UJIAN AKHIR SEMESTER MATA KULIAH

PEMBELAJARAN MESIN

Anggota Kelompok 6 :

1. Fransisca Stevanie Ekawati (09021382328127)
2. Nabilla Kesuma (09021382328132)
3. Zsa Zsa Aulia Az Zahrah (09021382328136)
4. Saravina Zharfa Kelana Putri (09021382328149)
5. Nabila Ayu Talita (09021382328158)

Kelas : Teknik Informatika Bilingual P1

Kode Program dan PPT

- Link Colab : Random Forest Klasifikasi Diabetes- Versi 2.ipynb
- Link PPT : https://www.canva.com/design/DAG6Xt36yk0/QcxZvD8kp2sSbzWvJDC5vQ/edit?utm_content=DAG6Xt36yk0&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton

1. Jelaskan tentang teori dasar dari tugas kelompok yang dibahas.

   Jawab:

   Tugas kelompok ini membahas penerapan *machine learning* untuk melakukan klasifikasi diabetes menggunakan algoritma Random Forest. Secara umum, tujuan dari klasifikasi ini adalah memprediksi apakah seseorang menderita diabetes atau tidak berdasarkan data kesehatan tertentu yang bersifat numerik, seperti kadar glukosa, indeks massa tubuh (BMI), tekanan darah, usia, dan variabel medis lainnya.

   a. Klasifikasi dalam Machine Learning

      Klasifikasi merupakan salah satu metode dalam *supervised learning*, yaitu pembelajaran mesin yang menggunakan data berlabel. Pada kasus ini, data memiliki label *Outcome*, yang menunjukkan kondisi seseorang:

- 0 → Tidak diabetes
- 1 → Diabetes

Model *machine learning* dilatih menggunakan data historis untuk mempelajari pola hubungan antara fitur (*variabel input*) dan label (kelas). Setelah proses pelatihan, model dapat digunakan untuk memprediksi kelas dari data baru yang belum pernah dilihat sebelumnya.

b. Dataset Diabetes

Dataset yang digunakan dalam tugas ini adalah dataset diabetes (Pima Indians Diabetes Dataset) yang berasal dari GitHub yang umum digunakan dalam penelitian klasifikasi medis. Dataset ini terdiri dari beberapa fitur numerik yang merepresentasikan kondisi kesehatan seseorang, antara lain:

- Pregnancies
- Glucose
- BloodPressure
- SkinThickness
- Insulin
- BMI
- DiabetesPedigreeFunction
- Age

Fitur-fitur tersebut dipilih karena memiliki keterkaitan langsung dengan faktor risiko diabetes. Dataset ini sangat cocok untuk eksperimen *machine learning* karena datanya terstruktur, numerik, dan memiliki label yang jelas.

c. Random Forest sebagai Metode Klasifikasi

Random Forest adalah algoritma *ensemble learning* yang bekerja dengan membangun banyak pohon keputusan (*decision tree*) dan menggabungkan hasil prediksinya. Setiap pohon dibuat dari sampel data

yang diambil secara acak (*bootstrap sampling*) serta menggunakan subset fitur yang berbeda pada setiap pemisahan node.

Prinsip dasar Random Forest adalah:

- Setiap pohon keputusan melakukan prediksi secara independen.
- Hasil akhir ditentukan melalui mekanisme *voting* (untuk klasifikasi).
- Kelas dengan jumlah suara terbanyak menjadi hasil prediksi model.

Adapun Keunggulan Random Forest antara lain:

- Mampu menangani data dengan hubungan non-linear.
- Lebih tahan terhadap *overfitting* dibandingkan *decision tree* tunggal.
- Dapat menangani data dengan banyak fitur.
- Memiliki performa yang baik pada dataset medis.

d. Preprocessing Data

Sebelum data diproses oleh model Random Forest, dilakukan tahap *preprocessing*, yang bertujuan meningkatkan kualitas data dan kinerja model. Tahapan *preprocessing* yang digunakan meliputi:

- Penanganan *missing value* menggunakan metode imputasi (misalnya median).
- Normalisasi data agar skala antar fitur lebih seragam.
- Pemisahan data latih dan data uji untuk menghindari bias evaluasi.

