LAPORAN AKHIR ASSURANCE OF LEARNING

MATA KULIAH DATA MINING

BINUS UNIVERSITY GANJIL 2024

KLASIFIKASI DAN PREDIKSI POTENSI DIABETES PADA PASIEN DI INDONESIA

Disusun oleh:

Pradipa Javier Fatah	2602158815	COMPUTER SCIENCE
Muhammad Alvin	2602135603	COMPUTER SCIENCE
Nathanael Sanliago Suhardjo	2602190772	COMPUTER SCIENCE
Ngui Su Ying	2602141051	COMPUTER SCIENCE
Nixen Jenio	2602159982	COMPUTER SCIENCE

UNIVERSITAS BINA NUSANTARA

JAKARTA

2024

DAFTAR ISI

BAB I
PENDAHULUAN

1.1 Latar Belakang

Dilansir dari laman *voaindonesia.com*, jumlah penderita diabetes di Indonesia terus meningkat. Hal ini didukung oleh data International Diabetes Federation (IDF) yang menunjukkan bahwa pada tahun 2021, Indonesia termasuk dalam lima negara dengan jumlah penderita diabetes tertinggi, yaitu sebanyak 19,5 juta jiwa. Sebanyak 45% hingga 58,8% dari penderita diabetes tidak terdiagnosis, terutama pada penyandang diabetes melitus tipe 2. Dalam 10 tahun terakhir, angka penderita diabetes meningkat secara signifikan, dan tren ini diprediksi akan terus berlanjut pada masa mendatang.

Pada tahun 2021, jumlah penderita diabetes di Indonesia tercatat sebanyak 19,47 juta jiwa. Angka ini diperkirakan akan terus meningkat setiap tahun. Pada tahun 2035, estimasi jumlah penderita diabetes secara global diproyeksikan mencapai 592 juta jiwa. Penyakit kronis ini juga menjadi salah satu penyebab kematian tertinggi di Indonesia. Berdasarkan data dari Institute for Health Metrics and Evaluation, diabetes menjadi penyebab kematian tertinggi ketiga di Indonesia pada tahun 2019, dengan angka kematian mencapai 57,42% per 100.000 penduduk. Oleh karena itu, jika permasalahan ini tidak segera ditangani, angka penderita diabetes akan terus meningkat setiap tahunnya.

Melakukan analisis dan penambangan data kesehatan dapat membantu mengidentifikasi deteksi dini diabetes, meningkatkan efektivitas pengelolaan penyakit, serta memperbaiki kualitas hidup pasien. Analisis ini dapat didasarkan pada rincian demografis, faktor gaya hidup, riwayat medis, pengukuran klinis, penggunaan obat, gejala, skor kualitas hidup, paparan lingkungan, dan perilaku kesehatan. Dengan menerapkan metode *data mining* pada kumpulan data, tenaga medis dapat memprediksi diagnosis diabetes dengan lebih akurat, sehingga membantu dalam pengambilan keputusan klinis. Langkah ini memungkinkan pasien untuk mengambil tindakan pencegahan lebih awal guna menghindari komplikasi di masa depan. Oleh karena itu, analisis dan penambangan data pada data kesehatan sangat penting untuk mendukung tenaga medis dan meningkatkan kualitas hidup individu.

1.2 Rumusan Masalah

1. Mengapa jumlah penderita diabetes di Indonesia terus meningkat, dan apa saja faktor-faktor yang mempengaruhinya?
2. Bagaimana penerapan metode *data mining* dapat membantu mendeteksi risiko diabetes secara dini?
3. Bagaimana *data mining* dapat meningkatkan efektivitas pengelolaan dan pengambilan keputusan klinis bagi penderita diabetes?
4. Bagaimana metode *machine learning*, seperti XGBoost dan Random Forest, dapat digunakan untuk memprediksi risiko diabetes secara lebih akurat?

1.3 Tujuan Penelitian

1. Menganalisis faktor-faktor yang memengaruhi peningkatan jumlah penderita diabetes di Indonesia.
2. Mengkaji manfaat penerapan metode *data mining* dalam mendeteksi dini risiko diabetes.
3. Mengidentifikasi cara *data mining* dapat meningkatkan efektivitas pengelolaan dan pengambilan keputusan klinis untuk penderita diabetes.
4. Mengembangkan dan mengevaluasi model prediksi risiko diabetes menggunakan algoritma *machine learning*, seperti XGBoost dan Random Forest, untuk meningkatkan akurasi deteksi dini dibandingkan metode konvensional.

1.4 Manfaat Penelitian

1. Bagi tenaga medis, memberikan wawasan untuk mendeteksi risiko diabetes lebih dini dan mengambil keputusan klinis secara tepat.
2. Bagi pasien, membantu pasien memahami faktor risiko dan mengambil langkah preventif untuk mengelola kondisi mereka secara lebih efektif.
3. Bagi peneliti, menyediakan data dan model prediktif yang dapat dijadikan referensi untuk penelitian lanjutan di bidang kesehatan, khususnya terkait diabetes.

