LAPORAN PROYEK SAINS DATA

Heart Attack Analysis & Prediction Dataset

Dosen Pengampu: Eka Mala Sari Rochman, S.Kom., M.Kom.



Disusun oleh:

Rahma Nurhaliza (210411100176)

**PROGRAM STUDI TEKNIK INFORMATIKA**

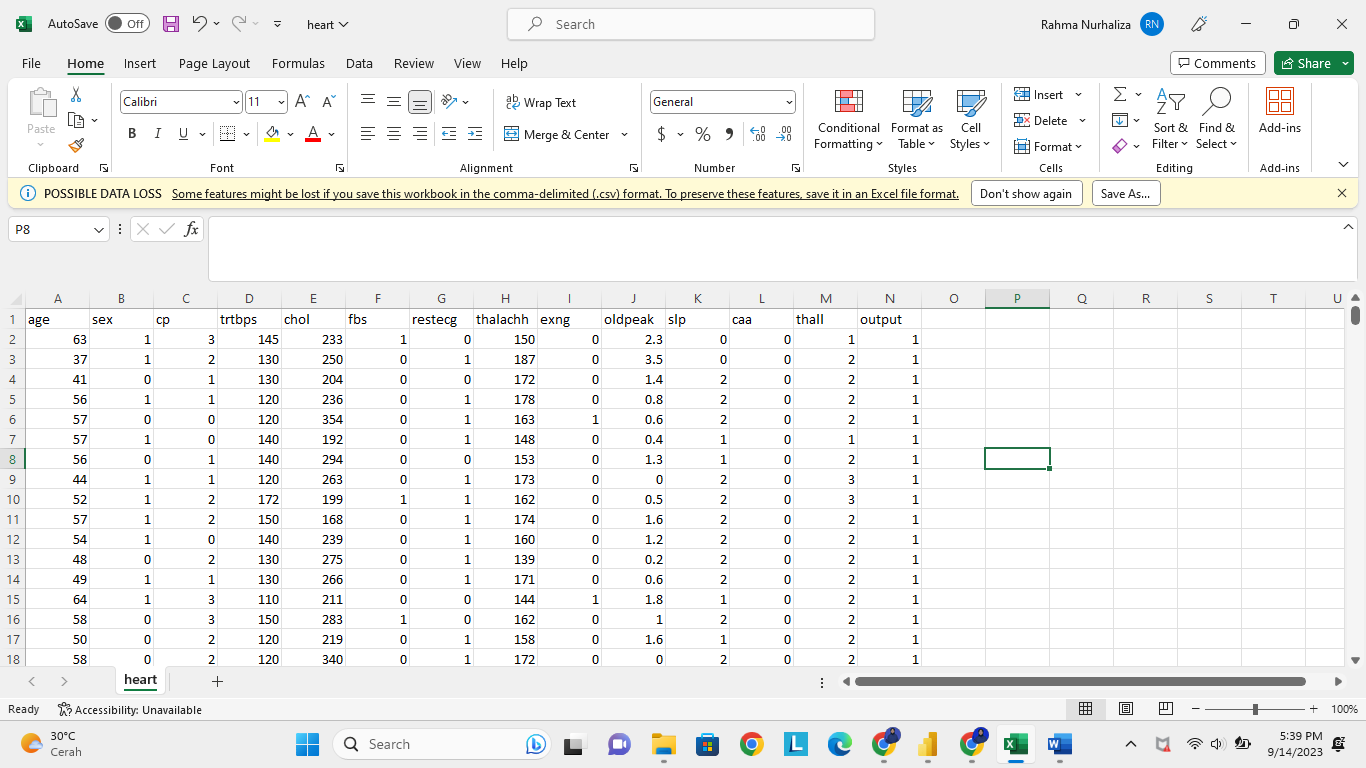
**FAKULTAS TEKNIK**

**UNIVERSITAS TRUNOYO MADURA**

**2023**

1. Data

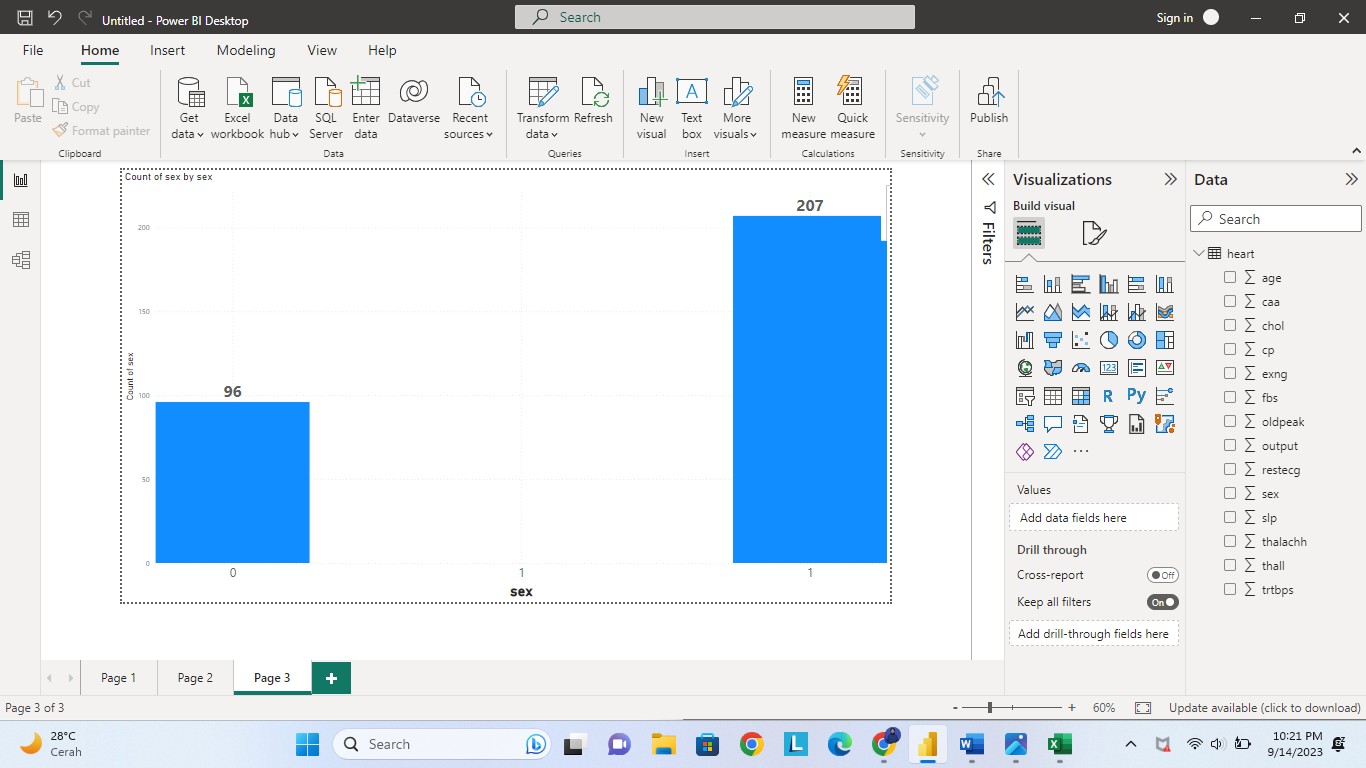
Data yang digunakan: <https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset> (Heart Attack Analysis & Prediction Dataset).

