LAPORAN PROYEK UAS

PREDIKSI PENYAKIT DIABETES MENGGUNAKAN ALGORITMA RANDOM FOREST

Disusun untuk memenuhi Ujian Akhir Semester Mata Kuliah Kecerdasan Buatan

Disusun oleh:

AISHA KAMIL A	2306015
GINA QURROTA A	2306029

Dosen Pengampu Mata Kuliah:

Leni Fitriani, ST. M.Kom.

INSTITUT TEKNOLOGI GARUT
JURUSAN ILMU KOMPUTER
PROGRAM STUDI TEKNIK INFORMATIKA

TAHUN AKADEMIK 2024/2025

DAFTAR ISI

DAFTAR GAMBAR

BAB I

PENDAHULUAN

1.1 LATAR BELAKANG

Penyakit diabetes mellitus adalah salah satu kondisi kronis yang semakin umum di seluruh dunia dan menjadi salah satu tantangan utama dalam pelayanan kesehatan. Menurut laporan dari WHO, lebih dari 422 juta individu di seluruh dunia menderita diabetes, dan jumlah ini diperkirakan akan terus meningkat, terutama di negara-negara yang sedang berkembang(*Diabetes*, n.d.). Pendeteksian dini memiliki peranan yang sangat penting agar komplikasi serius, seperti kerusakan pada saraf, jantung, dan ginjal, dapat dicegah. Akan tetapi, fakta di lapangan menunjukkan bahwa banyak orang yang menderita diabetes tidak didiagnosis lebih awal akibat keterbatasan dalam sumber daya medis dan rendahnya kesadaran masyarakat mengenai tanda-tanda awal penyakit ini.

Kemajuan teknologi kecerdasan buatan (AI) memberikan kesempatan yang signifikan untuk meningkatkan sistem dalam mendeteksi penyakit secara tepat dan efisien. Salah satu metode yang semakin populer dalam dunia medis adalah penggunaan algoritma pembelajaran mesin, terutama untuk melakukan klasifikasi risiko berdasarkan informasi kesehatan pasien(Ikram et al., 2022). Dalam hal ini, data historis pasien seperti kadar glukosa, tekanan darah, dan indeks massa tubuh dapat digunakan sebagai dasar untuk mengembangkan sistem prediksi yang canggih.

Algoritma Random Forest dipilih sebagai metode utama dalam proyek ini karena kemampuan yang dimilikinya untuk mengelola data yang rumit serta memberikan pemahaman yang penting tentang fitur-fitur yang paling berpengaruh pada keputusan klasifikasi. Di samping itu, penerapan teknik seperti SMOTE membantu sistem dalam menangani ketidakseimbangan kelas dalam data medis, yang seringkali menjadi masalah di dataset penyakit kronis seperti diabetes(Navarro-Cáceres et al., 2020).

1.2 RUMUSAN MASALAH

Rumusan masalah dalam penelitian ini mencakup cara untuk mengembangkan sistem prediksi diabetes menggunakan algoritma Random Forest yang dapat menghasilkan klasifikasi yang tepat. Selain itu, ada tantangan lain mengenai penanganan distribusi kelas yang tidak seimbang

dalam data yang digunakan, di mana sebagian besar informasi menunjukkan pasien tidak mengalami diabetes.

Jika masalah ketidakseimbangan data ini tidak ditangani, ada risiko bahwa model prediktif akan berpihak pada kelas mayoritas, yang dalam konteks medis sangat berbahaya karena dapat mengabaikan pasien yang sebenarnya terkena diabetes. Oleh sebab itu, perlu dilakukan percobaan penerapan SMOTE sebagai solusi untuk menyeimbangkan distribusi antar kelas(Navarro-Cáceres et al., 2020).

Selanjutnya, evaluasi menyeluruh terhadap kinerja sistem prediksi ini sangat diperlukan dengan menggunakan metrik seperti akurasi, presisi, recall, dan skor F1. Tujuannya adalah untuk memastikan bahwa model tidak hanya kuat dari sisi akurasi keseluruhan, tetapi juga efektif dalam mengidentifikasi pasien yang menderita diabetes secara akurat.

1.3 TUJUAN

Tujuan utama dari proyek ini adalah menciptakan sistem untuk memprediksi penyakit diabetes dengan menggunakan algoritma Random Forest. Sistem ini dirancang untuk mempermudah deteksi awal kemungkinan diabetes berdasarkan informasi pasien yang ada. Diharapkan, sistem ini dapat diadopsi oleh tenaga medis sebagai alat untuk mempercepat dan mempermudah proses diagnosis awal.

Tujuan lain dari proyek ini adalah untuk menilai dampak penerapan teknik SMOTE terhadap akurasi dalam klasifikasi pada dataset yang tidak seimbang. Teknik ini dianggap efektif dalam meminimalkan bias dalam model klasifikasi dan menghasilkan prediksi yang lebih seimbang antara pasien diabetes dan non-diabetes(*Diabetes*, n.d.; Ikram et al., 2022).

Selain itu, proyek ini juga bertujuan untuk menguji efektivitas fitur kesehatan tertentu dalam proses klasifikasi. Menurut studi terdahulu, fitur seperti kadar glukosa dan BMI sering kali menjadi faktor utama dalam memprediksi diabetes(Ikram et al., 2022), sehingga penting untuk memverifikasi hal ini melalui analisis nilai korelasi dan pentingnya fitur dari model yang ada(Ribeiro et al., 2016).

1.4 MANFAAT

Manfaat nyata dari inisiatif ini adalah adanya sistem yang dapat mendeteksi potensi diabetes lebih awal, yang mendukung dokter dan pasien dalam mengambil langkah pencegahan yang lebih efektif. Sistem ini dirancang untuk melakukan skrining awal dengan cepat, terutama di daerah yang tidak memiliki banyak fasilitas laboratorium.

