KLASIFIKASI PENYAKIT DIABETES MENGGUNAKAN METODE *NAÏVE BAYES CLASSIFIER*

SKRIPSI

Oleh:
DENIS ERLANGGA
NIM. 19650012

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2024

KLASIFIKASI PENYAKIT DIABETES MENGGUNAKAN METODE *NAÏVE BAYES CLASSIFIER*

SKRIPSI

Diajukan Kepada:
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang
Untuk Memenuhi Salah Satu Persyaratan Dalam
Memperoleh Gelar Sarjana Komputer (S.Kom)

Oleh:
DENIS ERLANGGA
NIM. 19650012

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM MALANG
2024

HALAMAN PERSETUJUAN

KLASIFIKASI PENYAKIT DIABETES MENGGUNAKAN METODE *NAÏVE BAYES CLASSIFIER*

SKRIPSI

Oleh:
DENIS ERLANGGA
NIM. 19650012

Telah Diperiksa dan Disetujui untuk Diuji:
Tanggal : 02 Mei 2024

Pembimbing I,	Pembimbing II,
Dr. Muhammad Faisal, M.T	Prof. Dr. Suhartono, M.Kom
NIP. 19740510 200501 1 007	NIP. 19680519 200312 1 001

Mengetahui,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Dr. Fachrul Kurniawan, M.MT, IPM
NIP. 19771020 200912 1 001

HALAMAN PENGESAHAN

KLASIFIKASI PENYAKIT DIABETES MENGGUNAKAN METODE *NAÏVE BAYES CLASSIFIER*

SKRIPSI

Oleh:
DENIS ERLANGGA
NIM. 19650012

Telah Dipertahankan di Depan Dewan Penguji Skripsi
dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
Untuk Memperoleh Gelar Sarjana Komputer (S.Kom)
Tanggal : 10 Juni 2024

Susunan Dewan Penguji

Ketua Penguji	: Hani Nurhayati, M.T NIP. 19780625 200801 2 006	( )
Anggota Penguji I	: Roro Inda Melani, M.T., M.Sc NIP. 19780925 200501 2 008	( )
Anggota Penguji II	: Dr. Muhammad Faisal, M.T NIP. 19740510 200501 1 007	( )
Anggota Penguji III	: Prof. Dr. Suhartono, M.Kom NIP. 19680519 200312 1 001	( )

Mengetahui dan Mengesahkan,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Dr. Fachrul Kurniawan, M.MT, IPM
NIP. 19771020 200912 1 001

PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan dibawah ini:

Nama	: Denis Erlangga
NIM	: 19650012
Program Studi	: Teknik Informatika
Fakultas	: Sains dan Teknologi
Judul Skripsi	: Klasifikasi Penyakit Diabetes Menggunakan Metode Naïve Bayes Classifier

Menyatakan dengan sebenarnya bahwa Skripsi yang saya tulis ini benar-banar merupakan hasil karya saya sendiri, bukan merupakan pengambil alihan data, tulisan atau pikiran orang lain yang saya akui sebagai hasil tulisan atau pikiran saya sendiri, kecuali dengan mencantumkan sumber cuplikan pada daftar pustaka.

Apabila dikemudian hari terbukti atau dapat dibuktikan skripsi ini merupakan hasil jiplakan, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Malang, 10 Juni 2024
Yang membuat pernyataan,

Denis Erlangga
NIM. 19650012

HALAMAN MOTTO

“Jika kamu melangkah mungkin ada hasil, mungkin tidak ada hasil.
Tapi jika kamu tidak melangkah, sudah pasti tidak ada hasil.
Takut gagal sama saja artinya dengan takut berhasil”

“Berjuanglah sedikit lebih keras.
Karena yang kamu hadapi bukan hanya pesaingmu, tapi juga sisa umur ibumu”

HALAMAN PERSEMBAHAN

Alhamdulillahirabbil'alamin. Dengan rasa syukur yang sebesar-besarnya saya menyampaikan terima kasih kepada ibu Yuniati dan kakak saya Baladhika Javis Erlangga atas doa, dukungan, kepercayaan, dan kasih sayang yang menjadi sumber motivasi dalam perjalanan menempuh pendidikan akademis saya selama ini sehingga dapat menyelesaikan tugas akhir skripsi ini dengan baik. Tolong agar tetap terus sehat dan selalu bahagia kapanpun dan dimanapun keluarga saya berada. Saya berdoa dan berterima kasih banyak kepada bapak saya alm bapak Kusairi yang telah memberikan banyak hal terbaik kepada saya selama masa hidupnya. Semoga semua harapan dan doa keluarga yang telah diberikan dapat menjadikan saya menjadi pribadi yang lebih baik, berkembang, dan sukses di masa yang akan datang.

Terima kasih kepada teman-teman yang telah membantu dalam bertukar pikiran dan memberikan dukungan penuh yang berharga hingga saat ini dan semoga seterusnya dapat tetap menjalin hubungan baik. Saya juga mengucapkan terima kasih yang mendalam kepada diri saya sendiri sebagai wujud rasa syukur dan penghargaan karena dapat menyelesaikan skripsi ini dengan baik meskipun dengan berbagai kendala yang ada saya dapat bertahan dan melewatinya. Pesan yang ingin saya tekankan yaitu tidak harus menjadi hebat untuk memulai sebuah mimpi, namun hanya perlu memulai untuk meraih mimpi yang hebat. Letak dari sebuah pencapaian bukan dilihat dari hasilnya, melainkan prosesnya. Karya ini saya persembahkan dengan harapan dapat bermanfaat bagi pihak yang sedang mengembangkan topik penelitian ini. Terima kasih atas dukungan dan doa yang terus menerus diberikan.

KATA PENGANTAR

Assalamualaikum Warahmatullahi Wabarakatuh.

Alhamdulillahirabbil'alamin, tidak ada ungkapan yang paling pantas untuk diungkapkan selain rasa syukur dan terima kasih kepada Allah Subhanahu Wa Ta'ala atas karunia dan rahmat-Nya sehingga saya dapat menyelesaikan skripsi ini dengan baik. Shalawat dan salam semoga tercurahkan kepada Nabi Muhammad Rasulullah Shallallahu 'Alaihi Wasallam yang telah memimpin kita semua dari kegelapan menuju jalan yang terang benderang.

Kerendahan hati dan penuh rasa syukur kebahagiaan ini, saya mempersembahkan skripsi yang telah saya susun ini untuk memenuhi salah satu persyaratan sarjana komputer (S.KOM) di Universitas Islam Negeri Maulana Malik Ibrahim Malang program studi Teknik Informatika Fakultas Sains dan Teknologi.

Berbagai kendala yang ditemui selama proses penyusunan skripsi ini dapat saya lewati berkat kegigihan dan kesabaran, sehingga skripsi ini dapat terselesaikan dengan baik. Serta peran dukungan dan arahan yang tak ternilai harganya yang telah diberikan kepada saya dari berbagai pihak juga membangun saya menjadi lebih giat dalam berproses penyusunan skripsi ini.

Dalam perjalanan penyusunan skripsi ini, saya senantiasa merasakan karunia Allah Subhanahu Wa Ta'ala. Segala pencapaian keberhasilan dan kemajuan yang saya raih tidak lepas dari limpahan rahmat-Nya yang begitu besar dan serta petunjuk-Nya yang selalu ada disaat saya membutuhkan. Saya merasa terhormat dan bersyukur atas kesempatan ini. Oleh sebab itu, saya ingin mengucapkan terimkasih kepada :

1. Prof. Dr. H. M. Zainuddin, M.A., selaku rektor Universitas Islam Negeri Maulana Malik Ibrahim Malang.
2. Prof. Dr. Sri Hariani, M.Si., selaku dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang.
3. Dr. Fachrul Kurniawan, M.MT., selaku ketua program studi Teknik Informatika Universitas Islam Negeri Maulana Malik Ibrahim Malang.
4. Dr. Muhammad Faisal, M.T., selaku dosen pembimbing I dan Prof. Dr. Suhartono, M.Kom., selaku dosen pembimbing II yang telah membimbing penulis untuk mengembangkan pemikiran dalam penyusunan skripsi ini hingga selesai.
5. Hani Nurhayati, M.T., selaku dosen penguji I dan Roro Inda Melani, M.T., M.Sc selaku dosen penguji II yang telah menguji, menasehati, serta memberikan saran untuk menjadikan penyusunan skripsi ini lebih baik lagi.
6. Seluruh dosen dan segenap staff program studi Teknik Informatika Universitas Islam Negeri Maulana Malik Ibrahim Malang yang telah memberikan segala ilmu dan wawasan semasa kuliah.
7. Teman-teman program studi Teknik Informatika Angkatan 2019 “ALIEN” yang telah menghabiskan waktu, memotivasi, dan berjuang bersama penulis semasa kuliah.
8. M Alfi Masykur Nazemi, Yusabbih Barqu F L, Rifqi Mufiddin, serta teman-teman dekat penulis yang senantiasa memberikan dukungan penuh dan menjadi teman perjalanan menempuh pendidikan mulai dari menjadi mahasiswa baru hingga lulus satu persatu.

9. Seseorang yang tidak bisa disebutkan namanya yang telah memberikan dukungan dan menjadi penyemangat penulis.
10. Semua pihak yang tidak dapat penulis sebutkan satu persatu yang telah membantu penulis dalam menyelesaikan skripsi ini.

Penulis menyadari masih banyak kekurangan dalam pembuatan skripsi ini karena keterbatasan pengetahuan dan pengalaman. Oleh karena itu, kritik dan saran yang membangun sangat diharapkan demi kemajuan penelitian selanjutnya. Akhir kata, semoga skripsi ini dapat bermanfaat bagi kita semua.

Wassalamualaikum Warahmatullahi Wabarakatuh.