Dataset yang digunakan dalam penilitian ini membahas mengenai penyakit serangan jantung yang didapatkan melalui Kaggle. Secara keseluruhan data ini terdiri dari 303 sampel dengan 13 atribut/parameter dan output sebagai target. terdapat 2 target yang meliputi 0 = lebih sedikit kemungkinan terkena penyakit serangan jantung, dan 1= lebih besar kemungkinan terkena penyakit jantung.

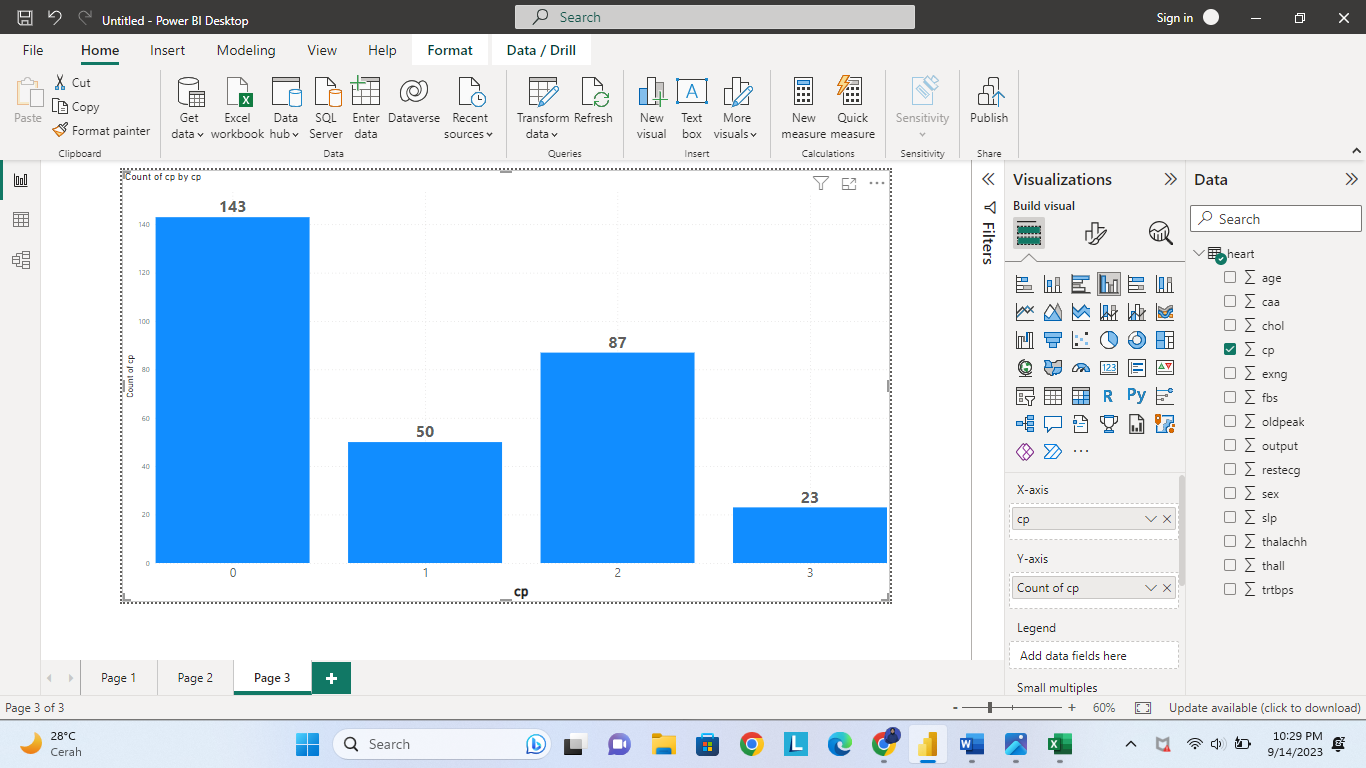
Parameter-parameter yang berkaitan dengan target yang ingin dicapai diantaranya meliputi

