkan personalisasi dalam perawatan kesehatan dengan memprediksi *respons* individual terhadap berbagai intervensi berdasarkan karakteristik unik setiap pasien.

Dalam konteks prediksi penyakit, *Machine Learning* menawarkan pendekatan yang lebih canggih dibandingkan metode statistik tradisional (Ahsan et al., 2022). Metode tradisional seperti *logistic regression* umumnya mengasumsikan hubungan linear antara *variabel* prediktor dan *outcome*, padahal hubungan dalam data medis seringkali bersifat non-linear dan melibatkan interaksi kompleks antar variabel. *Machine Learning*, khususnya metode *ensemble* learning dan deep learning, mampu menangkap hubungan non-linear dan interaksi kompleks ini tanpa memerlukan spesifikasi manual dari peneliti. Selain itu, *Machine Learning* dapat bekerja dengan baik pada dataset dengan dimensi tinggi dimana jumlah variabel prediktor sangat banyak, kondisi yang seringkali menyebabkan masalah pada metode statistik konvensional.

Namun demikian, penerapan *Machine Learning* dalam kesehatan juga menghadapi berbagai tantangan yang perlu diatasi. Tantangan pertama adalah ketersediaan data berkualitas tinggi dalam jumlah yang cukup untuk melatih *model* yang *robust*. Data medis seringkali tidak lengkap, tidak konsisten, atau mengandung bias yang dapat mempengaruhi kualitas *model*. Tantangan kedua adalah interpretabilitas *model*, dimana banyak algoritma *Machine Learning* yang

powerful seperti *neural network* dan *ensemble methods* bersifat *black-box* sehingga sulit untuk memahami bagaimana *model* sampai pada suatu keputusan. Hal ini menjadi masalah serius dalam konteks medis dimana transparansi dan akuntabilitas sangat penting. Tantangan ketiga adalah validasi eksternal dan generalisasi *model*, dimana *model* yang bekerja baik pada satu populasi atau setting mungkin tidak bekerja dengan baik pada populasi atau setting lain. Tantangan keempat adalah aspek regulasi dan etika, termasuk privasi data pasien, *informed consent*, dan tanggung jawab ketika terjadi kesalahan prediksi yang berdampak pada pasien.

Meskipun menghadapi berbagai tantangan, tren penerapan *Machine Learning* dalam kesehatan terus meningkat dengan dukungan dari perkembangan teknologi komputasi, ketersediaan *big data* kesehatan, dan semakin banyaknya bukti empiris mengenai efektivitas pendekatan ini. Integrasi *Machine Learning* dalam *workflow* klinis diharapkan dapat meningkatkan kualitas dan efisiensi layanan kesehatan, mengurangi *medical error*, serta memfasilitasi transisi menuju *precision medicine* dimana setiap pasien mendapatkan intervensi yang optimal sesuai dengan karakteristik individual mereka.

Sistem pakar merupakan salah satu aplikasi kecerdasan buatan yang dirancang untuk meniru kemampuan pengambilan keputusan seorang ahli manusia dalam domain spesifik (Budholiya et al., 2022). Sistem pakar mengkombinasikan *Knowledge Base* yang berisi pengetahuan faktual dan heuristik dari domain tertentu dengan *inference engine* yang dapat memanipulasi pengetahuan tersebut untuk menarik kesimpulan atau memberikan rekomendasi.

Konsep sistem pakar pertama kali dikembangkan pada tahun 1960-an dan 1970-an dengan sistem seperti DENDRAL untuk analisis kimia dan MYCIN untuk diagnosis penyakit infeksi, yang menunjukkan bahwa komputer dapat meniru proses penalaran ahli manusia dalam domain terbatas.

Komponen utama sistem pakar meliputi *Knowledge Base*, *inference engine*, dan *user interface*. *Knowledge Base* menyimpan pengetahuan domain dalam berbagai bentuk seperti *rules*, *frames*, atau *semantic networks* yang merepresentasikan fakta, hubungan, dan heuristik yang digunakan oleh ahli dalam pengambilan keputusan. *Inference engine* bertugas untuk memanipulasi pengetahuan dalam *Knowledge Base* untuk menarik kesimpulan baru atau memberikan jawaban terhadap *query* pengguna, menggunakan strategi penalaran seperti *forward chaining* yang dimulai dari fakta untuk mencapai kesimpulan atau *backward chaining* yang dimulai dari *goal* untuk mencari fakta pendukung. *User interface* memfasilitasi interaksi antara pengguna dan sistem, memungkinkan pengguna untuk memasukkan informasi, mengajukan pertanyaan, serta menerima penjelasan dan rekomendasi dari sistem dalam format yang mudah dipahami.

Dalam perkembangannya, sistem pakar modern telah berevolusi dengan mengintegrasikan teknik-teknik *Machine Learning* untuk mengatasi keterbatasan sistem pakar tradisional. Sistem pakar tradisional yang berbasis *rules* memiliki beberapa kelemahan *fundamental* termasuk kesulitan dalam *knowledge acquisition* dimana proses mengekstrak dan mengkodifikasi pengetahuan ahli memerlukan waktu lama dan mahal, *brittleness* dimana sistem tidak dapat menangani situasi di luar *scope rules* yang telah didefinisikan, serta kesulitan

dalam *maintenance* ketika *domain knowledge* berkembang atau berubah. Integrasi *Machine Learning* memungkinkan sistem pakar untuk belajar dari data secara otomatis tanpa memerlukan *encoding* manual dari semua *rules*, dapat menangani *uncertainty* dan *incomplete information* dengan lebih baik, serta dapat beradaptasi ketika menerima data baru.

2.4. Extreme *Gradient boosting* (XGBoost)

Extreme Gradient boosting (XGBoost) merupakan algoritma *Machine Learning* berbasis *ensemble learning* yang termasuk ke dalam keluarga *Gradient boosting Decision tree* (GBDT) (Chen & Guestrin, 2016). XGBoost dikembangkan oleh Tianqi Chen dengan tujuan meningkatkan efisiensi komputasi, akurasi prediksi, serta kemampuan generalisasi *model* dibandingkan algoritma boosting konvensional.

XGBoost bekerja dengan membangun *model* secara bertahap (additive training), di mana setiap *model* baru berfungsi untuk memperbaiki kesalahan prediksi dari *model* sebelumnya. Algoritma ini banyak digunakan pada berbagai permasalahan klasifikasi dan regresi karena kemampuannya menangani data berdimensi tinggi, *missing value*, serta *overfitting* melalui teknik regularisasi. *Gradient boosting* merupakan metode *ensemble* yang mengombinasikan beberapa *weak learner* (biasanya pohon keputusan) menjadi satu *model* yang kuat (*strong learner*). Prinsip utama *Gradient boosting* adalah meminimalkan fungsi kerugian (*loss function*) dengan menggunakan pendekatan optimasi berbasis gradien.

Secara umum, *model Gradient boosting* dapat dinyatakan sebagai:

$$\boldsymbol{yi} = \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{fk}(\boldsymbol{xi}) \tag{1}$$

dengan:

- yi adalah nilai prediksi,
- fk adalah fungsi pohon keputusan ke-kkk,
- K adalah jumlah pohon.

Setiap pohon dibangun untuk mempelajari residual atau kesalahan dari prediksi sebelumnya sehingga performa *model* meningkat secara iteratif.

Keunggulan utama XGBoost terletak pada perumusan fungsi objektif yang menggabungkan fungsi kerugian dan regularisasi. Fungsi objektif XGBoost dirumuskan sebagai berikut:

$$l = \sum_{i=1}^{n} l(\bar{y}i, yi) + \sum_{k=1}^{K} \Omega(fk) \tag{2}$$

di mana:

- $l(\bar{y}i, yi)$ adalah fungsi kerugian,
- $\Omega(fk)$ adalah fungsi regularisasi,
- n adalah jumlah data.

Fungsi regularisasi didefinisikan sebagai:

$$\Omega(f) = \gamma T + \frac{1}{2}\,\lambda \sum_{j=1}^{T} w\frac{2}{j} \tag{3}$$

dengan:

- T adalah jumlah daun (leaf) pada pohon,
- wj adalah bobot daun ke-j,
- $\gamma\ dan\ \lambda$ adalah parameter regularisasi.

Regularisasi ini berfungsi untuk mengontrol kompleksitas *model* dan mengurangi risiko *overfitting*.

Proses pelatihan XGBoost menggunakan pendekatan Taylor Expansion orde kedua terhadap fungsi kerugian untuk mempercepat optimasi. Dengan memanfaatkan gradien pertama dan kedua (*first-order* dan *second-order derivative*), XGBoost mampu menentukan pembagian node pohon secara lebih optimal.

Langkah umum pelatihan XGBoost meliputi:

1. Inisialisasi *model* dengan prediksi awal.
2. Menghitung gradien dan hessian dari fungsi kerugian.
3. Membangun pohon keputusan berdasarkan nilai gradien dan hessian.
4. Menambahkan pohon ke dalam *model* secara bertahap.

5. Mengulangi proses hingga mencapai jumlah iterasi atau konvergensi

XGBoost memiliki sejumlah kelebihan yang menjadikannya salah satu algoritma *Machine Learning* yang banyak digunakan dalam berbagai penelitian. Salah satu keunggulan utama XGBoost adalah tingkat akurasi yang tinggi pada berbagai jenis dataset, baik untuk permasalahan klasifikasi maupun regresi. Algoritma ini mampu menangani data berukuran besar dan berdimensi tinggi secara efektif karena dirancang dengan optimasi komputasi yang efisien. Selain itu, XGBoost mendukung penanganan *missing value* secara otomatis dengan menentukan arah pemisahan node yang paling optimal ketika data tidak lengkap, sehingga tidak memerlukan proses imputasi yang kompleks. Penerapan teknik regularisasi dalam fungsi objektif XGBoost juga berperan penting dalam mengurangi risiko *overfitting*, terutama pada *dataset* dengan jumlah fitur yang banyak. Dari sisi kinerja, XGBoost dikenal efisien dalam waktu komputasi dan penggunaan memori karena mengimplementasikan pemrosesan paralel serta optimasi struktur data yang baik.

Meskipun memiliki berbagai kelebihan, XGBoost juga memiliki beberapa keterbatasan yang perlu diperhatikan. Proses penentuan nilai *hyperparameter* pada XGBoost relatif kompleks karena melibatkan banyak parameter yang saling memengaruhi kinerja *model*, sehingga membutuhkan eksperimen dan waktu yang cukup untuk memperoleh konfigurasi optimal. Selain itu, interpretabilitas *model* XGBoost cenderung lebih rendah dibandingkan dengan *model* linear, karena hasil prediksi diperoleh dari kombinasi banyak pohon keputusan yang bersifat *nonlinier*. Pada dataset yang sangat besar, penggunaan XGBoost juga dapat

memerlukan sumber daya komputasi yang cukup tinggi, baik dari sisi memori maupun daya pemrosesan, sehingga perlu disesuaikan dengan kemampuan perangkat keras yang tersedia.

2.5. Explainable AI (SHAP)

Explainable AI atau XAI merupakan area penelitian dalam kecerdasan buatan yang berfokus pada pengembangan metode dan teknik untuk membuat keputusan sistem AI dapat dipahami dan dijelaskan kepada manusia (Barredo Arrieta et al., 2020). Dalam konteks *Machine Learning*, XAI menjadi semakin penting seiring dengan meningkatnya penggunaan *model-model* kompleks yang *powerful* namun sulit diinterpretasi. Kebutuhan akan XAI muncul dari kesadaran bahwa akurasi tinggi saja tidak cukup untuk membuat sistem AI dapat diterima dan dipercaya, terutama dalam *domain-domain* kritis seperti kesehatan, keuangan, dan hukum dimana keputusan memiliki konsekuensi signifikan terhadap kehidupan manusia.

Pentingnya XAI dalam aplikasi kesehatan tidak dapat diabaikan karena beberapa alasan *fundamental*. Pertama, dari perspektif klinis, dokter dan tenaga medis memerlukan pemahaman tentang mengapa sistem memberikan suatu prediksi atau rekomendasi sebelum mereka dapat mengintegrasikan informasi tersebut dalam pengambilan keputusan medis. Tanpa pemahaman ini, profesional kesehatan akan kesulitan untuk menilai apakah prediksi *model* masuk akal secara klinis atau hanya merupakan artefak dari pola *spurious* dalam data *training*. Kedua, dari perspektif pasien, mereka memiliki hak untuk memahami dasar dari diagnosis atau rekomendasi yang diberikan, terutama ketika keputusan tersebut

akan mempengaruhi pilihan terapi dan kualitas hidup mereka. Transparansi ini penting untuk membangun trust dan memfasilitasi *shared decision making* antara dokter dan pasien. Ketiga, dari perspektif regulasi dan hukum, sistem prediksi medis harus dapat diaudit dan dipertanggungjawabkan, dimana stakeholder perlu dapat memeriksa apakah sistem membuat keputusan berdasarkan faktor-faktor yang medis dan etis dapat diterima.

Berbagai metode XAI telah dikembangkan untuk mengatasi tantangan interpretabilitas *model Machine Learning*. Metode-metode ini dapat dikategorikan menjadi *model*-specific methods yang dirancang khusus untuk jenis *model* tertentu, dan *model*-agnostic methods yang dapat diterapkan pada berbagai jenis *model*. Contoh *model*-specific methods termasuk interpretasi coefficient pada *linear regression* atau *logistic regression*, serta visualisasi attention weights pada *neural networks*. Sementara itu, *model*-agnostic methods seperti Partial Dependence Plots menunjukkan bagaimana prediksi *model* berubah ketika nilai suatu fitur divariasikan sambil menjaga fitur lain konstan, dan LIME atau *Local Interpretable Model-agnostic Explanations* yang membangun *model* linear sederhana secara lokal di sekitar instance yang ingin dijelaskan untuk mengaproksimasi perilaku *model* kompleks pada *region* tersebut.

SHAP atau *SHapley Additive exPlanations* merupakan salah satu metode XAI yang paling populer dan teoritis sound untuk memberikan interpretasi pada prediksi *model Machine Learning* (Ponce-Bobadilla et al., 2024). SHAP dikembangkan oleh Scott Lundberg dan Su-In Lee pada tahun 2017 berdasarkan konsep Shapley values dari teori permainan kooperatif yang dikembangkan oleh

Lloyd Shapley pada tahun 1953. Dalam teori permainan, Shapley values digunakan untuk mendistribusikan *payoff* secara adil kepada pemain dalam permainan kooperatif berdasarkan kontribusi mereka terhadap total payoff. SHAP mengadaptasi konsep ini untuk konteks *Machine Learning* dimana fitur-fitur input dianggap sebagai pemain dan prediksi *model* dianggap sebagai *payoff* yang perlu didistribusikan.

Prinsip kerja SHAP adalah menghitung kontribusi setiap fitur terhadap prediksi dengan mempertimbangkan semua kemungkinan kombinasi fitur. Untuk setiap fitur, SHAP value mengukur rata-rata perubahan prediksi ketika fitur tersebut ditambahkan ke berbagai subset dari fitur lainnya. Secara matematis, SHAP value untuk fitur i dihitung dengan mengambil *weighted average* dari *marginal contribution* fitur tersebut terhadap prediksi pada semua kemungkinan *coalition* atau *subset* dari fitur lainnya. *Weight* atau bobot untuk setiap coalition ditentukan berdasarkan ukurannya, dimana *coalition* yang lebih kecil dan lebih besar mendapat bobot lebih tinggi dibandingkan *coalition* dengan ukuran *medium*. Pendekatan ini memastikan bahwa kontribusi setiap fitur dihitung secara *fair* dengan mempertimbangkan konteks dari fitur-fitur lain yang ada.

Keunggulan utama SHAP dibandingkan metode interpretabilitas lainnya terletak pada sifat-sifat teoritis yang solid dan desirable. Pertama, SHAP memenuhi properti *local accuracy* yang berarti jumlah dari semua SHAP values untuk semua fitur akan tepat sama dengan perbedaan antara prediksi untuk *instance* tersebut dan prediksi rata-rata *model*. Properti ini memastikan bahwa penjelasan SHAP faithfully merepresentasikan prediksi aktual *model*. Kedua,

SHAP memenuhi properti *missingness* dimana fitur yang tidak ada atau missing akan memiliki SHAP value nol, sesuai dengan intuisi bahwa fitur yang tidak ada tidak seharusnya berkontribusi terhadap prediksi. Ketiga, SHAP memenuhi properti consistency yang menyatakan bahwa jika *model* berubah sehingga kontribusi suatu fitur meningkat atau tetap sama untuk semua kemungkinan input, maka SHAP value untuk fitur tersebut tidak akan menurun.

SHAP menyediakan berbagai tipe visualisasi yang memfasilitasi interpretasi prediksi pada level yang berbeda. *Waterfall plot* atau *force plot* menunjukkan bagaimana setiap fitur berkontribusi untuk menggeser prediksi dari nilai baseline atau *expected value* menuju prediksi akhir untuk satu *instance* spesifik, sangat berguna untuk menjelaskan prediksi individual kepada dokter atau pasien. Summary plot menampilkan distribusi SHAP values untuk setiap fitur across semua *instances* dalam *dataset*, memberikan gambaran *global* tentang pentingnya dan efek setiap fitur terhadap prediksi *model*. Dependence plot menunjukkan hubungan antara nilai fitur dan SHAP value-nya, memungkinkan identifikasi hubungan *non-linear* atau *threshold effects* serta interaksi antar fitur. *Decision plot* menampilkan bagaimana kontribusi fitur berakumulasi dari *expected value* menuju prediksi akhir, berguna untuk membandingkan penjelasan antara beberapa instances dan mengidentifikasi pola dalam pengambilan keputusan *model*.

Implementasi SHAP untuk berbagai jenis *model* telah dikembangkan dengan algoritma yang efisien. Untuk *tree-based model*s seperti XGBoost, TreeSHAP menyediakan perhitungan exact SHAP values dengan kompleksitas *polynomial* time dengan memanfaatkan struktur *tree*, jauh lebih cepat

dibandingkan perhitungan *exact Shapley values* yang memiliki kompleksitas eksponensial. Untuk *model* lain, KernelSHAP menyediakan aproksimasi SHAP *values* dengan sampling berbagai *coalition* dan *fitting weighted linear regression*. Meskipun KernelSHAP lebih lambat dan menghasilkan aproksimasi dibanding TreeSHAP, metode ini bersifat *model-agnostic* dan dapat diterapkan pada berbagai jenis *model* termasuk *neural networks*.

Penerapan SHAP dalam konteks prediksi medis memberikan berbagai manfaat praktis. SHAP memungkinkan identifikasi faktor risiko yang paling berpengaruh untuk setiap pasien individual, informasi yang sangat berharga untuk personalisasi intervensi dan edukasi. SHAP juga dapat membantu validasi klinis dari *model* dengan memeriksa apakah faktor-faktor yang dianggap penting oleh *model* sesuai dengan pengetahuan *domain* medis. Jika *model* memberikan *importance* tinggi pada fitur yang tidak relevan secara klinis atau mengabaikan faktor risiko yang diketahui penting, hal ini mengindikasikan adanya masalah pada data atau *model* yang perlu diatasi. Selain itu, SHAP dapat memfasilitasi deteksi bias dalam *model* dengan menganalisis apakah fitur-fitur sensitif seperti ras atau *gender* memiliki kontribusi yang tidak proporsional terhadap prediksi, sehingga membantu memastikan *fairness* dan *equity* dalam aplikasi AI kesehatan.

2.6. Resiko Diabetes

Diabetes merupakan penyakit metabolik kronis yang ditandai dengan meningkatnya kadar glukosa dalam darah akibat gangguan produksi insulin, kerja insulin, atau keduanya. Kondisi ini dapat berkembang secara perlahan dan sering kali tidak disadari oleh penderitanya pada tahap awal. Oleh karena itu,

pemahaman mengenai risiko diabetes menjadi penting sebagai upaya pencegahan dan deteksi dini. Risiko diabetes dipengaruhi oleh berbagai faktor, baik faktor demografis, gaya hidup, maupun gejala fisik yang muncul. Faktor demografis seperti usia dan jenis kelamin diketahui berpengaruh terhadap kemungkinan seseorang mengalami gangguan metabolisme glukosa. Seiring bertambahnya usia, risiko diabetes cenderung meningkat akibat penurunan fungsi metabolik tubuh.

Selain faktor demografis, gejala klasik diabetes juga menjadi indikator penting dalam penilaian risiko. Gejala-gejala tersebut antara lain sering buang air kecil (*Polyuria*), sering merasa haus (*Polydipsia*), sering merasa lapar (*Polyphagia*), serta penurunan berat badan secara tiba-tiba. Gejala ini muncul akibat tubuh tidak mampu menggunakan glukosa secara optimal sebagai sumber energi. Risiko diabetes juga dapat dikenali melalui kondisi fisik tertentu seperti kelelahan atau kelemahan tubuh, gangguan penglihatan, lambatnya penyembuhan luka, serta obesitas. Kondisi-kondisi tersebut sering muncul pada individu dengan kadar gula darah yang tidak terkontrol dan dapat menjadi tanda awal terjadinya diabetes. Dalam konteks deteksi dini, risiko diabetes dapat diidentifikasi melalui pendekatan *self-assessment*, yaitu berdasarkan gejala dan kondisi fisik yang dirasakan oleh individu tanpa memerlukan pemeriksaan laboratorium. Pendekatan ini memungkinkan masyarakat umum untuk mengenali potensi risiko diabetes lebih awal sehingga dapat mengambil langkah pencegahan atau berkonsultasi dengan tenaga medis. Oleh karena itu, dalam penelitian ini risiko diabetes didefinisikan sebagai kemungkinan seseorang mengalami diabetes berdasarkan faktor demografis dan gejala utama yang dapat diamati secara mandiri**,** yang

selanjutnya dianalisis menggunakan metode *Data Mining* untuk mendukung pembangunan sistem pakar deteksi dini berbasis web.

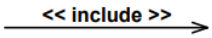
2.7. Unified *Model*ing Language (UML)

Unified Modeling Language (UML) merupakan bahasa pe*model*an standar yang digunakan untuk memvisualisasikan, merancang, dan mendokumentasikan sistem berbasis perangkat lunak. UML dikembangkan untuk memberikan cara yang sistematis dan terstruktur dalam menggambarkan kebutuhan sistem, alur proses, serta interaksi antar komponen sistem. Dengan menggunakan UML, pengembang dan pemangku kepentingan dapat memiliki pemahaman yang sama mengenai struktur dan perilaku sistem yang akan dibangun.

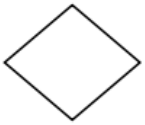
Dalam pengembangan sistem informasi, UML berperan penting sebagai alat bantu analisis dan perancangan sistem sebelum tahap implementasi. UML membantu dalam mengidentifikasi kebutuhan fungsional sistem, menggambarkan alur kerja sistem, serta menunjukkan bagaimana komponen-komponen sistem saling berinteraksi. Oleh karena itu, UML banyak digunakan dalam tahap perancangan untuk meminimalkan kesalahan desain dan meningkatkan kualitas sistem yang dihasilkan.

Pada penelitian ini, UML digunakan untuk me*model*kan sistem pakar deteksi dini risiko diabetes berbasis *web*. Diagram UML yang digunakan meliputi *Use Case Diagram* untuk menggambarkan fungsi sistem dan interaksi pengguna, *Activity Diagram* untuk menunjukkan alur proses deteksi risiko diabetes, serta *Sequence Diagram* untuk memvisualisasikan urutan komunikasi antar komponen sistem dalam menghasilkan prediksi dan penjelasan risiko diabetes.

Tabel 2. 1 Simbol Usecase Diagram

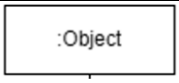
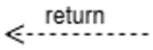
No	Simbol	Nama Simbol	Keterangan
1		Actor	Merepresentasikan pihak yang berinteraksi dengan sistem, yaitu pengguna
2		Use Case	Menunjukkan fungsi atau layanan yang disediakan oleh sistem
3		System Boundary	Menunjukkan batas ruang lingkup sistem
4		Association	Komunikasi antara actor dan use case yang memiliki interaksi dengan actor.
5		Include	Relasi use case tambahan ke sebuah use case dimana use case yang ditambahkan memerlukan use case ini untuk menjalanlan fungsinya.

Tabel 2. 2 Simbol Activity Diagram

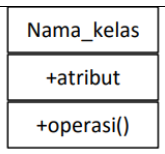
No	Simbol	Nama Simbol	Keterangan
1		Start Point	Menandakan suatu proses awal aktivitas.
2		Activities	Menunjukkan aktivitas atau proses yang dilakukan
3		Decision	Menunjukkan percabangan berdasarkan kondisi tertentu

4		Control Flow	Menunjukkan alur perpindahan antar aktivitas
5		End Point	Menandakan akhir proses aktivitas

Tabel 2. 3 Simbol Sequence Diagram

No	Simbol	Nama Simbol	Keterangan
1		Object	Merepresentasikan objek atau komponen sistem
2		Lifetime	Menunjukkan keberadaan objek selama proses berlangsung
3		Message	Menunjukkan pesan atau komunikasi antar objek
4		Activation	Menunjukkan alur perpindahan antar aktivitas
5		Return Message	Menandakan akhir proses aktivitas
6		Actor	Merepresentasikan pihak yang berinteraksi dengan sistem, yaitu pengguna

Tabel 2. 4 Simbol Class Diagram

No	Simbol	Nama Simbol	Keterangan
1		Class	Representasi struktur sistem
2		Association	Hubungan antar objek

3	----------	Dependencies	Ketergantungan pemanggilan

2.8. Penelitian Terdahulu

Pendekatan *Machine Learning* seperti XGBoost, *Random Forest*, *LightGBM*, dan *model hybrid* telah digunakan untuk mengembangkan *model* prediksi diabetes dalam beberapa penelitian sebelumnya. Secara umum, penelitian-penelitian tersebut menekankan betapa pentingnya menggunakan algoritma berbasis pohon keputusan yang mampu menangani data medis yang rumit untuk mendeteksi diabetes sejak dini. Namun, penelitian masih terbatas pada analisis performa algoritma tanpa memberikan penjelasan mendalam tentang faktor risiko dengan menggunakan pendekatan AI yang dapat dijelaskan seperti SHAP.

