**KLASIFIKASI DATA *HEART DISEASE* UNTUK MEMPREDIKSI KEBERADAAN PENYAKIT JANTUNG PADA PASIEN**

LAPORAN INI DISUSUN UNTUK MEMENUHI TUGAS UJIAN AKHIR SEMESTER MATA KULIAH DATA WAREHOUSE DAN DATA MINING

****

DISUSUN OLEH

FAKHIRAH DIVA MUSLIMAH 2110512042

NAFA AZZAHRA 2110512050

KEYSHA ALEA MAHARANI A.S 2110512057

MUHAMMAD HIDAYATULLAH 2110512066

ENRILE FADHLI FAHREZI 2110512076

**S1 SISTEM INFORMASI**

**FAKULTAS ILMU KOMPUTER**

**UNIVERSITAS PEMBANGUNAN NASIONAL VETERAN JAKARTA**

**2023**

**KATA PENGANTAR**

Puji syukur kami panjatkan kepada Tuhan Yang Maha Esa karena berkat rahmat dan karunia-Nya kami dapat menyelesaikan makalah yang berjudul “Klasifikasi Data Heart Disease untuk Memprediksi Keberadaan Penyakit Jantung Pada Pasien” ini. Kami juga mengucapkan terima kasih kepada Ibu Iin Ernawati, S.Kom., M.Si. selaku dosen pengampu mata kuliah Data Warehouse dan Data Mining yang telah membimbing kami dalam menyelesaikan makalah ini.

Makalah ini disusun untuk memenuhi ujian akhir semester mata kuliah Data Warehouse dan Data Mining. Selain itu, makalah ini juga bertujuan untuk membuka dan menambah wawasan bagi kami dan para pembaca tentang menggunakan metode klasifikasi dan algoritma Native Bayes dalam melakukan proses data mining.

Kami sebagai penulis menyadari bahwa makalah ini tentunya masih jauh dari kata sempurna. Oleh karena itu, kami mengharapkan kritik dan saran yang membangun dari seluruh pihak agar kami dapat menyempurnakan kekurangan pada makalah ini. Kami berharap makalah ini dapat memberi manfaat dan wawasan yang berguna bagi seluruh pembaca.

Jakarta, 10 Juni 2023

Penulis

**BAB I**

**PENDAHULUAN**

1. **Latar Belakang**
2. **Definisi**

Data menurut Drs. Jhon J. Longkutoy adalah suatu istilah majemuk dari fakta yang mengandung arti yang dihubungkan dengan kenyataan, simbol, gambar, angka, huruf yang menunjukan suatu ide, objek, kondisi atau situasi dan lainnya. Sedangkan menurut Williams dan Sawyer (2007) data terdiri dari fakta-fakta dan angka-angka yang diolah menjadi informasi. Menurut Sugiyono (2014:335-336) analisis data merupakan proses mencari dan menyusun secara sistematis data yang diperoleh dari hasil wawancara, observasi, dan dokumentasi, dengan cara mengorganisasikan data ke dalam kategori, menjabarkan ke dalam unit-unit, melakukan sintesa, menyusun ke dalam pola, memilih mana yang penting dan yang akan dipelajari, dan membuat kesimpulan sehingga mudah dipahami oleh diri sendiri dan orang lain. Dapat disimpulkan data adalah suatu kumpulan fakta, angka, informasi, atau instruksi yang disajikan dalam bentuk yang dapat diolah atau dipahami.

Data mining menurut Suntoro (2019) adalah proses untuk mendapatkan informasi yang berguna dari basis data yang besar dan perlu diekstraksi agar menjadi informasi baru dan dapat membantu dalam pengambilan keputusan. Sedangkan menurut Larose (2006), data mining adalah sebuah proses menemukan sesuatu bermakna dengan memilah data melalui repository dengan bantuan teknologi sosialisasi pola, statistik, serta matematika. Oleh karena itu data mining adalah proses penemuan pola, hubungan, dan informasi yang bermanfaat dari kumpulan data yang besar atau kompleks.

1. **Konsep**

Konsep data mining pada data melibatkan penggunaan teknik dan algoritma untuk mengeksplorasi dan menganalisis data dengan tujuan mengungkap pengetahuan yang tersembunyi atau pola yang signifikan. Berikut adalah beberapa konsep atau tahapan dalam data mining pada data:

1. Seleksi Data

Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang akan digunakan untuk proses data mining, disimpan dalam suatu berkas, terpisah dari basis data operasional.

1. Pre-processing/ Cleaning ( pemilihan data)

Sebelum proses data mining dapat dilaksanakan, perlu dilakukan proses cleaning pada data yang menjadi fokus KDD. Proses cleaning mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (tipografi). Juga dilakukan proses enrichment, yaitu proses "memperkaya" data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk KDD, seperti data atau informasi eksternal.

1. Transformasi

Coding adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses data mining. Proses coding dalam KDD merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data

1. Data Mining

Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma dalam data mining sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.

1. Interpretasi/Evaluasi

Pola informasi yang dihasilkan dari proses data mining perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses KDD yang disebut dengan interpretation. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya.

1. **Penerapan Data Mining**

Penerapan data mining dalam kehidupan sehari-hari adalah:

1. Kesehatan dan Perawatan Medis

Penerapan ini berhubungan dengan dataset yang kami pilih, yakni data mengenai penyakit jantung. Data mining dapat digunakan dalam hal kesehatan dan perawatan medis untuk mengidentifikasi pola penyakit, diagnosis, dan hasil pengobatan. Ini membantu dalam membuat model prediktif untuk mendeteksi risiko penyakit, memperbaiki prosedur medis, dan meningkatkan pengambilan keputusan klinis.

