KLASIFIKASI PENYAKIT DIABETES MENGGUNAKAN ALGORITMA DECISION TREE DAN K-NEAREST NEIGHBOUR (K-NN)

SKRIPSI

Diajukan Sebagai Syarat Untuk Menyelesaikan Pendidikan
Program Sarjana (S-1)
Pada Program Studi Sistem Informasi

Oleh:
Nama: Anggi Mika
Nim: 2102030042

FAKULTAS ILMU TEKNIK
PROGRAM STUDI SISTEM INFORMASI
UNIVERSITAS BINA INSAN KOTA LUBUKLINGGAU
2025

HALAMAN PENGESAHAN SKRIPSI

KLASIFIKASI PENYAKIT DIABETES MENGGUNAKAN ALGORITMA DECISION TREE DAN K-NEAREST NEIGHBOUR

Oleh:
Nama: Anggi Mika
NIM: 2102030042

Lubuklinggau, 2025

Pembimbing I	**Pembimbing II**
Lukman Hakim, M.Kom	**Satrianansyah, M.Kom**

Mengesahkan,
Dekan Fakultas Ilmu Teknik
Universitas Bina Insan

Dr. Rudi Kurniawan, ST.,M.Kom

HALAMAN PERSETUJUAN TIM PENGUJI SKRIPSI

Pada Hari.....tanggal.....bulan.....tahun 2025 telah dilaksanakan sidang Skripsi oleh Program Studi Sistem Informasi Fakultas Ilmu Teknik Universitas Bina Insan.

Nama : Anggi Mika
NIM : 2102030042
Judul Skripsi : Klasifikasi penyakit diabetes menggunakan Algoritma Decision Ttree Dan K-Nearest Neighbour (KNN)

Komisi Penguji

1. Ketua : Lukman Hakim, M.Kom (.....)
2. Sekretaris : Satrianansyah, M.Kom (.....)
3. Anggota : Harma Oktafia LW, M.Kom (.....)

Mengetahui,
Kepala Program Studi Sistem Informasi
Fakultas Ilmu Teknik
Universitas Bina Insan

Harma Oktafia Lingga Wijaya, M.Kom

HALAMAN MOTTO DAN PERSEMBAHAN

MOTTO :

- **Kelambatan bukanlah kelemahan, tapi cara unik orang bermalas-malasan mengejar target tepat waktu dengan kecerdasan.**
- **Ketika kata-kata tak lagi cukup, tindakanlah yang memberi makna.**
- **Dalam setiap langkah, temukan kebijaksanaan untuk terus berkembang.**
- **Keberhasilan terletak pada memberikan nilai positif dengan bijak dan tekun.**

Persembahan kepada :

- **Untuk Ibu dan Ayah,dua sosok yang tak pernah lelah memberikan cinta dan dorongan. Terima kasih atas kasih sayang, pengorbanan, dan petunjuk yang tak ternilai.**
- **Teman-teman, kalian adalah Sahabat sejati yang membuat perjalanan ini penuh warna. Segala jerih payah ini kupersembahkan untuk kalian, sebagai tanda terima kasih dan penghormatan yang tulus.**

HALAMAN PERNYATAAN

Saya yang bertanda tangan diba[illegible]:

Nama Mahasiswa : Anggi Mik[illegible]

NIM : 2102030042

Program Studi : Sistem Informasi

Menyatakan dengan sesungguhnya bahwa penelitian dan penulisan Skripsi yang saya susun sebagai persyaratan untuk memperoleh gelar Sarjana (S-1) Universitas Bina Insan, merupakan hasil kerja saya sendiri dan tidak menyuruh orang lain yang mengerjakannya. Ada bagian tertentu dalam penulisan skripsi ini yang saya kutip dari hasil karya orang lain dan telah saya tuliskan sumbernya secara jelas sesuai dengan norma, kaidah dan etika penulisan ilmiah.

Jika dikemudian hari ternyata terbukti bahwa penelitian dan tugas akhir ini bukan hasil kerja saya sendiri atau plagiat dalam bagian-bagian tertentu, maka saya bersedia dikenakan sanksi sesuai dengan peraturan perundungan yang berlaku.

Lubuklinggau, 2025

Penulis

Anggi Mika

NIM 2102030042

ABSTRACT

Diabetes is a chronic metabolic disease in which diabetic patients do not produce enough insulin or it can be said that the patient's body is unable to utilize insulin properly, causing blood sugar levels in the body to experience excessive amounts, this condition is often felt after complications occur in the body's organs. Patients are diagnosed with diabetes when their blood glucose levels exceed the normal value of the disease. Data mining is the process of obtaining useful information from large databases and needs to be extracted to become new information and can help in decision making, efforts to identify and manage diabetes are very important, considering the high number of patients who come with complaints related to this disease. In this context, especially the decision tree algorithm and K-Nearest Neighbor (KNN) offer an innovative approach to classifying diabetes risk based on patient data, for the need to collect data on diabetic patients who are expected to be able to carry out prevention. Therefore, applying classification techniques with data mining with the C4.5 and K-Nearest Neighbor (KNN) algorithms. Where the classification can achieve superior accuracy. Based on the test results, the Decision Tree method achieved an accuracy of 73.42 percent, a precision of 73.00 percent, a recall of 73.00 percent, and an F1-score of 72.00 percent. Meanwhile, the K-Nearest Neighbor (KNN) method achieved an accuracy of 74.05 percent, a precision of 74.00 percent, a recall of 74.00 percent, and an F1-score of 73.00 percent.

Keywords: Diabetes, Decision Tree, KNN, Method

ABSTRAK

Diabetes adalah penyakit met[illegible] yang kronis yang mana pesien penyakit diabetes tidak menghasilkan ju[illegible]lin yang cukup atau bisa dikatakan tubuh pasien tidak sanggup memanfa[illegible]ulin dengan baik sehingga menyebabkan gula darah di dalam tubuh mer[illegible]mlah yang berlebihan, kondisi ini sering kali dirasakan setelah komplikasi terjadi pada organ tubuh Pasien didiagnosa menderita penyakit diabetes pada saat kadar glukosa darahnya melebihi nilai normal Penyakit. Data mining adalah proses untuk mendapatkan informasi yang berguna dari basis data yang besar dan perlu diekstraksi agar menjadi informasi baru dan dapat membantu dalam pengambilan keputusan, upaya untuk mengidentifikasi dan mengelola diabetes sangat penting, mengingat tingginya jumlah pasien yang datang dengan keluhan terkait penyakit ini. Dalam konteks ini, khususnya algoritma *decision tree* dan K-Nearest Neighbour (KNN) menawarkan pendekatan yang inovatif untuk mengklasifikasikan risiko diabetes berdasarkan data pasien, untuk perlunya dilakukan pendataan pasien penderita diabetes yang diharapkan mampu untuk dilakukannya pencegahan. Oleh sebab itu menerapkan teknik klasifikasi dengan data mining dengan algoritma C4.5 dan K-Nearest Neighbour (KNN) . Dimana klasifikasi tersebut dapat mencapai ketelitian yang unggul. Berdasarkan hasil pengujian metode Decision didapatkan akurasi sebesar 73,42 persen, presisi 73.00 persen, recall 73,00 persen dan F1-Score 72,00 Persen sedangkan untuk K-Nearest Neighbour (KNN) didapatkan hasil pengujian berdasarkan akurasi sebesar 74,05 persen, presisi 74.00 persen, recall 74,00 persen dan F1-Score 73,00 Persen

Kata Kunci : Diabetes, Decision Tree, KNN, Metode

KATA PENGANTAR

Alhamdulillah puji dan [illegible]r Mahasiswa ucapkan kepada Allah SWT atas segala rahmat dan karu[illegible]yang telah memberikan kekuatan dan kesempatan, sehingga mahasiswi dapat menyelesaikan Skripsi ini dengan maksimal. Untuk diajukan sebagai syarat menyelesaikan pendidikan program Sarjana (S-1) Pada Program Studi Sistem Informasi Fakultas Ilmu Teknik Universitas Bina Insan. Sholawat beserta salam semoga tetap tercurahkan kepada bagi Nabi Muhammad SAW, keluarga, sahabat, serta umatnya hingga akhir zaman.

Selama proses penulisan dan penyusunan Skripsi ini, mahasiswa telah berusaha sebaik- baiknya untuk dapat menyelesaikan Skripsi ini baik tepat pada waktunya. Mahasiswa menyadari bahwa Skripsi ini tentunya masih jauh dari sempurna dan mungkin terdapat kesalahan baik sengaja maupun tidak sengaja, oleh sebab itu kritik dan saran yang membangun tentunya sangat diharapkan dari berbagai pihak.

Mahasiswa mengucapkan banyak terima kasih kepada pihak-pihak yang telah membantu selama proses penyelesaian Skripsi ini diantaranya yaitu:

1. Allah SWT atas segala rahmat dan karunia-Nya yang telah memberikan kekuatan dan kesempatan, sehingga mahasiswa dapat menyelesaikan Skripsi ini dengan maksimal.
2. Ibuku Misnawati dan ayahku Rikal Pefi yang telah banyak memberikan dukungan dan bantuannya dalam penulisan Skripsi ini.
3. Bapak Dr. H. Sardiyo, M.M. selaku Rektor Universitas Bina Insan

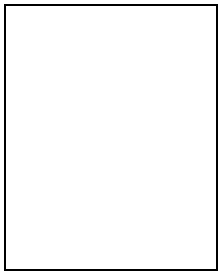
4. Bapak Dr. Muhammad Akbar, S.T., M.M selaku Wakil Rektor I Universitas Bina Insan.
5. Bapak Wakhid Nur M M.Pd., M.M selaku Wakil Rektor II Universitas Bina Insan
6. Bapak Dr. Rudi Kurniawan, ST.,M.Kom selaku Dekan Fakultas Universitas Bina Insan
7. Ibu Harma Oktafia Lingga Wijaya, M.Kom selaku Kepala Program Studi Sistem Informasi Fakultas Ilmu Teknik Universitas Bina Insan
8. Bapak Lukman Hakim, M.Kom selaku Pembimbing I yang telah banyak memberikan bimbingan dan arah dalam penulisan Skripsi ini.
9. Bapak Satrianansyah, M.Kom selaku Pembimbing II yang telah banyak memberikan bimbingan dan arah dalam penulisan Skripsi ini
10. Adik-ku Redho Minallah yang selalu mendukung dan selalu memberikan motivasi untukku.
11. Temanku Antika Yupi Yolanda yang selalu menemaniku dan memotivasiku

Akhir kata semoga penelitian ini dapat bermanfaat bagi untuk penelitian selanjutnya.

Lubuklinggau, 2025

Penulis

DAFTAR RIWAYAT HIDUP



Biodata

Nama	**: ANGGI MIKA**
Tempat / Tanggal Lahir	**: Desa Suro / 30 Juli 2003**
Jenis Kelamin	**: Perempuan**
Agama	**: Islam**
Alamat	**: Desa Suro, Kec. Muara Beliti Kab. Musi Rawas**

Pendidikan

- SD	**: SD Negeri Desa Suro**
- SMP	**: SMP Negeri Muara Beliti**
- SMA	**: SMA Negeri 2 Muara Beliti**

DAFTAR ISI

HALAMAN PERSETUJUAN [illegible]ENGUJI SKRIPSI ..... **ii**
HALAMAN MOTTO PERSE[illegible]N ..... **iii**
HALAMAN PERNYATAAN ..... **iv**
ABSTRACT. ..... **v**
ABSTRAK ..... **vi**
KATA PENGANTAR ..... **vii**
DAFTAR RIWAYAT HIDUPa ..... **ix**
DAFTAR ISI ..... **x**
DAFTAR TABEL ..... **xi**
DAFTAR GAMBAR ..... **xii**
DAFTAR LAMPIRAN ..... **xiv**

BAB I PENDAHULUAN **1**
1.1 Latar Belakang 1
1.2 Identifikasi Masalah 4
1.3 Rumusan Masalah 4
1.4 Batasan Masalah 4
1.5 Tujuan dan Manfaat 5
1.6 Sistematika Penulisan 7

BAB II TINJAUAN PUSTAKA **8**
2.1 Literatur 8
2.2 Penelitian Relevan 12
2.3 Kerangka Berpikir 15

BAB III METODOLOGI PENELITIAN **16**
3.1 Metodelogi Penelitian 16
3.2 Metode Pengumpulan Data 18
3.3 Metode Analisa 18
3.4 Tempat Penelitian 22
3.5 Waktu Penelitian 22

BAB IV HASIL DAN PEMBAHASAN **26**
4.1 Gambaran Umum 26
4.2 Hasil Penelitian 26
4.3 Pembahasan 69

BAB V KESIMPULAN DAN SARAN 84

5.1 Kesimpulan 84

5.2 Saran 85

DAFTAR PUSTAKA

DAFTAR LAMPIRAN

DAFTAR TABEL

Tabel 2.1 Penelitian Relevan 14
Tabel 3. 1 Jadwal Penelitian 23
Tabel 4.1 Dataset Penelitian 29
Tabel 4.2 Mengubah Data Menjadi Kategori 30
Tabel 4.3 Tabel Analisa Pelatihan 31
Tabel 4.4 Tabel Analisa Pengujian 32
Tabel 4.5 Klasifikasi Data Pengujian (Testing) 35

DAFTAR GAMBAR

Gambar 2.2 Kerangka Berpikir 17
Gambar 4.1 Pohon Keputusan 34
Gambar 4.2 Mounting Google 34
Gambar 4.3 Menentukan Path pada Data Set 36
Gambar 4.4 Pembacaan dan Seleksi Awal 37
Gambar 4.5 Memastikan Kualitas Data 38
Gambar 4.6 Visusalisasi Distribusi Data Usia 38
Gambar 4.7 Distribusi Nilai Glukosa 39
Gambar 4.8 Jumlah Observasi per Kehamilan 40
Gambar 4.9 Distribusi BMI Berdasarkan Usia 40
Gambar 4.10 Distribusi Kepadatan BMI 42
Gambar 4.11 Proporsi Jumlah Kehamilan 42
Gambar 4.12 Histogram Distribusi Dataset 42
Gambar 4.13 Matrik Korelasi Fitur 43
Gambar 4.14 Visualisasi Hubungan Glukosa dan Outcome Diabetes 44
Gambar 4.15 Code Pemisahan Fitur dan Target dari Data Base 44
Gambar 4.16 Code Pembagian Dataset 45
Gambar 4.17 Visusalisasi Distribusi Target 45
Gambar 4.18 Code Validasi 5-Fold 45
Gambar 4.19 Code Melatih Model Decision Tree 46
Gambar 4.20 Visualisasi Model 46
Gambar 4.21 Code latihan model 47
Gambar 4.22 Visuaisasi Model Telah dilatih 47
Gambar 4.23 Code Evaluasi Kinerja Model Decision Tree Data Pelatihan 48
Gambar 4.24 Distribusi Data Aktual dan Prediksi 49
Gambar 4.25 Code Menampilkan Matriks Kebingungan 49
Gambar 4.26 Visualisasikan Mmatriks Kebingungan 50
Gambar 4.27 Code Menghitung Matriks Kebingungan , Presisi dan Recall 51
Gambar 4.28 Code Visualisasi Tingkat Kepentingan Fitur dalam Decision Tree 51
Gambar 4.29 Receiver Operating Characteristic (ROC) Curve 52

Gambar 4.30 Pair Plot 53
Gambar 4.31 Analisis Lengkap Karakteriktik Pelatihan 54
Gambar 4.32 Gambar 4. 30 Analisis Lengkap Karakteriktik Pengujian 55
Gambar 4.33 Tingkat Kepenting[illegible]Numerik 57
Gambar 4.34 Hasil Akurasi Dat[illegible] 58
Gambar 4.35 Confunsion Matrik 59
Gambar 4.36 Grafix ROC (Receiver Operating Characteristic Curve) 61
Gambar 4.1 Pair plot Hubungan Antar Fitur 62
Gambar 4.38 Matrik Kebingungan KNN 80
Gambar 4.39 Klasifikasi Report KNN 81
Gambar 4.40 Kurva ROC KNN 81

DAF LAMPIRAN

1. Lembar pengajuan judul
2. Lembar Bimbingan skripsi Pembimbing I
4. Lembar Bimbingan skripsi Pembimbing II
5. Data Penelitian
6. Lembar perbaikan

BAB I

PENDAHULUAN

1.1 Latar Belakang

Diabetes adalah p[illegible] metabolis yang kronis di mana pesien penyakit diabetes tidak menghasilkan jumlah insulin yang cukup atau bisa dikatakan tubuh pasien tidak sanggup memanfaatkan insulin dengan baik sehingga menyebabkan gula darah di dalam tubuh mengalami jumlah yang berlebihan, kondisi ini sering kali dirasakan setelah komplikasi terjadi pada organ tubuh pasien didiagnosa menderita penyakit diabetes pada saat kadar glukosa darahnya melebihi nilai normal. Penyakit. diabetes adalah penyakit yang memiliki kompleksitas tinggi, perawatan medis yang berkelanjutan sangat dibutuhkan guna menurunkan dampak komplikasi dengan pengecekan glikemik [1].

Penyakit diabetes merupakan salah satu tantangan kesehatan masyarakat yang signifikan di seluruh dunia, termasuk di Indonesia. Data menunjukkan bahwa prevalensi diabetes terus meningkat, Diabetes tidak hanya berdampak pada kesehatan individu, tetapi juga mempengaruhi sistem kesehatan secara keseluruhan, mengakibatkan biaya perawatan yang tinggi dan mengurangi kualitas hidup pasien. Pengklasifikasian secara tepat orang-orang yang memiliki hasil tes laboratorium apakah positif atau negatif memiliki penyakit diabetes penting dilakukan untuk memperoleh penanganan yang tepat [2].

Data mining adalah sebuah metode untuk melakukan akuisisi pengetahuan sehingga dengan data mining, informasi-informasi implisit dan

berharga dari sebuah data dapat diekstrak. Data mining adalah proses untuk mendapatkan informasi yang berguna dari basis data yang besar dan perlu diekstraksi agar menjadi informasi baru dan dapat membantu dalam pengambilan keputusan [ ]. Pengertian data mining adalah analisa yang dilakukan secara otomatis pada data besar dan kompleks dengan tujuan untuk mendapatkan pola penting yang keberadaannya biasanya tidak disadari.