Malang, 10 Juni 2024

Penulis

DAFTAR ISI

HALAMAN PENGAJUAN ..... **ii**
HALAMAN PERSETUJUAN ..... **iii**
HALAMAN PENGESAHAN ..... **iv**
PERNYATAAN KEASLIAN TULISAN ..... **v**
HALAMAN MOTTO ..... **vi**
HALAMAN PERSEMBAHAN ..... **vii**
KATA PENGANTAR ..... **viii**
DAFTAR ISI ..... **xi**
DAFTAR GAMBAR ..... **xiii**
DAFTAR TABEL ..... **xiv**
ABSTRAK ..... **xv**
ABSTRACT ..... **xvi**
خلاصة ..... **xvii**
BAB I PENDAHULUAN **1**
1.1 Latar Belakang 1
1.2 Pernyataan Masalah 5
1.3 Tujuan Penelitian 5
1.4 Batasan Masalah 5
1.5 Manfaat Penelitian 6
BAB II STUDI PUSTAKA **7**
2.1 Penelitian Terkait 7
2.2 Diabetes Melitus 11
2.3 Dataset 15
2.4 Missing Value 16
2.5 Balancing Data 17
2.6 Scaling Data 18
2.7 Split Data 19
2.8 Machine Learning 19
2.9 Naïve Bayes Classifier 22
2.10 K-Fold Cross Validation 22
2.11 Confusion Matrix 24
BAB III DESAIN PENELITIAN **26**
3.1 Diagram Alir Penelitian 26
3.2 Akuisisi Data 27
3.2.1 Data Input 27
3.2.2 Identifikasi Data Fitur Atribut 28
3.2.3 Contoh Dataset Penyakit Diabetes 29
3.3 Preprocessing Data 30
3.3.1 Missing Value 30
3.3.2 Balancing Data 31
3.3.3 Scaling Data 32
3.4 Split Data 33
3.5 10-Fold Cros Validation 34

3.6 Implementasi Metode Naïve Bayes 34
3.7 Skenario Uji Coba 36
BAB IV UJI COBA DAN PEMBAHASAN 38
4.1 Preprocessing 38
4.1.1 Missing Value 38
4.1.2 Balancing Data 41
4.1.3 Scaling Data 43
4.2 Split Data 45
4.3 Hasil Uji Coba 46
4.3.1 Pengujian Rasio Perbandingan 90 : 10 46
4.3.2 Pengujian Rasio Perbandingan 80 : 20 50
4.3.3 Pengujian Rasio Perbandingan 75 : 25 54
4.3.4 Pengujian Rasio Perbandingan 70 : 30 58
4.4 Pembahasan 62
4.5 Integrasi Penelitian Dalam Tafsir Al-Qur'an 77
BAB V KESIMPULAN DAN SARAN 83
5.1 Kesimpulan 83
5.2 Saran 83
DAFTAR PUSTAKA

DAFTAR GAMBAR

Gambar 2. 1 10-Fold Cross Validation 23
Gambar 3. 1 Diagram Alir Penelitian 26
Gambar 3. 2 Data Input Jumlah Pasien 28
Gambar 3. 3 Implementasi Metode Naive Bayes 35
Gambar 4. 1 Missing Value Atribut 38
Gambar 4. 2 Perbandingan Jumlah Data Sebelum dan Sesudah Proses SMOTE. 42
Gambar 4. 3 Source Code Proses SMOTE 43
Gambar 4. 4 Source Code Proses Scaling 45
Gambar 4. 5 Source Code Proses Split Data 45
Gambar 4. 6 Perbandingan Data Actual dan Predicted dari Pengujian 90 : 10 ... 48
Gambar 4. 7 Confussion Matrix Pengujian 90 : 10 49
Gambar 4. 8 Perbandingan Data Actual dan Predicted dari Pengujian 80 : 20 52
Gambar 4. 9 Confussion Matrix Pengujian 80 : 20 53
Gambar 4. 10 Perbandingan Data Actual dan Predicted dari Pengujian 75 : 25 .. 56
Gambar 4. 11 Confussion Matrix Pengujian 75 : 25 57
Gambar 4. 12 Perbandingan Data Actual dan Predicted dari Pengujian 70 : 30 .. 60
Gambar 4. 13 Confussion Matrix Pengujian 70 : 30 61
Gambar 4. 14 Perbandingan Nilai Akurasi Tiap Pengujian 64
Gambar 4. 15 Perbedaan Hasil Prediksi Akurasi Antar Metode 67
Gambar 4. 16 Perbandingan Akurasi Dari 3 Dataset Dengan Naïve Bayes 69
Gambar 4. 17 Perbandingan 10Fold Cross Validation Pengujian 90 : 10 72
Gambar 4. 18 Perbandingan 10Fold Cross Validation Pengujian 80 : 20 73
Gambar 4. 19 Perbandingan 10Fold Cross Validation Pengujian 75 : 25 74
Gambar 4. 20 Perbandingan 10Fold Cross Validation Pengujian 70 : 30 75
Gambar 4. 21 Perbandingan 10Fold Cross Validation Tiap Pengujian 77

DAFTAR TABEL

Tabel 2. 1 Penelitian Terkait 10
Tabel 2. 2 Confusion Matrix 24
Tabel 3. 1 Identifikasi Data Fitur Atribut 29
Tabel 3. 2 Contoh Dataset Penyakit Diabetes 30
Tabel 3. 3 Data Memiliki Missing Value Sebelum Eliminasi 31
Tabel 4. 1 Jumlah Missing Value Pada Tiap Atribut 39
Tabel 4. 2 Jumlah Missing Value Pada Tiap Data 40
Tabel 4. 3 Jumlah Missing Value Pada Tiap Atribut Setelah Eliminasi Data 40
Tabel 4. 4 Contoh Data Setelah Proses Eliminasi 41
Tabel 4. 5 Contoh Data Setelah Proses SMOTE 43
Tabel 4. 6 Contoh Data Setelah Proses Scaling 44
Tabel 4. 7 Rasio Perbandingan Split Data 45
Tabel 4. 8 Hasil Pengujian Prediksi Pengujian 90 : 10 47
Tabel 4. 9 Hasil Confusion Matrix Pengujian 90:10 49
Tabel 4. 10 10-Fold Cross Validation Pengujian 90 : 10 50
Tabel 4. 11 Hasil Pengujian Prediksi Pengujian 80 : 20 51
Tabel 4. 12 Hasil Confusion Matrix Pengujian 80:20 53
Tabel 4. 13 10-Fold Cross Validation Pengujian 80 : 20 54
Tabel 4. 14 Hasil Pengujian Prediksi Pengujian 75 : 25 55
Tabel 4. 15 Hasil Confusion Matrix Pengujian 75:25 57
Tabel 4. 16 10-Fold Cross Validation Pengujian 75 : 25 58
Tabel 4. 17 Hasil Pengujian Prediksi Pengujian 70 : 30 59
Tabel 4. 18 Hasil Confusion Matrix Pengujian 70:30 61
Tabel 4. 19 10-Fold Cross Validation Pengujian 70 : 30 62
Tabel 4. 20 Perbandingan Hasil Akurasi Tiap Pengujian 64
Tabel 4. 21 Perbandingan 10 Fold Cross Validation Tiap Pengujian 76

ABSTRAK

Erlangga, Denis. 2024. **Klasifikasi Penyakit Diabetes Menggunakan Metode Naïve Bayes Classifier**. Skripsi. Program Studi Teknik Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing : (I) Dr. Muhammad Faisal, M.T (II) Prof. Prof. Dr. Suhartono, M.Kom.

Kata Kunci: *Klasifikasi, Penyakit Diabetes, Naïve Bayes Classifier*

Penelitian sebagai klasifikasi untuk penyakit diabetes yang mengimplementasikan metode dari bagian *machine learning* yaitu *Naïve Bayes classifier*, dari penerapan metode ini menghasilkan nilai ketepatan dalam prediksi data menggunakan bahasa pemrograman python. Data yang digunakan yaitu Pima Indians Diabetes Database yang diperoleh melalui situs publik. Isi dari dataset yakni pasien perempuan dengan usia diatas 21 tahun. Data diproses melalui tahapan eliminasi data, *balancing* data, dan *scaling* data lalu dilakukan pemisahan data dengan empat rasio perbandingan data latih dan data uji diantaranya 90:10, 80:20, 75:25, dan 70:30. Dari empat rasio perbandingan tersebut, yang menghasilkan performa model dalam memprediksi ketepatan data yang terbaik didapatkan dari rasio perbandingan 90:10 menghasilkan akurasi sebesar 80% yang dikategorikan baik, presisi pasien positif diabetes sebesar 88%, presisi pasien negatif diabetes sebesar 73%, recall pasien positif diabetes sebesar 74%, recall pasien negatif diabetes sebesar 87%, f1 score pasien positif diabetes sebesar 80%, dan f1 score pasien negatif diabetes sebesar 79%. Serta penerapan teknik validasi silang *k-fold cross validation* dengan nilai *k* sama dengan 10. Diperoleh nilai akurasi dari proses iterasi sepuluh kali menghasilkan nilai ketepatan prediksi data yang optimal dihasilkan dari nilai *k*=7 dengan nilai sebesar 84,06%.

ABSTRACT

Erlangga, Denis. 2024. **The Classification of Diabetes Using Naïve Bayes Classifier Method**. Thesis. Informatics Engineering Faculty of Science and Technology Universitas Islam Negeri Maulana Malik Ibrahim Malang. Advisor: (I) Dr. Muhammad Faisal, M.T. (II) Prof. Prof. Dr. Suhartono, M.Kom.

The research classifies diabetes by implementing a part of machine learning, the Naïve Bayes Classifier. The method implementation generates data prediction accuracy score using python programming. The data is from the Pima Indians Diabetes Database, taken from a public website. The dataset contains female patients above 21 years old. The data processing consists of data elimination, balancing, and scaling. Then, the data are separated into four ratios of training and testing data, namely 90:10, 80:20, 75:25, and 70:30. From the comparison of four ratios, the model performance generating the best data prediction accuracy is the ratio 90:10, leading to 80% accuracy, and is categorized as sufficient. In addition, the precision of positive and negative diabetes patients is 88% and 73%, respectively. The recall of positive and negative diabetes patients is 74% and 87%, respectively. The f1-score of positive and negative diabetes patients is 80% and 79%, respectively. The implementation of the K-Fold Cross Validation technique shows k = 10. The iteration process, conducted ten times, indicates an optimal data prediction accuracy of 84.06 with k=7.

Keywords*: Classification, Diabetes, Naïve Bayes Classifier*

خلاصة

إرلانغا، دينيس. 2024. **تصنيف مرض السكري باستخدام طريقة المصنف البايزي الساذج**. رسالة الماجستير. قسم الهندسة المعلوماتية، كلية العلوم والتكنولوجيا بجامعة مولانا مالك إبراهيم الإسلامية الحكومية مالانج. المشرف الأول: د. محمد فيصل، الماجستير. المشرف الثاني: أ. د. سوهارتونو، الماجستير.

الكلمات الرئيسية: تصنيف، مرض سكري، مصنف بايزي ساذج.