Dari segi teknologi, inisiatif ini memberikan sumbangan penting bagi pengembangan model pembelajaran mesin yang lebih tepat dan responsif terhadap isu data yang tidak seimbang. Hasil dari pengujian dan pengembangan dapat menjadi referensi bagi penelitian lain yang menggunakan metode serupa untuk penyakit kronis yang berbeda.

Dalam konteks akademik, proyek ini berperan penting dalam literatur mengenai penerapan pembelajaran mesin untuk bidang kesehatan, dan diharapkan dapat memperluas wawasan ilmu pengetahuan, terutama dalam penerapan algoritma ensemble seperti Random Forest dalam konteks medis(Ribeiro et al., 2016).

BAB II

`LANDASAN TEORI

2.1 KECERDASAN BUATAN (AI)

Kecerdasan Buatan atau yang dikenal dengan istilah Artificial Intelligence (AI) adalah suatu bidang dalam ilmu komputer yang bertujuan untuk menciptakan sistem yang bisa menjalankan tugas yang umumnya memerlukan kecerdasan manusia. Tugas-tugas ini meliputi pengambilan keputusan, analisis bahasa alami, pengenalan pola, serta klasifikasi dalam bidang medis. Di sektor kesehatan, AI telah mengalami perkembangan yang signifikan dan diterapkan dalam berbagai cara untuk mempercepat proses diagnosis serta meningkatkan ketepatan analisis data pasien.

Salah satu cara AI diterapkan dalam kehidupan sehari-hari adalah melalui sistem prediksi penyakit menggunakan machine learning. Sistem ini dilatih dengan data masa lalu untuk menemukan pola yang mungkin tidak disadari oleh manusia. Dalam situasi penyakit kronis seperti diabetes, AI mampu membantu dalam mengidentifikasi pasien yang memiliki risiko tinggi hanya dengan memanfaatkan data numerik sederhana seperti level glukosa dan tekanan darah(*Diabetes*, n.d.) (Ikram et al., 2022).

Dengan bertambahnya jumlah data kesehatan dan kerumitan informasi, AI menjadi alat yang crucial untuk meningkatkan keputusan medis yang berdasarkan bukti. Sistem pintar ini tidak hanya memaksimalkan efisiensi dalam diagnosis, tetapi juga membantu dalam merencanakan pengobatan dan memantau pasien secara langsung(Ikram et al., 2022).

2.2 ALGORITMA RANDOM FOREST

Algoritma Random Forest merupakan teknik pembelajaran kelompok yang membentuk sejumlah pohon keputusan dan mengkombinasikan hasilnya untuk menghasilkan prediksi akhir. Teknik ini terkenal dalam klasifikasi medis karena keandalannya, kemampuannya dalam menangani data anomali, serta keefektifannya dalam mengelola data numerik dan karakteristik non-linear. Setiap pohon dalam Random Forest dibuat berdasarkan sampel acak dari data pelatihan dan fitur, yang membantu mencegah terjadinya overfitting.

Dalam hal prediksi diabetes, Random Forest telah terbukti efektif karena dapat menilai karakteristik kesehatan yang paling berpengaruh terhadap klasifikasi. Selain itu, metode ini bisa menghasilkan fitur penting untuk memberikan pemahaman tentang mana yang paling

berpengaruh terhadap hasil prediksi. Ini sangat bermanfaat bagi profesional medis agar dapat memfokuskan perhatian pada parameter-parameter yang sangat penting dalam diagnosis awal(Ikram et al., 2022).

Kelebihan lainnya dari Random Forest adalah kemampuannya untuk menangani data medis yang sering kali tidak seimbang. Dengan adanya teknik seperti SMOTE, algoritma ini dapat memberikan hasil prediksi yang konsisten, bahkan dalam kumpulan data dengan distribusi kelas yang tidak merata(Navarro-Cáceres et al., 2020).

2.3 KONSEP EVALUASI MODEL

Evaluasi model adalah komponen krusial dalam pembelajaran mesin karena menentukan tingkat akurasi dan keandalan sistem yang dirancang untuk prediksi. Beberapa metrik penting yang sering digunakan meliputi akurasi, precision, recall, dan F1-score. Akurasi menunjukkan frekuensi prediksi model yang benar, namun bisa menyesatkan jika data tidak seimbang. Oleh sebab itu, precision dan recall digunakan untuk mengevaluasi kinerja model berdasarkan setiap kelas. Precision menunjukkan proporsi prediksi positif yang benar-benar positif, sedangkan recall mengukur seberapa efektif model dalam menemukan semua kejadian positif dalam data. Dalam situasi prediksi diabetes, recall sangat penting karena kesalahan dalam mengidentifikasi pasien yang benar-benar sakit dapat berakibat serius. F1-score, yang merupakan rata-rata harmonis dari precision dan recall, memberikan ukuran yang lebih seimbang untuk menilai kinerja(Navarro-Cáceres et al., 2020).

Selain itu, interpretabilitas model juga menjadi aspek yang sangat penting. Metode seperti LIME atau SHAP dapat dipakai untuk menjelaskan keputusan model kepada pengguna, terutama dalam bidang medis di mana sekadar akurasi tidak cukup. Keterbukaan dalam sistem AI adalah syarat utama agar model dapat diterima dan digunakan secara luas oleh tenaga kesehatan dan pasien(Ribeiro et al., 2016).

2.4 STUDI TERKAIT

Banyak penelitian mengenai penggunaan AI untuk memprediksi diabetes telah dilakukan dalam beberapa tahun terakhir. Dalam ulasannya, Alghamdi et al. (2022) menyimpulkan bahwa algoritma Random Forest dan XGBoost adalah di antara yang paling akurat untuk klasifikasi diabetes. Mereka juga menekankan pentingnya preprocessing data dan pemilihan fitur yang sesuai untuk meningkatkan performa model(Ikram et al., 2022).