Dalam konteks ini, khususnya algoritma D*ecision tree* dan *KNN* menawarkan pendekatan yang inovatif untuk mengklasifikasikan risiko diabetes berdasarkan data penderita diabetes. Guna menyikapi masalah ini, perlu adanya pendeteksian sejak dini penyakit diabetes. Deteksi sejak dini diharapkan dapat menurunkan resiko komplikasi pada pasien diabetes diwaktu mendatang guna menganalisa pasien pengidap penyakit Diabetes sejak dini, Pencatatan terhadap penyakit ini banyak dilakukan agar dapat dilakukan pencegahan. Salah satu yang pencatatan yang bisa dilakukan adalah dengan memanfaatkan teknik klasifikasi dengan metode algoritma *decision tree*.C4.5 dan *K-Nearest Neighbour* (KNN)

Klasifikasi merupakan sebuah proses untuk menciptakan fungsi atau model menjelaskan kelas pada data atau konsep guna untuk memprediksi kelas dari sebuah objek yang labelnya belum didapatkan. Pada penelitian ini, teknik klasifikasi dimanfaatkan untuk mengetahui mengkelompokkan penyebab terjangkit penyakit diabetes dan tidak terjangkit. Beberapa algoritma dapat digunakan untuk perhitungan proses klasifikasi. Algoritma klasifikasi diantaranya adalah *Decision Tree* C4.5 dan *K-Nearest Neighbour* (KNN)

Algoritma *decision tree* C4.5 dan *K-Nearest Neighbour* (KNN) dikenal karena kemampuannya untuk menghasilkan model yang mudah dipahami dan interpretat gga memudahkan tenaga medis dalam mengambil keputusan. D enggunakan data rekam medis pasien, model ini dapat membantu mengidentifikasi faktor risiko dan memprediksi kemungkinan seseorang menderita diabetes [5]. Oleh karena itu, penelitian ini bertujuan untuk menerapkan algoritma *decision tree C4.5* dan *K-Nearest Neighbou*r (KNN) dalam klasifikasi penyakit diabetes dengan harapan dapat memberikan kontribusi signifikan terhadap peningkatan diagnosis serta memperoleh wawasan yang lebih baik mengenai pola-pola risiko diabetes dalam upaya pencegahan dan pengelolaan diabetes di masyarakat selanjutnya

1.2 Identifikasi Masalah

Berdasarkan latar belakang diatas, maka penulis dapat mengidentifikasikan masalah yaitu kurangnya analisa pendeteksi diabetes menggunakan sistem berbasis data mining untuk Klasifikasi data pada penderita penyakit diabetes dengan metode algoritma *decision tree C4.5* dan *K-Nearest Neighbou*r (KNN)

1.3 Rumusan Masalah

Berdasarkan latar belakang dan identifikasi masalah diatas maka penulis merumuskan permasalahan **“**Bagaimana Klasifikasi penyakit diabetes menggunakan Algoritma *Decision Tree C4.5* dan *K-Nearest Neighbou*r (KNN)

1.4 Batasan Masalah

Agar penelitian ini lebih terfokus dan tidak meluas, beberapa batasan masalah yang diterapkan d elitian ini adalah sebagai berikut:

1. Lingkup Data Penelitia dibatasi pada data pasien diabetes yang diperoleh sebagai data primer dan dari data kaggle sebagai data sekunder
2. Fokus pada Algoritma *decision Tree* dan *K-Nearest Neighbou*r (KNN). dengan menggunakan Phyton untuk klasifikasi diabetes, tanpa membahas algoritma *machine learning* lainnya.
3. Variabel yang digunakan hanya variabel tertentu yang akan digunakan dalam model, seperti *Pregnancies, Glucose, Blood presure, Skin thickness, Insulin, BMI, Diabetes Pedigree Function, Age, dan Outcom*e, . Variabel lain yang mungkin relevan tidak akan dianalisis dalam penelitian ini.
4. Metode Evaluasi model akan terbatas pada metrik *akurasi, presisi, recal*l, dan *F1-score*.
5. Waktu dan sumber daya penelitian ini akan dilakukan dari sumber daya yang tersedia, sehingga mungkin ada keterbatasan dalam hal kedalaman analisis dan jumlah data yang digunakan.

1.5 Tujuan dan Manfaat Penelitian

1. Tujuan Penelitian

Penelitian ini bertujuan untuk:

a. Menerapkan algoritma *decision tree C4.5* dan *K-Nearest Neighbou*r (KNN) untuk mengklasifikasikan risiko diabetes

b. Mengidentifikasi faktor risiko untuk mengidentifikasi faktor-faktor risiko yang paling signifikan dalam pengembangan diabetes berdasarkan data pasien.

c. Meningkatkan akurasi untuk meningkatkan akurasi diagnosis diabetes dibandingkan dengan metode tradisional yang saat ini digunakan.

d. Mengklasifikasikan penyakit diabetes menggunakan metode algoritma *decision tree C4.5* dan *K-Nearest Neighbou*r (KNN)

2. Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:

a. Peningkatan Diagnosis Penelitian ini dapat membantu meningkatkan akurasi dan kecepatan diagnosis diabetes, sehingga memungkinkan penanganan yang lebih cepat dan efektif.

b. Penggunaan Data yang Efisien dengan menerapkan algoritma *decision tree C4.5* dan *K-Nearest Neighbou*r (KNN) dan memanfaatkan data pasien dengan lebih efisien untuk pengambilan keputusan klinis.

c. Menjadi media belajar bagi penulis dan juga pembaca terkait penggunaan metode *decision tree C4.5* dan *K-Nearest Neighbou*r (KNN) dalam melakukan klasifikasi.

d. Pencegahan komplikasi Dengan identifikasi faktor risiko yang lebih baik, tindakan pencegahan dapat diambil lebih awal untuk mengurangi kemungkinan komplikasi pada pasien diabetes.

1.6 Sistematika Penulisan

Dalam penulisan skripsi ini yang merupakan laporan dari hasil penelitian, direncanakan terdiri lima bab, masing-masing bab berisi:

BAB I : PENDAHULUAN

Dalam bab ini berisi latar belakang masalah, perumusan masalah, batasan masalah, tujuan dan manfaat penelitian, metode penelitian, dan sistematika penulisan.

BAB II : TINJAUAN PUSTAKA

Dalam bab ini berisi teori-teori yang mendasari masalah yang diteliti, diantaranya klasifikasi, diabeter, algoritma *decision tree C4.5, K-Nearest Neighbour* (KNN), data mining dan kerangka berpikir

BAB III : METODOLOGI PENELITIAN

Dalam bab ini berisi tentang gambaran metode penelitian, pengumpulan data, metode analisa, tempat penelitian dan jadwal penelitian

BAB IV : HASIL DAN PEMBAHASAN

Bab ini berisi tentang hasil dan pembahasan kode model dan perhitungan klasifikasi penyakit diabetes

BAB V : KESIMPULAN DAN SARAN

Bab ini berisi kesimpulan dari seluruh hasil penelitian dan saran-saran/ masukan-masukan yang berguna di masa yang akan datang

BAB II
TINJAUAN PUSTAKA

2.1 Literatur

2.1.1 Klasifikasi

Klasifikasi adalah proses katagori yang dilakukan dalam sekumpulan data kemudian membaginya dalam kelas kelas tertentu. Klasifikasi memberikan penilaian objek data untuk memasukannya kedalam kelas tertentu dari jumlah kelas yang tersedia. Klasifikasi dapat didefinisikan sebagai pekerjaan yang melakukan pelatihan/pembelajaran terhadap fungsi target yang memetakan setiap set atribut ke satu jumlah label kelas yang tersedia [4]. Klasifikasi adalah proses dari mencari suatu himpunan model (fungsi) yang dapat mendeskripsikan dan membedakan kelas-kelas data, dengan tujuan dapat menggunakan model tersebut untuk memprediksi kelas dari suatu objek yang mana kelasnya belum diketahui.

Klasifikasi diberikan sejumlah *record* yang dinamakan langkah pelatihan, yang terdiri dari beberapa atribut, salah satu atribut menunjukkan kelas untuk record yang dapat digunakan untuk menemukan model dari langkah pelatihan dan langkah pengujian

2.1.2 Diabetes

Diabetes adalah penyakit yang memiliki metabolisme yang ditandai dengan tingginya kadar glukosa darah [5]. Orang yang menderita penyakit diabetes memiliki peningkatan risiko masalah kesehatan serius hingga mengancam jiwa yang mengakibatkan biaya perawatan medis, penurunan kualitas hidup dan peningkatan kematian [6] Diabetes juga

dapat dikatakan sebagai sebuah penyakit metabolik yang disebabkan oeleh kurangnya hormon insulin atau ketidak mampuan tubuh dalam memanfaatkan insulin, [illegible] kadar glukosa atau kadar gula dalam darah tidak terkendali.

Mengemukakan bahwa pasien yang sedang mengalami gejala diabetes dianjurkan untuk lebih mengkontrol diri karena hal ini sangat penting untuk mencegah komplikasi akut dan mengurangi risiko komplikasi.

Diabetes merupakan masalah kesehatan masyarakat global yang signifikan, bertanggung jawab atas mortalitas dan morbiditas yang cukup besar di seluruh dunia dan menyebabkan kerugian ekonomi yang substansial [7].

Diabetes memberikan kontribusi sebagai salah satu penyebab kematian umum pada penderita penyakit jantung dan pembuluh darah, diabetes dapat dikelompokan menjadi beberapa bagian yaitu: [8]

1. Diabetes melitus tipe 1 meliputi diabetes disebabkan oleh penghancuran sel *B* pankreas baik oleh proses autoimun maupun idiopatik sehingga produksi insulin berkurang bahkan berhenti. Biasanya diabetes tipe ini terjadi pada usia kurang dari 20 tahun.
2. Diabetes melitus tipe 2 merupakan diabetes dengan kelainan metabolik yang ditandai dengan kadar glukosa darah yang tinggi dalam konteks resitensi insulin dan defisiensi insulin relatif. Diabetes tipe ini biasanya diderita pada usia dari 20 tahun.

3. Diabetes mellitus gestasional pada intoleransi glukosa pada masa pengenalan pertama selama kehamilan.
4. Jenis spesifik lainnya mencakup berbagai macam kondisi tidak umum, terutama bentuk diabetes yang ditentukan secara genetik atau diabetes yang terkait dengan penyakit lain atau penggunaan narkoba.

2.1.3 Algoritma C4.5

C4.5 merupakan salah satu algoritma pada *decision tree* berdasarkan informasi entropi. Algoritma ini menggunakan kriteria pemisahan yang diimodifikasi yang disebut dengan rasio *gain*. Tujuan dari algoritma ini adalah untuk menemukan beberapa hubungan antara variabel predikator dan kategori melalui pelatihan dan pembelajaran set pelatihan, kemudian menerapkan hubungan ini ke contoh, mengklasifikasi data dan menyelesaikan pengambilan keputusan. Algoritma C4.5 merupakan pembangunan algoritma ID3 dimana kekurangan yang dimiliki algoritma ID3 ditutupi oleh algoritma C4.5. Empat hal yang membedakan algoritma C4.5 dengan ID3 antara lain: tahan (robust) terhadap data noise, mampu mengenai variabel dengan tipe diskrit maupun kontinu, mampu mengenai variabel yang memiliki missing value, dan dapat memangkas cabang dari pohon keputusan [9].

Berikut langkah-langkah membangun pohon keputusan dengan algoritma C4.5

a. Menentukan nilai *Entropy*(S) untuk setiap nilai kriteria

Pada langkah ini, dilakukan pencarian nilai *entropy*(S), di mana *entropy*(S) berfungsi sebagai ukuran untuk menilai variasi setiap nilai

atribut kriteria terhadap atribut keputusan dalam suatu *dataset*. Semakin rendah nilai *entropy* (S), maka tingkat variasi dalam *dataset* semakin rendah, sebaliknya, semakin tinggi nilai *entropy* (S), maka tingkat variasinya menjadi lebih tinggi. Rumus matematika *entropy* (S)[12], adalah seperti berikut:

$$Entropy(S) = \sum_{i=1}^{n} - pi * log2(pi)$$

Dimana:

S = Jumlah kasus (*Sampling*)

n = Jumlah partisi S

pi = Proporsi dari Si terhadap S

Dalam proses pencarian nilai jika semua nilai *entropy (S)* kriteria terhadap atribut keputusan sama, maka nilai *entropy* = 1, dan jika hanya satu nilai kriteria yang tidak sama dengan 0, maka nilai *Entropy* = 0.

b. Menentukan nilai *Gain* (S,A) untuk setiap atribut *Gain* (S,A) adalah hasil dari pengurangan total nilai *entropy* dari nilai setiap atribut kriteria dikalikan dengan proporsi nilai atribut terhadap jumlah kasus. *Gain*(S,A) berperan sebagai ukuran efektivitas setiap atribut kriteria dalam proses klasifikasi data. Untuk mencari nilai *Gain* (S,A)[12], digunakan rumus sebagai berikut:

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{S} Entropy(Si)$$

Dimana

S = Jumlah kasus (*Sampling*) \

A = atribut

n = Jumlah partisi S

$|Si|$ = Jumlah kasus pada partisi ke-i

$|S|$ = Jumlah kasus dalam S

c. Membentuk node dengan aturan cabang berdasarkan *Gain* tertinggi

d. Mengulangi proses untuk masing-masing cabang

2.1.4 *Decision Tree*

Decision tree merupakan salah satu metode klasifikasi yang menggunakan representasi struktur pohon dimana setiap node mempresentasikan atribut, dimana cabangnya mempresentasikan nilai dari atribut, dan daun mempresentasikan kelasnya. *Decision tree* adalah sebuah struktur yang dapat digunakan untuk membagi kumpulan-kumpulan data yang besar menjadi himpunan-hinpunan *record* yang lebih kecil dengan menerapkan serangkaian aturan keputusan [10].

Decision tree adalah metode klasifikasi yang mudah untuk dipahami atau diintrepetasikan oleh manusia, sehingga metode ini menjadi salah satu metode yang cukup populer. *Decision tree* adalah teknik yang banyak digunakan untuk membangun model klasifikasi yang berdasarkan kumpulan data yang dikumpulkan. Ini adalah suatu teknik yang menciptakan pohon aturan dengan menghitung rasio keuntungan (*gain ratio*) yang memberikan bobot tertentu pada atribut yang terdapat dalam sebuah himpunan data.

Seorang peneliti mengatakan bahwa *decision tree* merupakan salah satu metode klasifikasi yang menggunakan representasi struktur pohon dimana setiap *node* merepresentasikan atribut, dimana cabangnya

merepresentasikan nilai dari atribut, dan daun merepresentasikan kelas. Akar dari *decision tree* disebut sebagai *root*. Pada *decision tree* terdapat 3 jenis *node,* yaitu *Root N* *rnal node,*dan *leaf node.* Setiap *leaf node* di pohon mewakili tes u* atribut, dan cabangnya mewakili setiap hasil tes (ya dan tidak berarti positif dan negatif).

2.1.5 *K-Nearest Neighbou*r (KNN)

Algoritma K-Nearest Neighbor (KNN) adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Data pembelajaran diproyeksikan ke ruang berdimensi banyak, dimana masing-masing dimensi merepresentasikan fitur dari data. Algoritma KNN termasuk metode yang menggunakan algoritma supervised .Prinsip kerja KNN adalah mencari jarak terdekat antara data yang akan dievaluasi dengan K tetangga (neighbor) terdekatnya dalam data pelatihan. Teknik ini termasuk dalam kelompok klasifikasi nonparametric. Pada proses pengklasifikasian, algoritma ini tidak menggunakan model apapun untuk dicocokkan dan hanya berdasarkan pada memori[11).

Algoritma KNN menggunakan klasifikasi ketetanggaan sebagai nilai prediksi dari data uji yang baru. Jarak yang digunakan adalah jarak Euclidean Distance. Jarak *Euclidean* adalah jarak yang paling umum digunakan pada data numerik. Nilai K yang terbaik untuk algoritma ini tergantung pada data. Secara umum, nilai K yang tinggi akan mengurangi efek noise pada klasifikasi. Kasus khusus dimana klasifikasi diprekdisikan

berdasarkan training data yang paling dekat disebut algoritma KNN. Adapun Tahapan Algoritma KNN sebagai berikut:

1. Tentukan parameter K
2. Hitung jarak antara data yang akan dievaluasi dengan semua pelatihan
3. Urutkan jarak yang terbentuk (dari terkecil ke terbesar).
4. Tentukan jarak terdekat sejumlah K
5. Pasangkan kelas yang bersesuaian
6. Cari jumlah kelas dari tetangga yang terdekat dan tetapkan kelas tersebut sebagai kelas data yang akan dievaluasi.

2.16 Data Mining

Data mining atau disebut juga dengan *Knowledge Discovery* in *Database* (KDD) merupakan aktivitas yang berkaitan dengan pengumpulan data, pemakaian data historis untuk menemukan pengetahuan, informasi, keteraturan, pola atau hubungan dalam data yang berukuran besar [12]. Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, ataupun algoritma yang ada dalam data mining sangat bervariasi.

Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan. Kemampuan data mining dalam menggali informasi dalam cakupan yang sangat besar ini menjadi kelebihan yang tidak perlu diragukan. Teknologi seperti ini biasanya digunakan untuk memprediksi berbagai hal dalam kehidupan dimana data mining mengotomatisasi proses pencarian

2.1.7 Bahasa Python

Python merupakan salah satu bahasa pemrograman yang banyak digunakan oleh perus besar maupun para *developer* untuk mengembangkan berba m aplikasi berbasis desktop, web dan mobile. Python diciptakan oleh Guido van Rossum di Belanda pada tahun 1990 dan namanya diambildari acara televisi kesukaan mengembangkan Python sebagaihobi, kemudian Python menjadi bahasa pemrograman yang dipakai secara luas dalam industri dan pendidikan karena sederhana[13]

2.2 Penelitian Relevan

Dalam proses penelitian ini penulis mencari teori-teori penelitian yang relavan untuk dapat mendukung penyelesaian penelitian ini berupa Jurnal-jurnal yang dapat dijadikan acuan terkait dengan masalah penyakit Diabetes, diantaranya sebagai berikut:

Tabel 2.1 Penelitian Relevan

No	Judul	Penulis	Hasil
1	Implementasi data mining untuk prediksi penyakit diabetes dengan algoritma C4.5	S. Ucha Putri, E. Irawan, F.Rizky, S.Tunas Bangsa,P.A.(2021) [4]	Dalam penelitian, data sampel yang digunakan adalah 49 data pasien penyakit diabetes dari RSUD. Dr. Djasmen saragih pematang siantar.
2	Deteksi penyakit diabetes mellitus menggunakan algoritma	A. Afifuddin and L. Hakim.(2023) [9]	Penelitian ini menggunakan algoritma C4.5 dengan datset berjumlah 2000 dan menggunakan 5 variabel, dataset ini menghasikan

	decision tree model untuk C4.5		aplikasi pendeksi penyakit diabetes
3	Optimalisasin *feature selection* untuk mendeteksi metode *decision tree*	A. [illegible] Heriansyah , ar[illegible]stuti, (2024) [6]	Implementasi metode *feature selection* menggunakan Information Gain pada model Decision Tree menghasilkan tingkat akurasi sebesar 72,25% menunjukkan bahwa hasil ini meningkat dibandingkan dengan beberapa penelitian terdahulu.
4	Klasifikasi Pasien Diabetes Mellitus Menggunakan Metode *Smooth Support Vector Machine* (Ssvm)	Rizky Adhi Nugroho1 , Tarno2 , Alan Prahutama (2017) [14]	Klasifikasi dengan metode SSVM yang terbaik didapatkan dengan nilai akurasi sebesar 0,9703. Selain akurasi, tingkat ketepatan akurasi dapat diketahui dari nilai sensitivity keakuratan klasifikasi pada kelas positif dan specificity yaitu keakuratan klasifikasi pada kelas negatif

5	Klasifikasi Diabetes Mellitus Menggunakan Support vector Machine (Studi Kasus: Puskesmas Modopuro, Mojokerto)	Andharini Dwi Cahyani1*) , Ari Ba... 019) [15]	Nilai akurasi cross validation terbaik yang didapatkan menggunakan kernel linear, polynomial, dan sigmoid masing-masing yaitu 62%, 64%, dan 54%. Berdasarkan hasil pengamatan kernel polynomial mendapatkan hasil akurasi yang lebih baik dari kernel lainnya
6	Klasifikasi Penyakit Diabetes Melitus Menggunakan *Adaboost Classifier*	Ginanjar Abdurrahma (2022) [20]	Hasil klasifikasi *Adaboost Clasifier* pada dataset setelah *imputing mean* diperoleh nilai akurasi sebesar 80.09 %, sedangkan untuk dataset setelah *imputing median* diperoleh nilai akurasi sebesar 76.19 %,

2.3 Kerangka Berfikir

Berikut dibawah ini adal gka berpikir yang menjadi acuan dalam metodologi penelitian ini

Kerangka Berpikir	
Rumusan Masalah	"Bagaimana Klasifikasi penyakit diabetes menggunakan Algoritma Decision Tree dan KNN"
Pengumpulan Data	Berasal dari data sekunder dan data primer dari tempat penelitian
Preprocesing	Label Encoding, Normalisasi & Correlation Matrix
Extraks Fitur	Correlation Analysis, Feature Importance
Pelatihan Model	Algoritma Decision Tree dan KNN
Evaluasi	Menghitung Akurasi dengan confision matrix, Classification report, Roc Curve, K-fold cross validation

Gambar 2.1 Kerangka Berpikir

BAB III
METODOLOGI PENELITIAN

3.1. Metode Penelitian

Penelitian ini menggunakan *decision tree* C4.5 untuk membangun sebuah pohon keputusan yang lebih mudah untuk dimengerti, fleksibel, dan menarik. *Decision tree* atau pohon keputusan adalah model untuk memprediksi sebuah struktur pohon atau hirarki untuk mengubah data menjadi pohon keputusan [20].

