

Perbandingan Metode Naïve Bayes dan KNN (K-Nearest Neighbor) dalam Klasifikasi Penyakit Diabetes

Tugas Akhir

diajukan untuk memenuhi salah satu syarat

memperoleh gelar sarjana

dari Program Studi Informatika

Fakultas Informatika

Universitas Telkom

1301162757

Iqmal Lendra Faisal Amien



Program Studi Sarjana Informatika

Fakultas Informatika

Universitas Telkom

Bandung

2022

LEMBAR PENGESAHAN

Perbandingan Metode Naïve Bayes dan KNN (K-Nearest Neighbor) dalam Klasifikasi Penyakit Diabetes

Comparison of Naïve Bayes and KNN (K-Nearest Neighbor) Methods in Classification of Diabetes

NIM : 1301162757

Iqmal Lendra Faisal Amien

Tugas akhir ini telah diterima dan disahkan untuk memenuhi sebagian syarat memperoleh gelar pada Program Studi Sarjana Informatika


Fakultas Informatika

Universitas Telkom


Bandung, 30 Agustus 2022

Menyetujui

Pembimbing I,


Widi Astuti, S.T., M.Kom.
00740046

Pembimbing II,


Dr. Kemas Muslim Lhaksamana,
S.T., M.ISD.
13820075

Ketua Program Studi
Sarjana Informatika,


Dr. Erwin Budi Setiawan, S.Si., M. T.
NIP: 00760045

LEMBAR PERNYATAAN

Dengan ini saya, Iqmal Lendra Faisal Amien, menyatakan sesungguhnya bahwa Tugas Akhir saya dengan judul “ Perbandingan Metode Naïve Bayes dan KNN (K-Nearest Neighbor) dalam Klasifikasi Penyakit Diabetes ” beserta dengan seluruh isinya adalah merupakan hasil karya sendiri, dan saya tidak melakukan penjiplakan yang tidak sesuai dengan etika keilmuan yang berlaku dalam masyarakat keilmuan. Saya siap menanggung resiko/sanksi yang diberikan jika di kemudian hari ditemukan pelanggaran terhadap etika keilmuan dalam buku TA atau jika ada klaim dari pihak lain terhadap keaslian karya,

Bandung, 30 Agustus 2022

Yang Menyatakan



Iqmal Lendra Faisal Amien

Perbandingan Metode Naïve Bayes dan KNN (K-Nearest Neighbor) dalam Klasifikasi Penyakit Diabetes

Iqmal Lendra Faisal Amien¹, Widi Astuti², Kemas Muslim Lahksamana³

¹Fakultas Informatika, Universitas Telkom, Bandung

hyounomi@students.telkomuniversity.ac.id

widiastuti@telkomuniversity.ac.id

kemasmuslim@telkomuniversity.ac.id

Abstrak

Diabetes merupakan salah satu penyakit yang mematikan di dunia. Faktor penyebab dari penyakit diabetes salah satunya adalah pola makan yang tidak teratur. Asupan gula yang dikonsumsi berlebihan dengan kurangnya aktivitas fisik sampai mengalami obesitas, mampu menaikkan kadar gula dalam tubuh. Selain itu, faktor keturunan juga berpengaruh terhadap penyakit diabetes. Oleh karena itu, diperlukan deteksi penyakit diabetes. Naïve dan juga K-Nearest Neighbor (KNN) dapat digunakan dalam pengklasifikasian penyakit diabetes. Pada penelitian ini didapatkan nilai akurasi tertinggi dari penggunaan metode K-Nearest Neighbor dengan K=5 sebesar 90%, lalu nilai Akurasi dari metode Naïve Bayes didapatkan sebesar 80%.

Kata kunci : Naïve Bayes, KNN, Diabetes

Abstract

Diabetes is one of the deadliest diseases in the world. One of the causes of diabetes is an irregular diet. Excessive intake of excessive sugar with lack of physical activity to experience, can raise sugar levels in the body. In addition, heredity also affects diabetes. Therefore, a system is needed to detect diabetes. Naïve bayes and also K-Nearest Neighbor (KNN) can be used as a classification of diabetes. In the results of this study, the highest accuracy value was obtained from the use of the K-Nearest Neighbor method with K = 5 of 90%, then the accuracy value of the Naïve Bayes method was obtained by 80%.

Keywords: Naïve Bayes, KNN, Diabetes

1. Pendahuluan

Latar Belakang

Diabetes termasuk ke dalam jenis/kelompok penyakit metabolik (metabolic disorder) yaitu suatu jenis penyakit yang ditandai dengan kenaikan kadar gula dalam darah. Penyakit metabolik seperti diabetes adalah penyakit genetik yang biasanya disebabkan oleh penyakit keturunan, lalu penyebab lainnya bisa dari racun makanan dan infeksi. Penyakit mempengaruhi destruksi sel beta, diabetagonik, dan olahraga yang kurang. Biasanya penyakit ini menyebabkan komplikasi kronis yang terjadi pada mata, ginjal, syaraf, dan pembuluh darah[1].

Menurut laporan The International Diabetes Federation, kenaikan jumlah penduduk yang terjangkit diabetes dari tahun 2003 hingga 2025 akan terus meningkat. Indonesia sendiri terdapat pada peringkat ke-5 dengan jumlah penderita diabetes sebesar 12.9 juta jiwa dan diperkirakan akan terus naik hingga peringkat ke-3 dunia pada tahun 2025[20].

Terdapat banyak metode klasifikasi pada supervised learning pada machine learning, diantaranya adalah K-Nearest Neighbor, Naïve Bayes, Support Vector Machine, Neural Network, Random Forest Classifier, Ada Boost Classifier, serta Quadratic Discriminant Analysis[10].

Dalam penelitian ini penulis menggunakan metode KNN dan Naïve Bayes untuk proses klasifikasinya, metode pengklasifikasian menggunakan KNN dan juga Naïve Bayes merupakan proses metode yang akan dikembangkan kedepannya untuk pengklasifikasian penyakit diabetes. Metode KNN memiliki beberapa kelebihan yaitu, pelatihan sangat cepat, sederhana, efektif pada data pelatihan yang besar. Sedangkan kekurangan KNN sendiri adalah nilai k bias harus ditentukan, komputasi kompleks, keterbatasan memori, dan dapat tertipu oleh atribut yang tidak relevan[3]. adapun metode Naïve Bayes yaitu metode klasifikasi yang menggunakan teori Bayes Probabilitas untuk memprediksi kelas yang tidak diketahui. Pada Naïve Bayes dataset di anggap paling independent[4]. Evaluasi tingkat keberhasilan akan didasarkan pada nilai akurasi, baik menggunakan algoritma KNN maupun Naïve Bayes menggunakan data pada UCI Machine Learning Pima Indian Diabetes, sehingga hasil yang bisa didapat dari klasifikasi yang akan dibandingkan, dan bisa digunakan dalam kebutuhan medis maupun pengklasifikasian dalam penyakit diabetes.

Ada beberapa permasalahan lainnya dalam penelitian diantaranya missing data, missing data merupakan informasi yang tidak tersedia untuk sebuah obyek oleh karena itu akan dilakukan penanganan missing data oleh nilai mean dan median.

Topik dan Batasannya

Hal utama dalam penelitian ini yang pertama yaitu membandingkan hasil dari penggunaan auto machine learning dengan menggunakan metode KNN dan Naïve Bayes hasil dengan akurasi tertinggi yang akan direkomendasikan dalam penggunaan klasifikasi dengan jumlah data yang telah ada.