البحث كتصنيف لمرض السكري الذي ينفذ طريقة من قسم أجهزة التعليم، وهي المصنف البايزي الساذج (*Naïve Bayes Classifier*)، من تطبيق هذه الطريقة ينتج قيمة الدقة في التنبؤ بالبيانات باستخدام لغة برمجة *Python*. البيانات المستخدمة هي قاعدة بيانات *Pima Indians Diabetes Database* التي تم الحصول عليها من خلال موقع ويب عام. محتوى مجموعة البيانات هو المرضى الإناث فوق سن 21 عاما. تمت معالجة البيانات من خلال مراحل حذف البيانات وموازنة البيانات وتحجيم البيانات، ثم تم فصل البيانات بأربع نسب مقارنة بين بيانات التدريب وبيانات الاختبار، بما في ذلك 90:10 و 80:20 و 75:25 و 70:30. من بين نسب المقارنة الأربعة، مما أدى إلى أداء النموذج في التنبؤ بأفضل دقة بيانات تم الحصول عليها من نسبة 90:10 مما أدى إلى دقة 80٪ والتي تم تصنيفها على أنها كافية، وكانت ثبات مرضى السكري الإيجابي 88٪، وثبات مرضى السكري السلبي 73٪، وكان استدعاء مرضى السكري الإيجابي 74٪، واستدعاء مرضى السكري السلبي 87٪، وكانت درجة ف1 لمرضى السكري الإيجابي 80٪، ودرجة ف1 لمرضى السكري السلبي 79٪. وكذلك تطبيق تقنية التحقق المتقاطع *K-Fold* بقيمة k تساوي 10. تم الحصول على قيمة الدقة من عملية التكرار عشر مرات مما أدى إلى قيمة الدقة للتنبؤ المثلى للبيانات الناتجة عن قيمة k = 7 بقيمة 84.06٪.

BAB I

PENDAHULUAN

1.1 Latar Belakang

Diabetes terjadi karena kondisi gangguan metabolisme dengan ditandai kadar gula darah atau glukosa mengalami peningkatan. Glukosa yang tinggi menyebabkan organ pankreas terganggu dan tidak dapat memproduksi cukup insulin sehingga menyebabkan kerusakan pada sel pankreas. Hormon insulin yang diproduksi oleh pankreas berfungsi untuk menjaga keseimbangan glukosa di dalam darah untuk mencegah terjadinya fluktuasi glukosa yang signifikan (Rachmawani & Oktarlina, 2017). Diperkirakan akan terdapat 537 juta penderita diabetes di seluruh dunia pada tahun 2021, dan angka tersebut akan meningkat menjadi 643 juta pada tahun 2030 dan 784 juta pada tahun 2045 (International Diabetes Foundation, 2021).

Penyebab terjadinya diabetes dapat berupa faktor genetik dan perilaku pola hidup. Faktor genetik diabetes dapat menurun pada anak dari orang tua dan pada faktor perilaku pola hidup memiliki keterkaitan pada yang dikonsumsi oleh individu karena memiliki kadar glukosa yang beragam pada tiap asupan. Glukosa akan meningkat setelah makan dan akan menurun jika tidak ada asupan makanan atau dua jam setelah asupan dicerna oleh tubuh. Olahraga yang tidak giat dan pilihan gaya hidup yang buruk berkontribusi terhadap berkembangnya penyakit diabetes. Menjadi tidak aktif meningkatkan peluang untuk terkena diabetes karena olahraga membakar glukosa, yang dibutuhkan tubuh untuk energi. sehingga sel-sel menjadi lebih sensitif terhadap insulin. Jika kadar glukosa berlebih menumpuk di

dalam tubuh akan memiliki dampak pada kesehatan dalam jangka waktu tertentu hingga mengakibatkan komplikasi pada tubuh penderita diabetes. Ketika penyakit yang sudah ada sebelumnya berkembang dan menyebar ke bagian tubuh yang lain, menjadikannya lebih parah dan kompleks, atau ketika penyakit baru muncul dapat timbul komplikasi.

Seiring berjalannya waktu jika tidak segera dicegah maupun diatasi, penyakit diabetes akan menyerang organ dan bagian tubuh dari mata sampai ujung kaki hingga menyebabkan kematian. Diabetes merupakan salah satu penyakit kronis diantara beberapa penyakit lain seperti penyakit kardiovaskular, kanker, dan pernapasan kronis yang tidak dapat menyebar ke orang lain. Penyakit kronis termasuk ke dalam penyakit tidak menular (PTM) karena tidak menularkan dari satu orang ke orang lain melalui kontak fisik atau udara (Dinas Kesehatan Riau, 2018). Penyakit tidak menular saat ini menjadi penyebab yang utama kematian di seluruh dunia yang mewakili 80% dari semua tiap kematian setiap tahunnya termasuk penyakit diabetes. Penyakit tidak menular telah merenggut nyawa lebih dari 41 juta orang setiap tahunnya dan sekitar 86% kematian dini akibat dari penyakit tidak menular terjadi pada negara yang memiliki tingkat pendapatan rendah hingga menengah (World Health Organization, 2023).

Masyarakat di dunia sebaiknya lebih waspada terhadap penyakit diabetes, terlebih penyebab diabetes yang masih dianggap sepele menyangkut perilaku pola hidup dan aktivitas fisik yaitu olahraga. Diabetes pada perempuan harus lebih diutamakan karena dapat menurunkan diabetes kepada anak. Anak-anak dapat mewarisi diabetes melalui gen orang tuanya jika salah satu atau keduanya mengidap

penyakit tersebut. Seseorang yang memiliki riwayat keluarga menderita diabetes lebih rentan dibandingkan seseorang yang tidak memiliki riwayat diabetes. (Sudaryanto et al., 2014). Jumlah angka kelahiran dari keluarga yang memiliki riwayat diabetes berpengaruh terhadap populasi diabetes di dunia. Setiap orang di seluruh dunia baik masyarakat modern maupun tradisional dapat berpotensi mengalami diabetes. Karena penyakit diabetes dapat terjadi pada semua orang, yang salah satunya pada perempuan suku Pima Indian. Suku ini merupakan suku asli Amerika yang tinggal di Amerika Serikat bagian Barat Daya dan Utara Meksiko. Pima Indian sebagian besar tinggal di wilayah Sungai Gila dan Sungai Salt di Arizona, AS.

Melihat tingginya angka kematian pada penderita diabetes di seluruh dunia, kondisi ini menjadi prihatin bagi semua masyarakat. Salah satu cara untuk mencegahnya dengan melakukan cek kesehatan untuk mendeteksi penyakit lebih dini yang dilakukan oleh tenaga medis yang ahli dalam bidangnya. Pengecekan kesehatan dapat membantu mendeteksi dini suatu penyakit agar pengobatan dapat dilakukan sesegera mungkin dan membantu mengurangi kesakitan dan komplikasi di masa yang akan datang sehingga diharapkan memiliki kehidupan yang sehat dan panjang (Deliana et al., 2023). Dalam hal ini ilmu pengetahuan sangat dibutuhkan untuk memperoleh wawasan yang luas. Segala urusan dari semua bidang tidak terlepas dari ilmu yang harus dipelajari dan telah diberikan Allah SWT seperti yang tercantum dalam Al Qur’an surah At-Talaq ayat 12 sebagai berikut.

ٱللَّهُ ٱلَّذِى خَلَقَ سَبْعَ سَمَٰوَٰتٍ وَمِنَ ٱلْأَرْضِ مِثْلَهُنَّ يَتَنَزَّلُ ٱلْأَمْرُ بَيْنَهُنَّ لِتَعْلَمُوٓا۟ أَنَّ ٱللَّهَ عَلَىٰ كُلِّ شَىْءٍ قَدِيرٌ وَأَنَّ ٱللَّهَ قَدْ أَحَاطَ بِكُلِّ شَىْءٍ عِلْمًا

"Allah-lah yang menciptakan tujuh langit dan seperti itu pula bumi. Perintah Allah berlaku padanya, agar kamu mengetahui bahwasanya Allah Maha Kuasa atas segala sesuatu, dan sesungguhnya Allah ilmu-Nya benar-benar meliputi segala sesuatu" (QS. At-Talaq : 12).

Dari ayat tersebut dapat dimengerti, bahwa Allah SWT telah memberikan kepada hambanya ilmu pengetahuan dari semua bidang salah satunya ilmu pengetahuan teknologi agar dapat dimanfaatkan sebaik-baiknya untuk keperluan hambanya dalam memudahkan kehidupan sehari-hari.

Ilmu pengetahuan teknologi dapat membantu kebutuhan medis dengan memanfaatkan penggunaan *machine learning. Naïve Bayes* adalah algoritma pembelajaran mesin yang sering digunakan. Ilmuwan Inggris Thomas Bayes menciptakan metode probabilitas dan kategorisasi statistik yang dikenal sebagai *Naïve Bayes*. Kadang-kadang disebut sebagai Teorema Bayes karena membuat prediksi tentang masa depan berdasarkan pengalaman masa lalu (Sihombing, 2021). *Naive Bayes* digunakan untuk menyelesaikan masalah prediksi berupa klasifikasi. Selain itu, *Naive Bayes* memiliki jumlah data latih (*training*) lebih kecil dibandingkan dengan data asli, sehingga mudah menentukan dalam perkiraan parameter yang diperlukan pada alur proses klasifikasi dengan nilai akurasi yang tinggi. Algoritma *Naïve Bayes* dapat melakukan perhitungan dengan cepat dengan akurasi yang tinggi menggunakan alur yang sederhana (Sinaga et al., 2022).

Pemanfaatan *machine learning* diharapkan mampu memudahkan dalam melakukan pengklasifikasian penyakit diabetes dengan efektif, sehingga seseorang dapat diketahui penyakit diabetes lebih awal agar dapat menekan angka kematian penyakit diabetes di Indonesia. Berdasarkan uraian yang dijelaskan, penelitian ini

mengimplementasikan metode *Naïve Bayes classifier* untuk mengklasifikasi penyakit diabetes pada dataset Pima Indians Diabetes Database.

1.2 Pernyataan Masalah

Dari uraian pada latar belakang penelitian ini, bisa dirumuskan pernyataan masalah dengan dasar penyusunan penelitian ini yaitu bagaimana mengetahui performa model klasifikasi penyakit diabetes menggunakan metode algoritma *Naïve Bayes classifier* pada data Pima Indians Diabetes Database?

1.3 Tujuan Penelitian

Penelitian ini bertujuan untuk mengetahui performa model dari klasifikasi penyakit diabetes menggunakan metode algoritma *Naïve Bayes classifier* yang dilakukan peneliti dengan hasil penelitian lain menggunakan metode berbeda yang telah dilakukan dalam penelitian terdahulu pada data yang sama yaitu Pima Indians Diabetes Database.