Selanjutnya, riset oleh Haq et al. (2020) menunjukkan bahwa penerapan teknik oversampling seperti SMOTE dapat secara signifikan memperbaiki kinerja model klasifikasi medis, terutama dalam mendeteksi kelas minoritas seperti diabetesi. Teknik ini efektif dalam memperbaiki distribusi data sehingga model tidak lagi cenderung pada kelas mayoritas(Navarro-Cáceres et al., 2020).

Dalam hal interpretabilitas, Ribeiro et al. (2016) memperkenalkan metode LIME yang memberikan penjelasan lokal tentang keputusan model, membantu pengguna yang tidak berpengalaman memahami hasil prediksi perangkat. Pendekatan ini sangat bermanfaat untuk membangun kepercayaan terhadap model dalam sistem prediksi berbasis AI di sektor kesehatan(Ribeiro et al., 2016).

BAB III

HASIL DAN PEMBAHASAN

3.1 BUSINESS UNDERSTANDING

Tingginya jumlah penderita diabetes yang seringkali tidak teridentifikasi sejak awal. Diabetes adalah penyakit jangka panjang yang bisa menyebabkan komplikasi serius jika tidak ditangani dengan benar(Febrian et al., 2024). Salah satu tantangan utama di bidang kesehatan adalah keterbatasan dalam mendeteksi penyakit ini secara cepat dan tepat, terutama di daerah yang kekurangan tenaga medis atau teknologi diagnostik yang memadai. Selain itu, data medis yang digunakan untuk klasifikasi sering kali tidak seimbang, di mana jumlah pasien yang tidak menderita diabetes jauh lebih banyak dibandingkan dengan mereka yang terkena, sehingga hal ini dapat mempengaruhi kinerja model prediksi.

Tujuan dari proyek ini adalah untuk mengembangkan sistem klasifikasi yang menggunakan pembelajaran mesin yang dapat memperkirakan kemungkinan seseorang menderita diabetes berdasarkan parameter kesehatan seperti kadar glukosa, tekanan darah, dan indeks massa tubuh. Algoritma yang dipakai yaitu Random Forest, yang dikenal dapat menangani jenis data yang kompleks, serta teknik SMOTE (Synthetic Minority Over-sampling Technique) yang digunakan untuk menyeimbangkan distribusi kelas agar model tidak condong ke kelompok mayoritas.

Pengguna sistem ini meliputi petugas kesehatan seperti dokter dan perawat, institusi kesehatan seperti rumah sakit serta laboratorium, dan juga para peneliti serta pembuat kebijakan di sektor kesehatan. Dengan adanya sistem ini, mereka bisa melakukan pemeriksaan awal terhadap pasien dengan cara yang lebih efisien dan tepat.

Penerapan kecerdasan buatan dalam proyek ini memberikan berbagai manfaat, antara lain mempercepat proses deteksi dini diabetes, meningkatkan akurasi hasil diagnosa, serta membantu dalam pengambilan keputusan yang lebih baik dalam perawatan pasien. Dengan sistem prediksi yang berbasis kecerdasan buatan ini, pemeriksaan medis menjadi lebih efisien, tepat, dan dapat mencakup lebih banyak orang, khususnya dalam program pencegahan dan pengendalian penyakit jangka panjang.

3.2 DATA UNDERSTANDING

Data yang digunakan dalam proyek ini berasal dari dataset publik yang dapat diakses di Kaggle: https://www. kaggle. com/datasets/akshaydattatraykhare/diabetes-dataset? resource=download

Dataset ini memiliki 8 fitur (atribut) dan 1 target klasifikasi. Adapun fitur-fitur tersebut adalah:

- Kehamilan: Total jumlah kehamilan yang pernah dialami oleh pasien.
- Glukosa: Tingkat glukosa plasma dalam uji toleransi glukosa selama dua jam.
- Tekanan Darah: Tinggi tekanan darah diastolik (dalam mm Hg).
- Ketebalan Kulit: Ketebalan lipatan kulit pada trisep (dalam mm).
- Insulin: Kadar insulin dalam serum setelah dua jam (dalam mu U/ml).
- Indeks Massa Tubuh: Rasio berat badan dalam kilogram terhadap kuadrat tinggi badan dalam meter.
- Fungsi Silisilah Diabetes: Skor probabilitas berdasarkan sejarah keluarga mengenai diabetes.
- Usia: Usia pasien dalam tahun.

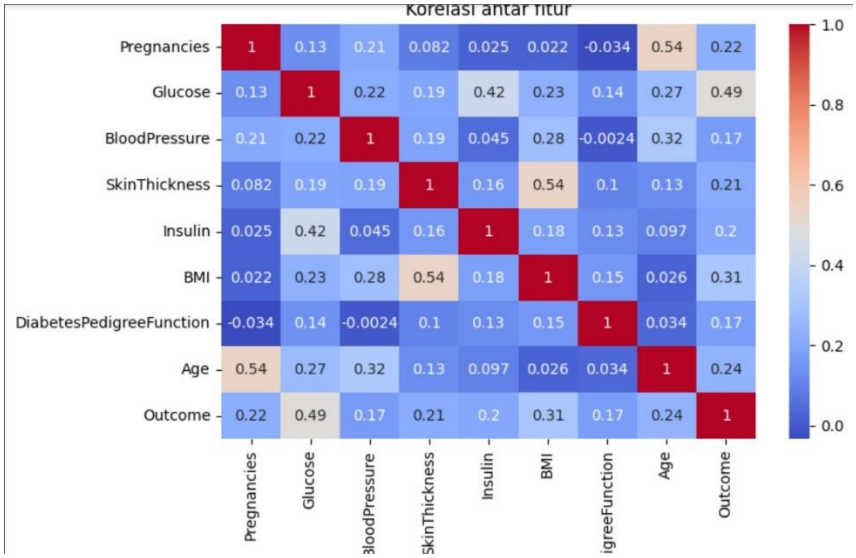
Target klasifikasi tersebut adalah Hasil, yakni label biner yang menunjukkan apakah pasien memiliki diabetes (1) atau tidak (0).