1. Analisis Sosial Media

Data mining juga dapat diterapkan dalam bidang sosial media untuk memahami tren, sentimen pengguna, dan preferensi mereka. Ini juga dapat digunakan untuk mengidentifikasi pola perilaku dan preferensi yang dapat digunakan untuk tujuan pemasaran atau analisis opini publik.

1. Rekomendasi Produk

Sistem rekomendasi yang berbasis data mining digunakan dalam platform e-commerce, streaming musik, dan aplikasi lainnya untuk memberikan rekomendasi produk atau konten yang disesuaikan dengan preferensi pengguna. Data mining digunakan untuk menganalisis pola konsumsi dan preferensi pengguna serta memprediksi item atau konten yang mungkin diminati.

1. Transportasi dan Logistik

Data mining juga dapat diterapkan dalam bidang transportasi dan logistik untuk mengoptimalkan rute pengiriman, mengelola inventaris, dan memprediksi permintaan. Analisis data dapat membantu mengidentifikasi pola permintaan, memperbaiki efisiensi operasional, dan mengurangi biaya logistik.

1. **Alasan Pemilihan Dataset**

Data yang penulis gunakan untuk analisis dan penelitian ini adalah dataset penyakit jantung. Alasan penulis memilih dataset ini adalah penyakit jantung merupakan salah satu penyakit paling umum dan penyebab utama kematian di banyak negara. Dengan menganalisis dataset tentang penyakit jantung, peneliti dapat mengidentifikasi faktor risiko, pola, dan tren yang berkaitan dengan penyakit ini. Informasi ini dapat digunakan untuk mengembangkan strategi pencegahan, diagnosis dini, dan pengobatan yang lebih efektif. Dataset ini juga mencakup informasi demografis, riwayat medis, faktor risiko, hasil tes, dan tindakan pengobatan dengan ketersediaan data yang melimpah karena telah dikumpulkan oleh berbagai institusi kesehatan dan lembaga penelitian yang memungkinkan peneliti untuk melakukan analisis yang mendalam dan menyeluruh.

Kami memilih dataset ini juga karena adanya kemajuan besar dalam teknologi dan metode analisis yang dapat digunakan untuk memahami penyakit jantung. Contohnya, kecerdasan buatan dan analisis big data telah menghasilkan perkembangan baru dalam pemodelan prediktif dan pemahaman lebih dalam tentang mekanisme penyakit. Dengan menggunakan dataset penyakit jantung, peneliti dapat menerapkan teknologi dan metode terbaru ini untuk mempelajari dan mengidentifikasi pola baru yang relevan dengan penyakit jantung.

1. **Alasan Pemilihan Metode**

Metode yang penulis gunakan adalah klasifikasi dengan menganalisis data untuk digunakan dalam memperkirakan akurasi. Kami memilih metode ini karena penulis dapat mengidentifikasi pola dan hubungan antara variabel dalam data yang akan membantu dalam mengungkap informasi yang berharga dan tersembunyi dalam dataset. Metode ini juga memungkinkan penulis untuk membuat prediksi berdasarkan data yang ada. Dengan membangun model klasifikasi yang baik dapat menggunakan klasifikasi tersebut untuk mengklasifikasikan data baru yang belum diketahui labelnya. Ini membantu dalam pengambilan keputusan yang lebih baik dan memberikan wawasan yang berharga untuk tindakan masa depan. Alasan lainnya dengan mengklasifikasikan data ke dalam kategori atau kelas yang telah ditentukan, penulis dapat mengurangi kompleksitas data dan menggambarkannya dengan cara yang lebih ringkas dan terstruktur. Hal ini berhubungan dengan dataset yang kami pilih karena memiliki data yang banyak.

1. **Alasan Pemilihan Algoritma**

Algoritma yang penulis gunakan dalam menganalisis dan memprediksi untuk mengambil kesimpulan adalah metode Naïve Bayes. Alasan kami memilih algoritma ini adalah karena relatif sederhana dalam konsep dan implementasi dan cocok dengan dataset yang kami gunakan. Algoritma ini mengasumsikan independensi kondisional antara fitur, yang mengurangi kompleksitas perhitungan. Oleh karena itu, algoritma ini dapat menghasilkan model yang cepat dan efisien, terutama pada data yang memiliki dimensi tinggi. Algoritma Naive Bayes cenderung tahan terhadap atribut yang tidak relevan atau atribut yang memiliki dampak kecil terhadap hasil klasifikasi. Karena asumsi independensi kondisional, atribut yang tidak relevan dapat diabaikan, dan metode ini masih mampu menghasilkan hasil yang baik. Naive Bayes juga bekerja dengan baik pada data yang memiliki atribut kategorikal atau diskrit. Dalam konteks klasifikasi, metode ini dapat menghasilkan model yang baik ketika fitur-fitur yang digunakan untuk klasifikasi adalah variabel kategorikal atau biner. Alasan-alasan itulah yang kami pertimbangkan untuk memilih algoritma ini dengan menyesuaikan metode dan dataset yang kami pilih juga.

1. **Tujuan**

Tujuan dari melakukan data mining pada dataset penyakit jantung adalah:

1. Mengetahui berapa banyak pasien yang terkena penyakit jantung pada masing-masing jenis kelamin.
2. Mengetahui berapa banyak pasien yang tidak terkena penyakit jantung pada masing-masing jenis kelamin.

**BAB II**

**PEMBAHASAN**

1. **Sumber dan Jumlah Data**

Data bersumber dari website Kaggle dengan URL <https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset> dan jumlah data sebesar 14 kolom dan baris isi data sebanyak 1025 baris. Data yang kami gunakan merupakan kumpulan data mengenai penyakit jantung yang berasal dari tahun 1988 dan terdiri dari beberapa database. Data ini kami gunakan untuk mengetahui pasien yang terkena dan tidak terkena penyakit jantung.