Tahapan ini penting karena kualitas input data sangat mempengaruhi hasil prediksi model *machine learning*.

e. Evalusasi Model

Untuk menilai kinerja model Random Forest, digunakan beberapa metrik evaluasi, antara lain:

1. Accuracy → mengukur persentase prediksi yang benar
2. Confusion Matrix → menunjukkan detail prediksi benar dan salah
3. Precision, Recall, dan F1-score → mengevaluasi kualitas klasifikasi

4. ROC-AUC → mengukur kemampuan model membedakan kelas positif dan negatif

Metrik-metrik ini digunakan untuk memastikan bahwa model tidak hanya akurat secara umum, tetapi juga andal dalam mendeteksi kasus diabetes.

f. Tujuan Penerapan Model

Penerapan Random Forest dalam tugas ini bertujuan untuk:

1. Mengimplementasikan algoritma Random Forest untuk melakukan klasifikasi penyakit diabetes menggunakan dataset Pima Indians Diabetes.
2. Menganalisis kinerja model Random Forest berdasarkan metrik evaluasi, yaitu akurasi, precision, recall, dan ROC-AUC.
3. Mengetahui tingkat efektivitas algoritma Random Forest dalam mendeteksi penyakit diabetes secara dini berdasarkan data pasien.

2. Jelaskan EDA (Exploratory Data Analysis) dari data yang dipakai.

Jawab:

a. Struktur dan Karakteristik Dataset

Dataset yang digunakan adalah Pima Indians Diabetes Dataset yang terdiri dari 768 data observasi dan 9 atribut, dengan 8 fitur prediktor dan 1 variabel target (*Outcome*). seluruh fitur bersifat numerik, sehingga sesuai untuk diterapkan pada algoritma *machine learning* berbasis statistik seperti Random Forest. Pemeriksaan awal menggunakan info() menunjukkan bahwa setiap kolom memiliki jumlah data non-null yang sama, sehingga secara eksplisit tidak ditemukan missing values.

Namun demikian, hasil analisis statistik deskriptif menunjukkan adanya nilai minimum 0 pada beberapa fitur medis, seperti Glucose, BloodPressure, SkinThickness, Insulin, dan BMI. secara klinis, nilai nol pada atribut - atribut tersebut tidak realistis dan tidak mempresentasikan kondisi fisiologis yang valid, sehingga diinterpretasikan sebagai *missing values implisit*. Temuan ini menunjukkan bahwa meskipun dataset tampak

lengkap secara struktur, tetap diperlukan tahapan pra proses lanjutan agar kualitas data tidak menurunkan performa model klasifikasi.

b. Statistik Deskriptif

Analisis statistik deskriptif dilakukan untuk memperoleh gambaran sebaran data, nilai rata-rata, serta tingkat variasi setiap fitur. Hasil analisis menunjukkan bahwa beberapa fitur, seperti Insulin dan Glucose, memiliki nilai standar deviasi yang relatif tinggi, yang mengindikasikan adanya variasi besar antar individu dalam dataset. Kondisi ini mencerminkan karakteristik data medis yang heterogen dan mendukung penggunaan algoritma *ensemble* seperti Random Forest, yang mampu menangani variasi data serta hubungan non-linear antar fitur dengan baik.

c. Distribusi Kelas Target (Outcome)

Variabel target *Outcome* memiliki dua kelas, yaitu 0 (tidak diabetes) dan 1 (diabetes). Hasil EDA menunjukkan bahwa terdapat 500 data (65,10%) pada kelas tidak diabetes dan 268 data (34,90%) pada kelas diabetes. Distribusi ini mengindikasikan adanya ketidakseimbangan kelas (*imbalanced data*), yang merupakan kondisi umum pada dataset medis.

