PROYEK SAINS DATA

**“ ANALISIS DATA DIABETES MENGGUNAKAN POWER BI “**

Dosen Pengampuh:

*Eka Mala Sari Rochman, S.Kom M.Kom*



**Oleh:**

Walid Rijal Awali (210411100143)

PROGRAM STUDI TEKNIK INFORMATIKA

FAKULTAS TEKNIK

UNIVERSITAS TRUNOJOYO MADURA

1. **Latar Belakang**

Ketidakmampuan tubuh untuk memanfaatkan insulin atau kekurangan hormon insulin menyebabkan kadar gula di dalam darah tidak dapat terkendali, yang menyebabkan diabetes melitus. Penyakit diabetes mellitus sekarang menyerang disemua kalangan usia, baik muda maupun lansia. Lebih dari 1,2 juta anak di seluruh dunia menderita penyakit diabetes mellitus. Diabetes mellitus tetap menjadi salah satu penyakit paling mematikan di dunia. Dunia mengalami 6,7 juta kematian akibat diabetes mellitus pada tahun 2021.

World Health Organization menyatakan bahwa jumlah orang yang menderita diabetes melitus terus meningkat hingga empat juta kali lebih besar dari tahun sebelumnya. Satu juta orang mati setiap tahun akibat diabetes melitus. Penyakit diabetes melitus dapat disembuhkan jika penderita menyadari dan mengikuti pengobatan sejak awal sebelum menjadi lebih parah. Oleh karena itu, masyarakat membutuhkan informasi tentang faktor-faktor yang menyebabkan diabetes melitus, dan juga gejala gejala yang terjadi.

1. **Analisis Data**

Data yang digunakan pada analisis ini merupakan dataset diabetes dengan record data sebanyak 100000, yang diperoleh dari halaman url berikut : <https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset?resource=download> .

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

Berikut merupakan hasil analisis dari fitur atau kolom yang terdapat pada dataset di atas antara lain:

1. Gender

Gender mengacu pada jenis kelamin biologis seseorang, yang dapat berdampak pada kerentanan mereka terhadap diabetes. Ada tiga kategori di dalamnya pria, wanita dan lainnya.

1. Age

Usia merupakan faktor penting karena diabetes lebih sering didiagnosis pada orang dewasa yang lebih tua. Usia berkisar antara 0-80 tahun dalam dataset kami.

1. Hypertension

Hipertensi adalah suatu kondisi medis di mana tekanan darah di arteri terus meningkat. Nilainya 0 atau 1 dimana 0 berarti tidak menderita hipertensi dan 1 berarti menderita hipertensi.

1. Heart Disease

Penyakit jantung adalah kondisi medis lain yang dikaitkan dengan peningkatan risiko terkena diabetes. Nilainya 0 atau 1 dimana 0 berarti tidak mengidap penyakit jantung dan angka 1 berarti mengidap penyakit jantung.

1. Smoking history

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 selamanya.

1. BMI

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 pada dataset adalah 10,16 hingga 71,55. BMI kurang dari 18,5 berarti kurus, 18,5-24,9 normal, 25-29,9 kelebihan berat badan, dan 30 atau lebih berarti obesitas.

1. HbA1c level

Kadar HbA1c (Hemoglobin A1c) merupakan ukuran rata-rata kadar gula darah seseorang selama 2-3 bulan terakhir. Kadar yang lebih tinggi menunjukkan risiko lebih besar terkena diabetes. Sebagian besar lebih dari 6,5% Tingkat HbA1c mengindikasikan diabetes.

1. Blood glucose level

Kadar glukosa darah mengacu pada jumlah glukosa dalam aliran darah pada waktu tertentu. Kadar glukosa darah yang tinggi merupakan indikator utama diabetes.

1. Diabetes

Diabetes merupakan variabel target yang diprediksi, dengan nilai 1 menunjukkan adanya diabetes dan 0 menunjukkan tidak adanya diabetes.

