PREDIKSI DIABETES DENGAN ALGORITMA RANDOM FOREST, LOGISTIC REGRESSION, DAN KNN

Nama : Endra Agustino

Nim : A11.2022.14614

Kelompok : A11.4517

FAKULTAS ILMU KOMPUTER

UNIVERSITAS DIAN NUSWANTORO

2025

PENDAHULUAN

RINGKASAN

Proyek ini bertujuan untuk membangun model prediktif yang dapat memprediksi apakah seseorang memiliki, berpotensi, atau tidak memiliki diabetes berdasarkan dataset yang digunakan, yang mencakup usia, jenis kelamin, kadar urea, (Cr), HbA1c, kolesterol, dan profil lipid (TG, HDL, LDL, VLDL), (BMI) dari seseorang. Dalam projek ini saya menggunakan supervised learning dengan beberapa algoritma klasifikasi seperti Random Forest, Logistic Regression, dan K-Nearest Neighbors diuji dan dibandingkan performanya berdasarkan akurasi, presisi, recall, dan f1-score. Proyek ini juga diharap dapat mencegah peningkatan jumlah penderita diabetes. Dengan menggunakan algoritma machine learning, prediksi risiko diabetes dapat dilakukan dengan lebih akurat untuk membantu tenaga medis dalam pengambilan keputusan.

MASALAH

Diabetes adalah salah satu masalah kesehatan yang umum dan berkembang di seluruh dunia. Prediksi dini terhadap penyakit ini dapat membantu dalam pencegahan dan pengelolaan yang lebih baik. Dataset yang digunakan berisi berbagai atribut medis yang dapat memengaruhi risiko seseorang terkena diabetes. melakukan diagnosis secara manual untuk setiap pasien membutuhkan waktu dan sumber daya yang besar. Oleh karena itu penggunaan data mining bisa sangat membantu pengambilan Keputusan yang lebih baik

TUJUAN

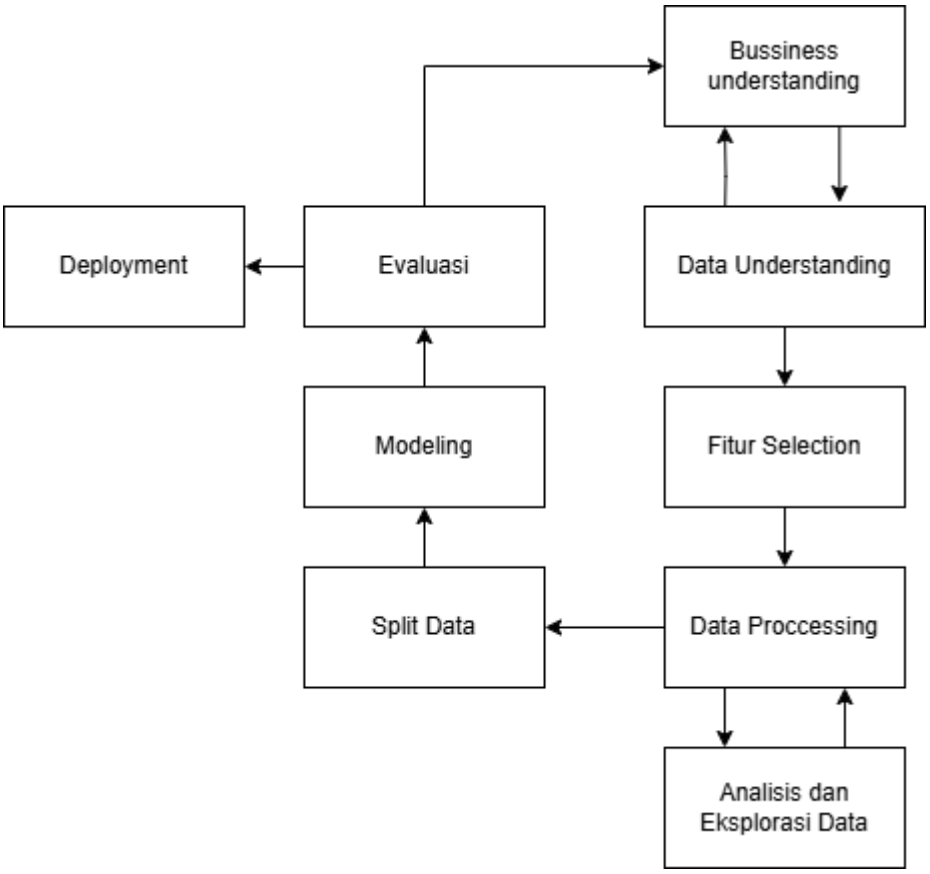
- **Memahami risiko diabetes**
  memahami bagaimana faktor-faktor ini berkontribusi pada kemungkinan seseorang mengidap diabetes.