Ketidakseimbangan kelas tersebut memiliki implikasi penting dalam proses klasifikasi, karena model berpotensi lebih berpihak pada kelas mayoritas jika hanya dievaluasi menggunakan metrik akurasi. Oleh karena itu, sejak tahap EDA telah ditetapkan bahwa evaluasi model perlu menggunakan metrik tambahan seperti precision, recall, F1-score, dan ROC-AUC, serta mempertimbangkan strategi penanganan ketidakseimbangan data pada tahap pemodelan.

d. Analisis Korelasi Antar Fitur

EDA juga mencakup analisis korelasi *Pearson* untuk mengidentifikasi hubungan linear antara fitur prediktor dan variabel target. Hasil analisis menunjukkan bahwa fitur Glucose memiliki korelasi positif

paling kuat terhadap variabel Outcome, diikuti oleh BMI, Age, dan Pregnancies. Temuan ini sejalan dengan pengetahuan medis, dimana kadar glukosa darah, indeks massa tubuh, dan usia merupakan indikator utama dalam risiko diabetes tipe 2. Analisis korelasi ini memberikan indikasi awal mengenai fitur-fitur yang berpotensi paling berpengaruh dalam proses klasifikasi, serta menjadi dasar bagi interpretasi model pada tahap analisis *feature importance* setelah proses pelatihan Random Forest.

e. Kesimpulan EDA

Secara keseluruhan, hasil Exploratory Data Analysis menunjukkan bahwa dataset memiliki struktur yang baik dan relevan untuk klasifikasi diabetes, namun juga mengandung beberapa tantangan penting, yaitu keberadaan nilai tidak realistis yang perlu ditangani melalui imputasi serta ketidakseimbangan pada kelas target. Selain itu, analisis korelasi memperlihatkan hubungan yang masuk akal secara klinis antara beberapa fitur utama dengan outcome diabetes. Dengan demikian, EDA menjadi landasan penting dalam menentukan strategi praproses data, pemilihan metrik evaluasi, serta penerapan algoritma Random Forest pada tahap pemodelan selanjutnya.

3. Jelaskan aplikasi yang dibuat disertai mekanisme yang digunakan.

Jawab:

Aplikasi yang dibuat merupakan aplikasi web prediksi diabetes yang dirancang untuk mempermudah pengguna dalam melakukan prediksi indikasi diabetes. Aplikasi ini dibangun menggunakan framework Streamlit, sehingga dapat dijalankan melalui browser tanpa instalasi khusus di sisi pengguna. Aplikasi ini berperan sebagai sistem pendukung keputusan (decision support system) yang memanfaatkan model machine learning untuk memberikan prediksi berdasarkan data pasien yang dimasukkan.

Mekanisme Kerja Aplikasi:

a. Antarmuka Pengguna (User Interface)

Aplikasi menyediakan antarmuka berbasis web yang sederhana dan interaktif sehingga mudah digunakan oleh pengguna. Antarmuka ini dibangun menggunakan Streamlit dan dapat diakses langsung melalui browser tanpa memerlukan instalasi tambahan. Pada halaman aplikasi, pengguna mengisi sebuah form input yang berisi parameter medis pasien, yaitu jumlah kehamilan, kadar glukosa, tekanan darah, ketebalan kulit, insulin, BMI, diabetes pedigree function, serta usia. Seluruh input disediakan dalam bentuk numerik agar sesuai dengan kebutuhan sistem.

Form ini berfungsi untuk memastikan bahwa semua data yang dibutuhkan oleh model tersedia dan lengkap sebelum proses prediksi dijalankan, sehingga aplikasi dapat bekerja secara konsisten dan minim kesalahan input.

b. Pengumpulan dan Pengolahan Data Input

Setelah pengguna menekan tombol Prediksi, seluruh data yang telah dimasukkan pada form akan dikumpulkan oleh sistem. Data tersebut kemudian disusun dalam format yang terstruktur dan sesuai dengan susunan fitur yang digunakan oleh model machine learning. Dengan mekanisme ini, data input dapat langsung diproses oleh model tanpa memerlukan tahapan tambahan, serta memastikan kesesuaian antara data yang dimasukkan pengguna dan data yang digunakan saat pengembangan model.

c. Proses Prediksi

Aplikasi kemudian menjalankan model machine learning yang telah dilatih sebelumnya untuk memproses data input pengguna. Model melakukan perhitungan berdasarkan pola yang telah dipelajari dan menghasilkan hasil prediksi. Dalam proses ini, model menentukan:

- Hasil prediksi kelas (terindikasi diabetes atau tidak), serta
- Nilai probabilitas yang menunjukkan tingkat kemungkinan pasien mengalami diabetes.