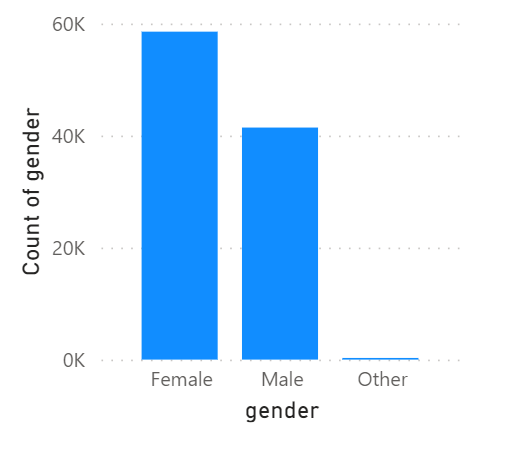
1. **Visualisasi data**

Pada gambar 2 kita dapat melihat visualisasi dari data diabetes yang akan dilakukan proses pengolahan data, terdapat beberapa visualisasi yang dilakukan terhadap beberapa parameter, contohnya visualisasi data jumlah penderita penyakit diabetes berdasarkan gender, jumlah penderita diabetes berdasarkan umur, dan visualisasi dari parameter lainnya.

Dari hasil visualisasi tersebut, memiliki hasil analisis yang dapat digunakan dalam melakukan pengoptimalan dalam pengolahan data, seperti terdapatnya missing value pada data tersebut, dan juga terjadinya ketidakseimbangan data yang memiliki pengaruh besar dalam hasil dari proses pengolahan atau penelitian yang akan dilakukan selanjutnya.

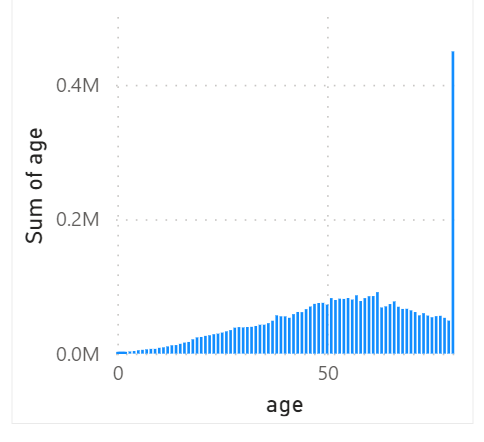
Berdasarkan hasil analisis yang telah dilakukan terdapat 9 fitur atau kolom pada data hepatitis yang akan digunakan berikut merupakan visualisasi data dari masing-masing fitur antara lain:

