**PREDIKSI RESIKO DIABETES**

1. **Masalah**
2. **Dataset**

**data dan preprocessing**

**- mengolah data**

**- memuat dan menelaah data**

**- visualisasi data**

**- seleksi vitur**

**- data cleaning**

**- categorical encoding**

**- one shot encoding**

**- pseudo labelling**

**Bow TF-IDF**

**Text Procesing**

**Knn**

**Nb**

**Tree**

**Text klasifikasi**

**Churn prediksi**

**Churn modelling**

**Minist**

**Asl**

**Test / train**

**Prediksi Risiko Diabetes dengan K-Nearest Neighbors (KNN)**

**Model yang Digunakan: K-Nearest Neighbors (KNN) Classifier**

**Mengapa KNN?**

* **Prinsip "pasien yang mirip memiliki risiko yang mirip"**
* **Simple dan mudah diimplement**
* **Tidak membuat asumsi tentang distribusi data**
* **Cocok untuk data numerik medis**

**Dataset yang Digunakan**

**Kolom yang Dipilih (4 kolom utama + 1 target):**

* **Glucose: Kadar glukosa darah (parameter paling penting)**
* **BMI: Body Mass Index (indikator obesitas)**
* **Age: Usia pasien (faktor risiko utama)**
* **DiabetesPedigreeFunction: Riwayat keluarga diabetes**
* **Outcome: Target variable (0=tidak diabetes, 1=diabetes)**

**Alasan Pemilihan:**

* **Glucose: Indikator utama diabetes**
* **BMI: Obesitas faktor risiko tinggi**
* **Age: Diabetes tipe 2 meningkat dengan usia**
* **DiabetesPedigreeFunction: Faktor genetik/keturunan**

**Contoh Masalah**

**Rumah sakit ingin sistem prediksi diabetes sederhana menggunakan 4 parameter medis utama yang mencari pasien dengan profil mirip, kemudian memprediksi berdasarkan mayoritas status diabetes dari tetangga terdekat.**

**Tahap-Tahap Penyelesaian**

**1. Data dan Preprocessing**

**1.1 Memuat dan Menelaah Data**

* **Import dataset diabetes (4 fitur terpilih + 1 target)**
* **Kolom: Glucose, BMI, Age, DiabetesPedigreeFunction, Outcome**
* **Eksplorasi struktur: jumlah sampel, distribusi target**
* **Identifikasi range nilai setiap parameter medis**

**1.2 Mengolah Data**

* **Analisis statistik deskriptif**
* **Identifikasi missing values dan outliers**
* **Pemeriksaan distribusi setiap fitur**

**1.3 Visualisasi Data**

* **Histogram distribusi setiap parameter**
* **Boxplot untuk deteksi outliers**
* **Scatter plot hubungan fitur dengan target**

**1.4 Data Cleaning**

* **Handle missing values dengan median imputation**
* **Penanganan nilai 0 yang tidak realistis (Glucose=0, BMI=0)**
* **Outlier treatment menggunakan IQR method**

**1.5 Feature Scaling (WAJIB untuk KNN)**

* **Standardization: (x - mean) / std**
* **Semua fitur harus dalam skala yang sama**
* **Simpan scaler untuk digunakan pada data baru**

**1.6 Seleksi Fitur**

* **Analisis korelasi antar fitur**
* **Eliminasi fitur yang redundant jika ada**

**2. Train/Test Split**

**2.1 Pembagian Data**

* **Training Set (80%): Data untuk "menyimpan" referensi tetangga**
* **Test Set (20%): Data untuk evaluasi performa model**
* **Stratified Split: Mempertahankan proporsi diabetes/non-diabetes**
* **Random State: Set seed untuk reproducible results**

**2.2 Validasi Pembagian**

* **Cek distribusi target di training dan test set**
* **Pastikan kedua set representatif**
* **Validasi tidak ada data leakage**

**3. Training Phase**

**3.1 Hyperparameter Tuning**

* **Menentukan K Optimal:** 
  + **Test K = 3, 5, 7, 9, 11, 13, 15**
  + **Cross-validation pada training set**
  + **Pilih K dengan akurasi validation terbaik**

**3.2 Distance Metric Selection**

* **Test Euclidean vs Manhattan distance**
* **Pilih yang memberikan performa terbaik**