Seluruh proses prediksi berjalan secara otomatis dan real-time, sehingga pengguna dapat langsung memperoleh hasil setelah menekan tombol prediksi tanpa perlu melakukan interaksi tambahan.

d. Penanganan Input Ekstrem

Untuk menjaga kestabilan dan keandalan hasil prediksi, aplikasi menerapkan mekanisme pembatasan nilai input agar tetap berada dalam rentang medis yang realistis. Jika pengguna memasukkan nilai yang terlalu ekstrem atau tidak wajar, sistem akan menyesuaikan nilai tersebut ke batas yang telah ditentukan.

Mekanisme ini bertujuan untuk:

- Mencegah kesalahan prediksi akibat input yang tidak sesuai,
- Menjaga konsistensi hasil prediksi,
- Memastikan aplikasi tetap memberikan hasil yang masuk akal dalam berbagai kondisi input.

e. Penyajian Hasil Prediksi

Setelah proses prediksi selesai, hasil ditampilkan secara langsung pada halaman aplikasi. Aplikasi menampilkan status pasien, yaitu apakah pasien terindikasi diabetes atau tidak, serta probabilitas diabetes dalam bentuk persentase. Hasil prediksi disajikan menggunakan indikator visual berupa warna dan pesan teks, sehingga informasi yang diberikan mudah dipahami oleh pengguna dan dapat membantu dalam pengambilan keputusan awal.

f. Kesimpulan

Aplikasi prediksi diabetes ini mengintegrasikan antarmuka web interaktif, mekanisme pengumpulan dan pengolahan input, serta model machine learning ke dalam satu sistem aplikasi yang utuh. Dengan mekanisme tersebut, pengguna dapat melakukan prediksi diabetes secara cepat, mudah, dan konsisten melalui aplikasi web.

4. Buat analisa dari hasil aplikasi yang dibuat.

Jawab:

Berdasarkan hasil pengujian dan evaluasi terhadap aplikasi prediksi diabetes yang telah dikembangkan, dilakukan analisis menyeluruh untuk menilai kinerja model, kemampuan prediksi, serta kelayakan aplikasi dalam membantu proses screening awal diabetes. Analisis tersebut mencakup evaluasi

performa model, interpretasi hasil prediksi, serta implikasi penggunaannya dalam konteks nyata.

a. Evaluasi performa model secara keseluruhan

Model prediksi diabetes menunjukkan performa yang cukup baik dengan nilai *accuracy* sebesar 75,52%, yang menandakan bahwa model mampu mengklasifikasikan data pasien dengan tingkat ketepatan yang stabil. Nilai tersebut menunjukkan bahwa sebagian besar data uji dapat diprediksi dengan benar, baik untuk kelas diabetes maupun non-diabetes, sehingga model memiliki kinerja yang layak untuk digunakan pada tahap awal analisis risiko.