Dataset ini mencakup 768 baris (entri data) dan 9 kolom. Format file yang digunakan adalah CSV (Comma-Separated Values), yang mudah untuk dibaca dan diproses dengan berbagai alat analisis data. Sebagian besar tipe data dalam dataset ini adalah numerik, baik dalam format float atau integer, sementara kolom Hasil bertipe integer dan berfungsi sebagai variabel target untuk klasifikasi.

Dengan pemahaman tentang data ini, langkah-langkah preprocessing seperti penanganan nilai nol (yang tidak sesuai secara medis), pengisian nilai yang hilang, serta analisis distribusi dapat dilakukan dengan lebih akurat untuk mendukung proses pemodelan yang efektif,(Dikan Ismafillah et al., 2023).

3.3 EXPLORATORY DATA ANALYSIS (EDA)

3.3.1 HEATMAP



Gambar 1 Heatmap

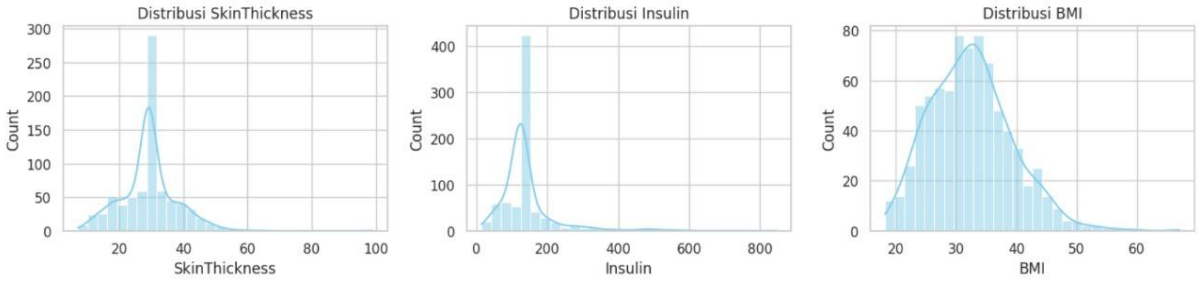
Gambar heatmap yang ditunjukkan sebelumnya mencerminkan matriks hubungan antar fitur dalam data diabetes yang sedang kamu teliti. Setiap kotak pada heatmap menggambarkan nilai korelasi Pearson antara sepasang fitur, di mana skala warnanya mulai dari biru (korelasi rendah atau negatif) hingga merah (korelasi tinggi atau positif). Nilai korelasi ini berada dalam rentang -1 hingga 1, di mana:

- *+1* menandakan korelasi positif sempurna (apabila satu fitur meningkat, yang lain juga ikut meningkat),
- *0* menandakan tidak adanya hubungan linear,
- *-1* menandakan korelasi negatif sempurna (satu fitur meningkat, yang lain menurun).

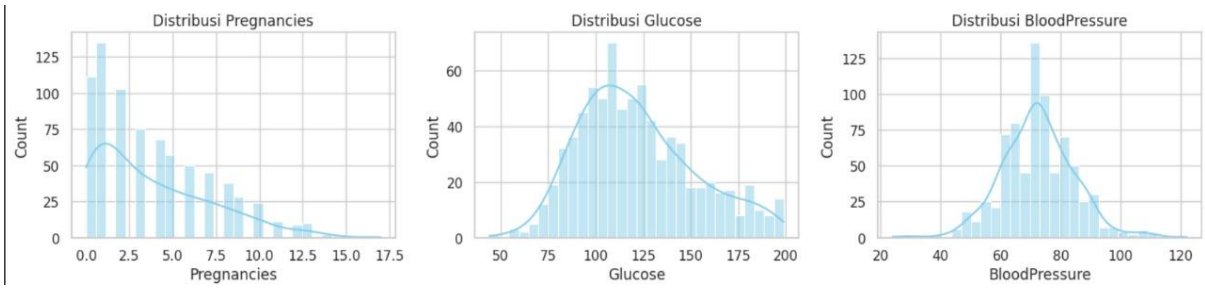
Fokus utama dari heatmap ini biasanya terletak pada baris dan kolom yang bertuliskan *Outcome*, karena ini merupakan target yang ingin diprediksi: apakah individu tersebut menderita diabetes (1) atau tidak (0). Dari heatmap terlihat bahwa fitur Glucose memiliki nilai korelasi paling tinggi terhadap Outcome (0. 49), yang menunjukkan bahwa semakin tinggi kadar glukosa seseorang, semakin besar kemungkinan mereka terdiagnosis diabetes. Fitur lainnya seperti BMI (0. 31), Age (0. 24), dan Pregnancies (0. 22) juga menunjukkan korelasi positif terhadap diabetes, meskipun tidak sekuat Glucose.

Di samping itu, heatmap ini juga berperan dalam mengidentifikasi *hubungan antar fitur*, seperti BMI dan SkinThickness yang memiliki nilai korelasi cukup tinggi (0. 54), yang mungkin menunjukkan adanya keterkaitan logis antara kedua variabel tersebut (misalnya, individu dengan persentase lemak tubuh lebih tinggi cenderung memiliki ketebalan kulit yang lebih besar). Heatmap ini sangat bermanfaat dalam tahap eksplorasi data karena dapat membantu kita menentukan fitur mana yang relevan, mendeteksi fitur yang redundan, serta memahami pola-pola tersembunyi dalam data sebelum tahap pemodelan dilakukan.