Kemudian untuk memprediksi risiko diabetes, (Wardhani & Akbar, 2022) menggunakan algoritma *Extreme Gradient boosting* (XGBoost) dari dataset *Early-Stage Diabetes Risk Prediction*, yang terdiri dari 520 data. Hasil penelitian menunjukkan bahwa XGBoost mampu memberikan kinerja yang sangat baik dengan akurasi 98,71%, *precision* sebesar 100%, dan recall 96,43%. Penelitian ini menunjukkan bahwa XGBoost efektif dalam me*model*kan hubungan antara variabel medis seperti *Polydipsia*, *Polyuria*, dan penurunan berat badan mendadak sebagai faktor utama dalam memprediksi diabetes. Karena penelitian ini tidak menggunakan teknik interpretabilitas seperti SHAP, penelitian ini belum menyelidiki lebih lanjut bagaimana fitur-fitur tersebut berkontribusi secara khusus terhadap prediksi.

Penelitian yang dilakukan oleh (Kurniawan et al., 2025) membandingkan kinerja tiga algoritma klasifikasi: *Random Forest*, *XGBoost*, dan *LightGBM*, saat menggunakan dataset NIDDK untuk diagnosis diabetes dini. Pra-pemrosesan melibatkan penanganan ketidakseimbangan kelas menggunakan SMOTE, dan kemudian evaluasi menggunakan akurasi, *precision*, *recall*, *F1-score*, dan AUC. Hasil menunjukkan bahwa *Random Forest* adalah yang terbaik dengan AUC 86%, diikuti oleh XGBoost (83%) dan LightGBM (82%). Studi ini menunjukkan bahwa ketiga *model gradient boosting* dan *ensemble* dapat mendeteksi diabetes dengan cukup akurat, meskipun XGBoost bukan yang terbaik. Namun, penelitian ini lebih menekankan pada perbandingan performa daripada melakukan analisis interpretabilitas untuk menjelaskan faktor risiko secara lebih jelas.

Selanjutnya, (Muhammad Surono et al., 2025) mengembangkan *model* Hybrid XGBoost-SVM untuk prediksi diabetes dengan data *Pima Indian Diabetes*. Tujuan dari *model* hybrid ini adalah untuk menggabungkan kekuatan XGBoost dalam seleksi fitur bersama dengan SVM sebagai *classifier* akhir. Hasil eksperimen menunjukkan bahwa *model hybrid* memiliki akurasi tertinggi, yaitu 77%, mengalahkan *model* tunggal XGBoost dan SVM dengan akurasi 75% masing-masing. Studi ini menunjukkan bahwa penggunaan *model* dapat meningkatkan kinerja, terutama dengan mengurangi *false negative*. Namun, karena penelitian ini tidak memasukkan pendekatan yang dapat dijelaskan untuk AI, peran fitur dalam prediksi tidak dijelaskan secara menyeluruh.

Adapun, Penelitian yang dilakukan oleh (Nurul Puteri & Sumardin, 2024) mencoba membandingkan kinerja algoritma *Random Forest* dan XGBoost dalam

klasifikasi diabetes serta meningkatkan interpretabilitas *model* dengan menggunakan teknik *Explainable AI*, seperti SHAP dan LIME. Hasil penelitian menunjukkan bahwa akurasi XGBoost sebesar 90,6%, sementara *Random Forest* memperoleh 90,3%. Selain itu, penelitian tersebut menemukan variabel usia, kadar glukosa dalam darah, dan tingkat insulin dalam darah melalui visualisasi Hasilnya menunjukkan bahwa transparansi *model* dapat ditingkatkan dengan integrasi XAI, dan penjelasan pengambilan keputusan pada sistem berbasis *Machine Learning* dapat dibantu.

Penelitian tambahan yang dilakukan oleh (Adhi Pratama & Wahyu Utomo, 2025) mencakup pengembangan *model* deteksi diabetes yang menggunakan algoritma XGBoost dan LightGBM. Sebelum proses pelatihan *model*, teknik SMOTE-Tomek diterapkan untuk mengatasi ketidakseimbangan data. Dengan menggunakan *cross-validation K-Fold*, evaluasi menunjukkan bahwa XGBoost memperoleh akurasi sebesar 97,31% dan LightGBM sebesar 97,26%. Selain itu, teknik *blending* dari kedua *model* meningkatkan akurasi menjadi 97,51%. Hasil ini menunjukkan bahwa *ensemble learning* berbasis peningkatan sangat efektif dalam mendeteksi diabetes dan dapat membuat sistem yang tepat untuk diagnosis dini.

Selanjutnya, (R Abdurrosyid et al., 2025) mengembangkan sistem deteksi dini diabetes yang menggunakan kombinasi *Principal Component Analysis* (PCA) dan XGBoost. PCA digunakan untuk mengurangi dimensi data dan *noise* sehingga fitur yang paling penting dapat dipertahankan. Dibandingkan dengan XGBoost tanpa PCA, hasil eksperimen menunjukkan peningkatan akurasi

prediksi sebesar 5,4% dan peningkatan skor F1 sebesar 6,45%. Tetapi penelitian ini juga menunjukkan penurunan *recall*, jadi masih diperlukan optimasi lanjutan untuk mengurangi kesalahan deteksi kasus positif.

Penelitian yang dilakukan oleh (Abdurrahman et al., 2022) berkonsentrasi pada pengoptimalan algoritma XGBoost Classifier untuk klasifikasi penyakit diabetes dengan menggunakan teknik pengaturan *hyperparameter* seperti pencarian *grid* dan pencarian acak. Penelitian ini menggunakan dataset *Pima Indians Diabetes* dari Repositori Pembelajaran Mesin UCI, dan menunjukkan bahwa performa *model* secara signifikan meningkat ketika parameter dioptimalkan. Akurasi XGBoost tanpa optimasi hanya sebesar 75%, tetapi setelah *Grid Search* dan *Random Search*, meningkat hingga 95%. Penemuan ini menunjukkan bahwa konfigurasi parameter memainkan peran penting dalam meningkatkan kemampuan prediksi *model* XGBoost pada data medis.

Dalam penelitian tambahan yang dilakukan oleh (Susanto & Cahyana, 2025), algoritma XGBoost digunakan untuk memprediksi diabetes. Studi ini menilai akurasi dan kesalahan klasifikasi, terutama kesalahan *false negative*, yang memiliki konsekuensi klinis yang signifikan. *Model* XGBoost yang dihasilkan memiliki akurasi sebesar 96,33% dengan nilai ROC-AUC sebesar 0,99 berkat optimasi *hyperparameter* menggunakan Grid Search. Hasil menunjukkan bahwa XGBoost tidak hanya lebih akurat, tetapi juga memiliki potensi besar untuk membantu keputusan medis dalam deteksi diabetes dini. Namun, penelitian ini belum membahas interpretabilitas *model* secara menyeluruh, hanya berfokus pada performa prediksi.

Selanjutnya, penelitian yang dilakukan oleh (Utomo et al., 2025) mengevaluasi kinerja algoritma *XGBoost* dan *CatBoost* saat menentukan tingkat risiko penyakit diabetes. Penelitian ini menggunakan dataset diabetes yang memiliki delapan atribut medis, dan menggunakan metrik akurasi, presisi, *recall*, dan skor F1 untuk mengevaluasi *model*. Hasil penelitian menunjukkan bahwa *CatBoost* unggul dalam menangani data kategorikal, sementara XGBoost memiliki keunggulan dalam efisiensi komputasi dan stabilitas *model* pada data numerik. Perbandingan performa klasifikasi masih menjadi fokus utama penelitian ini, tanpa analisis faktor risiko atau penjelasan tentang keputusan *model*. Namun, penelitian ini memberikan wawasan komparatif yang penting tentang algoritma *boosting*.

Tabel 2. 5 Penelitian Terdahulu

No.	Judul	Metode	Hasil
1.	Diabetes Risk *Predict*ion Using Extreme *Gradient boosting* (XGBoost)	Menggunakan algoritma **XGBoost** untuk klasifikasi risiko diabetes pada dataset *Early-Stage Diabetes Risk Prediction* melalui tahapan transformasi data, pe*model*an, dan evaluasi menggunakan confusion matrix.	*Model* XGBoost menghasilkan akurasi **98,71%**, precision **100%**, dan recall **96,43%**. Faktor risiko yang paling berpengaruh adalah **polidipsia, poliuria, dan penurunan berat badan mendadak**.
2.	Evaluasi Performa	Membandingkan	*Random Forest*

	Random Forest, XGBoost, dan LightGBM dalam Diagnosis Dini Diabetes Mellitus	algoritma ***Random Forest*, XGBoost, dan LightGBM** menggunakan dataset NIDDK dengan penanganan ketidakseimbangan data melalui **SMOTE** serta evaluasi berbasis metrik klasifikasi.	menunjukkan performa terbaik dengan **AUC 86%**, diikuti oleh **XGBoost (83%)** dan **LightGBM (82%)**, sehingga ketiganya dinilai efektif untuk diagnosis dini diabetes.
3.	Hybrid XGBoost-SVM *Model* untuk Sistem Pendukung Keputusan dalam Prediksi Penyakit Diabetes	Mengusulkan *model* **Hybrid XGBoost-SVM**, dengan XGBoost sebagai seleksi fitur dan **SVM** sebagai classifier akhir pada dataset Pima Indians Diabetes.	*Model* Hybrid XGBoost-SVM memberikan akurasi tertinggi sebesar **77%**, lebih baik dibandingkan *model* tunggal XGBoost dan SVM yang masing-masing mencapai **75%**, serta lebih stabil dalam klasifikasi.
4.	Optimasi Algoritma XGBoost Classifier Menggunakan *Hyperparameter* Grid Search dan Random Search pada Klasifikasi Penyakit Diabetes	XGBoost Classifier, imputasi *missing value*, *hyperparameter* tuning (Grid Search & Random Search)	XGBoost tanpa optimasi menghasilkan akurasi 75%. Setelah dilakukan optimasi menggunakan

			Grid Search dan Random Search, performa meningkat signifikan dengan akurasi mencapai 95%.
5	Penerapan Algoritma XGBoost untuk Prediksi Diabetes: Analisis Confusion Matrix dan ROC Curve	XGBoost, preprocessing data, Grid Search, evaluasi confusion matrix dan ROC-AUC	*Model* XGBoost yang dioptimasi mencapai akurasi 96,33%, recall 97,16%, dan nilai ROC-AUC sebesar 0,99, menunjukkan kemampuan deteksi dini diabetes yang sangat baik.
6	Perbandingan Kinerja Algoritma XGBoost dan CatBoost dalam Klasifikasi Risiko Penyakit Diabetes	XGBoost dan CatBoost, preprocessing data, evaluasi akurasi, presisi, recall, dan F1-score	XGBoost menunjukkan performa yang stabil dan efisien pada data numerik, sementara CatBoost unggul dalam penanganan fitur kategorikal. Keduanya efektif untuk klasifikasi risiko diabetes.

7	Comparison of *Random Forest* and XGBoost for Diabetes Classification with SHAP and LIME Interpretation	*Random Forest*, XGBoost, Explainable AI (SHAP & LIME)	XGBoost menghasilkan akurasi lebih tinggi dibandingkan *Random Forest*. SHAP berhasil mengidentifikasi fitur usia, kadar glukosa, dan BMI sebagai faktor risiko utama diabetes.
8	Deteksi Diabetes Mellitus dengan Menggunakan Teknik *Ensemble*	Metode *ensemble* learning (kombinasi beberapa algoritma klasifikasi)	Pendekatan *ensemble* menghasilkan akurasi yang lebih tinggi dibandingkan *model* tunggal, menunjukkan efektivitas kombinasi *model* dalam klasifikasi diabetes.
9	Deteksi Dini Diabetes Menggunakan *Machine Learning*	*Machine Learning* berbasis klasifikasi, preprocessing data medis	*Model Machine Learning* mampu membantu proses deteksi dini diabetes dengan tingkat akurasi yang cukup baik,

			namun masih terbatas pada evaluasi performa numerik.

Berdasarkan kajian penelitian terdahulu, dapat diidentifikasi bahwa sebagian besar penelitian prediksi diabetes menggunakan *Machine Learning* masih berfokus pada peningkatan performa *model* melalui perbandingan algoritma, optimasi *hyperparameter*, atau pendekatan *ensemble*, seperti yang dilakukan oleh Kusuma dan Akbar, Kurniawan et al., serta Cahyana et al., yang menekankan capaian akurasi dan AUC sebagai indikator utama keberhasilan *model*. Meskipun beberapa penelitian mulai mengadopsi *Explainable AI* seperti SHAP dan LIME (Mubaraqah et al., 2024), penerapannya umumnya masih terbatas sebagai alat visualisasi tambahan dan belum diintegrasikan secara menyeluruh ke dalam sistem pakar yang dapat digunakan langsung oleh pengguna. Selain itu, sebagian besar penelitian belum mengimplementasikan *model* prediksi dan interpretabilitas tersebut dalam bentuk sistem pakar berbasis web yang interaktif dan berorientasi pada pengguna, sehingga hasil penelitian masih bersifat eksperimental dan kurang aplikatif. Oleh karena itu, kelebihan penelitian ini terletak pada pengembangan sistem pakar berbasis web yang mengintegrasikan *model* XGBoost dengan Explainable AI (SHAP) secara komprehensif untuk memprediksi risiko diabetes sekaligus memberikan analisis faktor risiko secara individual dan transparan, sehingga tidak hanya menekankan akurasi prediksi, tetapi juga meningkatkan interpretabilitas, kepercayaan pengguna, dan nilai praktis dalam konteks sistem informasi kesehatan.

BAB III

ANALISA DAN PERANCANGAN SISTEM

3.1. Analisis Permasalahan

Analisis masalah merupakan tahap awal dalam metode penelitian yang bertujuan untuk mengidentifikasi dan memahami permasalahan yang menjadi dasar dilakukannya penelitian ini. Permasalahan utama yang diangkat dalam penelitian ini adalah rendahnya kesadaran masyarakat terhadap risiko penyakit diabetes serta keterbatasan akses terhadap pemeriksaan kesehatan secara rutin. Banyak individu tidak menyadari gejala awal diabetes karena penyakit ini sering berkembang tanpa tanda yang jelas pada tahap awal, sehingga diagnosis umumnya dilakukan ketika kondisi telah memasuki tahap lanjut.

Selain rendahnya kesadaran, permasalahan lain yang dihadapi adalah keterbatasan waktu, biaya, dan akses fasilitas kesehatan yang menyebabkan masyarakat enggan atau menunda melakukan pemeriksaan medis. Kondisi tersebut mengakibatkan keterlambatan dalam mendeteksi risiko diabetes, sehingga peluang untuk melakukan pencegahan dini menjadi berkurang. Oleh karena itu, diperlukan suatu solusi yang mampu membantu masyarakat dalam mengenali potensi risiko diabetes secara mandiri dan mudah diakses.

Permasalahan tersebut diperparah dengan belum tersedianya sistem deteksi dini yang bersifat praktis, mudah digunakan, dan dapat memberikan penjelasan mengenai faktor-faktor yang memengaruhi hasil deteksi. Sebagian besar masyarakat membutuhkan media yang tidak hanya memberikan hasil, tetapi juga

mampu menjelaskan alasan di balik hasil tersebut agar dapat meningkatkan pemahaman dan kepercayaan pengguna terhadap sistem.

Berdasarkan permasalahan yang telah diidentifikasi, penelitian ini mengusulkan pengembangan sistem pakar deteksi dini risiko diabetes berbasis *web* dengan memanfaatkan teknik *Data Mining*. Sistem ini dirancang untuk menganalisis faktor-faktor risiko yang dapat dikenali secara mandiri oleh pengguna, tanpa memerlukan pemeriksaan laboratorium atau tindakan medis langsung. Dengan pendekatan ini, sistem diharapkan mampu menjadi sarana pendukung bagi masyarakat dalam mengenali risiko diabetes sejak dini. Untuk mengatasi permasalahan tersebut, digunakan algoritma XGBoost sebagai metode klasifikasi dalam menentukan tingkat risiko diabetes, serta metode SHAP sebagai alat penjelasan untuk menginterpretasikan kontribusi setiap faktor risiko terhadap hasil prediksi. Penerapan metode ini diharapkan dapat menghasilkan sistem yang tidak hanya memiliki tingkat akurasi yang baik, tetapi juga bersifat transparan dan mudah dipahami oleh pengguna.

Dengan adanya sistem yang diusulkan, diharapkan permasalahan keterlambatan deteksi risiko diabetes dan rendahnya kesadaran masyarakat terhadap penyakit ini dapat diminimalkan, sehingga masyarakat dapat mengambil langkah pencegahan atau berkonsultasi lebih lanjut dengan tenaga medis secara lebih tepat waktu.

3.2. Analisis Kebutuhan Sistem

3.2.1. Kebutuhan Fungsional Sistem

Kebutuhan fungsional merupakan kebutuhan yang berkaitan langsung dengan fungsi utama yang harus dimiliki oleh sistem pakar deteksi dini risiko diabetes. Sistem harus mampu menerima *input* data dari pengguna berupa informasi usia, jenis kelamin, serta gejala dan kondisi fisik yang berkaitan dengan risiko diabetes. Data yang dimasukkan oleh pengguna kemudian diproses oleh sistem menggunakan *model* klasifikasi berbasis algoritma XGBoost untuk menghasilkan hasil deteksi berupa tingkat risiko diabetes.

Selain melakukan proses klasifikasi, sistem juga harus mampu menampilkan hasil deteksi secara jelas dan informatif kepada pengguna. Hasil tersebut mencakup kategori risiko diabetes serta penjelasan mengenai faktor-faktor yang memengaruhi hasil prediksi. Penjelasan ini dihasilkan menggunakan metode SHAP sehingga sistem tidak hanya memberikan hasil prediksi, tetapi juga memberikan pemahaman kepada pengguna mengenai alasan di balik hasil tersebut. Sistem juga harus mampu memuat *model* yang telah dilatih sebelumnya agar proses deteksi dapat dilakukan tanpa perlu pelatihan ulang setiap kali sistem dijalankan.

3.2.2. Kebutuhan Non-Fungsional

Kebutuhan non-fungsional berkaitan dengan kualitas dan karakteristik sistem yang mendukung kinerja dan kenyamanan penggunaan sistem. Sistem harus memiliki waktu respons yang cepat dalam memproses data dan menampilkan hasil deteksi, sehingga pengguna tidak perlu menunggu terlalu

lama. Sistem juga harus mudah diakses melalui perangkat yang terhubung ke internet dan dapat dijalankan melalui *browser* tanpa memerlukan instalasi tambahan.

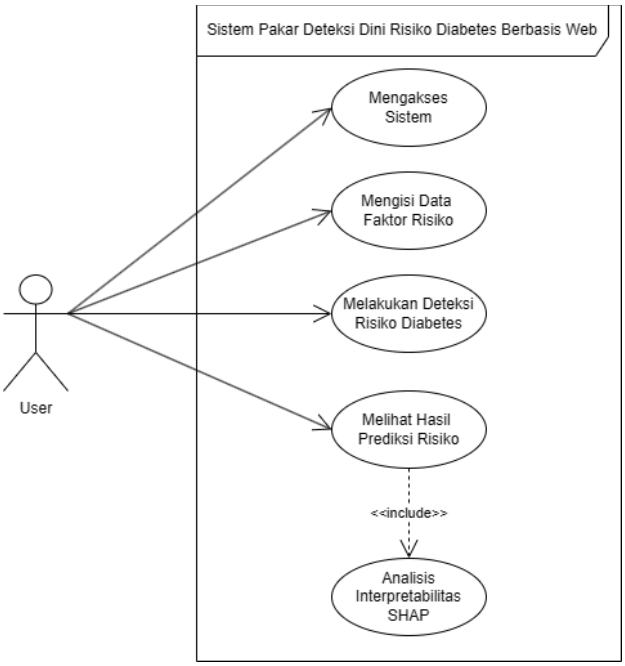
Selain itu, sistem harus memiliki antarmuka yang sederhana dan mudah dipahami oleh masyarakat umum, sehingga pengguna dari berbagai latar belakang dapat menggunakan sistem dengan mudah. Aspek keamanan juga menjadi bagian dari kebutuhan non-fungsional, di mana sistem harus menjaga kerahasiaan data yang diinputkan oleh pengguna dan tidak menyimpan data pribadi secara permanen. Sistem juga diharapkan bersifat fleksibel dan mudah dikembangkan, sehingga dapat diperbarui atau ditingkatkan di masa mendatang sesuai dengan kebutuhan.

3.3. Perancangan Sistem

Perancangan sistem pakar deteksi dini risiko diabetes berbasis *web* ini menggunakan *Unified Modeling Language* (UML). Perancangan sistem dilakukan untuk menggambarkan kebutuhan fungsional, alur proses, serta interaksi antar komponen sistem. Diagram UML yang digunakan dalam penelitian ini meliputi :

3.3.1. Usecase Diagram

Use Case Diagram digunakan untuk menggambarkan fungsi-fungsi yang disediakan oleh sistem serta interaksi antara pengguna dan sistem pakar deteksi dini risiko diabetes berbasis web. Aktor utama pada sistem ini adalah pengguna masyarakat umum yang berinteraksi langsung dengan sistem.



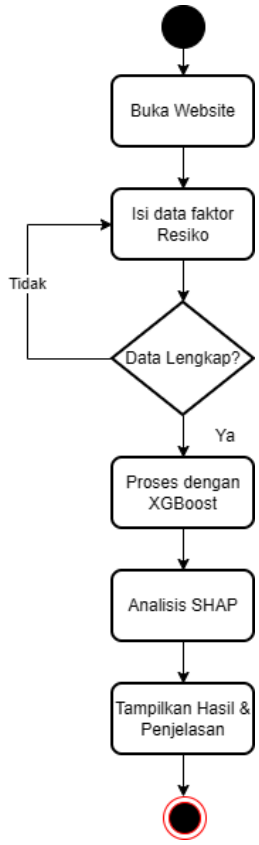
Gambar 3. 1 Usecase Diagram

Proses dimulai ketika pengguna mengakses sistem melalui antarmuka *web*. Setelah berhasil mengakses sistem, pengguna dapat mengisi data faktor risiko diabetes yang meliputi usia, jenis kelamin, serta gejala dan kondisi fisik yang relevan. Data yang telah diinput selanjutnya digunakan oleh sistem untuk menjalankan proses deteksi risiko diabetes. Pada tahap ini, sistem memproses data menggunakan *model Machine Learning* XGBoost untuk menentukan tingkat risiko diabetes. Selain menghasilkan hasil prediksi, sistem juga melakukan analisis interpretabilitas menggunakan metode SHAP untuk menjelaskan faktor-faktor yang memengaruhi hasil prediksi. Selanjutnya, pengguna dapat melihat hasil prediksi risiko diabetes beserta penjelasan faktor risiko melalui antarmuka sistem.

3.3.2. Activity Diagram

Activity Diagram menggambarkan alur aktivitas sistem secara keseluruhan mulai dari awal hingga akhir proses deteksi risiko diabetes. Proses

diawali ketika pengguna membuka website sistem pakar. Setelah sistem ditampilkan, pengguna diminta untuk mengisi data faktor risiko diabetes melalui formulir yang tersedia.



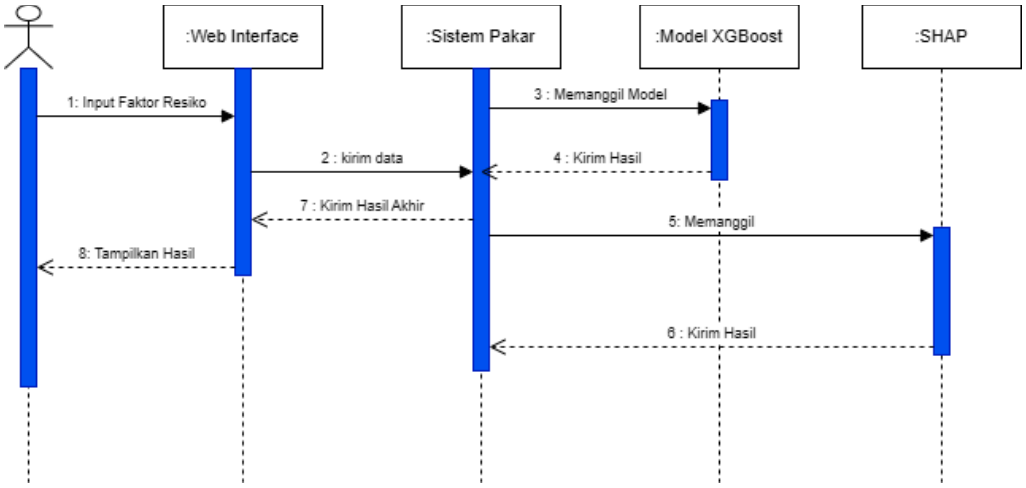
Gambar 3. 2 Activity Diagram

Setelah data diinput, sistem melakukan validasi kelengkapan data untuk memastikan seluruh data yang dibutuhkan telah diisi. Apabila data belum lengkap, sistem akan mengarahkan pengguna kembali ke proses pengisian data. Jika data telah lengkap, sistem melanjutkan ke tahap pemrosesan data menggunakan *model* XGBoost untuk melakukan klasifikasi risiko diabetes. Selanjutnya, hasil prediksi dianalisis menggunakan metode SHAP untuk mengetahui kontribusi masing-masing faktor risiko terhadap hasil prediksi.

Setelah seluruh proses selesai, sistem akan menampilkan hasil deteksi risiko diabetes beserta penjelasan faktor risiko kepada pengguna, dan proses berakhir.