Langkah- langkah yang digunakan dalam penelitian ini yaitu

1. Mendeskripsikan data penelitian, *preprocessing*, dan mempartisi data menjadi 80% data latih dan 20% data uji.
2. Mengklasifikasi status penyakit diabetes menggunakan metode *Decision Tree*
    a. Menghitung *entropy*
    b. Menghitung nilai *gain*
    c. *Gain* tertinggi akan dipilih untuk menjadi *root node* akar pohon
    d. Ulangi langkah perhitungan *entropy d*an *gain* sampai seluruh variabel predictor masuk dalam kelas. variabel prediktor yang sudah terpilih sebelumnya maka tidak diikutkan untuk perhitungan selanjutnya.
    e. Menghitung tingkat ketepatan klasifikasi.

3.2 Metode Pengumpulan Data

Metode pengumpulan data merupakan langkah penting dalam memperoleh data yang diperlukan untuk keperluan penelitian. Dalam

penelitian ini peneliti menggunakan metode pengumpulan data, berupa data sekunder berasal dari data kaggle yang mencakup 8 atribut, dan 1 label diantaranya: *Pregnancies, Bloodpresure, Skinthickness, Insulin, BMI, Diabetes Pedigree Age, dan Outcome* dengan jumlah data yang didapat adalah 768 record data ditambah dengan data primer yang berjumlah 19 data yang berkaitan dengan penyakit diabetes.

Adapun metode yang digunakan dalam pengumpuan data penelitian ini adalah

a. Metode Observasi

Peneliti melakukan observasi dengan mencatat data-data yang berkaitan dengan penelitian

b. Metode Dokumentasi

Penulis mencari dokumen-dokumen yang diperlukan dalam penulisan laporan penelitian ini dengan cara mendokumentasikan dokumen dan data tersebut.

c. Metode Study Pustaka

Metode Study Literature adalah metode pengumpulan data dengan membaca buku-buku, jurnal penelitian terdahulu yang berhubungan langsung dengan topik penelitian

3.3 Metode Analisa

Metode analisa yang digunakan dalam penelitian ini adalah sebagai berikut:

1. Ekpolarasi Data

Langkah awal ini bertujuan untuk memahami karakteristik dataset yang digunakan. Analisis dilakukan dengan Visualisasi Data yaitu membuat tabel untuk memahami distribusi data. menggunakan *correlation matrix* untuk melihat hubungan antar fitur, terutama antara fitur independen dan variabel target.

2. Pra-pemrosesan Data

Pra-pemrosesan dilakukan untuk memastikan data dalam kondisi siap untuk diolah. Tahap ini meliputi:

a. Penanganan nilai hilang:

Rata-rata (*mean*) atau median dari fitur.Prediksi berbasis regresi untuk nilai yang hilang.

b. Normalisasi Data

Data dinormalisasi agar setiap fitur memiliki skala yang seragam.

c. Deteksi dan penanganan *Outliers*:

Data yang berada diluar rentang normal diidentifikasi menggunakan *boxplot. Outliers* dapat dihapus atau digantikan jika dianggap mengganggu proses analisis.

3. Implementasi Algoritma *Decision Tree dan KNN*

Proses implementasi algoritma melibatkan:

a. Pemilihan kriteria pemisahan:

Gini Index: mengukur ketidakpastian pada simpul pohon.

Information Gain: mengukur seberapa banyak informasi yang diperoleh setelah pembagian. Data.

b. Pembentukan model:

Dataset dibagi menjadi data (*training set*) dan data uji (*testing set*) dengan perbandingan 80:20. Model dilatih menggunakan data latih untuk membentuk str hon.

c. Optimasi parameter:

Parameter seperti kedalaman maksimum pohon (*max_depth*) dan jumlah minimum data di setiap simpul (*min_samples_split*) optimalkan menggunakan pencarian grrid (*grid search*).

4. Evaluasi Model

Evaluasi dilakukan untuk mengukur performa model terhadap data uji dengan menggunakan metrik berikut:

a. *Confusion Matrix:*

Tabel yang menunjukkan jumlah prediksi benar dan salah pada kategori Positif (diabetes) dan Negatif (tidak diabetes).

b. Metrik Evaluasi:

Akurasi: Persentase prediksi benar dari total prediksi.P resisi: Kemampuan model mendeteksi kasus positif dengan benar (Prediksi Positif Benar / Total Prediksi Positif). Recall: Kemampuan model mendeteksi semua kasus positif (Prediksi Positif Benar / Total Kasus Positif). F1-Score: Rata-rata harmonis antara presisi dan recall, berguna saat data tidak seimbang.

5. Validasi Model

Untuk memastikan bahwa model tidak bias dan mampu memberikan hasil yang stabil, dilakukan validasi sebagai berikut:

a. K-Fold Cross Validation:

Dataset dibagi menjadi K subset. Model dilatih dan diuji sebanyak K kali, dengan setiap subset bergantian menjadi data uji. Hasil dari setiap iterasi dirata-rata untuk mendapatkan performa keseluruhan model.

b. Analisis Validasi:

Membandingkan performa antara data latih dan data uji untuk memastikan bahwa model tidak mengalami overfitting (terlalu cocok dengan data latih).

6. Interpretasi dan Analisis Hasil

Langkah terakhir adalah memahami hasil model untuk menjawab pertanyaan penelitian:

a. Analisis Fitur Penting:

Fitur-fitur yang memberikan kontribusi besar terhadap prediksi (berdasarkan struktur pohon) dianalisis.

b. Perbandingan Kinerja Model:

Kinerja model dibandingkan dengan penelitian terdahulu atau metode lain untuk melihat keunggulan dan kelemahannya.

c. Kesimpulan dan Rekomendasi:

Memberikan kesimpulan mengenai efektivitas algoritma decision tree dalam mengklasifikasikan penyakit diabetes. Memberikan saran penggunaan model untuk aplikasi praktis, seperti sistem pendukung keputusan di bidang medis.

3.4 Tempat Penelitian

Penelitian ini dilaksanakan di laboratorium Universitas Bina Insan jln Jend. Besar Soeharto Lubuk Ku[illegible]uklinggau

3.5 Waktu Penelitian

Dalam penyelesaian Penelitian skripsi ini, waktu penelitian yang dibutuhkan oleh Mahasiswa dari bulan Februari 2025 sampai dengan Juli 2025, dapat dilihat pada tabel berikut ini :

Tabel 3.1 Jadwal Penelitian

No	Jenis Kegiatan	Waktu Kegiatan																					
		Feb				Mar				Apr				Mei				Jun				Jul			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	Pengajuan Judul																								
2	Penyusunan Proposal																								
3	Bimbingan proposal																								
4	Pengumpulan Data																								
5	Seminar Proposal																								
6	Perbaikan Proposal																								
7	Pembuatan Kode Model																								
8	Penyusunan Skripsi																								
9	Bimbingan Skripsi																								
10	Sidang Skripsi																								

3.6 Alat dan Bahan

Dalam penyelesaian penelitian ini digunakan alat diantaranya :

1. Alat

a. Hardware:

a) Komputer/Laptop

b) Printer

c) *Sandisk* 32GB

b. Software:

a) Sistem operasi *windows* 11

b) *Mendeley Desktop*

c) *Microsoft Office*

d) *Google colab*

e) *Google Drive*

2. Bahan

Bahan yang digunakan dalam membuat penelitian ini yaitu:

a. Data Pasien diabetes

b. Printer

c. Tinta printer

d. Kertas A4 80 gram

e. Jurnal Penelitian

3.7 Metode Pengujian Sistem Dan Pengolahan Data

Berikut adalah metode pengujian sistem dan pengolahan data untuk penelitian klasifikasi penyakit diabetes ini peneliti menggunakan algoritma *decision tree*: C4.5 dan KNN

Metode Pengujian Sistem dilakukan dengan tahap

1. Uji Validitas dan Reliabilitas Data

   a. Validitas memastikan yang dikumpulkan akurat dan relevan. Gunakan teknik seperti validation untuk memeriksa apakah data merepresentasikan populasi yang dituju.

   b. Reliabilitas uji konsistensi pengukuran dengan menggunakan metode retest atau split-half untuk memastikan data memberikan hasil yang sama jika diuji kembali.

2. Pengujian Model

   a. Pembagian Data Bagi dataset menjadi dua bagian: data pelatihan (80%) dan data pengujian (20%).

   b. Pengujian Kinerja Model Setelah melatih model decision tree, uji model menggunakan data pengujian dan hitung metrik evaluasi seperti:

      a) Akurasi Proporsi prediksi yang benar.

      b) Presisi Proporsi prediksi positif yang benar.

      c) Recall Proporsi kasus positif yang terdeteksi.

      d) F1-Score Rata-rata harmonis dari presisi dan recall.

      e) Visualisasi Hasil Gunakan confusion matrix untuk menggambarkan performa model secara lebih jelas.

Sedangkan Metode Pengolahan Data dilakukan dengan cara

1. Pra-pemrosesan Data

   a. Pembersihan Data Identifikasi dan hilangkan nilai yang hilang atau outlier.

b. Transformasi Data Normalisasi atau standarisasi fitur untuk memastikan keselarasan data.

2. Analisis Eksploratori

a. Lakukan analisis des uk memahami distribusi data, hubungan antar fitur, dan faktor risiko yang berkontribusi terhadap diabetes.

b. Visualisasikan data menggunakan grafik seperti histogram, boxplot, dan scatter plot.

3. Pembangunan Model

a. Gunakan algoritma decision tree dan KNN untuk melatih model dengan data pelatihan.

b. Sesuaikan parameter seperti kedalaman pohon, minimum sampel untuk membagi, dan lainnya.

c. Evaluasi Model

Evauasi dilakukan dengan uji model dengan data pengujian, hitung metrik evaluasi, dan bandingkan hasil dengan model lain jika ada.

d. Gunakan teknik validasi silang untuk memastikan model tidak mengalami overfitting.

e. Analisis Fitur Penting

f. Identifikasi fitur-fitur yang paling berpengaruh terhadap keputusan model menggunakan metode seperti feature importance dari *decision tree.*

BAB IV
HASIL DAN PEMBAHASAN

4.1 Gambaran Umum

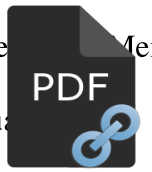
Penelitian ini dilakukan dengan menggunakan data diabetes sebagai data sekunder dan data observasi sebagai data Primer, penelitian ini dilakukan secara mandiri oleh peneliti di lingkungan Fakultas Ilmu Komputer, Program Studi Sistem Informasi, dengan menggunakan perangkat lunak Python, Google Colaboratory, dan berbagai pustaka machine learning untuk proses analisis data. Seluruh proses eksperimen dilakukan secara digital dan terkomputerisasi.

4.2 Hasil Penelitian

Penelitian ini bertujuan untuk mengklasifikasikan penyakit diabetes berdasarkan data medis menggunakan algoritma *Decision Tree* C4.5 Dan *KNN* dan Langkah-langkah utama yang dilakukan dalam proses eksperimen meliputi:

1. Pra-pemrosesan data: Mengganti nilai-nilai nol yang tidak logis dengan nilai median untuk fitur-fitur penting seperti *Pregnancies, Glucose, Bloodpresure, Skinthickness, Insulin, BMI, Diabetes Pedigree Function, Age, dan Outcom*e
2. Visualisasi data: Untuk memahami distribusi dan korelasi antar fitur.
3. Pemisahan data: Data dibagi menjadi data latih dan data uji dengan rasio 80:20.
4. Pelatihan model: Menggunakan pipeline dengan *Standard Scale*r dan *Decision Tree Classifier.*

5. Evaluasi model: Melalui akurasi, confusion matrix, ROC-AUC, dan classification report.
6. Monitoring hasil ekspe menggunakan MLflow untuk menyimpan model dan metrik evalu



Model juga berhasil disimpan dan dimuat kembali menggunakan joblib, yang membuktikan bahwa model dapat digunakan ulang (*deployable*) untuk keperluan implementasi di masa depan.

4.2.1 Pengambilan Data Set

Sumber informasi dalam penelitian ini berasal dari pada data data sekunder dan data primer yang berasal data publik sebagai data sekunder dan data dari Rumah Sakit Siti Aisyah Lubuklinggau sebagai data primer. Data tersebut mencakup dari beberapa atribut, diantaranya: *Pregnancies, Glucose, Bloodpresure, Skinthickness, Insulin, BMI, Diabetes Pedigree Function, Age, dan Outcom*e dengan jumlah data yang didapat adalah 787 data yang beasal 768 data sekunder ditambah dengan 19 data record data primer yang berasal dari observasi dari hal yang terkait.

Data yang digunakan dalam penelitian ini adalah data yang ada hubungan dengan data penyakit diabetes yang diderita. .Algoritma *Decision Tree* C4.5 dan *KNN* dengan bahasa Python digunakan untuk membuat model aturan dari kumpulan data yang Mahasiswa kumpulkan. Tabel 4.1 merupakan dataset yang digunakan dalam penelitian ini.

Tabel 4.1 Dataset Penelitian

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigree Function	Age	Outcome
6	98	58	33	190	34	0,43	43	0
2	112	75	32	0	35,7	0,148	21	0
2	108	64	0	0	30,8	0,158	21	0
8	107	80	0	0	24,6	0,856	34	0
7	136	90	0	0	29,9	0,21	50	0
6	103	72	32	190	37,7	0,324	55	0
1	71	48	18	76	20,4	0,323	22	0
0	117	0	0	0	33,8	0,932	44	0
4	154	72	29	126	31,3	0,338	37	0
5	147	78	0	0	33,7	0,218	65	0
10	111	70	27	0	27,5	0,141	40	1
7	179	95	31	0	34,2	0,164	60	0
4	148	60	27	318	30,9	0,15	29	1
5	96	74	18	67	33,6	0,997	43	0
2	88	58	26	16	28,4	0,766	22	0
1	125	50	40	167	33,3	0,962	28	1
3	84	72	32	0	37,2	0,267	28	0
5	86	68	28	71	30,2	0,364	24	0
0	125	96	0	0	22,5	0,262	21	0
1	551	123	0	121	21,5	0	66	1
3	132	170	0	87	26,0	0	55	1
2	288	161	0	44	21,5	0	50	1
1	223	118	0	92	21,5	0	52	1
1	186	101	0	182	37,9	0	53	1
0	294	187	0	105	49,7	0	55	1
0	230	128	0	87	35,8	0	66	1
2	201	106	0	153	21,5	0	44	1
2	204	142	0	0	21,5	0	44	1
2	346	113	0	212	28,6	0	49	1
1	106	188	0	33	22,4	0	49	1
0	142	326	0	88	27,6	0	57	1
1	227	92	0	142	43,5	0	57	1
1	481	170	0	0	34	0	45	1
0	425	140	0	0	40,6	0	54	1
2	118	136	0	87	24,7	0	63	1
3	114	144	0	96	40,6	0	67	1
2	120	118	0	77	23,6	0	36	1
0	110	96	0	143	32,2	0	45	1

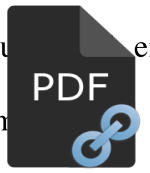
Guna mendapatkan hasil perhitungan menggunakan algoritma C4.5 *decision tree* dan *KNN*, preeprosesing sebelumnya perlu dilakukan transformasi data dengan mendiskritisasi variabel kontinu menjadi varibel katagorik untuk memudahkan dalam proses perhitungan sesuai dengan pada tabel 4.1, dimana kriteria jumlah melahirkan di bagi menjadi 4 bagian yaitu tidak pernah melahirkan, melahirkan 1 kali, 2 kali, 3 kali dan lebih

dari 4 kali, untuk kadar gula darah (*Glukosa*) bernilai rendah jika kriteria dibawah 70mgdl, akan bernilai normal jika kriteria antara 70 - 100mgdl, dan bernilai tinggi jika k rnilai lebih besar dari 100mgdl. Tekanan Darah (*Blood pressure*) dikatakan normal jika kriteria bernilai 65-75 mmHg, bernilai tinggi jika melebihi dari 75 mmHg., dan bernilai rendah jika dibawah 65 mmHg, Normal Ketebalan Kulit (*skin thickness*) seseorang akan dikatakan normal jika bernilai 4.90 – 21mm. Normal insulin berkisar antara 140-199 mgdl, bernilai tinggi jika lebih dari 200mgdl. Normal indek masa tubuh (bmi) seseorang berkisar antara 18,5-25. Normal dpfr seseorang bernilai jika lebih 2, dan bernilai tinggi jika kurang dari 2. untuk batas usia Batas usia muda adalah dari umur 10-45 tahun, dan diatas 45 tahun. Di katagori Tua dari Hasil mengubah data numerik menjadi data kategorikal, ditunjukan pada tabel 4.2

Variabel / Atribut	Kategori	Jumlah	Jumlah Positif	Jumlah Negati
Pregnancies	nulipara (0)	116	43	73
	primigravida (1)	141	35	106
	multigravida (2)	109	25	84
	multipara (3)	77	29	48
	grandemultipara (>3)	344	155	189
Glucose	rendah (<100 mg/dL)	192	14	178
	normal (100-125 mg/dL)	284	83	201
	tinggi (>125 mg/dL)	311	190	121
Blood Pressure	rendah (<65 mmHg)	201	46	155
	normal (65-75 mmHg)	278	97	181
	tinggi (>75 mmHg)	308	144	164
SkinThickness	rendah (<10 mm)	4	1	3
	normal (10-25 mm)	196	37	159
	tinggi (>25 mm)	587	249	338
Insulin	rendah (<140 mg/dL)	609	195	414
	normal (140-199 mg/dL)	88	44	44
	tinggi (>199 mg/dL)	90	48	42
BMI	rendah (<18.5)	4	0	4
	normal (18.5-25)	116	15	101
	tinggi (>25)	667	272	395
Diabetes Pedigree Function	rendah (>2)	4	3	1
	tinggi (<=2)	783	284	499
Age	muda (10-45)	655	215	440
	tua (>45)	132	72	60
Outcome	positif (1)	287	287	0
	negatif (0)	500	0	500

Tabel 4. 2 Mengubah Data Menjadi Kategori

Hasil transformasi data dibuat oleh proses komputer yang menentukan nilai entropi dari setiap atribu entropi yang diperoleh diproses untuk mendapatkan nilai konfirm



4.2.2 Pembagian Data Set

Kemudian, setelah semua operasi selesai dilakukan partisi data untuk membagi data ke dalam dua bagian secara acak, dimana dalam penelitian ini yang digunakan adalah data pelatihan (treen) sebesar 80% yaitu 629 data pengamatan dan data pengujian(Test) sebesar 20% yaitu 158 pengamatan

Tabel 4. 3 Tabel Analisa Pelatihan

Variabel / Atribut	Kategori	Jumlah	Jumlah Positif	Jumlah Negatif
Pregnancies	nulipara (0)	98	36	62
	primigravida (1)	116	28	88
	multigravida (2)	80	18	62
	multipara (3)	63	25	38
	grandemultipara (>3)	272	122	150
Glucose	rendah (<100 mg/dL)	150	11	139
	normal (100-125 mg/dL)	221	64	157
	tinggi (>125 mg/dL)	258	154	104
Blood Pressure	rendah (<65 mmHg)	152	34	118
	normal (65-75 mmHg)	225	77	148
	tinggi (>75 mmHg)	252	118	134
SkinThickness	rendah (<10 mm)	4	1	3
	normal (10-25 mm)	144	29	115
	tinggi (>25 mm)	481	199	282
Insulin	rendah (<140 mg/dL)	481	152	329
	normal (140-199 mg/dL)	72	34	38
	tinggi (>199 mg/dL)	76	43	33
BMI	rendah (<18.5)	2	0	2
	normal (18.5-25)	82	13	69
	tinggi (>25)	545	216	329
Diabetes Pedigree Function	rendah (>2)	4	3	1
	tinggi (<=2)	625	226	399
Age	muda (10-45)	518	169	349
	tua (>45)	111	60	51
Outcome	positif (1)	229	229	0
	negatif (0)	400	0	400