1. **Jumlah dan Nama Atribut atau Variabel**

Kolom atau atribut berjumlah 14 dengan nama masing-masing atributnya yaitu seperti sebagai berikut:

* *age* : umur.
* *sex* : jenis kelamin. Nilai 0 menandakan perempuan dan nilai 1 menandakan laki-laki.
* *cp* : terdapat empat macam sakit dada.
* *trestbps* : tekanan darah ketika sedang istirahat.
* *chol* : kadar kolesterol darah dalam mg/dl.
* *fbs* : kadar gula darah ketika berpuasa di atas 120 mg/dl atau tidak.
* *restecg* : hasil pemeriksaan electrocardiographic dimana nilai 0 artinya normal, nilai 1 artinya abnormalitas minor, dan nilai 2 artinya abnormalitas yang besar.
* *thalach* : detak jantung maksimal yang didapatkan saat tes fisik..
* *exang* : angina eksersional yaitu rasa nyeri atau ketidaknyamanan pada dada selama ataupun setelah aktivitas fisik..
* *oldpeak* = ST depression induced by exercise relative to rest depresi ST yang dihasilkan oleh aktivitas fisik relatif terhadap istirahat.
* *slope:* kemiringan segmen ST pada *electrocardiographic* selama puncak aktivitas fisik.
* *ca* : menggambarkan jumlah pembuluh utama (0-3) yang terwanakan oleh *fluoroscopy.*
* *thal* : thalassemia untuk menggambarkan hasil tes nuklir dengan nilai 0 artinya normal, nilai 1 artinya efek tetap, dan nilai 2 artinya efek yang dapat dibalik.
* *target* : berisikan nilai 0 artinya negatif penyakit jantung dan nilai 1 artinya positif penyakit jantung.

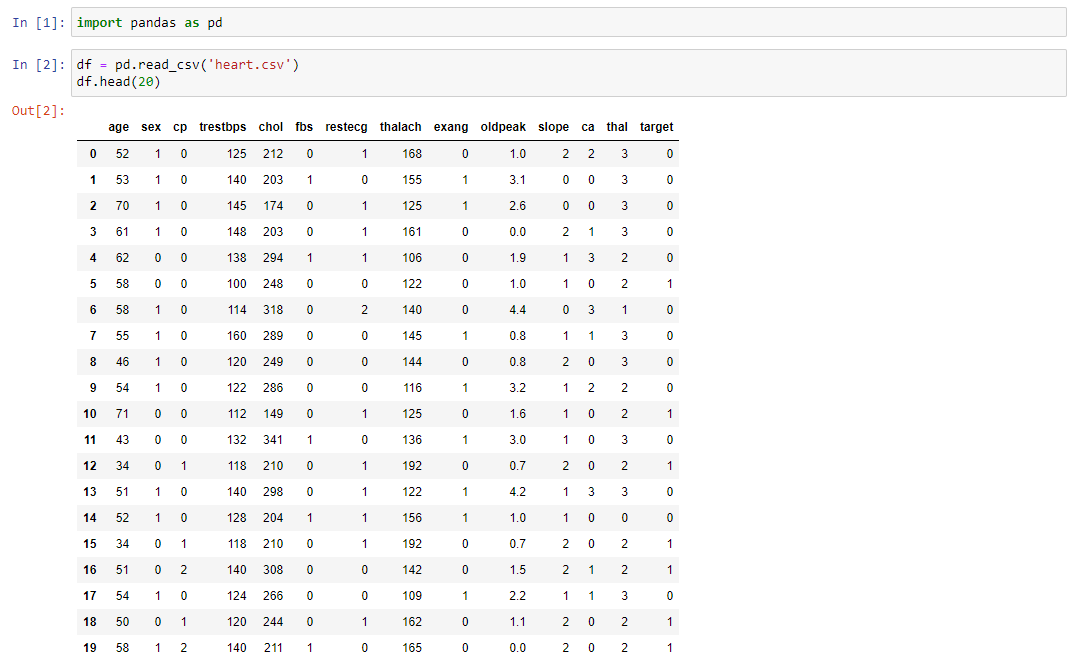
1. **Bentuk Data**

Bentuk data yang digunakan adalah data terstruktur (*structured data*) karena data yang dipakai yaitu *Heart Disease Dataset* terorganisir dalam sebuah format tabel yang sudah ditentukan atribut-atribut serta tipe datanya.