1. Age, usia pasien dalam memprediksi kemungkinan lebih besar atau lebih sedikit seseorang terkena serangan jantung maka usia dari pasien merupakan salah satu faktor penting yang mempengaruhi prediksi karena seiring bertambahnya usia maka tubuh mengalami berbagai perubahan fisiologis. Proses penuaan dapat menyebabkan perubahan dalam struktur dan fungsi jantung serta pembuluh darah, salah satunya adalah penumpukan plak dalam arteri (aterosklerosis) yang merupakan faktor resiko utama penyakit jantung.
2. Sex, Jenis kelamin pasien merupakan salah satu faktor penting yang mempengaruhi prediksi dikarenakan terdapat perbedaan hormonal antara laki-laki dan Perempuan dan terdapat perbedaan gejala penyakit jantung antara laki-laki dan Perempuan.
3. Cp, Chest Pain merupakan jenis nyeri dada, nyeri dada menjadi salah satu gejala paling umum dari penyakit jantung dan menjadi peringatan dini tentang resiko penyakit jantung yang lebih besar. Terdapat 4 jenis nyeri dada 0= Typical angina yaitu gejala umum nyeri dada dengan kemungkinan penyumbatan arteri koroner, 1= Atypical Angina yaitu gejala tidak rinci, kemungkinan penyumbatan lebih rendah, 2= Non-anginal pain yaitu rasa menususk dan sakit dalam jangka waktu panjang atau pendek, 3= Asymptomatic yaitu tidak menunjukan gejala penyakit.
4. Trestbps, Resting Blood Pressure (mm Hg) atau tekanan darah istirahat termasuk menjadi faktor penting dalam memprediksi karena jika tekanan darah istirahat tinggi (hipertensi), itu dapat menunjukkan adanya masalah dalam pembuluh darah sehingga mempengaruhi kerja jantung dengan meningkatkan beban kerja jantung.
5. Chol, Cholesterol (mg/dl) sangat diperlukan dalam menprediksi kemungkinan pasien terserang penyakit jantung dikarenakan kolesterol berperan dalam pengembangan aterosklerosis, yaitu penumpukan plak lemak di dalam arteri. Plak tersebut dapat mengakibatkan penyakit jantung. Semakin tinggi kadar chol semakin besar resiko penyakit jantung
6. Fbs, Fasting Blood Sugar, yaitu pengukuran besar gula darah keadaan puasa dalam satuan mg/dl. Parameter Fbs juga sangat diperlukan dalam meprediksi resiko pasien serangan jantung dikarenakan jika gula darah tinggi dalam keadaan puasa dapat menunjukkan adanya resiko diabetes atau kadar gula darah yang tergannggu, dan diabetes menjadi faktor resiko uatama penyakit jantung. Dalam dataset parameter Fbs terbagi menjadi 2 yaitu 0= kuran dari 120 mg/dl, 1= lebih dari 120 mg/dl.
7. Restecg, Resting Electrocardiographic Results, atau hasil elektrokardiografi istirahat. Data elektrokardiografi (EKG atau ECG) sangat diperlukan dalam meprediksi resiko pasien serangan jantung karena mencerminkan aktivitas listrik jantung dan digunakan untuk memantau kesehetan jantung seseorang. Restecg dapat mendeteksi adanya gangguan dalam aktivitas listrik jantung, seperti aritmia(denyut jantung yang tidak teratur), atau iskemia(kekurangan aliran darah ke jantung). Hasil restecg digunakan sebagai salah satu alat diagnosis untuk menilai Kesehatan jantung seseorang. Dalam dataset Restecg terbagi menjadi 3 yaitu 0= normal, 1= gelombang ST meningkat/ menurun lebih dari 0,5 mV, 2 = Venticular kiri mengalami hipertropi.
