Laporan Analisis Data dengan Model Decision Tree pada Dataset Heart

Nama : Ilham Akbar

NIM : 4112322005

1. Deskripsi Dataset

Dataset yang digunakan adalah dataset yang berkaitan dengan **penyakit jantung**. Dataset ini terdiri dari variabel demografis, hasil tes medis, dan faktor risiko lainnya yang dapat membantu memprediksi apakah seseorang berisiko memiliki penyakit jantung atau tidak. Alasan memilih dataset ini adalah karena penyakit jantung merupakan salah satu penyebab kematian utama, sehingga prediksi dan analisis data ini memiliki implikasi besar dalam hal pencegahan dan perawatan kesehatan.

Setiap baris dalam dataset merepresentasikan satu pasien, dengan beberapa fitur utama, antara lain:

a. **Age (Umur)** Menunjukkan usia pasien dalam tahun.

b. **Sex (Jenis Kelamin)** Menunjukkan jenis kelamin pasien, dengan nilai M untuk laki-laki dan F untuk perempuan.

c. **Chest Pain Type (Jenis Nyeri Dada)** Menggambarkan jenis nyeri dada yang dialami oleh pasien, dengan kategori berikut:
- **TA (Typical Angina)**: Nyeri dada khas yang berkaitan dengan angina.
- **ATA (Atypical Angina)**: Nyeri dada yang tidak khas atau tidak sepenuhnya cocok dengan angina.
- **NAP (Non-Anginal Pain)**: Nyeri dada yang bukan karena angina.
- **ASY (Asymptomatic)**: Tanpa gejala nyeri dada.

d. **RestingBP (Tekanan Darah Saat Istirahat)** Mengukur tekanan darah pasien dalam milimeter merkuri (mm Hg) saat pasien beristirahat.

e. **Cholesterol (Kolesterol)** Mengukur tingkat kolesterol serum dalam darah, dinyatakan dalam mg/dL.

f. **FastingBS (Gula Darah Saat Puasa)** Menunjukkan apakah gula darah saat puasa melebihi 120 mg/dL. Nilai 1 menandakan gula darah tinggi, sedangkan 0 berarti normal.

g. **MaxHR (Denyut Jantung Maksimum)** Denyut jantung maksimum yang dicapai pasien selama tes fisik atau aktivitas berat, dinyatakan dalam satuan BPM (beats per minute).

h. **Exercise Angina (Angina Saat Berolahraga)** Menunjukkan apakah pasien mengalami angina (nyeri dada) selama aktivitas fisik. Y menandakan angina, N berarti tidak ada angina.

i. **Oldpeak** Ukuran depresi segmen ST pada EKG selama tes latihan atau stres, diukur dalam milimeter.

j. **ST_Slope (Kemiringan Segmen ST)** Menggambarkan kemiringan segmen ST pada EKG selama tes latihan. Kategori:
- **Up (Menanjak)**: Cenderung normal dan menandakan pemulihan yang sehat.
- **Flat (Datar)**: Dapat mengindikasikan iskemia atau gangguan aliran darah.
- **Down (Menurun)**: Biasanya mengindikasikan risiko penyakit arteri koroner.

k. **Heart Disease** Variabel biner yang menunjukkan apakah seorang pasien didiagnosis memiliki penyakit jantung atau tidak.

0: Tidak ada penyakit jantung (negative)

1: Ada penyakit jantung (positive)

Variabel ini merupakan target utama dari analisis prediktif. Nilai 1 menunjukkan bahwa pasien memiliki penyakit jantung, sementara nilai 0 menunjukkan pasien tidak memilikinya. Model prediksi akan dilatih untuk mengklasifikasikan pasien berdasarkan fitur lain, dengan tujuan untuk memprediksi kemungkinan diagnosis penyakit jantung.

2. Pra-Pemrosesan Data

a. Penanganan Data Kategori

Beberapa kolom dalam dataset ini berupa variabel kategori, seperti Sex, ChestPainType, RestingECG, ExerciseAngina, dan ST_Slope. Untuk memudahkan dalam pemrosesan model, fitur ini diubah menjadi bentuk numerik.

b. Normalisasi Data

Agar model lebih efektif, fitur numerik seperti Age, RestingBP, Cholesterol, MaxHR, dan Oldpeak dinormalisasi. Normalisasi ini dilakukan untuk memastikan semua variabel memiliki skala yang seimbang, sehingga model Decision Tree tidak terlalu berat pada variabel dengan skala besar.

c. Visualisasi Distribusi

Dilakukan Visualisasi Distribusi Heart Disease untuk memahami proporsi pasien dengan dan tanpa penyakit jantung, dan juga memvisualisasikan distribusi usia berdasarkan variabel Heart Disease untuk melihat apakah ada pola usia tertentu yang berisiko lebih tinggi.

d. Korelasi

Peta korelasi atau Heatmap dibuat untuk memahami hubungan antara variabel-variabel dalam dataset. Korelasi ini menunjukkan bagaimana variabel numerik dan kategori berpengaruh terhadap variabel target (Heart Disease), yang sangat berguna dalam pemilihan fitur utama untuk model. Pada Visual Heatmap ini kita dapat menentukan pengaruh varibel fitur terhadap variabel target dengan rentang nilai antara -1 sampai 1, semakin mendekati -1 menandakan bahwa variabel fitur berpengaruh secara negatif (berkebalikan) dengan variabel target, dan semakin mendekati 1 menandakan bahwa variabel fitur berpengaruh secara positif (selaras) dengan variabel target.