KNN baik diterapkan penggunaannya dalam jumlah data yang besar, metode ini juga cukup sederhana untuk dipahami dan mudah diimplementasikan, selain itu metode KNN baik digunakan untuk masalah klasifikasi. Proses data mining dilakukan untuk mempreprocessing data terlebih dahulu agar data bersih dan layak untuk digunakan.

Batasan penelitian yang dilakukan yaitu :

- Menggunakan metode klasifikasi KNN dan Naïve Bayes
- Melakukan Analisis hasil dari klasifikasi untuk akurasi tertinggi dan terendah
- Menggunakan Teknik Data Mining untuk melakukan Preprocessing pada data Pima Indian Diabetes
- Penangan Missing Value

Tujuan

Adapun beberapa tujuan dari penelitian ini diantaranya :

- Mendapatkan Hasil Optimum dari Klasifikasi KNN dan Naïve Bayes pada data uji dan data testing yang disajikan.
- Mendapatkan Perbandingan Hasil dari Klasifikasi KNN dan Naïve Bayes.
- Dapat dimanfaatkan untuk pendeteksian penyakit Diabetes.
- Menangani Missing Value

Organisasi Tulisan

Pada selanjutnya merupakan Studi Terkait yang menjelaskan studi apa saja yang dipaparkan didalam sini, Sistem yang dibangun tahapan dalam perancangan sistem, Evaluasi Analisis dari pengerjaan sistem, Kesimpulan Hasil dan juga saran yang diberikan.

2. Studi Terkait

Dalam beberapa penelitian sebelumnya, terdapat beberapa metode klasifikasi yang telah diimplementasikan. akurasi terbesar yang didapat yaitu ada pada metode Naive Bayes dengan nilai akurasi tertinggi 91,9%[2]. Adapun metode KNN dengan akurasi tertinggi dari K=3 dengan akurasi 86.61%[16]. Sedangkan 2 metode lainnya decision tree dan random forest, hanya mendapatkan akurasi sebesar 77,86% untuk random forest. Dan 77.21% untuk decision tree [5]. Oleh karena itu dalam pengklasifikasian penyakit diabetes penulis menggunakan metode KNN dan Naive Bayes untuk menemukan hasil paling optimum.

Pada dataset dari Pima Indian Diabetes diketahui ada beberapa missing value yang perlu ditangani menggunakan proses imputasi. Menurut Penelitian[17], berikut ini ada beberapa pilihan proses imputasi di antaranya menggunakan Mean, Median, NN, EM, Regression, dll. Berdasarkan penelitian yang dilakukan oleh Singh, S., & Prasad, J. dinyatakan bahwa, imputasi menggunakan mean merupakan salah satu metode termudah untuk mengatasi missing value[18]. Lalu pada penelitian berikutnya, menggunakan metode imputasi median yang dikombinasikan dengan penanganan outlier menggunakan median dan classifier tertentu mampu meningkatkan performansi sistem[19]. Oleh sebab itu, penanganan missing value pada penelitian ini akan menerapkan proses imputasi mean dan median, yang diharapkan dapat meningkatkan nilai akurasi pada proses klasifikasi yang diterapkan kedepannya.

Lebih jauh mengenai klasifikasi, berikut beberapa jurnal penelitian yang penulis cari :

Penelitian yang dilakukan oleh A. Sharmila Agnal, dan E. Saraswathi, Assistant Professor, Departement of Computer Science and Engineering, dari SRM Institute of Science and Technology, Chennai, India. Dengan judul "Analyzing Diabetic Data Using Naive-Bayes Classifier". Penelitian berikut ini melakukan analisis terhadap data diabetes dengan menggunakan metode Naive Bayes Classifier. Ditekankan pada penelitiannya dimana hasil yang didapatkan dengan akurasi 70% [1].

Penelitian dilakukan oleh peneliti Bagus Setya Rintyarna dari jurusan Teknik Informatika dengan judul "Klasifikasi Penyakit Diabetes Dengan Hidden Naïve Bayes". Menggunakan Hidden Naïve Bayes atau Naïve

Bayes Classifier yang sudah ditingkatkan akurasi. Dan didapatkan akurasi rata-rata yang dicapainya hingga 91,9% [2].

Penelitian dilakukan oleh Siti Mutro, Abidatul Izzah, Arrie Kurniawardhani, Mukhamad Masrur, dari Fakultas Teknik Unipdu, dengan judul Penelitian “Optimasi Teknik Klasifikasi Modified K-Nearest Neighbor Menggunakan Algoritma Genetika”. Pada penelitian ini dilakukan perubahan perhitungan dalam proses klasifikasi algoritma KNN dengan menggunakan algoritma Genetik. Agar nilai bias-k dalam algoritma KNN bisa di optimasi oleh algoritma GA yang dapat mengoptimalkan nilai bias-k. dan didapat akurasi rata-rata 80% [3].

Pada penelitian “Perbandingan Akurasi Klasifikasi Penyakit Diabetes Menggunakan Algoritma Adaboost random Forest dan Adaboost-decision Tree dengan imputasi median dan KNN”. Didapatkan hasil adaboost random forest tertinggi dengan akurasi 77.86%. Lalu Adaboost-Decision Tree dengan imputasi median dengan akurasi sebesar 77.21% [5].

Penelitian yang dilakukan oleh M. Syukri Mastafa, & I Wayan Simpen, dengan judul penelitian “Implementasi Algoritma K-Nearest Neighbor (KNN) Untuk Memprediksi Pasien Terkena Penyakit Diabetes Pada Puskesmas Manyampa Kabupaten Bulukumba.”. Penelitian yang dilakukan untuk melakukan pengujian terhadap kemungkinan seorang pasien baru pada puskesmas Manyampa kabupaten Bulukumba dapat terkena penyakit diabetes militus atau tidak. Dengan menggunakan klasifikasi K-Nearest Neighbor lalu didapat akurasi tertinggi dengan nilai akurasi sebesar 68.30% [12].

Pada penelitian sebelumnya dilakukan oleh Riri N,D., Heru W,H., dan Aji P,W., juga membandingkan metode KNN dan Naive Bayes dengan judul “Perbandingan Kinerja Metode Naive Bayes dan K-Neares Neighbor untuk klasifikasi Artikel Berbahasa Indonesia”. Disini dikatakan bahwa penggunaan metode Naive Bayes dipilih karena metode tersebut dapat menghasilkan akurasi yang maksimal dengan data latih yang sedikit. Sedangkan metode KNN dipilih karena metode tersebut tangguh terhadap data noise, hasil yang diberikan untuk metode Naive Bayes yaitu 70% sedangkan untuk KNN didapatkan akurasi cukup rendah dengan akurasi 40% [15].

Adapun penelitian yang dilakukan oleh Y C Tapidingan, D Paseru dengan judul penelitian “Comparative Analysis of Classification Methods of KNN and Naive Bayes to Determine Stress Level of Junior High School Students”. Didapatkan hasil tertinggi KNN dengan K = 3 mencapai 86.61%. Sedangkan untuk metode Naive Bayes didapatkan nilai mencapai 87.40% [16].

2.1 Diabetes

Penyakit diabetes merupakan penyakit metabolic (*metabolic disorder*) yang ditandai dengan kadar glukosa darah (gula darah). Diabetes merupakan penyakit silent killer, dikatakan seperti itu karena sering tidak disadari oleh penderitanya. Saat diketahui sudah terjadi berbagai macam komplikasi [1]. Diabetes dibagi menjadi beberapa tipe, diantaranya ada tipe 1 dan tipe 2 [6].