1.4 Batasan Masalah

Terdapat beberapa batasan-batasan masalah yang diterapkan agar dapat memfokuskan arah penelitian terhadap tujuan yang ingin diraih sebagai berikut.

1. Penelitian ini menggunakan data sekunder untuk umum yang diperoleh melalui platform University of California, Irvine (UCI) *Machine Learning Repository* : Pima Indians Diabetes Database.
2. Klasifikasi yang dilakukan ditujukan untuk pasien perempuan penyakit diabetes pada Pima Indians Diabetes Database yang berusia diatas 21 tahun.
3. Parameter yang digunakan berjumlah 9 fitur atribut.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat mampu memberikan manfaat kepada pembaca terkait topik yang dibahas.

1. Membantu para peneliti yang akan mendalami lebih jauh topik bahasan penelitian ini.
2. Memberikan pengetahuan wawasan mengenai metode *Naïve Bayes classifier* kepada pembaca.
3. Memberikan wawasan mengenai penyakit diabetes kepada pembaca.

BAB II

STUDI PUSTAKA

2.1 Penelitian Terkait

Dalam penelitian Agatsa et al (2020) yang berjudul "*Klasifikasi Pasien Pengidap Diabetes menggunakan Metode Support Vector Machine*" menerapkan algoritma *machine learning* dengan dataset yang digunakan merupakan Pima Indians Diabetes database, yang diperoleh dari UCI *Machine Learning Repository* Databases dengan perolehan data berjumlah 768 pasien dengan atribut berjumlah delapan diantaranya riwayat jumlah kehamilan, kadar glukosa setelah 2 jam makan, tekanan darah, ketebalan lipatan pada kulit trisep, insulin, indeks masa tubuh, riwayat keluarga diabeted dan umur. Sebanyak 614 data digunakan sebagai data latih (*training*) dan 154 data digunakan sebagai data uji (*testing*). Data pasien dilakukan proses klasifikasi menjadi dua kelas yaitu positif diabetes dan negatif diabetes. Data pasien melalui tahap *preprocessing* menggunakan *Min-max* normalisasi yang selanjutnya dilakukan pengujian akurasi dengan menggunakan *k-fold cross validation* pada *training* dan validasi *support vector machine*. Didapatkan hasil evaluasi seberapa baik model yang telah dibangun dengan akurasi sebesar 77,92% (Agatsa et al., 2020).

Dalam penelitian Robbani et al (2020) yang berjudul "*Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma C4.5*" mengimplementasikan *machine learning* dengan *repository public* memiliki data berjumlah 768 dan delapan atribut. Penelitian yang dilakukukan melakukan seleksi pemilihan atribut hanya yang penting saja untuk memfokuskan data. Atribut-atribut yang digunakan diantaranya

seperti glokosa, *blood pressure*, BMI dan *age*. Sedangkan atribut atau fitur yang tidak digunakan dalam penelitian diantaranya *pregnancies* jumlah riwayat kehamilan dianggap tidak memiliki pengaruh pada penyakit diabetes, insulin, *skin thickness* merupakan ketebalan pada kulit untuk pengobatan luka diabetes, dan silsilah keluarga diabetes (*diabetes pedigree function*) dinilai tidak memiliki pengaruh besar terhadap penyakit diabetes serta di dalam data mengandung *noise*. Klasifikasi pada *supervised learning* menggunakan algoritma C4.5 menghasilkan penelitian dengan akurasi sebesar 74.08% (Robbani et al., 2022).

Selain itu terdapat algoritma *Naïve Bayes* yang dapat digunakan dalam klasifikasi seperti yang dilakukan dalam penelitian Paramitha et al (2023) yang berjudul "*Klasifikasi Penyakit Stroke Menggunakan Metode Naïve Bayes*" mengimplementasikan algoritma *machine learning* dengan dataset pasien penyakit stroke sebanyak 200 pasien. Kelas positif stroke dan negatif stroke menggunakan sepuluh variabel bebas. Variabel-variabel tersebut memiliki tujuh variabel dengan tipe data kategorik dan tiga data dengan tipe data numerik. Dilakukan *preprocessing* data dengan pengecekan data hilang atau *missing value* dan data duplikat lalu mengubah data numerik atau angka menjadi data kategorik beberapa interval. Pembagian untuk data *training* dan data *testing* menggunakan rasio 60:40, 70:30, 80:20, dan 90:10. Penggunaan metode *Naïve Bayes classifier* ini berdasar pada asumsi nilai variabel saling bebas apabila diberikan nilai *output* sehingga penelitian ini memperoleh hasil evaluasi model menggunakan *confusion matrix* pada data penyakit stroke pembagian data 80:20 menghasilkan akurasi yang lebih tinggi sebesar 80% (Paramitha et al., 2023).

Sementara pada algoritma *Naïve Bayes* juga digunakan dalam penelitian A'yuniyah et al (2022) yang berjudul "*Implementasi Algoritma Naïve Bayes Classifier (NBC) untuk Klasifikasi Penyakit Ginjal Kronik*" mengimplementasikan algoritma *machine learning* untuk proses klasifikasi dengan dataset penyakit ginjal kronik (PGK) *Dataset* dengan 22 atribut. Dilakukan tahap *preprocessing* data dengan *cleaning* data dan *noise* berupa membersihkan data *non-relevan*, *record* yang hilang, *invalid*, atau salah ketik. Setelah itu mentransformasi data agar lebih sederhana. Diperoleh evaluasi nilai tingkat hasil akurasi sebesar 96.43%, dan hasil recall sebesar 93.18%, hasil presisi sebesar 93.02% dan AUC sebesar 93.2% dengan perbandingan 7:3. Sehingga dapat disimpulkan kinerja NBC dalam klasifikasi data PGK tergolong dalam sangat baik (A'yuniyah et al., 2022).

Dari penelitian terkait seperti yang telah dilakukan sebelumnya pada tabel 2.1 menggunakan data Pima Indians Diabetes Database dan juga menggunakan metode *Naïve Bayes*, di nilai relevan dengan penyusunan penelitian ini menggunakan Pima Indians Diabetes Database serta algoritma *Naïve Bayes classifier* untuk memperoleh estimasi akurasi nilai yang baik. Sehingga penelitian ini bertujuan mengklasifikasikan penyakit diabetes, dengan mengimplementasikan salah satu metode bagian dari machine learning yaitu algoritma *Naïve Bayes*. Penelitian ini memperoleh data dari repositori publik dengan perolehan data pasien sebanyak 768 dan memiliki delapan atribut serta dua label *outcome*.

Tabel 2. 1 Penelitian Terkait

No	Nama Penulis	Metode Penelitian	Objek Penelitian	Hasil Penelitian	Penelitian yang akan dilakukan
1	Agatsa et al, 2020	Support Vector Machine	Pasien Perempuan Pima Indians Diabetes Database	Hasil penelitian dengan klasifikasi yang dilakukan menghasilkan akurasi sebesar 77,92%.	Objek penelitian dan validasi silang yang digunakan sama, namun peneliti menggunakan metode berbeda yaitu Naïve Bayes. Serta melakukan pembagian data *training* dan data *testing* ke dalam beberapa model berbeda.
2	Robbani et al, 2022	Algoritma C4.5	Pasien Perempuan Pima Indians Diabetes Database	Hasil penelitian dengan klasifikasi yang dilakukan menghasilkan akurasi 74,08%.	Objek penelitian yang digunakan sama, Namun peneliti menggunakan metode berbeda yaitu Naïve Bayes dengan menambahkan K-Fold CV. Serta melakukan pembagian data *training* dan data *testing* ke dalam beberapa model berbeda.
3	Paramitha et al, 2023	Naïve Bayes	Pasien Perempuan dan Laki-laki Brain Stroke Prediction Dataset	Hasil penelitian dengan klasifikasi yang dilakukan menghasilkan akurasi sebesar 74.08%	Objek penelitian yang digunakan berbeda yaitu menggunakan Pima Indians Diabetes Database, namun menggunakan metode algoritma sama-sama Naïve Bayes dengan menambahkan K-Fold CV. Serta menggunakan proporsi pembagian model data *training* dan data *testing* berbeda.

No	Nama Penulis	Metode Penelitian	Objek Penelitian	Hasil Penelitian	Penelitian Yang Akan Dilakukan
4	A'yuniyah et al, 2022	Naïve Bayes	Penyakit ginjal kronik (PGK) atau Chronic Kidney Disease Dataset	Hasil penelitian dengan klasifikasi yang dilakukan menghasilkan akurasi 96.43%.	Objek penelitian yang digunakan berbeda yaitu menggunakan Pima Indians Diabetes Database, namun menggunakan metode algoritma sama-sama Naïve Bayes dengan menambahkan K-Fold CV. Serta menggunakan proporsi pembagian model data *training* dan data *testing* berbeda.

2.2 Diabetes Melitus

Pada tahun 2021 terdapat 537 juta pasien diabetes di dunia dan jumlah tersebut diprediksi akan terus terjadi peningkatan jumlah menjadi sebanyak 643 juta orang pada 2030, dan sebanyak 784 juta orang pada 2045 (International Diabetes Foundation, 2021). Diabetes memiliki penyebutan nama lain seperti diabetes melitus, kencing manis, dan penyakit gula. Diabetes melitus adalah kondisi yang terjadi peningkatan glukosa atau kadar gula dalam darah yang disebabkan oleh gangguan metabolisme pada tubuh (Wahyuni, 2022). Glukosa yang meningkat memiliki dampak yang buruk bagi kesehatan seseorang.

Terjadinya peningkatan glukosa dalam tubuh ada kaitannya dengan hormon insulin yang dapat terjadi dan dialami oleh semua orang kalangan berbagai usia. Kondisi kelebihan glukosa yang meningkat berlebihan atau hiperglikemia dapat terjadi karena menurunnya produksi insulin (Hermayanti & Nursiloningrum, 2017). Cara respon tubuh dalam mengatur kadar gula (glukosa) dapat terpengaruh jika seseorang memiliki diabetes. Gula darah merupakan bagian yang termasuk vital

atau fatal bagi kesehatan tubuh karena merupakan suatu sumber energi bagi sel dan jaringan. Glukosa dapat diproses menjadi zat-zat yang diperlukan tubuh.