1. **Jenis Pemodelan atau Pembelajaran**

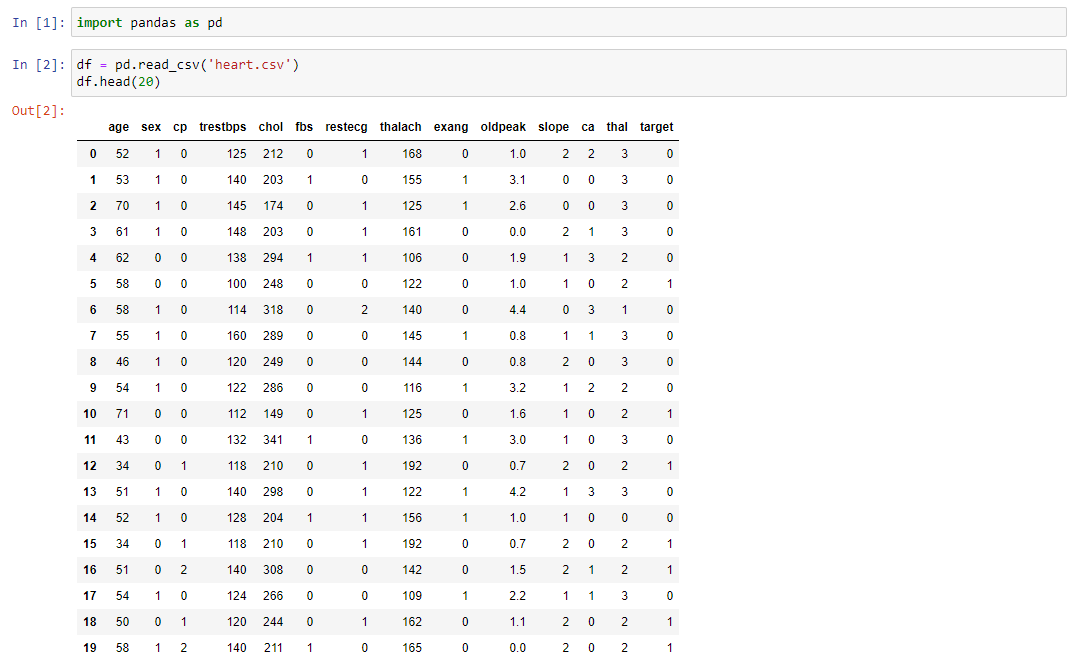
Jenis pemodelan data yang digunakan adalah *supervised learning* dikarenakan tujuannya adalah untuk mengembangkan model yang dapat mempelajari pola dari data yang outputnya sudah diketahui lalu menggunakan model tersebut untuk memprediksi output yang tepat. Dengan menggunakan dataset yang memiliki label target yang diketahui, maka bisa dilakukan pembangun model yang dapat memprediksi nilai target berdasarkan atribut yang ada.

1. **Tahap Kegiatan**
2. **Data Understanding dan Preprocessing**
3. Import library



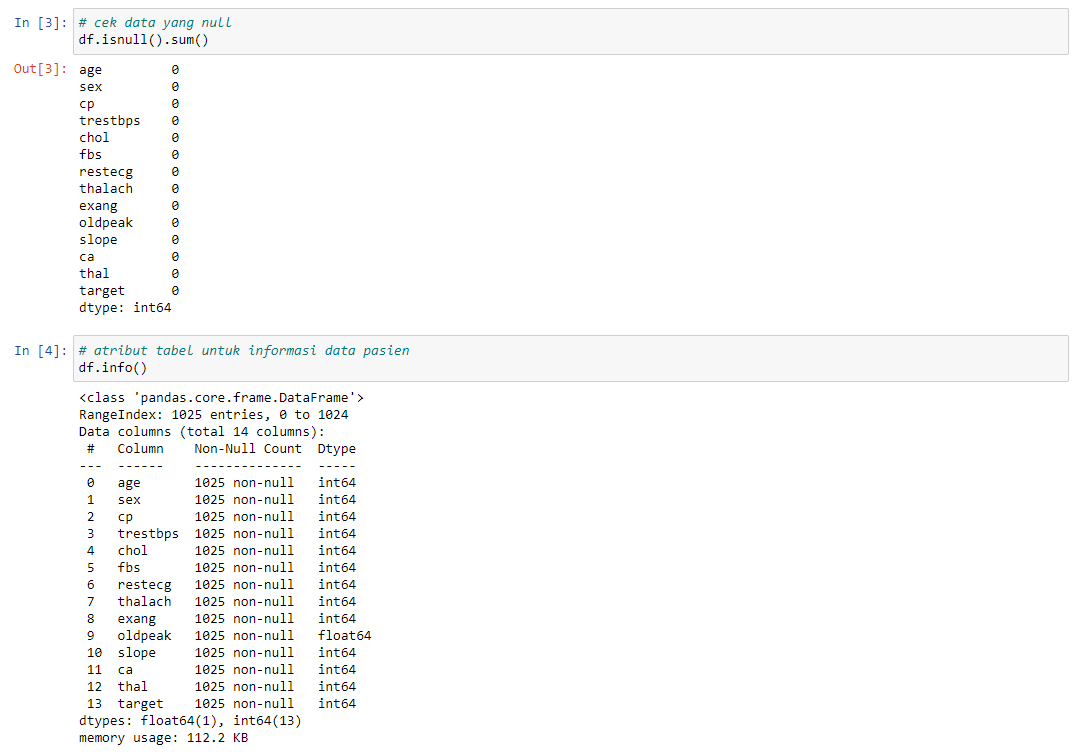
Kode di atas digunakan untuk mengimport library pandas dan memberikan alias ‘pd’ agar lebih mudah digunakan dalam kode selanjutnya. Dengan menggunakan kode diatas, maka kita akan dapat menggunakan berbagai fitur dan fungsionalitas yang ada pada library pandas.