Tabel 4. 4 Tabel Analisa Pengujian

Variabel / Atribut		Jumlah	Jumlah Positif	Jumlah Negatif
Pregnancies		18	7	11
	(1)	25	7	18
	(2)	29	7	22
	(3)	14	4	10
	grandemultipara (>3)	72	33	39
Glucose	rendah (<100 mg/dL)	42	3	39
	normal (100-125 mg/dL)	63	19	44
	tinggi (>125 mg/dL)	53	36	17
Blood Pressure	rendah (<65 mmHg)	49	12	37
	normal (65-75 mmHg)	53	20	33
	tinggi (>75 mmHg)	56	26	30
SkinThickness	rendah (<10 mm)	0	0	0
	normal (10-25 mm)	52	8	44
	tinggi (>25 mm)	106	50	56
Insulin	rendah (<140 mg/dL)	128	43	85
	normal (140-199 mg/dL)	16	10	6
	tinggi (>199 mg/dL)	14	5	9
BMI	rendah (<18.5)	2	0	2
	normal (18.5-25)	34	2	32
	tinggi (>25)	122	56	66
Diabetes Pedigree Function	rendah (>2)	0	0	0
	tinggi (<=2)	158	58	100
Age	muda (10-45)	137	46	91
	tua (>45)	21	12	9
Outcome	positif (1)	58	58	0
	negatif (0)	100	0	100

4. 3 Mengklasfikasi Status Penyakit Diabetes dengan Metode *Decision Tree*

Dalam penelitian yang dilakukan memakai data yang diujikan berjumlah 158 data . Setelah informasi yang diperlukan tersedia, fitur yang diperlukan dari penelitian ini ditetapkan yaitu pemeriksaan pendahuluan terhadap pasien, meliputi: Kehamilan, glukosa, tekanan darah, ketebalan kulit, insulin, BMI, riwayat diabetes, usia dan hasil. Sekaligus diprediksi apakah kondisi pasien positif ataupun negatif, yakni variabel hasil. Berikut adalah hasil yang didapat setelah dilakukannya klasifikasi berdasarkan atribut di atas dan diolah menggunakan perhitungan manual C4.5. Kalkulasi yang didapatkan pada kolom dibawah berikut.

kolom entropi dan gain pencarian 1 adalah:

Entropy Data Pengujian

= ((-58/158) * IMLOG2(58/158) + (-100/158) * IMLOG2(100/158)) = **0.948**

Entropy pada atribut melahir*ancies)*

0 = ((-7/18) * IMLOG2(7/18) + (-11/18) * IMLOG2(11/18)) = 0.964

1 = ((7/25) * IMLOG2(7/25) + (-18/25) * IMLOG2(18/25)) = 0.855

2= ((-7/29) * IMLOG2(7/29) + (-22/29) * IMLOG2(22/29)) = 0.797

3 = ((-4/14) * IMLOG2(4/14) + (-10/14) * IMLOG2(10/14)) = 0.863

>3 = ((-33/72) * IMLOG2(33/72) + (-39/72) * IMLOG2(39/72)) = 0.995

Gain total yang didapat = (0.948) - (((18/158) * (0.964) + (((25/158) * (0.855)

+ (((29/158) * (0.863) + (((14/158) * (0.864) + (((72/158) * (0.995))) = **0.027**

Entropy total pada attribut glukose

Rendah=((-3/42) * IMLOG2(3/42)) + ((-39/42) * IMLOG(39/42))= 0.371

Normal = ((-19/63) * IMLOG2(19/63)) + ((-(44/63) * IMLOG2((44/63)) = 0.883

Tinggi = ((-36/53) * IMLOG2(36/53) + (-17/53) * IMLOG2(17/53)) = 0.905

Gain total = (0.948) - (((42/158) * (0.371) + ((63/158) * (0.883) + ((53/158)

* (0.905))) = **0.1939**

Entropy total pada atribut blood pressure

Rendah = ((-12/49) * IMLOG2(12/49)) + ((-37/49) * IMLOG2(37/49)) = 0.803

Normal = ((-20/53) * IMLOG2(20/53) + ((-33/53) * IMLOG2(33/53))= 0.956

Tinggi = ((-26/56)*IMLOG2(24/56) + ((-30/56) * IMLOG2(30/56)) = 0.996

Gain total = (0.948) - (((49/158 * (0.803) + ((53/158) * (0.956) + (56/158) *

(0.996))) = **0.0255**

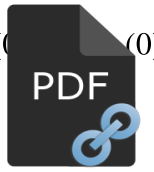
Entropy total pada attribut skin thicness

Rendah= ((-0/0) * IMLOG2(0/0) + ((-0/0) * IMLOG2(0/0)) =0

Normal = ((-8/52) * IMLOG2(8/52) + ((-44/52) * IMLOG2(44/52)) = 0.619

Tinggi = ((-50/103) * IMLOG2(50/103) + ((-56/103) * IMLOG2(56/103)) = 0.998

Gain total = (0.948) - ((((0) + ((52/158) * (0.619) + ((103/158) * (0.998)) = **0.0752**



Entropy total pada attribut insulin

Rendah = ((-43/128) * IMLOG2(43/128) + ((-85/128) * IMLOG2(85/128)) = 0.921

Normal = ((-10/16) * IMLOG2(10/16) + ((-6/16) * IMLOG2(6/16)) = 0.954

Tinggi = ((-5/14) * IMLOG2(4/14) + ((-9/14) * IMLOG2(9/14)) = 0.940

Gain total = (0.948) - ((128/158) * (0.921) + ((16/158) * (0.954) + ((14/158) * (0.940)) = **0.0224**

Entropy total pada attribut BMI

Rendah = ((-0/2) * IMLOG2(0/2) + ((-2/2) * IMLOG22/2)) = 0

Normal = ((-2/34) * IMLOG2(3/34) + ((-32/35) * IMLOG2(32/34)) = 0.323

Tinggi = ((-56/122) * IMLOG2(56/122) + ((-66/122) * IMLOG2(66/122)) = 0.995

Gain total = (0.948) - ((2/158) * (0.) + ((34/158) * (0.323)+ ((122/158) * (0.995)))) = **0.1106**

Entropy total pada attribut dpfr

Rendah = ((-0/0) * IMLOG2(0/0) + ((-0/0) * IMLOG2(0/0)) = 0

Tinggi = ((-58/158) * IMLOG2(58/158)+((-100/158)*IMLOG2(100/158)) = 0.948

Gain total = (0.948) - ((((0/158) * (0) + ((54/158) * (0.948)))) = **0.000**

Entropy total pada attribut age

Muda = ((-46/137) * IMLOG2(46/137) + ((-91/137) * IMLOG2(91/137)) = 0.921

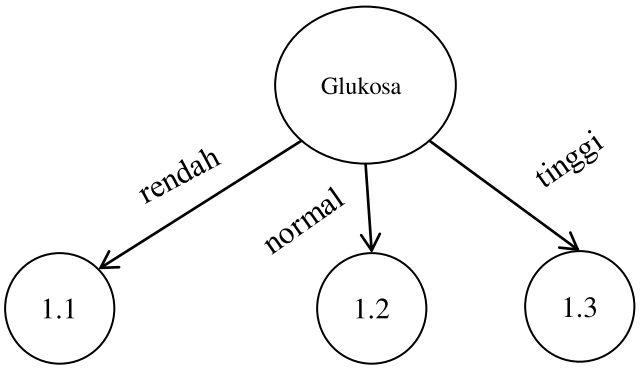
Tua = ((-15/24) * IMLOG2(15/24) + ((-9/24) * IMLOG2(9/24)) = 0.985

Gain total = (0.948) - (((137/158) * (0.921) + ((24/158) * (0.985))) = **0.019**

Tabel 4. 5 Klasifikasi Data Pengujian (Testing)

Variabel / Atribut		Jumlah	Jumlah Positif	Jumlah Negatif	Entropi	Gain
Pregnancies		18	7	11	0.964	0.027
	pri	25	7	18	0.855	
	mu	29	7	22	0.797	
	n	14	4	10	0.863	
	grand	72	33	39	0.995	
Glucose	rendah (<100 mg/dL)	42	3	39	0.371	0.1939
	normal (100-125 mg/dL)	63	19	44	0.883	
	tinggi (>125 mg/dL)	53	36	17	0.905	
Blood Pressure	rendah (<65 mmHg)	49	12	37	0.803	0.0255
	normal (65-75 mmHg)	53	20	33	0.956	
	tinggi (>75 mmHg)	56	26	30	0.996	
SkinThickness	rendah (<10 mm)	0	0	0	0	0.0752
	normal (10-25 mm)	52	8	44	0.619	
	tinggi (>25 mm)	106	50	56	0.998	
Insulin	rendah (<140 mg/dL)	128	43	85	0.921	0.0224
	normal (140-199 mg/dL)	16	10	6	0.954	
	tinggi (>199 mg/dL)	14	5	9	0.94	
BMI	rendah (<18.5)	2	0	2	0	0.1106
	normal (18.5-25)	34	2	32	0.323	
	tinggi (>25)	122	56	66	0.995	
Diabetes Pedigree Function	rendah (>2)	0	0	0	0	0
	tinggi (<=2)	158	58	100	0.948	
Age	muda (10-45)	137	46	91	0.921	0.0191
	tua (>45)	21	12	9	0.985	
Outcome	positif (1)	58	58	0	0	
	negatif (0)	100	0	100	0	

Setelah melakukan perhitungan entropi kemudian dilakukan Perhitungan *gain* sampai variabel terhitung keseluruhan. Dari hasil perhitungan tersebut dapat dilihat pada tabel bahwa hasil *entropy* dan *gain* tertinggi adalah pada atribut Glukosa yaitu sebesar 0.1939 ,maka dapat diartikan bahwa atribut Glukosa menjadi akar dari node ke-1. Pohon keputusan untuk node 1 dapat dilihat pada gambar berikut



Gambar 4. 2 Pohon Keputusan

Tahap berikutnya akan dilakukan perhitungan *entropy* dan juga *gain* untuk semua cabang dari variabel Glukosa sebagai akar pohon yang mana memiliki tiga katagori y gori Rendah dengan 42 kasus, Normal dengan 63 kasus dan Ting 53 kasus .

Selanjutnya melakukan perhitungan *entropy* dan juga *gain* dengan cara yang sama untuk seluruh cabang pohon keputusan. Proses perhitungan pohon keputusan dihentikan jika semua data sampel berada dalam kelas yang sama,tidak ada lagi atribut yang akan dilakukan partisi, atau tidak ada data sampel lagi yang akan diuji. Dari perhitungan yang dilakukan maka akan terbentuk sebuah model pengkondisian untuk menentukan klasifikasi pada penyakit diabetes.

4.3.1 Perancangan Arsitektur *Decision Tree*

Selanjutnya dilakukan perancangan arsitektur klasifikasi penyakit diabetes dengan model *Decision tree* C4.5 dengan menggunakan mesin learning google colabs dengan langkah sebagai berikut

1. Pustaka MLflow diinstal. MLflow sangat penting untuk melacak eksperimen, mengelola model, dan memantau siklus hidup Machine Learning. Setelah instalasi, berbagai pustaka Python diimpor untuk tugas-tugas spesifik, Manipulasi Data: pandas (sebagai pd) dan numpy (sebagai np) digunakan untuk mengelola dan menganalisis dataset, termasuk penanganan nilai yang hilang. Pemodelan dan Evaluasi: Dari sklear berfungsi membagi data menjadi set pelatihan dan pengujian, *Decision Tree* Classifier adalah algoritma klasifikasi inti yang akan digunakan. sklearn.metrics menyediakan berbagai metrik evaluasi seperti

accuracy_score, classification_report, confusion_matrix, roc_curve, dan auc untuk menilai performa model secara komprehensif. Visualisasi: matplotlib.pyplot (sebag dan seaborn (sebagai sns) diimpor untuk membuat plot dan vis data serta hasil model yang menarik. Integrasi dan Utilitas: google.colab.drive memungkinkan akses ke Google Drive (jika menggunakan Google Colab). Pipeline dari sklearn.pipeline membantu membangun alur kerja yang terintegrasi. Standard Scaler dari sklearn.preprocessing digunakan untuk standardisasi fitur. joblib berguna untuk menyimpan dan memuat model terlatih. mlflow.sklearn menyediakan integrasi khusus antara MLflow dan scikit-learn. Terakhir, json diimpor untuk menangani data dalam format JSON, mungkin untuk menyimpan metrik atau konfigurasi.

2. Melakukan mounting Google Drive ke lingkungan Google Colab agar notebook dapat mengakses file data dan menyimpan hasil proyek yang berada di Drive, diawali dengan pesan indikator "Mounting Google Drive..." dan diakhiri dengan konfirmasi "Google Drive mounted successfully." setelah otorisasi pengguna.

```
print("Mounting Google Drive...")
drive.mount('/content/drive') # Mengaitkan (mount) Google
Drive ke lingkungan Colab agar file dapat diakses.
print("Google Drive mounted successfully.")
print("Google Drive mounted successfully.")
```

Gambar 4. 2 Mounting Google Colab

3. Menentukan path ke file dataset

```
file_path = '/content/drive/MyDrive/diabetes_shuffled.xlsx'
```

Gambar 4. 3 Menentukan Path pada Data Set

4. Melakukan pembacaan dan inspeksi awal dataset yang disimpan di Google Drive. Pertama memberikan notifikasi proses. Kemudian membaca file Excel dari lokasi yang [illegible]an oleh file_path ke dalam Data Frame pandas bernama data. S[illegible], serangkaian perintah print dan fungsi Data Frame digunakan untuk inspeksi awal: data.info() menampilkan ringkasan Data Frame (jumlah entri non-null, tipe data, penggunaan memori), data.head() menunjukkan lima baris pertama untuk memahami struktur data, data.describe() menyajikan statistik deskriptif kolom numerik (rata-rata, standar deviasi, min, max, kuartil), data.isnull().sum() menghitung dan menampilkan jumlah nilai yang hilang di setiap kolom, dan data.dtypes menampilkan tipe data setiap kolom. Proses ini esensial untuk memahami karakteristik dataset sebelum melanjutkan ke preprocessing atau pemodelan.

```
Loading dataset and performing initial inspection...
Informasi Dataset Awal:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Glucose                   768 non-null    int64
 1   BloodPressure             768 non-null    int64
 2   SkinThickness             768 non-null    int64
 3   Insulin                   768 non-null    int64
 4   BMI                       768 non-null    float64
 5   DiabetesPedigreeFunction  768 non-null    float64
 6   Age                       768 non-null    int64
 7   Outcome                   768 non-null    int64
 8   Pregnancies               768 non-null    int64
dtypes: float64(2), int64(7)
```

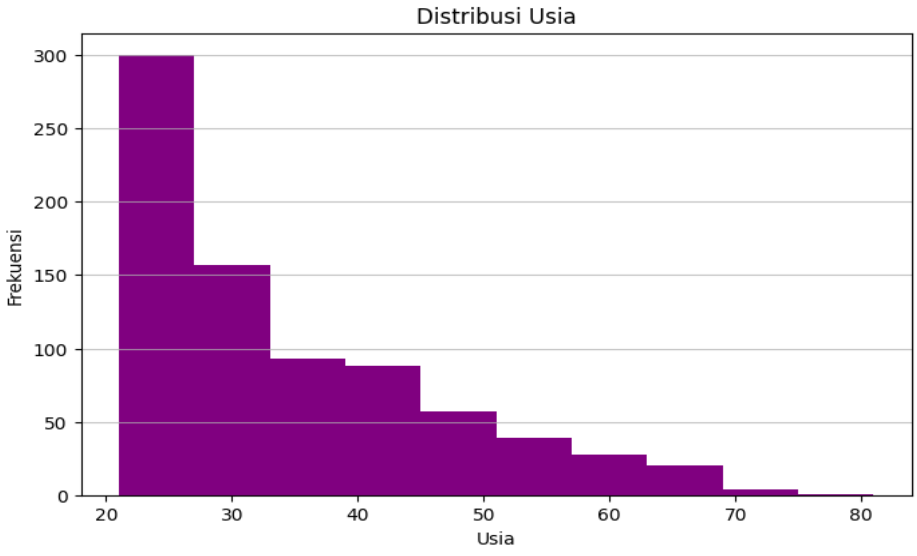
Gambar 4. 4 Pembacaan dan Seleksi Awal

5. Tahapan penting dalam data preprocessing untuk menangani nilai nol (0) dan nilai hilang (NaN) dalam dataset. Pertama, sebagai indikator proses. Berdasarkan pemahaman domain penyakit diabetes, nilai 0 pada kolom-kolom spesifik seperti 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', dan 'BMI' dianggap sebagai nilai hilang yang tidak valid secara medis. Oleh karena itu, digunakan untuk mengiterasi setiap kolom tersebut dan

menggantinya semua nilai 0 di. Setelah penggantian ini, baris data.fillna. Tahap ini krusial untuk memastikan kualitas data sebelum pemodelan.

