**LAPORAN TUGAS**

**ANALISIS DATA PENYAKIT DIABETES**

**MENGGUNAKAN POWER BI**

**Sebuah gambar berisi teks, lambang, simbol, lingkaran

Deskripsi dibuat secara otomatis**

**DISUSUN OLEH :**

**TAUFIQU REZA YOGA P 210411100014**

**PRODI TEKNIK INFORMATIKA**

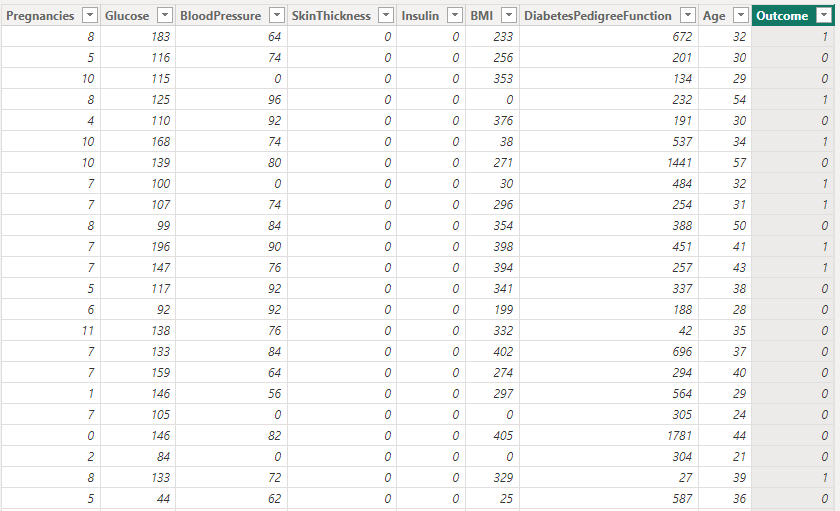
**JURUSAN TEKNIK INFORMATIKA**

**FAKULTAS TEKNIK**

**UNIVERSITAS TRUNOJOYO MADURA**

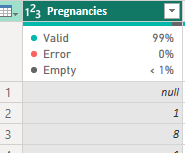
**2023**

**Analisa data diabetes (Cleaning Data)**



Tahap pertama yang harus dilakukan setelah kita mencari dataset di Kaggle yakni melakukan Cleaning data. Dimana di proses Cleaning data ini kita harus melakukan seleksi fitur apakah fitur yang ada pada dataset tersebut bisa digunakan atau tidak, serta mengecek apakah ada missing value atau data kosong pada setiap fitur. Jika dirasa ada fitur yang memiliki missing value yang banyak bisa dilakukan remove atau penghapusan fitur. Pada dataset diabetes ini hanya ada satu missing value pada fitur Pregnancies maka kita harus mengisi data tersebut tidak perlu melakukan penghapusan fitur. Untuk fitur yang digunakan yaitu semua fitur digunakan dikarenakan tidak ada missing value pada fitur yang lain.

**Missing Value**



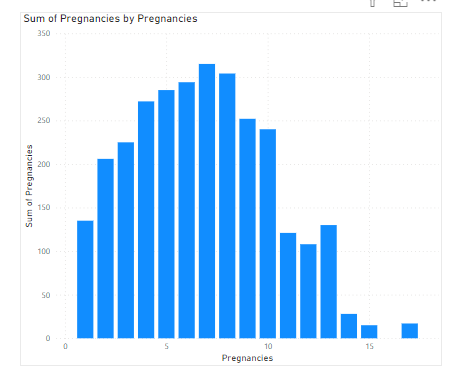
Missing value terdapat pada fitur Pregnancies dimana di situ terdapat 1 data yang kosong. Maka kita harus mengisi data tersebut dengan null atau nan.

**Visualisasi Dataset Diabetes**

Berikut ini adalah fitur-fitur yang ada pada dataset diabetes :

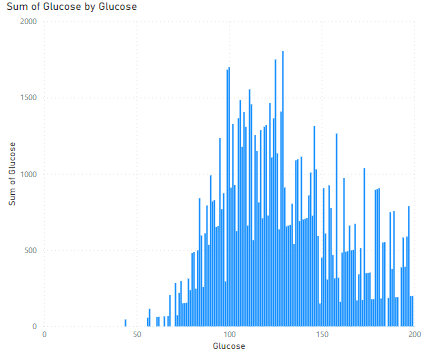
1. Pregnancies

Dalam dataset penyakit diabetes, kolom "pregnancies" mengacu pada jumlah kehamilan yang pernah dialami oleh individu yang terdapat dalam dataset tersebut. Kolom ini mencatat berapa kali seorang individu pernah hamil selama hidupnya.