Diabetes tipe-1 biasa terdapat pada anak atau remaja. Faktor terkena diabetes tipe-1 biasanya infeksi virus atau reaksi auto-imun dimana rusaknya system kekebalan tubuh, yang merusak sel-sel penghasil insulin. Untuk bertahan hidup, insulin harus diberikan dari luar dengan cara disuntikan.

Dari keseluruhan penderita diabetes, jumlah penderita diabetes tipe-2 yang paling banyak, diperkirakan sekitar 90-99%. Penderita diabetes faktor utamanya dari cara hidup atau lifestyle yang kurang sehat dan tidak baik. Biasanya tipe ini mengenai orang dewasa namun diabetes ini juga dapat terjadi kepada remaja. Oleh karena itu istilah diabetes tipe-2 dianggap lebih cocok.

Diabetes tipe-2 dideteksi berkembang sangat lambat, dapat bertahun-tahun. Sehingga tanda-tandanya sering kali tidak jelas. Penderita biasanya memiliki Riwayat keturunan diabetes, lalu gejala lainnya biasanya dirasakan cepat Lelah, berat badan turun drastis walaupun banyak makan, dan kesemutan di tungkai. Bahkan penderita bisa saja tidak merasakan ada perubahan.

2.2 Data Mining

Data Mining merupakan proses penggunaan statistika, kecerdasan bauta, matematika, dan machine learning, untuk mengidentifikasi informasi yang bermanfaat [9]. Data mining didefinisikan sebagai proses dalam penemuan pola dalam sebuah data. Dari tugas yang dilakukan data mining dikelompokkan menjadi beberapa bagian yaitu deskripsi, estimasi, prediksi, klasifikasi, clustering dan asosiasi. Terdapat beberapa proses dalam data mining, diantaranya Preparation data pada Langkah ini, data dipilih, lalu di bersihkan, dan dilakuakn preprocessed data. Penggunaan algoritma data mining, dilakuakn untuk menggali data yang terintegrasi mempermudah menentukan identifikasi informasi yang bernilai. [14].

2.3 KNN (*K-Nearest Neighbour*)

Algoritma K-NN adalah algoritma ketetangga terdekat, yang dihitung dari nilai jarak pada pengujian data testing dengan data training, dari nilai terkecil ketetanggaan terdekat. Tujuan algoritma ini untuk pengklasifikasian objek baru berdasarkan atribut dan training samples. Pada metode ini nilai K pada K-Nearest Neighbor (K-NN) merupakan K-data terdekat dari data uji. K-Nearest Neighbor dapat menghasilkan klasifikasi yang baik jika menggunakan data dalam jumlah yang besar [12].

Adapun tahapan dari algoritma K-Nearest Neighbor (K-NN) dijelaskan sebagai berikut :

1. Persiapan data training dan testing
2. Menentukan nilai-k
3. Menghitung jarak data testing ke setiap data training.

Data training dihitung menggunakan rumus penghitung jarak Euclidean sebagai berikut :

$$\text{Euclidean Distance} = \sqrt{(x - x_{\text{terbaru}})^2 + (y - y_{\text{terbaru}})^2}$$

x = nilai dari atribut sebelumnya

x_{terbaru} = nilai dari atribut yang akan diklasifikasikan

y = nilai atribut kedua dari sebelumnya

y_{terbaru} = nilai dari atribut kedua yang baru

K-Nearest Neighbor juga memiliki beberapa kelebihan yaitu, data pelatihan cepat, dan efektif jika digunakan dengan data pelatihan yang besar. Adapun kekurangan pada metode ini yaitu nilai K-bias yang harus ditentukan sehingga hasil dari setiap K berbeda-beda. Untuk memperoleh hasil yang terbaik harus menggunakan semua atribut atau hanya menggunakan atribut yang sudah pasti saja. Pemilihan jumlah K ditetapkan oleh peneliti, pemilihan nilai K tentu mempengaruhi tingkat akurasi prediksi yang dikerjakan.

2.4 Naïve Bayes

Naïve bayes merupakan metode yang menggunakan data yang telah ada sebelumnya atau metode yang memprediksi dari kejadian di masa lampau. Kekurangan metode ini yaitu kurang mendukung atribut yang bernilai continuous (numerik). Secara singkat algoritma naïve bayes classification adalah pengklasifikasi kumpulan data statistika yang mana untuk memprediksi semua probabilitas tiap anggota suatu class. Dikarenakan data dalam penelitian kali ini berbentuk kontinu oleh karena akan digunakan metode Distribusi Gaussian yang biasa digunakan untuk data bertipe numerik. Berikut merupakan Rumus yang digunakan pada metode Naïve Bayes :

Distribusi Gaussian

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \exp \frac{-(x-\mu)^2}{2\sigma^2}$$

$$\mu = \frac{\sum_i^n x_i}{n}$$

$$\sigma = \sqrt{\frac{\sum_i^n (x_i - \mu)^2}{n - 1}}$$

g = Gaussian

μ = mean (nilai rata – rata)

σ = standar deviasi

Akurasi

Akurasi Merupakan tingkat kedekatan antara nilai prediksi dengan nilai actual.

$$AKURASI = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

Presisi

Presisi didefinisikan sebagai rasio item relevan yang dipilih terhadap permintaan tersebut.

$$PRESISI = \frac{TP}{TP + FP}$$

Recall

Recall, sebagai rasio dari item relevan yang dipilih terhadap total jumlah item relevan yang tersedia

$$RECALL = \frac{TP}{TP + FN}$$

F-Measure

F-Measure Adalah harmonic mean antara nilai presisi dan recall, F-Measure Biasa disebut F1-Score.

$$F - Measure = 2 \left(\frac{Presisi \times Recall}{Presisi + Recall} \right)$$

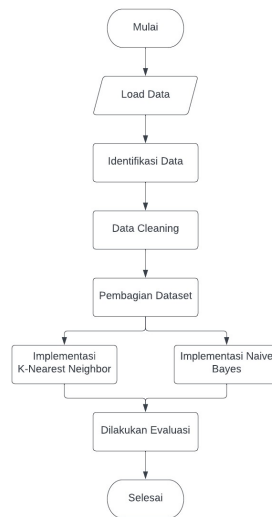
2.5 Missing Value Imputation

Missing Value Imputation merupakan proses penggantian data yang hilang dimana data bernilai 0 dan tidak sesuai dengan atribut yang ada sehingga harus dilakukan penggantian attribute namun harus dilihat dari dataset terlebih dahulu, ada beberapa alasan terjadinya missing value diantaranya :

1. Interview recording error yang terjadi karena kelalaian saat pengumpulan data
2. Respondent inability error, responden tidak mampu menjawab pertanyaan
3. Unwillingness Respondent error, dikarenakan responden tidak ingin memberikan jawaban misalnya usia, pekerjaan, alamat, dll.

3. Sistem yang Dibangun

Sistem yang dibangun untuk penelitian ini, dibentuk sesuai kebutuhan proses metode KNN dan Naïve Bayes dimana dimulai dari Mengumpulkan dataset, splitting atau membagi dataset, implementasi metode KNN dan Naïve Bayes dengan penerapan machine learning maka alur model yang dikerjakan pada berikut ini :



Gambar 1 . Flowchart sistem yang dibangun

1. Load Data

Melakukan input data dari UCI Machine Learning menggunakan dataset Pima Indian Diabetes, dengan jumlah 765 data Pasien dari Pima Indian Diabetes, terdapat beberapa attribute diantaranya, Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age, Outcome. Attribute diatas akan dibagi-bagi lagi menjadi jadi data test, dan data train.