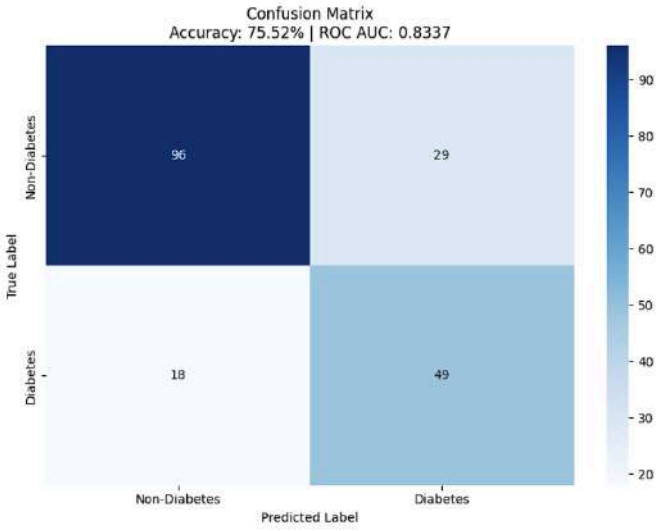
```
FINAL RANDOM FOREST RESULT:
Accuracy: 0.7552 (75.52%)
ROC AUC: 0.8337

Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.77      0.80       125
           1       0.63      0.73      0.68        67

    accuracy                           0.76       192
   macro avg       0.74      0.75      0.74       192
weighted avg       0.77      0.76      0.76       192
```

b. Kemampuan model dalam membedakan kelas (ROC-AUC)

Nilai ROC-AUC sebesar 0,8337 menunjukkan bahwa model memiliki kemampuan yang kuat dalam membedakan antara pasien diabetes dan non-diabetes. Nilai tersebut mengindikasikan bahwa model mampu memberikan probabilitas prediksi yang baik di berbagai ambang keputusan (*threshold*), sehingga tidak hanya bergantung pada satu batas klasifikasi tertentu.

c. Analisis kesalahan prediksi melalui confusion matrix

Berdasarkan confusion matrix, model berhasil mengklasifikasikan 96 data non-diabetes secara benar (*true negative*) dan 49 data diabetes secara benar (*true positive*). Namun, masih terdapat 29 *false positive* dan 18 *false negative*. Jumlah *false negative* yang relatif rendah menunjukkan bahwa model cukup efektif dalam mendeteksi pasien diabetes, sehingga risiko pasien diabetes yang tidak terdeteksi dapat diminimalkan.



d. Analisis sensitivitas (recall) dan spesifisitas

Model menghasilkan nilai *recall* (*sensitivity*) sebesar 73,13% untuk kelas diabetes, yang berarti sebagian besar pasien diabetes berhasil terdeteksi oleh sistem. Nilai *recall* tersebut sangat penting dalam konteks screening kesehatan, karena tujuan utama adalah mengidentifikasi pasien yang berisiko sedini mungkin. Selain itu, nilai *specificity* sebesar 76,80% menunjukkan bahwa model juga mampu mengenali pasien non-diabetes dengan cukup baik.

```
Metrik Detail:
Sensitivity/Recall (Diabetes): 0.7313 (73.13%)
Specificity (Non-Diabetes): 0.7680 (76.80%)
Precision (Diabetes): 0.6282 (62.82%)
```