8. Thalach, yaitu Maximum Heart Rate Achieved" atau "Detak Jantung Maksimum yang Dicapai". Thalach adalah parameter yang penting dalam memprediksi risiko penyakit jantung karena thalach mencerminkan kemanpuan jantung untung mencapai detak jantung maksimum selama aktivitas fisik atau stress. Jantung yang sehat memiliki kapasitas untuk meningkatkan denyut jantung saat diperlukan selama aktivitas fisik atau stress. Sedangkan detak jantung yang rendah atau anomali dalam respons jantung selama aktivitas fisik dapat menjadi tanda peringatan yang memengaruhi prediksi risiko seseorang terkena penyakit jantung.
9. Exang, "Exercise Induced Angina” atau nyeri akibat olahraga. Parameter Exang juga sangat diperlukan dalam meprediksi resiko pasien serangan jantung dikarenakan angina merupakan nyeri dada atau ketidaknyamanan yang terjadi ketika jantung tidak mendapatkan cukup oksigen yang dibutuhkan, biasanya selama aktivitas fisik atau stres. Jika seseorang melaporkan adanya angina yang dipicu oleh latihan (Exang), itu bisa menjadi tanda bahwa mereka memiliki masalah jantung yang mendasari. Dalam dataset parameter Exang terbagi menjadi 2 yaitu 0= tidak nyeri, 1=nyeri
10. Oldpeak, yaitu besar segemen ST dari olahraga relative terhadap kondisi istirahat sangat diperlukan dalam meprediksi resiko pasien serangan jantung dikarenakan tingkat depresi ST selama latihan adalah petunjuk tentang sejauh mana iskemia miokard terjadi dan dapat membantu dalam penentuan risiko penyakit jantung.
11. Slope, yaitu besar kemiringan segmen ST pada kondisi latihan pucak atau maksimum.parameter tersebut sangat diperlukan dalam meprediksi resiko pasien serangan jantung dikarenakan Slope dari segmen ST selama tes stres jantung mengindikasikan sejauh mana perubahan dalam segmen ST terjadi selama aktivitas fisik. Jika ada perubahan yang signifikan dalam segmen ST (seperti depresi ST atau elevasi ST), itu dapat menjadi indikasi adanya iskemia miokard, yang berarti otot jantung mungkin tidak mendapatkan cukup pasokan darah dan oksigen selama Latihan, sehingga dapat membantu dalam menentukan resiko penyakit jantung. Dalam dataset slope terbagi menjadi 3 yaitu angka 0= tidak adanya perubahan segemen ST yang signifikan pada EKG pasien, angka 1= menunjukkan elevasi segemen ST pada EKG pasien, angka 2 = Mengindikasi depresi segmen ST pada EKG pasien.
12. Ca, Jumlah Coronary Arteries atau jumlah pembuluh darah adalah parameter yang penting dalam memprediksi risiko penyakit jantung koroner karena umlah pembuluh darah utama yang tersumbat adalah tanda yang jelas dari penyakit jantung koroner. Ketika satu atau lebih pembuluh darah utama yang memasok darah ke jantung mengalami penyumbatan (stenosis) akibat penumpukan plak, itu dapat mengganggu aliran darah ke jantung dan menyebabkan angina atau serangan jantung.Parametes Caa terbagi menjadi 5 yaitu 0(normal), 1(penyumbatan Sebagian ringan), 2(penyumbatan Sebagian), 3(penyumpatan parah), dan 4(penyumbatan Ekstensif).
13. Thall, Breakdown for thalium test, yaitu status Kesehatan jantung. Thalium digunakan dalam pemindaian nuklir jantung untuk membantu memprediksi dan mendiagnosis masalah jantung, terutama pada pasien yang memiliki risiko serangan jantung atau mengalami gejala jantung yang tidak jelas. Status jantung yang dibagi menjadi 4 diantaranya, 0 = tidak diketahui, 1 = cacat tetap, 2 = normal, 3 = cacat reversibel.