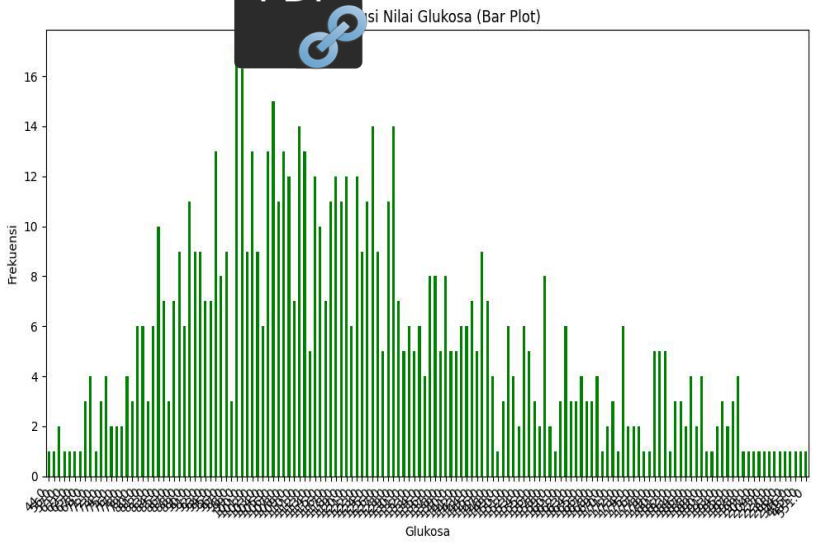
Gambar 4. 5 Memastikan Kualitas Data

6. Melakukan visualisasi distribusi usia dalam dataset menggunakan histogram: mengatur ukuran plot; membuat histogram kolom 'Age' dengan warna ungu; masing-masing memberi judul, label sumbu x, dan label sumbu y; menambahkan grid pada sumbu y; dan menampilkan histogram tersebut, membantu memahami sebaran usia pada dataset



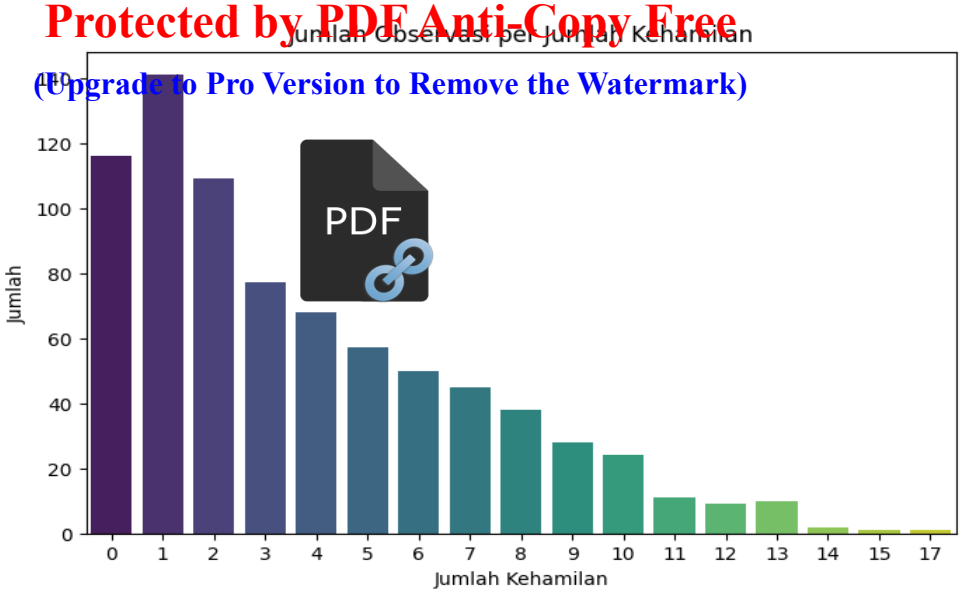
Gambar 4. 6 Visusalisasi Distribusi Data Usia

7. Menghasilkan bar plot distribusi nilai glukosa: dengan menghitung frekuensi setiap nilai 'Glucose', mengurutkannya, lalu membuat bar plot berwarna hijau;



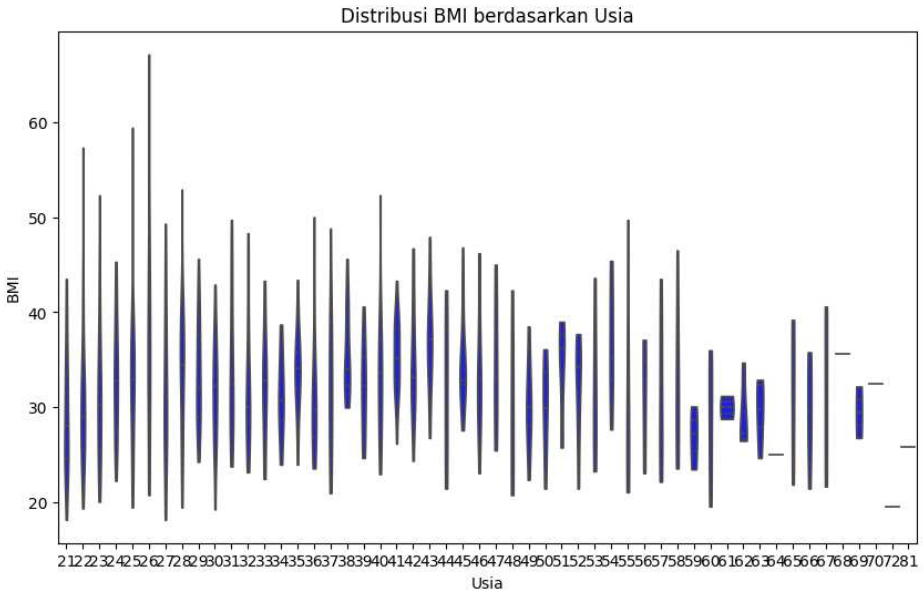
Gambar 4. 7 Distribusi Nilai Glukosa

8. Membuat count plot untuk visualisasi jumlah observasi berdasarkan jumlah kehamilan mengatur ukuran figure; menggunakan seaborn untuk menghasilkan count plot dari kolom mengatur judul plot serta label sumbu X dan Y; terakhir, plt.show() menampilkan plot tersebut, memberikan gambaran visual mengenai frekuensi setiap kategori jumlah kehamilan dalam dataset.



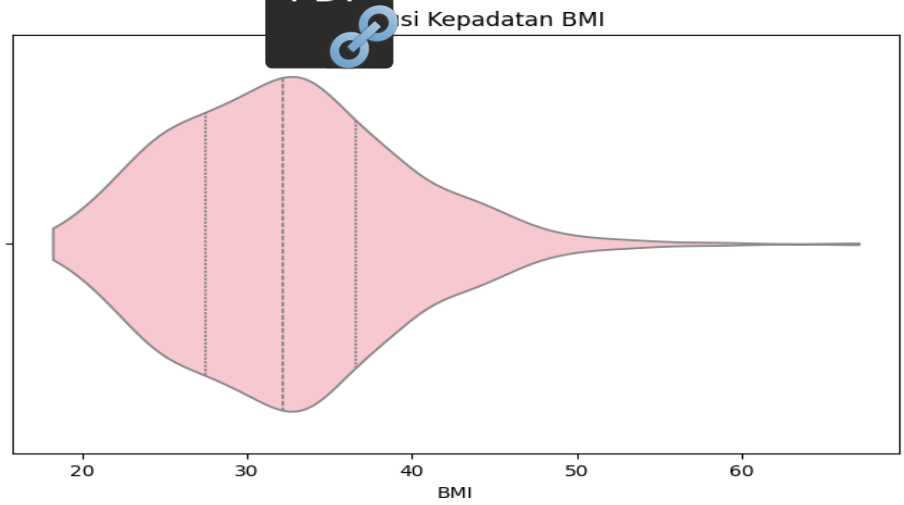
Gambar 4. 8 Jumlah Observasi per kehamilan

9. Menghasilkan violin plot untuk memvisualisasikan distribusi Body Mass Index (BMI) berdasarkan kelompok usia: yang berguna untuk melihat kepadatan probabilitas dan sebaran data BMI pada berbagai kategori usia.



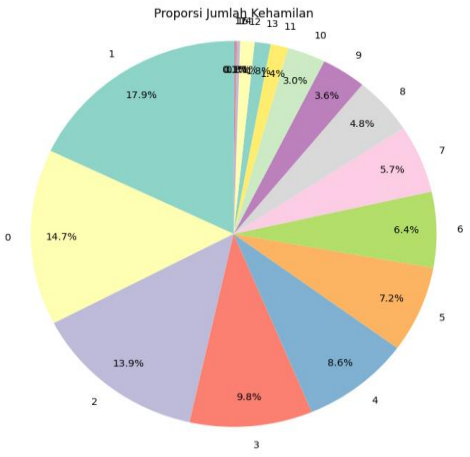
Gambar 4. 9 Distribusi BMI Berdasarkan Usia

10 Menghasilkan violin plot untuk memvisualisasikan distribusi kepadatan Body Mass Index (BMI) yang berguna untuk melihat bentuk distribusi, kepadatan, dan sebara MI secara sekilas.



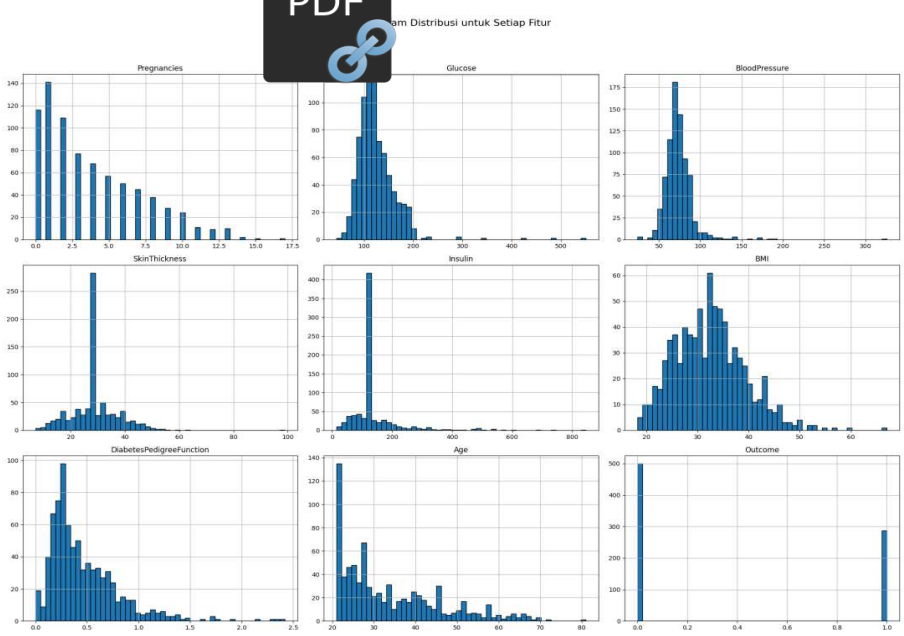
Gambar 4. 10 Distribusi Kepadatan BMI

11. Menghasilkan pie chart yang menunjukkan proporsi jumlah kehamilan: mengatur ukuran figure index mendapatkan data frekuensi dan label kategori kehamilan; colors membuat pie chart dengan label, warna, persentase otomatis, sudut awal, dan posisi persentase mengatur judu memastikan bentuk lingkaran sempurna; dan menampilkan plot tersebut.



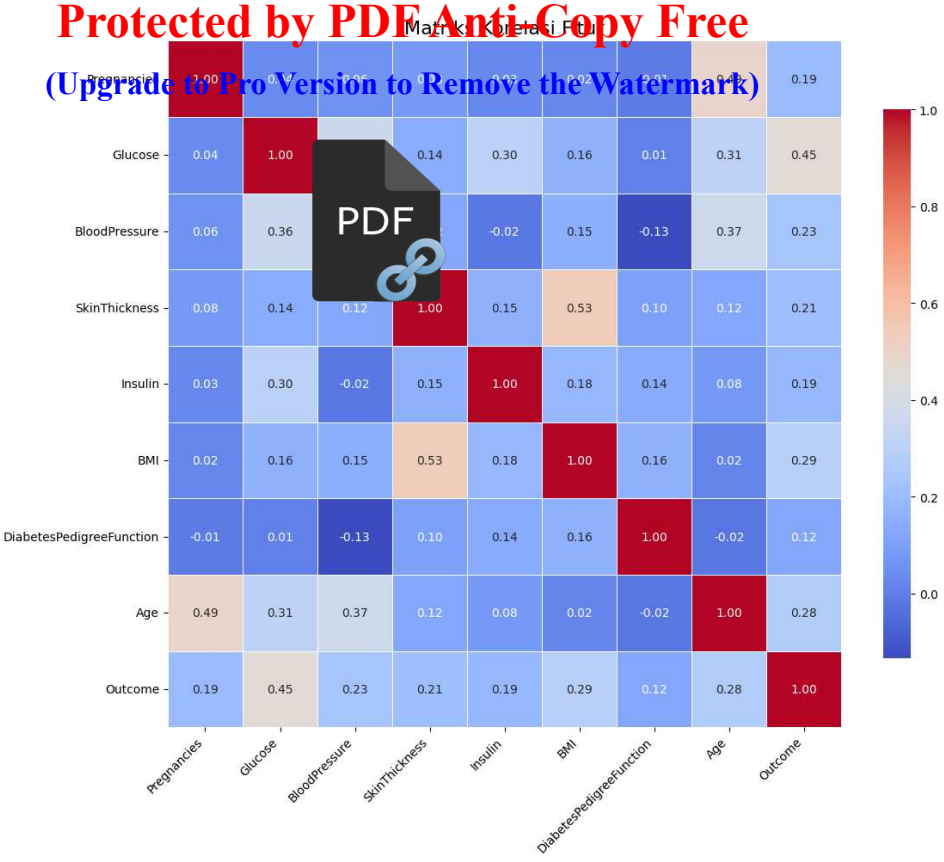
Gambar 4. 3 Proporsi Jumlah Kehamilan

12 Menghasilkan histogram distribusi untuk setiap fitur numerik dalam dataset:



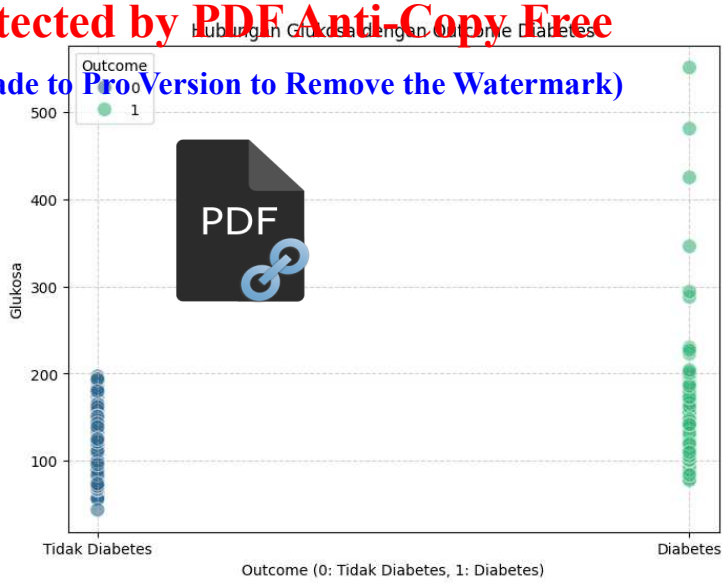
Gambar 4. 4 Histogram Distribusi Data aset

12. Menganalisis dan memvisualisasikan matriks korelasi antar fitur dalam dataset menghitung matriks korelasi; menampilkan korelasi fitur 'Outcome' dengan fitur lain secara menurun, mengatur ukuran figure; membuat heatmap dari matriks korelasi dengan anotasi



Gambar 4. 5 Matrik Korelasi Fitur

13. Menghasilkan scatter plot untuk memvisualisasikan hubungan antara tingkat glukosa dan outcome diabetes mengatur ukuran figure; membuat scatter plot dengan 'Outcome' pada sumbu X dan 'Glucose' pada sumbu Y, menggunakan warna berbeda untuk setiap outcome, transparansi 0.6, dan ukuran titik 100; menampilkan plot tersebut, membantu memahami pola glukosa pada pasien diabetes dan non-diabetes.



Gambar 4. 6 Visualisasi Hubungan Glukosa dan Outcome Diabetes

14. Melakukan pemisahan fitur dan variabel target dari dataset: memberikan notifikasi proses, membuat Data Frame X yang berisi semua kolom fitur dengan menghapus kolom 'Outcome' dari data yang merupakan langkah esensial sebelum melatih model *Machine Learning*.

```
print("\nSeparating features (X) and target (y)...")
X = data.drop('Outcome', axis=1) # Memisahkan fitur (semua
kolom kecuali 'Outcome') ke dalam variabel X.
y = data['Outcome'] # Memisahkan variabel target ('Outcome')
ke dalam variabel y.
print("Features and target separated.")
```

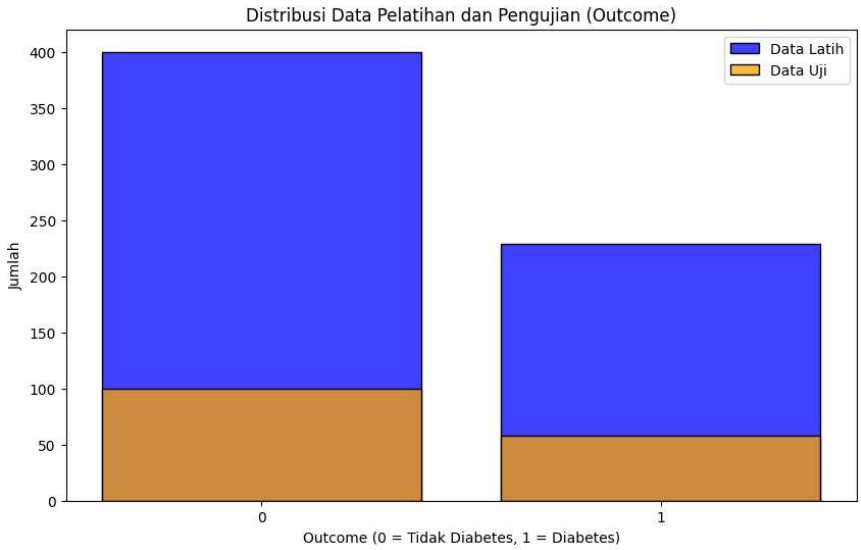
4. 7 Code Pemisahan Fitur dan Target dari Data Base

15. Melakukan pembagian dataset menjadi set pelatihan dan pengujian: menetapkan 20% data untuk pengujian, random_etap sama di kedua set (pelatihan dan pengujian), yang sangat penting untuk dataset dengan kelas yang tidak seimbang.

```
print("\nSplitting dataset into training and testing sets...")
# Membagi X dan y menjadi set pelatihan dan pengujian.
# test_size=0.2 be      20% data akan digunakan untuk
pengujian.
# random_state=42      an hasil pembagian konsisten
(reproducible).
# stratify=y memastikan proporsi kelas 'Outcome' sama di
data pelatihan dan pengujian.
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42, stratify=y)
```

Gambar 4. 8 Code Pembagian Dataset

16. Memvisualisasikan distribusi variabel target ('Outcome') pada data pelatihan dan pengujian menggunakan histogram



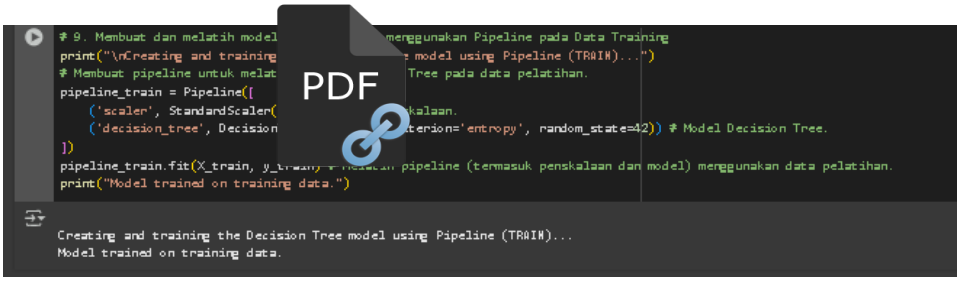
Gambar 4. 9 Visusalisasi Distribusi Target

17. Melakukan validasi silang (cross-validation) 5-fold pada seluruh dataset (x, y) untuk mendapatkan estimasi kinerja model yang lebih robust:

```
Performing Cross-Validation...
Akurasi Rata-Rata dari Cross-Validation (5-fold): 0.7129 ± 0.0281
Cross-Validation completed.
```

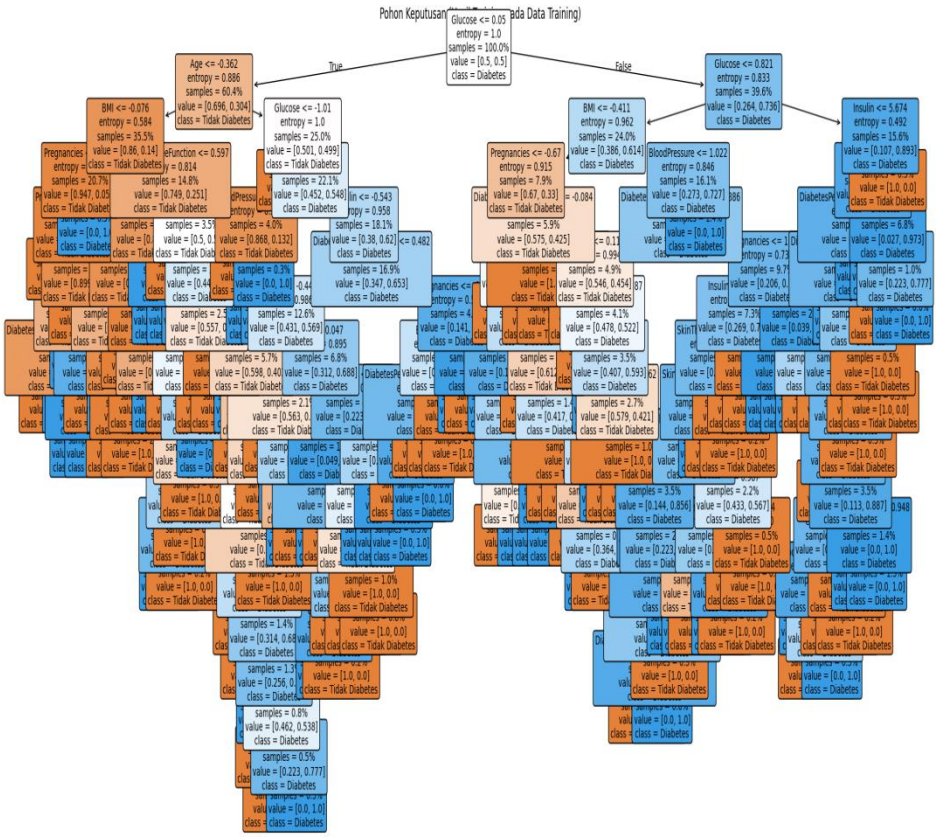
Gambar 4. 10 Code Validasi 5-Fold

18. Membuat dan melatih model *Decision Tree* menggunakan pipeline pada data pelatihan



Gambar 4. 11 Code Melatih Model *Decision Tree*

19. Memvisualisasikan model *Decision Tree* yang telah dilatih menggunakan data pelatihan menampilkan visualisasi pohon keputusan secara rinci, membantu memahami logika keputusan model.