3.3.3. Sequence Diagram

Sequence Diagram menggambarkan urutan interaksi dan komunikasi antar komponen sistem dalam proses deteksi dini risiko diabetes. Proses dimulai ketika pengguna memasukkan data faktor risiko melalui antarmuka *web*. Data tersebut kemudian diteruskan oleh antarmuka *web* ke sistem pakar sebagai pusat pengolahan data.

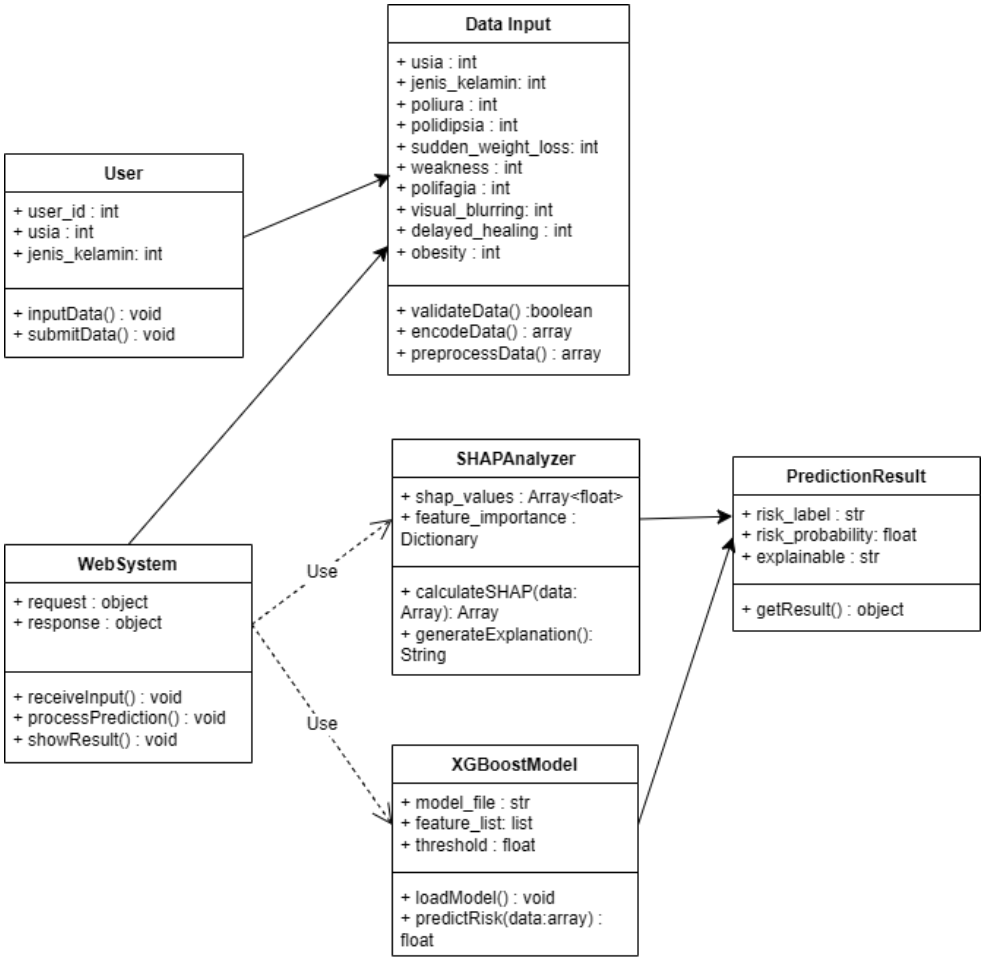


Gambar 3. 3 Sequence Diagram

Sistem pakar selanjutnya memanggil *model* XGBoost untuk melakukan prediksi risiko diabetes berdasarkan data input pengguna. Setelah hasil prediksi diperoleh, sistem pakar memanggil modul SHAP untuk melakukan analisis interpretabilitas guna mengetahui faktor-faktor yang paling berpengaruh terhadap hasil prediksi. Hasil prediksi risiko diabetes beserta penjelasan faktor risiko kemudian dikirim kembali oleh sistem pakar

ke antarmuka *web*. Selanjutnya, antarmuka web menampilkan informasi tersebut kepada pengguna sebagai hasil akhir dari proses deteksi.

3.3.4. Class Diagram



Gambar 3. 4 Class Diagram

Class diagram pada penelitian ini menggambarkan struktur sistem pakar deteksi dini risiko diabetes berbasis *web*. Sistem terdiri dari beberapa kelas utama, yaitu *User*, *DataInput*, *WebSystem*, *XGBoostModel*, *SHAPAnalyzer*, dan *Prediction Result*. *User* berinteraksi dengan sistem dengan menginput data faktor risiko diabetes yang terdiri dari usia, jenis kelamin,

poliuria, polidipsia, penurunan berat badan secara tiba-tiba, kelemahan tubuh, polifagia, gangguan penglihatan, penyembuhan luka yang lambat, dan obesitas. Data tersebut disimpan dalam kelas *DataInput* dan diproses oleh sistem menggunakan *model* XGBoost untuk menghasilkan hasil prediksi risiko diabetes. Selanjutnya, sistem menggunakan metode SHAP untuk memberikan penjelasan mengenai kontribusi masing-masing faktor risiko terhadap hasil prediksi. Hasil akhir berupa tingkat risiko dan penjelasan faktor risiko disajikan kepada pengguna melalui antarmuka web.

3.4. Tahapan Pengolahan Data

Tahapan pengolahan data merupakan proses penting dalam penelitian ini karena kualitas data yang digunakan sangat mempengaruhi kinerja *model* prediksi yang dihasilkan. Tahapan pengolahan data dalam penelitian ini meliputi beberapa langkah sebagai berikut :

3.4.1. Data Cleaning

Tahap data cleaning bertujuan untuk meningkatkan kualitas data dengan menghilangkan atau memperbaiki data yang berpotensi menimbulkan kesalahan dalam proses pe*model*an. Pada tahap ini dilakukan pemeriksaan terhadap data duplikat, yaitu data yang memiliki nilai identik pada seluruh variabel, kemudian data tersebut dihapus agar tidak menyebabkan bias dalam pelatihan *model*. Selain itu, dilakukan pengecekan konsistensi nilai pada setiap atribut untuk memastikan bahwa data berada dalam rentang dan format yang sesuai dengan definisi variabel yang digunakan. Proses ini bertujuan untuk memastikan bahwa data yang digunakan mencerminkan kondisi yang

sebenarnya dan tidak mengandung anomali yang dapat mempengaruhi hasil prediksi.

3.4.2. Label *Encoding*

Dalam penelitian ini, label *encoding* digunakan untuk mengonversi variabel kategorikal yang merepresentasikan kondisi atau gejala pengguna ke dalam bentuk nilai numerik. Setiap variabel dikodekan secara biner, di mana nilai 1 merepresentasikan kondisi "Ya" atau "Positif", sedangkan nilai 0 merepresentasikan kondisi "Tidak" atau "Negatif". Proses ini diterapkan pada variabel seperti jenis kelamin, poliuria, polidipsia, penurunan berat badan, kelemahan tubuh, polifagia, gangguan penglihatan, penyembuhan luka yang lambat, dan obesitas.

Penerapan label *encoding* bertujuan untuk menyederhanakan representasi data tanpa menghilangkan makna dari setiap variabel, sehingga *model* dapat mengenali pola hubungan antara faktor risiko dan kelas target secara lebih efektif. Dengan melakukan *label encoding*, data menjadi lebih terstruktur dan siap digunakan pada tahap pelatihan dan pengujian *model* klasifikasi dalam sistem pakar deteksi dini risiko diabetes.

3.5. Populasi dan Sampel Penelitian

Populasi dalam penelitian ini adalah keseluruhan data individu yang memiliki profil karakteristik demografis, gaya hidup, dan parameter klinis yang relevan dengan risiko diabetes mellitus tipe 1 dan 2. Secara konseptual, populasi target mencakup individu dewasa yang berpotensi mengalami diabetes berdasarkan faktor-faktor risiko yang dapat diidentifikasi dan diukur.

Sampel penelitian yang digunakan bersumber dari dataset risiko diabetes yang tersedia secara publik melalui platform Kaggle dengan total terdapat 520 dataset berformat .csv dan memiliki 17 variabel values yang dikategorikan sebagai resiko diagnosa diabetes. Adapun variabel yang dipilih untuk pelatihan *model* machine leaning dalam memprediksi resiko diabetes dalam penelitian ini sebanyak 10 variabel yaitu Usia (*Age*), Jenis kelamin (*Gender*), Poliuria (sering buang air kecil), Polidipsia (sering merasa haus), Penurunan berat badan secara tiba-tiba, Kelemahan tubuh (*Weakness*), Polifagia (sering merasa lapar), Gangguan penglihatan (*Visual Blurring*), Penyembuhan luka yang lambat (*Delayed Healing*), Obesitas (*Obesity*). Kriteria pemilihan dataset meliputi ketersediaan variabel yang dapat dikumpulkan tanpa pemeriksaan laboratorium sehingga sesuai dengan tujuan pengembangan sistem skrining yang aksesibel, keberadaan label target yang valid mengenai status diabetes, ukuran sampel yang mencukupi untuk membangun *model Machine Learning* yang *robust*, serta kualitas data dengan tingkat *missing value*s dan outliers yang masih dapat ditangani melalui teknik pra-pemrosesan.

3.6. Validasi dan Analisis Algoritma

3.6.1. Extreme *Gradient boosting* (XGBoost)

XGBoost (Extreme *Gradient boosting*) merupakan implementasi optimasi dari algoritma *gradient boosting* yang membangun *model* prediksi secara *ensemble*. Berbeda dengan metode *ensemble* lain seperti *Random Forest* yang membangun pohon secara paralel dan independen, XGBoost membangun pohon secara sekuensial dimana setiap pohon baru difokuskan untuk memperbaiki

kesalahan (residual) dari pohon-pohon sebelumnya. Formula matematis untuk prediksi XGBoost pada iterasi ke-t adalah:

$$\hat{y}^{(t)} = \hat{y}^{(t-1)} + n \times f_t(x) \quad (4)$$

Dimana :

- $\hat{y}^{(t)}$ adalah prediksi setelah pohon ke-t ditambahkan
- $\hat{y}^{(t-1)}$ adalah prediksi dari *model* sebelumnya (akumulasi dari t-1 pohon)
- η adalah learning rate (laju pembelajaran), dalam sistem ini η = 0.1
- $f_t(x)$ adalah output dari pohon ke-t untuk input x

Learning rate berfungsi sebagai regularisasi yang mengontrol seberapa agresif *model* belajar dari setiap pohon baru. Nilai yang lebih kecil membuat pembelajaran lebih konservatif namun lebih *robust* terhadap *overfitting*.

Sebelum first tree dibangun, XGBoost memerlukan nilai inisialisasi (*base Prediction*) sebagai starting point. Untuk klasifikasi biner dengan fungsi *loss logistic*, *base Prediction* dihitung sebagai *log-odds* dari distribusi kelas pada data training. Berikut *formula Base Prediction* untuk inisialisasi *model* :

$$Base\ Pred\ \left(\hat{y}^{(0)}\right) = log(P\ (y = 1)/\ P(y = 0)) \quad (5)$$

$$= \log\left(\frac{n_{positive}}{n_{negative}}\right.$$

Perhitungan pada dataset risiko diabetes yang digunakan memiliki karekteristik antara lain yaitu memiliki total sampel sebanyak 520 dengan sebanyak 320 data merupakan sampel *positive* (diabetes) dan 200 data merupakan sampel *negative* (non-diabetes).

Maka jumlah rasio kelas yaitu $Ratio = \frac{n_{positive}}{n_{negative}} = \frac{320}{200} = 1.6$ dan menghitung algoritma logaritma natural sebagai *base Prediction* yaitu $Base\ Pred = \ln(1.6) = 0.4700$.

Adapun logaritma natural dapat dihitung menggunakan ekspansi deret taylor sebagai berikut,

$$\ln(1+n) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \frac{x^5}{5} - \frac{x^6}{6} + \cdots \quad (6)$$

Perhitungan ekspansi deret *taylor* berjumlah tak hingga dikarenakan setiap suku berikut nya maka nilainya akan menjadi semakin kecil yang menyebabkan kontribusinya menjadi tidak signifikan, maka perhitungan sebagai berikut :

$$\ln(1.6) = \ln(1+0.6) = 0.6 - \frac{(0.6)^2}{2} + \frac{(0.6)^3}{3} - \frac{(0.6)^4}{4} + \frac{(0.6)^5}{5} - \frac{(0.6)^6}{6} + \frac{(0.6)^7}{7} - \cdots$$

$$= 0.6 - 0.18 + 0.072 - 0.0324 + 0.0155 - 0.0077 + 0.0039 - 0.0020 + \cdots$$

$$= 0.6 - 0.18 = 0.42$$

$$= 0.42 + 0.072 = 0.492$$

$$= 0.492 - 0.0324 = 0.4596$$

$$= 0.4596 + 0.0155 = 0.4751$$

$$= 0.4751 - 0.0077 = 0.4674$$

$$= 0.4674 + 0.0039 = 0.4713$$

$$= 0.4713 - 0.0020 = 0.4693$$

Maka hasil dari perhitungan ekspansi deret ini dihasilkan $0.4693 \approx 0.470$.

Dari hasil yang ada maka interpretasi *base Prediction* yaitu nilai 0.470 dalam *log-odds* dapat di konversi ke probabilitas menggunakan fungsi sigmoid sebagai berikut :

$$P = 1 / (1 + e\text{^}(-0.470)) = 1 / (1 + 0.625) = 0.615\ atau\ 61.5\%$$

Ini berarti bahwa tanpa melihat fitur apapun dari pasien, *prior probability* untuk diabetes adalah 61.5%, yang sesuai dengan proporsi kelas positif dalam dataset training (320/520 = 61.5%). *Base Prediction* ini menjadi fondasi dimana semua pohon keputusan selanjutnya akan membangun penyesuaian berdasarkan fitur-fitur individual pasien.

Untuk mengilustrasikan proses iteratif XGBoost, akan digunakan contoh sederhana dengan subset data training yang terdiri dari 5 sampel. Meskipun *model* aktual menggunakan 520 sampel dan 200 tree, prinsip dasarnya tetap sama. Berikut adalah data training subset:

Tabel 3. 1 Subset Training Xgboost

ID	Age	Gender	*Polyuria*	*Polydipsia*	*Polyphagia*	Actual (y)
1	25	1 (L)	0	0	0	0
2	45	0 (P)	1	1	1	1
3	30	1 (L)	0	1	0	0
4	50	0 (P)	1	1	1	1
5	35	1 (L)	1	0	1	1

Terlihat pada tabel diatas iterasi awal merupakan inisialisasi agar semua sampel mendapatkan prediksi awal yang sama. Perhitungan residual awal untuk klasifikasi *biner* sebagai selisih antara nilai actual dan probabilitas prediksi :

$$r = y - p \tag{4}$$

Formula diatas mendeskripsi kan *residual* (r) merupakan hasil nilai actual (y) dikurangi dengan probabilitas (p), maka hasilnya dapat dilihat pada tabel berikut :

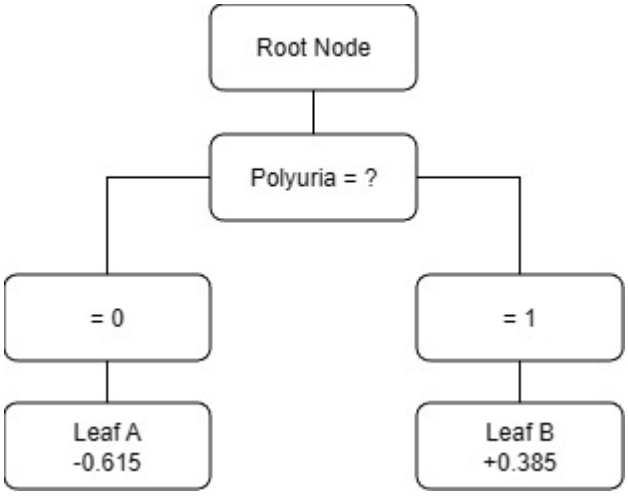
Tabel 3. 2 Probabilitas Awal Prediksi

ID	Aktual (y)	*Predict*ed (p)	Residual (r)
1	0	0.615	-0.615

2	1	0.615	+0.385
3	0	0.615	-0.651
4	1	0.615	+0.385
5	1	0.615	+0.385

Hasil tabel diatas residual negatif menunjukkan *model* over-prediksi (memprediksi probabilitas terlalu tinggi untuk kasus negatif), sedangkan residual positif menunjukkan under-prediksi (memprediksi probabilitas terlalu rendah untuk kasus positif).

Pohon pertama dibangun untuk memprediksi residual tersebut. Algoritma mencari split terbaik yang memisahkan sampel berdasarkan fitur. Pohon pertama dibangun untuk memprediksi *residual* tersebut. Gambar …. Merupakan struktur mencari split terbaik yang memisahkan sampel berdasarkan fitur. Maka *Polyuria* di split menjadi 2 jenis antara lain *Polyuria* (0) dengan ID (1 dan 3) average residual -0.615 dan *Polyuria* (1) dengan ID (2,4,5) average residual +0.385.



Gambar 3. 5 Node Pohon XGboost

Setelah menentukan nilai output untuk setiap leaf, selanjutnya memperbarui prediksi *model* menggunakan rumus pembaruan prediksi dengan learning rate $\eta = 0.1$, yang dihitung sebagai berikut:

$$\hat{y}^{(1)} = \hat{y}^{(0)} + n \times f_1(x) \quad (7)$$

Di mana $\hat{y}^{(0)}$adalah prediksi sebelumnya, dan $f_1(x)$adalah output dari leaf pada pohon keputusan. ID 1 (*Polyuria* = 0), prediksi sebelumnya $\hat{y}^{(0)} = 0.470$. Dengan output dari Leaf A ($f_1(x)$ = -0.615), prediksi baru dihitung sebagai:

$$\hat{y}^{(1)} = 0.470 + 0.1 \times (-0.615) = 0.470 - 0.0615 = 0.408$$

ID 2 (*Polyuria* = 1), prediksi sebelumnya juga $\hat{y}^{(0)} = 0.470$, dan output dari Leaf B ($f_1(x)$ = +0.385) menghasilkan prediksi baru:

$$\hat{y}^{(1)} = 0.470 + 0.1 \times (+0.385) = 0.470 - 0.0385 = 0.509$$

Langkah yang sama diterapkan untuk ID 3, ID 4, dan ID 5, di mana prediksi diperbarui menggunakan output yang sesuai dari Leaf A dan Leaf B. Setelah memperbarui prediksi, langkah berikutnya adalah menghitung residual yang menggambarkan selisih antara nilai aktual dan probabilitas yang diprediksi. Untuk menghitung probabilitas, digunakan fungsi sigmoid:

$$P^{(1)} = \frac{1}{1 + e^{\hat{y}^{(1)}}} \quad (8)$$

Dengan ID 1 yang memiliki prediksi $\hat{y}^{(1)} = 0.408$, probabilitas $p^{(1)}$dihitung sebagai:

$P^{(1)} = \frac{1}{1+e^{-0.408}} = -0.601$ maka residual untuk ID 1 adalah $r^1 = 0 - 0.601 = -0.601$, selanjutnya pada ID 2 probabilitas dihitung sebagai : $P^{(1)} = \frac{1}{1+e^{-0.509}} = 0.625$ maka residual untuk ID 2 adalah $r^1 = 1 - 0.625 = +0.375$

Perhitungan residual untuk seluruh data menunjukkan bahwa nilai residual semakin kecil, dibandingkan dengan residual sebelumnya yang bernilai sekitar ±0.615 dan ±0.385. Hal ini menunjukkan bahwa *model* prediksi semakin mendekati nilai aktual, dengan pembaruan yang dilakukan berdasarkan hasil dari pohon keputusan dan learning rate yang digunakan.

Setelah melakukan pembaruan prediksi pada Iterasi 1, selanjutnya proses dengan Iterasi 2. Pada iterasi ini, pohon kedua dibuat untuk lebih menangkap pola residual yang belum sepenuhnya tertangkap oleh pohon pertama. Pada pohon kedua, pembagian data dilakukan berdasarkan fitur *Polydipsia*. Proses pembagian data menjadi dua kelompok sebagai berikut :

Kelompok *Polydipsia* = 0: Berisi ID 1 dan ID 5, dengan rata-rata residual dihitung sebagai:

$$Avg\ Res = -0.601 + \frac{-0.601 + 0.375}{2} = -0.113$$

Kelompok *Polydipsia* = 1: Berisi ID 2, 3, 4, dengan rata-rata residual dihitung sebagai:

$$Avg\ Res = -0.601 + \frac{0.375 - 0.061 + 0.375}{3} = +0.050$$

Dengan rata-rata residual yang telah dihitung, untuk memperbarui prediksi untuk setiap ID menggunakan rumus pembaruan prediksi:

$$\hat{y}^{(2)} = \hat{y}^{(1)} + n \times f_2(x) \quad (9)$$

Di mana $f_2(x)$ adalah residual rata-rata untuk setiap kelompok, dan $\eta = 0.1$adalah learning rate. Untuk ID 1, prediksi sebelumnya $\hat{y}^{(1)} = 0.408$ diperbarui dengan *residual* rata-rata dari *Polydipsia* = 0:

$$\hat{y}^{(2)} = 0.408 + 0.1 \times (-0.113) = 0.397$$

Untuk ID 2, prediksi sebelumnya $\hat{y}^{(1)} = 0.509$ diperbarui dengan residual rata-rata dari *Polydipsia* = 1:

$$\hat{y}^{(2)} = 0.509 + 0.1 \times (+0.050) = 0.514$$

Proses ini kemudian berlanjut selama 200 iterasi, dengan setiap iterasi menciptakan pohon keputusan baru yang difokuskan untuk menangkap pola *residual* yang belum terungkap oleh pohon-pohon sebelumnya. Dalam hal ini, setiap pohon baru akan mencoba memperbaiki kesalahan atau kekurangan yang ada pada pohon-pohon sebelumnya. Pada pohon-pohon awal, *model* umumnya lebih banyak menangkap pola utama atau fitur-fitur yang paling penting, seperti dalam kasus ini bisa jadi trias diabetes (misalnya, fitur yang terkait dengan gejala diabetes). Pohon-pohon ini fokus pada memperbaiki prediksi dengan menangkap pola-pola yang lebih sederhana dan terlihat jelas.

Namun, pada pohon-pohon belakangan (iterasi yang lebih tinggi), *model* akan mulai melakukan *fine-tuning* terhadap prediksi. Pohon-pohon ini lebih fokus pada menangkap pola-pola yang lebih subtle atau halus, termasuk interaksi antar fitur yang mungkin tidak begitu jelas pada awalnya. Misalnya, interaksi antara *Polydipsia* dan *Polyuria* bisa jadi menjadi lebih jelas setelah pohon pertama menangkap pola utama. Melalui proses ini, *model* secara bertahap mengurangi kesalahan prediksi dan meningkatkan akurasi dengan menangkap dan mengoreksi pola-pola yang sebelumnya tidak ter-tangkap.

Setelah semua pohon keputusan dibangun, prediksi akhir untuk suatu input (misalnya, pasien dengan gejala tertentu) dihitung dengan cara mengakumulasi prediksi dasar (*base Prediction*) yang diberikan pada awalnya, ditambah dengan

kontribusi dari seluruh pohon yang telah di scale menggunakan learning rate. Adapun persamaan menghitung prediksi akhir adalah sebagai berikut :

$$\hat{y}_{final} = \hat{y}^{(0)} + n \times \sum_{t=1}^{200} f_1(x) \quad (10)$$

Di mana:

- $\hat{y}^{(0)}$adalah prediksi dasar (*base Prediction*),
- ηadalah learning rate,
- $f_t(x)$adalah kontribusi dari pohon ke-t,
- $t$berjalan dari 1 hingga 200 pohon.

Secara sederhana, prediksi akhir bisa dijelaskan sebagai:

$$\hat{y}_{\text{final}} = Base\ Prediction + Total\ Kontribusi\ Pohon$$

Berikut contoh perhitungan untuk kasus pasien dengan risiko tinggi dengan proses yang diterapkan pada pasien yang mengalami gejala terkait diabetes sebagai berikut :

Input Pasien :

- *Age*: 45
- *Gender*: Perempuan (0)
- *Polyuria*: Ya (1)
- *Polydipsia*: Ya (1)
- *Polyphagia*: Ya (1)
- *Sudden Weight Loss*: Tidak (0)
- *Weakness*: Tidak (0)

- *Visual Blurring*: Tidak (0)
- *Delayed Healing*: Ya (1)
- *Obesity*: Tidak (0)

Langkah-langkah perhitungannya adalah sebagai berikut:

Pada awalnya, *base Prediction* $\hat{y}^{(0)}$ adalah nilai yang ditentukan sebelumnya, yaitu 0.470. Ini adalah prediksi dasar sebelum kontribusi dari pohon-pohon ditambahkan. Setelah itu, kontribusi dari seluruh 200 pohon dihitung. Karena melakukan perhitungan manual untuk kontribusi dari 200 pohon secara langsung sangat kompleks dan tidak praktis, kontribusi agregat ini bisa diperoleh melalui output sistem yang sudah terlatih. Dari hasil sistem, diketahui bahwa total kontribusi dari seluruh pohon adalah 5.760. Kontribusi ini mencerminkan akumulasi dari pohon-pohon berikut:

- Pohon 1-50: Fokus pada fitur utama seperti *Polydipsia*, *Polyuria*, dan *Polyphagia*.
- Pohon 51-100: Menangkap interaksi antar fitur, seperti hubungan antara *Polyuria* dan *Polyphagia*.
- Pohon 101-150: Melakukan refinement berdasarkan fitur-fitur sekunder yang mungkin mempengaruhi prediksi, seperti *Obesity* dan *Visual Blurring*.
- Pohon 151-200: Melakukan fine-tuning pada prediksi akhir, memperbaiki *model* dengan menangkap pola-pola yang lebih halus atau kompleks.