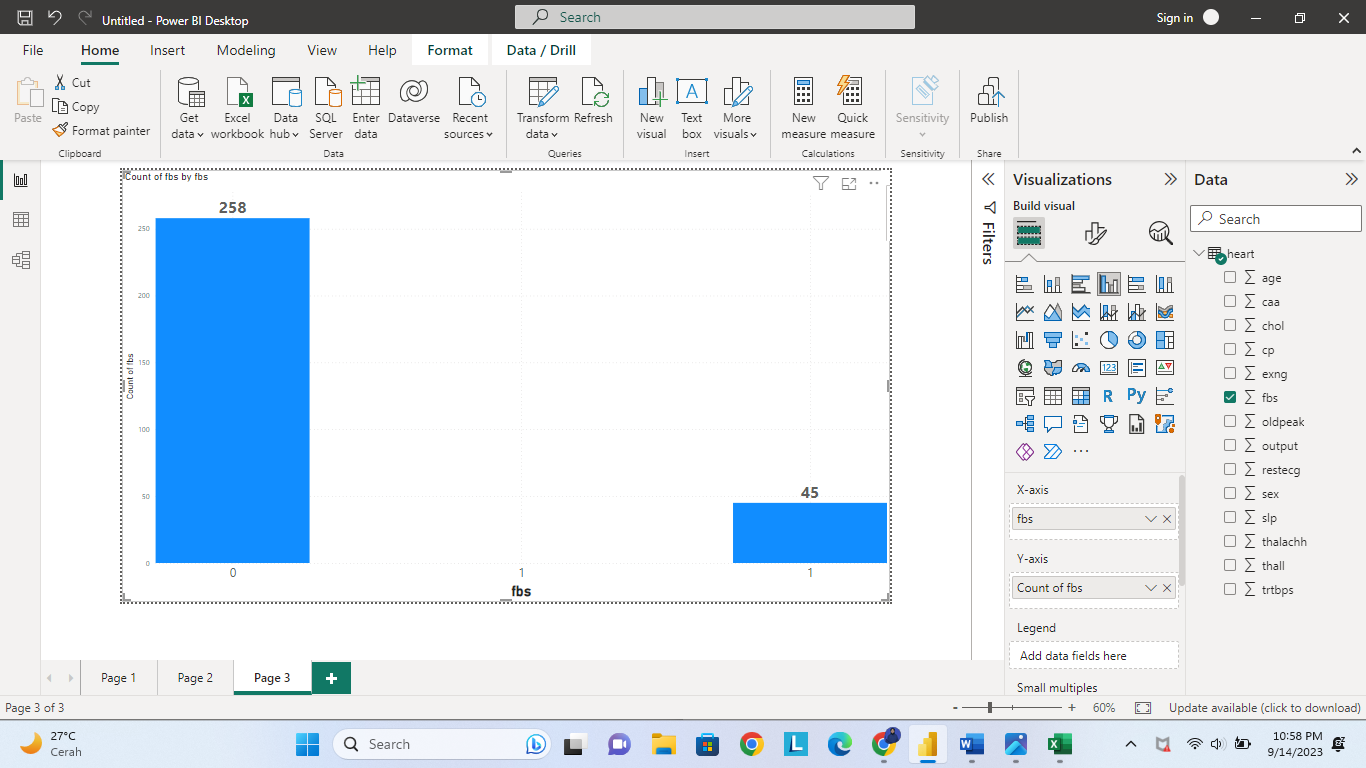
Analisis Sebaran Data

Pembagian Berdasarkan Gender

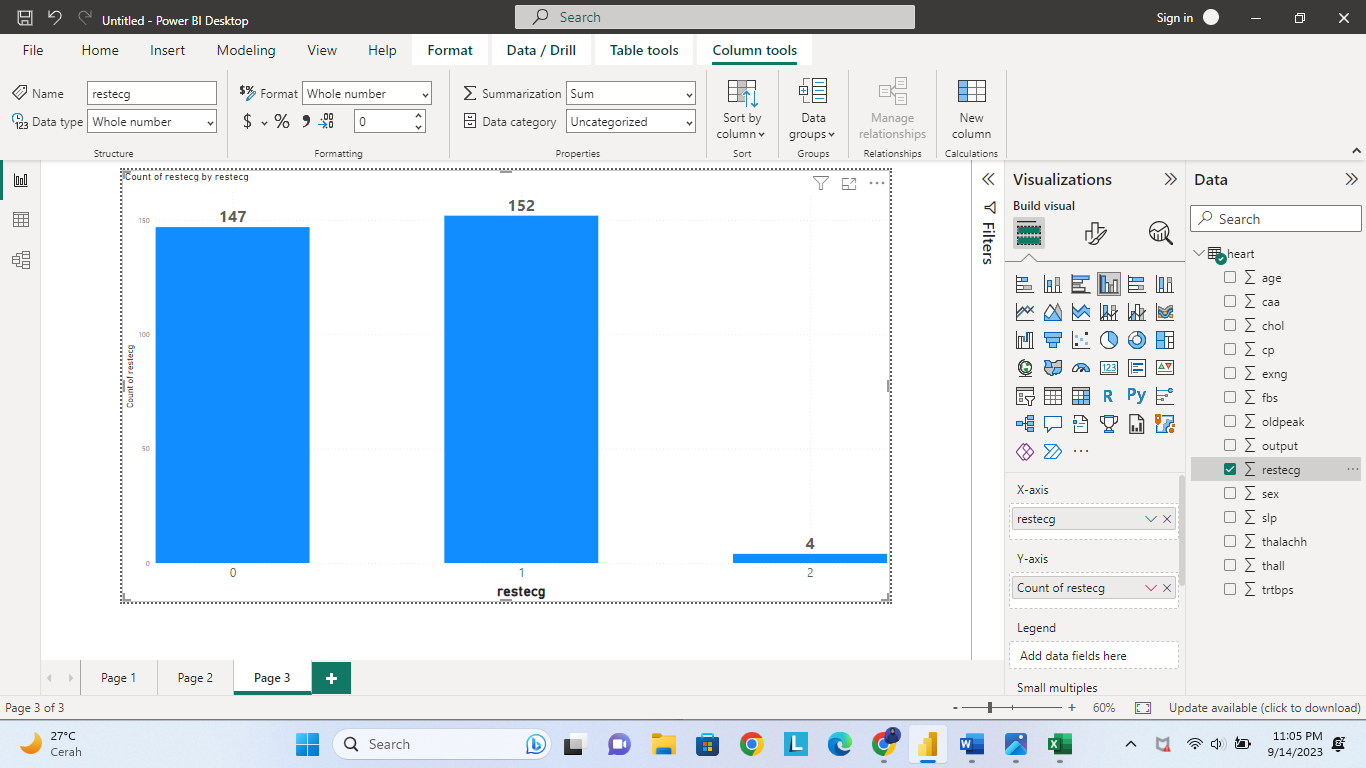
Apabila dirincikan dalam jenis kelamin maka terdapat 2 variabel yang dihasilkan yaitu 0 = Perempuan dan 1 = Laki-Laki. Dari total jumlah 303 orang diantaranya 96 orang berjenis kelamin Perempuan dan 207 orang beserjenis kelamin Laki-Laki. Jika dilihat dari data yang dihasilkan maka Laki-Laki lebih rentan terkena penyakit serangan jantung.