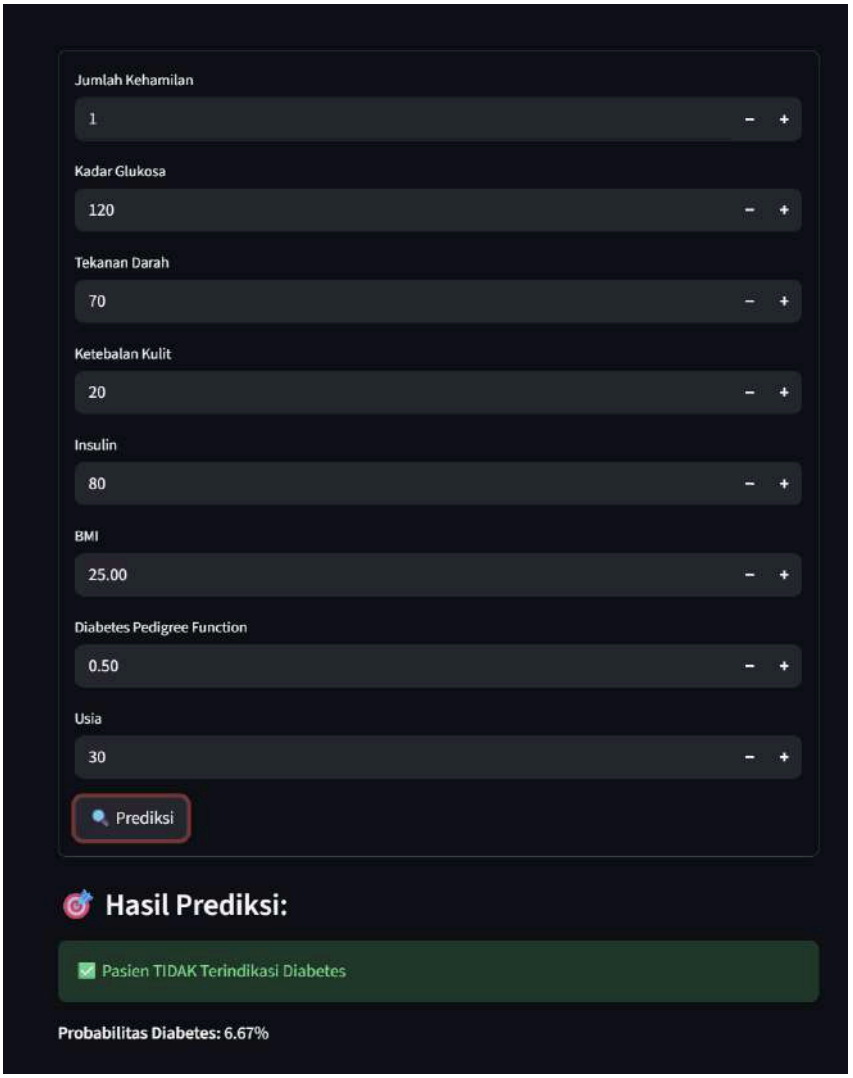
e. Analisis precision pada prediksi diabetes

Precision untuk kelas diabetes sebesar 62,82%, yang menunjukkan bahwa dari seluruh prediksi diabetes yang dihasilkan, sekitar 63% merupakan prediksi yang benar. Meskipun masih terdapat *false positive*, kondisinya masih dapat diterima dalam sistem screening awal, karena lebih aman melakukan pemeriksaan lanjutan daripada melewatkan pasien yang sebenarnya berisiko diabetes.

f. Validasi aplikasi melalui pengujian test case

Validasi aplikasi dilakukan melalui pengujian langsung pada antarmuka pengguna (UI) menggunakan data pasien. Pada salah satu *test case*, aplikasi memprediksi pasien tidak terindikasi diabetes dengan

probabilitas 6,67%, yang sesuai dengan karakteristik input berisiko rendah. Hasilnya menunjukkan bahwa aplikasi mampu memproses data dengan benar, menampilkan output yang konsisten, serta berjalan stabil tanpa kesalahan sistem, sehingga layak digunakan sebagai alat bantu screening awal diabetes.



g. Kelayakan dan keterbatasan aplikasi

Secara keseluruhan, aplikasi prediksi diabetes yang dikembangkan layak digunakan sebagai alat bantu screening awal, khususnya ketika diuji melalui antarmuka pengguna (UI). Hasil pengujian menunjukkan bahwa aplikasi mampu menerima input data dengan baik, melakukan proses prediksi secara konsisten, serta menghasilkan output yang sesuai dengan

kondisi data uji, sehingga secara fungsional aplikasi berjalan dengan stabil dan tidak menunjukkan potensi kesalahan sistem.

Keterbatasan utama aplikasi tidak terletak pada mekanisme prediksi atau implementasi model, melainkan pada nilai akurasi yang masih tergolong sedang serta keterbatasan dataset yang digunakan. Ukuran dan karakteristik dataset dapat mempengaruhi kemampuan model dalam menangkap pola yang lebih kompleks, sehingga berpotensi membatasi peningkatan performa prediksi. Oleh karena itu, hasil prediksi yang dihasilkan aplikasi tidak dimaksudkan sebagai diagnosis medis final, melainkan sebagai informasi pendukung untuk membantu pengguna atau tenaga medis dalam melakukan pemeriksaan lanjutan dan pengambilan keputusan awal.