2. Identifikasi Data

Dilakukan beberapa visualiasi, untuk dilakukan identifikasi data apakah data tersebut mempunyai missing value atau tidak, identifikasi dilakukan untuk setiap atribut yang ada pada data.

3. Data Cleaning

Pada Proses ini setelah dilakukan identifikasi data, terdapat atribut yang masih memiliki missing value, diantaranya Glucose, BloodPressure, SkinThickness, Insulin, dan BMI, maka missing value diganti dengan nilai median dan mean dari masing-masing attribute.

4. Pembagian dataset

Dilakukan pembagian dataset dari attribute yang disajikan menjadi 70% Training, dan 30% Testing

5. Penerapan Klasifikasi KNN dan Naïve Bayes

Dilakukan klasifikasi dengan penerapan machine learning pada data yang telah disediakan menggunakan dua metode KNN dan juga Naïve Bayes, dengan output Akurasi, Presisi, Recall, dan F1-Score

6. Dilakukan Evaluasi

Setelah mendapatkan hasil percobaan dari K-Nearest Neighbor dari K=1 sampai K=9 dengan masing-masing 5 percobaan. Adapun metode naïve bayes yang masing-masing dilakukan 5 percobaan, untuk setiap penggunaan missing value imputation mean dan median, maupun tidak diterapkan missing value imputation didapatkan hasil paling tinggi yaitu K=5 = 90% dari seluruh seluruh k yang paling optimum

Setelah mendapatkan hasil, dari percobaan metode KNN dan Naïve Bayes, dilakukan masing-masing 5 percobaan untuk setiap kasus, diantaranya terdapat 3 kasus yaitu, Tanpa penerapan missing value imputation, penerapan missing value imputation median, dan penerapan missing value imputation mean, pada percobaan di atas peneliti melakukan percobaan manual dengan melakukan running program berulang-ulang, sehingga ditemukan hasil akurasi tertinggi dari percobaan tersebut. Adapun pembagian data train untuk setiap

percobaan, data train dibagi menjadi 90%, 80%, dan 70%, lalu dilakukan perhitungan persentase missing value untuk setiap data train yang dibagi.

Kasus pertama yang dapat dilihat dari table 1, tidak menerapkan missing value imputation. Didapatkan hasil tertinggi untuk metode KNN, $K=1$ hingga $K=9$ terdapat pada $K=7$, yang dilakukan pada percobaan ke-3, dengan persentase data train 80% dan data testing 20% yang didapatkan hasil akurasi sebesar 75%. Sedangkan untuk metode Naïve Bayes sendiri didapatkan hasil akurasi tertinggi sebesar 77% dengan data train 80% dan data testing 20%. Pada proses ini terdapat missing value pada data training dengan persentase 48.70% dari 614 jumlah data train. dapat dilihat pada tabel 1.

Pada kasus kedua yang dapat dilihat pada tabel 2, yang menerapkan missing value imputation mean, didapatkan hasil tertinggi untuk metode KNN dengan $K=6$ dengan pembagian data train 80% dan data testing 20% dimana terdapat missing value pada data train yang sebelum diganti menjadi nilai mean, dengan persentase missing value sebesar 48.53% dari 614 jumlah data train, didapatkan hasil akurasi tertinggi sebesar 85%. Lalu metode Naïve Bayes mendapatkan hasil akurasi paling tinggi diantara percobaan lainnya didapatkan akurasi 80% dengan pembagian data train 70% dan data testing 30%, terdapat missing value pada data train dengan persentase 49.16% dari 537 data train.

Lalu pada kasus ketiga yang dapat dilihat pada tabel 3, menerapkan missing value imputation median, penggunaan metode KNN mendapatkan hasil akurasi paling tinggi dari seluruh percobaan, yaitu pada $K=5$ dengan data train 70% dan data testing 30% dengan persentase missing value pada data train sebesar 50.28%, dan didapatkan hasil akurasi 90%, sedangkan metode Naïve Bayes mendapatkan hasil akurasi tertinggi 78% dengan persentase data train 70% dan data testing 30%. Setelah setiap scenario atau kasus dilakukan percobaan maka dilihat confusion matrix untuk nilai akurasi tertinggi dan nilai akurasi paling rendah.

Jadi jumlah percobaan yang dilakukan pada penelitian kali ini dengan masing-masing kasus dilakukan running program sebanyak 5 kali, yang dimana terdapat metode K-NN dari $K=1$ sampai $K=9$ dan juga metode naïve bayes, sehingga total percobaan dilakukan sekitar 150 kali percobaan, dengan nilai akurasi tertinggi sebesar 90% untuk metode knn, dan 80% untuk metode naïve bayes.

4. Evaluasi

4.1 Hasil Pengujian dari K-Nearest Neighbour dan Naïve Bayes

Dilakukan beberapa percobaan dalam Penelitian ini yaitu kasus pertama tidak diterapkan proses missing value imputation pada data dan didapatkan hasil sebagai berikut :

Table 1. Analisis Data Train, Data Testing, hasil KNN, dan Naïve Bayes

No	Missing Value Imputation	Persentasi Data Train	Jumlah Data Train	Persentasi Missing Value Dari Data Train	Jumlah Missing Value Pada Data Train	Hasil K-Nearest Neighbor Tertinggi	Hasil Naïve Bayes Tertinggi
1	None	90%	691	50.07%	346	K3 = 72%	56%
2	None	80%	614	48.70%	299	K7 = 75%	77%
3	None	70%	537	48.97%	263	K9 = 74%	73%

Pada table 1 dapat dilihat bahwa hasil tertinggi untuk masing-masing metode tanpa menggunakan Missing Value Imputation, didapatkan hasil tertinggi dengan metode Naïve Bayes dengan skor akurasi 77% menggunakan 80% data train dan 20% data testing, sedangkan untuk K-Nearest Neighbor didapatkan hasil tertinggi dengan skor 75% dari K = 7 dengan 80% data train dan 20% data testing.

Lalu ada kasus kedua dengan diterapkan missing value imputation Mean pada dataset dan didapatkan hasil sebagai berikut :

Table 2. Analisis Data Train, Data Testing, hasil KNN, dan Naïve Bayes Menggunakan Missing Value Imputation Mean

No	Missing Value Imputation	Persentasi Data Train	Jumlah Data Train	Persentasi Missing Value Dari Data Train	Jumlah Missing Value Pada Data Train	Hasil K-Nearest Neighbor Tertinggi	Hasil Naïve Bayes Tertinggi
1	Mean	90%	691	48.62%	336	K3 = 83%	75%
2	Mean	80%	614	48.53%	298	K6 = 85%	79%
3	Mean	70%	537	49.16%	264	K4 = 82%	80%

Hasil tertinggi yang didapat Ketika diterapkan missing value imputation median sebesar 90% dari metode K-Nearest Neighbor dengan K = 5 dan jumlah data training sebesar 70% dan data testing sebesar 30%, sedangkan untuk metode Naïve Bayes masih unggul pada table 2 dengan akurasi 80%.