- **Membangun model prediksi**
  membangun model untuk memprediksi kelas diabetes seseorang. Tiga kelas target yang digunakan dalam dataset ini adalah Diabetic (seseorang yang mengidap diabetes) dan Non-Diabetic (seseorang yang tidak mengidap diabetes) Predicted Diabetic (Kemungkinan mengidap diabetes).

- **Evaluasi dan pengujian model**
  model akan diuji menggunakan data yang belum pernah dilihat sebelumnya untuk mengukur akurasi prediksi. Evaluasi akan dilakukan dengan menggunakan beberapa metrik, seperti akurasi, precision, recall, dan F1-score

- **Manfaat bagi kesehatan masyarakat**
  Dengan membangun model yang akurat, kita dapat membantu pihak medis atau individu untuk melakukan deteksi dini terhadap diabetes.

ALUR PENYELESAIAN



DATASET

- **ID**
  Identifikasi unik untuk setiap pasien di dataset. Digunakan untuk membedakan rekam medis setiap individu.

- **No_Pation**
  Nomor pasien adalah penanda tambahan yang bisa digunakan untuk referensi lebih lanjut.

- **Gender**
  Jenis kelamin pasien, diwakili oleh 'M' (Male) untuk laki-laki dan 'F' (Female) untuk perempuan.

- **AGE (tahun)**
  Usia pasien dalam tahun. Memberikan informasi tentang umur pasien yang relevan untuk analisis risiko diabetes.

- **Urea (mg/dl)**
  Tingkat urea dalam darah, digunakan untuk memantau fungsi ginjal yang mungkin terpengaruh oleh diabetes.

- **Cr ()**
  Rasio kreatinin (Creatinine Ratio), indikator kesehatan ginjal dan fungsi metabolik.

- **HbA1c**
  Hemoglobin A1C, persentase gula darah terikat pada hemoglobin. Indikator utama untuk memantau diabetes selama periode waktu tertentu (biasanya 2–3 bulan).

- **Chol**
  Kolesterol total dalam darah, yang meliputi semua jenis kolesterol (HDL, LDL, dan lainnya).

- **TG**
  Trigliserida, salah satu jenis lemak dalam darah. Tingginya kadar ini dapat meningkatkan risiko penyakit jantung, terutama pada pasien diabetes.

- **HDL**
  High-Density Lipoprotein, dikenal sebagai kolesterol "baik" karena membantu mengurangi risiko penyakit jantung.

- **LDL**
  Low-Density Lipoprotein, dikenal sebagai kolesterol "jahat" karena kontribusinya terhadap penumpukan lemak di arteri.

- **VLDL**
  Very-Low-Density Lipoprotein, tipe lain dari kolesterol yang dapat memengaruhi metabolisme lipid.

- **BMI**
  Body Mass Index, rasio berat badan terhadap tinggi badan pasien, digunakan untuk menentukan apakah seseorang berada pada kategori berat badan sehat, kurang, atau obesitas.