Ketika seseorang terdiagnosis penyakit diabetes di dalam tubuhnya terjadi kekurangan produksi insulin dari pankreas atau ketidakmampuan tubuh pada individu untuk menggunakan insulin yang dihasilkan pankreas secara efektif sehingga organ pankreas akan terganggu dan kesulitan menjaga keseimbangan kadar glukosa. Hormon insulin yang diproduksi oleh pankreas berfungsi menjaga keseimbangan glukosa dalam darah untuk mencegah terjadinya fluktuasi glukosa yang signifikan (Rachmawani & Oktarlina, 2017). Hal ini karena insulin memiliki fungsi memindahkan glukosa ke dalam sel agar dapat dihasilkan energi. Jika kadar glukosa sangat tinggi menyebabkan organ pankreas menjadi kurang peka terhadap merespon glukosa. Dalam pembahasan lain penyakit diabetes dapat terjadi jika sel tidak memberikan respon sesuai pada insulin saat tubuh tidak lagi dapat menghasilkan cukup insulin yang diperlukan untuk mempertahankan glukosa agar darah normal (Indrayanti et al., 2017)

Seseorang yang mengalami diabetes dapat memiliki beberapa gejala yang timbul diantaranya seperti sering buang air kecil (BAK), sering kehausan, sering lapar, dan penglihatan menjadi kabur atau rabun. Gejala lain yang mungkin juga dikeluhkan diantaranya mudah lelah, mudah lemas, sering merasa kesemutan, sering merasa gatal, impotensi pada pria, serta keputihan pada perempuan (Suardana et al., 2015). Tiap orang dapat memiliki gejala diabetes yang berbeda-beda dipengaruhi oleh kondisi tubuh. Disamping itu diabetes melitus dapat terjadi karena faktor genetik dan perilaku hidup.

Kadar glukosa juga memiliki kaitan dengan apa yang dikonsumsi oleh individu. Pada dasarnya tiap individu memiliki kadar glukosa yang beragam dan akan terjadi peningkatan setelah proses makan dan kemudian kembali seperti semula dalam kurung waktu sekitar dua jam. Peningkatan glukosa didukung dengan perilaku gaya hidup kurang sehat dan pasif berolahraga. Kurangnya aktifitas fisik akan berisiko terkena diabetes karena jika melakukan olahraga glukosa akan dibakar atau diubah menjadi energi yang diperlukan oleh tubuh, sehingga sel – sel menjadi lebih sensitif terhadap insulin.

Pola hidup yang tidak dikendalikan pastinya memiliki dampak negatif terhadap tubuh. Adanya glukosa yang tinggi mengakibatkan adanya penumpukan glukosa di dalam tubuh. Apabila kadar glukosa dibiarkan menumpuk di dalam tubuh akan berdampak serius pada kesehatan seseorang dalam jangka waktu tertentu hingga mengakibatkan komplikasi (Lestari et al., 2021). Komplikasi dapat terjadi saat penyakit yang sudah dimiliki menjadi berkembang lebih rumit dan parah sehingga muncul adanya penyakit baru. Diabetes melitus memiliki risiko komplikasi kardiovaskular seperti serangan jantung dan stroke hingga tiga kali lebih tinggi dengan dibandingkan orang yang tidak memiliki diabetes (Aini et al., 2020).

Pasien diabetes umum ditemui yang sudah memiliki komplikasi karena baru melakukan pengecekan kesehatan. Hal ini terjadi karena pasien meremehkan gejala-gejala yang telah timbul sebelumnya. Penyakit diabetes memiliki jangka waktu pengobatan yang panjang. Pada dasarnya hal ini dapat dicegah dan dihindari dengan melakukan pengecekan kesehatan lebih dini untuk membantu mendeteksi

dini suatu penyakit agar pengobatan dapat dilakukan sesegera mungkin dan membantu mengurangi kesakitan dan komplikasi di masa yang akan datang sehingga diharapkan memiliki kehidupan yang sehat dan panjang (Deliana et al., 2023). Selain itu untuk pasien yang belum terdiagnosis diabetes akan mendapatkan sosialisasi cara mengatur pola perilaku hidup agar mencegah timbulnya penyakit diabetes di waktu yang akan datang.

Jika diabetes tidak segera diatasi, akan menyerang fungsi organ dan bagian tubuh mulai dari mata sampai ujung kaki. Diabetes termasuk dalam penyakit kronis yang dapat menyebabkan kematian. Penyakit kronis merupakan penyakit yang dapat dialami seseorang dalam kurung waktu yang lama, bahkan hingga seumur hidup. Pada awalnya penyakit kronis tidak akan muncul gejala, seiring bertambahnya waktu akan muncul gejala setelah pernyakitnya bertambah parah dan menyebar. Disamping itu, penyakit kronis tidak dapat disebarkan melalui kontak fisik maupun udara karena termasuk dalam penyakit tidak menular (PTM) (Dinas Kesehatan Riau, 2018). Diantara beberapa penyakit menular seperti penyakit jantung, stroke, kanker, pernapasan kronis dan juga diabetes merupakan penyakit yang dialami dengan pengobatan jangka panjang, karena organ dalam tubuh telah tersebar penyakit sehingga tidak dapat bekerja dengan baik dan rusak.

Masyarakat sering menyepelekan penyebab timbulnya PTM karena gaya hidup yang kurang sehat seperti minuman alkohol, makanan tidak bergizi, merokok, dan kurang aktivitas fisik dapat memperburuk keadaan. Pemahaman masyarakat yang kurang akan pentingnya menjaga kesehatan dapat perlahan membutuh diri sendiri. Saat ini penyakit tidak menular telah menjadi penyebab yang utama

terhadap kematian di seluruh dunia. Sebesar 80% dari semua kematian di seluruh dunia setiap tahunnya disumbang dari penyakit tidak menular termasuk diabetes. Lebih dari 41 juta orang meninggal yang diakibatkan oleh penyakit tidak menular setiap tahunnya. Secara keseluruhan kematian PTM di dunia, 77% terjadi di negara dengan ekonomi rendah hingga menengah. Dalam satu tahun terdapat 17 juta orang meninggal dunia karena PTM sebelum berusia 70 tahun. Sekitar 86% kematian dini akibat PTM terjadi pada negara yang memiliki tingkat pendapatan rendah hingga menengah (World Health Organization, 2023)

2.3 Dataset

Suatu dataset berisikan informasi yang tersusun secara sistematis dan dikelompokkan berdasarkan karakteristik untuk tujuan tertentu. Beberapa jenis dataset yang sering ditemui seperti data numerik, data kategori, data teks serta waktu dan tanggal. Peneliti dapat mengumpulkan kumpulan data dari berbagai sumber, termasuk *file*, *database*, dan data observasi. Kumpulan data dapat digunakan dalam berbagai bidang seperti ilmu komputer, ilmu data, statistika, dan *machine learning* untuk melakukan analisis, pengolahan, dan pengembangan model. Dataset juga memiliki peran yang penting dalam sebuah penelitian yang mengolah data. Dataset merupakan kumpulan data yang dapat digunakan sebagai bahan percobaan penelitian (Yuliska & Syaliman, 2020). Sebagai contoh dalam bidang *machine learning*, dataset merupakan bagian dari proses pelatihan sebuah model. Dataset pelatihan berisi contoh-contoh data yang digunakan oleh algoritma *machine learning* untuk memahami pola dan membuat prediksi atau keputusan yang relevan. Dataset yang berkulitas juga akan memiliki pengaruh terhadap pada

pembentukkan model dan hasil akurasi menjadi tinggi. Seiring dengan pertumbuhan teknologi dan pengumpulan data, peran dataset semakin penting dalam pengembangan dan penelitian berbagai aplikasi teknologi informasi. Selain itu kegunaan lain dari dataset adalah untuk mengamati pengaruh dan keterkaitan hubungan antar variabel.

Pada penelitian ini penulis menggunakan data berupa dataset yang didapatkan dari dataset publik melalui UCI *Machine Learning Repository* : Pima Indians Diabetes Database. Dataset berisi 768 pasien dari keturunan Indian Pima. Keseluruhan dataset adalah perempuan berumur setidaknya 21 tahun. Terdapat sembilan atribut yang ada dalam dataset diantaranya yaitu *pregnancies* atau riwayat kehamilan, *glucose* atau glukosa pada tubuh, *blood pressure* atau tekanan darah, *skin thickness* atau ketebalan kulit, kadar insulin, BMI atau masa indeks tubuh, *diabetes pedigree function* atau riwayat diabetes keluarga, dan label *outcome*.

2.4 Missing Value

Permasalahan yang sering berkaitan dengan penggunaan dataset salah satunya yaitu adanya *missing value*. *Missing value* merupakan informasi yang tidak ditemukan atau tidak tersedia dari sebuah objek atau kasus (Rahmat et al., 2017). Hal ini dapat terjadi karena kurangnya informasi mengenai objek atau sampel, kesulitan dalam menemukannya, atau tidak adanya informasi mengenai objek atau sampel tersebut. Secara umum nilai yang hilang tidak menjadi masalah bagi keseluruhan kumpulan data jika nilai tersebut mewakili, misalnya 1% dari total atau jumlah yang sangat kecil. Jika jumlah persentase dari data hilang cukup banyak terutama pada kolom data yang hilang berjumlah dua atau lebih, maka perlu

dilakukan perlakukan atau tindakan untuk mengatasinya. Nilai kosong atau *null* pada dataset dapat dilakukan proses dengan dua cara, yaitu dengan menghilangkan baris yang mengandung nilai *null* atau mengganti nilai *null* berdasarkan *mean*, *median*, atau nilai statistik lainnya dalam kelas data yang sama (Ramadhan, 2021).

2.5 Balancing Data

Balancing data merupakan proses mengatasi ketidakseimbangan dataset terhadap kelas yang memiliki jumlah sampel tidak seimbang. Ketidakseimbangan dapat terjadi saat antara kelas satu dengan yang lain memiliki jumlah sampel yang sedikit dibandingkan dengan kelas lainnya atau dikatakan tidak seimbang. Ketidakseimbangan data akan memiliki pengaruh terhadap hasil penelitian yang dilakukan. Kelas minoritas cenderung akan diabaikan menjadi bias oleh *machine learning* dalam klasifikasi sehingga hasil penelitian kurang maksimal. Proses penyeimbangan dapat dilakukan maupun tidak tergantung pada penelitian yang dilakukan dan dataset yang digunakan. *Balancing* data termasuk ke dalam tahap *preprocessing* pada penelitian. Perlakuan penyeimbangan data terhadap dataset yang sesuai dapat memberikan hasil yang baik terhadap hasil penelitian. Manfaat yang diperoleh dari melakukan penyeimbangan data diantaranya meningkatkan kinerja model dan mendapatkan hasil evaluasi yang maksimal.