4. Bagi masyarakat umum, meningkatkan kesadaran akan pentingnya pengelolaan gaya hidup untuk mencegah risiko diabetes.
5. Bagi peneliti selanjutnya, membuka peluang untuk pengembangan model yang lebih kompleks atau eksplorasi algoritma lain dalam mendeteksi dan mengelola penyakit kronis.

BAB II
TINJAUAN PUSTAKA

2.1 Diabetes

Diabetes salah satu penyakit kronis yang disebabkan oleh kadar gula darah yang melebihi batas normal manusia pada umumnya. Kondisi ini terjadi ketika tubuh sudah tidak mampu mengambil glukosa dan diubah menjadi energi, sehingga terjadi gula menumpuk dalam aliran darah tubuh. Jika pengidap diabetes tidak mengontrol konsumsi gula tiap harinya, penyakit akan semakin parah dan menyebabkan kerusakan pada berbagai organ dan jaringan tubuh.

2.1.1 Tipe Diabetes

a. Diabetes tipe 1

   Pada tipe ini, sistem imun tubuh akan menyerang dirinya sendiri yaitu penyakit autoimun, sehingga pankreas tidak bisa menghasilkan insulin dan memerlukan suntikan insulin secara rutin.

b. Diabetes tipe 2

   Diabetes tipe ini, tubuh tidak membuat cukup insulin atau merespon insulin secara normal sehingga gula menumpuk di dalam darah.

c. Diabetes gestasional

   Diabetes yang terjadi pada ibu hamil, yang disebabkan oleh perubahan hormon selama kehamilan.

d. Prediabetes

   Jenis diabetes ini, ketika kondisi gula dalam darah lebih tinggi dari normal namun tidak setinggi penderita diabetes tipe 2.

2.1.2 Penyebab Diabetes

Diabetes disebabkan oleh tingginya kadar gula darah di dalam tubuh. Kadar gula darah normal pada umumnya berada di bawah 100 mg/dL. Namun, jika kadar gula darah mencapai 100–125 mg/dL, kondisi ini sudah termasuk dalam kategori prediabetes. Seseorang dapat dikategorikan sebagai penderita diabetes apabila kadar gula darahnya mencapai 126 mg/dL atau lebih. Kadar gula

darah yang tinggi, yang dikenal sebagai hiperglikemia, merupakan faktor utama yang sangat memengaruhi diabetes.

2.1.3 Faktor Diabetes

Faktor diabetes terbagi menjadi dua kategori, yaitu tipe 1 dan tipe 2. Faktor diabetes tipe 1 disebabkan oleh keturunan, usia, geografis, dan faktor lainnya. Faktor diabetes tipe 1 sering kali terdeteksi pada anak-anak berusia 4–7 tahun dan 10–14 tahun. Faktor geografis juga memengaruhi, karena orang yang tinggal di daerah jauh dari garis khatulistiwa cenderung mendapatkan paparan sinar matahari lebih sedikit. Kekurangan cahaya matahari ini dapat menyebabkan defisiensi vitamin D, yang berperan dalam memicu penyakit autoimun. Faktor lainnya termasuk konsumsi susu sapi terlalu dini, air minum yang mengandung natrium nitrat, konsumsi sereal atau gluten sebelum usia 4–7 bulan, serta riwayat menderita penyakit kuning saat lahir.

Faktor diabetes tipe 2 disebabkan oleh berat badan berlebih atau obesitas, kurangnya aktivitas fisik, ras, usia, dan riwayat diabetes selama kehamilan. Ras kulit hitam memiliki angka pengidap diabetes tipe 2 yang lebih tinggi dibandingkan ras kulit putih. Selain itu, usia juga merupakan faktor penting, dengan diabetes tipe 2 lebih sering ditemukan pada individu berusia di atas 45 tahun.

2.1.4 Gejala Diabetes

Gejala diabetes berbeda-beda pada setiap pengidapnya, tergantung tingkat keparahan dan jenis penyakit gula yang diidap. Gejala-gejala berikut biasanya dialami oleh pengidap diabetes pada umumnya. Pengidap diabetes biasanya mudah sekali lelah, frekuensi buang air kecil meningkat, rasa haus yang meningkat, mengalami gangguan penglihatan, infeksi tubuh, berat badan yang terus menurun, dan adanya keton dalam urine.

2.2 Data Mining dalam Prediksi Diabetes

Menurut laman voaindonesia.com, jumlah penderita diabetes di Indonesia terus meningkat. Data dari International Diabetes Federation (IDF) tahun 2021 menunjukkan bahwa Indonesia termasuk dalam lima negara dengan jumlah penderita diabetes tertinggi,

yaitu sebanyak 19,5 juta penderita. Apabila tidak dilakukan penanganan sejak dini, jumlah penderita diabetes diprediksi akan terus meningkat secara signifikan.