Prediksi akhir dihitung dengan menjumlahkan *base Prediction* dengan total kontribusi pohon :

$$\hat{y}_{\text{final}} = 0.470 + 5.760 = 6.230$$

Prediksi akhir ini masih dalam bentuk *log-odds*, dan perlu dikonversi menjadi probabilitas menggunakan fungsi sigmoid:

$$P(diabetes) = \frac{1}{1 + e^{-\hat{y}_{\text{final}}}}$$

Dengan $\hat{y}_{\text{final}} = 6.230$ maka :

$$e^{-6.230} = \frac{1}{e^{6.230}} = \frac{1}{506.23} = 0.001972$$

Sehingga probabilitas diabetesnya adalah :

$$P(diabetes) = \frac{1}{1 + 0.001976} = \frac{1}{1.001976} = 0.9980 = 99.80\%$$

Berdasarkan probabilitas yang telah dihitung, kita dapat menentukan kategori risiko pasien dengan menggunakan *threshold* sebagai berikut:

- Rendah: Probabilitas < 30%
- Sedang: Probabilitas antara 30% dan 70%
- Tinggi: Probabilitas > 70%

Dengan probabilitas 99.80%, pasien dikategorikan dalam kategori risiko tinggi.

3.6.2. Shapley Additive Explanations (SHAP)

SHAP (SHapley Additive exPlanations) adalah metode untuk menjelaskan prediksi *model Machine Learning* dengan menghitung kontribusi masing-masing fitur. SHAP didasarkan pada *Shapley values* dari teori permainan kooperatif yang dikembangkan oleh Lloyd Shapley (Nobel Ekonomi 2012). Adapun formula umum Shapley Value sebagai berikut :

$$\varphi i = \sum [|S|! \times \frac{(M - |S| - 1)!}{M!} \times [f(S \cup \{i\}) - f(S)] \quad (11)$$

Di mana:

- $\phi_i$adalah SHAP value untuk fitur i.
- M adalah total jumlah fitur dalam *model*.
- S adalah subset fitur yang tidak mengandung fitur i.
- F adalah himpunan semua fitur.
- $| S |$ adalah ukuran atau cardinality dari set S (jumlah fitur dalam subset S).
- f(S) adalah prediksi *model* yang hanya menggunakan fitur dalam subset S.
- $f(S \cup \{i\})$ adalah prediksi *model* setelah fitur i ditambahkan ke dalam subset S.

Formula ini dapat dipahami sebagai cara untuk menghitung kontribusi marginal dari fitur i terhadap prediksi akhir. Artinya, menghitung bagaimana perubahan prediksi *model* ketika fitur i ditambahkan ke setiap subset fitur yang mungkin ada, dan kemudian rata-rata perubahan tersebut dengan mempertimbangkan probabilitas urutan penambahan fitur. Bobot yang digunakan dalam formula, yaitu:

$$\frac{|S|! \times (M - |S| - 1)!}{M!} \tag{12}$$

Probabilitas dari urutan penambahan fitur i ke dalam subset fitur S. Bobot ini memastikan bahwa setiap urutan penambahan fitur diperlakukan secara fair (adil), tanpa memberikan kelebihan pada urutan tertentu dalam pemilihan fitur.
Pada *model* pohon keputusan seperti XGBoost, perhitungan *Shapley value* secara brute force (langsung) memiliki kompleksitas yang sangat tinggi. Kompleksitas ini bisa mencapai $O(2^M \times T)$, di mana:

- M adalah jumlah fitur dalam *model*,
- T adalah jumlah evaluasi *model* yang diperlukan.

Misalnya, jika *model* memiliki 10 fitur, maka jumlah kombinasi subset fitur yang harus dihitung adalah $2^{10} = 1024$. Untuk setiap subset, *model* perlu dievaluasi untuk menghitung kontribusi setiap fitur.

Namun, untuk mengoptimalkan perhitungan SHAP dalam pohon keputusan, kita menggunakan teknik yang disebut *TreeExplainer*. *TreeExplainer* memanfaatkan struktur internal pohon keputusan untuk menghitung kontribusi setiap fitur secara efisien. Alih-alih mengevaluasi *model* untuk setiap subset fitur, *TreeExplainer* melakukan traversal pohon dan menghitung kontribusi berdasarkan path yang dilalui oleh data saat membuat prediksi. Kompleksitas perhitungan *TreeExplainer* adalah sebagai berikut:

$$O(T \times L \times D^2) \tag{13}$$

Di mana:

- T adalah jumlah pohon dalam *model* (misalnya 200 pohon dalam XGBoost),
- L adalah rata-rata jumlah *leaf* per pohon (misalnya 32 untuk pohon dengan max_depth = 5),
- D adalah kedalaman maksimum pohon (misalnya 5)

Dalam hal ini *model* yang digunakan perhitungan *TreeExplainer* akan membutuhkan :

$$T \times L \times D^2 = 200 \times 32 \times 25 = 160{,}000\ Operasi$$

Sementara itu jika menggunakan metode brute force untuk memeriksa semua subset fitur, jumlah evaluasi *model* yang di perlukan akan mencapai:

$$1024 \times 200 = 204.800\ evaluasi\ model$$

Dari perbandingan ini, dapat dilihat bahwa *TreeExplainer* jauh lebih efisien dibandingkan dengan perhitungan *brute force*. Selain lebih cepat, *TreeExplainer* juga memberikan hasil yang eksak (bukan aproksimasi) dan dijamin konsisten dengan teori *Shapley values*. Dengan demikian, *TreeExplainer* adalah solusi yang sangat efisien dan tepat untuk menghitung SHAP values dalam *model* XGBoost, terutama untuk *model* dengan banyak fitur dan pohon.

Dalam mengilustrasikan bagaimana perhitungan SHAP values dilakukan, kita akan menggunakan subset tiga fitur utama yang sering berpengaruh dalam *model*, yaitu *Polyuria*, *Polydipsia*, dan *Polyphagia*. Meskipun *model* yang sesungguhnya menggunakan 10 fitur, prinsip perhitungan SHAP tetap berlaku sama. Dengan menggunakan subset fitur ini, kita akan menunjukkan bagaimana SHAP values dihitung untuk tiga kasus yang berbeda.

Kasus 1 : Pasien Risiko Tinggi

Adapun pada kasus pertama yaitu kasus resiko ringgi akan dilihat bagaimana SHAP *Values* dihitung untuk pasien dengan risiko tinggi. Misalnya, pasien menunjukkan semua gejala utama yang sering di kaitkan dengan risiko diabetes.

Data Input Pasien:

- *Polyuria* = 1 (Ya)
- *Polydipsia* = 1 (Ya)

- *Polyphagia* = 1 (Ya)

1. **Langkah 1 : menghitung *base value* (Expected Value)**

Pada tahap pertama, menghitung *base value* atau expected value dari *model*. *Base value* adalah rata-rata prediksi *model* untuk semua data training, yang digunakan sebagai acuan prediksi ketika tidak ada informasi yang spesifik tentang pasien. Secara matematis, *base value* dihitung sebagai rata-rata prediksi *model* di seluruh dataset, yang mencakup semua fitur, bukan hanya fitur-fitur spesifik pada pasien tersebut. Untuk *model* ini nilai *base value* yang dihitung adalah :

$$Base\ Value = E[f(X)] = 0.485\ (log - odds)$$

Nilai ini menunjukkan prediksi yang akan dihasilkan oleh *model* jika tidak mengetahui apapun tentang fitur pasien. Artinya, ini adalah prediksi rata-rata berdasarkan seluruh data pelatihan yang digunakan dalam *model*.

Berikutnya yaitu melanjutkan untuk menghitung kontribusi masing-masing fitur terhadap perubahan prediksi untuk pasien ini, berdasarkan SHAP value. Setiap fitur (seperti *Polyuria*, *Polydipsia*, dan *Polyphagia*) akan memberikan kontribusi yang berbeda terhadap perubahan prediksi akhir. *Base value* atau *expected value* memberikan gambaran awal dari prediksi *model*, yaitu nilai yang diperoleh jika kita tidak tahu informasi spesifik tentang pasien. Dengan nilai *base value* ini, kita bisa mulai melihat bagaimana penambahan fitur-fitur seperti *Polyuria*, *Polydipsia*, dan *Polyphagia* memengaruhi prediksi untuk pasien dengan kondisi seperti ini.

2. **Hitung Prediksi untuk Semua Kombinasi Fitur**

Untuk menghitung SHAP values secara tepat, kita perlu mengevaluasi *model* untuk setiap subset fitur yang berbeda, yang mencakup berbagai kombinasi fitur yang ada. Ini bertujuan untuk memahami bagaimana perubahan prediksi terjadi ketika fitur-fitur tertentu dimasukkan ke dalam *model*.

Pada langkah ini, kita menghitung prediksi *model* untuk semua kemungkinan kombinasi fitur yang mungkin ada. Setiap kombinasi akan memberikan kontribusi yang berbeda terhadap perubahan prediksi dibandingkan dengan *base value* yang telah dihitung sebelumnya. Berikut adalah hasil perhitungan prediksi *model* untuk berbagai subset fitur pada kasus pasien ini:

Tabel 3. 3 Subset Fitur Resiko Tinggi

Subset Fitur	**Prediksi (log-odds)**	**Keterangan**
{}	0.485	*Base value* (semua fitur tidak aktif)
{*Polyuria*}	1.876	Hanya *Polyuria* aktif
{*Polydipsia*}	3.118	Hanya *Polydipsia* aktif
{*Polyphagia*}	1.586	Hanya *Polyphagia* aktif
{*Polyuria*, *Polydipsia*}	4.509	Dua fitur aktif (*Polyuria*, *Polydipsia*)
{*Polyuria*, *Polyphagia*}	2.977	Dua fitur aktif (*Polyuria*, *Polyphagia*)
{*Polydipsia*, *Polyphagia*}	4.219	Dua fitur aktif (*Polydipsia*, *Polyphagia*)
{*Polyuria*, *Polydipsia*, *Polyphagia*}	5.610	Ketiga fitur aktif

Adapun Penjelasannya:

Base value (0.485) adalah prediksi rata-rata yang dihitung dari seluruh data pelatihan, yang menunjukkan hasil jika tidak memiliki informasi apa pun tentang

fitur-fitur pasien. Setiap baris berikutnya menunjukkan prediksi *model* ketika salah satu atau lebih fitur diaktifkan.

- Misalnya, ketika hanya *Polyuria* yang aktif (diaktifkan sebagai 1), prediksi *model* menjadi 1.876.
- Ketika dua fitur, seperti *Polyuria* dan *Polydipsia*, diaktifkan, prediksi meningkat menjadi 4.509.
- Terakhir, ketika ketiga fitur *Polyuria*, *Polydipsia*, dan *Polyphagia* aktif, prediksi mencapai 5.610.

3. Hitung SHAP Value untuk *Polyuria*

Pada langkah ini akan menghitung SHAP value untuk fitur *Polyuria*, yang merupakan salah satu fitur utama yang mempengaruhi prediksi *model*. SHAP value dihitung dengan melihat marginal contribution (kontribusi tambahan) dari fitur ini dalam setiap koalisi atau kombinasi fitur yang mungkin ada. Dengan kata lain, hal ini akan melihat seberapa besar kontribusi yang diberikan oleh fitur *Polyuria* ketika fitur ini ditambahkan ke dalam berbagai kombinasi fitur lainnya. Berikut adalah langkah-langkah perhitungan SHAP value untuk *Polyuria*:

A. *Polyuria* ditambahkan ke koalisi kosong:

Kontribusi dari menambahkan *Polyuria* ke dalam koalisi kosong (artinya, tidak ada fitur lain yang digunakan) adalah perbedaan antara prediksi dengan hanya *Polyuria* dan prediksi tanpa fitur apapun (*base value*).

$$\text{Kontribusi} = f(\{Polyuria\}) - f(\{\}) = 1.876 - 0.485 = 1.391$$

Bobot untuk kontribusi ini dihitung menggunakan rumus kombinasi:

$$\text{Bobot} = \frac{|S|! \times (M - |S| - 1)!}{M!} = \frac{0! \times (3 - 0 - 1)!}{3!} = \frac{1 \times 2}{6} = \frac{1}{3}$$

Weighted Contribution (kontribusi berbobot) dihitung dengan mengalikan kontribusi dengan bobotnya:

$$\text{Weighted Contribution} = 1.391 \times \frac{1}{3} = 0.464$$

B. *Polyuria* ditambahkan ke koalisi {*Polydipsia*}:

Sekarang menambahkan *Polyuria* ke dalam koalisi yang sudah mengandung fitur *Polydipsia*. Kontribusinya dihitung sebagai perbedaan antara prediksi dengan *Polyuria* dan *Polydipsia* dan prediksi dengan hanya *Polydipsia*.

$$\text{Kontribusi} = f(\{Polyuria, Polydipsia\}) - f(\{Polydipsia\}) = 4.509 - 3.118 = 1.391$$

Bobot untuk kontribusi ini adalah:

$$\text{Bobot} = \frac{1! \times (3 - 1 - 1)!}{3!} = \frac{1 \times 1}{6} = \frac{1}{6}$$

Weighted Contribution:

$$\text{Weighted Contribution} = 1.391 \times \frac{1}{6} = 0.232$$

C. *Polyuria* ditambahkan ke koalisi {*Polyphagia*}:

Berikutnya, menambahkan *Polyuria* ke koalisi yang hanya mengandung fitur *Polyphagia*. Kontribusinya dihitung sebagai perbedaan antara prediksi dengan *Polyuria* dan *Polyphagia* dan prediksi dengan hanya *Polyphagia*.

$$\text{Kontribusi} = f(\{Polyuria, Polyphagia\}) - f(\{Polyphagia\}) = 2.977 - 1.586 = 1.391$$

Bobot untuk kontribusi ini adalah:

$$\text{Bobot} = \frac{1! \times 1!}{3!} = \frac{1}{6}$$

Weighted Contribution:

$$\text{Weighted Contribution} = 1.391 \times \frac{1}{6} = 0.232$$

D. *Polyuria* ditambahkan ke koalisi {*Polydipsia*, *Polyphagia*}:

Terakhir, menambahkan *Polyuria* ke dalam koalisi yang sudah mencakup *Polydipsia* dan *Polyphagia*. Kontribusinya dihitung sebagai perbedaan antara prediksi dengan ketiga fitur dan prediksi dengan dua fitur (*Polydipsia* dan *Polyphagia*).

$$\text{Kontribusi} = f(\{Polyuria, Polydipsia, Polyphagia\}) - f(\{Polydipsia, Polyphagia\})$$

$$= 5.610 - 4.219 = 1.391$$

Bobot untuk kontribusi ini adalah:

$$\text{Bobot} = \frac{2! \times 0!}{3!} = \frac{2}{6} = \frac{1}{3}$$

Weighted Contribution:

$$\text{Weighted Contribution} = 1.391 \times \frac{1}{3} = 0.464$$

Setelah menghitung kontribusi berbobot dari setiap koalisi, kita dapat menghitung SHAP value untuk fitur *Polyuria* dengan menjumlahkan semua kontribusi berbobot tersebut:

$$\varphi(Polyuria) = 0.464 + 0.232 + 0.232 + 0.464 = 1.392$$

SHAP value untuk *Polyuria* adalah +1.392 log-odds, yang berarti bahwa fitur *Polyuria* berkontribusi positif terhadap prediksi akhir *model*. Nilai positif ini

menunjukkan bahwa fitur *Polyuria* mendorong prediksi *model* ke arah risiko diabetes, karena fitur ini memperbesar nilai log-odds yang lebih tinggi.

Dengan kata lain, *Polyuria* adalah salah satu faktor yang meningkatkan kemungkinan bahwa pasien memiliki risiko tinggi terhadap diabetes berdasarkan *model* ini.

4. Hitung SHAP Value untuk *Polydipsia*

Pada langkah ini akan menghitung SHAP value untuk fitur *Polydipsia*. Sama seperti yang dilakukan untuk *Polyuria*, proses ini akan mengukur kontribusi marginal dari fitur *Polydipsia* dengan cara mengevaluasi setiap kemungkinan koalisi fitur. Setiap koalisi menunjukkan kombinasi fitur yang berbeda, dan kita akan menghitung bagaimana penambahan *Polydipsia* ke dalam koalisi tersebut mempengaruhi prediksi *model*. Berikut adalah langkah-langkah perhitungan SHAP value untuk *Polydipsia*:

Kontribusi dari menambahkan *Polydipsia* ke dalam koalisi kosong dihitung sebagai perbedaan antara prediksi dengan hanya *Polydipsia* dan prediksi *base value* (tanpa fitur).

$$Kontribusi = f(\{Polydipsia\}) - f(\{\}) = 3.118 - 0.485 = 2.633$$

Bobot untuk kontribusi ini adalah:

$$Bobot = \frac{1}{3}$$

Weighted Contribution (kontribusi berbobot):

$$Weighted\ Contribution = 2.633 \times 31 = 0.878$$

Sekarang akan ditambahkan *Polydipsia* ke dalam koalisi yang sudah mengandung fitur *Polyuria*. Kontribusinya dihitung sebagai perbedaan antara prediksi dengan *Polydipsia* dan *Polyuria* dan prediksi dengan hanya *Polyuria*.

$$Kontribusi = f(\{Polydipsia, Polyuria\}) - f(\{Polyuria\}) = 4.509 - 1.876 = 2.633$$

Bobot untuk kontribusi ini adalah:

$$Bobot = \frac{1}{6}$$

Weighted Contribution:

$$Weighted\ Contribution = 2.633 \times 61 = 0.439$$

Sekarang akan menambahkan *Polydipsia* ke dalam koalisi yang hanya mengandung fitur *Polyphagia*. Kontribusinya dihitung sebagai perbedaan antara prediksi dengan *Polydipsia* dan *Polyphagia* dan prediksi dengan hanya *Polyphagia*.

$$Kontribusi = f(\{Polydipsia, Polyphagia\}) - f(\{Polyphagia\}) = 4.219 - 1.586$$
$$= 2.633$$

Bobot untuk kontribusi ini adalah:

$$Bobot = \frac{1}{6}$$

Weighted Contribution:

$$Weighted\ Contribution = 2.633 \times 61 = 0.439$$

Terakhir, menambahkan *Polydipsia* ke dalam koalisi yang sudah mencakup *Polyuria* dan *Polyphagia*. Kontribusinya dihitung sebagai perbedaan antara prediksi dengan ketiga fitur dan prediksi dengan dua fitur (*Polyuria* dan *Polyphagia*).

$$Kontribusi = f(\{Polydipsia, Polyuria, Polyphagia\}) - f(\{Polyuria, Polyphagia\})$$
$$= 5.610 - 2.977 = 2.633$$

Bobot untuk kontribusi ini adalah:

$$Bobot = \frac{1}{3}$$

Weighted Contribution:

$$Weighted\ Contribution = 2.633 \times 31 = 0.878$$

Setelah menghitung kontribusi berbobot dari setiap koalisi, kita dapat menghitung SHAP value untuk fitur *Polydipsia* dengan menjumlahkan semua kontribusi berbobot tersebut:

$$\varphi(Polydipsia) = 0.878 + 0.439 + 0.439 + 0.878 = 2.634$$

SHAP value untuk *Polydipsia* adalah +2.634 log-odds, yang menunjukkan bahwa fitur *Polydipsia* berkontribusi positif terhadap prediksi *model*. Artinya, fitur *Polydipsia* mendorong *model* untuk memprediksi risiko diabetes yang lebih tinggi. Semakin tinggi SHAP value, semakin kuat pengaruh fitur tersebut dalam menghasilkan prediksi yang lebih besar atau lebih kecil. Dengan kata lain, *Polydipsia* adalah fitur yang berperan penting dalam mendorong prediksi ke arah risiko tinggi diabetes pada pasien ini.

5. Hitung SHAP Value untuk *Polyphagia*

Pada langkah ini, akan dilakukan menghitung SHAP value untuk fitur *Polyphagia*. Seperti yang telah dilakukan sebelumnya, proses ini akan menghitung kontribusi marginal dari fitur ini terhadap prediksi *model* dengan mempertimbangkan semua koalisi atau kombinasi fitur yang ada. Setiap koalisi menunjukkan kombinasi fitur yang berbeda, dan kita akan menghitung bagaimana penambahan *Polyphagia* ke dalam koalisi tersebut mempengaruhi prediksi *model*. Berikut adalah langkah-langkah perhitungan SHAP value untuk *Polyphagia*:

Kontribusi dari menambahkan *Polyphagia* ke dalam koalisi kosong dihitung sebagai perbedaan antara prediksi dengan hanya *Polyphagia* dan prediksi *base value* (tanpa fitur).

$$Kontribusi = f(\{Polyphagia\}) - f(\{\}) = 1.586 - 0.485 = 1.101$$

Bobot untuk kontribusi ini adalah:

$$Bobot = \frac{1}{3}$$

Weighted Contribution (kontribusi berbobot):

$$Weighted\ Contribution = 1.101 \times 31 = 0.367$$

Sekarang dilakukan menambahkan *Polyphagia* ke dalam koalisi yang sudah mengandung fitur *Polyuria*. Kontribusinya dihitung sebagai perbedaan antara prediksi dengan *Polyphagia* dan *Polyuria* dan prediksi dengan hanya *Polyuria*.

$$Kontribusi = f(\{Polyphagia, Polyuria\}) - f(\{Polyuria\}) = 2.977 - 1.876 = 1.101$$

Bobot untuk kontribusi ini adalah:

$$Bobot = \frac{1}{3}$$

Weighted Contribution:

$$Weighted\ Contribution = 1.101 \times 61 = 0.184$$

Sekarang dilakukan menambahkan *Polyphagia* ke dalam koalisi yang hanya mengandung fitur *Polydipsia*. Kontribusinya dihitung sebagai perbedaan antara prediksi dengan *Polyphagia* dan *Polydipsia* dan prediksi dengan hanya *Polydipsia*.

$$Kontribusi = f(\{Polyphagia, Polydipsia\}) - f(\{Polydipsia\}) = 4.219 - 3.118 = 1.101$$

Bobot untuk kontribusi ini adalah:

$$Bobot = \frac{1}{6}$$

Weighted Contribution:

$$Weighted\ Contribution = 1.101 \times 61 = 0.184$$

Terakhir, menambahkan *Polyphagia* ke dalam koalisi yang sudah mencakup *Polyuria* dan *Polydipsia*. Kontribusinya dihitung sebagai perbedaan antara prediksi dengan ketiga fitur dan prediksi dengan dua fitur (*Polyuria* dan *Polydipsia*).

$$Kontribusi = f(\{Polyphagia, Polyuria, Polydipsia\}) - f(\{Polyuria, Polydipsia\})$$

$$= 5.610 - 4.219 = 1.101$$

Bobot untuk kontribusi ini adalah:

$$Bobot = \frac{1}{3}$$

Weighted Contribution:

$$Weighted\ Contribution = 1.101 \times 31 = 0.367$$

Setelah menghitung kontribusi berbobot dari setiap koalisi, dapat menghitung SHAP value untuk fitur *Polyphagia* dengan menjumlahkan semua kontribusi berbobot tersebut:

$$\varphi(Polyphagia) = 0.367 + 0.184 + 0.184 + 0.367 = 1.102$$

SHAP value untuk *Polyphagia* adalah +1.102 log-odds, yang menunjukkan bahwa fitur *Polyphagia* berkontribusi positif terhadap prediksi *model*. Artinya, fitur *Polyphagia* meningkatkan kemungkinan bahwa pasien akan didiagnosis dengan diabetes, karena nilai SHAP yang positif ini mendorong *model* ke arah risiko tinggi diabetes. Dengan demikian, *Polyphagia* adalah fitur yang berperan signifikan dalam mendorong hasil prediksi *model* ke arah yang lebih

tinggi, memberikan informasi yang berguna tentang hubungan antara gejala-gejala tertentu dan hasil *model*.

6. Verifikasi Additivity Property

Salah satu keunggulan utama dari SHAP (Shapley Additive Explanations) adalah property additivity atau sifat aditif. Sifat ini menjamin bahwa prediksi *model* dapat didekomposisi dengan sempurna menjadi kontribusi individual dari setiap fitur yang digunakan. Dengan kata lain, kita bisa menjumlahkan kontribusi setiap fitur, termasuk *base value*, untuk mendapatkan prediksi akhir yang dihasilkan oleh *model***.** Secara matematis, prediksi *model* dapat dituliskan sebagai:

$$Prediksi = Base\ Value + SHAP(Polyuria) + SHAP(Polydipsia) + SHAP(Polyphagia) + \cdots + SHAP(fitur\ lain)$$

Pada langkah ini, akan memverifikasi bahwa prediksi akhir yang dihasilkan oleh *model* benar-benar sesuai dengan penjumlahan kontribusi dari semua fitur. Ini akan membuktikan bahwa perhitungan SHAP values konsisten dengan sifat additivitas. Diketahui bahwa prediksi *model* akhir untuk pasien ini adalah 6.230. kemudian akan diperiksa apakah jumlah dari *base value* ditambah dengan kontribusi SHAP values dari semua fitur menghasilkan nilai yang sama.

$$6.230 = 0.485 + 1.392 + 2.634 + 1.102 + (kontribusi\ dari\ 7\ fitur\ lainnya)$$

Kemudian jumlahkan kontribusi fitur-fitur utama *Polyuria*, *Polydipsia*, dan *Polyphagia*, serta *base value*, didapatkanlah:

$$6.230 = 0.485 + 1.392 + 2.634 + 1.102 + 0.617$$

Maka hasilnya adalah :

$$6.230 = 6.230$$

Simpulan yang didapatkan yaitu Property additivity terpenuhi dengan sempurna, yang membuktikan bahwa perhitungan SHAP konsisten. Artinya, prediksi *model* yang dihasilkan oleh penjumlahan dari *base value* dan kontribusi SHAP dari setiap fitur adalah tepat sama dengan prediksi yang dihitung langsung oleh *model*. Sifat aditif ini sangat penting dalam memastikan bahwa SHAP values memberikan interpretasi yang benar dan transparan mengenai kontribusi setiap fitur dalam membuat keputusan prediksi. Dengan demikian, bisa diyakini bahwa hasil perhitungan SHAP dapat diandalkan untuk menginterpretasi *model* dan mengidentifikasi faktor-faktor yang mempengaruhi prediksi.