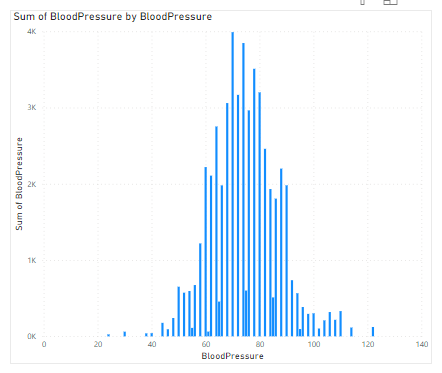
1. Glucose

Data glucose dalam konteks penyakit diabetes mengacu pada tingkat glukosa (gula) dalam darah seseorang. Glukosa adalah jenis gula yang merupakan sumber utama energi bagi tubuh manusia. Tingkat glukosa dalam darah dapat berfluktuasi sepanjang hari dan dipengaruhi oleh berbagai faktor seperti makanan yang dikonsumsi, aktivitas fisik, dan metabolisme tubuh.



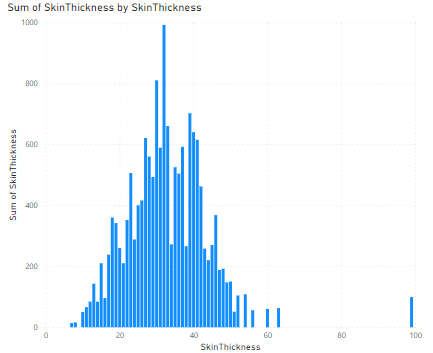
1. BloodPressure

Data "Bloodpressure" (tekanan darah) dalam konteks dataset penyakit diabetes mengacu pada pengukuran tekanan darah pasien yang termasuk dalam dataset tersebut. Tekanan darah adalah pengukuran penting dalam analisis penyakit diabetes karena tekanan darah yang tinggi (hipertensi) adalah salah satu faktor risiko yang dapat berkontribusi pada perkembangan diabetes tipe 2 atau dapat menjadi komplikasi dari diabetes.



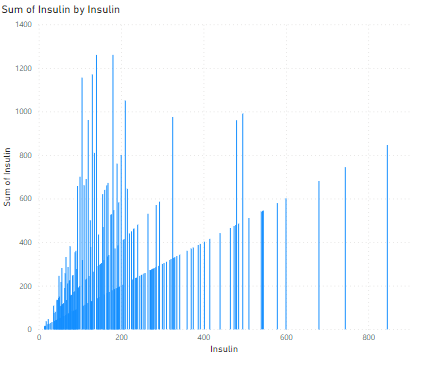
1. SkinThickness

Dalam konteks data penyakit diabetes, "Skinthickness" (tebal kulit) adalah salah satu fitur atau kolom yang mengukur ketebalan lipatan kulit pada pasien. Ketebalan lipatan kulit dapat diukur dalam milimeter (mm) dan sering digunakan sebagai salah satu atribut dalam analisis diabetes, terutama dalam pemodelan prediktif untuk memahami hubungannya dengan risiko diabetes atau faktor-faktor lain yang terkait dengan kondisi ini.



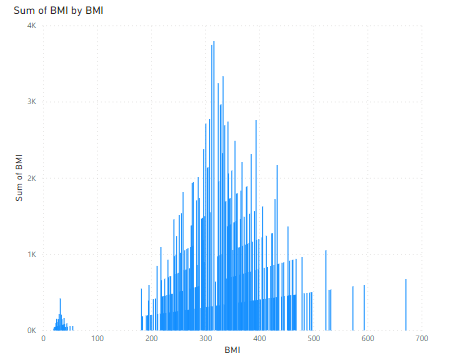
1. Insulin

Dalam konteks data penyakit diabetes, "insulin" adalah salah satu atribut atau fitur yang mungkin ada dalam dataset. Ini mengacu pada tingkat atau jumlah hormon insulin yang ada dalam darah seseorang. Insulin adalah hormon yang diproduksi oleh pankreas dan memiliki peran penting dalam mengatur kadar gula darah.