Berdasarkan Pembagian Tipe Nyeri Dada (cp)

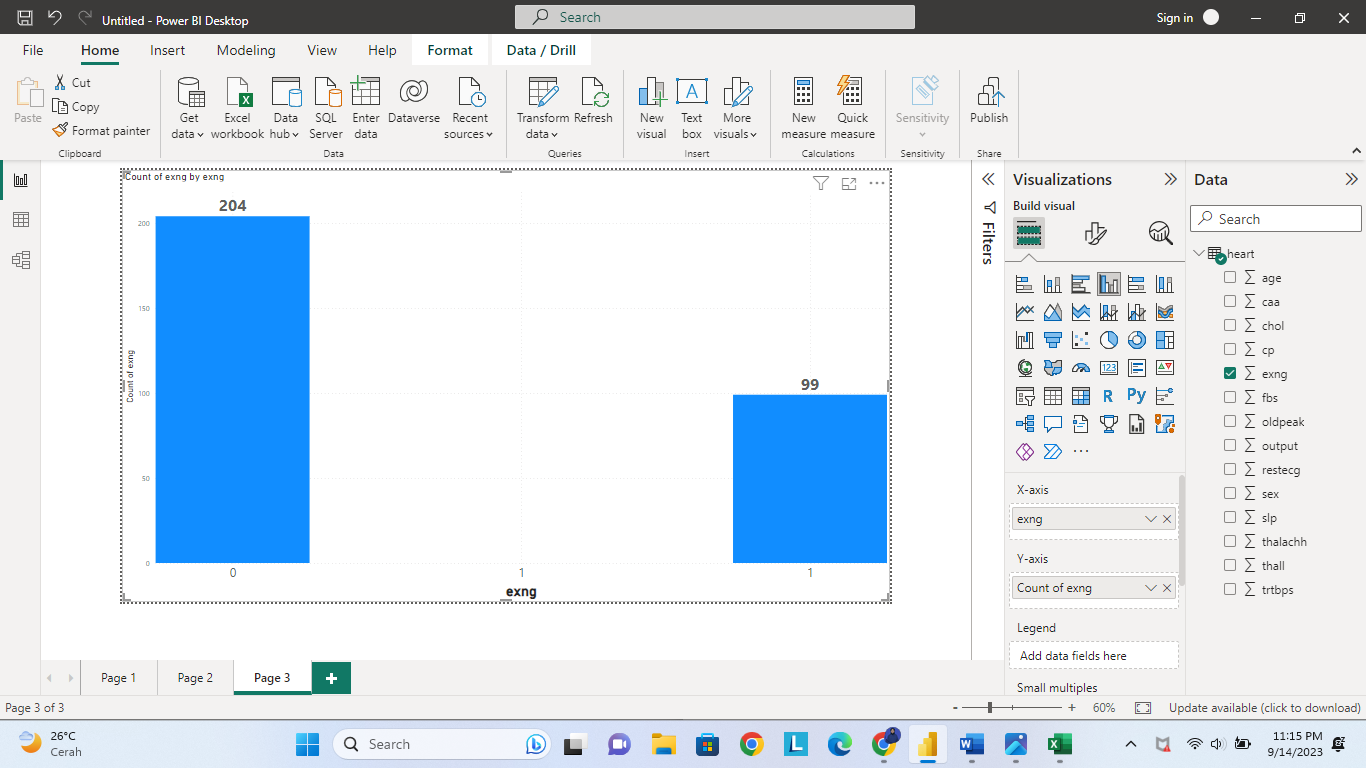
Dari sebaran data nyeri dada (cp) dapat dilihat untuk pasien yang mengidap tipe 0 (Typical Angina) terdapat 143 orang. Pasien yang mengidap tipe nyeri dada 1 (Atypical Angina) berjumlah 50 orang. Pasien yang mengidap tipe 2 (Non-Anginal Pain) adalah 87 orang dan Sisanya 23 orang tidak menunjukkan gejala penyakit Tipe 3 (Asymptomatic). Dapat dilihat juga bahwa orang yang mengalami nyeri dada tipe 0 (Typical Angina) adalah yang tertinggi sedangkan orang yang memiliki nyeri dada tipe 3 (Asymptomatic) adalah yang terendah. Penderita nyeri dada tipe 0 hampir 50% lebih banyak dari semua tipe nyeri dada yang dialami oleh pasien

Berdasarkan Nilai Jumlah Gula Darah (fbs)