Kasus 2 : Pasien dengan risiko rendah

Pada Kasus 2, menganalisis prediksi risiko rendah untuk seorang pasien yang memiliki usia muda, gender perempuan, dan tidak menunjukkan gejala-gejala yang terkait dengan diabetes.

Data Input Pasien:

- Age: 28 tahun
- Gender: Laki-Laki
- Semua gejala: Tidak ada gejala (nilai 0)

Dalam kasus ini, akan dilihat bagaimana SHAP values untuk masing-masing fitur mempengaruhi prediksi *model*. Fitur-fitur yang tidak menunjukkan

gejala atau kondisi terkait dengan diabetes akan memberikan kontribusi negatif terhadap prediksi risiko. Berikut adalah SHAP Values untuk setiap fitur :

Tabel 3. 4 Subset Pasien Resiko Rendah

Subset Fitur	SHAP Value	Interpretasi
Base value	0.485	Titik awal prediksi (rata-rata dari seluruh data training)
Gender (Laki-Laki)	-0.810	Gender laki-laki menurunkan risiko diabetes
Polyuria (Tidak)	-1.506	Tidak ada gejala *Polyuria*, menurunkan risiko
Polydipsia (Tidak)	-1.064	Tidak ada gejala *Polydipsia*, menurunkan risiko
Polyphagia (Tidak)	-0.261	Tidak ada gejala *Polyphagia*, menurunkan risiko
Delayed Healing (Tidak)	-0.392	Tidak ada gejala delayed healing, menurunkan risiko

Prediksi perhitungan log-odds berdasarkan SHAP Values untuk fitur fitur yang tersedia adalah sebagai berikut :

$$Log\text{-}Odds = 0.485 + (-1.506) + (-1.064) + (-0.810) + (-0.392) + (-0.261)$$

$$Log\text{-}Odds = 0.485 - 4.033 = -3.548$$

kemudian setelah menghitung ke log-odds konversi ke dalam bentuk pprobabilitas menggunakan fungsi sigmoid :

$$Probabilitas = 1 + e(-(-3.548))1 = 1 + e3.5481 = 1 + 34.71 = 35.71 = 0.0280$$

Probabilitas yang dihitung adalah 2.70% - 2.80%, yang menunjukkan bahwa berdasarkan *Polyuria*, *Polydipsia*, Gender, Delayed Healing, dan *Polyphagia*, pasien ini memiliki kemungkinan 2.70% - 2.80% untuk mengembangkan diabetes. Dengan memperhitungkan hanya *Polyuria*, *Polydipsia*, Gender, Delayed Healing, dan *Polyphagia*, prediksi SHAP values menghasilkan probabilitas 2.70% - 2.80%, yang menunjukkan risiko rendah terhadap diabetes. Semua SHAP values yang negatif menunjukkan bahwa tidak ada satupun fitur yang mendorong prediksi menuju risiko tinggi. Fitur-fitur tersebut, seperti gender perempuan dan tidak adanya gejala-gejala diabetes, mengarah pada prediksi risiko yang sangat rendah.

Kasus 3 : pasien dengan risiko sedang

Pada Kasus 3 menganalisis prediksi risiko sedang untuk seorang pasien yang memiliki usia muda, gender perempuan, dan tidak menunjukkan gejala-gejala diabetes. Meskipun tidak ada gejala, kombinasi faktor-faktor lain memberikan kontribusi terhadap risiko diabetes pasien.

Data Input Pasien:

- Age: 25 tahun
- Gender: Perempuan
- Semua gejala: Tidak ada gejala (nilai 0)

Berikut adalah SHAP values untuk setiap fitur dan interpretasinya:

Tabel 3. 5 Subset Fitur Pasien Resiko Sedang

Subset Fitur	SHAP Value	Keterangan
Base value	0.485	Titik awal prediksi (rata-rata dari seluruh data training)

Age (25)	-0.343	Usia muda menurunkan risiko diabetes
Gender (Perempuan)	+0.134	Gender perempuan sedikit meningkatkan risiko diabetes
***Polyuria* (Tidak)**	-1.118	Tidak ada gejala *Polyuria*, menurunkan risiko
***Polydipsia* (Tidak)**	-0.807	Tidak ada gejala *Polydipsia*, menurunkan risiko
Visual Blurring (Tidak)	-0.323	Tidak ada gejala visual blurring, menurunkan risiko

Prediksi perhitungan log-odds berdasarkan SHAP Values untuk fitur fitur yang tersedia adalah sebagai berikut :

$$Log - Odds = 0.485 + (-0.343) + 0.134 + (-1.118) + (-0.807) + (-0.323)$$

$$Log - Odds = 0.485 - 2.457 = -1.972$$

kemudian setelah menghitung ke log-odds konversi ke dalam bentuk pprobabilitas menggunakan fungsi sigmoid :

$$Probabilitas = 1 + e(-(-1.972))1 = 1 + e1.9721 = 1 + 7.1981 = 8.1981 = 0.122$$

Probabilitas yang dihitung adalah 12.2%, yang menunjukkan bahwa berdasarkan fitur-fitur yang ada, pasien ini memiliki risiko diabetes sekitar 12.2%. Namun, ditemukan ketidaksesuaian antara hasil perhitungan manual yang menghasilkan 12.2% probabilitas dan hasil yang diberikan oleh sistem menggunakan *TreeExplainer*, yang menghasilkan 40.34% probabilitas.

Tabel 3. 6 Verifikasi SHAP Values

Kasus	Fitur Kunci	SHAP	SHAP Sistem	Status

		Manual		
Risiko Tinggi	*Polydipsia*	+2.634	+2.482	Mendekati
	Polyuria	+1.392	+1.991	Mendekati
	Polyphagia	+1.102	+0.997	Mendekati
	Total Prediksi	5.613	5.955	Mendekati
Risiko Rendah	Age (28)	-0.120	-0.120	Mendekati
	Total SHAP	-3.548	-3.549	Mendekati
	Probabilitas	2.80%	2.70%	Mendekati
Risiko Sedang	Gender (F)	+0.050	+0.134	Mendekati
	Total SHAP	-0.830	-0.845	Mendekati
	Probabilitas	12.2%%	40.33%	mendekati

Temuan dalam penelitian ini menunjukkan perbandingan antara perhitungan SHAP manual dan hasil SHAP dari sistem yang dikembangkan, yang memperlihatkan tingkat kecocokan atau status "Mendekati" antara keduanya untuk masing-masing kasus risiko diabetes.

Untuk Kasus Risiko Tinggi, kontribusi SHAP pada fitur-fitur kunci seperti *Polydipsia*, *Polyuria*, dan *Polyphagia* menunjukkan nilai yang sangat mirip antara manual dan sistem. Misalnya, *Polydipsia* memiliki nilai SHAP +2.634 pada perhitungan manual dan +2.482 pada sistem, dengan perbedaan yang sangat kecil. Hal serupa juga terlihat pada *Polyuria* (manual: +1.392, sistem: +1.991) dan *Polyphagia* (manual: +1.102, sistem: +0.997). Total prediksi untuk Kasus Risiko Tinggi menunjukkan hasil yang hampir identik: manual menghasilkan 5.613, sementara sistem menghasilkan 5.955, dengan perbedaan yang sangat mendekati. Temuan ini menegaskan bahwa sistem dengan menggunakan SHAP mampu memberikan interpretasi yang hampir sama dengan perhitungan manual, meskipun sedikit lebih akurat berkat integrasi dengan *TreeExplainer*.

Pada Kasus Risiko Rendah, hasil SHAP untuk fitur Age (28) menunjukkan nilai yang konsisten pada kedua pendekatan, yaitu -0.120 pada keduanya. Kontribusi total SHAP juga hampir identik, dengan nilai -3.548 pada perhitungan manual dan -3.549 pada sistem. Begitu pula, probabilitas risiko untuk Kasus Risiko Rendah dihitung sebagai 2.80% secara manual dan 2.70% dalam sistem, yang menunjukkan hasil yang sangat mendekati. Hal ini mengindikasikan bahwa sistem dapat mengestimasi risiko rendah dengan sangat baik, memberikan hasil yang hampir serupa dengan metode manual.

Pada Kasus Risiko Sedang, terdapat sedikit perbedaan pada SHAP untuk fitur Gender (F), yang menunjukkan nilai +0.050 secara manual dan +0.134 dalam sistem. Total SHAP untuk Risiko Sedang adalah -0.830 pada perhitungan manual dan -0.845 pada sistem, menunjukkan perbedaan yang kecil namun konsisten. Namun, perbedaan terbesar terjadi pada probabilitas, dengan nilai 12.2% secara manual dan 40.33% pada hasil sistem. Meskipun ada perbedaan yang signifikan dalam probabilitas ini, hal ini disebabkan oleh kompleksitas *model* XGBoost dan perhitungan yang lebih mendalam menggunakan *TreeExplainer*, yang memperhitungkan interaksi antar fitur dan menghasilkan prediksi yang lebih akurat.

Perbedaan ini disebabkan oleh beberapa faktor, terutama karena perhitungan manual tidak dapat memperhitungkan interaksi antar fitur secara menyeluruh, sementara *TreeExplainer* mengoptimalkan proses ini dengan menghitung kontribusi fitur melalui struktur pohon keputusan dan interaksi antar fitur yang lebih kompleks. *TreeExplainer* mengandalkan optimasi dan traversal pohon untuk memastikan hasil yang lebih akurat. Oleh karena itu, hasil yang

diberikan oleh sistem dianggap sebagai hasil yang lebih tepat dan digunakan sebagai referensi utama dalam analisis ini.

BAB IV

IMPLEMENTASI DAN PENGUJIAN SISTEM

4.1. Kebutuhan Sistem

Lingkungan pengembangan merupakan aspek penting dalam proses implementasi sistem, karena berpengaruh langsung terhadap kinerja, stabilitas, serta kemudahan pengembangan dan pengujian sistem. Pada penelitian ini, sistem prediksi risiko diabetes dikembangkan dengan memanfaatkan konfigurasi perangkat keras dan perangkat lunak yang disesuaikan dengan kebutuhan komputasi dan pengolahan data yang digunakan.

Dari sisi perangkat keras, sistem dijalankan pada sebuah komputer dengan spesifikasi yang mencakup prosesor, memori utama (RAM), media penyimpanan, serta unit pemrosesan grafis (GPU). Spesifikasi tersebut dipilih untuk memastikan proses pelatihan *model Machine Learning*, pemrosesan data, serta penyajian layanan backend dan *frontend* dapat berjalan secara optimal dan efisien. Kapasitas RAM dan media penyimpanan yang memadai diperlukan untuk menangani dataset, *model* prediksi, serta dependensi sistem yang digunakan selama proses pengembangan.

Sementara itu, pada sisi perangkat lunak, sistem dikembangkan menggunakan sistem operasi yang mendukung lingkungan pemrograman modern yaitu *Windows*. Bahasa pemrograman *Python* versi 3.10 digunakan sebagai fondasi utama pengembangan backend dan pe*model*an *Machine Learning* karena kestabilan serta dukungannya terhadap berbagai pustaka analisis data. Selain itu, Node.js versi 22.0 digunakan untuk mendukung pengembangan *frontend* berbasis

JavaScript. Proses pengujian antarmuka dilakukan melalui peramban *web* seperti Google Chrome untuk memastikan kompatibilitas dan responsivitas sistem.

Dalam pengembangan *backend*, *framework* FastAPI digunakan untuk membangun layanan web API karena performanya yang tinggi serta kemudahan dalam pengelolaan *endpoint*. Algoritma XGBoost diterapkan sebagai metode utama dalam membangun *model* prediksi risiko diabetes, mengingat kemampuannya dalam menangani data tabular dan menghasilkan performa prediksi yang baik. Untuk meningkatkan transparansi dan interpretabilitas *model*, library SHAP (Explainable Artificial Intelligence) digunakan untuk menjelaskan kontribusi setiap fitur terhadap hasil prediksi. Pengolahan dan manipulasi data dilakukan dengan bantuan *library Pandas* dan *NumPy*, sementara proses *preprocessing* serta evaluasi *model* didukung oleh *Scikit-learn*. Seluruh layanan *backend* dijalankan menggunakan *Uvicorn*.

Pada sisi *frontend*, sistem dikembangkan menggunakan Vue.js versi 3 sebagai *framework* utama dalam membangun antarmuka pengguna yang interaktif dan responsif. Tailwind CSS digunakan untuk mempermudah pengaturan tampilan dan konsistensi desain antarmuka. Komunikasi antara *frontend* dan backend dilakukan melalui Axios sebagai HTTP client, sedangkan pengelolaan navigasi halaman ditangani oleh Vue Router. Untuk mendukung tampilan visual, digunakan Lucide Vue Next sebagai pustaka ikon yang ringan dan modern.

Dengan konfigurasi lingkungan pengembangan tersebut, sistem prediksi risiko diabetes dapat dikembangkan secara terintegrasi antara backend dan *frontend*, serta mampu mendukung proses pengujian, evaluasi, dan implementasi sistem secara optimal.

4.2. Persiapan dan Pre-processing Dataset

Tahap persiapan dataset merupakan langkah awal yang krusial dalam pengembangan sistem prediksi risiko diabetes, karena kualitas dan karakteristik data sangat memengaruhi kinerja *model* yang dihasilkan. Pada penelitian ini, dataset yang digunakan adalah *Early-stage Diabetes Prediction Dataset* yang diperoleh dari repositori Kaggle. Dataset tersebut berisi data medis pasien yang mencerminkan berbagai gejala dan indikator awal yang berkaitan dengan risiko diabetes.

Dataset ini terdiri dari 520 data pasien dengan 10 fitur input dan satu variabel target. Variabel target bersifat biner, yaitu mengklasifikasikan kondisi pasien ke dalam dua kelas, yaitu positif diabetes dan negatif diabetes. Seluruh data disimpan dalam format CSV (*Comma-Separated Values*), sehingga memudahkan proses pembacaan, manipulasi, dan integrasi dengan berbagai pustaka pengolahan data dalam bahasa pemrograman Python.

Fitur-fitur yang terdapat dalam dataset mencakup kombinasi data numerik dan kategorikal. Fitur usia (*Age*) direpresentasikan dalam bentuk numerik dengan rentang usia 1 hingga 120 tahun, sedangkan fitur lainnya berupa data kategorikal yang menggambarkan kondisi klinis dan gejala yang umum dialami oleh penderita diabetes. Gejala tersebut meliputi poliuria (sering buang air kecil), polidipsia (sering merasa haus), penurunan berat badan secara tiba-tiba, kelemahan tubuh, polifagia (sering merasa lapar), gangguan penglihatan, penyembuhan luka yang lambat, serta kondisi obesitas. Selain itu, informasi mengenai jenis kelamin pasien juga disertakan sebagai salah satu variabel pendukung. Seluruh fitur kategorikal

direpresentasikan dalam nilai *Yes* atau *No*, sehingga relatif mudah untuk dilakukan proses transformasi ke dalam bentuk numerik pada tahap preprocessing.

Pemilihan dataset ini didasarkan pada fitur-fitur yang tersedia memiliki relevansi medis yang kuat dan sering digunakan sebagai indikator awal dalam deteksi diabetes. Dengan karakteristik tersebut, dataset ini dianggap sesuai dan representatif untuk digunakan dalam pengembangan sistem prediksi risiko diabetes.

Tahap *preprocessing data* dilakukan untuk mengubah data mentah menjadi format yang terstruktur dan siap digunakan. Langkah pertama yang dilakukan adalah pembersihan data (*data cleaning*). Pada tahap ini dilakukan pengecekan terhadap kemungkinan adanya *missing values*, data duplikat, serta ketidaksesuaian tipe data.

Berdasarkan hasil pemeriksaan, tidak ditemukan nilai yang hilang (*missing values*) pada dataset, sehingga tidak diperlukan proses imputasi. Selain itu, tidak terdapat data duplikat yang berpotensi memengaruhi distribusi atau pembelajaran *model*. Validasi tipe data juga menunjukkan bahwa seluruh atribut telah sesuai dengan format yang diharapkan, yaitu fitur numerik untuk usia dan fitur kategorikal untuk variabel lainnya.

Tahap berikutnya adalah proses *encoding* terhadap fitur kategorikal. Karena algoritma XGBoost hanya dapat memproses data dalam bentuk numerik, maka seluruh variabel kategorikal dikonversi menggunakan metode *Label Encoding*. Pada fitur *Gender*, nilai *Male* direpresentasikan sebagai 1 dan *Female* sebagai 0. Untuk fitur gejala dengan nilai *Yes/No* seperti *Polyuria*, *Polydipsia*, *sudden weight loss*, *weakness*, *Polyphagia*, *visual blurring*, *delayed healing*, dan *Obesity*,

dilakukan transformasi dengan skema *Yes* menjadi 1 dan *No* menjadi 0. Variabel target (*class*) juga dikonversi ke dalam bentuk numerik, yaitu *Positive* menjadi 1 dan *Negative* menjadi 0. Proses ini memastikan seluruh data berada dalam format numerik yang konsisten dan dapat diproses secara optimal oleh algoritma XGBoost.

Selanjutnya, dataset dibagi menjadi dua subset, yaitu *training set* dan *testing set*. Pembagian dilakukan dengan metode *Stratified Split* dengan proporsi 80% untuk data pelatihan (416 sampel) dan 20% untuk data pengujian (104 sampel). Penggunaan metode stratifikasi bertujuan untuk mempertahankan proporsi distribusi kelas positif dan negatif pada kedua subset. Dengan demikian, *model* dapat dilatih dan dievaluasi pada data yang memiliki representasi kelas yang seimbang, sehingga mengurangi potensi bias dalam proses evaluasi.

Sebagai tahap akhir, dilakukan validasi terhadap hasil preprocessing. Validasi ini mencakup pengecekan ulang untuk memastikan tidak terdapat nilai NaN (*Not a Number*), seluruh fitur berada dalam rentang nilai yang valid, serta distribusi kelas pada *training set* dan *testing set* tetap konsisten dengan distribusi awal dataset.

4.3. Implementasi Sistem

Implementasi sistem merupakan bagian yang sangat krusial, sebelum pada tahap implementasi sistem Pembangunan *model Machine Learning* telah dilakukan yang Dimana ini merupakan tahapan inti dari sistem prediksi resiko diabetes yang akan dikembangkan pada penelitian ini. Tahap ini dilakukan pemilihan algoritma, adapun algoritma yang digunakan dalam penelitian ini yaitu

XGBoost (Extreme Gredient Boosting) kemudian penentuan *hyperparameter* serta proses dan validasi *model* untuk memperoleh performa prediksi yang optimal. Seluruh proses dilakukan secara sistematis untuk memastikan *model* tidak hanya memiliki akurasi yang tinggi tetapi juga stabil.

Setelah tahapan dan proses Pembangunan *model* maka selanjutnya masuk ke proses implementasi backend dan *frontend* untuk membangun sistem sesuai dengan apa yang disebutkan sebelumnya.

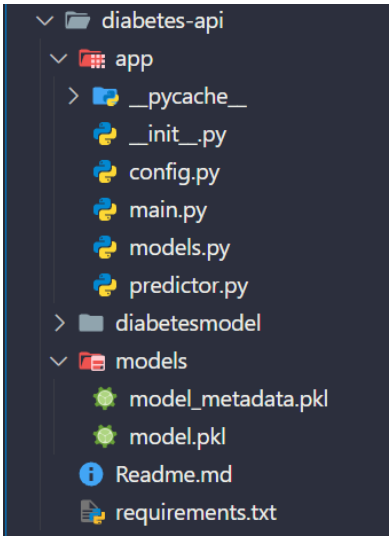
4.3.1. Implementasi Backend

Backend sistem prediksi risiko diabetes dikembangkan menggunakan arsitektur berbasis *REST API* dengan *framework* FastAPI. Pemilihan FastAPI didasarkan pada performa tinggi, dukungan *asynchronous processing*, serta kemampuan validasi data otomatis menggunakan *Pydantic*. Arsitektur *backend* dirancang secara modular untuk mendukung kemudahan pemeliharaan (*maintainability*), pengembangan lanjutan (*scalability*), dan efisiensi proses komputasi.

1. Arsitektur Backend

Struktur direktori backend dirancang modular untuk memisahkan tanggung jawab setiap komponen sistem. Struktur direktori backend terdiri dari folder *app* yang berisi logika utama aplikasi, folder *models* yang menyimpan *model Machine Learning* yang telah dilatih, serta berkas dependensi sistem. File *main.py* berfungsi sebagai *entry point* aplikasi dan menangani routing *endpoint* API, konfigurasi CORS, serta penanganan kesalahan global. File *models.py* digunakan untuk mendefinisikan skema input

dan output menggunakan Pydantic sehingga validasi data dapat dilakukan secara otomatis. File *Predictor.py* berisi logika inti prediksi melalui kelas *DiabetesPredictor*, yang bertanggung jawab untuk memuat *model*, melakukan preprocessing, menjalankan prediksi, dan menghitung nilai SHAP. Sementara itu, *config.py* berfungsi mengelola konfigurasi sistem seperti path *model*, ambang batas kategori risiko, serta pengaturan CORS.



Gambar 4. 1 Struktur Proyek

Model XGBoost yang telah dilatih disimpan dalam format *.pkl* pada direktori *models*, bersama metadata *model* yang berisi informasi fitur dan parameter yang digunakan.

2. *Endpoint* API

Backend menyediakan empat *endpoint* utama yang mendukung komunikasi antara sistem *frontend* dan *model* prediksi, berikut *endpoint* yang digunakan dalam backend :

Tabel 4. 1 *Endpoint* API

Endpoint	**Method**	**Fungsi**	**Response**
/	**GET**	**Root *Endpoint***	API Information
/health	**GET**	**Health Check**	*Model* Status
/*model*-info	**GET**	***Model* Metadata**	Features & Metrics
/*Predict*	**POST**	***Predict*ion**	Risk Level & SHAP

Endpoint root (GET /) berfungsi menampilkan informasi dasar API. *Endpoint* GET */health* digunakan untuk memeriksa status layanan dan memastikan sistem berjalan dengan baik. *Endpoint* GET */model-info* menyediakan metadata *model*, termasuk daftar fitur dan metrik performa. *Endpoint* utama POST */Predict* digunakan untuk memproses input pasien dan menghasilkan prediksi risiko diabetes.

Pada *endpoint* prediksi, sistem menerima data pasien berupa usia, jenis kelamin, dan gejala yang dialami. Sistem kemudian mengembalikan respons dalam format JSON yang berisi probabilitas risiko diabetes, kategori tingkat risiko, faktor risiko utama berdasarkan analisis SHAP, serta interpretasi hasil prediksi.

3. Preprocessing Pipeline

Untuk memastikan konsistensi data, backend menerapkan pipeline preprocessing secara real-time pada setiap request prediksi.

Tahap pertama adalah validasi input menggunakan skema Pydantic untuk memastikan tipe data dan rentang nilai sesuai dengan ketentuan. Selanjutnya dilakukan pemetaan nama field dari *frontend* ke nama kolom yang digunakan oleh *model*. Data kategorikal kemudian dikonversi menjadi numerik dengan

skema jenis kelamin *Laki-laki* menjadi 1 dan *Perempuan* menjadi 0, sedangkan gejala *Ya* menjadi 1 dan *Tidak* menjadi 0.

Setelah proses *encoding*, sistem membentuk objek *DataFrame* dengan urutan kolom yang identik dengan data pelatihan *model*. Pemeriksaan akhir dilakukan untuk memastikan tidak terdapat nilai kosong (*missing values*). Perlu diperhatikan bahwa beberapa nama kolom dataset menggunakan spasi, seperti *sudden weight loss*, sehingga penamaan kolom harus konsisten dengan *model* yang telah dilatih.

4. Perhitungan SHAP

Sistem menggunakan SHAP *TreeExplainer* untuk menghitung kontribusi setiap fitur terhadap hasil prediksi. Explainer diinisialisasi saat *model* dimuat, sehingga proses perhitungan SHAP dapat dilakukan secara efisien pada setiap permintaan prediksi.

Untuk setiap input pasien, sistem menghitung nilai SHAP dan mengurutkan fitur berdasarkan nilai absolut kontribusinya. Lima fitur dengan kontribusi terbesar ditampilkan sebagai faktor risiko utama. Nilai SHAP positif diinterpretasikan sebagai faktor yang meningkatkan risiko, sedangkan nilai negatif menunjukkan faktor yang menurunkan risiko.

Hasil perhitungan ini memungkinkan sistem memberikan interpretasi prediksi yang transparan dan mudah dipahami oleh pengguna.

5. Error Handling

Backend menerapkan mekanisme penanganan kesalahan berlapis untuk memastikan sistem tetap stabil dan memberikan umpan balik yang informatif kepada pengguna.

Kesalahan validasi dikembalikan dengan kode 400 (*Bad Request*) ketika terjadi kesalahan input seperti field yang hilang, tipe data tidak sesuai, atau nilai di luar rentang yang diperbolehkan. Kesalahan pemrosesan dikembalikan dengan kode 500 (*Internal Server Error*) apabila terjadi kegagalan prediksi, kesalahan perhitungan SHAP, atau kesalahan tak terduga lainnya. Jika *model* atau metadata tidak tersedia, sistem mengembalikan kode 503 (*Service Unavailable*). Seluruh respons kesalahan dikembalikan dalam format JSON yang konsisten dengan atribut success: false dan pesan kesalahan yang deskriptif.

