Prediksi Penyakit Jantung
dengan menggunakan Random Forest

Nama : Vicola Nanda Pratama

NIM : A11.2022.14240

Kelas : Penambangan Data 4504

Program Studi Teknik Informatika

Jurusan Ilmu Komputer

Universitas Dian Nuswantoro Semarang

2024

a. Deskripsi Singkat

Penyakit jantung merupakan salah satu penyebab utama kematian di dunia. Deteksi dini terhadap risiko penyakit jantung dapat membantu dalam pengambilan keputusan medis yang lebih cepat dan tepat. Proyek ini bertujuan untuk membangun model prediksi penyakit jantung menggunakan metode Random Forest dan membandingkan performanya dengan Decision Tree dan K-Nearest Neighbors (KNN).

b. Masalah dan Tujuan yang ingin dicapai

a. Permasalahan :

1. Bagaimana mengolah dataset penyakit jantung untuk menghasilkan prediksi yang akurat?

2. Metode mana yang dapat memberikan performa terbaik dalam resiko penyakit jantung?

b. Tujuan :

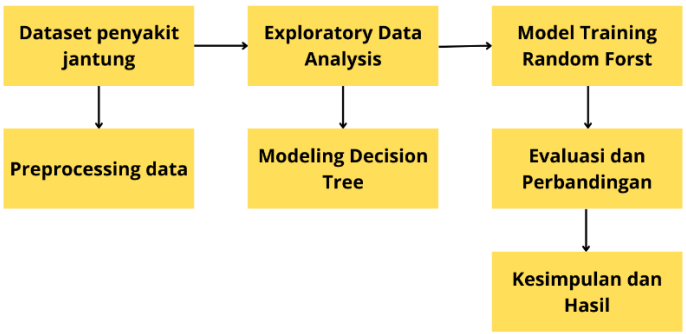
1. Mengolah dataset penyakit jantung untuk membangun model prediksi.

2. Mengimplementasikan metode Random Forest, Decision Tree, dan KNN untuk memprediksi.

3. Membandingkan performa dari masing-masing model dan melakukan evaluasi keunggulan Random Forest

c. Alur/Tahapan/Eksperimen

1. Bagan



2. Pengumpulan Data

a. Dataset diperoleh dari sumber terpercaya yang berisi informasi tentang atribut yang dapat digunakan untuk memprediksi obesitas.

b. Memuat dataset ke dalam program menggunakan library Pandas (pd.read_csv()).

3. Eksplorasi Data (EDA)

a. Melihat beberapa baris pertama dataset dengan head() untuk memahami struktur data.

b. Menggunakan fungsi info() untuk melihat tipe data dari setiap kolom dan apakah ada nilai yang hilang.

c. Melakukan analisis statistik deskriptif menggunakan describe() untuk memahami distribusi nilai pada setiap kolom.

d. Membuat visualisasi data menggunakan Seaborn dan Matplotlib

4. Persiapan Data

a. Memisahkan dataset menjadi fitur (X) dan target (y), di mana kolom target adalah kolom yang menunjukkan kelas obesitas.

b. Mengonversi data kategorikal menjadi numerik menggunakan One-Hot Encoding atau Label Encoding.

c. Melakukan normalisasi fitur numerik menggunakan MinMaxScaler agar semua nilai berada dalam rentang 0-1.

d. Menangani data tidak seimbang (jika ada) dengan teknik seperti oversampling menggunakan SMOTE atau undersampling.

5. Pembagian Data

a. Membagi dataset menjadi training set (80%) dan testing set (20%) menggunakan fungsi train_test_split dari Scikit-learn.

b. Data training digunakan untuk melatih model, sementara data testing digunakan untuk evaluasi.

6. Pembangunan Model

a. Model 1: Random Forest

- Membuat model Random Forest menggunakan RandomForestClassifier dari Scikit-learn.

- Menyesuaikan parameter seperti jumlah pohon (n_estimators), kedalaman maksimum (max_depth), dan lainnya.

- Melatih model menggunakan data training.

b. Model 2: Decision Tree

- Membuat model Decision Tree menggunakan DecisionTreeClassifier.

- Menyesuaikan parameter seperti kedalaman maksimum (max_depth).

- Melatih model menggunakan data training.

c. Model 3: K-Nearest Neighbors (KNN)

- Membuat model KNN menggunakan KNeighborsClassifier.

- Menentukan jumlah tetangga terbaik (n_neighbors) dengan pengujian parameter.

- Melatih model menggunakan data training.

7. Evaluasi Model

a. Menggunakan data testing untuk mengevaluasi performa setiap model.

b. Menghitung metrik evaluasi seperti:

- Akurasi menggunakan accuracy_score.