Lalu pada kasus ketiga diterapkan missing value imputation Median pada dataset dan didapatkan hasil sebagai berikut :

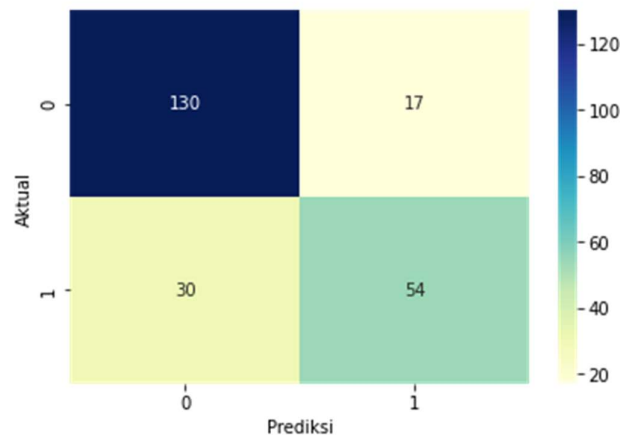
Table 3. Analisis Data Train, Data Testing, hasil KNN, dan Naïve Bayes Menggunakan Missing Value Imputation Median

No	Missing Value Imputation	Persentasi Data Train	Jumlah Data Train	Persentasi Missing Value Dari Data Train	Jumlah Missing Value Pada Data Train	Hasil K-Nearest Neighbor Tertinggi	Hasil Naïve Bayes Tertinggi
1	Median	90%	691	47.90%	331	K5 = 83%	68%
2	Median	80%	614	49.84%	306	K1 = 85%	75%
3	Median	70%	537	50.28%	270	K5 = 90%	78%

Dari hasil ketiga percobaan diatas didapatkan skor tertinggi dari K = 5 dengan jumlah data train 70% dan data testing 30%.

Adapun hasil pengujian kali ini terdapat Confussion Matrix KNN, dan Naïve Bayes, lalu ada hasil pengujian dari KNN, dan Naïve Bayes, lalu dari data tersebut ditentukan best case, dan worst casenya, best case mempunyai nilai akhir akurasi yang tinggi sedangkan worst case mempunyai nilai akurasi yang sangat rendah.

Confusion matrix



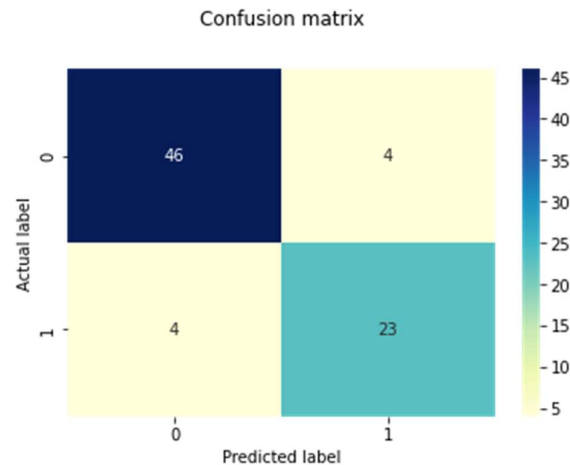
Gambar 2. Confusion Matrix Naïve Bayes Best Case

Pada Gambar 1 Merupakan Hasil Confusion Matrix dari Naïve Bayes best case, yang digunakan untuk menghitung hasil dari presisi, Recall, F1-Score, dan Akurasi dari table 1.

Table 4. Naïve Bayes Best Case

Status Pasien	Missing Value Imputation	Presisi	Recall	F1-Score	Akurasi
0	Mean	0.81	0.88	0.85	0.80
1		0.76	0.64	0.70	

Dari confusion matrix pada Gambar 1. Didapat data Presisi, Recall, F1-Score dan Akurasi pada Table 1, dari Gambar 1. Untuk data pasien sehat nilai confusion matrixnya menggunakan ini TP = 54, TN = 130, FP = 17, dan FN = 30. Dari hasil analisis diatas accuracy tertinggi pada metode naïve bayes di dapat 80% dengan imputation missing value menggunakan mean.

**Gambar 3. Confusion Matrix K-Nearest Neighbour (KNN) K = 7 Best Case**

Pada Gambar 2. Datas akan digunakan untuk menghitung presisi, recall, F1-Score, dan Akurasi pada Table 2.

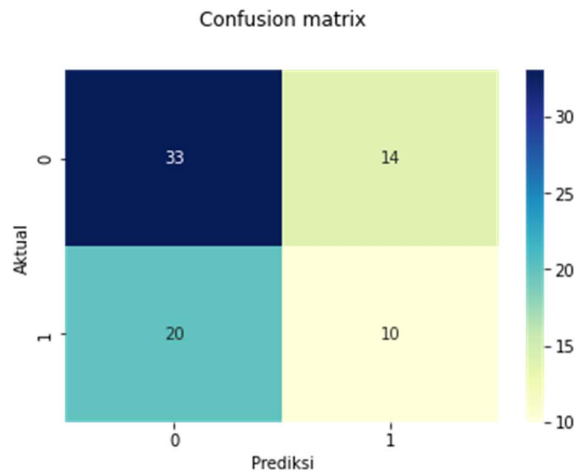
Table 5. K-Nearest Neighbour (KNN) K = 7 Best Case

Status Pasien	Missing Value Imputation	Presisi	Recall	F1-Score	Akurasi
0	Median	0.92	0.92	0.92	0.90
1		0.85	0.85	0.85	

Dari confusion matrix pada Gambar 2. data pasien sehat nilai confusion matrixnya menggunakan ini TP = 23, TN = 46, FP = 4, dan FN = 4. Untuk Kasus Terbaik (Best Case) berdasarkan data tersebut nilai akurasi dari K-Nearest Neighbour (KNN) K= 5 yang menggunakan missing value imputation Median.

sedangkan untuk Naïve Bayes dengan menggunakan missing value imputation Median dengan status pasien sehat(0) hasil presisi mencapai 80%, recall mencapai 88%, dan F1-Score mencapai 85%, lalu untuk status pasien Diabetes(1) presisi mencapai 76%, Recall 64%, F1-Score 70%, dengan total akurasi 80%.

Lalu untuk worstcase sebagai berikut :



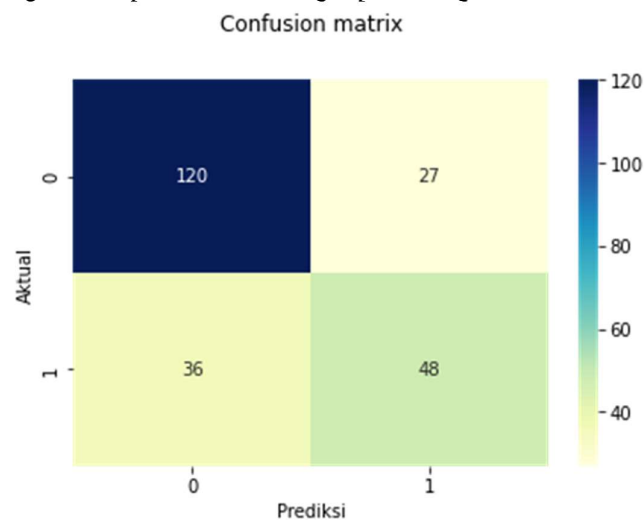
Gambar 4. Confusion Matrix Naïve Bayes Worst Case

Pada Gambar 3. Data confusion matrix akan digunakan untuk menghitung presisi, recall, f1-score, dan akurasi pada Table 3.

Table 6. Naïve Bayes Worst Case

Status Pasien	Missing Value Imputation	Presisi	Recall	F1-Score	Akurasi
0	Median	0.62	0.70	0.66	0.56
1		0.42	0.33	0.37	

Dari confusion matrix pada Gambar 3. data pasien sehat nilai confusion matrixnya menggunakan ini TP = 10, TN = 33, FP = 14, dan FN = 10. Didapatkan perbandingan data tanpa menggunakan missing value imputation dengan menggunakan missing value imputation mean dengan perbandingan 24%.



Gambar 5. Confusion Matrix K-Nearest Neighbour Worst Case

Pada Gambar 4. Data dari confusion matrix diatas akan digunakan pada Table 4.