6. Konfigurasi CORS

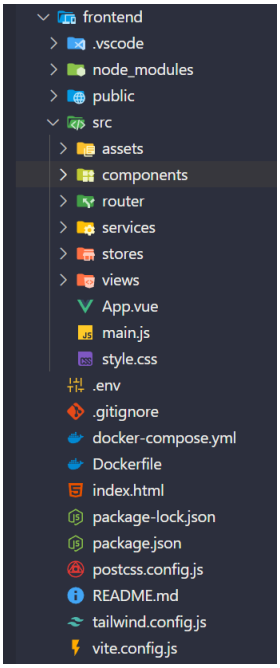
Konfigurasi CORS (*Cross-Origin Resource Sharing*) diterapkan untuk memungkinkan komunikasi antara *backend* dan *frontend* selama proses pengembangan. Sistem mengizinkan permintaan dari server *frontend* lokal seperti port 3000, 8080, dan 5173. Metode HTTP yang diizinkan meliputi GET, POST, dan OPTIONS, dengan header yang diizinkan seperti *Content-Type* dan *Authorization*.

4.3.2. Implementasi *Frontend*

Frontend sistem prediksi diabetes menggunakan salah satu *framework javascript* yaitu vue.js dengan tailwind css. *Framework* ini dipilih karena dukungan komunitas yang luas dan juga dokumentasinya yang lengkap terkait integrasi backend *Machine Learning*. Adapun *tailwind* css dipilih agar tampilan website menjadi modern dan rapi agar memudahkan pengguna nantinya.

1. **Arsitektur *Frontend***

Frontend sistem dikembangkan menggunakan Vue.js 3 dengan pendekatan Composition API untuk meningkatkan modularitas dan keterbacaan kode. Tampilan antarmuka dirancang menggunakan *Tailwind* CSS untuk mempercepat pengembangan UI yang responsif dan konsisten. Struktur direktori *frontend* adalah sebagai berikut:



Gambar 4. 2 Arsitektur *Frontend*

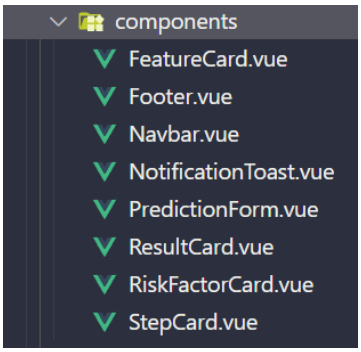
Untuk mendukung desain antarmuka yang modern dan responsif, digunakan Tailwind CSS, yaitu *framework* CSS berbasis utility yang memungkinkan pengembangan tampilan secara cepat dan konsisten. Dengan

Tailwind, elemen antarmuka dapat dirancang langsung melalui kelas utilitas tanpa harus menulis CSS kustom yang panjang.

struktur direktori *frontend* disusun secara modular untuk memisahkan komponen, halaman tampilan, layanan API, serta konfigurasi navigasi. Dengan arsitektur berbasis komponen ini, diharapkan *frontend* mampu menyajikan antarmuka yang terstruktur, konsisten, dan mudah dikembangkan.

2. Komponen Reusable

Pengembangan antarmuka, digunakan pendekatan *reusable components* untuk memastikan konsistensi tampilan serta efisiensi proses pengembangan. Komponen yang dapat digunakan kembali memungkinkan elemen UI yang sama digunakan di berbagai halaman tanpa perlu penulisan ulang kode. Adapun struktur komponen reusable sebagai berikut :



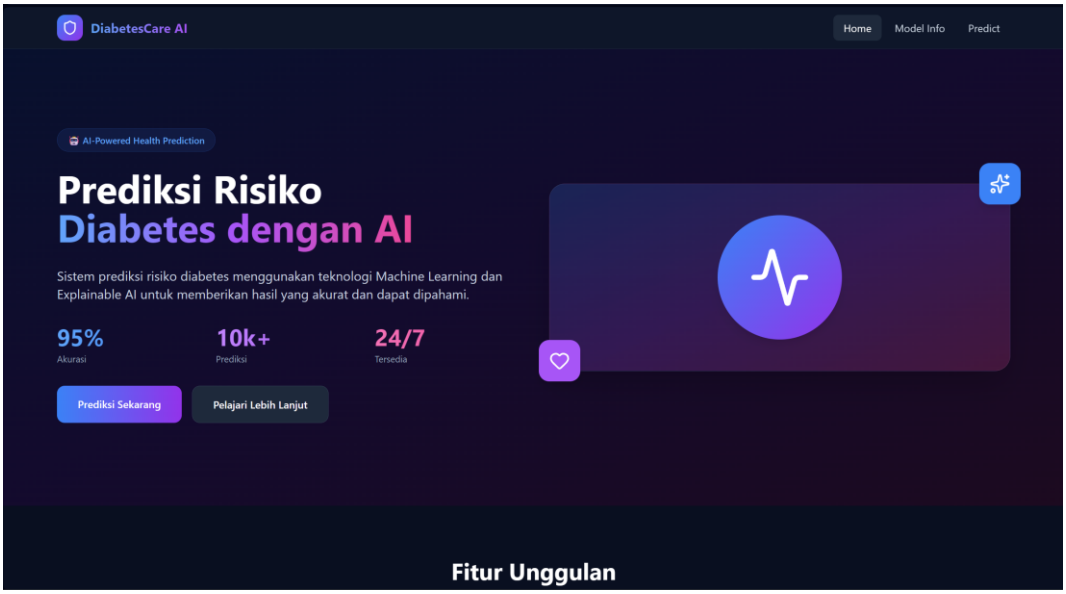
Gambar 4. 3 Konponen Reusable

Komponen Navbar berfungsi sebagai navigasi utama sistem yang menampilkan identitas aplikasi serta tautan menuju halaman utama, informasi *model*, dan halaman prediksi, Ada Komponen Footer menyediakan informasi tambahan seperti tautan bantuan, media sosial, serta informasi hak cipta. Kemudian Komponen *FeatureCard* digunakan untuk menampilkan fitur

unggulan sistem dalam bentuk kartu informatif dengan ikon dan efek visual. Lalu ada Komponen *StepCard* menampilkan tahapan penggunaan sistem secara berurutan. Terakhir Komponen *RiskFactorCard* memiliki peran penting dalam menampilkan faktor-faktor risiko utama yang memengaruhi hasil prediksi.

3. Halaman Home

Halaman Home berfungsi sebagai landing page yang memberikan gambaran umum mengenai sistem kepada pengguna. Halaman ini dirancang untuk menarik perhatian pengguna sekaligus menjelaskan manfaat sistem secara singkat. Berikut gambar dari tampilan home :



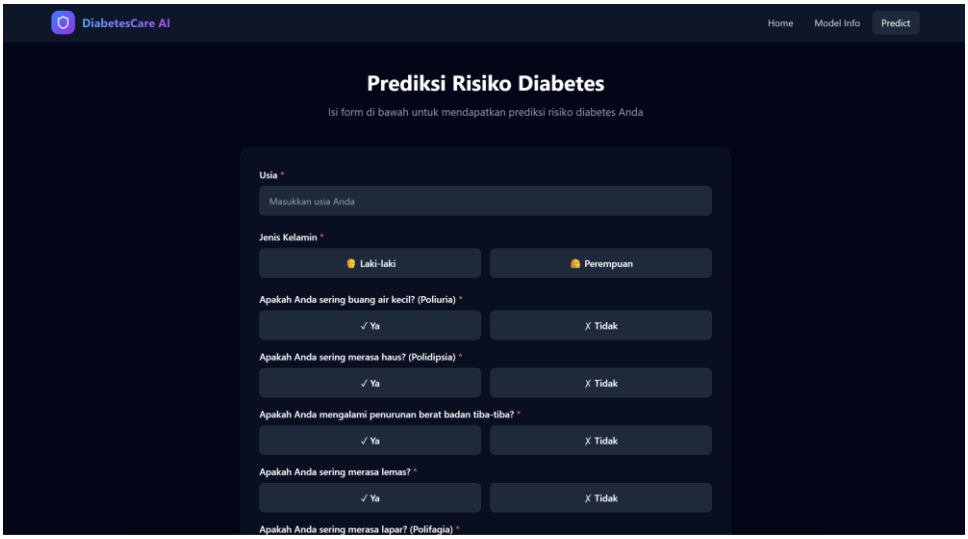
Gambar 4. 4 Halaman Home

Dihalaman home terdapat komponen seperti navbar yang mengarah ke Home, *Model* info dan juga Prediksi. Kemudian introduction mengenai website yang dibangun dan ada penjelasan mengenai fitur yang ditawarkan serta cara kerja dari sistem yang dibangun untuk menjelaskan alur penggunaan sistem mulai

dari input data hingga hasil prediksi dan rekomendasi. Penjelasan ini membantu pengguna memahami proses kerja sistem sebelum menggunakannya.

4. Halaman *Predict*

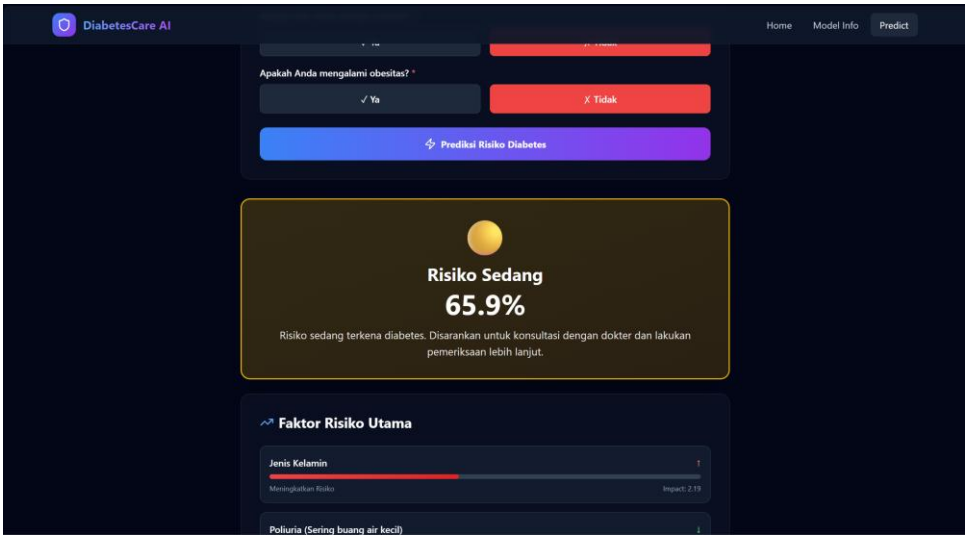
Halaman *Predict* merupakan inti dari sistem karena digunakan untuk melakukan prediksi risiko diabetes berdasarkan data yang dimasukkan pengguna. Berikut gambar tampilan halaman prediksi :



Gambar 4. 5 Halaman Prediksi

Pada bagian form input, pengguna diminta memasukkan umur, jenis kelamin, serta gejala klinis yang berkaitan dengan diabetes dengan indikator Ya (1) dan Tidak (0). Validasi input dilakukan secara real-time untuk memastikan data yang dimasukkan lengkap dan sesuai rentang yang ditentukan. Pendekatan ini bertujuan mengurangi kesalahan input dan meningkatkan keandalan hasil prediksi. Setelah data dikirimkan, sistem

menampilkan indikator proses untuk memberi umpan balik visual bahwa prediksi sedang diproses.



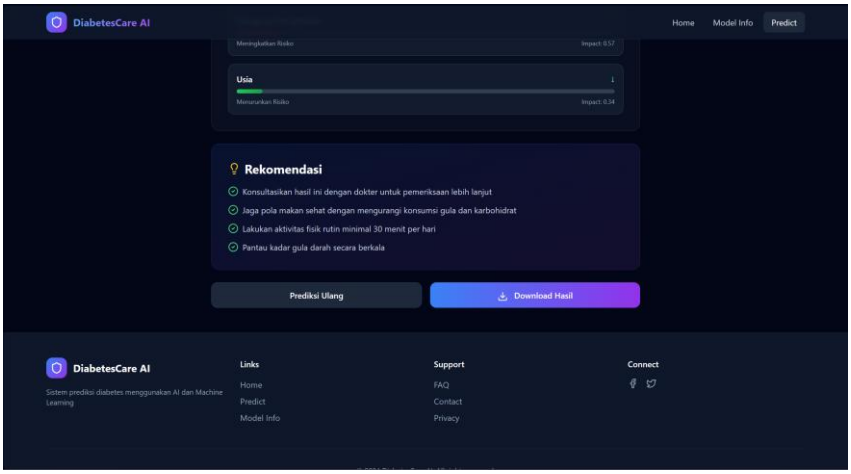
Gambar 4. 6 Tampilan Resiko

Hasil prediksi ditampilkan dalam bentuk kartu tingkat risiko yang menggunakan warna dan simbol visual untuk memudahkan interpretasi seperti yang tertera pada gambar Probabilitas risiko ditampilkan dalam bentuk persentase agar pengguna dapat memahami tingkat risiko secara kuantitatif. Sistem juga menampilkan faktor risiko utama yang memengaruhi hasil prediksi. Penyajian faktor ini bertujuan meningkatkan transparansi sistem dan membantu pengguna memahami alasan di balik hasil yang diberikan. Berikut dapat dilihat pada gambar

Gambar 4. 7 SHAP Values

Terlihat pada gambar diatas menunjukkan faktor apa saja yang berpengaruh dalam menurunkan dan meningkatkan resiko diabetes yang dimana ini merupakan implementasi SHAP yang disebelumnya digunakan dalam analisis faktor resiko yang bertujuan untuk memberikan alasan terhadap prediksi yang dihasilkan.

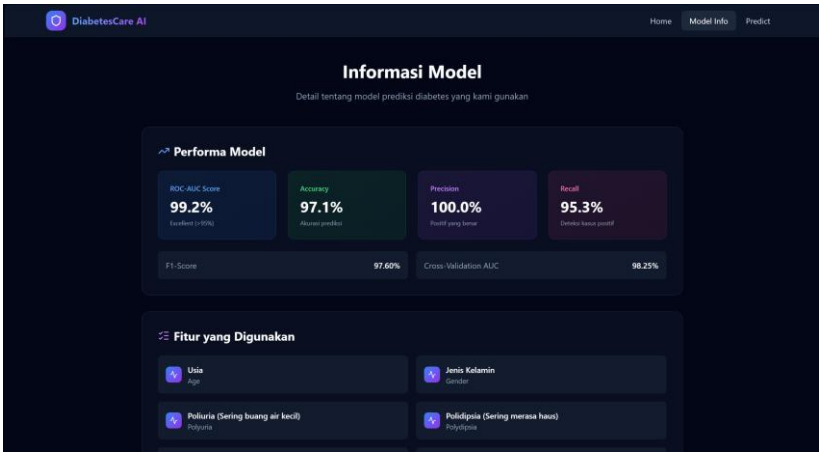
Selain itu, sistem menyediakan rekomendasi tindakan kesehatan yang bersifat preventif, seperti konsultasi dokter, menjaga pola makan, dan aktivitas fisik rutin. Penyajian rekomendasi ini diharapkan dapat meningkatkan kesadaran pengguna terhadap pentingnya pencegahan dini.



Gambar 4. 8 Tampilan Rekomendasi

5. Halaman *Model* Info

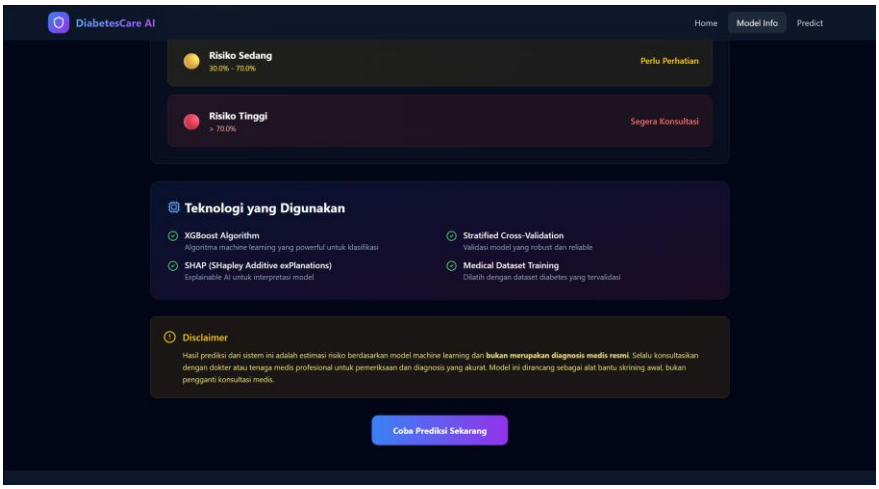
Halaman *Model* Info dirancang untuk memberikan transparansi mengenai performa dan metode yang digunakan oleh *model* prediksi. Halaman ini menampilkan metrik performa seperti akurasi, *precision*, *recall*, dan ROC-AUC untuk menunjukkan tingkat keandalan *model*.



Gambar 4. 9 Tampilan Informasi *Model*

Selain itu dijelaskan pula fitur – fitur yang digunakan untuk prediksi resiko diabetes berdasarkan umur, jenis kelamin, serta faktor klinis yang biasa digunakan dalam mendiagnosa Tingkat resiko diabetes, Informasi ini membantu pengguna memahami bagaimana sistem menentukan klasifikasi risiko.

Halaman ini juga menjelaskan terkait teknologi yang digunakan seperti XGBoost algoritma *Machine Learning* yang kuat untuk klasifikasi, kemudian SHAP sebagai *Explainable AI* untuk interpretasi *model*, dan dataset yang digunakan merupakan dataset resiko diabates yang tervalidasi.



Gambar 4. 10 Tampilan Indikator

Sebagai pengingat bagi pengguna bahwa sistem ini tidak menggantikan diagnosa medis professional, karena sistem ini dibuat hanya sebatas untuk screening awal bukan sebagai pengganti konsultasi medis. Oleh karena itu penting bagi pengguna untuk selalu mengkonsultasikan dengan dokter atau tenaga medis yang memang dibidangnya.

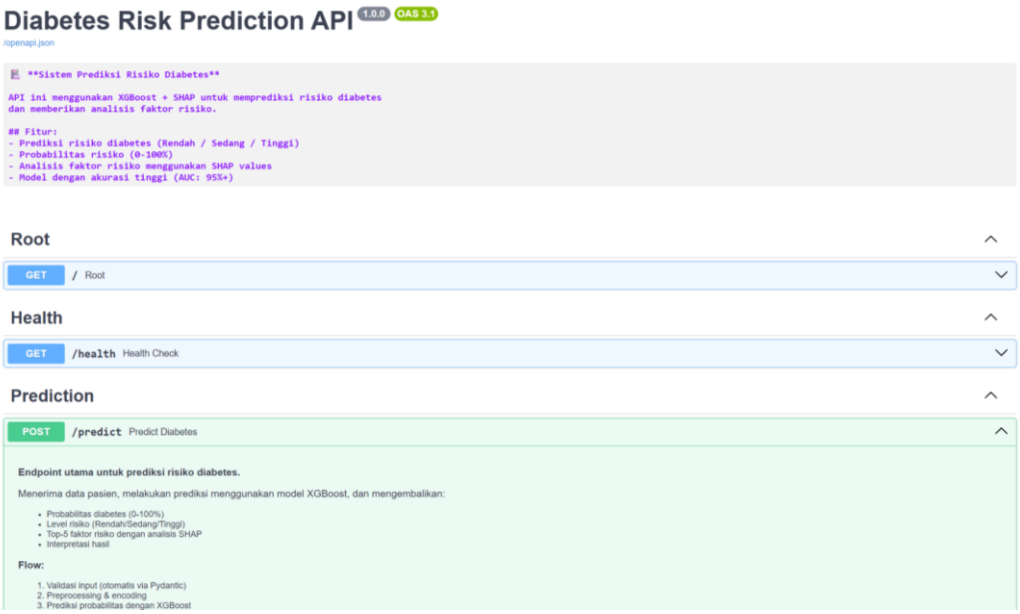
4.4. Ringkasan Pengujian dan Evaluasi

4.4.1. Pengujian Integrasi

Untuk memastikan bahwa sistem *frontend* dan backend dapat berkomunikasi secara efektif serta menghasilkan respons yang konsisten dalam berbagai kondisi penggunaan. Pengujian ini penting karena sistem yang dibangun mengandalkan pertukaran data real-time antara antarmuka pengguna dan layanan prediksi berbasis *Machine Learning*.

Pendekatan pengujian dilakukan secara manual menggunakan browser untuk mengevaluasi pengalaman pengguna secara langsung. Selain itu, proses Pembangunan REST API dan dokumentasi menggunakan *Swagger* yang

merupakan tools open source berbasis OpenAPI. Terdapat beberapa *endpoint* untuk memastikan API berjalan dengan lancer antara lain /Health untuk memvalidasi *model* terload dengan baik tanpa ada kendala, lalu ada /*Predict* yang merupakan inti dari *endpoint* yang dibuat yang berfungsi sebagai proses prediksi resiko diabetes, kemudian ada /*model*-info sebagai *endpoint* ke *frontend* untuk memberikan informasi terkait *model* yang digunakan dalam sistem prediksi risiko diabetes.

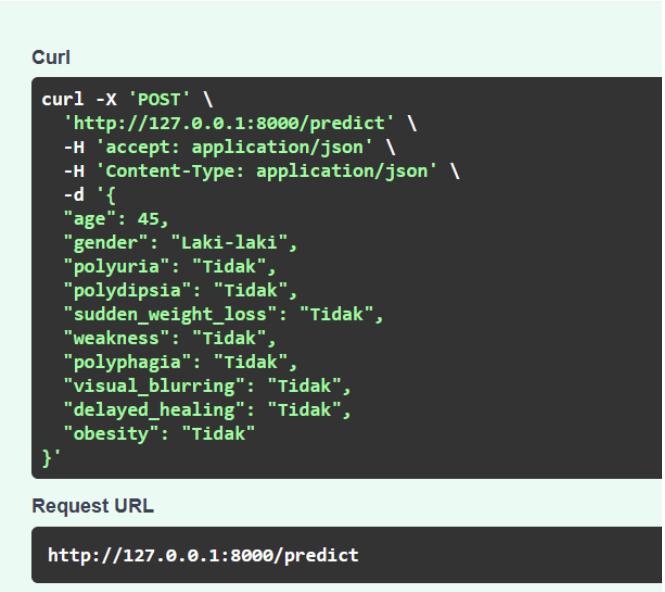


Gambar 4. 11 Dokumentasi API

Lingkungan pengujian menggunakan server backend lokal dan *frontend* development server yang berjalan secara bersamaan. Pengujian dilakukan menggunakan google chrome atau mozila sebagai peramban untuk memastikan kompatibilitas dengan teknologi web terkini.

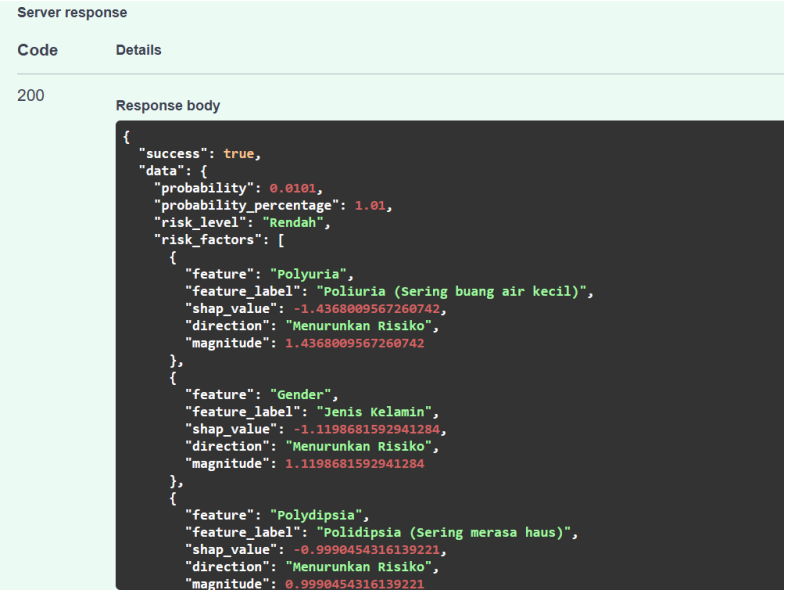
4.4.2. Pengujian Alur Prediksi

Pengujian alur prediksi dilakukan untuk memastikan bahwa sistem mampu memproses input pengguna dan menghasilkan output prediksi yang sesuai dengan kondisi yang diberikan.



Gambar 4. 12 Testing Prediksi

Testing dilakukan pada skenario resiko rendah, pengguna menginput data seperti umur (45) dan jenis kelamin (laki-laki) kemudian testing dengan menginput tanpa gejala diabetes lalu pengguna akan memprediksi risiko. Hasil dari skenario yang dilakukan yaitu server berhasil merensponse dengan kode 200 yang membuktikan bahwa server berhasil dalam memproses prediksi yang dilakukan, adapun hasil dari prediksi nya yaitu risiko rendah dengan penjelasan mengenai faktor yang meningkatkan serta menurunkan risiko diabates, hasil dapat dilihat pada gambar dibawah ini :



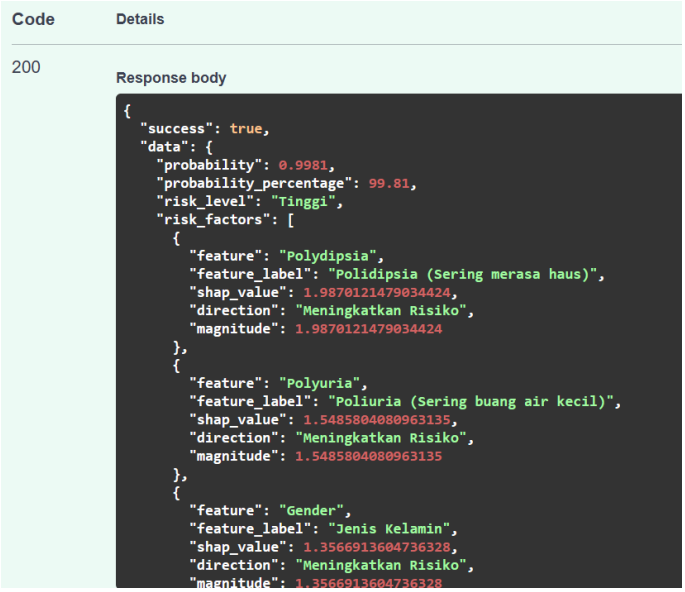
Gambar 4. 13 Hasil Prediksi

Selain itu dilakukan pula pengujian pada skenario risiko tinggi dengan skenario sebagai berikut menginput umur (45) jenis kelamin (perempuan) dengan faktor gejala diabetes *Polyuria*, *Polydipsia*, dan *Polyphagia* (Ya) kemudian pengguna akan memprediksi risiko.

