# Implementasi Model Klasifikasi untuk Deteksi Risiko Diabetes Menggunakan Algoritma K-Nearest Neighbors (KNN)

SIANE SANTOSO

A11.2022.13956

1. **Deskripsi Singkat**

Diabetes adalah penyakit kronis yang memengaruhi jutaan orang di seluruh dunia, menyebabkan berbagai komplikasi kesehatan serius apabila tidak terdiagnosis dan ditangani sejak dini. Salah satu tantangan besar dalam pencegahan diabetes adalah kurangnya alat prediktif yang dapat membantu tenaga medis dan masyarakat untuk mendeteksi risiko diabetes dengan mudah dan cepat berdasarkan data klinis yang tersedia.

Dalam eksperimen ini, saya menggunakan algoritma K-Nearest Neighbors (KNN), sebuah metode yang sederhana namun efektif untuk masalah klasifikasi, dalam membangun model prediksi risiko diabetes. Dataset yang digunakan diambil dari Kaggle, berisi informasi demografis dan klinis yang relevan seperti usia, jenis kelamin, riwayat kesehatan, dan berbagai parameter fisik lainnya yang berkaitan dengan kondisi diabetes. Setiap fitur dalam dataset ini merupakan atribut numerik, sehingga memungkinkan analisis yang efisien menggunakan metode berbasis jarak seperti KNN.

1. **Masalah dan Tujuan yang Ingin Dicapai**
2. Masalah

* Diabetes, sebagai salah satu penyebab utama kematian dan kecacatan, membutuhkan perhatian serius dalam hal deteksi dini. Namun, akses ke alat pemeriksaan yang lebih komprehensif sering kali terbatas, terutama di daerah dengan sumber daya kesehatan terbatas.
* Salah satu solusi yang potensial adalah membangun model prediksi yang efisien dan dapat diakses dengan mudah oleh masyarakat. Model ini diharapkan dapat memperkirakan risiko diabetes berdasarkan data klinis yang umum, memungkinkan skrining dini yang efisien dan dapat membantu dalam memberikan informasi awal kepada individu yang mungkin berisiko terkena diabetes.

1. Tujuan

* Membangun model prediksi risiko diabetes dengan algoritma K-Nearest Neighbors (KNN) berdasarkan data numerik dari atribut klinis.
* Menyederhanakan proses deteksi dini melalui penyediaan aplikasi berbasis Streamlit, yang dapat digunakan oleh tenaga kesehatan maupun masyarakat untuk memprediksi risiko diabetes hanya dengan memasukkan beberapa informasi klinis.
* Meningkatkan pemahaman terhadap pengaruh faktor-faktor klinis tertentu (seperti usia, tekanan darah, dan indeks massa tubuh) terhadap risiko diabetes dengan cara yang mudah diakses, terukur, dan berbasis data.

1. **Kerangka Eksperimen**
2. Persiapan Dataset
3. Pengunduhan dan Eksplorasi Dataset

Dataset yang digunakan akan diambil dari Kaggle, dengan tujuan memanfaatkan informasi demografis dan klinis yang sudah tersedia terkait diabetes. Eksplorasi data awal dilakukan untuk memahami struktur dan karakteristik dataset.

1. Analisis Awal Data

Langkah ini mencakup pemeriksaan distribusi data untuk setiap fitur, melihat korelasi antar fitur, dan mengidentifikasi pola tertentu yang bisa muncul pada data individu yang memiliki dan tidak memiliki diabetes.

1. Pembersihan Data

Dalam proses ini, nilai yang hilang atau ekstrem (outlier) akan ditangani dengan metode seperti imputasi nilai median atau pembuangan data yang sangat ekstrim, agar kualitas data lebih baik dan performa model dapat ditingkatkan.

1. Preprocessing Data
2. Normalisasi Data

Karena KNN sensitif terhadap skala fitur, maka seluruh fitur numerik akan dinormalisasi ke rentang yang sama menggunakan metode seperti Min-Max Scaling. Ini bertujuan agar jarak antar data tidak didominasi oleh satu fitur tertentu yang memiliki rentang lebih besar.

1. Pembagian Data

Dataset akan dibagi menjadi data latih dan data uji dengan rasio tertentu, untuk melatih model sekaligus mengukur performa pada data yang tidak terlihat saat pelatihan.

1. Pemodelan dengan algoritma K-Nearest Neighbors (KNN)
2. Pemilihan Parameter K Optimal

Melalui metode pencarian parameter seperti Grid Search atau Cross-Validation, akan dicari nilai K yang memberikan performa terbaik. Nilai K yang optimal akan membantu dalam menghindari masalah overfitting atau underfitting pada model KNN.

1. Pelatihan Model KNN

Setelah parameter optimal ditentukan, model KNN akan dilatih pada data latih. Algoritma KNN akan menghitung jarak antara data baru dan data latih untuk mengklasifikasikan individu sebagai berisiko atau tidak berisiko diabetes berdasarkan K tetangga terdekat.

1. Evaluasi Model
2. Prediksi pada Data Uji: Setelah pelatihan selesai, model diuji pada data uji untuk mengevaluasi keakuratan prediksinya.
3. Pengukuran Akurasi: Menggunakan metrik akurasi untuk mengevaluasi kinerja model dalam mengklasifikasikan individu dengan atau tanpa diabetes.
4. **Penjelasan Dataset**
5. Sumber Data

Dataset yang digunakan diambil dari platform Kaggle, yang menyediakan data terkait diabetes dari berbagai sumber. Dataset ini memuat data klinis serta informasi yang relevan untuk penelitian dan pengembangan model prediksi risiko diabetes.

1. Penjelasan Atribut dalam Dataset
2. Gender

Jenis kelamin individu, biasanya dikodekan sebagai 0 untuk pria dan 1 untuk wanita, atau pengkodean lainnya.

1. Age

Usia individu dalam tahun, yang merupakan salah satu faktor risiko penting untuk diabetes.

1. Hypertension

Indikator hipertensi, dengan 0 menunjukkan tidak ada riwayat hipertensi dan 1 menunjukkan ada riwayat hipertensi. Hipertensi berkaitan erat dengan peningkatan risiko diabetes.

1. heart\_disease

Indikator riwayat penyakit jantung, dengan 0 untuk tidak ada riwayat dan 1 untuk ada riwayat. Penyakit jantung sering kali berhubungan dengan kondisi diabetes yang kurang terkendali.

1. smoking\_history

Riwayat merokok individu, yang biasanya dikodekan dalam beberapa kategori seperti tidak pernah merokok, mantan perokok, atau perokok aktif. Riwayat merokok seringkali dianggap sebagai faktor risiko tambahan untuk diabetes.

1. Bmi

Indeks massa tubuh, yang merupakan ukuran hubungan antara berat dan tinggi badan. BMI tinggi sering dikaitkan dengan risiko diabetes yang lebih tinggi.

1. HbA1c\_level

Tingkat Hemoglobin A1c, yang menunjukkan kadar gula darah rata-rata dalam tiga bulan terakhir. HbA1c yang tinggi merupakan salah satu indikasi kuat dari diabetes.

1. blood\_glucose\_level

Tingkat glukosa darah saat ini, yang bisa menjadi indikasi risiko diabetes apabila kadarnya terlalu tinggi atau rendah.

1. Diabetes

Label target, dengan 0 untuk individu tanpa diabetes dan 1 untuk individu dengan diabetes. Label ini yang akan menjadi target prediksi dalam model.