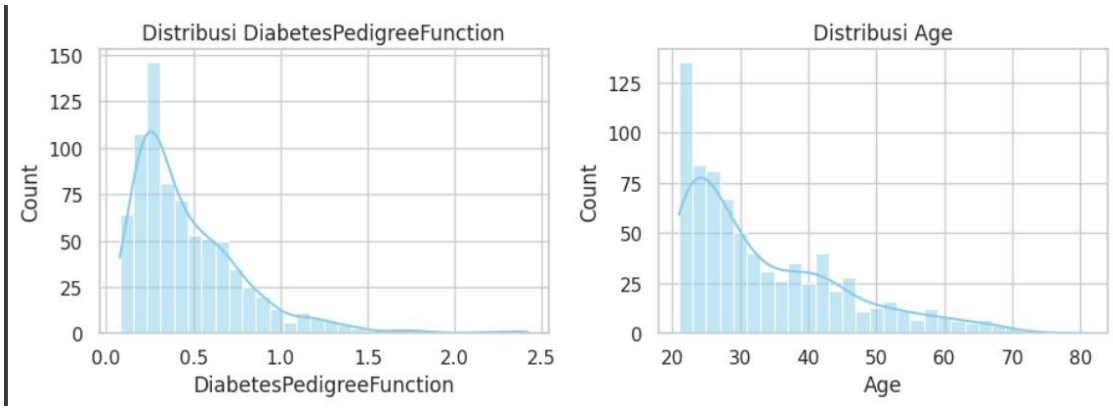
3.3.2 HISTOGRAM



Gambar 2 Histogram (1)



Gambar 3 Histogram (2)



Gambar 4 Histogram (3)

Berikut adalah penjelasan mengenai setiap histogram yang ditampilkan. Setiap grafik menunjukkan penyebaran data untuk satu fitur numerik dalam dataset diabetes, disertai kurva KDE (Kernel Density Estimate) untuk menggambarkan bentuk distribusi dengan lebih halus.

1. Distribusi Kehamilan
    - Distribusinya memiliki condongan ke kanan.
    - Sebagian besar pasien mengalami kehamilan antara 0 hingga 5 kali.

- Terdapat beberapa outlier dengan jumlah kehamilan lebih dari 10 kali, tetapi jumlahnya sangat terbatas.

Ini menunjukkan bahwa mayoritas pasien adalah wanita dengan jumlah kehamilan yang cukup rendah.

2. Distribusi Glukosa
    - Bentuk distribusinya hampir normal, meskipun sedikit condong ke kanan.
    - Puncak frekuensinya terletak di kisaran 100 hingga 125.
    - Masih ada nilai yang sangat tinggi (lebih dari 180), yang mungkin mewakili pasien dengan risiko diabetes yang tinggi.

Glukosa adalah fitur yang sangat penting dalam klasifikasi diabetes, dan distribusinya terlihat cukup menyebar.

3. Distribusi Tekanan Darah
    - Distribusi ini cukup seimbang dan mendekati bentuk normal.
    - Sebagian besar nilai tekanan darah pasien berada pada rentang 60 hingga 80 mmHg.
    - Ada sedikit nilai ekstrem (di atas 100), tetapi tidak signifikan.

Distribusi ini mencerminkan bahwa tekanan darah pasien umumnya stabil, meskipun beberapa cenderung tinggi.

4. Distribusi Ketebalan Kulit
    - Terdapat lonjakan signifikan di sebuah nilai tertentu (~32), yang mungkin merupakan hasil imputasi (nilai median).
    - Distribusinya condong ke kanan, dan tampaknya tidak alami — ini sering terjadi setelah proses imputasi.

Perlu diwaspadai karena hasil distribusi ini bisa terpengaruh oleh nilai pengisian data yang hilang.

5. Distribusi Insulin
    - Distribusinya sangat condong ke kanan, dengan banyak nilai rendah dan beberapa outlier yang besar (hingga 800).
    - Ada lonjakan di sekitar 150 hingga 200, yang kemungkinan juga disebabkan efek imputasi.

Ini adalah fitur yang rentan terhadap outlier dan bisa sangat mempengaruhi model jika tidak dinormalisasi atau diperbaiki lebih lanjut.

6. Distribusi Indeks Massa Tubuh (BMI)

- Distribusinya cukup seimbang, dengan sedikit condongan ke kanan.
- Sebagian besar nilai BMI terletak di rentang 30 hingga 40, yang termasuk dalam kategori kelebihan berat badan atau obesitas.

Ini memperkuat hubungan bahwa obesitas memiliki koneksi kuat dengan risiko diabetes.

7. Distribusi Fungsi Keturunan Diabetes
   - Distribusi ini sangat condong ke kanan.
   - Sebagian besar nilai terdapat di bawah 0. 5, dengan beberapa outlier lebih dari 2. 0.

Ini berarti sebagian besar pasien tidak memiliki riwayat keluarga yang kuat terkait diabetes, meskipun beberapa memiliki risiko genetik yang tinggi.

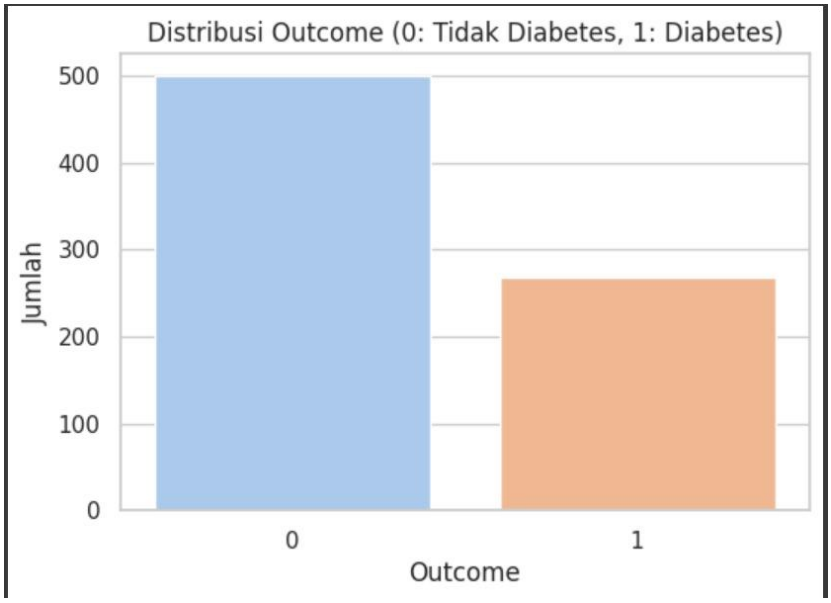
8. Distribusi Usia
   - Distribusi usia juga memiliki condongan ke kanan.
   - Mayoritas pasien berusia antara 20 hingga 40 tahun, dengan penurunan yang tajam di usia lebih lanjut.
   - Masih ada pasien yang berusia di atas 60 hingga 70 tahun, tetapi jumlahnya sedikit.