1. Gender



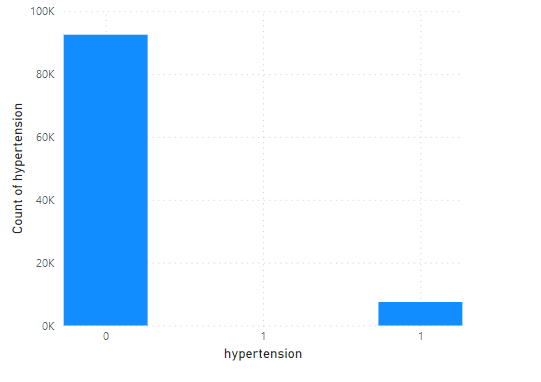
Gambar Grafik Data Gender

1. Age



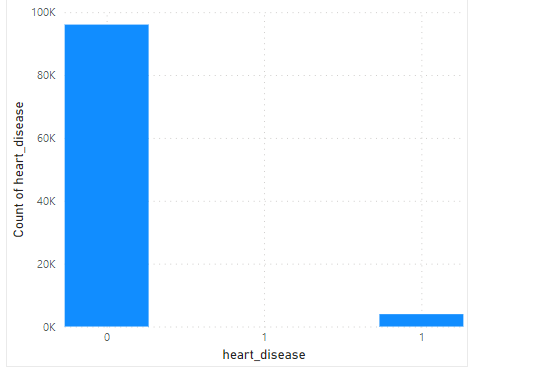
Gambar Grafik Data Umur

1. Hypertension



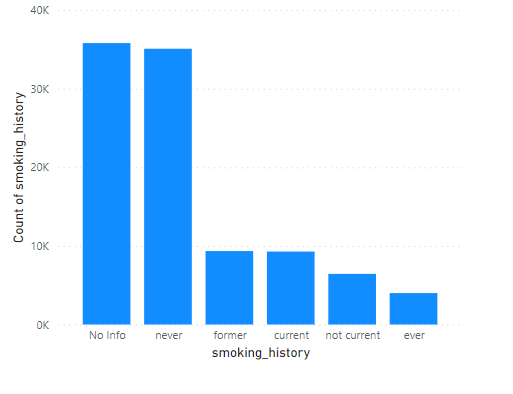
Gambar Grafik Data Hypertensi

1. Heart Disease



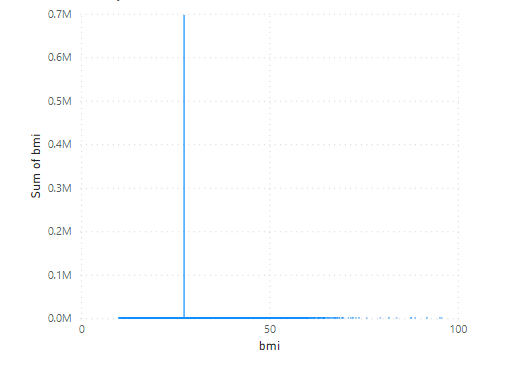
Gambar Grafik Data Heart Disease

1. Smoking history



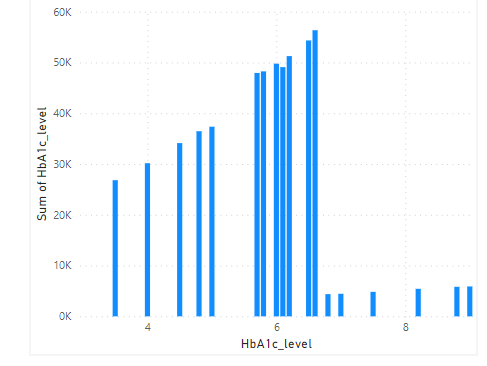
Gambar Grafik Data smoking History

1. BMI



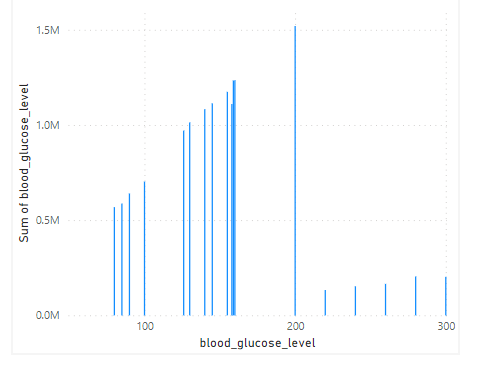
Gambar Grafik Data BMI

1. HbA1c level



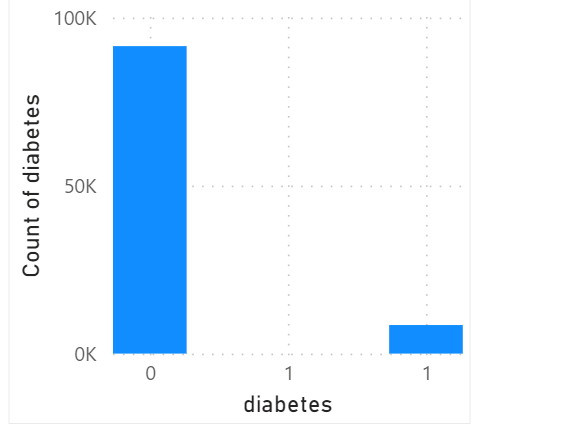
Gambar Grafik Data HbA1c level

1. Blood glucose level



Gambar Grafik Data Blood glucose level

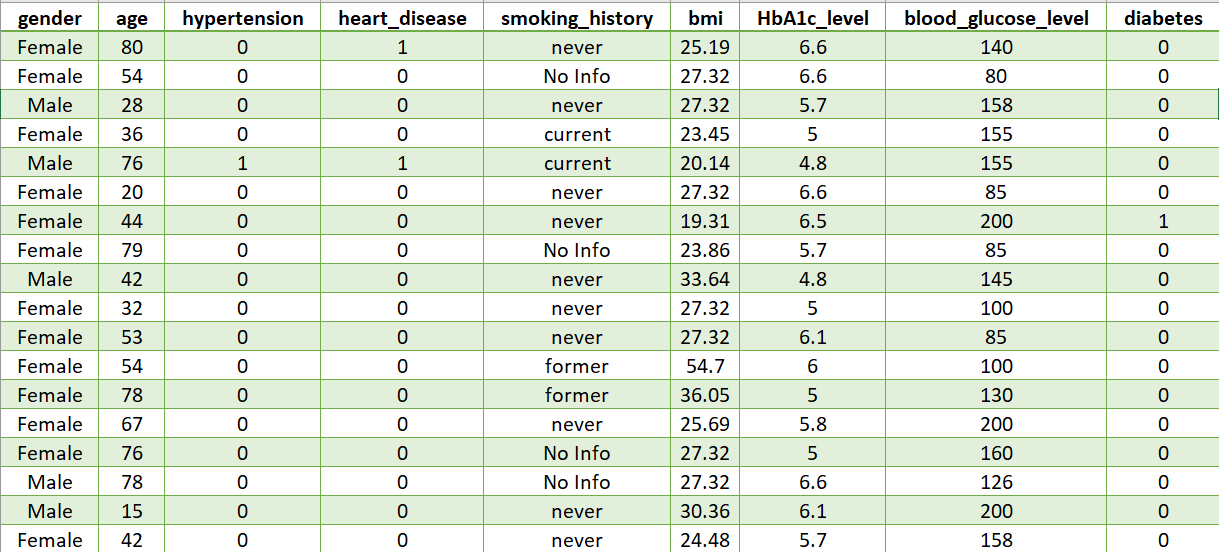
1. Diabetes



Gambar Grafik Data Diabetes

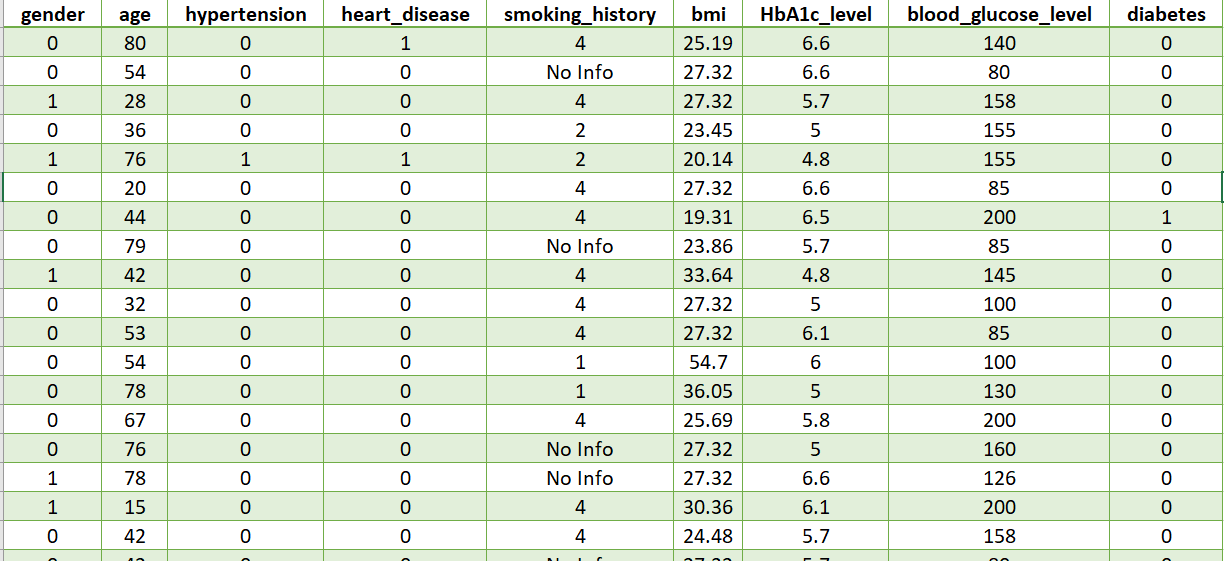
1. **Prepocessing Data**
2. Data Transformation

Data diubah atau digabung ke dalam format yang sesuai untuk diproses dalam data mining. Beberapa metode data mining membutuhkan format data yang khusus sebelum bisa diaplikasikan. Sebagai contoh beberapa metode standar seperti analisis asosiasi dan clustering hanya bisa menerima input data kategorikal. Karenanya data berupa angka numerik 10 yang berlanjut perlu dibagi-bagi menjadi beberapa interval. Proses ini sering disebut transformasi data. Pada dataset yang digunakan terdapat beberapa data kategorikal yang berupa teks agar dapat tersebut dapat dibaca oleh komputer maka kita ganti data kategorikal tersebut dengan metode biner yaitu 0 dan 1. Berikut merupakan gambar tabel yang belum dilakukan data transformation:



