



UNIVERSITAS SYIAH KUALA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM



IMPROVING DIABETES DETECTION USING K-FOLD CROSS VALIDATION AND FEATURE SELECTION

SYAUKAS RAHMATILLAH
MUHAMMAD ALI MURTAZA
PROGRAM STUDI KECERDASAN BUATAN

MATA KULIAH PROGRAMMING FOR DATA SCIENCE AND ARTIFICIAL
INTELLIGENCE



PENDAHULUAN

- **Masalah Global:** Diabetes Mellitus (DM) adalah penyakit metabolism kronis yang menyebabkan sekitar 1,6 juta kematian tahunan menurut WHO.
- **Komplikasi:** Jika tidak terdeteksi dini, dapat menyebabkan kerusakan jantung, stroke, gagal ginjal, dan kebutaan.
- **Peran Teknologi:** Machine Learning (ML) menjadi alat kuat untuk mengidentifikasi pola tersembunyi dalam data klinis untuk prediksi dini



RECAL GAP

- **Kelemahan Model Saat Ini:** Banyak model ML mencapai akurasi tinggi (~97%) tetapi memiliki Recall yang sangat rendah (0,39–0,49),.
- **Penyebab:** Ketidakseimbangan kelas dalam dataset (65% non-diabetes vs 35% diabetes) membuat model bias terhadap kelas mayoritas.,
- **Risiko Medis:** Kesalahan negatif (False Negative) sangat berbahaya karena menunda pengobatan yang menyelamatkan nyawa bagi pasien yang sebenarnya sakit.



TUJUAN PENELITIAN

- **Mengatasi Ketidakseimbangan Kelas:** Menggunakan teknik SMOTE untuk menghasilkan sampel sintetis.
- **Optimalisasi Fitur:** Menerapkan Recursive Feature Elimination (RFE) untuk menghilangkan kebisingan data.
- **Stabilitas Model:** Menggunakan 5-Fold Cross Validation untuk memastikan performa yang konsisten dan mencegah overfitting.,



DATASET & PREPROCESSING

- **Dataset:** Pima Indian Diabetes Dataset (PIDD) dengan 768 catatan medis pasien wanita.
- **Imputasi Data:** Mengganti nilai nol yang tidak logis (pada Glukosa, BMI, dll.) dengan nilai median agar lebih tahan terhadap outlier.
- **Standarisasi:** Menerapkan Z-score normalization agar semua fitur memiliki skala yang sama untuk mempercepat konvergensi algoritma.



METODOLOGI - SMOTE

- **Fungsi:** Menyeimbangkan jumlah sampel antara pasien sehat dan diabetes.
- **Cara Kerja:** Menciptakan sampel sintetis baru di sepanjang garis antara titik data minoritas dan tetangganya (k -nearest neighbors).
- **Hasil:** Dataset menjadi seimbang dengan masing-masing kelas memiliki 500 sampel,.



METODOLOGI – SELEKSI FITUR (RFE)

- **Fungsi:** Menghapus fitur yang redundan dan meningkatkan efisiensi komputasi.
- **Fitur Terpilih (6 Fitur):** Pregnancies, Glucose, BloodPressure, BMI, DiabetesPedigreeFunction, dan Age.
- **Fitur Dibuang:** SkinThickness dan Insulin karena daya prediksi rendah dan banyaknya data yang hilang.



PELATIHAN & VALIDASI MODEL

- **Algoritma yang Diuji:** Logistic Regression, Random Forest, Gradient Boosting, SVM, Decision Tree, dan Naive Bayes.
- **5-Fold Stratified Cross-Validation:** Membagi data menjadi 5 bagian untuk diuji secara bergantian guna menjamin hasil yang stabil dan tidak bias terhadap satu pembagian data saja.



HASIL EKSPERIMENT - MODEL TERBAIK

- Random Forest terpilih sebagai model terbaik dengan performa luar biasa:
 - **Recall:** 0,976 (Mendeteksi 97,6% pasien diabetes dengan benar).
 - **Akurasi:** 0,972.
 - **ROC-AUC:** 0,994 (Hampir sempurna dalam membedakan kelas).
- Stabilitas: Standar deviasi yang sangat rendah (<0,02) menunjukkan model konsisten di berbagai subset data



ANALISIS MATRIKS KEBINGUNGAN (CONFUSION MATRIX)

- Pada data uji yang seimbang:
 - True Positives: 96 (Sakit terdeteksi sakit).
 - False Negatives: Hanya 4 (Sakit terdeteksi sehat).
- Penurunan drastis pada False Negative ini sangat krusial untuk keselamatan pasien di lingkungan klinis.



PERBANDINGAN DENGAN LITERATUR

- **Peningkatan Recall:** Kerangka kerja ini mencapai peningkatan Recall sebesar 149% dibandingkan studi Ibrahim et al. (0,976 vs 0,39),.
- **Keunggulan:** Integrasi SMOTE, RFE, dan K-Fold terbukti jauh lebih efektif daripada hanya menggunakan algoritma standar tanpa penyeimbangan data,.



IMPLIKASI KLINIS & KESIMPULAN

- **Kesimpulan:** Mengatasi redundansi fitur dan ketidakseimbangan data secara bersamaan adalah kunci untuk membangun CDSS yang andal.
- **Dampak Klinis:** Intervensi dini yang lebih akurat, pencegahan komplikasi jangka panjang, dan efisiensi biaya perawatan kesehatan.
- **Pekerjaan Masa Depan:** Eksplorasi Deep Learning dan integrasi biomarker tambahan seperti HbA1c.



UNIVERSITAS SYIAH KUALA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

SEKIAN

TERIMA KASIH