Oleh karena itu, penerapan *data mining* dalam prediksi diabetes dapat menjadi solusi untuk mengatasi permasalahan tingginya angka penderita diabetes di Indonesia. Teknologi ini dapat membantu mendeteksi penyakit diabetes secara dini, mendukung pengambilan keputusan klinis oleh tenaga medis, dan memungkinkan pasien mengambil langkah pencegahan untuk menghindari komplikasi. Dengan model prediksi yang akurat, efektivitas pengelolaan penyakit diabetes dapat meningkat, sehingga kualitas hidup pasien menjadi lebih baik.

2.3 Himpunan Data

Untuk melakukan penelitian tentang memprediksi diabetes, dimulai dari mengumpulkan data yang relevan. Himpunan data untuk penelitian ini dilansir dari laman Kaggle.com dengan judul "Diabetes Predictive Analytics", selain itu untuk membantu penelitian ini, data juga dilansir dari Chatgpt.com. Setelah sumber data melalui proses pengumpulan, data integrasi menjadi satu dataset. Data akan di praproses, dimulai dengan melengkapi yang kosong, menentukan format, dan menghapus atau menambahkan data.

Data akan diolah untuk proses pelatihan dan validasi data. Terdapat 1879 data pada kumpulan data, 80% dari data yaitu 1503 data akan masuk ke dalam proses pelatihan data dan sisanya sebanyak 20% yaitu 375 data akan masuk ke dalam proses validasi model. Proses pelatihan pada data memberikan gambaran algoritma pada *machine learning*, sehingga dapat memvalidasi hasil *machine learning*.

2.4 Atribut Data

2.4.1 Informasi Pasien

- **PatientID**: Identifikasi unik yang diberikan kepada setiap pasien (6000 hingga 7878)

2.4.2 Rincian Demografis

- **Age**: Usia pasien berkisar dari 20 hingga 90 tahun.
- **Gender**: Jenis kelamin pasien, di mana 0 mewakili Laki-laki dan 1 mewakili Perempuan.
- **Ethnicity**: Etnisitas pasien, dikodekan sebagai berikut:
    - 0: Kaukasia
    - 1: Afrika Amerika
    - 2: Asia
    - 3: Lainnya
- **SocioeconomicStatus**: Status sosial ekonomi pasien, dikodekan sebagai berikut:
    - 0: Rendah
    - 1: Menengah
    - 2: Tinggi
- **EducationLevel**: Tingkat pendidikan pasien, dikodekan sebagai berikut:
    - 0: Tidak ada
    - 1: Sekolah Menengah
    - 2: Sarjana
    - 3: Lebih Tinggi

2.4.3 Faktor Gaya Hidup

- **BMI**: Indeks Massa Tubuh pasien, berkisar dari 15 hingga 40.
- **Smoking**: Status merokok, di mana 0 menunjukkan Tidak dan 1 menunjukkan Ya.
- **AlcoholConsumption**: Konsumsi alkohol mingguan dalam satuan, berkisar dari 0 hingga 20.
- **PhysicalActivity**: Aktivitas fisik mingguan dalam jam, berkisar dari 0 hingga 10.

- **DietQuality**: Skor kualitas diet, berkisar dari 0 hingga 10.
- **SleepQuality**: Skor kualitas tidur, berkisar dari 4 hingga 10.

2.4.4 Riwayat Medis

- **FamilyHistoryDiabetes**: Riwayat keluarga dengan diabetes, di mana 0 menunjukkan Tidak dan 1 menunjukkan Ya.
- **GestationalDiabetes**: Riwayat diabetes gestasional, di mana 0 menunjukkan Tidak dan 1 menunjukkan Ya.
- **PolycysticOvarySyndrome**: Kehadiran sindrom ovarium polikistik, di mana 0 menunjukkan Tidak dan 1 menunjukkan Ya.
- **PreviousPreDiabetes**: Riwayat pre-diabetes sebelumnya, di mana 0 menunjukkan Tidak dan 1 menunjukkan Ya.
- **Hypertension**: Kehadiran hipertensi, di mana 0 menunjukkan Tidak dan 1 menunjukkan Ya.

2.4.5 Pengukuran Klinis

- **SystolicBP**: Tekanan darah sistolik, berkisar dari 90 hingga 180 mmHg.
- **DiastolicBP**: Tekanan darah diastolik, berkisar dari 60 hingga 120 mmHg.
- **FastingBloodSugar**: Tingkat gula darah puasa, berkisar dari 70 hingga 200 mg/dL.
- **HbA1c**: Tingkat Hemoglobin A1c, berkisar dari 4,0% hingga 10,0%.
- **SerumCreatinine**: Tingkat kreatinin serum, berkisar dari 0,5 hingga 5,0 mg/dL.
- **BUNLevels**: Tingkat Urea Nitrogen Darah, berkisar dari 5 hingga 50 mg/dL.
- **CholesterolTotal**: Tingkat kolesterol total, berkisar dari 150 hingga 300 mg/dL.
- **CholesterolLDL**: Tingkat kolesterol LDL (*Low-density lipoprotein*), berkisar dari 50 hingga 200 mg/dL.
- **CholesterolHDL**: Tingkat kolesterol HDL (*High-density lipoprotein*), berkisar dari 20 hingga 100 mg/dL.
- **CholesterolTriglycerides**: Tingkat trigliserida, berkisar dari 50 hingga 400 mg/dL.