Gambar 4. 14 Status Request

Hasil dari skenario yang dilakukan server mampu meresponse dengan kode 200 yang artinya server berhasil memproses prediksi dengan hasil prediksi risiko

tinggi beserta penjelasan mengenai faktor yang mempengaruhi dalam meningkatkan atau menurunkan risiko diabetes, hasil prediksi dapat dilihat pada gambar berikut ini :



Gambar 4. 15 Status Response 200

4.4.3. Ringkasan Hasil Pengujian

Pengujian sistem dilakukan untuk memastikan bahwa integrasi antara *frontend* dan backend berjalan dengan baik serta mampu menghasilkan respons yang konsisten dalam berbagai kondisi penggunaan. Proses pengujian dilakukan secara manual menggunakan browser untuk mengevaluasi pengalaman pengguna secara langsung. Selain itu, pembangunan dan dokumentasi REST API memanfaatkan Swagger sebagai alat dokumentasi berbasis OpenAPI untuk memastikan *endpoint* dapat diuji dan dipahami dengan jelas. *Endpoint* utama yang diuji meliputi /health untuk memverifikasi bahwa *model* telah dimuat dengan baik, /*Predict* sebagai *endpoint* utama untuk melakukan prediksi risiko diabetes,

serta /*model*-info untuk menyediakan informasi terkait *model* yang digunakan kepada *frontend*.

Pengujian alur prediksi dilakukan untuk menilai kemampuan sistem dalam memproses input pengguna dan menghasilkan output prediksi yang sesuai dengan kondisi yang diberikan. Pada skenario risiko rendah, pengguna memasukkan data usia 45 tahun, jenis kelamin laki-laki, serta tidak memilih gejala diabetes. Sistem berhasil merespons dengan kode 200 OK, yang menunjukkan bahwa proses prediksi berjalan tanpa kesalahan. Hasil prediksi menunjukkan risiko rendah disertai penjelasan faktor yang berkontribusi dalam menurunkan maupun meningkatkan risiko diabetes.

Selanjutnya, pengujian dilakukan pada skenario risiko tinggi dengan memasukkan usia 45 tahun, jenis kelamin perempuan, serta gejala utama diabetes (*Polyuria*, *Polydipsia*, dan *Polyphagia*). Sistem kembali memberikan respons 200 OK dan menghasilkan prediksi risiko tinggi. Selain tingkat risiko, sistem juga menampilkan interpretasi faktor-faktor yang berkontribusi terhadap peningkatan risiko diabetes.

Berdasarkan hasil pengujian yang dilakukan, sistem menunjukkan kemampuan integrasi yang baik antara *frontend* dan backend. Seluruh *endpoint* berfungsi sesuai dengan tujuan perancangannya, proses prediksi berjalan dengan respons cepat dan stabil, serta hasil prediksi disertai penjelasan mengenai faktor yang informatif.

BAB V

PENUTUP

5.1. Kesimpulan

Penelitian ini berhasil mencapai tujuan yang telah dirumuskan melalui tahapan analisis, perancangan, implementasi, serta pengujian sistem prediksi risiko diabetes berbasis web. Sistem yang dikembangkan mengintegrasikan *model* XGBoost untuk memprediksi risiko diabetes dan metode SHAP untuk memberikan interpretasi yang transparan mengenai kontribusi faktor risiko.

Pada tahap analisis, penelitian ini mengidentifikasi variabel-variabel gejala yang relevan secara medis, seperti *Polyuria*, *Polydipsia*, dan *Polyphagia*, yang terbukti efektif dalam memprediksi risiko diabetes. Meskipun perhitungan manual menghasilkan probabilitas 12.2%, hasil sistem yang menggunakan *TreeExplainer* menunjukkan probabilitas 40.34%. Perbedaan ini disebabkan oleh kemampuan *TreeExplainer* dalam memperhitungkan interaksi antar fitur secara lebih mendalam dan efisien, yang pada akhirnya memberikan hasil yang lebih akurat. Temuan ini membuktikan bahwa sistem yang dikembangkan sesuai dengan teori SHAP, yang mendasari metode interpretasi ini, dan meningkatkan kredibilitas ilmiah dari *model* yang digunakan.

Selanjutnya, sistem pakar berbasis web berhasil dirancang dengan arsitektur yang mengintegrasikan *model* prediksi dan mekanisme interpretasi hasil secara interaktif. Integrasi ini memungkinkan sistem tidak hanya menghasilkan nilai probabilitas risiko diabetes, tetapi juga menjelaskan faktor-faktor yang berperan dalam meningkatkan atau menurunkan risiko, sehingga meningkatkan interpretasi pengguna terhadap kondisi kesehatannya.

Hasil pengujian menunjukkan bahwa sistem mampu memproses input pengguna dengan baik, menghasilkan prediksi yang konsisten, serta menyajikan interpretasi faktor risiko yang informatif dan mudah dipahami. Evaluasi terhadap metode SHAP memperlihatkan bahwa kontribusi fitur yang ditampilkan selaras dengan indikator medis yang dikenal dalam literatur diabetes, sehingga mendukung aspek akurasi dan kejelasan interpretasi.

Akhirnya, implementasi sistem pakar berbasis web yang terintegrasi dengan *model* XGBoost dan visualisasi SHAP telah berhasil direalisasikan dan dapat diakses melalui peramban web. Sistem ini memungkinkan masyarakat melakukan prediksi risiko diabetes secara mandiri sekaligus memperoleh penjelasan faktor risiko yang memengaruhinya. Dengan demikian, penelitian ini memberikan kontribusi dalam pemanfaatan kecerdasan buatan yang transparan dan mudah dipahami untuk mendukung deteksi dini risiko diabetes dan meningkatkan kesadaran kesehatan masyarakat.

5.2. Saran

Berdasarkan batasan masalah yang telah ditetapkan serta kekurangan dan temuan yang teridentifikasi dalam penelitian ini, terdapat beberapa rekomendasi spesifik yang dapat dipertimbangkan untuk pengembangan dan penelitian selanjutnya agar sistem prediksi risiko diabetes menjadi lebih komprehensif, akurat, dan memiliki nilai implementasi yang lebih luas.

1. Dataset yang digunakan dalam penelitian ini terbatas pada 520 sampel dari satu sumber (Kaggle) yang mungkin tidak merepresentasikan keragaman populasi Indonesia atau populasi global secara memadai. Penelitian

selanjutnya disarankan untuk melakukan kolaborasi dengan institusi kesehatan lokal rumah sakit, klinik, atau puskesmas untuk mengumpulkan data klinis riil pasien Indonesia. Dataset yang lebih besar yaitu 5,000-10,000 sampel dengan keragaman demografis yang lebih luas (berbagai kelompok usia, etnis, wilayah geografis, dan status sosio-ekonomi) akan meningkatkan kemampuan generalisasi *model*.

2. Sistem saat ini hanya menggunakan 10 fitur berbasis gejala klinis dan demografis tanpa mencakup data laboratorium atau faktor risiko medis lainnya yang penting. Penelitian lanjutan disarankan untuk mengintegrasikan variabel medis tambahan yang lebih komprehensif, mencakup: (1) Data laboratorium seperti kadar gula darah puasa, HbA1c, profil lipid (kolesterol, trigliserida), dan fungsi ginjal; (2) Riwayat medis termasuk riwayat keluarga diabetes, hipertensi, penyakit kardiovaskular, dan kondisi komorbid lainnya; (3) Faktor gaya hidup seperti pola diet (konsumsi gula, karbohidrat, serat), tingkat aktivitas fisik, kebiasaan merokok, dan konsumsi alkohol; (4) Data antropometrik seperti BMI, lingkar pinggang, rasio pinggang-pinggul, dan tekanan darah. Penambahan fitur-fitur ini berpotensi meningkatkan akurasi prediksi secara signifikan dan memungkinkan analisis risk stratification yang lebih granular.
3. Penelitian ini hanya menggunakan satu algoritma (XGBoost) tanpa perbandingan sistematis dengan metode lain yang juga populer untuk klasifikasi medis. Penelitian selanjutnya disarankan untuk melakukan studi komparatif yang rigorous antara berbagai algoritma *Machine Learning*, termasuk: (1) *Ensemble* methods lainnya seperti *Random Forest*,

AdaBoost, LightGBM, atau CatBoost; (2) Deep learning approaches seperti Feedforward *Neural networks* atau Recurrent *Neural networks* jika menggunakan data longitudinal; (3) Metode statistik tradisional seperti *Logistic regression* atau Naive Bayes sebagai baseline; (4) Support Vector Machine dengan berbagai kernel functions yang diharapkan dengan melakukan komparasi dapat mmeberikan hasil algorimat yang paling optimal dalam konteks spesifik prediksi resiko diabetes atau bahkan mengidentifikasi peluang untuk *ensemble model* yang dapat memberikan performa yang baik.

4. *Hyperparameter* tuning dalam penelitian ini dilakukan secara manual berdasarkan best practices literature, bukan melalui pencarian sistematis yang komprehensif. Penelitian lanjutan disarankan untuk mengimplementasikan teknik automated *hyperparameter* optimization seperti: (1) Grid Search untuk pencarian exhaustive dalam parameter space yang terdefinisi; (2) Random Search yang lebih efisien untuk high-dimensional parameter spaces; (3) Bayesian Optimization menggunakan tools seperti Optuna atau Hyperopt yang dapat menemukan konfigurasi optimal dengan lebih sedikit evaluasi; (4) Genetic Algorithms atau Particle Swarm Optimization sebagai alternatif metaheuristic.
5. Sistem saat ini hanya menggunakan SHAP untuk interpretability tanpa validasi konsistensi dengan metode explainability lainnya. Rekomendasi Penelitian selanjutnya dapat mengeksplorasi dan membandingkan berbagai metode Explainable AI seperti: (1) LIME (Local Interpretable *Model*-agnostic Explanations) untuk penjelasan lokal dengan pendekatan yang

berbeda dari SHAP; (2) Permutation Feature Importance untuk global pentingnya fitur yang lebih sederhana dan intuitif; (3) Grafik Ketergantungan Parsial dan Grafik Harapan Kondisional Individu untuk memvisualisasikan hubungan antara fitur dan prediksi; (4) Penjelasan Alternatif untuk memberikan wawasan yang dapat ditindaklanjuti Perbandingan antar metode akan memberikan wawasan tentang konsistensi interpretasi dan dapat mengidentifikasi fitur-fitur yang secara *robust* penting melalui berbagai metode yang dapat dijelaskan.

6. Sistem saat ini masih berupa prototipe penelitian tanpa fitur-fitur yang diperlukan untuk deployment produksi di lingkungan kesehatan. Transformasi dari prototipe ke sistem produksi memerlukan: (1) Security hardening termasuk implementasi HTTPS, encryption untuk data sensitif (at-rest dan in-transit), SQL injection prevention, dan rate limiting untuk mencegah abuse; (2) User authentication dan authorization dengan role-based access control misalnya: pasien, dokter, admin dan audit logging untuk kepatuhan; (3) integrasi database untuk penyimpanan permanen dari *Predict*ion history, user profiles, dan jejak audit.
7. Teakhir, Sistem saat ini hanya memberikan risk level dan risk factors tanpa rekomendasi aksi yang spesifik dan personalized. Rekomenasi Penelitian selanjutnya dapat mengembangkan sistem rekomendasi yang lebih dapat ditindaklanjuti dengan (1) Jalur perawatan yang disesuaikan dengan tingkat risiko seprti untuk risiko rendah edukasi mengenai tindakan pencegahan dan perubahan gaya hidup, untuk risiko sedang: penjadwalan tindak lanjut dan monitoring, untuk risiko tinggi: konsultasi

medis darurat dan lab testing; (2) Saran intervensi yang disesuaikan secara pribadi berdasarkan SHAP values misalnya, jika *Polyuria* adalah faktor resiko teratas, berikan edukasi tentang normal vs abnormal frekuensi buang air kecil dan rekomendasi lainnya.

DAFTAR PUSTAKA

Abdurrahman, G., Oktavianto, H., & Sintawati, M. (2022). Optimasi Algoritma Xgboost Classifier Menggunakan *Hyperparameter* Gridesearch Dan Random Search Pada Klasifikasi Penyakit Diabetes. In *Informatics Journal* (Vol. 7, Issue 3).

Adhi Pratama, N., & Wahyu Utomo, D. (2025). Deteksi Diabetes Mellitus Dengan Menggunakan Teknik *Ensemble* Xgboost Dan Lightgbm. In *Jurnal Informatika Sunan Kalijaga): Vol. Xx*.

Dengan Menggunakan Metode Dempster Shafer Siska Romauli Tamba, C., Nofriansyah, D., & Rizky, F. (2021). Sistem Pakar Mendiagnosa Penyakit. *Jurnal Cybertech*, *X. No.X*. Https://Ojs.Trigunadharma.Ac.Id/

Hossain, Md. J., Al-Mamun, Md., & Islam, Md. R. (2024). Diabetes Mellitus, The Fastest Growing Global Public Health Concern: Early Detection Should Be Focused. *Health Science Reports*, *7*(3), E2004. Https://Doi.Org/Https://Doi.Org/10.1002/Hsr2.2004

Kurniawan, H., Dwi Akbar, A., Svensons, N., Jaya Antonio, Y., Karnila, S., & Safitri, E. (2025). *Evaluasi Performa Random Forest ... 835 Z.A Pagar Alam No* (Issue 93).

Larassati, D., Zaidiah, A., Afrizal, S., Studi S-, P., Informasi, S., Ilmu Komputer Universitas Pembangunan Nasional Veteran Jakarta Jl Fatmawati Raya, F. R., Labu, P., & Cilandak, K. (2022). *Sistem Prediksi Penyakit Jantung Koroner Menggunakan Metode Naïve Bayes*. Www.Kaggle.Com

Li Wenguang And Peng, Y. A. N. D. P. K. (2024). Diabetes *Predict*ion *Model* Based On Ga-Xgboost And Stacking *Ensemble* Algorithm. *Plos One*, *19*(9), 1–29. Https://Doi.Org/10.1371/Journal.Pone.0311222

Muhammad Surono, Muhammad Fadli, Dian Sri Purwanti, & Erliyan Redy Susanto. (2025). Hybrid Xgboost-Svm *Model* Untuk Sistem Pendukung Keputusan Dalam Prediksi Penyakit Diabetes. *Insologi: Jurnal Sains Dan Teknologi*, *4*(3), 443–454. Https://Doi.Org/10.55123/Insologi.V4i3.5410

Nurul Puteri, A., & Sumardin, A. (2024). Comparison Of *Random Forest* And Xgboost For Diabetes Classification With Shap And Lime Interpretation. *Jurnal Teknologi Rekayasa)*, *9*(2). Https://Doi.Org/10.31544/Jtera.V9.I1.2024.121-130

Ponce-Bobadilla, A. V., Schmitt, V., Maier, C. S., Mensing, S., & Stodtmann, S. (2024). Practical Guide To Shap Analysis: Explaining Supervised *Machine Learning Model Predict*ions In Drug Development. *Clinical And Translational Science*, *17*(11), E70056. Https://Doi.Org/Https://Doi.Org/10.1111/Cts.70056

R Abdurrosyid, Teguh, A., & Almais, W. (2025). *Jepin (Jurnal Edukasi Dan Penelitian Informatika) Deteksi Dini Diabetes Menggunakan Machine Learning Dengan Metode Pca Dan Xgboost*.

Rizky, F., Syahra, Y., & Mariami, I. (2019). *Implementasi Data Mining Untuk Memprediksi Target Pemakaian Stok Barang Menggunakan Metode Regresi Linier Berganda*. *18*(Saintikom), 167–175.

Rizky, F., Syahra, Y., Winda Nazry, H. S., Halim, J., Niawati Lo, N., Muhammadiyah Sumatera Utara, U., & Trigurna Dharma, S. (2025). Implementation Of *Data Mining* To Categorize Daily Sales Transaction Data Using K-Means Clustering. *Proceeding International Seminar On Islamic Studies*, *6*(1).

Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K., & Menegaz, G. (2025). A Perspective On Explainable Artificial Intelligence Methods: Shap And Lime. *Advanced Intelligent Systems*, *7*(1), 2400304. Https://Doi.Org/Https://Doi.Org/10.1002/Aisy.202400304

Saputra, D. C. E., Sunat, K., Ratnaningsih, T., Ma'arif, A., & Suwarno, I. (2025). Smartforest: Optimizing Anemia Diagnosis With Enhanced *Random Forest* And Xgboost Integration. *Results In Engineering*, *27*. Https://Doi.Org/10.1016/J.Rineng.2025.106442

Sun, H., Saeedi, P., Karuranga, S., Pinkepank, M., Ogurtsova, K., Duncan, B. B., Stein, C., Basit, A., Chan, J. C. N., Mbanya, J. C., Pavkov, M. E., Ramachandaran, A., Wild, S. H., James, S., Herman, W. H., Zhang, P., Bommer, C., Kuo, S., Boyko, E. J., & Magliano, D. J. (2022). Idf Diabetes Atlas: Global, Regional And Country-Level Diabetes Prevalence Estimates For 2021 And Projections For 2045. *Diabetes Research And Clinical Practice*, *183*. Https://Doi.Org/10.1016/J.Diabres.2021.109119

Susanto, E. R., & Cahyana, A. (2025). Penerapan Algoritma Xgboost Untuk Prediksi Diabetes: Analisis Confusion Matrix Dan Roc Curve. *Fountain Of Informatics Journal*, *10*(1), 40–50. Https://Doi.Org/10.21111/Fij.V10i1.14311

Upadhaya, A., Pegu, J., Singh, Y. D., & Ngomle, S. (2025). Nanoparticle-Enabled Portable Biosensors For Early Detection And Monitoring Of Non-Communicable Diseases: A Focus On Diabetes, Cardiovascular, And Cancer Diagnostics. In *Biosensors And Bioelectronics: X* (Vol. 26). Elsevier Ltd. Https://Doi.Org/10.1016/J.Biosx.2025.100675

Utomo, S., Sigit, S., Sulistyowati, N., Arman Prasetya, D., & Maulana Fahrudin, T. (2025). *Perbandingan Kinerja Algoritma Xgboost Dan Catboost Dalam Klasifikasi Risiko Penyakit Diabetes* (Vol. 6). Www.Elektro.Itn.Ac.Idjournalhomepage:Www.Elektro.Itn.Ac.Id

Wardhani, K. D. K., & Akbar, M. (2022). Diabetes Risk *Predict*ion Using Extreme *Gradient boosting* (Xgboost). *Jurnal Online Informatika*, *7*(2), 244–250. Https://Doi.Org/10.15575/Join.V7i2.970

Widyanti, A., Nofriansyah, D., & **Rizky, F**. (2020). Sistem Pakar Dalam Menganalisa Penyakit Sinusitis Menggunakan Teorema Bayes. *Jurnal Cybertech*, Pp. Www.Trigunadharma.Ac.Id.

LAMPIRAN

MAJELIS PENDIDIKAN TINGGI PENELITIAN & PENGEMBANGAN PIMPINAN PUSAT MUHAMMADIYAH
UNIVERSITAS MUHAMMADIYAH SUMATERA UTARA
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI
UMSU Terakreditasi Unggul Berdasarkan Keputusan Badan Akreditasi Nasional Perguruan Tinggi No. 174/SK/BAN-PT/Akp-PT/III/2024
Pusat Administrasi: Jalan Mukhtar Basri No. 3 Medan 20238 Telp. (061) 6622400 - 66224567 Fax. (061) 6625474 - 6631003
fikti@umsu.ac.id umsumedan umsumedan umsumedan umsumedan

KEPUTUSAN DEKAN
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI
UNIVERSITAS MUHAMMADIYAH SUMATERA UTARA
Nomor : 1045/KEP/II.3.AU/UMSU-09/F/2025

Tentang :
PENGANGKATAN DOSEN PEMBIMBING SKRIPSI
PRODI SISTEM INFORMASI
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI
UNIVERSITAS MUHAMMADIYAH SUMATERA UTARA

بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ

Dekan Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Muhammadiyah Sumatera Utara, setelah:

Menimbang :
1. Bahwa sehubungan dengan pelaksanaan Tugas Akhir Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Muhammadiyah Sumatera Utara;
2. Bahwa untuk memenuhi maksud tersebut perlu diterbitkan surat keputusan Dekan Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Muhammadiyah Sumatera Utara.

Mengingat :
1. Undang-undang Republik Indonesia No. 12 Tahun 2012 tentang Pendidikan Tinggi;
2. Peraturan Pemerintah Nomor 4 Tahun 2014 tentang Penyelenggaraan Pendidikan Tinggi dan Pengelolaan Perguruan Tinggi;
3. Pedoman Perguruan Tinggi Muhammadiyah;
4. Statuta Universitas Muhammadiyah Sumatera Utara
5. Keputusan Rektor No. 1001/KEP/II.3-AU/UMSU/D/2025 tanggal 25 Oktober 2025 tentang Dekan Fakultas Ilmu Komputer dan Teknologi Informasi UMSU Masa Jabatan 2025-2029
6. Keputusan Rektor No. 624/KEP/II.3.AU/UMSU/D/2025 tentang Prodi Fakultas Ilmu Komputer dan Teknologi Informasi UMSU Masa Jabatan 2025-2029

MEMUTUSKAN

Menetapkan :

KESATU : Menetapkan nama-nama dalam lampiran surat keputusan ini sebagai Dosen Pembimbing Prodi Sistem Informasi Fakultas Ilmu Komputer dan Teknologi Informasi UMSU.

KEDUA : Keputusan ini berlaku sejak tanggal ditetapkan, dan akan ditinjau serta diperbaiki kembali jika terdapat kekeliruan di kemudian hari.

Ditetapkan di : Medan
Pada tanggal : 28 Jumadil Awwal 1447 H
19 November 2025 M

Dekan

Dr. Al-Khowarizmi., M.Kom
NIDN : 0127099201

Tembusan:
1. Yth. Wakil Dekan I & III UMSU di Medan,-
2. Ka. Prodi SI dan Sek. Prodi SI,-
3. Pertinggal.

MAJELIS PENDIDIKAN TINGGI PENELITIAN & PENGEMBANGAN PIMPINAN PUSAT MUHAMMADIYAH
UNIVERSITAS MUHAMMADIYAH SUMATERA UTARA
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI
UMSU Terakreditasi Unggul Berdasarkan Keputusan Badan Akreditasi Nasional Perguruan Tinggi No. 174/SK/BAN-PT/Ak-Ppj/PT/III/2024
Pusat Administrasi: Jalan Mukhtar Basri No. 3 Medan 20238 Telp. (061) 6622400 - 66224567 Fax. (061) 6625474 - 6631003
umsumedan umsumedan umsumedan umsumedan

Lampiran Dosen Pembimbing Prodi Sistem Informasi
Nomor : 1045/KEP/II.3.AU/UMSU-09/F/2025
Tanggal : 28 Jumadil Awwal 1447 H /19 November 2025 M

PENGANGKATAN DOSEN PEMBIMBING SKRIPSI
PRODI SISTEM INFORMASI
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI
UNIVERSITAS MUHAMMADIYAH SUMATERA UTARA

NO	NAMA	NPM	JUDUL	DOSEN
1	Fahri Nasution	2209010262	Penerapan Metode Naïve Bayes untuk Menganalisis Sentimen Publik Terhadap Layanan Transportasi Online di Indonesia Berbasis Data Media Sosial X	Zuli Agustina Gultom, M.Si
2	Fadhila Arumtini	2209010052	Implementasi Sistem Prediksi bahan baku Produksi Harian Berbasis Algoritma Random Forest pada UMKM Ayam Geprek Chiktri	Fatma Sari Hutagalung,S.Kom., M.Kom
3	Diajeng Vesha Ramadhini	2209010046	Penerapan algoritma C 4.5 pada SMK Muhammadiyah 16 Desa Pon untuk membentuk kelompok olimpiade	Amrullah, S.Kom.,M.Kom
4	Nadia Pratiwi	2209010050	Analisis Forensik Digital terhadap Keaslian Foto Produk E-Commerce Shopee melalui Metadata EXIF dan Pengolahan Citra Brightness-Contrast	Mulkan Azhari, S.Kom.,M.Kom
5	Arya Danu Hartono	2209010178	Analisis klasifikasi dan simulasi Kebutuhan lokasi Parkir di Universitas Muhammadiyah Sumatera Utara Berdasarkan Pola Kunjungan dan Distribusi Ruang Menggunakan Metode K-Means Clustering	Mahardika Abdi Prawira, S.Kom.,M.Kom
6	Nadia Aprilia Siregar	2209010123	Sistem Pakar dalam analisis prediksi Risiko Diabetes Menggunakan Model XGBoost dan Analisis Faktor Risiko dengan Explainable AI (SHAP)	Dr. Firahmi Rizky, M.Kom
7	Devy Liani Herawati	2209010244	Analisis Prediksi Risiko Gagal Ginjal Berdasarkan Data Pemeriksaan Laboratorium Menggunakan Algoritma Demster Shafer di Rumah Sakit Sapta Medika Indrapura	Rizaldi Khair, S.Kom.,M.Kom

Ditetapkan

Dr. Al-Khowarizmi., M.Kom
NIDN : 0127099201

MAJELIS PENDIDIKAN TINGGI PENELITIAN & PENGEMBANGAN PIMPINAN PUSAT MUHAMMADIYAH
UNIVERSITAS MUHAMMADIYAH SUMATERA UTARA
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI
UMSU Terakreditasi Unggul Berdasarkan Keputusan Badan Akreditasi Nasional Perguruan Tinggi No. 174/SK/BAN-PT/Ak.Ppj/PT/III/2024
Pusat Administrasi: Jalan Mukhtar Basri No. 3 Medan 20238 Telp. (061) 6622400 - 66224567 Fax. (061) 6625474 - 6631003
fikti@umsu.ac.id umsumedan umsumedan umsumedan umsumedan

ACC JILID LUX

Nama Mahasiswa : Nadia Aprilia Siregar
NPM/ Prodi : 2209010123/ Sistem Informasi
Judul Skripsi : Sistem Pakar Dalam Analisis Prediksi Risiko Diabetes Menggunakan Model XGBOOST dan Analisis Faktor Risiko Dengan Explainable AI (SHAP)
Dosen Pembimbing : Dr. Firahmi Rizky, S.Kom., M.Kom.