1. Load data



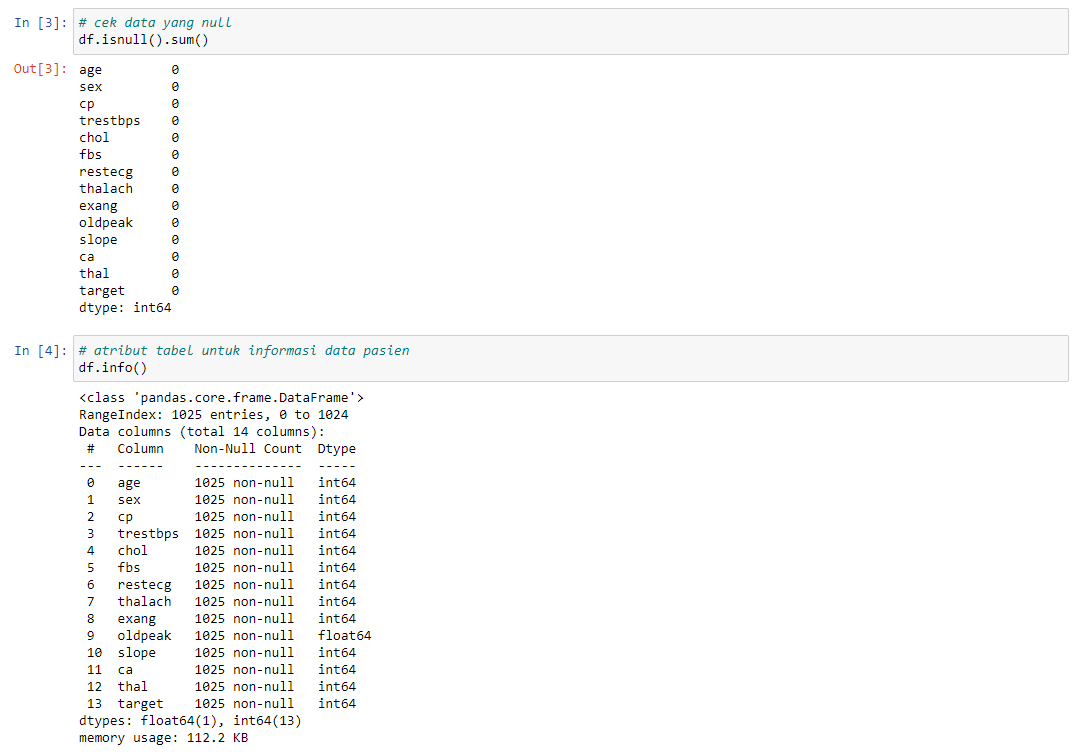
Kode di atas menggunakan library pandas untuk membaca file dataset yang kami gunakan yaitu ‘heart.csv’ dan menyimpannya ke dalam dataframe yang diberi nama ‘df’. Lalu ‘df.head(20)’ digunakan untuk mencetak 20 baris pertama dari dataset tersebut.

1. Cek missing values



Kode di atas merupakan kode untuk mengecek missing values pada data. Method df.isnull().sum() digunakan untuk memeriksa ada atau tidaknya nilai yang hilang pada kolom-kolom dataset.

1. Menampilkan informasi dataset



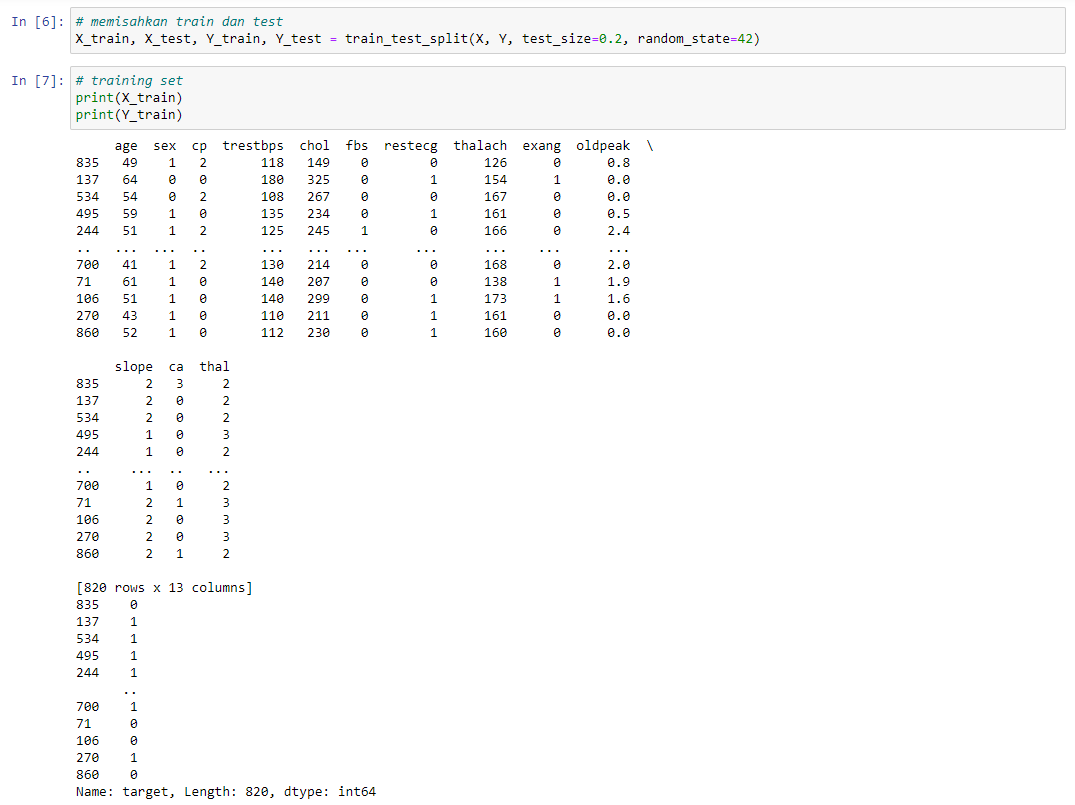
Kode di atas merupakan kode untuk menampilkan informasi umum tentang dataset. Method df.info() akan menampilkan informasi berupa jumlah baris, jumlah kolom, dan tipe data dari setiap kolom.