Gambar 4. 20 Visualisasi Model

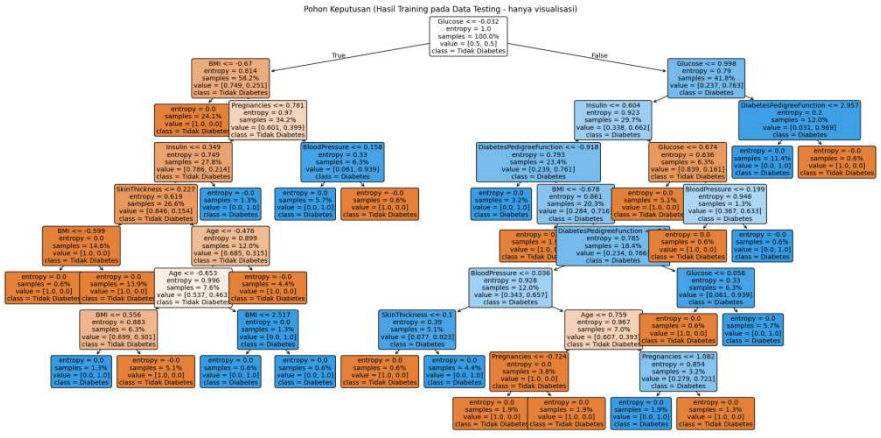
20. Membuat dan melatih model *Decision Tree* menggunakan pipeline khusus pada data pengujian , bukan untuk evaluasi utama, melainkan hanya untuk tujuan v i atau analisis pohon yang terbentuk dari subset data penguji roti bahwa hasilnya ditujukan untuk inspeksi pohon dari data yang belum pernah dilihat model utama.

```
# ---- Membuat dan melatih model Decision Tree pada Data Testing ----
print("\nTraining Decision Tree hanya pada Data Testing (untuk visualisasi/analisis)...")
# Membuat pipeline baru khusus untuk melatih model pada data pengujian.
# Ini dilakukan untuk tujuan visualisasi pohon yang terbentuk hanya dari data pengujian,
# BUKAN sebagai bagian dari alur evaluasi model utama.
pipeline_test = Pipeline([
    ('scaler', StandardScaler()), # Langkah penskalaan.
    ('decision_tree', DecisionTreeClassifier(criterion='entropy', random_state=42)) # Model Decision Tree.
])
pipeline_test.fit(X_test, y_test) # Melatih pipeline pada data pengujian.
print("Model trained on test data (for visualization only).")
```

```
Training Decision Tree hanya pada Data Testing (untuk visualisasi/analisis)...
Model trained on test data (for visualization only).
```

Gambar 4. 12 Code latihan model

21. Memvisualisasikan model *Decision Tree* yang sebelumnya telah dilatih khusus menggunakan data pengujian (bukan bagian dari alur evaluasi utama, melainkan hanya untuk inspeksi) memberikan gambaran bagaimana model berlogika pada data yang belum pernah dilihat model utama.



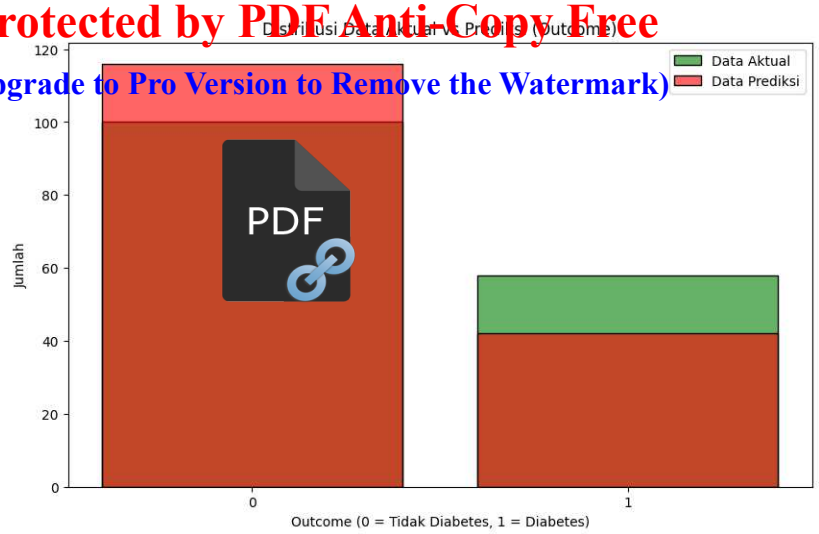
Gambar 4. 13 Visuaisasi Model Telah dilatih

22. Mengevaluasi kinerja model *Decision Tree* yang sebelumnya dilatih menggunakan data pelatihan), kemudian diterapkan untuk memprediksi outcome pada dat[illegible]ujian (X_test). memberikan notifikasi menghasilkan prediks[illegible]berdasarkan X_test, menghitung akurasi model dengan membandingkan y_pred dengan y_test yang aktual; menampilkan nilai akurasi terhitung; dan menyajikan laporan klasifikasi lengkap yang meliputi presisi, recall, f1-score, dan support untuk setiap kelas, memberikan gambaran detail tentang performa model pada data yang belum pernah dilihat sebelumnya.

```
Evaluating the model trained with TRAIN set...

Akurasi pada Data Uji: 0.7342

Laporan Klasifikasi:
              precision    recall  f1-score   support

           0       0.75      0.87      0.81       100
           1       0.69      0.50      0.58        58

    accuracy                           0.73       158
   macro avg       0.72      0.69      0.69       158
weighted avg       0.73      0.73      0.72       158
```

Gambar 4. 14 Code Evaluasi Kinerja Model *Decision Tree* Data Pelatihan

23. Menganalisis dan memvisualisasikan distribusi outcome aktual pada data pengujian: pertama, kode mencetak jumlah aktual kasus non-diabetes dan diabetes dari y_test, diikuti dengan jumlah prediksi non-diabetes dan diabetes dari y_pred, untuk memberikan gambaran kuantitatif perbandingan; kemudian, mengatur ukuran figur membuat histogram distribusi aktual membuat histogram distribusi prediksi (merah), keduanya tanpa estimasi kepadatan kernel (KDE) dan dengan bins diskrit;



Gambar 4. 15 Distribusi Data Aktual dan Prediksi

24. Menghitung dan menampilkan matriks kebingungan untuk mengevaluasi kinerja model memberikan notifikasi; menghitung matriks kebingungan dengan membandingkan nilai aktual (y_test) dengan nilai prediksi (y_pred), kemudian menampilkan matriks tersebut, yang merupakan tabel ringkasan kinerja klasifikasi yang menunjukkan jumlah True Positives, True Negatives, False Positives, dan False Negatives.

```
Matriks Kebingungan
print("\nCalculating Confusion Matrix...")
conf_matrix = confusion_matrix(y_test, y_pred) # Menghitung
matriks kebingungan.
print('\nMatriks K
```

Matriks Kebingungan:

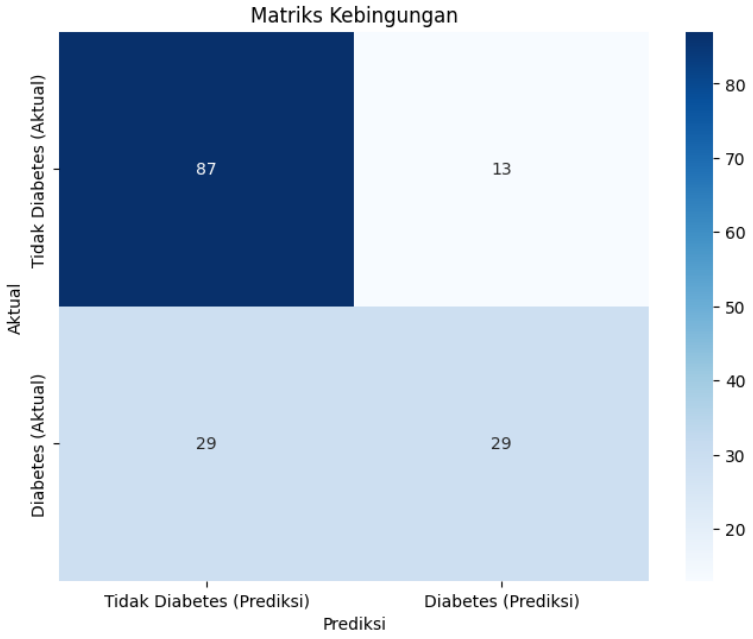
[[87 13]

[29 29]]

Gambar 4. 16 Code Menampilkan Matriks Kebingungan

25. Memvisualisasikan matriks kebingungan yang telah dihitung: menggunakan seaborn untuk membuat heatmap dari conf_matrix,

dengan anotasi angka serta label sumbu X dan Y yang informatif (xticklabels, yticklabels) untuk "Tidak Diabetes" dan "Diabetes" pada prediksi dan aktual; [illegible]i label umum pada sumbu; menetapkan judul plot; menam[illegible]ualisasi tersebut, yang memudahkan interpretasi kinerja klasifikasi model.



Gambar 4. 17 Visualisasikan Mmatriks Kebingungan

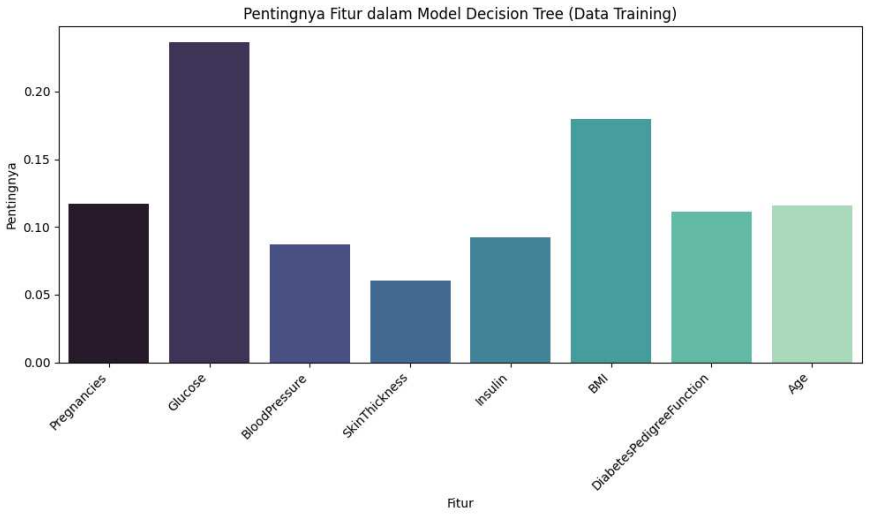
26. Menghitung dan menyajikan matriks kebingungan beserta presisi dan recall untuk setiap kelas dalam format tabel yang mudah dibaca. Pertama, ia menghitung confusion_matrix dari y_test dan y_pred, mengekstraksi nilai True Positives (TP), True Negatives (TN), False Positives (FP), dan False Negatives (FN). Selanjutnya, classification_report digunakan untuk mendapatkan presisi dan recall per kelas. Nilai-nilai ini kemudian disusun ke dalam DataFrame pandas secara manual, mengonversi presisi dan recall ke format persentase. Terakhir, tabulate digunakan untuk mencetak DataFrame tersebut

sebagai tabel yang diformat dengan baik, memvisualisasikan performa model dengan jelas, termasuk perbandingan prediksi "Iya" dan "Tidak" dengan nilai aktual, s si dan recall masing-masing kelas.

```
--- Output Tabel Matriks K ngan Presisi dan Recall dalam Persen ---
|             | TRUE (iya  idak)   |   True (Tidak).1 | Class Precision   |
|-------------|--------------|----------------|------------------|-------------------|
| Pred. Iya   | 29           | 13             |                0 | 69.05%            |
| Pred. Tidak | 29           | 87             |                0 | 75.00%            |
| Pred. Tidak | 0            | 0              |                0 | 0.00%             |
| Class Recall | 50.00%      | 87.00%         |                0 | 0.00%             |
```

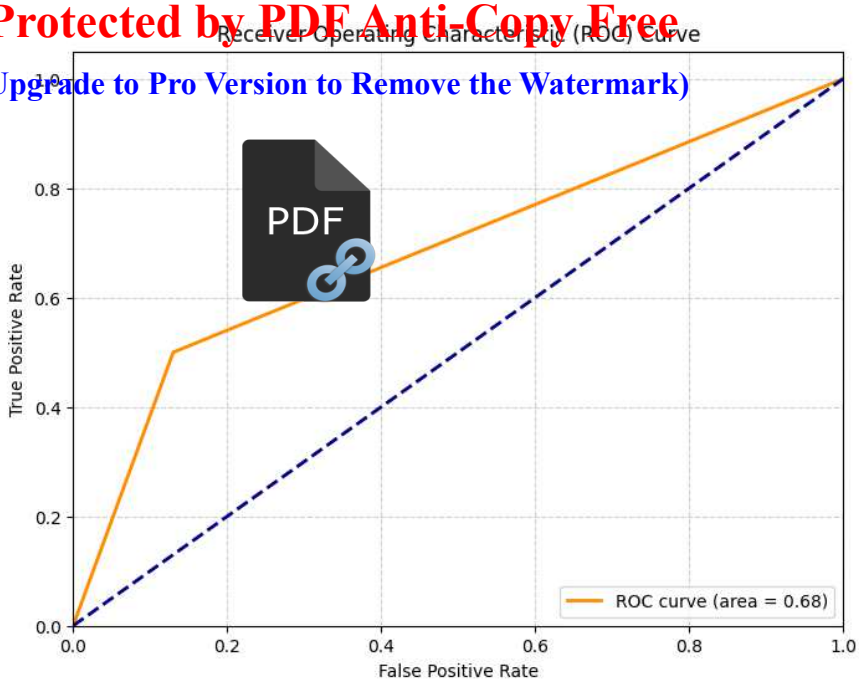
Gambar 4. 18 Code Menghitung Matriks Kebingungan , Presisi dan Recall

27. Menghitung dan memvisualisasikan tingkat kepentingan (feature_importances_) setiap fitur dalam model Decision Tree yang dilatih pada data pelatihan: yang membantu mengidentifikasi fitur mana yang paling berpengaruh dalam prediksi model.



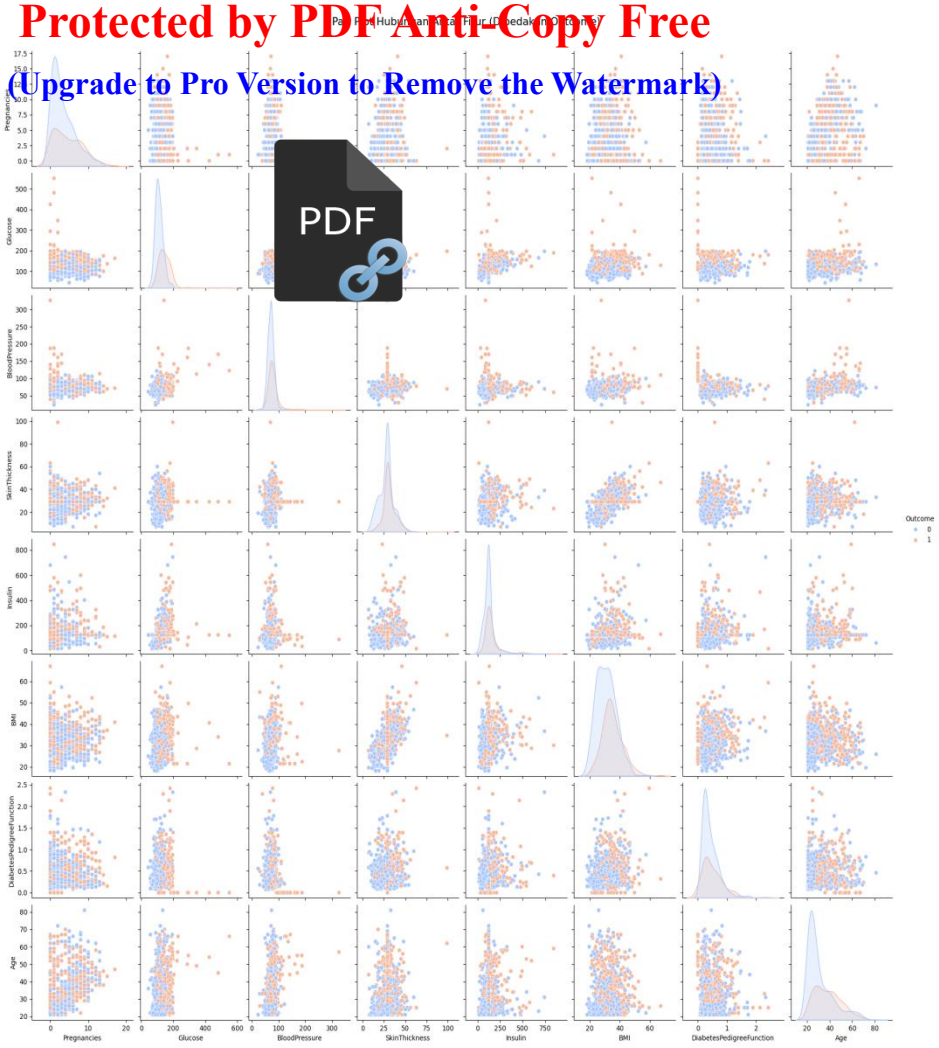
Gambar 4. 19 Code Visualisasi Tingkat Kepentingan Fitur dalam *Decision Tree*

28. Menghasilkan *Receiver Operating Characteristic (ROC) Curve* dan menghitung Area Under Curve (AUC) untuk mengevaluasi kinerja klasifikasi modelyang merupakan metrik penting untuk menilai kemampuan model membedakan antara kelas positif dan negatif.



Gambar 4. 209 Receiver Operating Characteristic (ROC) Curve

29. Menyimpan *pipeline* model yang telah dilatih (termasuk *scaler* dan model *Decision Tree*) ke dalam file
30. Menghasilkan *pair plot* untuk memvisualisasikan hubungan antar setiap pasangan fitur dalam dataset, dibedakan berdasarkan *outcome* diabetes:



Gambar 4.30 Pair Plot

31. Mengimplementasikan pemantauan performa model *Machine Learning* menggunakan MLflow dengan memulai sebuah *run* untuk mencatat berbagai aspek eksperimen. memungkinkan pelacakan komprehensif terhadap konfigurasi dan hasil model.