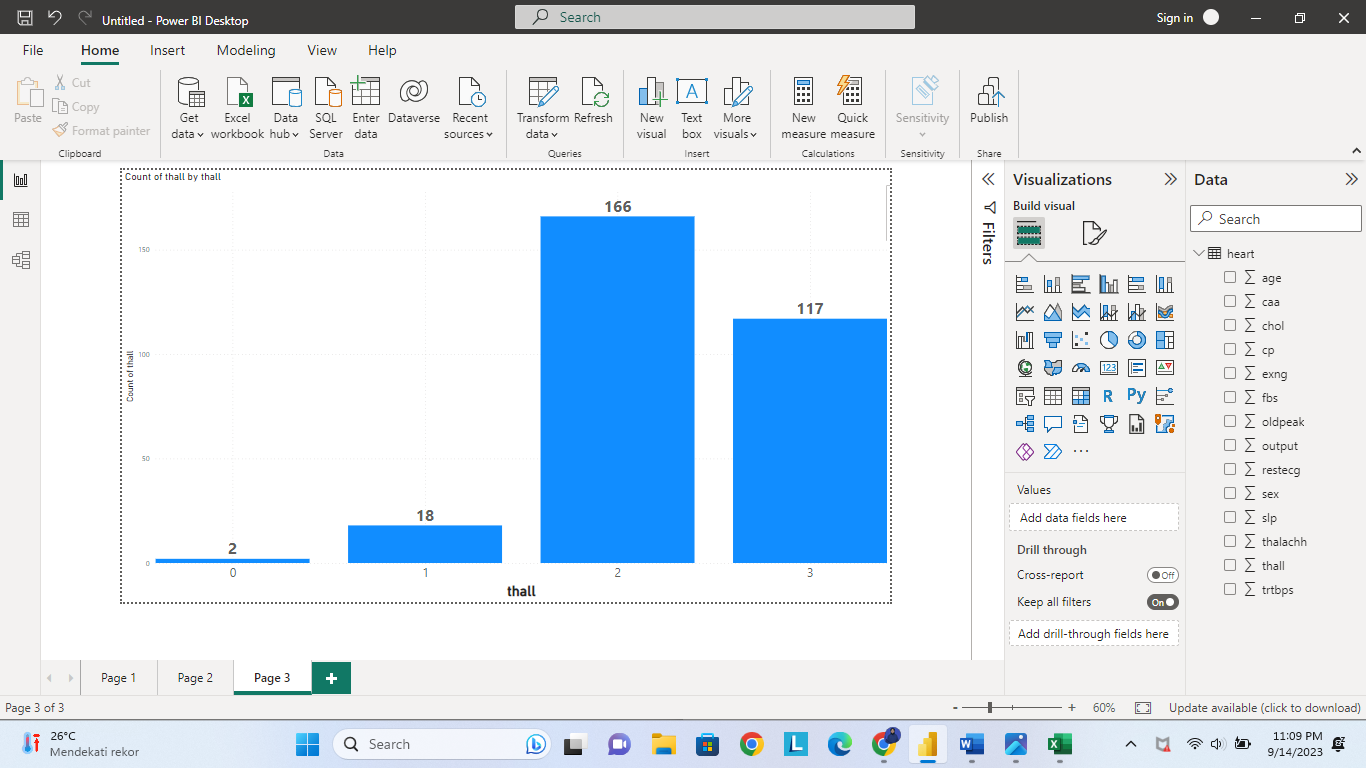
Jumlah gula darah dalam satuan mg/dl, lebih besar atau kurang dari 120mg/dl. (0 = kurang dari 120mg/dl, 1 = lebih dari 120mg/dl). Data yang dihasilkan menunjukkan bahwa terdapat 258 orang yang jumlah gula darahnya kurang dari 120mg/dl (0) dan 45 orang yang jumlah gula darahnya lebih dari 120mg/dl (1). Jumlah gula darah yang kurang dari 120mg/dl secara signifikan lebih tinggi.

Berdasarkan Denyut Jantung (restecg)

ECG (Restecg) yaitu hasil elektrokardiografi yang dimana 0 = Normal, 1 = gelombang ST meningkat / menurun, 2 = Venticular kiri mengalami hipertrofi. Dari pemaparan data tersebut, terdapat 147 orang yang hasil elektrokardiografi kesehatan jantungnnya Normal (0). Ada 151 orang yang gelombang detak jantungnya meningkat / menurun (1) dan terdapat 4 orang yang venticular kiri pada jantungnya mengalami hipertrofi (Peningkatan dan Pertumbuhan sel otot pada jantung).

Berdasarkan Nyeri Dada Saat Berolahraga (exng)

Pada dataset pengolahan data terdapat variabel Exang yaitu Nyeri dada setelah berolahraga. Dengan keterangan : 0 ( tidak nyeri), 1 (nyeri). Berdasarkan data yang dihasilkan, orang yang mengalami nyeri dada (1) setelah olahraga adalah sebanyak 99 orang, daan 204 orang tidak mengalami nyeri dada setelah berolahraga (0) hal ini tergantung pada kondisi fisik setiap orang yang berbeda-beda.

Berdasarkan Status Kesehatan Jantung (thall)

Dari pemaparan data yang dihasilkan, terdapat 303 jumlah total orang dengan kondisi status jantung yang berbeda. Ada 166 orang dengan status kesehatan jantung yang normal (2), 117 orang dengan status kesehatan jantung cacat reversibel (kondisi dimana cacat jantung yang masih bisa dilakukan tindak operasi dan berpotensi normal kembali setelah melakukan operasi). 18 orang dengan kondisi cacat tetap (irreversibel) tidak dapat sembuh total. dan 2 orang dengan status jantung yang tidak diketahui dikarenakan adanya beberapa faktor seperti penyumbatan arteri yang menyebabkan kondisi jantung tidak menentu dari waktu ke waktu.