2.4.6 Obat-obatan

- **AntihypertensiveMedications**: Penggunaan obat antihipertensi, di mana 0 menunjukkan Tidak dan 1 menunjukkan Ya.
- **Statins**: Penggunaan statin, di mana 0 menunjukkan Tidak dan 1 menunjukkan Ya.
- **AntidiabeticMedications**: Penggunaan obat antidiabetes, di mana 0 menunjukkan Tidak dan 1 menunjukkan Ya.

2.4.7 Gejala dan Kualitas Hidup

- **FrequentUrination**: Kehadiran sering buang air kecil, di mana 0 menunjukkan Tidak dan 1 menunjukkan Ya.
- **ExcessiveThirst**: Kehadiran rasa haus yang berlebihan, di mana 0 menunjukkan Tidak dan 1 menunjukkan Ya.
- **UnexplainedWeightLoss**: Kehadiran penurunan berat badan yang tidak dapat dijelaskan, di mana 0 menunjukkan Tidak dan 1 menunjukkan Ya.
- **FatigueLevels**: Tingkat kelelahan, berkisar dari 0 hingga 10.
- **BlurredVision**: Kehadiran penglihatan kabur, di mana 0 menunjukkan Tidak dan 1 menunjukkan Ya.
- **SlowHealingSores**: Kehadiran luka yang sembuh dengan lambat, di mana 0 menunjukkan Tidak dan 1 menunjukkan Ya.
- **TinglingHandsFeet**: Kehadiran kesemutan di tangan atau kaki, di mana 0 menunjukkan Tidak dan 1 menunjukkan Ya.
- **QualityOfLifeScore**: Skor kualitas hidup, berkisar dari 0 hingga 100.

2.4.8 Paparan Lingkungan dan Pekerjaan

- **HeavyMetalsExposure**: Paparan logam berat, di mana 0 menunjukkan Tidak dan 1 menunjukkan Ya.
- **OccupationalExposureChemicals**: Paparan bahan kimia berbahaya dalam pekerjaan, di mana 0 menunjukkan Tidak dan 1 menunjukkan Ya.

- **WaterQuality**: Kualitas air, di mana 0 menunjukkan Baik dan 1 menunjukkan Buruk.

2.4.9 Perilaku Kesehatan

- **MedicalCheckupsFrequency**: Frekuensi pemeriksaan medis per tahun, berkisar dari 0 hingga 4.
- **MedicationAdherence**: Skor kepatuhan terhadap pengobatan, berkisar dari 0 hingga 10.
- **HealthLiteracy**: Skor literasi kesehatan, berkisar dari 0 hingga 10.

2.4.10 Informasi Diagnosa

- **Diagnosis**: Status diagnosis untuk Diabetes, di mana 0 menunjukkan Tidak dan 1 menunjukkan Ya.
- 6. **LifestyleScore** (*float*): Skor gabungan gaya hidup:
    - Formula: $\text{DietQuality} + \text{SleepQuality} - \text{Smoking} - (\text{AlcoholConsumption}/2)$
  7. **DiabetesRiskScore** (*float*): Skor risiko diabetes:
    - Formula: $\text{BMI} + \text{HbA1c} + (\text{FamilyHistoryDiabetes} \times 5) + (\text{Hypertension} \times 2)$
  8. **TreatmentRecommendation** (*int*): Rekomendasi awal untuk pengobatan:
    - 0: Tidak perlu pengobatan (risiko rendah)
    - 1: Lifestyle Modification (risiko sedang)
    - 2: Obat (risiko tinggi)
    - 3: Obat dan Pemantauan Ketat (risiko sangat tinggi)
  9. **MedicationPriorityScore** (*float*): Skor prioritas pengobatan:
    - Formula: $(\text{BloodPressureRisk} \times 2) + (\text{CardiovascularRisk} \times 3) + \text{DiabetesRiskScore}$

BAB III
METODOLOGI PENELITIAN

3.1 XGBoost Classifier

XGBoost (*eXtreme Gradient Boosting*) adalah algoritma *machine learning* berbasis boosting yang dirancang untuk menghasilkan model yang efisien, cepat, dan skalabel. Algoritma ini sering digunakan untuk menyelesaikan berbagai masalah, baik klasifikasi maupun regresi. XGBoost merupakan pengembangan dari metode *boosting*, dengan optimasi khusus untuk menangani kumpulan data berukuran besar. Dalam penelitian ini, XGBoost Classifier digunakan sebagai salah satu metode utama.

3.2 Random Forest Classifier

Random Forest adalah algoritma *ensemble learning* yang dikembangkan oleh Leo Breiman dan Adele Cutler. Metode ini menggabungkan output dari beberapa pohon keputusan (*decision tree*) untuk meningkatkan akurasi prediksi. Algoritma ini mampu beradaptasi dengan baik pada masalah klasifikasi maupun regresi.