32. Mendefinisikan fungsi entropy untuk mengukur ketidakmurnian distribusi kelas dan information_gain untuk menghitung reduksi entropi yang dicapai oleh pembagian data berdasarkan atribut. Selanjutnya, fungsi buat_tabel_analisis mengaplikasikan skema kategorisasi atribut numerik yang telah didefinisikan (kategori), menghitung jumlah sampel

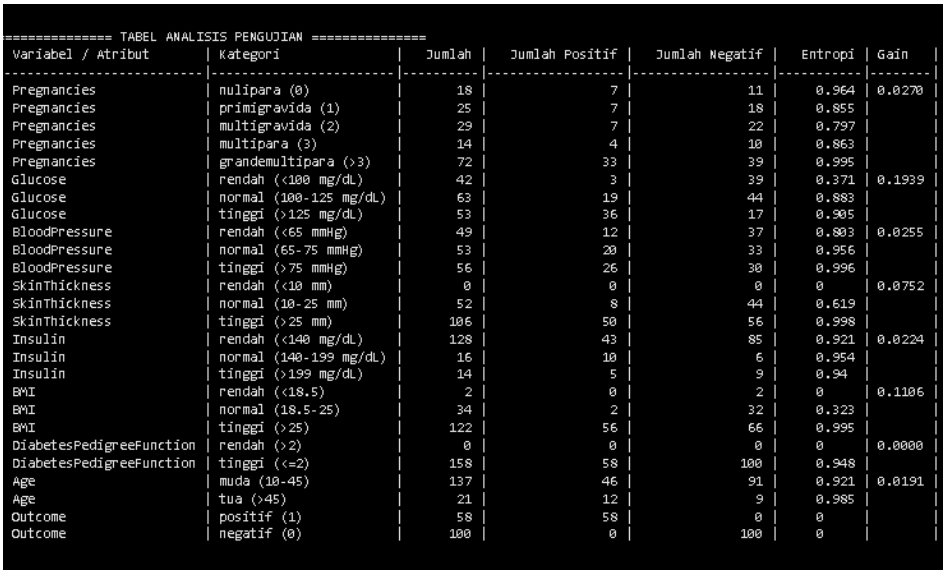
positif/negatif serta entropi untuk setiap kategori per atribut dalam dataset (pelatihan, pengujian, atau gabungan), menghitung Information Gain untuk setiap [PDF] dan kemudian mengompilasi serta menampilkan semua [PDF]sis tersebut dalam format tabel yang rapi menggunakan pandas dan tabulate, memberikan wawasan mendalam mengenai relevansi dan kekuatan prediktif setiap fitur terhadap variabel target dalam konteks *Decision Tree*.

33. Memanggil fungsi buat_tabel_analisis yang telah didefinisikan sebelumnya, meneruskan fitur (X_train) dan target (y_train) dari set data pelatihan, serta label "Pelatihan", untuk menghasilkan dan menampilkan tabel analisis lengkap yang mencakup jumlah kategori, jumlah sampel positif dan negatif, entropi, serta *Information Gain* untuk setiap fitur, yang khusus dihitung berdasarkan karakteristik data pelatihan.

```
============== TABEL ANALISIS PELATIHAN ==============
| Variabel / Atribut       | Kategori              |   Jumlah |   Jumlah Positif |   Jumlah Negatif |   Entropi | Gain
|--------------------------|-----------------------|----------|------------------|------------------|-----------|-------
| Pregnancies              | nulipara (0)          |       98 |               36 |               62 |     0.949 | 0.0272
| Pregnancies              | primigravida (1)      |      116 |               28 |               88 |     0.797 |
| Pregnancies              | multigravida (2)      |       80 |               18 |               62 |     0.769 |
| Pregnancies              | multipara (3)         |       63 |               25 |               38 |     0.969 |
| Pregnancies              | grandemultipara (>3)  |      272 |              122 |              150 |     0.992 |
| Glucose                  | rendah (<100 mg/dL)   |      150 |               11 |              139 |     0.378 | 0.1518
| Glucose                  | normal (100-125 mg/dL)|      221 |               64 |              157 |     0.868 |
| Glucose                  | tinggi (>125 mg/dL)   |      258 |              154 |              104 |     0.973 |
| BloodPressure            | rendah (<65 mmHg)     |      152 |               34 |              118 |     0.767 | 0.0297
| BloodPressure            | normal (65-75 mmHg)   |      225 |               77 |              148 |     0.927 |
| BloodPressure            | tinggi (>75 mmHg)     |      252 |              118 |              134 |     0.997 |
| SkinThickness            | rendah (<10 mm)       |        4 |                1 |                3 |     0.811 | 0.0267
| SkinThickness            | normal (10-25 mm)     |      144 |               29 |              115 |     0.725 |
| SkinThickness            | tinggi (>25 mm)       |      481 |              199 |              282 |     0.978 |
| Insulin                  | rendah (<140 mg/dL)   |      481 |              152 |              329 |     0.9   | 0.0243
| Insulin                  | normal (140-199 mg/dL)|       72 |               34 |               38 |     0.998 |
| Insulin                  | tinggi (>199 mg/dL)   |       76 |               43 |               33 |     0.987 |
| BMI                      | rendah (<18.5)        |        2 |                0 |                2 |     0     | 0.0244
| BMI                      | normal (18.5-25)      |       82 |               13 |               69 |     0.631 |
| BMI                      | tinggi (>25)          |      545 |              216 |              329 |     0.969 |
| DiabetesPedigreeFunction | rendah (>2)           |        4 |                3 |                1 |     0.811 | 0.0029
| DiabetesPedigreeFunction | tinggi (<=2)          |      625 |              226 |              399 |     0.944 |
| Age                      | muda (10-45)          |      518 |              169 |              349 |     0.911 | 0.0201
| Age                      | tua (>45)             |      111 |               60 |               51 |     0.995 |
| Outcome                  | positif (1)           |      229 |              229 |                0 |     0     |
| Outcome                  | negatif (0)           |      400 |                0 |              400 |     0     |
```

Gambar 4. 21 Analisis Lengkap Karakteriktik Pelatihan

34. Menanggil fungsi buat_tabel_analisis yang telah didefinisikan sebelumnya, meneruskan fitur (X_test) dan target (y_test) dari set data pengujian, serta l Pengujian", untuk menghasilkan dan menampilkan tabel a ngkap yang mencakup jumlah kategori, jumlah sampel positif dan negatif, entropi, serta *Information Gain* untuk setiap fitur, yang khusus dihitung berdasarkan karakteristik data pengujian.



========== TABEL ANALISIS PENGUJIAN ==========

Variabel / Atribut	Kategori	Jumlah	Jumlah Positif	Jumlah Negatif	Entropi	Gain
Pregnancies	nulipara (0)	18	7	11	0.964	0.0270
Pregnancies	primigravida (1)	25	7	18	0.855	
Pregnancies	multigravida (2)	29	7	22	0.797	
Pregnancies	multipara (3)	14	4	10	0.863	
Pregnancies	grandemultipara (>3)	72	33	39	0.995	
Glucose	rendah (<100 mg/dL)	42	3	39	0.371	0.1939
Glucose	normal (100-125 mg/dL)	63	19	44	0.883	
Glucose	tinggi (>125 mg/dL)	53	36	17	0.905	
BloodPressure	rendah (<65 mmHg)	49	12	37	0.803	0.0255
BloodPressure	normal (65-75 mmHg)	53	20	33	0.956	
BloodPressure	tinggi (>75 mmHg)	56	26	30	0.996	
SkinThickness	rendah (<10 mm)	0	0	0	0	0.0752
SkinThickness	normal (10-25 mm)	52	8	44	0.619	
SkinThickness	tinggi (>25 mm)	106	50	56	0.998	
Insulin	rendah (<140 mg/dL)	128	43	85	0.921	0.0224
Insulin	normal (140-199 mg/dL)	16	10	6	0.954	
Insulin	tinggi (>199 mg/dL)	14	5	9	0.94	
BMI	rendah (<18.5)	2	0	2	0	0.1106
BMI	normal (18.5-25)	34	2	32	0.323	
BMI	tinggi (>25)	122	56	66	0.995	
DiabetesPedigreeFunction	rendah (>2)	0	0	0	0	0.0000
DiabetesPedigreeFunction	tinggi (<=2)	158	58	100	0.948	
Age	muda (10-45)	137	46	91	0.921	0.0191
Age	tua (>45)	21	12	9	0.985	
Outcome	positif (1)	58	58	0	0	
Outcome	negatif (0)	100	0	100	0	

Gambar 4. 22 Analisis Lengkap Karakteriktik Pengujian

35. Melakukan beberapa operasi manajemen data dan *preprocessing*: pertama, ia menyimpan tiga tabel analisis (tabel_pelatihan, tabel_pengujian, tabel_final) ke dalam file CSV di Google Drive; kemudian, ia memuat ulang dataset utama, membersihkan nilai nol dan hilang dengan imputasi median seperti sebelumnya; selanjutnya, dataset tersebut dibagi menjadi tiga subset berdasarkan kategori kadar glukosa (rendah, normal, tinggi) menggunakan batas nilai tertentu; dan terakhir, setiap subset data glukosa ini disimpan ke dalam file Excel terpisah di

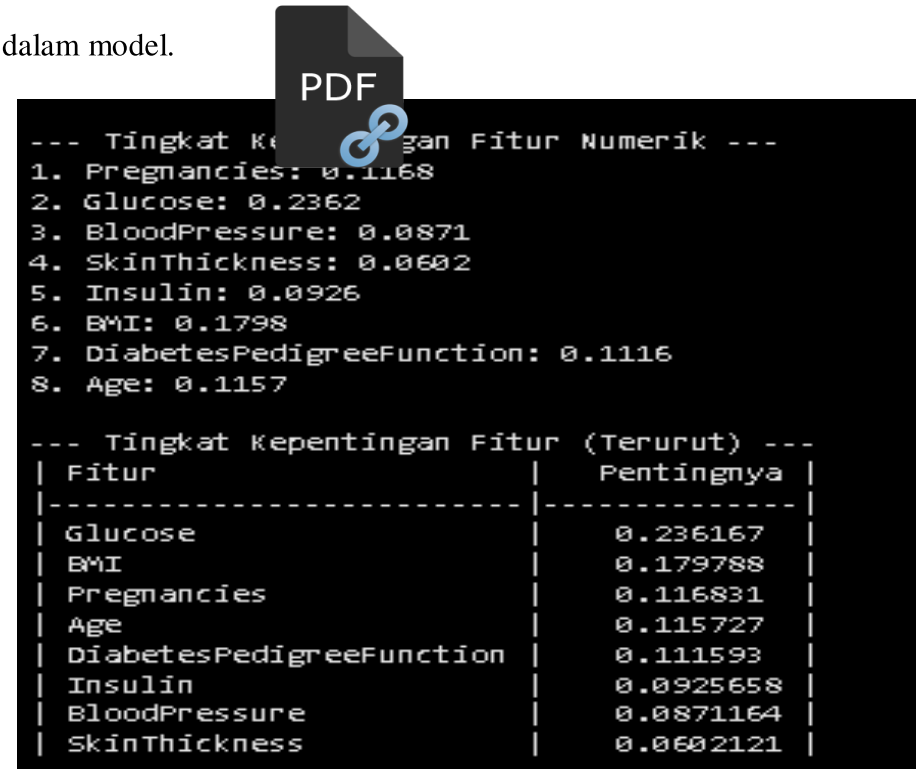
Google Drive, disertai pencetakan jumlah entri dalam setiap kategori untuk konfirmasi.

36. Mendefinisikan kateg[illegible] mengategorikan fitur numerik (misalnya, Pregnancies, Glucos[illegible] menjadi kategori diskrit berdasarkan rentang nilai spesifik, di mana setiap kategori memiliki nama deskriptif dan fungsi lambda yang menentukan kondisi keanggotaan. Selain itu, fungsi entropy diimplementasikan untuk menghitung entropi Shannon, yang mengukur tingkat ketidakmurnian distribusi dua kelas (positif/negatif) dalam suatu subset data, dengan nilai 0 jika subset murni atau kosong, yang merupakan komponen dasar untuk perhitungan *Information Gain* dalam algoritma pohon keputusan.

37. Mengimpor tabulate dan kemudian memanggil fungsi buat_tabel_analisis tiga kali untuk menghasilkan tabel analisis terpisah yang merinci entropi dan *information gain* fitur-fitur pada subset data pasien dengan kadar glukosa rendah, normal, dan tinggi. Setelah tabel tabel_rendah, tabel_normal, dan tabel_tinggi tersebut dibuat, masing-masing disimpan ke dalam file Excel (.xlsx) di Google Drive, disertai pesan konfirmasi, yang memfasilitasi analisis terperinci terhadap pentingnya fitur dalam konteks subpopulasi glukosa yang berbeda.

38. Menampilkan tingkat kepentingan fitur numerik yang telah dihitung dari model *Decision Tree* dalam dua format: pertama, sebagai daftar berurut sederhana yang mencetak setiap nama fitur dan nilai kepentingannya dengan empat desimal; dan kedua, sebagai DataFrame pandas yang diurutkan secara menurun berdasarkan nilai kepentingan, kemudian

dicetak dalam format tabel yang rapi menggunakan tabulate, memfasilitasi identifikasi cepat fitur-fitur yang paling berpengaruh dalam model.



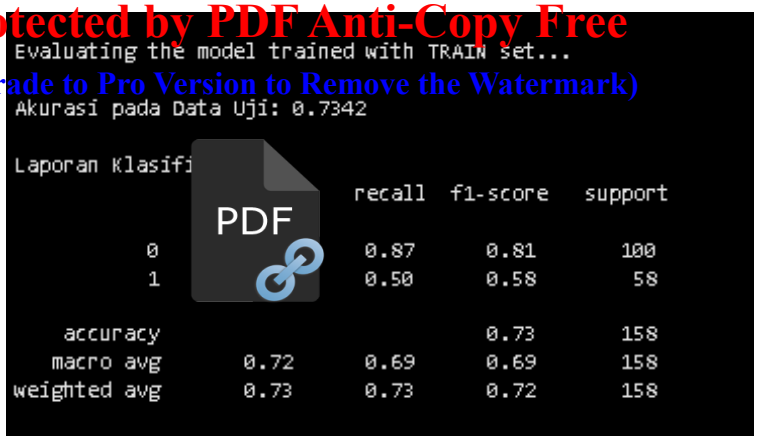
Gambar 4. 23 Tingkat Kepentingan Fitur Numerik

39. mengimplementasikan modul pengujian model interaktif yang meminta pengguna memasukkan nilai untuk setiap fitur. Hasil prediksi dan probabilitas kemudian ditampilkan kepada pengguna, memungkinkan pengujian model secara langsung dan *real-time* dengan data baru.

4.3.2 Evaluasi *Decision Tree*

1. Training Data

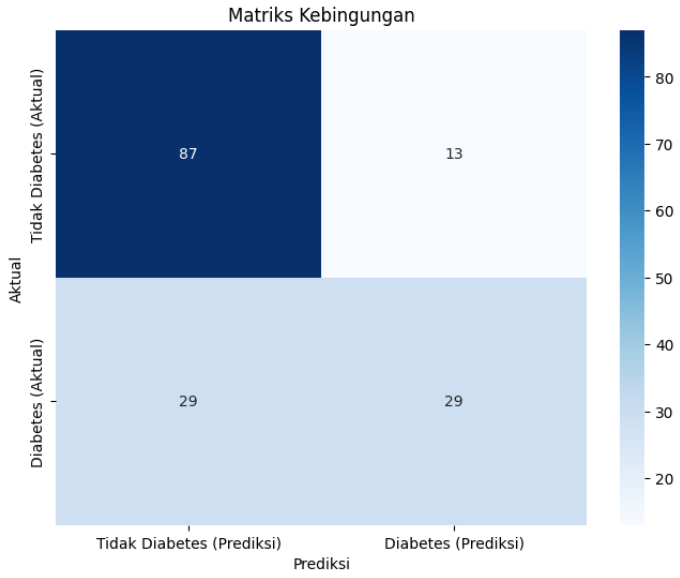
Pada tahap ini data yang sudah dipersiapkan akan kita lakukan proses pelatihan data supaya lebih akurat dalam melakukan prediksi. Berikut ini merupakan hasil dari training data yang Dimana penulis mendapatkan nilai hasil akurasi pada data uji 0.7342.



Gambar 4. 24 Hasil Akurasi Data Training

2. Confunsion matrix

Berikut ini merupakan grafik confunsion matrik hasil dari uji training data.

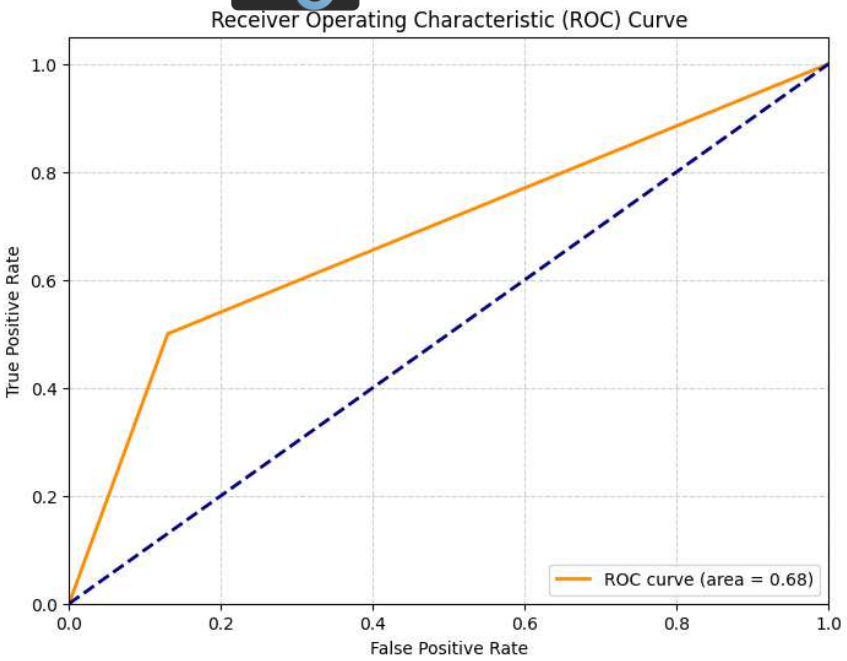


Gambar 4. 25 Confunsion Matrik

Pada gambar confunsioin matrix diatas dapat dijelaskan bahwa pada diabetes (actual) mendapatkan 29 dan diabetes prediksi mendapatkan 29. Pada Tidak diabetes (actual) mendapatkan 87 dan tidak diabetes prediksi mendapatkan 13.

3. Grafik ROC (*Receiver Operating Characteristic Curve*)

Grafix ROC (*Receiv rating Characteristic Curve*) hasil dari uji dataset diabetes.