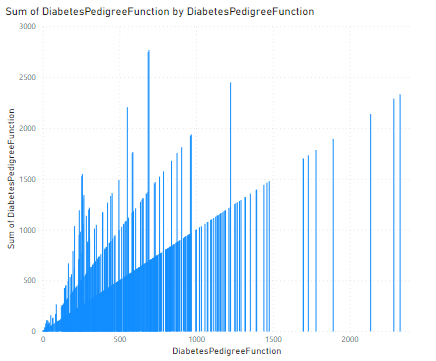
1. BMI

BMI (Body Mass Index) adalah sebuah ukuran yang digunakan untuk mengevaluasi hubungan antara berat badan seseorang dan tinggi badannya. BMI biasanya digunakan sebagai indikator untuk menentukan apakah seseorang berada dalam kategori berat badan yang sehat, kurang berat badan, kelebihan berat badan, atau obesitas.



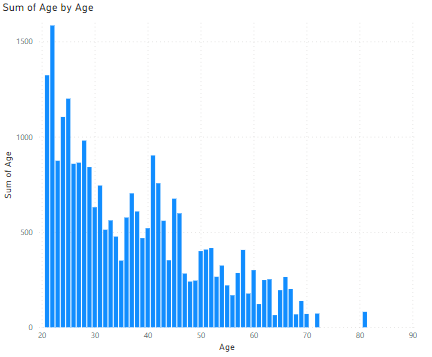
1. DiabetesPedigreeFunction

DiabetesPedigreeFunction adalah suatu angka yang menggambarkan sejauh mana riwayat keluarga seseorang memiliki diabetes dan seberapa besar kemungkinan genetiknya terpengaruh oleh penyakit ini. Nilai-nilai dalam kolom ini biasanya berkisar dari 0 hingga 2, di mana nilai yang lebih tinggi menunjukkan kemungkinan genetik yang lebih besar terkait dengan diabetes.



1. Age

Dalam konteks data penyakit diabetes, "age" (usia) adalah salah satu atribut atau kolom yang menggambarkan usia individu dalam dataset tersebut. Usia adalah informasi penting dalam analisis data diabetes karena usia seseorang dapat menjadi faktor risiko atau faktor prediktor dalam pengembangan penyakit diabetes.



**Normalisasi Data**

Normalisasi data adalah proses pengolahan data dalam analisis statistik dan ilmu data yang bertujuan untuk mengubah nilai-nilai dalam dataset agar memiliki rentang atau skala yang seragam. Tujuan normalisasi adalah untuk menghilangkan perbedaan dalam skala dan besaran antar variabel, sehingga memudahkan perbandingan dan analisis data. Normalisasi data dapat menjadi penting dalam beberapa konteks, seperti dalam pembuatan model pembelajaran mesin, analisis data statistik, dan pengolahan data. Berikut adalah beberapa alasan mengapa normalisasi data diperlukan:

**Perbandingan yang Adil**: Normalisasi memungkinkan perbandingan yang adil antar variabel. Tanpa normalisasi, variabel dengan skala besar dapat mendominasi perhitungan atau analisis, meskipun mereka mungkin tidak lebih penting secara substansial.

**Stabilitas Algoritma**: Dalam pembuatan model pembelajaran mesin, normalisasi dapat membantu algoritma konvergensi lebih cepat dan meningkatkan kinerja mereka. Algoritma yang peka terhadap skala variabel mungkin memiliki kesulitan mengatasi data yang tidak dinormalisasi.

**Visualisasi yang Lebih Baik**: Dalam visualisasi data, normalisasi dapat membantu grafik atau plot menjadi lebih informatif. Skala yang seragam membuatnya lebih mudah bagi pembaca untuk menginterpretasikan grafik.

**Mengurangi Redundansi**: Dalam beberapa kasus, normalisasi dapat membantu mengurangi dimensi atau fitur dalam data. Ini dapat mengurangi redundansi informasi dan memperbaiki efisiensi analisis.