1. Memisah atribut dan target



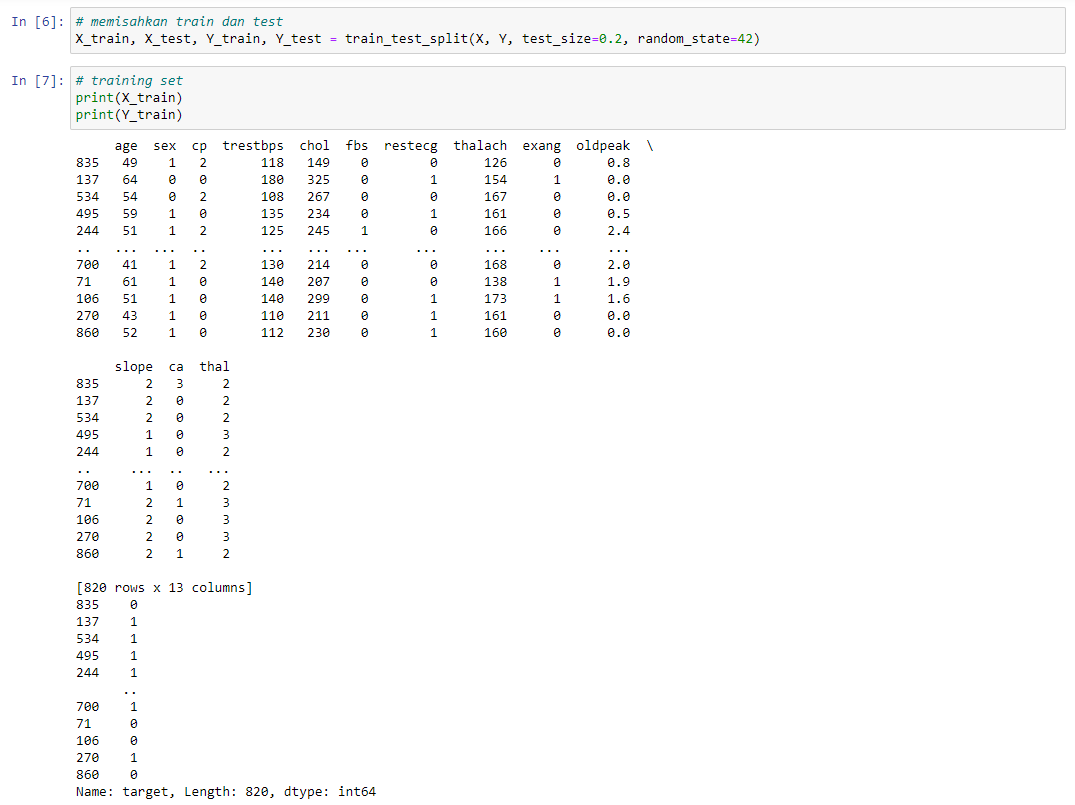
Kode tersebut merupakan kode untuk membuat prediksi dengan data uji. Pada tahap ini akan dilakukan pemisahan atribut yang akan digunakan sebagai variabel X dan target yang akan diprediksi sebagai variabel Y.

1. Membagi data menjadi train set dan test set

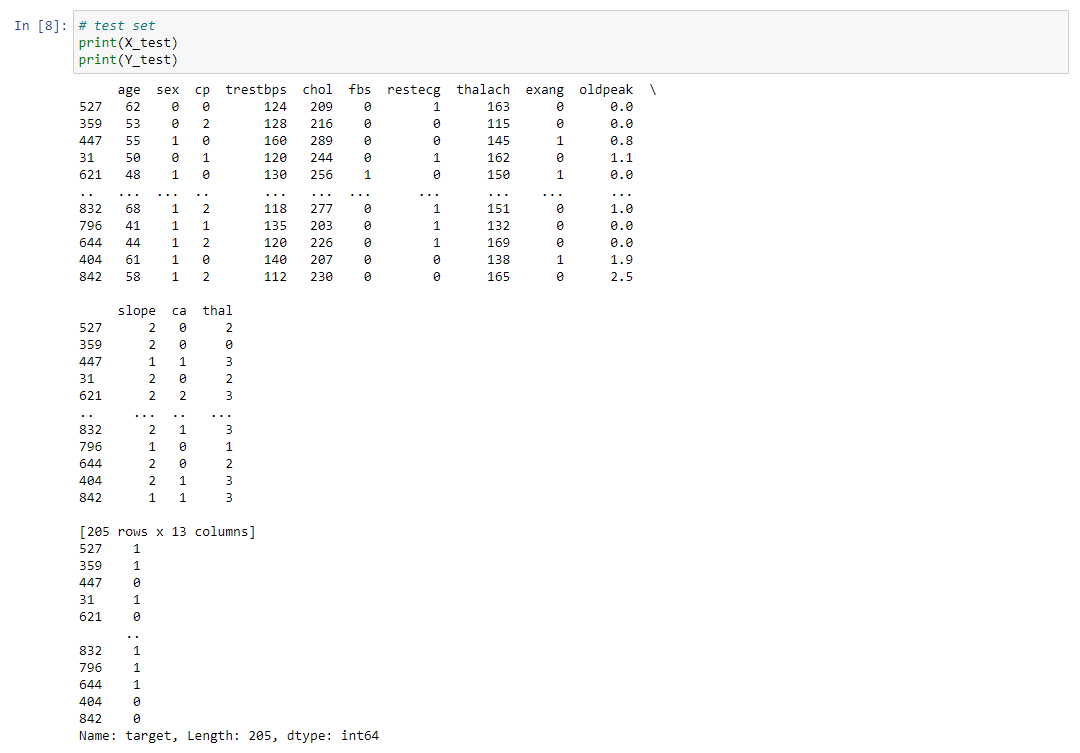


Kode tersebut merupakan kode untuk melakukan pembagian data menjadi train set (X\_train, Y\_train) dan test set (X\_test, Y\_test) menggunakan method train\_test\_split(). test\_size diatur menjadi 0.2 berarti proporsi data yang akan dijadikan data pengujian adalah sebanyak 20% dari jumlah keseluruhan dataset. selanjutnya, random\_state diatur menjadi 42.