Ini menunjukkan bahwa dataset ini didominasi oleh pasien yang berada dalam usia produktif. Fitur-fitur seperti Kehamilan, Insulin, Usia, dan Fungsi Keturunan Diabetes cenderung condong ke kanan. Fitur Tekanan Darah dan BMI cenderung seimbang atau mendekati distribusi normal. Beberapa distribusi (terutama Ketebalan Kulit dan Insulin) kemungkinan dipengaruhi oleh proses imputasi median, yang terlihat dari lonjakan mendadak di satu nilai tertentu.

Kelas yang tidak seimbang dapat menjadikan model pembelajaran mesin condong terhadap kelas yang lebih banyak. Akibatnya, model sering kali memberikan prediksi yang lebih mengarah ke kelas dominan, meskipun tingkat akurasi keseluruhan tampak tinggi. Namun, kinerja pada kelas minoritas (dalam kasus ini, pasien yang terkena diabetes) bisa sangat buruk, padahal deteksi pada kelompok ini sangatlah krusial.

Untuk menangani ketidakseimbangan ini, diterapkan metode SMOTE (Synthetic Minority Over-sampling Technique) yang secara otomatis menambah data sintetis ke kelas minoritas sehingga distribusi antar kelas menjadi seimbang. Setelah SMOTE diterapkan, dilakukan pemeriksaan ulang terhadap distribusi label dan hasilnya menunjukkan bahwa jumlah data pada kedua kelas telah setara. Dengan cara ini, model yang dibangun diharapkan bisa belajar dengan lebih proporsional dan memberikan prediksi yang lebih akurat untuk kedua kelas.

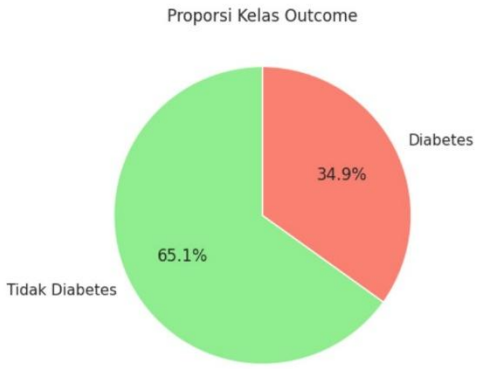
3.3.3 BARCHART



Gambar 5 Barchart

Grafik di atas menunjukkan distribusi outcome terkait diabetes, di mana 0 mewakili individu yang tidak terkena diabetes, sedangkan 1 mewakili individu yang terkena diabetes. Dari grafik tersebut, terlihat bahwa jumlah individu yang tidak memiliki diabetes (0) mencapai lebih dari 500 orang, ditandai dengan warna biru muda yang mencolok. Sementara itu, jumlah individu yang mengalami diabetes (1) jauh lebih sedikit, dengan total sekitar 200 orang, yang diwakili oleh warna oranye. Hal ini mengindikasikan bahwa dalam sampel yang dianalisis, proporsi individu yang tidak menderita diabetes jauh lebih tinggi dibandingkan dengan mereka yang menderita diabetes.

3.3.4 PIE CHART



Gambar 6 Pie Chart

Pie chart ini menunjukkan bahwa sebagian besar populasi (65.1%) tidak menderita diabetes, sementara 34.9% lainnya memiliki kondisi diabetes. Ini memberikan gambaran visual yang jelas tentang distribusi status kesehatan terkait diabetes dalam populasi yang dianalisis.

3.3.5 INSIGHT AWAL DARI POLA DATA

Hasil dari eksplorasi awal pada dataset diabetes memperlihatkan beberapa pola dan karakteristik penting yang memberikan wawasan awal tentang struktur data. Sebagian besar fitur numerik dalam dataset menunjukkan distribusi yang tidak merata (skewed), terutama pada fitur-fitur seperti Pregnancies, Insulin, Age, dan DiabetesPedigreeFunction, yang memiliki nilai skewness positif atau right-skewed. Ini menunjukkan bahwa mayoritas pasien memiliki nilai rendah, tetapi ada beberapa yang memiliki nilai yang sangat tinggi (outlier). Di sisi lain, fitur seperti BloodPressure dan BMI menunjukkan distribusi yang hampir simetris atau normal, yang memperlihatkan penyebaran data yang lebih konsisten pada atribut tersebut.

Hasil dari visualisasi korelasi antar fitur (heatmap) menunjukkan bahwa Glucose memiliki korelasi tertinggi dengan variabel target Outcome (r = 0. 49), diikuti oleh BMI (r = 0. 31) dan Age (r = 0. 24). Ini menunjukkan bahwa kadar gula darah, indeks massa tubuh, dan usia memberikan pengaruh yang penting terhadap status diabetes pasien. Temuan ini diperkuat oleh evaluasi model Random Forest, di mana ketiga fitur tersebut juga memiliki nilai feature importance tertinggi, menandakan bahwa baik dari segi statistik maupun algoritma, fitur-fitur ini paling relevan dalam proses prediksi.