Gambar 10 Data Diabetes Sebelum Transformasi

Terdapat data kategorik pada kolom gender dan smoking history, pada gender nilai female akan diganti dengan 0 dan male diganti dengan 1 sementara pada smoking history nilai not current akan diganti 0, nilai former diganti dengan 1, nilai current diganti dengan nilai 2, nilai ever diganti dengn nilai 3 dan nilai never diganti 4. Berikut merupakan gambar tabel setelah dilakukan data transformation:



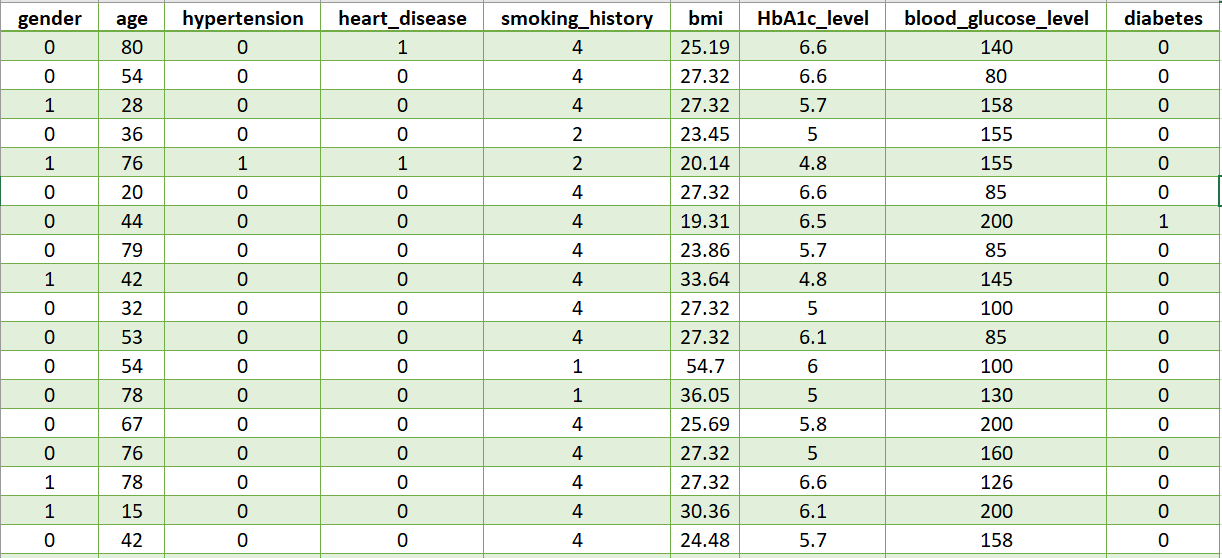
Gambar 11 Data Diabetes Sesudah Transformasi

1. Missing Value

Missing value merupakan istilah yang digunakan dalam analisis data untuk merujuk kepada nilai yang hilang atau tidak ada dalam suatu dataset. Missing value sering terjadi dalam metode klasifikasi dikarenakan informasi tentang obyek tidak diberikan, sulit dicari atau memang informasi tersebut tidak ada. Hal ini menyebabkan menurunnya keakuratan dan kualitas data pada saat data diolah. Data yang digunakan terdapat beberapa *missing value* yang harus diisikan berdasarkan kriteria data yang ada. Berikut merupakan tabel kolom dan jumlah nilai *missing value* yang ada:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| NO | Kolom | Jumlah Nilai Kosong | Tipe data | Mode |
| 1 | Smoking history | 35816 | kategorikal | 4 |

Pada data yang digunakan terdapat *missing value* pada kolom smoking history yang termasuk ke dalam kategorikal, maka data yang kosong tersebut akan diisi dengan nilai mode (nilai yang sering muncul) untuk mengisi baris yang terdapat *missing value* tersebut seperti pada rekap tabel diatas. Berikut merupakan gambar tabel setelah dilakukan proses pengisian data pada *missing value*:



Gambar Data Diabetes Setelah Pengisian Missing Value

1. Normalisasi Data

Min-Max Normalization Normalization atau normalisasi data merupakan bagian dari data transformasi, yaitu teknik mengubah data menjadi nilai yang lebih mudah dipahami (Patel, 2011). Tujuan normalisasi data adalah mendapatkan bobot yang sama dari semua atribut data dan tidak bervariasi atau hasil dari pembobotan tersebut tidak terdapat atribut yang lebih dominan atau dianggap lebih utama dari pada yang lain (Mustakim dkk, 2015).