Table 7. K-Nearest Neighbour (KNN) K = 1 Worst Case

Status Pasien	Missing Value Imputation	Presisi	Recall	F1-Score	Akurasi
0	Mean	0.77	0.82	0.79	0.73
1		0.64	0.57	0.60	

Dari confusion matrix pada Gambar 3. data pasien Tidak Diabetes nilai confusion matrixnya menggunakan ini TP = 120, TN = 48, FP = 36, dan FN = 27, sedangkan untuk data pasien Diabetes nilai Confusion Matrix menggunakan data berikut ini TP = 48, TN = 120, FP = 27, FN = 36. Menggunakan missing value Imputation Mean.

5. Kesimpulan

Dari hasil penelitian di atas, didapatkan hasil optimum dari masing-masing klasifikasi. Klasifikasi KNN mendapatkan nilai akurasi tertinggi 90% dari nilai K = 5 dengan missing value imputation Median, sedangkan klasifikasi Naïve Bayes memiliki akurasi tertinggi 80% dengan penerapan missing value imputation Mean. Ketika melakukan data cleaning, ada beberapa data yang memiliki missing value. Oleh karena itu, dilakukan penggantian data dengan penerapan missing value imputation menggunakan median dan juga mean.

Dapat dilihat perbandingan hasil dari Missing Value Imputation pada dataset di atas yang bertujuan untuk mengoptimalkan hasil yang didapatkan. Untuk missing value imputation Median, metode klasifikasi KNN mendapatkan akurasi tertinggi dengan nilai 90%, sedangkan untuk metode naïve bayes mendapatkan akurasi sebesar 78%. Sedangkan untuk missing value imputation mean, pada klasifikasi Naïve Bayes cukup mendapatkan akurasi tinggi dengan nilai skor akurasi sebesar 80%, sedangkan penerapan missing value imputation median terhadap klasifikasi KNN dengan K=5 mendapatkan akurasi tertinggi di antara K lainnya.

Daftar Pustaka

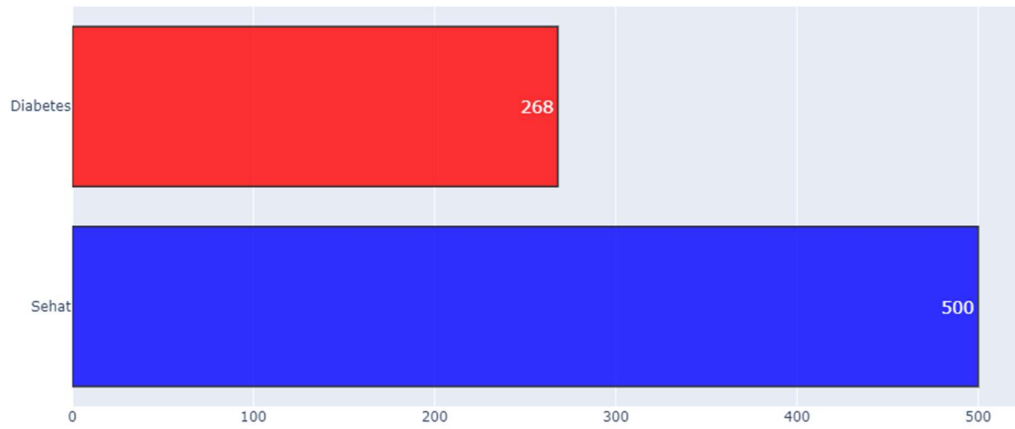
- [1] Agnal, S. A. & Saraswathi, E. (2020). Analyzing Diabetic Data Using Naïve Bayes Classifier : Assistant Professor, Department of Computer Science and Engineering, *SRM Institute of Science and Technology, Chennai, India*. 7(4), 2687–2698.
- [2] Bagus, S. R. (2018) Klasifikasi Penyakit Diabetes Dengan Hidden Naïve Bayes, Jurusan Teknik Informatika, *Universitas Muhammadiyah Jember*, Indonesia.
- [3] Siti, M., Abidatul, I., Arrie, K. & Mukhamad, M. (2014). Optimasi Teknik Klasifikasi Modified K Nearest Neighbor Menggunakan Algoritma Genetika : Jurusan Sistem Informasi, Fakultas Teknik, *Universitas Pesantren tinggi Darul ‘Ulum, Jombang, Indonesia*. 0216-9037.
- [4] Syaputri, A. W., Irwandi, E., & Mustakim, M. (2020). Naïve Bayes algorithm for classification of student major’s specialization. *Journal of Intelligent Computing & Health Informatics*, 1(1), 17.
- [5] Hidayat, T., Anelia, S.S., Pratiwi, U. R., S, N., & Prasvita, S. D. (2021). Perbandingan Akurasi Klasifikasi Diabetes Menggunakan Algoritma AdaBoost Random Forest dan Adaboost Decision Tree dengan imputasi median dan KNN. *Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)*.
- [6] Balakrishnan, V., & Kaur, W. (2019). ScienceDirect String-based Multinomial Naïve Bayes for Emotion Detection String-based Multinomial Naïve Bayes for Emotion Detection among Facebook Diabetes Community among Facebook Diabetes Community. *Procedia Computer Science*, 159, 30–37.
- [7] Fatimah, R. N. (2015). Diabetes Melitus Tipe 2. *Medical Faculty, Lampung University* 4, 86–95.
- [8] Fernanda, S. I., Ratnawati, D. E., & Adikara, P. P. (2017). Identifikasi Penyakit Diabetes Mellitus Menggunakan Metode Modified K- Nearest Neighbor (MKNN). *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, 1(6), 507–513.
- [9] Hestiana, D. W. (2017). Faktor-Faktor yang berhubungan dengan kepatuhan dalam Pengelolaan Diet pada Pasien Rawat Jalan Diabetes Mellitus Tipe 2 Di Kota Semarang. *Jurnal of Health Education*, 138-145.
- [10] Argina, M. A (2020). Penerapan Metode Klasifikasi K-Nearest Neighbor pada Dataset Penderita Penyakit Diabetes. *Indonesia Journal of Data and Science*, 2715-9930
- [11] Meilina, P. (2015). Penerapan Data Mining Dengan Metode Kalsifikasi Menggunakan Decision Tree dan Regresi : Sistem Komputer, dan Teknk Informatika. *Universitas Muhammadiyah Jakarta*. 7(1), 11-20.
- [12] Mustafa, M. S., & Simpen, I. W. (2019). Implementasi Algoritma K-Nearest Neighbor (KNN) Untuk Memprediksi Pasien Terkena Penyakit Diabetes Pada Puskesmas Manyampa Kabupaten Bulukumba. *Prosiding Seminar Ilmiah Sistem Informasi Dan Teknologi Informasi, VIII*(1), 1–10.
- [13] Ridwan, A. (2020). Penerapan Algoritma Naïve Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus. *Jurnal SISKOM-KB (Sistem Komputer Dan Kecerdasan Buatan)*, 4(1), 15–21.
- [14] Yunita, F. (2016). Sistem Klasifikasi Penyakit Diabetes Mellitus Menggunakan Metode K-Nearest Neighbor (K-NN). *Selodang Mayang: Jurnal Ilmiah Badan Perencanaan Pembangunan Daerah Kabupaten Indragiri Hilir*, 2(1), 223–230.
- [15] Devita, N. R., Herwanto, W. H., & Wibawa, P.A., (2017). Perbandingan Kinerja Metode Naïve Bayes dan K-Nearest Neighbor untuk Klasifikasi Artikel Berbahsa Indonesia. *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIK)*. 2355-7699.
- [16] Tapidingan, C.Y., & Paseru, D. (2020). Comparative Analysis of Classification Methods of KNN and Naïve Bayes to Determine Stress Level of Junior High School Students. *Indonesia Journal of Informations Systems (IJIS)*.