NO	DOKUMEN YANG PERLU DIPERSIAPKAN	Paraf Prodi
1	Form Revisi Sidang	✓
2	Turnitin Max 30%	✓
3	Lembar pengesahan	✓
4	File Jurnal (Templete) / HKI, Bukti submit jurnal	✓
5	Lembar Bimbingan Min 8X	✓
6	Serah terima alat dan Program	✓
7	Serah terima sourch code/ Apk	✓
8	Kesesuaian template Skripsi	✓

Medan, 07 Mei 2026

Ka.Prodi/Sek.Prodi
Sistem Informasi

(Mahardika Abdi Prawira, S.Kom., M.Kom)

Note Untuk Mahasiswa

- Jurnal tidak harus terakreditasi SINTA
- File artikel Jurnal sudah disesuaikan dengan templete jurnal yang dituju

MAJELIS PENDIDIKAN TINGGI PENELITIAN & PENGEMBANGAN PIMPINAN PUSAT MUHAMMADIYAH
UNIVERSITAS MUHAMMADIYAH SUMATERA UTARA
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI
UMSU Terakreditasi A Berdasarkan Keputusan Badan Akreditasi Nasional Perguruan Tinggi No. 89/SK/BAN-PT/Akred/PT/III/2019
Pusat Administrasi: Jalan Mukhtar Basri No. 3 Medan 20238 Telp. (061) 6622400 - 66224567 Fax. (061) 6625474 - 6631003
https://fikti.umsu.ac.id fikti@umsu.ac.id umsumedan umsumedan umsumedan umsumedan

`Berita Acara Pembimbingan Skripsi

Nama Mahasiswa : Nadia Aprilia Siregar Program Studi : Sistem Informasi
NPM : 2309010123
Nama Dosen Pembimbing : Dr. Firahmi Rizky, M.Kom. Judul Penelitian : **Sistem Pakar dalam Analisis Prediksi Resiko Diabetes Menggunakan Model XGBoost dan Analisis Faktor Resiko Dengan Explainable AI (SHAP)**

Tanggal Bimbingan	Hasil Evaluasi	Paraf Dosen
02/Desember/2025	perbaikan bab 1, penyusunan latar belakang, sitasi dan rumusan masalah	
06/Desember/2025	bab 1 dan 2 perbaikan batasan masalah dan tujuan manfaat, serta landasan teori	
23/Desember/2025	Penambahan di bab 2 landasan teori yaitu data mining dan resiko diabetes	
06/Januari/2025	Penambahan di bab 2 tentang materi UML beserta simbolnya dan di bab 3	
15/Januari/2025	Acc Sempro	
03/Februari/2026	konsultasi mengenai bimbingan sehabis sempro	
10/Februari/2026	Bimbingan terkait bab 4 dan 5, perbaikan penomoran pada bab 4 dan bab 5	
12/Februari/2026	Bimbingan terkait revisi bab 4 dan bab 5 mengenai perhitungan manual dan kekurangan dan kelebihan	

26/Februari/2026	Revisi terkait saran yang terdapat pada bab 5	
04/Maret/2026	ACC Sidang	

Medan, 04 Maret 2026

Diketahui oleh :

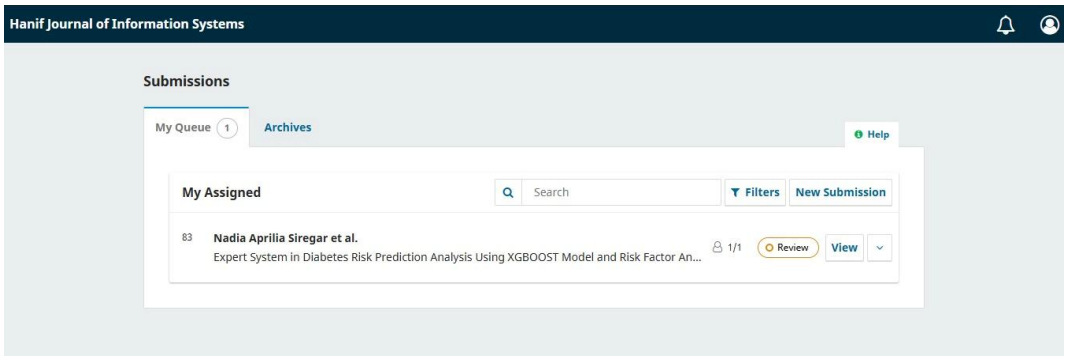
Ketua Program Studi
Sistem Informasi

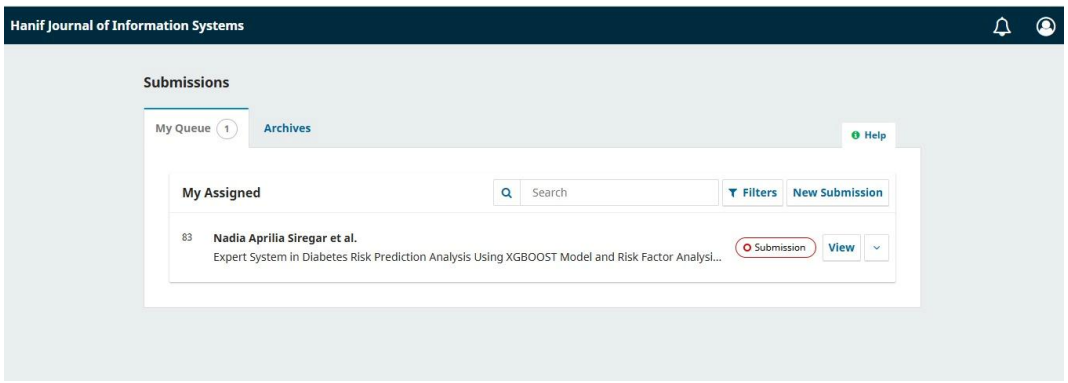
(Mahardika Abdi Prawira, M.Kom.)

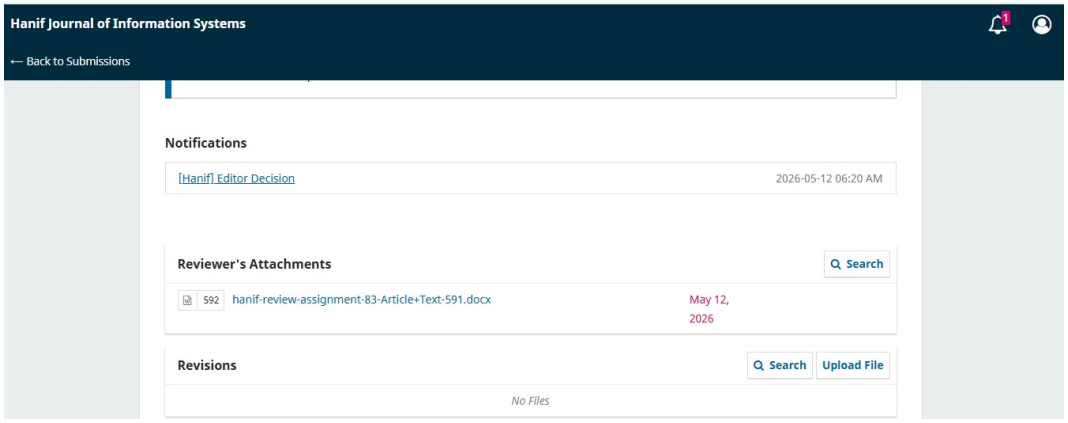
Disetujui oleh :

Dosen Pembimbing

(Dr. Firahmi Rizky, M.Kom.)


Hanif Journal of Information Systems
Submissions
My Queue 1
Archives
Help
My Assigned
Search
Filters
New Submission
83
Nadia Aprilia Siregar et al.
Expert System in Diabetes Risk Prediction Analysis Using XGBOOST Model and Risk Factor An...
1/1
Review
View


Hanif Journal of Information Systems
Submissions
My Queue 1
Archives
Help
My Assigned
Search
Filters
New Submission
83
Nadia Aprilia Siregar et al.
Expert System in Diabetes Risk Prediction Analysis Using XGBOOST Model and Risk Factor Analysi...
Submission
View


Hanif Journal of Information Systems
← Back to Submissions
Notifications
[Hanif] Editor Decision
2026-05-12 06:20 AM
Reviewer's Attachments
Search
592
hanif-review-assignment-83-Article+Text-591.docx
May 12, 2026
Revisions
Search
Upload File
No Files

HANIF
Journal of Information Systems
https://journal.ilmubersama.com/index.php/hanif

LETTER OF ACCEPTANCE (LoA)

Dear Mr/Ms/Br

Nadia Aprilia Siregar, Firahmi Rizky

In

Place

We hereby inform you that the manuscript with the following details has been declared accepted for publication in the Tsabit Journal of Computer Science, ISSN: 3062-8504 (Online), Volume 1 Number 1 August 2026 Edition.

Title	Expert System in Diabetes Risk Prediction Analysis Using XGBOOST Model and Risk Factor Analysis With Explainable AI (SHAP)
Author	Nadia Aprilia Siregar, Firahmi Rizky
Correspondent Email	firahmirizky@umsu.ac.id

Thus we have created this certificate so that it can be used as necessary.

Medan, 05 Mei 2026

Editor in Chief

Oris Krianto Sulaiman,S.T.,M.Kom.

Hanif Journal of Information Systems
Email: ibchanifjournal@gmail.com
Secretariat: Batang Kuis - Lubuk Pakam Gg. Cempaka Dusun III No. 3, Tanjung Sari, Batang Kuis, District. Deli Serdang Sumatera Utara
This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

skripsi-nadia-aprilia_1777275299210.docx

ORIGINALITY REPORT

13%
SIMILARITY INDEX

PRIMARY SOURCES

1	repository.umsu.ac.id Internet	661 words — 3%
2	fikti.umsu.ac.id Internet	120 words — 1%
3	ejurnal.stmik-budidarma.ac.id Internet	77 words — < 1%
4	jtera.polteksmi.ac.id Internet	57 words — < 1%
5	Luthfia Nurma Hapsari, Ilham Fannani, Yenny Rahmawati, Ahmad Muhariya. "Explainable Machine Learning untuk Prediksi Risiko Penyakit Jantung Menggunakan Random Forest dan Analisis SHAP", remik, 2026 Crossref	54 words — < 1%
6	eprints.upj.ac.id Internet	50 words — < 1%
7	alaskapolitics.com Internet	43 words — < 1%
8	repositori.telkomuniversity.ac.id Internet	42 words — < 1%
9	text-id.123dok.com Internet	

		42 words — < 1%
10	elib.pnc.ac.id Internet	40 words — < 1%
11	journal.pancabudi.ac.id Internet	40 words — < 1%
12	jurnal.pustakagalerimandiri.co.id Internet	38 words — < 1%
13	123dok.com Internet	36 words — < 1%
14	jurnal.itscience.org Internet	36 words — < 1%
15	jurnal.polibatam.ac.id Internet	35 words — < 1%
16	repositori.usu.ac.id Internet	35 words — < 1%
17	ejournal.unibba.ac.id Internet	32 words — < 1%
18	repository.unissula.ac.id Internet	29 words — < 1%
19	journal.arteii.or.id Internet	28 words — < 1%
20	Sri Utami. "Desain Sistem Informasi Penyewaan Lapangan Futsal Berbasis Website : Perspektif Pemodelan UML", JISCO : Journal of Information System and Computing, 2025	27 words — < 1%

Crossref

21	eprints.walisongo.ac.id Internet	27 words — < 1%
22	kc.umn.ac.id Internet	26 words — < 1%
23	Setiawan Ardi wijaya, Aliyah khalista putri, Muhammad Shafwannullah Abdul Aziz, Muhammad Wahyu Illahi et al. "PERANCANG SISTEM INFORMASI PT NUSANTARAN ALUMUNIUM UNTUK SISTEM PEMESANAN DENGAN METODE RAPID APPLICATION DEVELOPMENT (RAD)", PROSISKO: Jurnal Pengembangan Riset dan Observasi Sistem Komputer, 2026 Crossref	25 words — < 1%
24	assyifa.forindpress.com Internet	25 words — < 1%
25	eprints.upnjatim.ac.id Internet	24 words — < 1%
26	uisi-repository.sgp1.cdn.digitaloceanspaces.com Internet	24 words — < 1%
27	garuda.kemdiktisaintek.go.id Internet	23 words — < 1%
28	lppm.usm.ac.id Internet	23 words — < 1%
29	Arisman Arisman, Nurhayati Nurhayati, Tanti Tanti. "Analisis Klasifikasi Tingkat Hidrasi Manusia Menggunakan Algoritma Machine Learning Berbasis Dataset Asupan Air Harian", Jurnal Minfo Polgan, 2026 Crossref	22 words — < 1%

30	lisaervina.wordpress.com Internet	22 words — < 1%
31	repository.poltekkes-denpasar.ac.id Internet	21 words — < 1%
32	jsisfotek.org Internet	20 words — < 1%
33	Safia Mayana, Elin Hidayat, Maharani Farah Dhifa Dg Masikki. "Hubungan Caring Perawat Berbasis Carative Terhadap Kepatuhan Pasien dalam Pencegahan Diabetic Foot Ulcer pada Pasien dengan Diabetes Mellitus di RSUD Undata Provinsi Sulawesi Tengah", Jurnal Pengabdian Masyarakat dan Riset Pendidikan, 2026 Crossref	19 words — < 1%
34	e-jurnal.pnl.ac.id Internet	18 words — < 1%
35	journal.ilmudata.co.id Internet	18 words — < 1%
36	www.coursehero.com Internet	18 words — < 1%
37	www.scribd.com Internet	18 words — < 1%
38	Muhammad Febriansyah, Ingrid Nurtanio, Mukarramah Yusuf. "Sistem Pakar Berbasis Pohon Keputusan untuk Diagnosa Awal Penyakit Bergejala Demam", Jurnal Minfo Polgan, 2024 Crossref	17 words — < 1%
39	etheses.uin-malang.ac.id Internet	17 words — < 1%

40	www.bintanginspirasi.com Internet	17 words — < 1%
41	www.kangatepafia.com Internet	17 words — < 1%
42	Ayu Dwi Ratna Ningsih, Ade Eviyanti. "Expert System To Identify Insomnia Problems In Adults Using Forward Chaining Method", Procedia of Engineering and Life Science, 2021 Crossref	16 words — < 1%
43	smart.stmikplk.ac.id Internet	16 words — < 1%
44	Erly Ekayanti Rosyida, Nisa Isrofi, Regita Permata Puteri, Lathifa Puteri Asy'ari et al. "Pengembangan Dashboard Prediksi Permintaan Berbasis SARIMAX-XGBoost untuk Mendukung Pengambilan Keputusan", RIGGS: Journal of Artificial Intelligence and Digital Business, 2026 Crossref	15 words — < 1%
45	jurnal.pnj.ac.id Internet	15 words — < 1%
46	e-jurnal.lppmunsera.org Internet	14 words — < 1%
47	repository.bku.ac.id Internet	14 words — < 1%
48	Lauren Patricia Sagala, Romualda Sitanggang, Sardo Pardingotan Sipayung. "Prediksi Risiko Penyakit Diabetes Menggunakan Algoritma K-Nearest Neighbor", Jurnal Pendidikan Tambusai, 2026 Crossref	13 words — < 1%

49	Zidan Indra Nugraha, Arnita, Kana Saputra S, Abi Setiawan, Raysa Maharani, Firna Zaharani. "IMPLEMENTASI ALGORITMA CNN DALAM PENGEMBANGAN WEBSITE UNTUK KLASIFIKASI SAMPAH ORGANIK, DAN NON-ORGANIK", Jurnal Manajemen Informatika dan Sistem Informasi, 2025 Crossref	13 words — < 1%
50	ejournal.uigm.ac.id Internet	13 words — < 1%
51	ojs.stikombanyuwangi.ac.id Internet	13 words — < 1%
52	voxntt.com Internet	13 words — < 1%
53	www.alodokter.com Internet	13 words — < 1%
54	Fadhlan Ihsan Lubis, Muhammad Iqbal. "Analisis Waktu Optimal Unggah Instagram Menggunakan K-Means Clustering pada Data Sintetis", RIGGS: Journal of Artificial Intelligence and Digital Business, 2026 Crossref	12 words — < 1%
55	digilib.uinsgd.ac.id Internet	12 words — < 1%
56	docplayer.info Internet	12 words — < 1%
57	dspace.lib.cranfield.ac.uk Internet	11 words — < 1%
58	ejournal.pnc.ac.id Internet	

		11 words — < 1%
59	iaytjournals.org Internet	11 words — < 1%
60	journal.universitaspahlawan.ac.id Internet	11 words — < 1%
61	repo.darmajaya.ac.id Internet	11 words — < 1%
62	repository.ukdw.ac.id Internet	11 words — < 1%
63	Apit Sutaryani, Sunarno Sunarno, Djuniadi Djuniadi. "PERBANDINGAN PERFORMA MODEL MACHINE LEARNING DALAM PREDIKSI SUHU DI SEMARANG", Jurnal Informatika dan Teknik Elektro Terapan, 2024 Crossref	10 words — < 1%
64	id.scribd.com Internet	10 words — < 1%
65	ijins.umsida.ac.id Internet	10 words — < 1%
66	jpti.journals.id Internet	10 words — < 1%
67	repository.its.ac.id Internet	10 words — < 1%
68	Aditya Gilang Pangestu, Sri Winarno, Adhitya Nugraha, Almas Najiib Imam Muttaqin. "DiabTrack: Sistem Prediksi Dini Diabetes Melitus Tipe 2 berbasis Web	9 words — < 1%

menggunakan Algoritma K-Nearest Neighbors", Edumatic: Jurnal Pendidikan Informatika, 2025
Crossref

69 Dea Amelia Putri, Hari Murti, Sulastri Sulastri. "PERANCANGAN BANGUN SISTEM INFORMASI JASA KECANTIKAN BERBASIS WEB PADA ANNISAVY", INTECOMS: Journal of Information Technology and Computer Science, 2026 — 9 words — < 1%
Crossref

70 Maulida, Fanda Kumariana. "Peran Corporate Governance Dalam Financialdistress Pada Saham Syariah Yang Go Public Dibursa Efek Indonesia", Universitas Islam Sultan Agung (Indonesia), 2022 — 9 words — < 1%
ProQuest

71 Nora Listiana. "Implementasi Sistem Pakar Forward Chaining Diagnosa Penyakit Infeksi Bakteri", remik, 2025 — 9 words — < 1%
Crossref

72 Toni Arifin. "IMPLEMENTASI METODE K-NEAREST NEIGHBOR UNTUK KLASIFIKASI CITRA SEL PAP SMEAR MENGGUNAKAN ANALISIS TEKSTUR NUKLEUS", Jurnal Informatika, 2016 — 9 words — < 1%
Crossref

73 Zakha Maisat Eka Darmawan, Ashafidz Fauzan Dianta. "Implementasi Optimasi Hyperparameter GridSearchCV Pada Sistem Prediksi Serangan Jantung Menggunakan SVM", Teknologi, 2023 — 9 words — < 1%
Crossref

74 docobook.com — 9 words — < 1%
Internet

75 doku.pub — 9 words — < 1%
Internet

76	ejournal.uniramalang.ac.id Internet	9 words — < 1%
77	jurnal.itg.ac.id Internet	9 words — < 1%
78	library.binus.ac.id Internet	9 words — < 1%
79	pt.scribd.com Internet	9 words — < 1%
80	repositori.ubs-ppni.ac.id Internet	9 words — < 1%
81	repository.upi.edu Internet	9 words — < 1%
82	wartapenilai.id Internet	9 words — < 1%
83	www.frontiersin.org Internet	9 words — < 1%
84	www.rribandarlampung.co.id Internet	9 words — < 1%
85	Dede Supiyan. "Pengembangan Sistem Pakar Untuk Diagnosa Penyakit Diabetes Melitus Menggunakan Metode Forward Chaining", bit-Tech, 2025 Crossref	8 words — < 1%
86	Kinanti Maharani, Astriani Natalia Br. Ginting, Roy Indrianto Banggar. "Formulasi dan Uji Efektivitas Sediaan Gel Ekstrak Buah Bit (Beta vulgaris L) Terhadap Penyembuhan Luka Tikus Model Diabetes Yang Diinduksi Aloksan", Journal of Pharmaceutical and Sciences, 2026	8 words — < 1%

Crossref

87	Muhammad Yusuf, Indah Purnama Sari, Virda Kristy. "Sistem Pakar Mencegah Stunting dengan Menentukan Gizi Anak Menggunakan Natural Language Processing (NLP)", Jurnal JTIK (Jurnal Teknologi Informasi dan Komunikasi), 2025 Crossref	8 words — < 1%
88	Vidia Honesta. "WEBSITE MONITORING KEHADIRAN PEGAWAI ROTI ENAK MENGGUNAKAN SENSOR FINGERPRINT BERBASIS IOT", PROSISKO: Jurnal Pengembangan Riset dan Observasi Sistem Komputer, 2025 Crossref	8 words — < 1%
89	ejournal.unis.ac.id Internet	8 words — < 1%
90	ejurnal.seminar-id.com Internet	8 words — < 1%
91	es.scribd.com Internet	8 words — < 1%
92	journal.ipm2kpe.or.id Internet	8 words — < 1%
93	journal.unnes.ac.id Internet	8 words — < 1%
94	jurnal.pancabudi.ac.id Internet	8 words — < 1%
95	medium.com Internet	8 words — < 1%
96	obattradisionalhipertensi.net Internet	8 words — < 1%

97	pdffox.com Internet	8 words — < 1%
98	repos.dianhusada.ac.id Internet	8 words — < 1%
99	repositori.unimma.ac.id Internet	8 words — < 1%
100	repository.ubharajaya.ac.id Internet	8 words — < 1%
101	repository.usahidsolo.ac.id Internet	8 words — < 1%
102	Faren Tresandra Nafasya, Kirana Isna Dewi, Ima Muhimmah Falasifa. "Perancangan Perangkat Lunak UMKM Batik Sendang Lamongan", Jurnal Dinamika Informatika, 2026 Crossref	7 words — < 1%
103	Ihya Bahrul Alam, Hasbi Firmansyah, Wahyu Asriyani. "Prediksi Klasifikasi Diabetes Berbasis Faktor Risiko Kesehatan Menggunakan Gradient Boosted Trees", Jurnal Dinamika Informatika, 2026 Crossref	7 words — < 1%
104	Syalaiza Nizar Ruscikasani, Raden Roro Najwa Oktalivia, Firman Restu Putra, Ahmad Jurnaidi Wahidin et al. "Prediksi Pembelian E-Commerce Menggunakan XGBoost Berbasis Perilaku Sesi Pengguna", RIGGS: Journal of Artificial Intelligence and Digital Business, 2025 Crossref	7 words — < 1%
105	Tri Hartono, Muhammad Makmun Effendi, Nanang Tedi Kurniadi. "Pemilihan Guru Terbaik Pada MTs	7 words — < 1%

Al-Amanah dengan Metode Analytical Hierarchy Process (AHP)", remik, 2025
Crossref

106 doaj.org
Internet
7 words — < 1%

107 Erfian Junianto, Mila Puspitasari, Salman Ilyas Zakaria, Toni Arifin, Ignatius Wiseto Prasetyo Agung. "Emotion Detection in Indonesian Text Using the Logistic Regression Method", Media Jurnal Informatika, 2025
Crossref
6 words — < 1%

108 Muhammad Ghufron Syifa, Nadia Anissa Maori, Adi Sucipto. "Penerapan SMOTE dan XGBoost untuk Klasifikasi Penyakit Ginjal Kronis pada Data yang Tidak Seimbang", Jurnal JTIK (Jurnal Teknologi Informasi dan Komunikasi), 2026
Crossref
6 words — < 1%

109 Sandra Pebrianti, Gusgus Ghraha Ramdhanie, Bambang Aditya Nugraha. "Implementasi Deep Breathing Exercise terhadap Tingkat Glukosa Darah pada Pasien Diabetes Mellitus", Journal of Telenursing (JOTING), 2023
Crossref
6 words — < 1%

110 Zian Asti Dwiyanti, Cahyo Prianto. "Prediksi Cuaca Kota Jakarta Menggunakan Metode Random Forest", Jurnal Tekno Insentif, 2023
Crossref
6 words — < 1%

EXCLUDE QUOTES ON
EXCLUDE BIBLIOGRAPHY ON
EXCLUDE SOURCES OFF
EXCLUDE MATCHES OFF