Gambar 4. 26 Grafix ROC (Receiver Operating Characteristic Curve)

Pada gambar Grafix ROC (*Receiver Operating Characteristic Curve*) dapat dijelaskan bahwa pada true positive rate memperoleh nilai 0.5 hingga naik menjadi 1.0. AUC *(Area Under Curve*): 0.6850 ROC Curve generated and AUC calculated

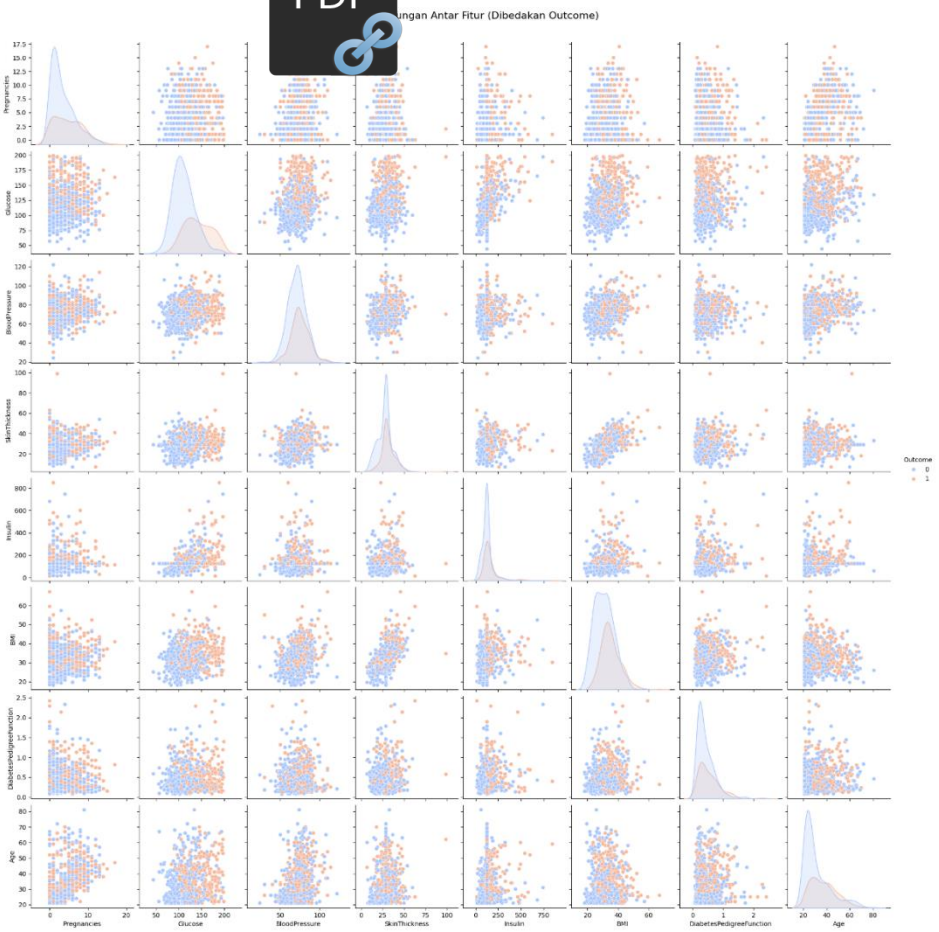
4. Analisis Visualisasi Data

Untuk memahami karakteristik data dan hubungan antar fitur, dilakukan visualisasi menggunakan pair plot. Visualisasi ini bertujuan untuk menampilkan distribusi masing-masing fitur serta hubungan antar pasangan fitur, yang dibedakan berdasarkan kelas *Outcome,* yaitu:

- *Outcome* = 0: Pasien tidak menderita diabetes

- *Outcome* = 1: Pasien menderita diabetes

Pair plot pada Gambar menunjukkan hasil hubungan antar fitur sebagai berikut:



Gambar 4. 27 Pair plot Hubungan Antar Fitur

Adapun Analisis hasil visualisasi adalah sebagai berikut

a. Fitur Glucose (Kadar Glukosa)

a) Terlihat paling membedakan antara pasien diabetes dan non-diabetes.

b) Titik merah (Outcome = 1) cenderung berkumpul pada nilai glukosa yang tinggi.

c) Ini menunjukkan bahwa kadar glukosa darah memiliki korelasi kuat dengan diabetes.

b. Fitur BMI (Body dex)

a) Pasien den MI tinggi (di atas 30) cenderung memiliki Outcome = 1.

b) Distribusi BMI untuk pasien diabetes berbeda signifikan dibandingkan dengan non-diabetes.

c. Fitur Age (Usia)

a) Pasien dengan usia di atas 40 tahun cenderung lebih banyak memiliki Outcome = 1.

b) Hal ini mendukung asumsi bahwa risiko diabetes meningkat seiring bertambahnya usia.

d Fitur Pregnancies (Jumlah Kehamilan)

a) Pasien dengan jumlah kehamilan lebih tinggi juga cenderung lebih berisiko, meskipun pengaruhnya tidak sekuat glukosa atau BMI.

b. .Fitur Insulin, SkinThickness, dan BloodPressure

a) Tiga fitur ini menunjukkan sebaran data yang cenderung tumpang tindih antar kelas.

b) Hal ini mengindikasikan bahwa fitur-fitur tersebut tidak memiliki perbedaan distribusi yang mencolok antara pasien diabetes dan non-diabetes secara visual.

f. Hubungan antar fitur (korelasi)

a) Sebagian besar scatter plot tidak menunjukkan pola linear yang kuat antar fi

b) Ini mempe utusan penggunaan algoritma seperti Decision Tree, yang tidak bergantung pada korelasi linear, melainkan membagi data berdasarkan ambang batas (threshold) secara hierarkis.

Dari visualisasi ini dapat disimpulkan bahwa fitur Glucose, BMI, Pregnancies dan Age merupakan fitur yang paling berpengaruh dalam membedakan antara pasien yang menderita diabetes dan yang tidak. Visualisasi ini membantu dalam memahami struktur data dan mendukung proses klasifikasi menggunakan algoritma Decision Tree pada tahap selanjutnya.

4.4. Mengklasfikasi Status Penyakit Diabetes dengan Metode *K-Nearest Neighbor (KNN)*

Adapun Tahapan Algoritma *KNN* sebagai berikut:

1. Tentukan parameter K.
2. Hitung jarak antara data yang akan dievaluasi dengan semua pelatihan.
3. Urutkan jarak yang terbentuk (dari terkecil ke terbesar).
4. Tentukan jarak terdekat sejumlah K.
5. Pasangkan kelas yang bersesuaian.
6. Cari jumlah kelas dari tetangga yang terdekat dan tetapkan kelas tersebut sebagai kelas data yang akan dievaluasi.

Berikut adalah hasil yang didapat setelah dilakukannya klasifikasi berdasarkan atribut diolah menggunakan *K-Nearest Neighbor (KNN)* sebagai berikut.

1. Inisialisasi Algoritma K

   Algoritma KNN diinisialisasi dengan parameter N Neighbor sama dengan K dimana K=5, berarti model akan menggunakan 5 tetangga terdekat untuk menentukan kelas dari data baru yang diberikan

   print("\nPerforming Cross-Validation (KNN)...")

   knn_pipeline = Pipeline([

   ('scaler', StandardScaler()),

   ('knn', KNeighborsClassifier(n_neighbors=5, weights='distance'))

2. Hasil inisialisasi data digunakan untuk melatih model KNN dengan data latih yaitu x_train dan y_train

   rint("\nTraining KNN model on training data...")

   knn_pipeline.fit(X_train, y_train)

   print("KNN model trained.")

3. Membuat Prediksi untuk menghasilkan prediksi berdasarkan pada data uji kemudian dibandingakan dengan data sebenarnya yaitu y_test untuk mengevauasi ferforma model

   y_pred_knn = knn_pipeline.predict(X_test)

   accuracy_knn = accuracy_score(y_test, y_pred_knn)

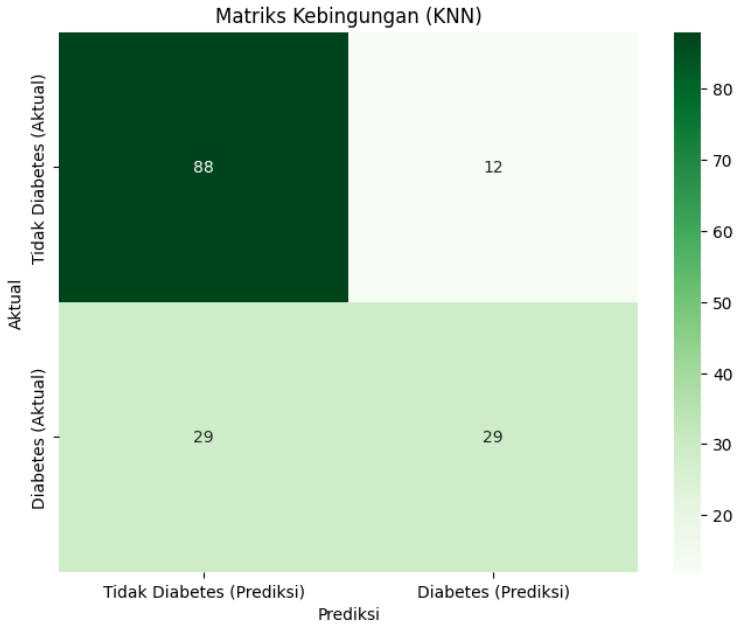
   print(f"\nAkurasi KNN pada Data Uji: {accuracy_knn:.4f}")

   print('\nLaporan Klasifikasi (KNN):\n', classification_report(y_test,

   y_pred_knn))

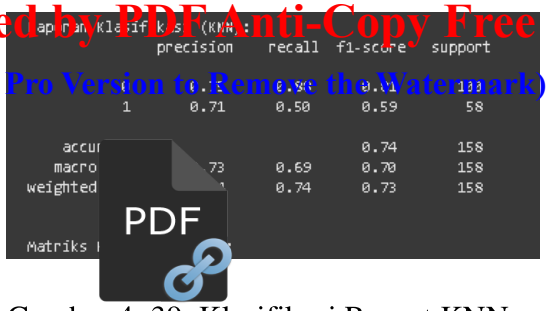
4. Kemudian dilakukan evaluasi model yaitu confusion matrik atau matrik kebingungan adalah alat untuk mengevaluasi ferforma klasifikasi, matrik kebingungan menunjuk[illegible]ibusi prediksi model terhadap data aktual. Pada penelitian ini ma[illegible]gunan menunjukkan 88 adalah perdiksi benar untuk kelas 0 atau true negative, 12 adalah prediksi salah dimana kelas 0 diprediksi kelas 1 atau False positive, 29 adalah prediksi salah dimana kelas 1 di prediksi kelas 0 atau false negative dan 29 adalah prediksi benar untuk kelas 1 yaitu True positive

```
conf_matrix_knn = confusion_matrix(y_test, y_pred_knn)
print('\nMatriks Kebingungan KNN:\n', conf_matrix_knn)
```



Gambar 4. 38 Matrik Kebingungan KNN

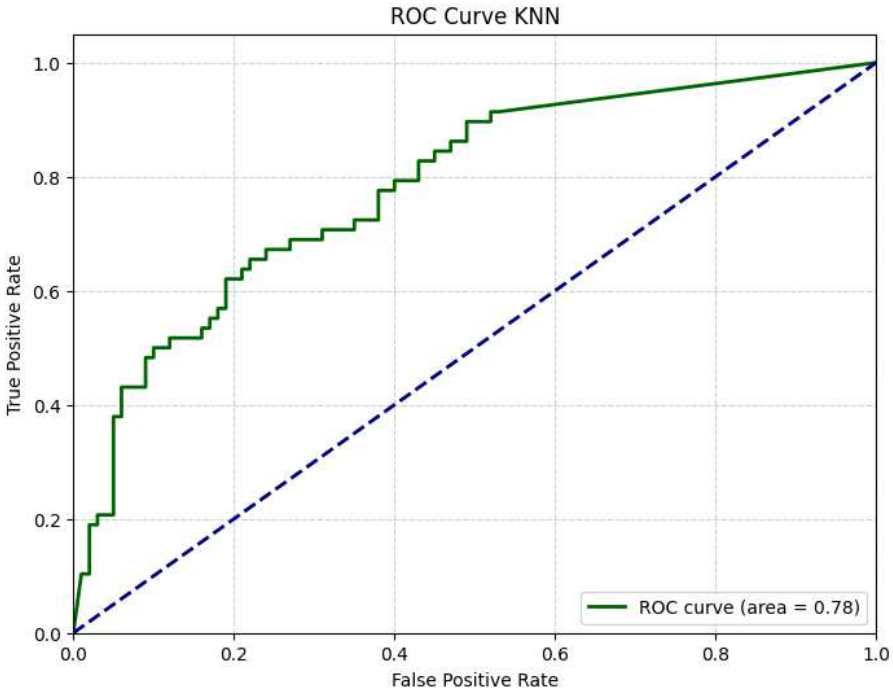
5. Melihat Klasifikasi report yang memberikan matrik penting diantaranya Presisi, recall dan f1-score untuk masing masing kelas



Gambar 4. 39 Klasifikasi Report KNN

Dilihat dari hasil pada gambar diatas untuk kelas 0 untuk presisi adalah 75% artinya dari semua presisi kelas 0 sebanyak 75% adalah benar dan kelas 1 sebanyak 71 \% pesen adalah benar

6. Menampilkan Kurva ROC dan kalkulasi AUC KNN



Gambar 4. 40 Kurva ROC

4.5 Pembahasan

1. Hasil penelitian ini menunjukkan bahwa algoritma *Decision Tree* dan *KNN* dapat digunakan secara efektif untuk mengklasifikasikan kondisi pasien terkait penyakit diabetes. Berdasarkan evaluasi terhadap dataset sebanyak 787 data pasien, didapatkan nilai akurasi sebesar 73,42 %, untuk

Dicision tree dan 74,05% yang tergolong tinggi dan menunjukkan kemampuan model dalam melakukan klasifikasi secara akurat.

2. Model *Dicision tree* m[illegible]eh nilai AUC sebesar 68,50% dan *KNN* sebesar 78,18% yang [illegible]dikasikan kemampuan model dalam membedakan kelas diabetes dan non-diabetes dengan baik. Hal ini diperkuat oleh hasil confusion matrix dan ROC Curve yang menunjukkan bahwa model memiliki kemampuan prediksi positif yang cukup tinggi. Meskipun tidak sempurna, tingkat kesalahan (false positive/false negative) masih dalam batas yang wajar untuk aplikasi awal klasifikasi medis berbasis data mining.

   Secara keseluruhan, penelitian ini menunjukkan bahwa algoritma *Decision Tree* dan *KNN* dapat digunakan sebagai alat bantu dalam mendeteksi risiko diabetes secara dini, dengan hasil klasifikasi yang cukup akurat. Model ini memberikan pondasi yang kuat bagi pengembangan sistem pendukung keputusan medis, yang di masa mendatang dapat diintegrasikan lebih lanjut dengan sistem informasi rumah sakit dan digunakan oleh tenaga kesehatan untuk diagnosis awal yang lebih cepat dan efisien.

3. Perbandingan antara Metode *Decision tree* dan *KNN* dari hasil yang didapatkan adalah

   Decision Tree - Akurasi Data Uji: 0.7342

   KNN - Akurasi Data Uji: 0.7405

   Decision Tree - AUC: 0.6850

   KNN - AUC: 0.7818

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpuan

Berdasarkan hasil penelitian dan pembahasan, dapat disimpulkan bahwa Penelitian ini berhasil membangun model klasifikasi penyakit diabetes menggunakan algoritma *Decision Tree.* Berdasarkan data diabetes yang di teliti berasal data publik sebagai data sekunder dan data tempat penelitian sebagai data primer mampu mengklasifikasikan kondisi pasien apakah termasuk penderita diabetes atau tidak dengan hasil akurasi sebesar 73,42% untuk metode *Decision tree* dan 74,05% untuk Metode *KNN* , yang tergolong tinggi dan cukup baik dalam mendukung pengambilan keputusan awal oleh tenaga medis. Algoritma *Decision Tree* dan *KNN* dipilih karena memiliki kelebihan dalam hal interpretasi hasil, sehingga alur pengambilan keputusan dapat ditelusuri secara logis melalui struktur pohon keputusan. Sistem klasifikasi ini menunjukkan potensi sebagai alat bantu diagnosis awal berbasis data historis medis, khususnya dalam mendeteksi kemungkinan pasien mengidap diabetes berdasarkan sejumlah parameter medis seperti kadar glukosa, tekanan darah, tingkat melahirkan, riwayat ,usia, BMI, dan lainnya.

5.2 Saran

Berikut ini adalah beberapa yang dapat dipertimbangkan untuk penelitian selanjutnya :

1. Perluasan jumlah dan peningkatan kualitas dataset sangat disarankan untuk meningkatkan generalisasi dan akurasi model, sehingga hasil prediksi dapat lebih representatif terhadap populasi pasien yang lebih luas.
2. Penggunaan algoritma pembanding seperti Random Forest, Support Vector Machine dapat dipertimbangkan untuk menguji dan membandingkan performa klasifikasi terhadap dataset yang sama.
3. Sistem ini dapat dikembangkan lebih lanjut dengan diintegrasikan ke dalam sistem informasi rumah sakit sebagai alat bantu diagnosis otomatis bagi tenaga medis dalam menangani pasien secara lebih cepat dan akurat.
4. Penelitian lanjutan juga dapat mengkaji aspek feature importance dari model *Decision Tree* untuk mengetahui fitur medis mana yang paling berpengaruh dalam klasifikasi, sehingga dapat menjadi fokus utama dalam upaya pencegahan atau penanganan penyakit diabetes.

DAFTAR PUSTAKA

[1] F. M. Hana, "Klasifika nderita Penyakit Diabetes Menggunakan Algoritma Decision Tree J. *SISKOM-KB (Sistem Komput. dan Kecerdasan Buatan)*, vol. pp. 32–39, 2020, doi: 10.47970/siskom-kb.v4i1.173.

[2] M. M. S. Jogo, M. K. Biddinika, and A. Fadlil, "Klasifikasi Penyakit Diabetes dengan Algoritma Decision Tree dan Naïve Bayes," *Resist. (Elektronika Kendali Telekomun. Tenaga List. Komputer) Vol.*, vol. 6, no. 2, pp. 113–118, 2023.

[3] A. K. Febianto and C. A. Sugianto, "Optimalisasi Algortima Klasifikasi Ensemble Menggunakan Algortima Genetika Untuk Prediksi Resiko Diabetes," *Brahmana J. Penerapan Kecerdasan Buatan*, vol. 5, no. 2, pp. 205–213, 2024.

[4] S. Ucha Putri, E. Irawan, F. Rizky, S. Tunas Bangsa, P. A. -Indonesia Jln Sudirman Blok No, and S. Utara, "Implementasi Data Mining Untuk Prediksi Penyakit Diabetes Dengan Algoritma C4.5," *Januari*, vol. 2, no. 1, pp. 39–46, 2021.

[5] N. Nurussakinah and M. Faisal, "Klasifikasi Penyakit Diabetes Menggunakan Algoritma Decision Tree," *J. Inform.*, vol. 10, no. 2, pp. 143–149, 2023, doi: 10.31294/inf.v10i2.15989.

[6] A. Pameka, R. Heriansyah, and L. W. Astuti, "Optimalisasi Feature Selection Untuk Mendeteksi Penyakit Diabetes Mellitus Menggunakan Metode Decision Tree," *JUPITER J. Penelit. ...*, pp. 589–599, 2024, [Online]. Available: https://jurnal.polsri.ac.id/index.php/jupiter/article/view/8632%0Ahttps://jurnal.polsri.ac.id/index.php/jupiter/article/download/8632/3291

[7] P. Patricia, Algoritma Decision Tree Dan Particle Swarm Klasifikasi Data Penyakit Diabetes Menggunakan Algoritma Decision Tree Dan Particle Swarm. 2024

[8] S. Pokhrel, "No TitleΕΛΕΝΗ," *Αγαη*, vol. 15, no. 1, pp. 37–48, 2024

[9] A. Afifuddin and L. Hakim, "Deteksi Penyakit Diabetes Mellitus Menggunakan Algoritma Decision Tree Model Arsitektur C4.5," *J. Krisnadana*, vol. 3, no. 1, pp. 25–33, 2023, doi: 10.58982/krisnadana.v3i1.470.s

[10] B. A. R. P. Wahyu, A. F. Farozi, C. P. Mahendra, and R. K. Hapsari, "Klasifikasi Penderita Penyakit Diabetes Berdasarkan Decision Tree," *INTEGER J. Inf. Technol.*, vol. 8, no. 1, pp. 80–89, 2023.

[11] *https://repository.uin-uska.ac.id/15906/7/7%20BAB%20II_2018403SIF.pdf*

[12] T. Syamsudin, T. Handhayani,) Muhammad, and I. Syaifudin, "Jurnal Ilmu Komputer dan Sistem Informasi Perbandingan Klasifikasi Penyakit Diabetes Menggunakan Metode Machine Learning," pp. 1–7, 2021, [Online]. Available: https://www.kaggle.com/datasets/nanditapore/healthcar

[13] N. Nurussakinah and M. Farsal, "Klasifikasi Penyakit Diabetes Menggunakan Algoritma Decision Tree," *J. Inform.*, vol. 10, no. 2, pp. 143–149, 2023, doi: 10.31294/inf.v10i2.15989.

[14] Rizky Adhi Nugroho1 , Tar Alan Prahutama "Klasifikasi Pasien Diabetes Militrus Menggunakan S *Support Vector Machine* (Ssvm)"Jurnal Gausian, Vol. 6,no.3, pp 4 017)

[15] Andharini Dwi Cahyani asuki "Klasifikasi Diabetes Mellitus Menggunakan Support ve ine (Studi Kasus: Puskesmas Modopuro, Mojokerto)" J. Rekayasa, Vol. 6,no.3, pp 174-182 (2019)

[16]Ginanjar Abdulrahman "Klasifikasi Penyakit Diabetes Melitus Menggunakan Adaboost Classifier"" J. Justindo, Vol. 7, no.1, pp 59-66 (2022)