- [17] Nguyen, T. T., & Tsoy, Y. (2015). A kernel PLS based classification method with missing data handling. *Statistical Papers*, 58(1), 211-225.
- [18] Singh, S., & Prasad, J. (2013). Estimation of missing values in the data mining and comparison of imputation methods. *Mathematical Journal of Interdisciplinary Sciences*, 1(2), 75-90.
- [19] Maniruzzaman, M., Rahman, M. J., Al-MehediHasan, M., Suri, H. S., Abedin, M. M., El-Baz, A., & Suri, J. S. (2018). Accurate diabetes risk stratification using machine learning: Role of missing value and outliers. *Journal of Medical Systems*, 42.
- [20] Yoon, K. H., et al., "Epidemic Obesity and Type 2 Diabetes in Asia", *Lancet* 2006 Division of Endocrinology and Metabolism, College of Medicine, Catholic University of Korea.

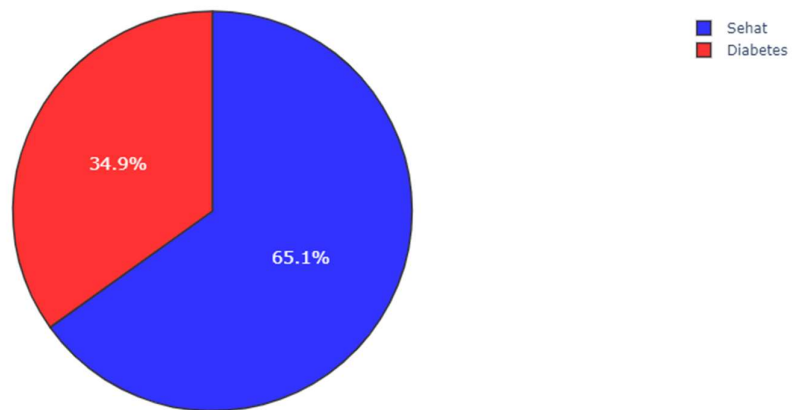
Lampiran

Data Pasien Diabetes

Data Pasien Sehat dan Diabetes

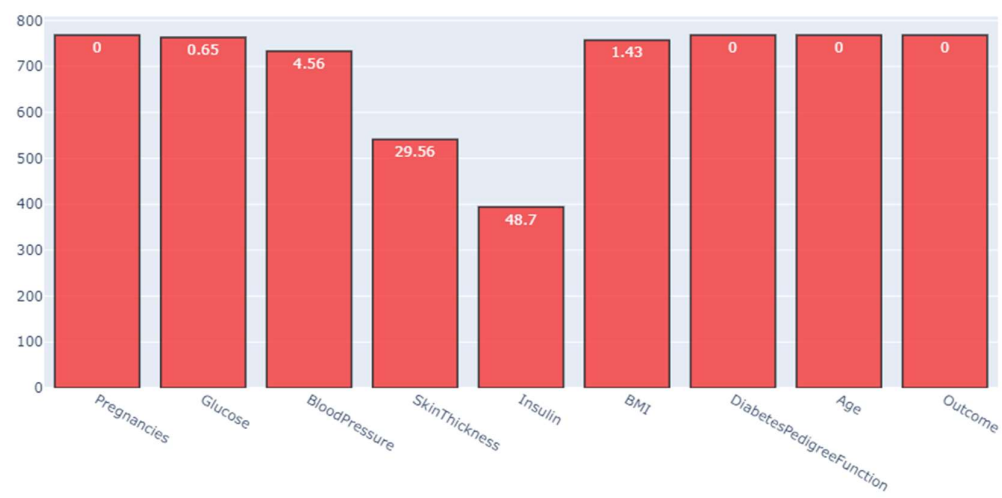


Persentasi Pasien Sehat dan Diabetes

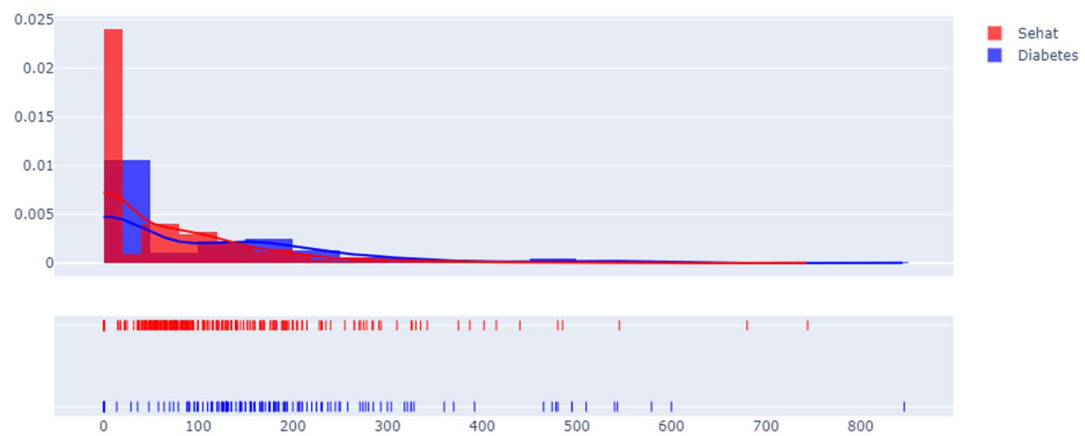


Missing Value

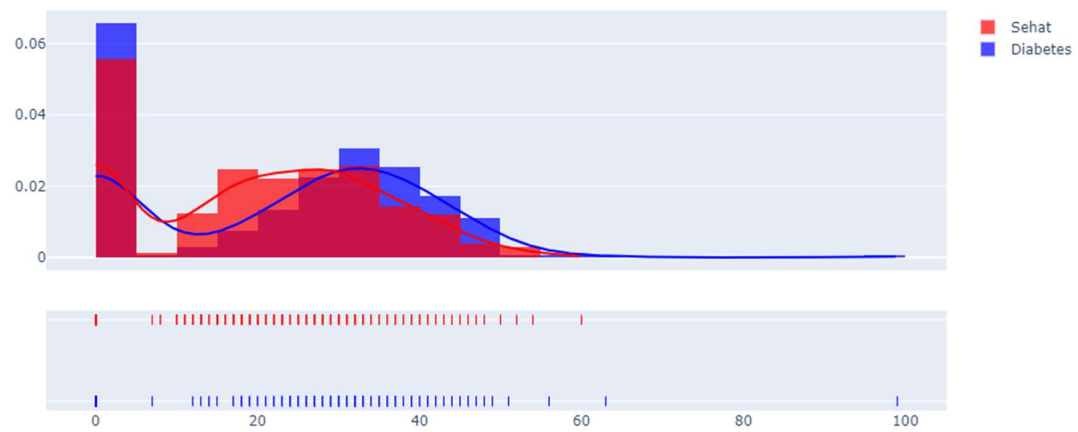
Missing Values (count & %)