Teknik yang dapat digunakan dalam mengatasi ketidakseimbangan data yaitu *oversampling*. Hal ini digunakan untuk menangani ketidakseimbangan kelas pada dataset, khususnya pada masalah klasifikasi. Teknik tersebut memiliki tiga sub teknik diantaranya *random oversampling*, *synthetic minority oversampling technique* (SMOTE), dan *adaptive synthetic sampling* (ADASYN). *Random*

oversampling merupakan suatu teknik dengan mengulang atau menduplikat sampel yang ada secara acak atau random pada kelas yang minoritas hingga jumlahnya sama dengan kelas pada mayoritas (Aryanti et al., 2023). SMOTE merupakan teknik yang dapat meningkatkan jumlah sampel pada kelas minoritas melalui pembuatan sampel sintetis berdasarkan sampel-sampel yang sudah ada dari data yang digunakan (Sulistiyowati & Jajuli, 2020). Sedangkan ADASYN merupakan teknik adaptif untuk menghasilkan sampel data sintetis pada kelas minoritas dibentuk oleh distribusi data secara random pada kelas mayoritas agar mengurangi bias dan memberikan sampel sintetis lebih sulit (Magnolia et al., 2023).

2.6 Scaling Data

Scalling data merupakan proses penskalaan nilai-nilai yang ada pada dataset agar memiliki nilai dengan *range* ukuran yang sama. Standarisasi diperlukan untuk memudahkan *machine learning* dalam melakukan klasifikasi pada objek penelitian. Penskalaan data bertujuan untuk memastikan setiap atribut atau variabel memiliki pengaruh yang seimbang terhadap analisis atau model yang penelitian. *Scaling* data dapat mentransformasi pada data kategori menjadi numerik. Pada umumnya data numerik pada tiap kolom atribut memiliki ukuran arau *range* nilai tidak sama, sehingga diperlukan suatu penskalaan data agar range nilai numerik tersebut dengan yang lain tidak jauh berbeda. *Scaling* data dapat dilakukan sesuai dengan kebutuhan penelitian yang memaksimalkan hasilnya. Beberapa cara yang dapat dilakukan untuk *scaling* data diantaranya *min-max scaling*, normalisasi, dan salah satu cara yang umum digunakan yaitu standarisasi (*standard scaler*).

2.7 Split Data

Split data merupakan proses pembagian atau pemisahan dataset yang digunakan dalam pengujian objek. Pemisahan data dibagi menjadi dua diantaranya yaitu data latih (*training*) dan data uji (*testing*). Hal ini bertujuan untuk memaksimalkan evaluasi kinerja model dan mencegah model terlalu kompleks (*overfitting*) atau model terlalu sederhana (*underfitting*). *Split* data pada *machine learning* membagi data latih lebih besar daripada data uji. Pembagian tersebut dapat disesuaikan dengan kebutuhan penelitian. Saat pembagian dari *training* data dan *testing* data, menginisialisasi dari test_size guna masing-masing uji coba untuk ukuran nilai pembagian data uji. Umumnya yang sering digunakan diantaranya seperti (0,1) 90:10, (0,2) 80:20, (0,25) 75: 25, (0,3) 70:30, dan (0,4) 60:40 dari keseluruhan objek data penelitian. Selain itu menggunakan bilangan yang *random* sebagai random_state dengan melakukan inisialisasi *random number generator* (RNG). Fungsi random_state dapat membantu dalam konsistensi hasil pengujian apabila dilakukan perulangan agar nilai menjadi tetap. Karena jika tidak menggunakan random_state dapat memberikan hasil yang acak saat dilakukan pengujian ulang.

2.8 Machine Learning

Machine learning merupakan suatu bagian dari kecerdasan buatan (*artificial intelligence*) yang memiliki tujuan untuk mengenali atau memahami data dan mengonversinya ke dalam suatu model (Kusuma, 2020). Pembelajaran mesin atau *machine learning* banyak digunakan untuk mempelajari atau menirukan penalaran seperti yang dilakukan oleh manusia untuk membantu menyelesaikan suatu

masalah dengan membangun model yang dibantu oleh bantuan algoritma. Algoritma pada *machine learning* menggunakan teknik statistik guna menemukan pola dari kumpulan data yang berjumlah banyak. Adanya *machine learning* memungkinkan perangkat komputer dapat mengerjakan pengolahan data dengan memberikan hasil yang maksimal secara otomatis. Menggunakan *machine learning* pada komputer dapat membantu membangun model untuk tahap *input-output* tanpa harus memasukan data secara berulang.

Umumnya *machine learning* sering ditemukan pada kasus seperti prediksi, *clustering* dan khususnya klasifikasi. Klasifikasi merupakan pengelompokkan data dengan jumlah banyak berdasarkan kategori dan tujuan tertentu seperti manusia membedakan benda yang satu dengan yang lainnya. Lalu setelahnya digunakan untuk mendeteksi suatu data yang diuji (Abu Ahmad, 2017)

Machine learning dapat diterapkan ke dalam bermacam bidang yang salah satunya bidang kesehatan. Sebagai contoh dokter dapat mendiagnosis suatu penyakit dalam waktu yang singkat tanpa menghabiskan waktu lama. Adanya *machine learning* dapat membantu suatu pekerjaan contohnya proses klasifikasi suatu penyakit pada bidang kesehatan menjadi lebih mudah dalam mengetahui jenis penyakitnya dan memberikan hasil klasifikasi dalam bentuk berupa gambar yang lebih optimal.

Terdapat beberapa macam dari *machine learning* yang diantaranya yaitu *supervised learning*, *unsupervised leaning*, dan *reinforcement learning*. *Supervised learning* membangun fungsi *input-output* berdasar pada data yang ada untuk dipelajari algoritma berdasarkan data *training* yang diberi label dengan tujuan

untuk generalisasi data *input* (Kusuma, 2020). Sedangkan *unsupervised learning* tidak dilakukan pemberian sebuah label dari kumpulan data serta tidak membutuhkan data *training* (Syuhada et al., 2021). Sedangkan *reinforcement learning* berada diantara *supervised learning* dan *unsupervised learning*. *Reinforcement learning* digunakan pada sejumlah data dengan ukuran besar yang dibagi ke dalam dua bagian yang tidak diberi label dan diberi label dimana konsepnya menyelesaikan suatu tujuan dengan tanpa ada pemberitahuan perangkat komputer dengan secara eksplisit apabila tujuan itu sudah tercapai (Roihan et al., 2020)

Metode *supervised learning* memiliki data latih dari data *input* digunakan untuk membangun suatu model dimana selanjutnya digunakan dalam memprediksi data *output*. Selain itu terdapat data uji yang digunakan guna menguji akurasi sistem yang telah dibangun. *Supervised learning* memiliki beberapa ciri khas berupa adanya proses pelatihan, pembelajaran, dan pengujian. Di antara algoritma yang populer dalam pembelajaran yang diawasi termasuk *neural networks, random forest, decision trees,* SVM, *Naïve Bayesian, back-propagation*, dan *linear regression* (Roihan et al., 2020). Dalam penelitian ini mengimplementasikan salah satu algotitma yang terdapat dalam *supervised learning* yaitu *Naïve Bayes classifier* untuk menyeleksi objek dengan karakteristik tertentu untuk membedakan objek satu dengan yang lain dapat digunakannya metode pada klasifikasi berdasarkan dataset yang diperoleh dari dataset publik melalui UCI *Machine Learning Repository* : Pima Indians Diabetes Database.

2.9 Naïve Bayes Classifier

Naïve Bayes classifier, sering dikenal dengan Teorema Bayes, adalah teknik klasifikasi yang mengandalkan metode statistik dan probabilitas. Ini dikembangkan oleh ilmuwan Inggris bernama Thomas Bayes dan memprediksi sebuah peluang masa depan dengan menganalisis pengalaman masa lalu. Keuntungan pendekatan *Naïve Bayes* salah satunya lebih akurat dibandingkan model pengklasifikasi lainnya dan hanya memerlukan sedikit data pelatihan untuk menentukan nilai parameter. (Sihombing, 2021).

Naïve Bayes merupakan algoritma yang sering dipakai untuk aplikasi data mining karena beberapa manfaatnya, termasuk kemudahan penggunaan, pemrosesan data yang cepat, dan pemanfaatan struktur yang efisien dan mudah. (Hadna et al., 2016). Dengan teknik ini, probabilitas sederhana dengan menghitung peluang dan menjumlahkan frekuensi serta kombinasi nilai dari kumpulan data dikategorikan.

2.10 K-Fold Cross Validation

K-fold cross validation yaitu teknik validasi silang untuk mengukur kinerja metode algoritma dengan memisah atau membagi sampel atau data secara *random* dan mengelompokkannya sebanyak jumlah nilai k. Pteroses *k-fold cross validation* memiliki iterasi yang digunakan untuk validasi data pada setiap iterasi *fold*, sedangkan data yang lainnya digunakan untuk data pelaltihan. Proses ini dilakukan berulang sebanyak jumlah *k*, setiap *fold* yang digunakan untuk data validasi secara satu persatu atau bergantian. Pengujian *k-fold cross validation* dapat meninjau suatu

model yang memiliki akurasi tinggi karena data dibagi secara *random* menjadi *k* partisi agar dapat komposisi model yang tepat.

Pada proses *k-fold cross validation* penelitian ini, data diujikan sebanyak sepuluh kali pada setiap data yang dipakai untuk data latih (*training*) dan data uji (*testing*). Pengujian data dilakukan pemecahan dan diulang dengan posisi dari data uji pada tiap iterasi yang berbeda. Hal ini memiliki tujuan untuk memperoleh hasil yang optimal. Pengujian sebanyak *10-fold cross validation* dapat ditampilkan seperti pada gambar 2.1 berikut.

Iterasi Ke-	10-Fold Cross Validation									
1	1	2	3	4	5	6	7	8	9	10
2	1	2	3	4	5	6	7	8	9	10
3	1	2	3	4	5	6	7	8	9	10
4	1	2	3	4	5	6	7	8	9	10
5	1	2	3	4	5	6	7	8	9	10
6	1	2	3	4	5	6	7	8	9	10
7	1	2	3	4	5	6	7	8	9	10
8	1	2	3	4	5	6	7	8	9	10
9	1	2	3	4	5	6	7	8	9	10
10	1	2	3	4	5	6	7	8	9	10

Gambar 2. 1 10-Fold Cross Validation

Iterasi nilai *k* dinyatakan sepuluh kali sehingga data akan diacak sebanyak sepuluh kali. Pada iterasi yang pertama, dari data uji diletakkan pada posisi yang awal, sementara iterasi kedua dari data uji diletakkan pada posisi yang kedua hingga berlanjut sampai iterasi pada kesepuluh. Nilai parameter dari *k* akan menghasilkan suatu akurasi dari penerapan yang dilakukan.

