Nama : Muhammad Alif al Husain

NIM : A11.20222.14718

Tugas2

Langkah CRISP-DM Klasifikasi Penyakit Diabetes

Dataset : https://www.kaggle.com/code/nurkharisaumami/klasifikasi-penyakit-diabetes/input

1. Business Understanding.

   Tujuan utama dari tahap ini adalah mengklasifikan apakah data pasien tersebut terdeteksi memiliki diabetes atau tidak, berdasarkan hasil lab pasien seperti jumlah kehamilan (Pregnancies), kadar glukosa (Glucose), tekanan darah (BloodPressure), ketebalan kulit (SkinThickness), kadar insulin (Insulin), indeks massa tubuh (BMI), fungsi silsilah diabetes (DiabetesPedigreeFunction), dan usia (Age).Masalah yang ingin dipecahkan : mendeteksi diabetes secara dini untuk membantu dokter dalam pengambilan Langkah awal mengenai pengobatan atau perawatan yang tepat bagi pasien tersebut.

2. Data Attributes.

   Pregnancies : Jumlah Kehamilan.
   Glucose : Kadar glukosa dalam darah.
   BloodPressure : Tekanan darah diastolik (mm Hg).
   SkinThickness : Ketebalan kulit (mm).
   Insulin : Kadar insulin (mu U/ml).
   BMI : Indeks massa tubuh (kg/m²).
   DiabetesPedigreeFunction : Fungsi yang mengukur kemungkinan diabetes berdasarkan riwayat keluarga.
   Age: Usia pasien (tahun).
   Output Variable Target (1) Jika pasien terdeteksi Diabetes, (0) Jika tidak.

3. Data Preparation.

   Pada tahap ini, kita akan membersihkan dan menyiapkan data untuk diolah lebih lanjut.
   Data Splitting : Membagi dataset menjadi training set dan test set untuk evaluasi. Biasanya dilakukan dengan rasio 80:20 atau 70:30.
   Handling Imbalanced Data : Jika data diabetes tidak seimbang (jumlah pasien yang memiliki diabetes jauh lebih sedikit atau lebih banyak daripada yang tidak), kita bisa menggunakan teknik seperti oversampling (SMOTE) atau undersampling untuk menyeimbangkan kelas.

4. Modeling.

   Pada tahap ini akan memilih dan melatih model untuk klasifikasi diabetes.
   Algoritma yang akan digunakan :
   K-Nearest Neighbors(KNN) : Algoritma non parametrik yang sederhana namun cukup efektif.

5. Evaluation.

   Setelah model dilatih, kita akan mengevaluasi performanya.

Confusion Matrix: Untuk melihat jumlah True Positive, True Negative, False Positive, dan False Negative dari model.
Metrics:
Akurasi: Rasio prediksi yang benar dari semua prediksi.
Precision: Persentase prediksi positif yang benar-benar positif.
Recall: Persentase kasus sebenarnya positif yang terdeteksi oleh model.
F1 Score: Rata-rata harmonis antara precision dan recall.
Cross Validation: Menggunakan k-fold cross-validation untuk memastikan model tidak overfitting dan performa stabil di seluruh dataset.

6. Deployment.

Setelah model memiliki performa yang baik, langkah terakhir adalah deployment ke dalam sistem produksi.

Langkah Deployment:

Mengintegrasikan model ke dalam aplikasi berbasis web, mobile, atau sistem klinis yang bisa digunakan dokter.

APIs: Model bisa diekspos melalui REST API yang bisa diakses oleh aplikasi lain.

Model Monitoring: Memantau performa model secara berkala, dan melakukan retraining jika ditemukan penurunan performa seiring waktu (misalnya karena distribusi data yang berubah).

Ringkasan Langkah CRISP-DM untuk Kasus Diabetes

Business Understanding: Mendeteksi diabetes berdasarkan data pasien.

Data Understanding: Mengeksplorasi fitur medis, menangani nilai yang hilang.

Data Preparation: Imputasi nilai nol, scaling, membagi data.

Modeling: Melatih model klasifikasi seperti Logistic Regression atau Random Forest.

Evaluation: Mengukur akurasi, precision, recall, dan F1 score, memastikan model stabil dengan cross-validation.

Deployment: Mengimplementasikan model ke sistem klinis untuk membantu deteksi dini diabetes.

Langkah-langkah ini memastikan kita dapat menerapkan data mining dengan baik untuk mendeteksi diabetes pada pasien, serta mendukung keputusan medis berbasis data.