Pencetakan data atribut dan target dari train set:



Pencetakan data atribut dan target dari test set:



1. **Classification dan Model Evaluation**
2. Membuat model klasifikasi



Pembuatan model klasifikasi dilakukan menggunakan algoritma Naive Bayes Gaussian dengan modul GaussianNB dari sickit-learn. Model dilatih dengan menggunakan method fit() pada data train set (X\_train dan Y\_train).



Selanjutnya, dilakukan perhitungan jumlah nilai dalam kolom target yaitu jumlah pasien yang terkena atau tidak terkena penyakit jantung. Perhitungan dilakukan menggunakan method value\_counts().

1. Melakukan prediksi



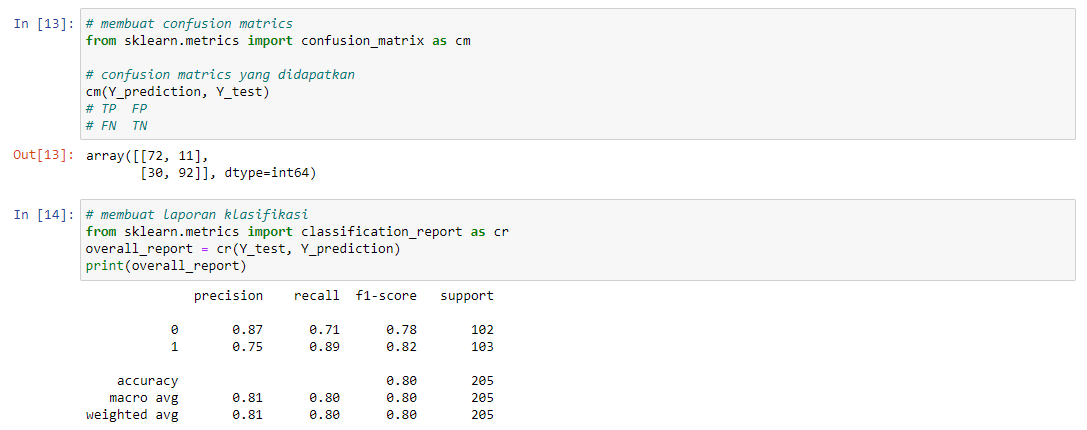
Kode tersebut merupakan kode untuk membuat prediksi menggunakan method predict() pada model yang telah dilatih untuk memprediksi target berdasarkan data pengujian (X\_test).

1. Mengukur akurasi



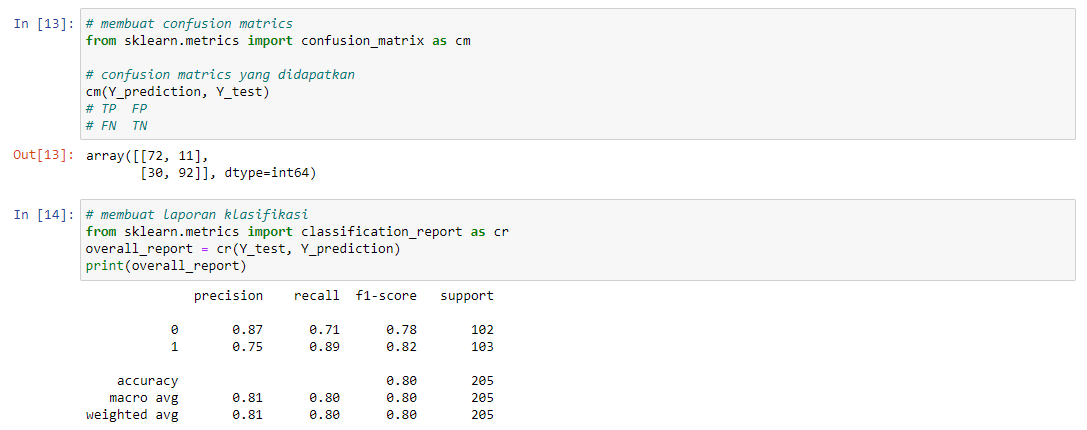
Kode di atas merupakan kode untuk menghitung akurasi model menggunakan method accuracy\_score dari model sklearn.metrics. Perhitungan akurasi dilakukan dengan membandingkan prediksi (Y\_prediction) dengan target sebenarnya (Y\_test) dan didapatkan hasil akurasi sebesar 0,8 atau 80%.

1. Membuat confusion matrix



Kode tersebut akan membuat confusion matrix menggunakan fungsi confusion matrix() dari modul sklearn.metrics. Tahap ini dilakukan dengan membandingkan prediksi (Y\_prediction) dengan target sebenarnya (Y\_test). Output yang dihasilkan adalah sebuah confusion matrix yang menggambarkan jumlah prediksi yang benar dan salah untuk setiap kelas.