2.11 Confusion Matrix

Confusion matrix dalam bahasa indonesia disebut juga dengan matriks kebingungan. *Confusion matrix* umumnya digunakan sebagai mengukur tingkat kinerja sebuah model yang sudah dibangun. Evaluasi model pada *confusion matrix* mengklasifikasi sebuah prediksi dengan berdasarkan tingkat seberapa dekat kesesuaian dengan nilai aktual dari data. Hasil pengujian *confusion matrix* memiliki dua kelas label yaitu label kelas positif dan label kelas negatif (Putra & Wibowo, 2020). *Confusion matrix* melakukan perbandingan hasil dari klasifikasi data uji berdasarkan data latih dengan data sebenarnya. Bentuk c*onfussion matrix* dengan dua kelas label yaitu *positive* dan *negative* ditampilkan pada tabel 2.2 sebagai berikut.

Tabel 2. 2 Confusion Matrix

Kelas Asli	Kelas Prediksi	
	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

True positive (TP) merupakan jumlah observasi yang seharusnya termasuk ke dalam kelas positif dan diprediksi dengan benar pada kelas positif. *False positive* (FP) adalah jumlah observasi yang seharusnya tidak termasuk ke dalam label kelas positif, namun diprediksi pada kelas positif. *False negative* (FN) adalah jumlah observasi yang seharusnya termasuk ke dalam label kelas positif, namun tidak diprediksi pada kelas positif. *True negative* (TN) adalah jumlah observasi

seharusnya tidak termasuk ke dalam kelas positif dan diprediksi dengan benar sebagai bukan kelas positif.

Untuk mengukur suatu suatu model dapat menggunakan beberapa metode dalam *confussion matrix* seperti akurasi, presisi, recall, dan f1 score. Dari nilai yang didapatkan digunakan sebagai gambaran semakin tinggi dari nilai semakin baik model yang dapat dihasilkan. Untuk memudahkan dalam pemahaman maka hasil akan dikalikan 100% sehingga akan ditampilkan dalam bentuk presentase. Akurasi merupakan presentasi ketepatan kedekatan antara nilai prediksi dengan nilai sebenarnya setelah dilakukan pengujian. Semakin tinggi nilai akurasi yang didapatkan akan semakin baik terhadap hasil nilai evaluasi. Akurasi, presisi, recall dan f1 score dapat dituliskan melalui persamaan-persamaan sebagai berikut.

$$Akurasi = \frac{TP + TN}{TP + FP + TN + FN} x100 \quad (2.1)$$

$$Presisi = \frac{TP}{TP + FP} x100 \quad (2.2)$$

$$Recall = \frac{TP}{TP + FN} x100 \quad (2.3)$$

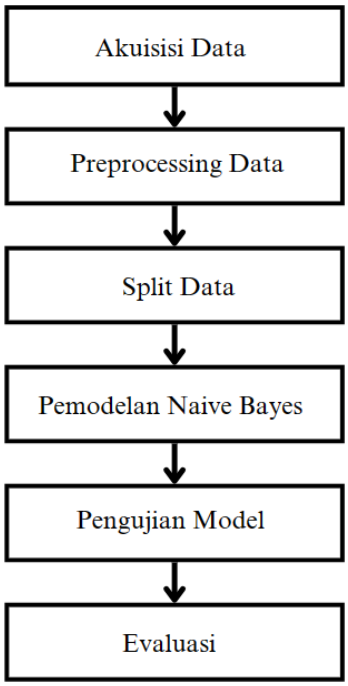
$$F1\ Score = 2x \frac{presisi\ x\ recall}{presisi + recall} x100 \quad (2.4)$$

BAB III

DESAIN PENELITIAN

3.1 Diagram Alir Penelitian

Tahapan penelitian berisi rangkaian tahapan yang bertujuan agar alur penelitian dapat tersusun secara terstruktur dan efektif. Tahap-tahap yang dilakukan melalui beberapa alur untuk mendapatkan hasil penelitian. Tahapan penelitian yang digunakan ditampilkan pada gambar 3.1 sebagai berikut.



Gambar 3. 1 Diagram Alir Penelitian

Identifikasi masalah berisikan rangkaian analisis mengenai permasalahan yang akan diangkat pada penelitian ini. Selanjutnya akan dilanjutkan dengan studi literatur pada penelitian-penlitian yang berkaitan dengan masalah sebagai penambah wawasan mengenai data dan metode. Akuisisi data dilakukan untuk pemahaman data yang digunakan untuk bahan penelitian. Pemisahan data

dilakukan menjadi data latih (*training*) dan data uji (*testing*). Data selanjutnya diolah dengan metode klasifikasi yaitu *Naïve Bayes classifier* berdasarkan data latih. Lalu dilakukan pengujian dan evaluasi model untuk mendapatkan hasil akurasi klasifikasi.

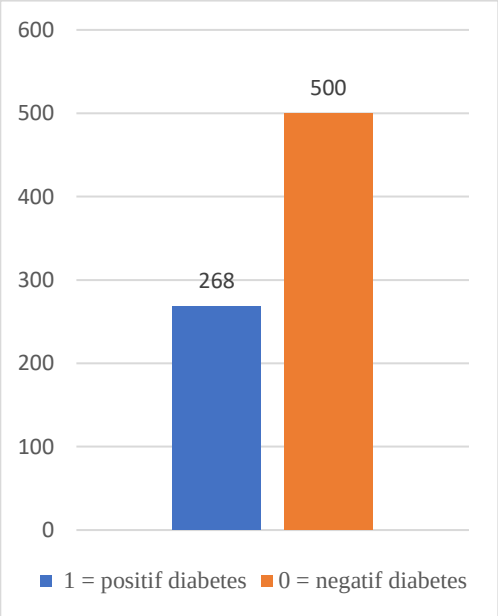
3.2 Akuisisi Data

Tahap ini berisikan informasi mengenai data *input* yang akan digunakan dalam penelittian. Pemahaman data berfungsi agar peneliti maupun pembaca dapat memahami bagian dari dataset yang digunakan. Contoh dataset merupakan sampel dari data asli sesungguhnya yang akan digunakan dalam melakukan klasifikasi penyakit diabetes. Akuisisi data dilakukan untuk mempersiapkan sebelum melakukan penelitian untuk menghindari kekeliruan dalam *input* data.

3.2.1 Data Input

Data input merupakan bahan yang digunakan dalam penelitian ini untuk diolah menggunakan metode. Penelitian menggunakan data sekunder yang diperolehnya dari UCI *Machine Learning Repository* : Pima Indians Diabetes Database. Pemilik dataset diabetes yaitu National Institutes of Diabetes and Digestive and Kidney *Diseases*. Dataset Pima Indians Diabetes Database disumbangkan oleh Vincent Sigillito Research Center, RMI Group Leader Applied Phisics Laboratory The John Hopkins University. Di dalam dataset Pima indians Diabetes Database berisi data pasien sebanyak 768 pasien perempuan. Masing-masing pasien setidaknya sudah berusia 21 tahun keatas dari keturunan suku Indian Pima. Suku Indian Sungai Gila di Arizona Selatan, AS, telah menjadi rumah bagi

suku Indian Pima sejak tahun 1965. Pada usia 35 tahun, suku ini memiliki risiko diabetes terbesar yang tercatat di seluruh dunia (50%). Terdapat sembilan fitur atribut pada penyakit diabetes perempuan di Pima Indian dengan 8 variabel dan 1 label *outcome*. Pada label *outcome* berisikan 268 data pasien positif diabetes dan 500 data pasien negatif seperti yang ditampilkan pada gambar 3.2 sebagai berikut.



Gambar 3. 2 Data Input Jumlah Pasien

Perbandingan jumlah data di atas merupakan data asli yang didapatkan untuk bahan penelitian. Setelah itu data diolah kembali melalui proses *preprocessing.*

3.2.2 Identifikasi Data Fitur Atribut

Identifikasi data fitur atribut berisi pemahaman data yang sangat penting yang mendalam tentang data yang digunakan membantu memahami karakteristik dasar dari dataset. Di samping itu memiliki tujuan memberikan pemahaman mengenai data yang digunakan dalam penelitian dan informasi terkait variabel yang ada pada fitur atribut dataset. Proses pemahaman data mencakup pengumpulan, eksplorasi,

dan struktur data. Variabel atribut yang terdapat pada Pima Indians Diabetes Database diantaranya yaitu *pregnancies, glucose, blood pressure, skin thickness,* insulin, BMI, *diabetes pedigree function, age* dan *outcome* seperti yang ditampilkan pada tabel 3.1 sebagai berikut.

Tabel 3. 1 Identifikasi Data Fitur Atribut

No	Nama Atribut	Deskripsi
1	Pregnancies	Jumlah kehamilan/kelahiran pada perempuan
2	Glucose	Kadar glukosa dalam darah 2 jam setelah makan
3	BloodPressure	Tekanan darah
4	SkinThickness	Ketebalan lipatan kulit trisep
5	Insulin	Kadar insulin dalam dalam 2 jam setelah makan
6	BMI	Tingkat ideal tubuh
7	DiabetesPedigreeFunction	Riwayat diabetes dalam keluarga
8	Age	Umur
9	Outcome	Kelas variabel yang terdiri dari dua kategori 0 untuk tidak penyakit diabetes dan 1untuk penyakit diabetes.

Masing-masing atribut pada Pima Indians Diabetes Database memiliki keterkaitan sebagai indikator penyakit diabetes. Atribut tersebut diambil melalui pendekatan medis dengan sampel darah, riwayat diabetes, dan juga fisik tubuh. Sehingga pendekatan yang dilakukan meliput eksternal dan internal pada pasien.

3.2.3 Contoh Dataset Penyakit Diabetes

Contoh dataset merupakan kumpulan data dengan jumlah banyak yang digunakan untuk keperluan analisis, pengolahan, atau pengembangan model dalam berbagai bidang. Dataset dapat berupa kumpulan pengukuran, observasi, atau informasi lain yang diorganisir dalam suatu format yang dapat diolah. Dataset dapat digunakan dalam berbagai bidang seperti *machine learning*, statistika, dan ilmu

data. Jumlah data yang digunakan dapat memberikan pengaruh pada hasil klasifikasi dari prediksi. Contoh dataset berjumlah lima baris pertama dari dataset Pima Indians Diabetes Database terdapat pada tabel 3.2 sebagai berikut.