Selain itu, ditemukan nilai-nilai yang tidak logis seperti nol pada kolom medis (contohnya tekanan darah = 0), yang telah berhasil ditangani melalui proses imputasi dengan menggunakan nilai median. Distribusi kelas pada label Outcome juga tidak seimbang, dengan jumlah pasien non-diabetes lebih banyak dibandingkan pasien diabetes. Untuk mengatasi hal ini, digunakan teknik SMOTE (Synthetic Minority Over-sampling Technique) yang berhasil menyeimbangkan jumlah data antara kedua kelas dan memastikan model tidak berpihak pada mayoritas.

Secara keseluruhan, eksplorasi awal ini menunjukkan bahwa dataset memiliki quality yang cukup baik untuk digunakan dalam pemodelan prediktif, dengan beberapa fitur kunci yang telah diidentifikasi dan strategi pengelolaan data yang tepat untuk menjaga kevalidan analisis.

3.4 DATA PREPARATION

Langkah awal adalah membersihkan data, di mana nilai-nilai yang tidak logis seperti nol pada atribut medis seperti Glukosa, Tekanan Darah, Ketebalan Kulit, Insulin, dan BMI diidentifikasi dan dianggap sebagai nilai yang hilang. Selanjutnya, dilakukan imputasi pada nilai-nilai yang hilang tersebut dengan menggunakan metode median agar distribusi data tetap terjaga dan tidak terpengaruh oleh outlier. Selain itu, data juga diperiksa untuk kemungkinan adanya baris yang duplikat, meskipun dalam kasus ini tidak ditemukan jumlah yang signifikan.

Karena semua fitur dalam dataset berupa angka, tahap pengkodean untuk data kategorikal tidak diperlukan.

Tahap berikutnya adalah melakukan normalisasi atau standardisasi pada data numerik. Tujuan dari langkah ini adalah agar semua fitur (kolom data) memiliki skala atau rentang nilai yang serupa, sehingga tidak ada fitur yang mendominasi model hanya karena memiliki nilai yang jauh lebih besar. Dalam penelitian ini, metode StandardScaler digunakan, yaitu teknik yang mengubah nilai setiap fitur menjadi bentuk dengan rata-rata nol dan deviasi standar satu.

Terakhir, proses pemisahan data dilakukan menjadi dua bagian, yakni set data pelatihan dan set data pengujian dengan perbandingan 80:20. Pemisahan ini menerapkan metode stratified sampling untuk mempertahankan proporsi kelas target tetap seimbang di kedua subset, sehingga evaluasi model menjadi lebih representatif dan tidak bias terhadap distribusi label.

3.5 MODELING

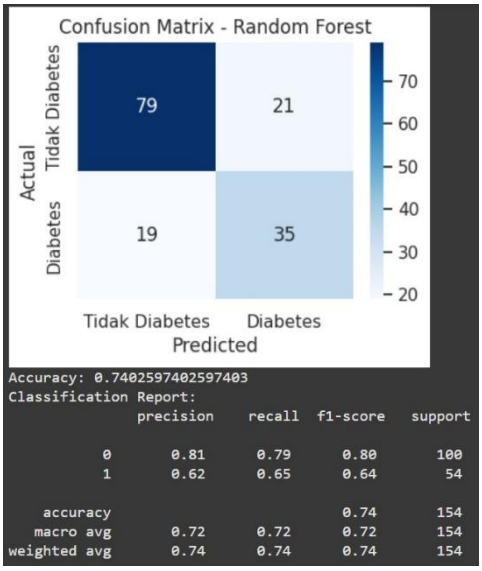
Pada fase ini, dilakukan pemodelan untuk menciptakan sistem klasifikasi yang dapat meramalkan kemungkinan seseorang menderita diabetes berdasarkan atribut-atribut yang ada dalam dataset. Algoritma yang diterapkan dalam penelitian ini adalah Random Forest Classifier.

3.5.1 PEMILIHAN ALGORITMA

Penggunaan Random Forest didasari oleh beberapa alasan. Random Forest adalah algoritma pembelajaran ensemble yang mengintegrasikan sejumlah pohon keputusan untuk menghasilkan prediksi yang lebih akurat dan stabil(Faisal & Budi Santoso, 2025). Algoritma ini memiliki keunggulan dalam mengatasi dataset dengan fitur numerik, mampu menangani hubungan non-linear antar atribut, dan kurang peka terhadap outlier serta nilai yang hilang. Selain itu, Random Forest juga dapat menawarkan wawasan penting tentang fitur mana yang paling berpengaruh terhadap prediksi, sehingga sangat membantu dalam interpretasi hasil model.

Hasil evaluasi model mencakup akurasi, confusion matrix, serta laporan klasifikasi (yang mencakup precision, recall, dan f1-score) untuk memberikan gambaran keseluruhan mengenai kinerja model dalam mengklasifikasikan pasien diabetes. Berdasarkan hasil yang diperoleh, Random Forest terbukti mampu memberikan kinerja prediktif yang baik dan konsisten pada dataset ini.

3.6 EVALUASI



Gambar 7 Confussion Matrix

Berdasarkan analisis yang dilakukan menggunakan model Random Forest, metrik kinerja yang diperoleh adalah sebagai berikut:

Tingkat akurasi model mencapai 74%, yang menunjukkan bahwa 74 dari 100 prediksi berhasil diklasifikasikan dengan akurat.

Confusion Matrix menunjukkan:

- 79 pasien yang tidak terdiagnosis diabetes dikenali dengan tepat (True Negative),
- 21 pasien yang seharusnya tidak diabetes salah diprediksi sebagai diabetes (False Positive),
- 35 pasien diabetes teridentifikasi dengan baik (True Positive),
- 19 pasien diabetes salah terklasifikasi sebagai bukan diabetes (False Negative)(*Diabetes*, n.d.).

Precision:

- Untuk kategori 0 (Tidak Diabetes): 0. 81 → dari seluruh prediksi "tidak diabetes", 81% adalah akurat.
- Untuk kategori 1 (Diabetes): 0. 62 → dari seluruh prediksi "diabetes", hanya 62% yang tepat.

