***Diabetes Production Dataset***

*Diabetes production dataset* merupakan kumpulan data medis dan demografi dari pasien, beserta status diabetes mereka (positif atau negatif). Data tersebut mencakup fitur-fitur seperti usia, jenis kelamin, indeks massa tubuh (IMT), hipertensi, penyakit jantung, riwayat merokok, kadar HbA1c, dan kadar glukosa darah. Kumpulan data ini dapat digunakan untuk membangun model pembelajaran mesin guna memprediksi diabetes pada pasien berdasarkan riwayat medis dan informasi demografi mereka. Hal ini dapat berguna bagi para profesional perawatan kesehatan dalam mengidentifikasi pasien yang mungkin berisiko terkena diabetes dan dalam mengembangkan rencana perawatan yang dipersonalisasi. Selain itu, kumpulan data tersebut dapat digunakan oleh para peneliti untuk mengeksplorasi hubungan antara berbagai faktor medis dan demografi serta kemungkinan terkena diabetes.

File diabetes\_prediction\_dataset.csv berisi data medis dan demografi pasien beserta status diabetes mereka, baik positif maupun negatif. File tersebut terdiri dari berbagai fitur seperti usia, jenis kelamin, indeks massa tubuh (IMT), hipertensi, penyakit jantung, riwayat merokok, kadar HbA1c, dan kadar glukosa darah. Kumpulan data tersebut dapat digunakan untuk membangun model pembelajaran mesin yang dapat memprediksi kemungkinan diabetes pada pasien berdasarkan riwayat medis dan detail demografi mereka.

*Kolom Variabel pada Dataset*

* **Jenis kelamin** mengacu pada jenis kelamin biologis individu, yang dapat memengaruhi kerentanan mereka terhadap diabetes. Ada tiga kategori di dalamnya: pria, wanita, dan lainnya.
* **Usia** merupakan faktor penting karena diabetes lebih sering didiagnosis pada orang dewasa yang lebih tua. Usia berkisar antara 0-80 dalam kumpulan data kami.
* **Hipertensi** adalah kondisi medis di mana tekanan darah di arteri terus meningkat. Hipertensi memiliki nilai 0 atau 1, di mana 0 menunjukkan mereka tidak memiliki hipertensi dan untuk 1 berarti mereka memiliki hipertensi.
* **Penyakit jantung** adalah kondisi medis lain yang dikaitkan dengan peningkatan risiko terkena diabetes. Memiliki nilai 0 atau 1, di mana 0 menunjukkan mereka tidak memiliki penyakit jantung dan untuk 1 berarti mereka memiliki penyakit jantung.
* **Riwayat merokok** juga dianggap sebagai faktor risiko diabetes dan dapat memperburuk komplikasi yang terkait dengan diabetes. Dalam kumpulan data kami, kami memiliki 5 kategori yaitu tidak saat ini, sebelumnya, Tidak Ada Info, saat ini, tidak pernah, dan pernah.
* **BMI (Indeks Massa Tubuh)** adalah ukuran lemak tubuh berdasarkan berat dan tinggi badan. Nilai BMI yang lebih tinggi dikaitkan dengan risiko diabetes yang lebih tinggi. Kisaran BMI dalam kumpulan data adalah dari 10,16 hingga 71,55. BMI kurang dari 18,5 berarti berat badan kurang, 18,5-24,9 berarti normal, 25-29,9 berarti kelebihan berat badan, dan 30 atau lebih berarti obesitas.
* **Kadar HbA1c (Hemoglobin A1c)** adalah ukuran kadar gula darah rata-rata seseorang selama 2-3 bulan terakhir. Kadar yang lebih tinggi menunjukkan risiko lebih besar terkena diabetes. Umumnya lebih dari 6,5% Kadar HbA1c menunjukkan diabetes.
* **Kadar glukosa darah** mengacu pada jumlah glukosa dalam aliran darah pada waktu tertentu. Kadar glukosa darah yang tinggi merupakan indikator utama diabetes.
* **Diabetes** adalah variabel target yang diprediksi, dengan nilai 1 menunjukkan adanya diabetes dan 0 menunjukkan tidak adanya diabetes.

Pada eksplorasi data dua dimensi, terkhusus pada statistika deskriptif atau analisis deskriptif, dengan menggunakan scatter plot, hanya bisa menggunakan data numerik. Begitu juga dengan menghitung nilai korelasinya.

Maka, pada eksplorasi data dua dimensi kali ini, saya memilih 3 pasangan variabel, age dan BMI, age dan blood glucose level, BMI dan HbA1c level.

***Age and* *BMI***

A graph with numbers and dots

Description automatically generated with medium confidence

Nilai Korelasi: **0.1422869**

Berdasarkan *scatter plot* di atas, dapat dilihat bahwa plot x dan y tampak naik, dan posisi plotnya tampak sangat renggang, sangat berjarak. Maka, dapat diketahui bahwa terdapat kecenderungan korelasi positif, tetapi sangat kecil.

Nilai koefesien korelasi bernilai positif, maka kedua variabel mempunyai hubungan searah. Namun, hubungannya lemah. Berdasarkan nilai koefisiennya, nilai korelasi antara *age* dan *blood glucose level* bernilai lemah positif.

Kesimpulan: *Age* dan *BMI* memiliki hubungan yang positif, tetapi nilainya sangat lemah.

***Age and* *Blood Glucose Level***

A graph with numbers and dots

Description automatically generated

Nilai Korelasi: **0.05513575**

Berdasarkan *scatter plot* di atas, dapat dilihat bahwa plot x dan y tampak naik, dan posisi plotnya tampak sangat renggang, sangat berjarak. Maka, dapat diketahui bahwa terdapat kecenderungan korelasi positif, tetapi sangat kecil.

Nilai koefesien korelasi bernilai positif, maka kedua variabel mempunyai hubungan searah. Namun, hubungannya sangat lemah. Berdasarkan nilai koefisiennya, nilai korelasi antara *age* dan *blood glucose level* bernilai sangat lemah positif.

Kesimpulan: *Age* dan *Blood Glucose Level* memiliki hubungan yang positif, tetapi nilainya sangat lemah.

***BMI and HbA1c Level***

A graph of data

Description automatically generated

Nilai Korelasi: **0.1864351**

Berdasarkan *scatter plot* di atas, dapat dilihat bahwa plot x dan y tampak naik, dan posisi plotnya tampak sangat renggang, sangat berjarak. Maka, dapat diketahui bahwa terdapat kecenderungan korelasi positif, tetapi sangat kecil.

Nilai koefesien korelasi bernilai positif, maka kedua variabel mempunyai hubungan searah. Namun, hubungannya lemah. Berdasarkan nilai koefisiennya, nilai korelasi antara *BMI* dan *HbA1c level* bernilai lemah positif.

Kesimpulan: *BMI* dan *HbA1c level* memiliki hubungan yang positif, tetapi nilainya sangat lemah.

*Dokumentasi RStudio*

A screenshot of a computer

Description automatically generated