3. Pembuatan dan Evaluasi Model Decision Tree

Model yang digunakan adalah Decision Tree Classifier dari pustaka scikit-learn di Python. Langkah-langkah yang dilakukan adalah sebagai berikut:

a. **Split Data**: Dataset dibagi menjadi 70% data training dan 30% data testing untuk melatih dan menguji performa model.

b. **Training Model**: Model Decision Tree dilatih menggunakan data training.

c. **Evaluasi Model**: Model dievaluasi menggunakan metrik-metrik berikut:
   - **Akurasi**: Persentase prediksi yang benar pada keseluruhan model sebesar 0.78.
   - **Presisi**: Persentase ketepatan model dalam memprediksi positif sebagai benar sebesar 0.87.
   - **Recall**: Ukuran sensitivitas model terhadap kasus positif yang sesungguhnya sebesar 0.73.
   - **F1 Score**: keseimbangan Presisi dan Recall pada keseluruhan model pada kelas **0** sebesar 0.75 dan kelas **1** sebesar 0.79

Setelah pelatihan, hasil evaluasi model menunjukkan nilai akurasi, presisi, dan recall yang baik, yang mengindikasikan bahwa model dapat mengenali pola dalam data dengan cukup efektif.

4. Visualisasi Decision Tree dan Analisis Hasil

Visualisasi struktur Decision Tree menunjukkan bagaimana data dibagi pada setiap node. Dalam setiap node, variabel yang memiliki pengaruh paling signifikan digunakan untuk memisahkan data, dan akhirnya, node ini mengarah pada prediksi Heart Disease.

Visualisasi juga memungkinkan kita untuk melihat fitur-fitur yang paling berkontribusi dalam model, berikut ini adalah 5 Variabel paling berkontribusi:

- **ST_Slope_Up**: 0.4052
- **Oldpeak**: 0.1578
- **MaxHR**: 0.0903
- **Cholesterol**: 0.0623
- **Age**: 0.0556

Confusion Matrix

Confusion Matrix digunakan untuk menunjukkan klasifikasi yang benar dan salah dari model, membantu kita memahami kesalahan prediksi yang terjadi. Berikut rinciannya :

- **True Positive (TP)**: 120
  Model memprediksi "Heart Disease" dan benar-benar "Heart Disease".
- **True Negative (TN)**: 94
  Model memprediksi "No Heart Disease" dan benar-benar "No Heart Disease".
- **False Positive (FP)**: 18
  Model memprediksi "Heart Disease" tetapi sebenarnya "No Heart Disease".
- **False Negative (FN)**: 44
  Model memprediksi "No Heart Disease" tetapi sebenarnya "Heart Disease".

5. Rekomendasi Berdasarkan Temuan

Berdasarkan analisis ini, ada beberapa rekomendasi yang dapat dipertimbangkan:

- **Penerapan dalam Kesehatan**: Model ini dapat diterapkan dalam screening awal untuk memprediksi risiko penyakit jantung berdasarkan data demografis dan hasil pemeriksaan. Hal ini memungkinkan tindakan preventif dilakukan lebih awal pada pasien berisiko.
- **Pengembangan Model Lebih Lanjut**: Perlu dipertimbangkan metode ensemble, seperti Random Forest atau Gradient Boosting, yang dapat memperbaiki performa prediksi dengan menggabungkan beberapa model Decision Tree.
- **Penambahan Data**: Data yang lebih banyak dan beragam, misalnya dari berbagai rumah sakit atau wilayah, dapat memperbaiki generalisasi model.
- **Interpretasi untuk Pasien**: Analisis ini dapat menjadi dasar untuk komunikasi dengan pasien, memberikan pemahaman lebih dalam tentang faktor-faktor risiko yang dapat mereka kelola (misalnya, diet untuk mengontrol kolesterol).

Kesimpulan

Dengan analisis dan model Decision Tree ini, model berhasil mengidentifikasi beberapa variabel yang berhubungan erat dengan risiko penyakit jantung. Visualisasi dan evaluasi model memberikan wawasan yang berharga dalam memahami bagaimana variabel-variabel seperti ST_Slope, Oldpeak, dan MaxHR memainkan peran penting dalam memprediksi adanya penyakit jantung. Rekomendasi yang disarankan dapat membantu dalam pengembangan model prediktif lebih lanjut dan memberikan manfaat nyata dalam bidang medis.