1. Report klasifikasi

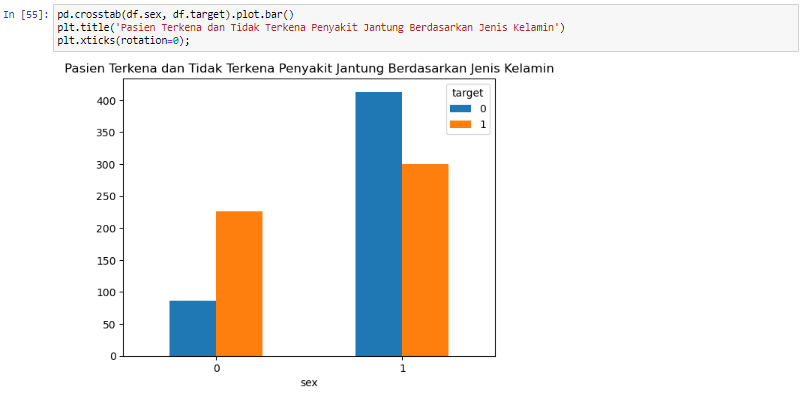


Selanjutnya, terdapat hasil laporan klasifikasi *(classification report)* menggunakan fungsi classification\_report() dari modul sklearn.metrics dengan membandingkan target sebenarnya (Y\_test) dengan prediksi (Y\_prediction). *Classification report* mencakup beberapa metrik seperti *precision, recall, f1-score,* dan *support* untuk setiap kelas.

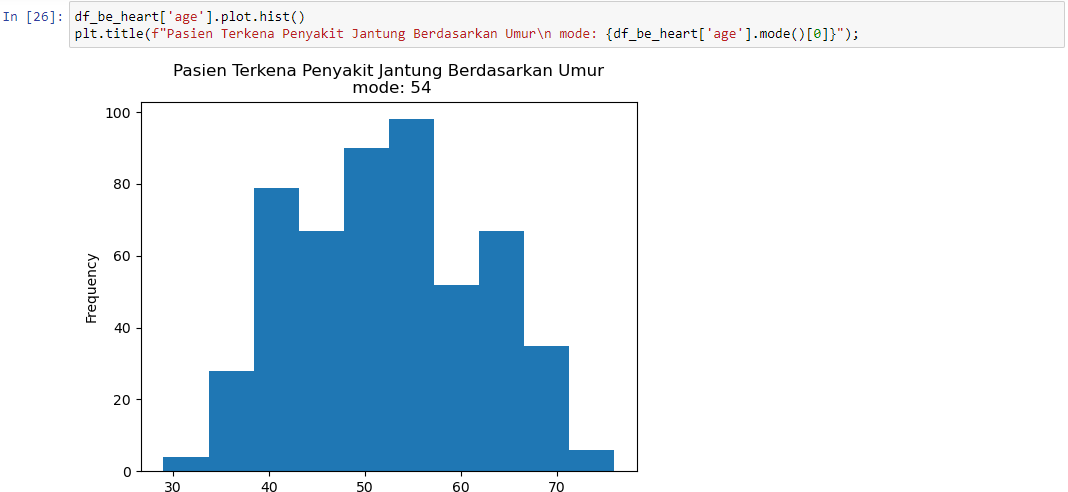
1. **Data Visualization**
2. Data Pasien yang Terkena dan Tidak Terkena Penyakit Jantung



1. Data Pasien yang Terkena dan Tidak Terkena Penyakit Jantung Berdasarkan Jenis Kelamin



1. Data Pasien yang Terkena dan Tidak Terkena Penyakit Jantung Berdasarkan Umur



**BAB III**

**PENUTUP**

1. **Kesimpulan**

Berdasarkan hasil penelitian yang diperoleh, dapat ditarik kesimpulan bahwa pasien yang terkena penyakit jantung lebih banyak berjenis kelamin laki-laki dan usia pasien yang terkena penyakit jantung paling banyak berumur 54 tahun. Klasifikasi data pasien ini mencapai tingkat akurasi untuk *recall* sebesar 89%, *precision* sebesar 75% dan rata-rata akurasi sebesar 80%. Akurasi menunjukkan angka yang cukup tinggi, oleh karena itu dapat disimpulkan bahwa sistem pengklasifikasian data *heart disease* untuk memprediksi keberadaan penyakit jantung pada pasien dapat diimplementasikan menggunakan metode *Naive Bayes*.

1. **Saran**

Pada makalah ini, terdapat pengklasifikasian data *heart disease* untuk memprediksi keberadaan penyakit jantung pada pasien menggunakan metode *Naive bayes*. Oleh karena itu, saran yang dapat kami berikan kepada peneliti selanjutnya adalah adanya klasifikasi lain menggunakan data ini dengan objek prediksi dan metode yang berbeda, sehingga memungkin dapat ditemukannya pola lain dalam melakukan data mining.

**DAFTAR PUSTAKA**

Fikrillah, H., & Kurniadi, D. (2019). Rekomendasi Pemilihan Program Studi Menggunakan Algoritma Naïve Bayes. *Jurnal Algoritma*, 43.

Jeremy, A., Christanti, V., & Mulyawan, B. (2018). Opinion Mining Untuk Ulasan Produk Dengan Klasifikasi Naive Bayes. *Jurnal Ilmu Komputer Dan Sistem Informasi*, 9.

Moriesta, E., & Ibrahim, A. (2017). Analisis Penyaringan Email Spam Menggunakan Metode Naive Bayes. 47.

Nurholiq, A., & Saryono, O. (2019). Analisis Pengendalian Kualitas (Quality Control) dalam Meningkatkan Kualitas Produk. Jurnal Ekonologi Ilmu Manajemen. *Jurnal Ekonologi Ilmu Manajemen*, 395.

Pratiwi, R. (2016). Prediksi Rating Film Menggunakan Metode Naïve Bayes. *Jurnal Teknik Elektro (ISSN 1411-0059)*, 60–63.

Siska, Y. (2019). Penerapan Data Mining dengan Algoritma Naïve Bayes Clasifier untuk Mengetahui Tingkat Kepuasan Pelanggan terhadap Pelayanan Servis Mobil (Studi Kasus: Katamso Service). *Majalah Ilmiah INTI*.