1. Data Cleaning

Data Cleaning merupakan suatu proses mempersiapkan data untuk di analisis dengan menghapus atau memodifikasi data yang tidak benar, tidak lengkap tidak relevan, diduplikasi, atau diformat dengan tidak benar. Data cleaning adalah langkah penting dalam pengolahan data karena data yang tidak valid atau tidak relevan dapat menghasilkan hasil analisis yang tidak akurat atau tidak dapat diandalkan. Pada dataset mengenai penyakit serangan jantung ini tidak terdapat parameter yang perlu untuk di hapus karena semua aparameter atau atribut sangat diperlukan dalam memprediksi besar atau kecilnya seseorang terkena penyakit serangan jantung agar menghasilkan hasil analisis yang akurat.

1. Transformasi Data

Transformasi data merupakan proses untuk mengubah struktur atau karakteristik data agar lebih cocok digunakan untuk analisis atau pemodelan. Hal ini dilakukan untuk memenuhi kebutuhan spesifik analisis yang akan dilakukan dan untuk meningkatkan kualitas data. Transformasi data mencakup berbagai tindakan yang dapat dilakukan pada data mentah untuk menghasilkan data yang lebih relevan dan bermakna. Salah satu metode yang digunakan dalam proses transformasi data adalah Normaliasi. Normalisasi merupakan suatu proses mengubah data ke dalam skala yang seragam atau dalam rentang tertentu. Ini umumnya digunakan ketika ada variabel-variabel dalam dataset yang memiliki satuan atau skala yang berbeda. Contoh teknik normalisasi meliputi Min-Max scaling dan Z-score normalization. Pada dataset mengenai penyakit serangan jantung ini tidak terdapat data yang tidak perlu di normalisasi karena semua data telah seragam, dan sudah dalam bentuk numerik.

1. Missing Value & Cara mengatasi

Missing Value atau nilai yang menghapus adalah istilah yang digunakan dalam analisis data untuk menggambarkan situasi ketika data yang seharusnya ada dalam suatu variabel atau kolom tidak tersedia atau tidak terisi lengkap dalam dataset. Nilai yang hilang dapat disebabkan oleh berbagai alasan, seperti kesalahan input, sensor yang gagal merekam data, atau ketidaklengkapan dalam pengumpulan data. Kehadiran nilai yang hilang dapat memiliki dampak signifikan pada analisis data dan harus dikelola dengan hati-hati. Terdapat beberapa cara yang digunkan untuk mengatasi data missing value diantaranya:

1. Penghapusan data, yaitu dengan menghapus baris atau kolom yang mengandung nilai yang hilang. Penghapusan data hanya cocok digunakan jika nilai yang hilang relative kecil dan tidak terlalu mempengaruhi secara signifikan hasil dari analisis.
2. Imputasi Sederhana, yaitu dengan Teknik pengisisan nilai yang hilang dengan nilai yang diestimasi berdasarkan data yang ada. beberapa metode imputasi diantaranya yaitu Mean (mengganti nilai yang hilang dengan nilai rata-rata dari seluruh data dalam variabel tertentu), Median (mengganti nilai yang hilamg dengan nilai media dari seluruh data dalam variabel tersebut), dan Modus (Mengganti nilai yang hilang dengan nilai modus, atau nilai yang paling sering muncul dari seluruh data dalam variabel tersebut).
3. Menggunakan Machine learning yaitu dengan memanfaatkan metode pengisisan atau prediksi nilai yang hilang dengan memanfaat algoritma machine learning.
4. Balancing Data & Cara Mengatasi

Balancing data adalah proses mengatasi masalah ketidakseimbangan kelas dalam sebuah dataset. Ketidakseimbangan kelas terjadi ketika jumlah sampel dalam setiap kelas dalam dataset tidak seimbang, artinya satu atau beberapa kelas memiliki jumlah sampel yang jauh lebih sedikit atau jauh lebih banyak dibandingkan dengan kelas lainnya. Masalah ini dapat memengaruhi kinerja model pembelajaran mesin, terutama dalam tugas klasifikasi di mana model harus memprediksi kelas dari data yang tidak seimbang. Berikut beberapa teknik yang digunakan untuk menyeimbangkan data

1. Oversampling: yaitu dengan duplikasi atau penambahan ulang sampel dari minoritas ke kelas mayoritas sehingga jumlahnya menjadi seimbang.
2. Undersampling: yaitu dengan melibatkan pengapusan atau pengurangan sampel dari kelas mayoritas sehingga jumlahnya menjadi seimbang dengan kelas minoritas.
3. SMOTE (synthetic minority over-sampling technique): Teknik SMOTE merupakan Teknik yang paling banyak digunakan untuk mengatasi data yang tidak seimbang. Tekniknya mirip dengan oversampling yang menduplikasi data dari kelas minoritas sehingga jumlahnya sama dengan jumlah kelas mayoritas. Tetapi, Teknik SMOTE tidak hanya dengan melakukan duplikasi data yang sama, tetapi akan membuat sampel baru yang menyerupai data asli dari kelas minoritas sehingga kelas minoritas menjadi jauh lebih beragam.