- **CLASS**
  Kategori kondisi diabetes pasien:

  - **Diabetes**: Pasien telah terdiagnosis diabetes.
  - **Non-Diabetes**: Pasien tidak menderita diabetes.
  - **Predict-Diabetic**: Pasien memiliki kemungkinan besar untuk menderita diabetes

EDA

- Memeriksa Struktur dan Tipe Data sesuai
- Menangani missing data
- Memvisualisasikan distribusi data
- Mengidentifikasi korelasi antar fitur
- Menangani outlier
- Transformasi data (normalisasi, encoding)
- Mempersiapkan data untuk model (split data)

FEATURE DATASET

Gender, age, urea, cr, HbA1c, Chol, TG, HDL, LDL, VLDL, BMI

PROSES LEARNING/MODELING

Proses learning atau modeling adalah tahap di mana kita menggunakan dataset yang telah diproses untuk melatih model pembelajaran mesin. Pada proyek ini, tujuan utamanya adalah untuk memprediksi apakah seseorang berisiko mengidap diabetes berdasarkan fitur-fitur medis. Proses ini mencakup pemilihan model, pelatihan model, serta evaluasi model yang digunakan.

- **Pemilihan model**
  Projek ini menggunakan 3 model, yaitu : Random Forest Classifier, Logistic Regression, KNN

- **Pembagian data**
  Data dibagi menjadi 2, yaitu training set dan test set

- **Pelatihan model**
  - **Random Forest Classifier**
    menggunakan beberapa pohon keputusan. Setiap pohon akan "belajar" dari subset data dan fitur, lalu memberikan prediksi berdasarkan mayoritas keputusan dari pohon-pohon tersebut.
  - **Logistic Regression**
    melatih model dengan menghitung hubungan antara fitur dan probabilitas hasil untuk masing-masing kelas (misalnya, Diabetic atau Non-Diabetic). Fungsi yang digunakan adalah fungsi logistik (sigmoid) untuk mengubah hasil menjadi probabilitas antara 0 dan 1.
  - **K-Nearest Neighbors (KNN)**
    mencari k tetangga terdekat untuk data yang ingin diprediksi. Prediksi dibuat berdasarkan mayoritas label dari tetangga terdekat tersebut.

- **Menyimpan model**
  model yang telah dilatih disimpan ke dalam file untuk digunakan kembali tanpa perlu melatihnya ulang.

- **Pengujian model**
  menggunakan data uji (test set) untuk menguji kinerja model. Model akan memprediksi nilai target (misalnya, kelas diabetes) pada data uji yang tidak pernah dilihat selama pelatihan.

- **Evaluasi model**
  model yang telah dilatih dievaluasi berdasarkan akurasi dan laporan klasifikasi yang dihasilkan untuk **data training** dan **data testing**.

PERFORMA MODEL

- **Model Random Forest**

  Akurasi model training Random Forest: 1.00
  Akurasi model testing Random Forest: 0.98

- **Logistic Regression**

  Akurasi model training Logistic Regression: 0.93
  Akurasi model testing Logistic Regression: 0.94

- **K-Nearest Neighbors**
  Akurasi model training K-Nearest Neighbors: 0.91
  Akurasi model testing K-Nearest Neighbors: 0.90

DISKUSI HASIL DAN KESIMPULAN

Model Random Forest menunjukkan performa terbaik dengan akurasi yang lebih tinggi dibandingkan dengan Logistic Regression dan K-Nearest Neighbors. Akurasi pengujian menunjukkan bahwa model yang lebih kompleks seperti Random Forest lebih baik dalam memprediksi kelas diabetes dibandingkan model yang lebih sederhana.

Dengan menggunakan berbagai model pembelajaran mesin, kita dapat memprediksi kemungkinan seseorang mengidap diabetes dengan akurasi yang cukup baik. Random Forest merupakan model yang paling efektif untuk dataset ini dan dapat digunakan untuk membantu dalam deteksi dini penyakit diabetes.