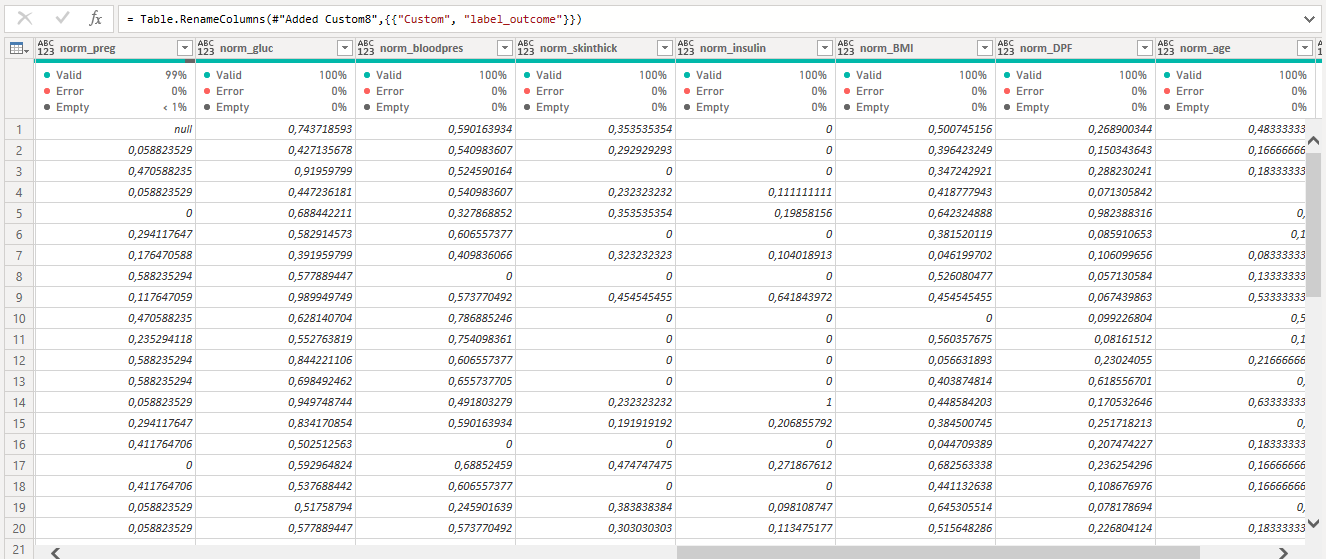
Terdapat beberapa teknik normalisasi yang umum digunakan, termasuk:

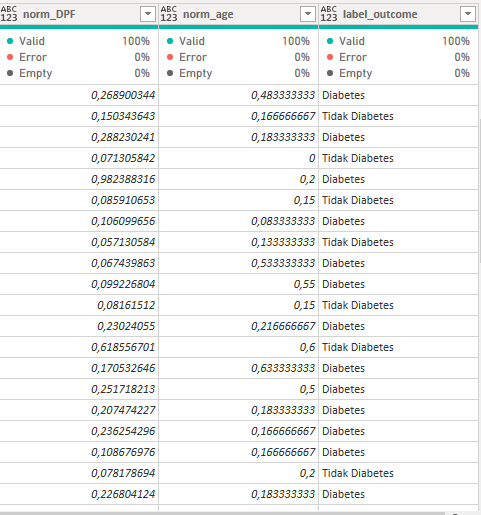
**Min-Max Scaling**: Mengubah nilai-nilai dalam dataset ke dalam rentang antara 0 dan 1 dengan rumus (x - min) / (max - min), di mana x adalah nilai asli, min adalah nilai minimum dalam dataset, dan max adalah nilai maksimum dalam dataset.

**Z-Score Standardization:** Mengubah nilai-nilai dalam dataset ke distribusi normal standar dengan rumus (x - mean) / std, di mana x adalah nilai asli, mean adalah rata-rata, dan std adalah deviasi standar dalam dataset.

**Robust Scaling:** Normalisasi yang tahan terhadap outlier dengan menggunakan median dan kuartil dalam perhitungan.

Pilihan teknik normalisasi tergantung pada karakteristik dataset dan tujuan analisis kita. Tujuan utama normalisasi adalah membuat data lebih mudah diinterpretasikan dan digunakan dalam analisis data. Pada dataset diabetes ini kita menggunakan Teknik Min-Max Scaling untuk normalisasi data.