Setiap nilai dalam data dikurangkan dengan nilai paling kecil dan dibagi dengan nilai paling besar kurang nilai paling kecil, sehingga skala rentang nilai yang didapatkan berada pada 0,0 hingga 1,0. Min-max melakukan transformasi linear pada data, menggunakan nilai minimum dan nilai maksimum. Normalisasi min-max mempertahankan hubungan antar nilai data asli (Jain, 2011 dalam Petra, 2015). Dengan demikian Min-Max Normalisasi secara matematis dapat dituliskan dalam persamaan berikut:

Keterangan:

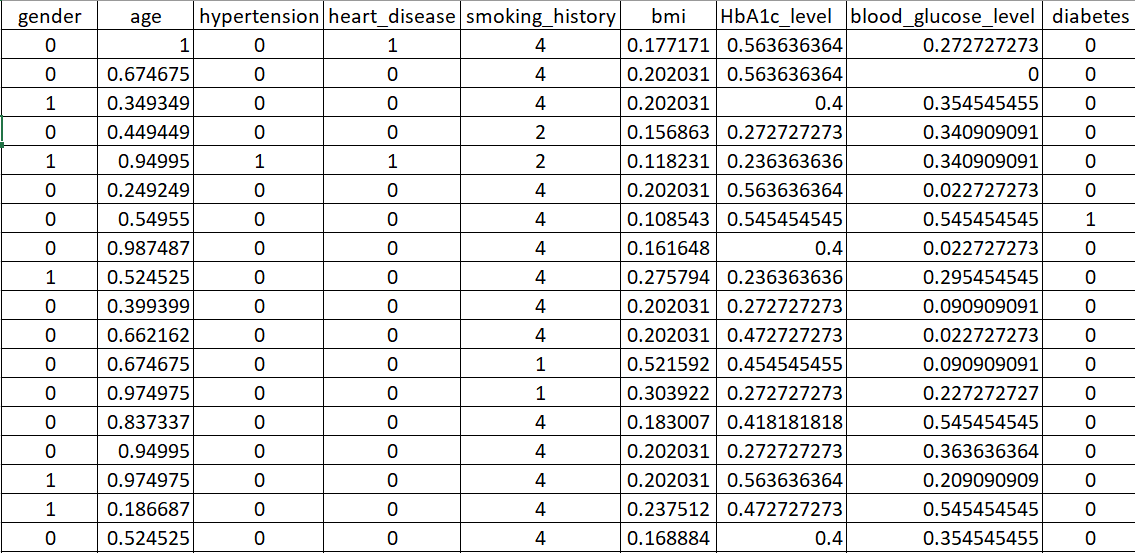
*Xnew* : nilai baru yang akan ditambahkan

*Xold* : nilai asli pada kolom.

*Xmin* : nilai paling rendah pada kolom.

*Xmax* : nilai paling tinggi pada kolom.

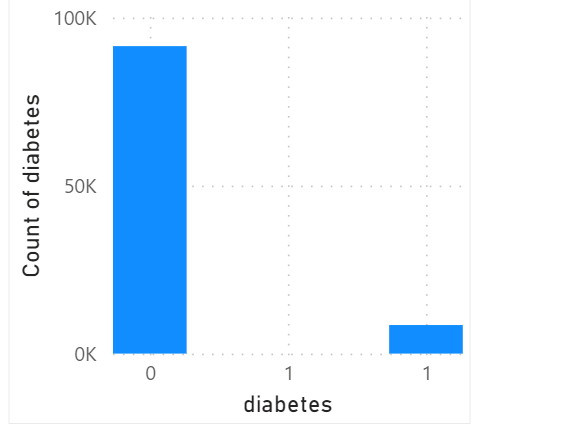
Berikut merupakan hasil perhitungan normalisasi pada setiap kolom data diabetes menggunakan *min max scaller:*