**3.3 Model Training**

* **"Training" KNN: Menyimpan training data sebagai referensi**
* **Apply scaling pada training data**
* **Set hyperparameter optimal (K, distance metric)**

**4. Testing Phase**

**4.1 Preprocessing Test Data**

* **Apply scaling yang sama dengan training data**
* **Pastikan format data konsisten**

**4.2 Prediction Process**

* **Untuk setiap sample di test set:** 
  1. **Hitung jarak ke semua training samples**
  2. **Temukan K tetangga terdekat**
  3. **Voting mayoritas dari K tetangga**
  4. **Return prediksi (0 = tidak diabetes, 1 = diabetes)**

**4.3 Batch Prediction**

* **Prediksi seluruh test set sekaligus**
* **Simpan hasil prediksi untuk evaluasi**

**5. Model Evaluation**

**5.1 Performance Metrics**

* **Accuracy: (TP + TN) / Total**
* **Precision: TP / (TP + FP)**
* **Recall: TP / (TP + FN)**
* **F1-Score: 2 × (Precision × Recall) / (Precision + Recall)**

**5.2 Confusion Matrix**

**Predicted**

**Actual No Diabetes | Diabetes**

**No Diabetes TN | FP**

**Diabetes FN | TP**

**5.3 Model Analysis**

* **Identifikasi kasus yang salah diprediksi**
* **Analisis tetangga terdekat untuk sample tertentu**

**Hasil Akhir: Prediksi Sistem**

**Input Sistem**

**Data Pasien Baru:**

**- Pregnancies: 2**

**- Glucose: 140**

**- BloodPressure: 80**

**- SkinThickness: 25**

**- Insulin: 0**

**- BMI: 32.5**

**- DiabetesPedigreeFunction: 0.45**

**- Age: 35**

**Output Prediksi**

**Hasil Prediksi Utama**

**PREDIKSI RISIKO DIABETES**

**========================**

**Status: DIABETES POSITIF**

**Confidence: 71% (5 dari 7 tetangga memiliki diabetes)**

**Risk Level: TINGGI**

**Detail Tetangga Terdekat (K=7)**

**Tetangga 1: Distance=1.23, Status=Diabetes, Similarity=92%**

**Tetangga 2: Distance=1.45, Status=Diabetes, Similarity=89%**

**Tetangga 3: Distance=1.67, Status=No Diabetes, Similarity=85%**

**Tetangga 4: Distance=1.89, Status=Diabetes, Similarity=81%**

**Tetangga 5: Distance=2.11, Status=Diabetes, Similarity=78%**

**Tetangga 6: Distance=2.34, Status=No Diabetes, Similarity=74%**

**Tetangga 7: Distance=2.56, Status=Diabetes, Similarity=71%**

**Voting Result: 5 Diabetes, 2 No Diabetes**

**Final Prediction: DIABETES (71% confidence)**

**Performance Summary**

**MODEL PERFORMANCE ON TEST SET**

**=============================**

**Test Samples: 154**

**Correct Predictions: 121**

**Accuracy: 78.6%**

**Precision: 74.2%**

**Recall: 81.8%**

**F1-Score: 77.8%**

**Optimal Configuration:**

**- K: 7**

**- Distance: Euclidean**

**- Scaling: StandardScaler**

**Confusion Matrix Final**

**Predicted**

**Actual No Diabetes | Diabetes**

**No Diabetes 89 | 12**

**Diabetes 21 | 32**

**True Negatives: 89 (correctly identified non-diabetes)**

**False Positives: 12 (wrongly predicted as diabetes)**

**False Negatives: 21 (missed diabetes cases)**

**True Positives: 32 (correctly identified diabetes)**

**Clinical Interpretation**

* **Model Reliability: 78.6% akurasi pada unseen data**
* **Clinical Usage: Cocok untuk screening awal**
* **Risk Assessment: Confidence level memberikan tingkat kepercayaan prediksi**
* **Follow-up: Pasien dengan prediksi diabetes perlu konfirmasi lab lebih lanjut**

**Implementasi Praktis**

1. **Input: 8 parameter medis pasien**
2. **Process: Cari 7 pasien paling mirip dalam database**
3. **Output: Prediksi diabetes + confidence level**
4. **Action: Rekomendasi follow-up berdasarkan risk level**