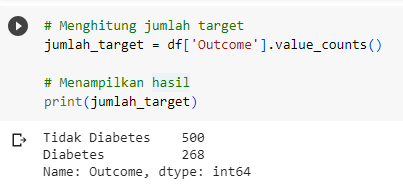




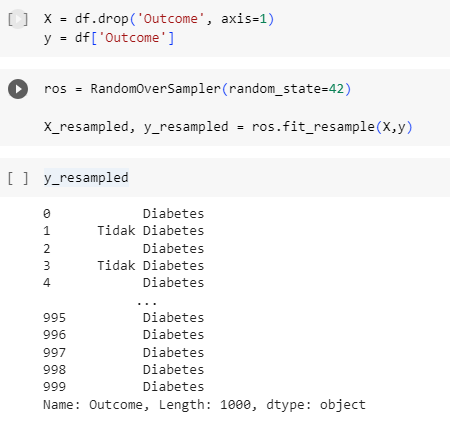
**Balancing Data**

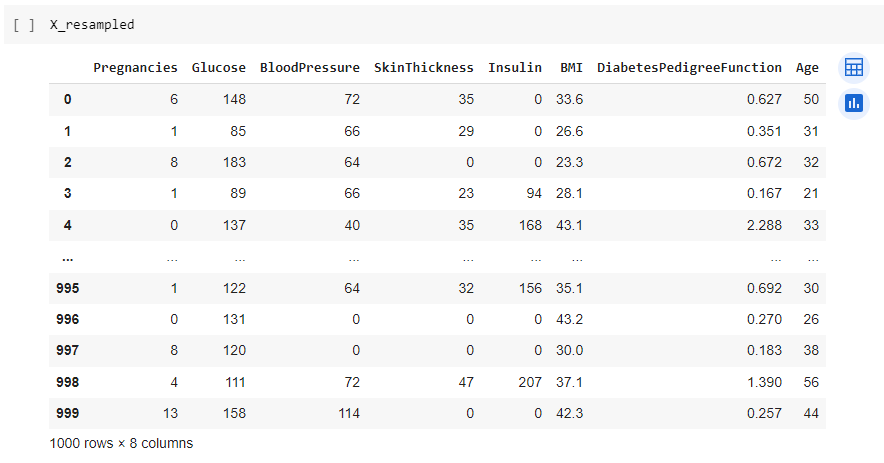
Balancing data adalah proses untuk menciptakan keseimbangan atau distribusi yang setara antara berbagai kategori atau kelas dalam sebuah dataset. Dalam konteks analisis data dan pembelajaran mesin (machine learning), balancing data diperlukan ketika terdapat ketidakseimbangan yang signifikan antara kelas-kelas atau kategori-kategori target dalam dataset.

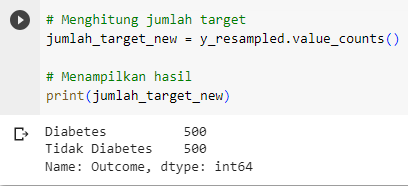
Biasanya, masalah ketidakseimbangan terjadi ketika satu atau beberapa kelas memiliki jumlah sampel yang jauh lebih banyak atau lebih sedikit dibandingkan dengan kelas lainnya. Ini dapat mengakibatkan model atau algoritma pembelajaran mesin menjadi tidak efektif dalam mengidentifikasi dan memprediksi kelas yang kurang umum atau minoritas. Untuk balancing data pada dataset diabetes ini kita menggunakan python dengan Teknik RandomOverSampler.



Berikut gambar diatas adalah hasil dataset diabetes jumlah target yang tidak seimbang maka butuh diseimbangkan.







Berikut adalah hasil dari balancing data. Dimana data yang awalnya tidak seimbang yakni dengan pasien yang terindikasi diabetes ada 500 sedangkan yang terindikasi tidak diabetes 268. Maka setelah dilakukan proses balancing data maka data tersebut menjadi balance seimbang yakni menjadi pasien yang terkena diabetes 500 sedangkan yang tidak diabetes juga 500.