Random Forest menggunakan teknik *bagging*, di mana data latih diambil secara acak untuk membangun beberapa pohon keputusan. Setelah itu, prediksi akhir diperoleh berdasarkan hasil pemungutan suara dari seluruh pohon. Hal ini membuat Random Forest memiliki akurasi tinggi dan mampu mengurangi risiko *overfitting*.

3.3 Decision Tree Classifier

Decision Tree Regressor merupakan metode *machine learning* yang bertujuan untuk memprediksi nilai kontinu berdasarkan serangkaian aturan keputusan yang dibangun dalam bentuk struktur pohon. Tahap awal dari metode ini dimulai dengan membagi dataset menjadi subset-subset yang semakin kecil berdasarkan nilai dari fitur-fitur yang ada. Pemilihan fitur yang paling baik dilakukan dengan memilih kriteria yang paling efektif dalam memisahkan nilai target di dalam setiap subset, seperti *Mean Squared Error* (MSE) atau *Mean Absolute Error* (MAE). Proses ini akan terus berlanjut untuk membangun struktur pohon keputusan, di mana setiap simpul merupakan keputusan berdasarkan fitur-fitur tertentu dan setiap cabang mewakili kemungkinan nilai

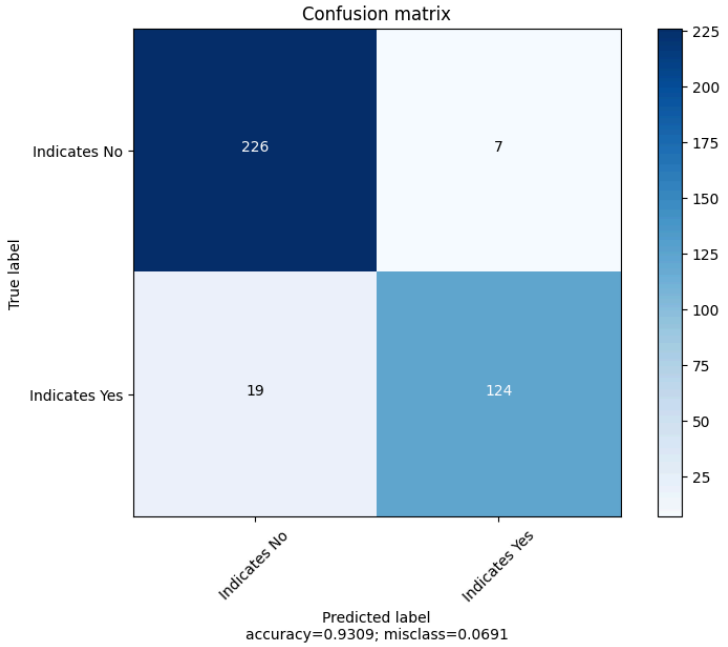
dari fitur tersebut. Kemudian, data baru dapat diprediksi dengan dijalankan melalui pohon keputusan yang sudah dibuat. Prediksi akhir diberikan berdasarkan nilai rata-rata atau median dari nilai target di simpul terminal yang sesuai.

Decision Tree Classifier adalah algoritma *machine learning* yang digunakan untuk tugas klasifikasi dengan struktur berbentuk pohon keputusan. Pohon ini terdiri dari simpul yang merepresentasikan fitur, cabang yang menunjukkan aturan keputusan, dan daun yang berisi kelas atau label hasil prediksi. Cara kerjanya dimulai dengan memilih fitur yang paling signifikan untuk memisahkan data berdasarkan kriteria seperti entropi atau gini index. Dataset kemudian dibagi secara rekursif ke dalam subset yang lebih kecil pada setiap simpul hingga mencapai simpul daun, di mana setiap daun mewakili sebuah kategori atau label. Proses ini menghasilkan model berbentuk hirarki aturan keputusan, yang dapat digunakan untuk memprediksi label data baru dengan menelusuri jalur dari akar ke daun berdasarkan nilai-nilai fitur input.