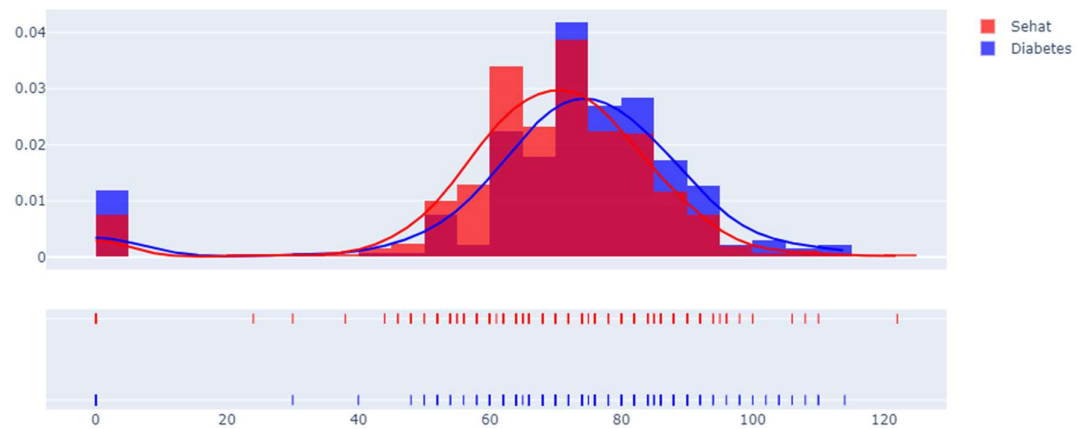
Insulin

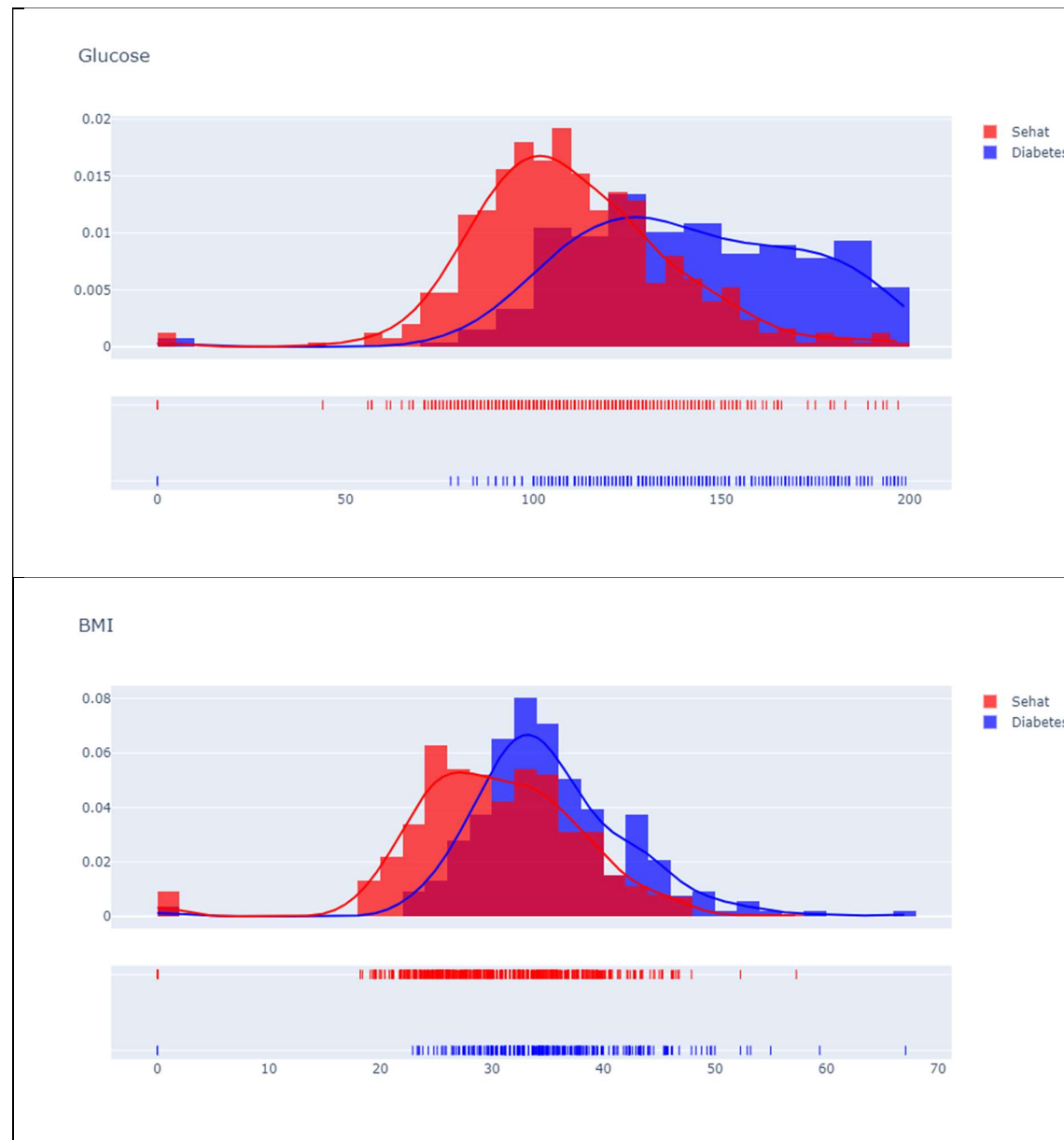


SkinThickness



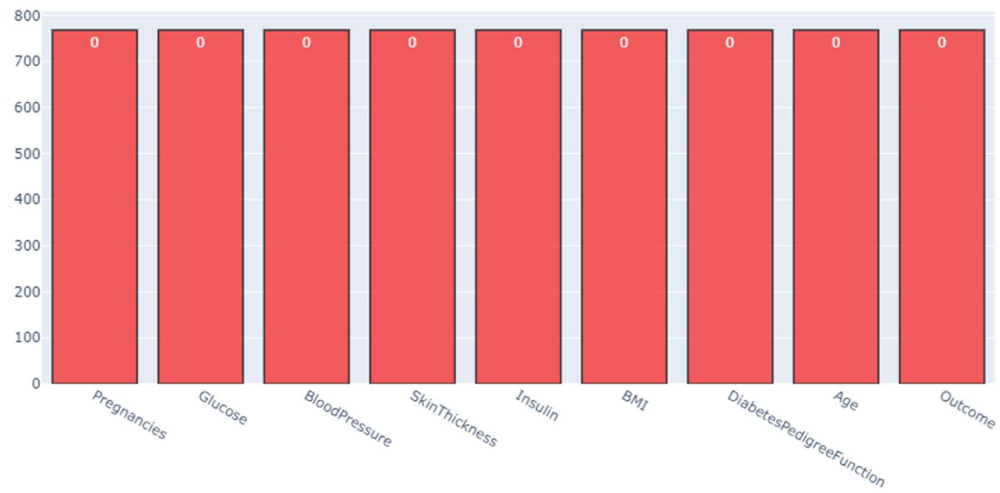
BloodPressure





Setelah Mengisi Missing Value dengan nilai mean dan median

Missing Values (count & %)



In [41]: x_train

Out[41]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
312	2	155.0	74.0	17.0	96.0	26.6	0.433	27
457	5	86.0	68.0	28.0	71.0	30.2	0.364	24
92	7	81.0	78.0	40.0	48.0	46.7	0.261	42
204	6	103.0	72.0	32.0	190.0	37.7	0.324	55
535	4	132.0	NaN	NaN	NaN	32.9	0.302	23
...
252	2	90.0	80.0	14.0	55.0	24.4	0.249	24
582	3	132.0	80.0	NaN	NaN	34.4	0.402	44
722	1	149.0	68.0	29.0	127.0	29.3	0.349	42
104	2	85.0	65.0	NaN	NaN	39.6	0.930	27
475	0	137.0	84.0	27.0	NaN	27.3	0.231	59

891 rows x 8 columns