- Precision, Recall, F1-Score menggunakan classification_report.

- Confusion Matrix untuk melihat detail prediksi benar dan salah.

c. Membandingkan hasil evaluasi dari Random Forest, Decision Tree, dan KNN.

8. Analisis Performa Model

a. Membuat grafik perbandingan akurasi antara Random Forest, Decision Tree, dan KNN.

b. Menampilkan confusion matrix untuk masing-masing model sebagai visualisasi performa.

c. Menganalisis Feature Importance dari Random Forest untuk melihat fitur mana yang paling berpengaruh dalam prediksi.

9. Diskusi dan Kesimpulan

a. Menjelaskan performa model secara detail:

- Random Forest memiliki akurasi tertinggi dibandingkan Decision Tree dan KNN.

- Fitur yang paling penting dalam prediksi obesitas berdasarkan analisis Feature Importance.

b. Menyimpulkan bahwa Random Forest adalah model yang optimal untuk dataset ini berdasarkan evaluasi dan analisis.

d. Penjelasan Datasets

Dataset yang digunakan berasal dari Kaggle, yaitu Heart Disease Dataset. Dataset ini terdiri dari beberapa fitur seperti usia, jenis kelamin, tekanan darah, kolesterol, dan sebagainya, serta satu target (label) yaitu risiko penyakit jantung.

sumber : https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset

Nama Atribut	Tipe Data	Deskripsi
Age	Continuous	Umur pasien dalam tahun
Sex	Categorical	Jenis kelamin pasien
Chest Pain (cp)	Categorical	Jenis nyeri dada (1 = typical angina, dll.)
Resting Blood Pressure (trestbps)	Continuous	Tekanan darah istirahat (mm Hg)
Cholesterol (chol)	Continuous	Kadar kolesterol serum dalam mg/dL
Fasting Blood Sugar (fbs)	Binary	Gula darah puasa > 120 mg/dL (1 = true, 0 = false)
Resting ECG (restecg)	Categorical	Hasil EKG istirahat
Max Heart Rate (thalach)	Continuous	Denyut jantung maksimal pasien
Exercise-Included Angina (exang)	Binary	Nyeri dada saat olahraga (1 = ya, 0 = tidak)
Oldpeak	Continuous	Depresi ST setelah olahraga
Slope	Categorical	Kemiringan segmen ST
Number of Major Vessels (ca)	Categorical	Jumlah pembuluh darah utama yang divisualisasi
Thalassemia (thal)	Categorical	Jenis thalassemia
Target	Binary	Diagnosis penyakit jantung (1 = ada, 0 = tidak ada)

e. Timeline Eksperimen

No	Tahapan	1	2	3	4	5	6	7	8	
1	Perencanaan dan Persipan data									
2	Eksplorasi Data dan Preproccessing									
3	Pemilihan Model dan Eskperimen awal									
4	Pengoptimalan Model									
5	Implementasi sistem prediksi									
6	Evaluasi dan Validasi									
7	Dokumentasi dan Penyelesaian Projek									

Penjelasan Timeline :

1. Pengumpulan Data dan Persiapan Data (Minggu 1-2)

    a. Mengunduh dataset penyakit jantung dari Kaggle dan memuatnya ke dalam program menggunakan Pandas.

b. Memeriksa struktur data, mendeteksi nilai yang hilang, dan melakukan imputasi jika diperlukan.

2. Eksplorasi Data dan Preprocessing (Minggu 2-3)

a. Menganalisis fitur penting dengan visualisasi data menggunakan Seaborn.

b. Melakukan encoding pada fitur kategori dan normalisasi data numerik dengan MinMaxScaler.

3. Pemilihan Model dan Eksperimen Awal (Minggu 3-4)

a. Membagi data menjadi training dan testing set.

b. Melatih beberapa model awal (Logistic Regression, Random Forest) dan mengevaluasi akurasi awal.

4. Pengoptimalan Model (Minggu 4-5)

a. Melakukan hyperparameter tuning dengan GridSearchCV untuk model terbaik.

b. Menggunakan cross-validation untuk memastikan performa konsisten.

5. Implementasi Sistem Prediksi (Minggu 5-6)

a. Membuat pipeline prediksi dengan model terbaik.

b. Mengembangkan antarmuka sederhana untuk menerima input pengguna.

6. Evaluasi dan Validasi Model (Minggu 6-7)

a. Mengevaluasi model menggunakan precision, recall, F1-score, dan AUC-ROC.

b. Menyusun confusion matrix untuk analisis performa prediksi.

7. Dokumentasi dan Penyelesaian Proyek (Minggu 7-9)

a. Menyusun laporan akhir dan dokumentasi teknis.

b. Melakukan presentasi hasil proyek dan menampilkan demo sistem.