BAB IV
HASIL DAN PEMBAHASAN

4.1 Hasil Prediksi Model XGBoost Classifier

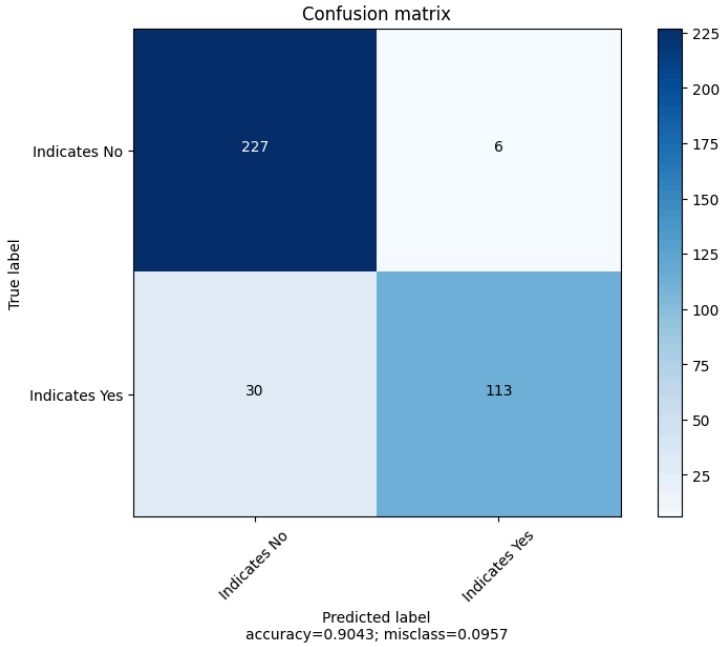
Hasil analisis model XGBoost Classifier menunjukkan performa yang sangat baik dengan nilai ROC AUC sebesar 0.9551, yang menunjukkan kemampuan model membedakan kelas positif dan negatif secara efektif. Nilai KS score yang tinggi (0.8549) dan p-value yang kecil menegaskan bahwa model dapat memisahkan distribusi kedua kelas dengan jelas. Berdasarkan *output*, model menunjukkan *precision* 0.92 dan *recall* 0.97 untuk kelas negatif (0), serta *precision* 0.95 dan *recall* 0.87 untuk kelas positif (1). Meskipun *recall* untuk kelas positif sedikit lebih rendah, akurasi keseluruhan mencapai 93%, dengan F1-score yang tinggi (0.95 untuk kelas negatif dan 0.91 untuk kelas positif), mengindikasikan keseimbangan yang baik antara *precision* dan *recall*.



Gambar 4.1 Confusion Matrix dari Model XGBoost Classifier

4.2 Hasil Prediksi Model Random Forest Classifier

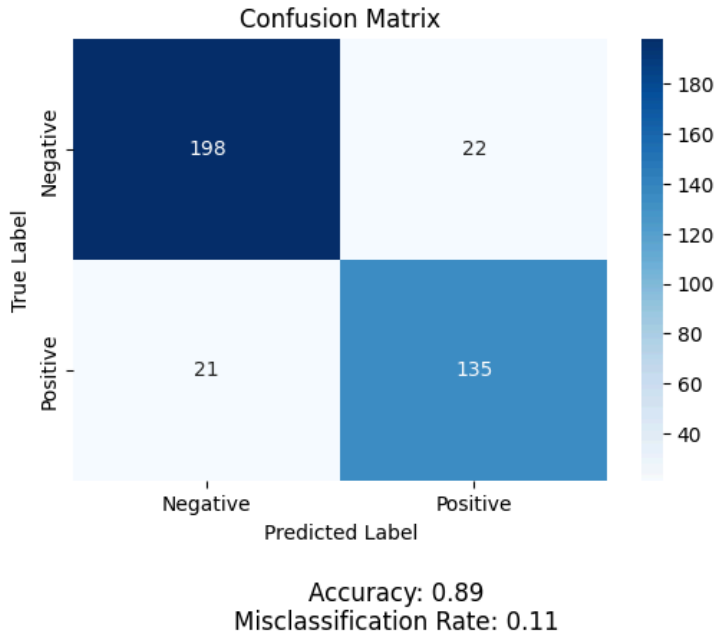
Model Random Forest Classifier menunjukkan performa yang baik dengan AUC sebesar 0.9444, yang menunjukkan kemampuan diskriminasi yang sangat baik antara kelas positif dan negatif. *Precision* untuk kelas positif tinggi, namun *recall* sedikit lebih rendah, menunjukkan model cenderung menghindari kesalahan positif tetapi terkadang melewatkan beberapa kasus positif. Untuk kelas negatif, model menunjukkan keseimbangan yang baik antara *precision* dan *recall*, dengan akurasi keseluruhan mencapai 91%. Nilai F1-Score yang tinggi menandakan keseimbangan yang baik antara akurasi dan deteksi. Perbaikan seperti penyesuaian *threshold* atau teknik *oversampling* pada kelas positif dapat meningkatkan deteksi kasus positif. Penyesuaian *threshold* atau teknik *oversampling* pada kelas minoritas.



Gambar 4.2 Confusion Matrix dari Model Random Forest Classifier

4.3 Hasil Prediksi Model Decision Tree Classifier

Model DecisionTreeClassifier menunjukkan hasil yang cukup baik dengan *accuracy* sebesar 0.89. *Precision* dan *recall* untuk kelas 0 (negatif) cukup tinggi, yaitu 0.90 dan 0.90, sementara untuk kelas 1 (positif), *precision* adalah 0.86 dan *recall* 0.87. Ini menunjukkan bahwa model dapat mengenali kelas 1 dengan baik, namun ada sedikit kekurangan dalam hal *precision* pada kelas 1, yang berarti model terkadang salah mengklasifikasikan beberapa kasus positif sebagai negatif. Selain itu, f1-score untuk kelas 1 adalah 0.86, yang masih cukup baik namun dapat ditingkatkan lebih jauh.