Recall:

- Kategori 0: 0. 79 → dari total pasien yang sebenarnya tidak diabetes, 79% teridentifikasi.
- Kategori 1: 0. 65 → dari semua pasien yang benar-benar menderita diabetes, hanya 65% terdeteksi(Teknika & Ria Supriyatna, n.d.).

F1-Score:

- Kategori 0: 0. 80 → menunjukkan keseimbangan yang baik antara precision dan recall.
- Kategori 1: 0. 64 → kinerja model dalam mengenali pasien diabetes perlu ditingkatkan.
- Rata-rata Macro dan Rata-rata Terbobot dari seluruh metrik berada di antara 0. 72 hingga 0. 74, merefleksikan kinerja yang cukup konsisten meskipun terdapat perbedaan dalam performa antar kategori.

BAB IV

KESIMPULAN DAN SARAN

4.1 KESIMPULAN

Berdasarkan hasil pemodelan menggunakan algoritma Random Forest pada dataset diabetes, diperoleh kinerja model sebagai berikut, Akurasi model mencapai 74%, yang menunjukkan bahwa model ini memiliki kemampuan prediksi yang cukup memadai secara umum. Model dapat lebih efektif mengklasifikasikan pasien yang tidak terdiagnosis diabetes (precision: 0. 81, recall: 0. 79). Namun, kinerjanya dalam mengidentifikasi pasien yang menderita diabetes masih tergolong rendah (precision: 0. 62, recall: 0. 65), yang menandakan masih ada kasus diabetes yang tidak terdeteksi dengan tepat (false negative yang cukup tinggi). Dari evaluasi ini, dapat disimpulkan bahwa tujuan proyek untuk menciptakan model prediksi diabetes telah tercapai pada tingkat dasar, tetapi masih perlu ditingkatkan, terutama dalam hal sensitivitas terhadap penderita diabetes.

Model ini menggunakan teknik SMOTE untuk mengatasi masalah ketidakseimbangan data, sehingga distribusi kelas menjadi lebih seimbang selama pelatihan. Model cukup efektif dalam mengenali pasien yang tidak mengalami diabetes, dengan nilai f1-score yang tinggi untuk kelas tersebut. Algoritma Random Forest secara alami dapat menangani variabel numerik dan tidak begitu rentan terhadap outlier. Recall yang rendah pada kelas 1 (pasien diabetes), sehingga model masih berisiko melewatkan beberapa pasien diabetes (false negative). Proses penyetelan hyperparameter yang optimal belum dilakukan. Masih menggunakan satu algoritma (Random Forest), sehingga belum ada analisis perbandingan dengan model lain. Fitur yang dimanfaatkan masih terbatas pada atribut standar dari dataset; tidak ada fitur tambahan dari sumber lain.

4.2 SARAN

Hasil evaluasi menunjukkan bahwa nilai precision dan recall pada kelas 1 (pasien diabetes) masih cukup rendah. Ini menandakan bahwa model belum optimal dalam mendeteksi pasien diabetes, dan sering terjadi kesalahan dalam memprediksi siapa yang sebenarnya menderita diabetes. Hal ini penting untuk diperhatikan, terutama jika model ini akan digunakan di bidang kesehatan atau medis. Oleh karena itu, disarankan untuk mempertimbangkan perbandingan model Random Forest dengan algoritma lain, seperti Logistic Regression, Support Vector Machine (SVM), atau XGBoost. Tujuannya adalah untuk mengetahui apakah model lain dapat

menghasilkan kinerja yang lebih baik, khususnya dalam meningkatkan kemampuan model untuk mengidentifikasi penderita diabetes. Dengan membandingkan berbagai algoritma, kita dapat memilih model yang paling sesuai dan memberikan kinerja yang lebih seimbang.

DAFTAR PUSTAKA

Diabetes. (n.d.). Retrieved July 5, 2025, from https://www.who.int/news-room/fact-sheets/detail/diabetes

Dikan Ismafillah, Tatang Rohana, & Yana Cahyana. (2023). Analisis algoritma pohon keputusan untuk memprediksi penyakit diabetes menggunakan oversampling smote. *INFOTECH : Jurnal Informatika & Teknologi*, *4*(1), 27–36. https://doi.org/10.37373/infotech.v4i1.452

Faisal, M., & Budi Santoso, I. (2025). Algoritma Random Forest dan Synthetic Minority Oversampling Technique (SMOTE) untuk Deteksi Diabetes. In *Jurnal Informatika Sunan Kalijaga)* (Vol. 10, Issue 2). MEI.

Febrian, M. G., Ferdiansyah, R., Nugraha, E. A., Satriatama, D., & Kusumastuti, R. (2024). *PREDIKSI RISIKO DIABETES MENGGUNAKAN ALGORITMA DECISION TREE DENGAN APLIKASI RAPID MINER*.

Ikram, R., Khan, A., Zahri, M., Saeed, A., Yavuz, M., & Kumam, P. (2022). Extinction and stationary distribution of a stochastic COVID-19 epidemic model with time-delay. *Computers in Biology and Medicine*, *141*, 105115. https://doi.org/10.1016/J.COMPBIOMED.2021.105115

Navarro-Cáceres, M., Sánchez-Jara, J. F. M., Quietinho Leithardt, V. R., & García-Ovejero, R. (2020). Assistive model to generate chord progressions using genetic programming with artificial immune properties. *Applied Sciences (Switzerland)*, *10*(17). https://doi.org/10.3390/app10176039

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should i trust you?" Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-August-2016*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Teknika, J., & Ria Supriyatna, A. (n.d.). Teknika 17 (1): 163-172 Prediksi Penyakit Diabetes Menggunakan Algoritma Random Forest. *IJCCS*, *x, No.x*, 1–5.