Gambar 13 Data Diabetes Setelah Proses Normalisasi

1. Balancing Data

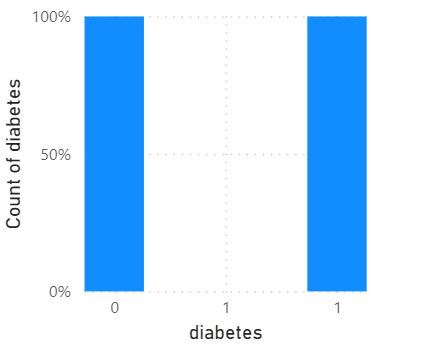
Imbalance class adalah kondisi distribusi antar kelas yang tidak seimbang pada suatu dataset, dimana salah satu kelasnya memiliki jumlah data yang sangat besar (kelas mayoritas) dibanding kelas lainnya (kelas minoritas). Perbedaan jumlah data yang besar antar kelas dapat mengakibatkan model klasifikasi sering tidak dapat memprediksikan kelas minoritas dengan tepat sehingga banyak data tes yang seharusnya berada pada kelas minoritas diprediksikan salah oleh model klasifikasi. Dataset yang digunakan pada analisis kali ini belum *balance* dikarenakan jumlah data diabetes yaitu 0 dan 1 tidak seimbang seperti pada gambar di bawah ini:



Gambar 14 Grafik Sebaran Data Diabetes Sebelum Balancing

Untuk mengatasi permasalahan imbalance class, salah satu metode yang digunakan adalah sampling. Metode sampling melakukan modifikasi terhadap distribusi data antar kelas mayoritas dan kelas minoritas pada dataset training untuk menyeimbangkan jumlah data tiap kelas. Salah satu metode sampling yang sering digunakan adalah *Syntetic Minority Over-sampling Technique* (SMOTE) [19].

SMOTE adalah metode over-sampling dimana data pada kelas minoritas diperbanyak dengan menggunakan data sintetik yang berasal dari replikasi data pada kelas minoritas. Berikut merupakan sebaran data setelah dilakukan balancing:



Gambar 15 Grafik Sebaran Data Diabetes Sesudah Balancing

REFERENSI :

N. A. Sari, “Sistem Pakar Mendiagnosa Penyakit Demam Berdarah Menggunakan Metode Certainty Factor,” *Encycl. Oper. Res. Manag. Sci.*, Pp. 160–160, 2013, Doi: 10.1007/978-1-4419-1153-7\_200031.

N. M. Putry, “Komparasi Algoritma Knn Dan Naïve Bayes Untuk Klasifikasi Diagnosis Penyakit Diabetes Mellitus,” *Evolusi  J. Sains Dan Manaj.*, Vol. 10, No. 1, 2022, Doi: 10.31294/Evolusi.V10i1.12514.

Asroni, B. M. Respati, And S. Riyadi, “Penerapan Algoritma C4.5 Untuk Klasifikasi Jenis Pekerjaan Alumni Di Universitas Muhammadiyah Yogyakarta,” J. Ilm. Fak. Tek. Univ. Muhammadiyah, Vol. 21, No. 2, Pp. 158–165, 2018.

M. R. Ritonga, Solikhun, M. R. Lubis, And A. P. Windarto, “Sistem Pakar Diagnosa Gejala Awal Penyakit Akibat Virus Pada Anak Berbasis Mobile Dengan Forward Chaining,”

A. Najib, D. Nurcahyono, And R. P. P. Setiawan, “Klasifikasi Diagnosa Penyakit Diabetes Mellitus (Dm) Menggunakan Algoritma C4.4,” J. Sains Terap. Teknol. Inf., Vol. 11, No. 2, 2019

Alasadi, S. A., & Bhaya, W. S. (2017). Review Of Data Preprocessing Techniques In Data Mining. Journal Of Engineering And Applied Sciences, 12(16), 4102-4107.

Ilmudatapy. 3 Metode Normalisasi Data (Feature Scaling) Di Python. Diakses Pada 14 September 2023 Dari <Https://Ilmudatapy.Com/Metode-Normalisasi-Data/>

A. A. D. Sulistyawati and M. Sadikin, “Penerapan Algoritma K-Medoids Untuk Menentukan Segmentasi Pelanggan,” *Sistemasi*, vol. 10, no. 3, p. 516, 2021, doi: 10.32520/stmsi.v10i3.1332.

N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002

G. Nguyen, A. Bouzerdoum, and S. L. Phung, “Learning pattern classification tasks with imbalanced data sets,” Fac. Informatics - Pap., Jan. 2009

E. Sutoyo and M. A. Fadlurrahman, “Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Television Advertisement Performance Rating Menggunakan Artificial Neural Network,” J. Edukasi dan Penelit. Inform., vol. 6, no. 3, p. 379, 2020, doi: 10.26418/jp.v6i3.42896.