Gambar 4.3 Confusion Matrix dari Model Decision Tree Classifier

BAB V
EVALUASI

5.1. Evaluasi

5.1.1 Evaluasi Model XGBoost Classifier

Model *XGBoost Classifier* menunjukkan kinerja yang sangat baik, dengan ROC AUC sebesar 0.9551, yang menandakan kemampuan luar biasa model dalam membedakan antara kelas positif dan negatif. KS score sebesar 0.8549 dengan p-value yang sangat kecil juga menguatkan bahwa model mampu memisahkan prediksi untuk kedua kelas dengan sangat jelas.

Pada kelas negatif (0), model memiliki *precision* sebesar 0.92 dan recall sebesar 0.97, yang berarti model sangat andal dalam mengenali data negatif dengan tingkat kesalahan yang sangat rendah. Untuk kelas positif (1), *precision*-nya bahkan lebih tinggi, yaitu 0.95, menunjukkan sebagian besar prediksi positif benar adanya. Namun, *recall* untuk kelas ini sedikit lebih rendah, yaitu 0.87, yang mengindikasikan adanya beberapa kasus positif yang masih terlewat.

Dengan akurasi keseluruhan sebesar 93%, model ini terbukti sangat efektif dalam mengenali pola data secara umum. Nilai F1-score yang tinggi pada kedua kelas juga menunjukkan keseimbangan yang baik antara ketepatan prediksi (*precision*) dan kemampuan mendeteksi data yang benar (*recall*). Perlu ditingkatkannya *recall* pada kelas positif dengan menyesuaikan *threshold* prediksi atau menggunakan teknik seperti *oversampling* untuk menangani ketidakseimbangan data pada kedua kelas.

5.1.2 Evaluasi Model Random Forest Classifier

Hasil evaluasi model *Random Forest Classifier* menunjukkan kinerja yang sangat baik, dengan AUC mencapai 0.9444. Ini menunjukkan bahwa model dapat dengan efektif membedakan antara kelas positif dan negatif. Nilai KS score yang tinggi, yaitu 0.8441, dengan p-value yang sangat kecil, juga menunjukkan bahwa distribusi antara kedua kelas terpisah dengan jelas, menandakan kualitas

diskriminasi yang sangat baik. Precision untuk kelas positif mencapai 0.95, yang berarti hampir semua prediksi positif benar, meskipun *recall*-nya sedikit lebih rendah (0.79), yang menunjukkan bahwa model mungkin melewatkan beberapa kasus positif. Di sisi lain, untuk kelas negatif, model tampil sangat baik dengan precision 0.88, *recall* 0.97, dan F1-score 0.93. Dengan akurasi keseluruhan sebesar 91%, model ini sangat handal untuk aplikasi yang membutuhkan prediksi yang akurat untuk kedua kelas. Namun, rendahnya *recall* untuk kelas positif menunjukkan adanya ruang untuk perbaikan, misalnya dengan menurunkan *threshold* prediksi atau menggunakan teknik *oversampling* untuk meningkatkan deteksi kasus positif. Secara keseluruhan, model sudah sangat seimbang antara *precision* dan *recall*, namun ada potensi untuk meningkatkan deteksi kasus positif agar performa model menjadi lebih optimal, terutama untuk mendeteksi kasus-kasus yang lebih penting.

5.1.3 Evaluasi Model Decision Tree Classifier

Meskipun akurasi sudah baik di angka 0.89, namun masih ada ruang untuk perbaikan, terutama untuk meningkatkan keseimbangan antara precision dan recall. Dengan menggunakan hyperparameter tuning, pruning, atau feature engineering, kinerja model dapat ditingkatkan lebih jauh. Selain itu, dalam kasus ketidakseimbangan kelas, penggunaan teknik seperti SMOTE atau metode ensemble lainnya seperti Random Forest dapat membantu meningkatkan kemampuan model untuk mengenali kelas minoritas. Secara keseluruhan, meskipun model ini sudah cukup baik, beberapa perbaikan masih bisa dilakukan untuk meningkatkan hasil prediksi, terutama untuk kelas positif.

BAB VI
KESIMPULAN

6.1. Kesimpulan

Berdasarkan evaluasi terhadap tiga model yang digunakan (XGBoost, Random Forest, dan Decision Tree Classifier), dapat disimpulkan bahwa XGBoost Classifier memberikan hasil terbaik dengan akurasi 93%, ROC AUC 0.9551, dan keseimbangan yang sangat baik antara precision dan recall pada kedua kelas. Walaupun recall untuk kelas positif sedikit lebih rendah (0.87), hal ini masih bisa diperbaiki dengan teknik seperti penyesuaian threshold atau oversampling untuk meningkatkan deteksi kasus positif.

Random Forest Classifier juga menunjukkan hasil yang sangat baik dengan akurasi 91% dan ROC AUC 0.9444. Precision untuk kelas positif cukup tinggi (0.95), namun recallnya lebih rendah (0.79), yang berarti beberapa kasus positif tidak terdeteksi. Model ini bekerja sangat baik dalam mendeteksi kelas negatif, tetapi untuk kelas positif, masih ada potensi perbaikan melalui balancing data atau penyesuaian threshold.

Decision Tree Classifier memberikan hasil yang cukup baik dengan akurasi 89%. Precision dan recall untuk kedua kelas cukup seimbang, namun performanya masih sedikit di bawah XGBoost dan Random Forest. Model ini dapat ditingkatkan dengan melakukan tuning hyperparameter, pruning, atau menggabungkannya dalam metode ensemble seperti Random Forest untuk hasil yang lebih optimal.

Secara keseluruhan, **XGBoost Classifier** adalah model terbaik untuk kasus ini. Namun, untuk meningkatkan deteksi kelas positif, penyesuaian threshold, balancing data, dan teknik lainnya masih diperlukan, agar performa pada kelas negatif tetap optimal. Hal ini penting, terutama jika deteksi kelas minoritas memiliki dampak yang signifikan.

DAFTAR PUSTAKA

Fadli, R. (2024, Agustus 27). *Apa itu Diabetes? Gejala, Penyebab, dan Pengobatan*. Halodoc. Diakses pada 3 Desember 2024, dari https://www.halodoc.com/kesehatan/diabetes?srsltid=AfmBOorGUbladmKj-8UMeWrWxbHLj9O_AYpkjQIKDmlHrymmvJYgAAKH

Nareza, M. (2024, November 4). *Diabetes - Gejala, Penyebab, dan Pengobatan*. Alodokter. Diakses pada 3 Desember 2024, melalui https://www.alodokter.com/diabetes

Litha, Y. (2024, November 21). *Jumlah Penderita Diabetes di Indonesia Terus Meningkat*. VOA. Diakses pada 3 Desember 2024, melalui https://www.voaindonesia.com/a/jumlah-penderita-diabetes-di-indonesia-terus-meningkat/7870777.html

Universitas Gadjah Mada. (2023, January 16). *Diabetes Penyebab Kematian Tertinggi di Indonesia: Batasi dengan Snack Sehat Rendah Gula – Direktorat Pengembangan Usaha*. Direktorat Pengembangan Usaha. Diakses pada 3 Desember 2024, melalui https://ditpui.ugm.ac.id/diabetes-penyebab-kematian-tertinggi-di-indonesia-batasi-dengan-snack-sehat-rendah-gula/

CNN Indonesia. (2021, December 6). *Indonesia Masuk 5 Besar Negara Kasus Diabetes Tertinggi di Dunia*. CNN Indonesia. Diakses pada 3 Desember 2024, melalui https://www.cnnindonesia.com/gaya-hidup/20211206080008-255-730258/indonesia-masuk-5-besar-negara-kasus-diabetes-tertinggi-di-dunia

Fahmi, Q. (2023, Januari 4). *Banyak Masyarakat yang Tidak Terdiagnosis Diabetes Melitus*. Universitas Muhammadiyah Jakarta. Diakses pada 3 Desember 2024, melalui https://umj.ac.id/kabar-kampus/2023/01/banyak-masyarakat-yang-tidak-terdiagnosis-diabetes-melitus/

Saraswati, M. R. (2022, Agustus 5). *Diabetes Melitus Adalah Masalah Kita*. Kemenkes - Direktorat Jenderal Pelayanan Kesehatan. Diakses pada 3 Desember 2024, melalui https://yankes.kemkes.go.id/view_artikel/1131/diabetes-melitus-adalah-masalah-kita

Artanti, P., Masdar, H., & Rosdiana, D. (2015). *Angka kejadian diabetes melitus tidak terdiagnosis pada masyarakat kota Pekanbaru* (Doctoral dissertation, Riau University).

Geriatri.id. (2024, April 18). *Diabetes dengan Komplikasi Penyebab Kematian Tertinggi Ketiga di Indonesia*. Geriatri - Lansia Sehat Bahagia. Diakses pada 3 Desember 2024, melalui https://www.geriatri.id/artikel/2182/diabetes-dengan-komplikasi-penyebab-kematian-tertinggi-ketiga-di-indonesia?page=1

XGBoost Contributors. (n.d.). *XGBoost Documentation*. Diakses pada 3 Desember 2024, melalui https://xgboost.readthedocs.io/en/stable/

Machine Learning for Engineers. (n.d.). *XGBoost Classifier*. Diakses pada 3 Desember 2024, melalui https://apmonitor.com/pds/index.php/Main/XGBoostClassifier#:~:text=XGBoost%20(eXtreme%20Gradient%20Boosting)%20is,types%20of%20machine%20learning%20problems

IBM. (n.d.). *Apa itu random forest?* Diakses pada 3 Desember 2024, melalui https://www.ibm.com/id-id/topics/random-forest

Geeksforgeeks. (n.d.). ML | Extra Tree Classifier for Feature Selection. Diakses pada 4 Desember 2024, melalui https://www.geeksforgeeks.org/ml-extra-tree-classifier-for-feature-selection/

Geeksforgeeks. (2024, Mei 17). Decision Tree. Diakses pada 4 Desember 2024, melalui https://www.geeksforgeeks.org/decision-tree/