Tabel 3. 2 Contoh Dataset Penyakit Diabetes

No	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
1	1	89	66	23	94	28,1	0,167	21	0
2	3	78	50	32	88	31	0,248	26	1
3	2	197	70	45	543	30,5	0,158	53	1
4	1	189	60	23	846	30,1	0,398	59	1
5	1	103	30	38	83	43,3	0,183	33	0

Contoh dataset ini merupakan representasi kecil dari jumlah keseluruhan dataset yang digunakan. Dataset yang diperoleh untuk penelitian ini terdapat 768 pasien perempuan yang berusia diatas 21 tahun.

3.3 Preprocessing Data

Tahap dari *preprocessing* data adalah tahap perlakuan pada dataset yang akan digunakan bertujuan agar memaksimalkan kualitas data. Ada beberapa teknik yang dapat digunakan dalam *preprocessing* data yaitu penanganan *missing value, balancing* data, dan *scaling* data.

3.3.1 Missing Value

Data yang hilang atau *missing value* dapat mempengaruhi kinerja evaluasi model pada klasifikasi. Hal ini dapat diatasi dengan menerapkan perlakukan agar dapat memaksimalkan hasil penelitian. Jumlah *missing value* yang besar berasal dari banyak faktor berkaitan dengan informasi dari sampel pasien. Hal ini bisa jadi

karena pasien tidak memberikan informasi hingga kerusakan atau kesalahan teknis saat pengambilan informasi. Pada Pima Indians Diabetes Database jumlah data yang memiliki *missing value* berkisar hingga ratusan data pasien dan sebagian pasien memiliki *missing value* satu hingga lima atribut. Jika terdapat nilai *missing value* salah satu cara yang dapat dilakukan dengan menghilangkan baris yang mengandung *missing value* (Ramadhan, 2021). Sehingga pada penelitian ini dilakukan eliminasi data untuk mengatasi *missing value*. Contoh data yang memiliki *missing value* ditampilkan pada tabel 3.3 sebagai berikut

Tabel 3. 3 Data Memiliki Missing Value Sebelum Eliminasi

No	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
5	3	78	50	32	88	31	0.248	26	1
…	…	…	…	…	…	…	…	…	…
764	10	101	76	48	180	32.9	0.171	63	0
765	2	122	70	27	0	36.8	0.34	27	0
766	5	121	72	23	112	26.2	0.245	30	0
767	1	126	60	0	0	30.1	0.349	47	1
768	1	93	70	31	0	30.4	0.315	23	0

3.3.2 Balancing Data

Dalam penelitian ini untuk mengatasi ketidakseimbangan data menggunakan teknik *oversampling* yaitu *Synthetic Minority Oversampling Technique* (SMOTE). Teknik ini dapat meningkatkan jumlah sampel pada dataset yang memiliki kelas

tidak seimbang dimana terdapat kelas minoritas dan kelas mayoritas. SMOTE dapat melakukan pembuatan sampel sintetis berdasarkan sampel yang sudah ada dari kelas minoritas untuk diseimbangkan jumlahnya menjadi seimbang dengan kelas mayoritas. Dalam penelitian Heranova (2019), SMOTE termasuk teknik pada *oversampling* yang populer digunakan untuk data *imbalance* (Heranova, 2019). Penggunaan teknik ini diharapkan dapat menyeimbangkan data yang tidak seimbang pada Pima Indians Diabetes Database.

Dalam dataset Pima Indians Diabetes Database terdapat 268 pasien positif diabetes dan 500 pasien negatif diabetes sebelum dilakukan *cleaning* data. Data yang mengandung *noise* seperti *missing value* dieliminasi sehingga tersisa 512 data pasien perempuan yang berumur diatas 21 tahun. Diantara keseluruhan kelas data memiliki *outcome* dengan nilai 1 sebagai positif diabetes dan 0 sebagai negatif terhadap penyakit diabetes. Jumlah pasien yang memiliki diabetes sebanyak 169 pasien, sedangkan yang tidak memiliki diabetes berjumlah 343 pasien.

3.3.3 Scaling Data

Scalling data adalah proses standarisasi nilai pada atribut dataset agar mempunyai skala yang sama. Hal ini bertujuan untuk memastikan setiap atribut atau variabel memiliki pengaruh yang seimbang terhadap analisis atau model yang penelitian. Strategi penskalaan nilai diperlukan karena data numerik biasanya memiliki rentang nilai yang tidak sama agar terdapat perbedaan pada data dengan numerik dengan data lainnya. Penerapan *scalling* data pada penelitian ini menggunakan jenis penskalaan *standard scaler*.

3.4 Split Data

Untuk membagi dataset menjadi data yang akan bersifat subjektif dan data yang akan dianalisis, diperlukan pemisahan data. Prosedur pemisahan data dimulai dengan data diolah, atau dibagi, menjadi data pelatihan dan pengujian. Data pelatihan digunakan untuk melatih model, sedangkan data pengujian digunakan untuk menguji beberapa model terlatih yang dapat membuat prediksi. Empat rasio perbandingan alternatif, termasuk 90:10, 80:20, 75:25, dan 70:30, digunakan dengan kumpulan data pelatihan dan pengujian.

Pemberian nilai *input* inisialisasi pada *random number generator* (RNG) sebagai random_state bertujuan agar pengulangan program yang dijalankan menghasilkan *output* tetap. Penelitian yang dilakukan Sholihah & Hermawan (2023) dengan judul "*Implementation Of Random Forest And Smote Methods For Economic Status Classification In Cirebon City*" menggunakan nilai 42 sebagai random_state pada salah satu pengujiannya (Sholihah & Hermawan, 2023). Selain itu pada penelitian yang dilakukan Felice et al (2023) yang berjudul *"Brain Stroke Prediction Using Random Forest Method with Tuning Parameter"* juga menggunakan nilai bilangan 42 sebagai inisialisasi random_state (Felice et al., 2023). Dalam penelitian Octaviary (2022) yang berjudul *"Deteksi Awal Penyakit Gagal Jantung Berdasarkan Faktor Risiko Menggunakan Metode Naive Bayes"* menyatakan karena nilai acak dapat menghasilkan hasil keluaran yang tetap pada tiap kali sistem program dilakukan kembali, maka nilai masukan yang dimasukkan dalam keadaan_acak untuk pengujian menggunakan nilai 42. (Octaviary, 2022). Penelitian ini saya mencoba mengunakan nilai *input* 42 pada random_state dengan

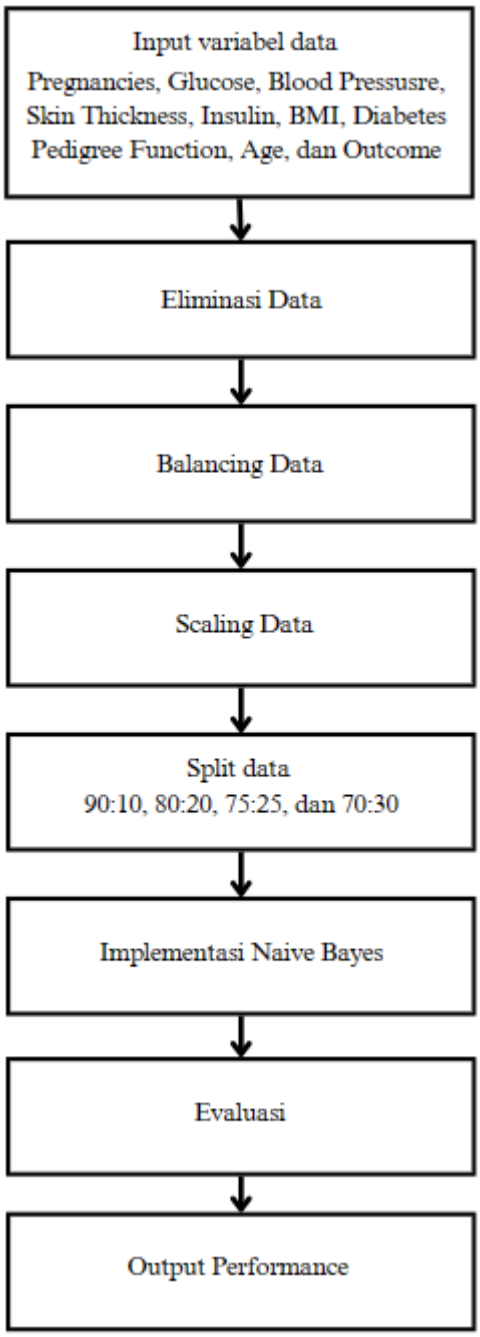
harapan dapat memberikan nilai *output* yang sama pada saat pengulangan uji coba dilakukan kembali.

3.5 10-Fold Cros Validation

Dalam menentukan keakuratan nilai k yang ideal dalam prediksi, penelitian ini juga menggunakan prosedur validasi yang silang dengan *k-fold cross validation*. Dalam penelitian, nilai *k* yang ideal akan menghasilkan akurasi yang sangat baik. Untuk menyelesaikan sepuluh pengulangan, prosedur penentuan nilai *k* diterapkan pada tiap data, termasuk training data dan testing data. Dalam hal ini, nilai *k* diatur ke iterasi sepuluh.

3.6 Implementasi Metode Naïve Bayes

Penelitian ini mengimplementasikan metode algoritma *Naïve Bayes*. Terdapat beberapa alur dalam penelitian ini. Dalam pengaplikasian algoritma *Naïve Bayes* terdapat proses dari *training* dan *testing* data menggunakan aplikasi Google *Colaboratory*. Dalam penerapan *Naïve Bayes* dilakukan pengujian data lalu dihitung probabilitas kelas dari atribut yang diuji. Setelah itu dilanjutkan dengan menghitung probabilitas hasil akhir dari data uji. Hasil akhir akan mengklasifikasikan data ke kelas 1 maupun 0. Dengan syarat diperoleh nilai yang tertinggi untuk menentukan hasil keputusan. Diagram alir dari metode *Naïve Bayes* ditampilkan seperti gambar 3.3 berikut.



Gambar 3. 3 Implementasi Metode Naive Bayes

Proses langkah-langkah implementasi metode *Naïve Bayes* pada diagram di atas yaitu seperti :

1. Pengumpulan data Pima Indians Diabetes Database dengan sembilan atribut di dalamnya.

2. Melakukan eliminasi data terhadap data yang memiliki *missing value* pada atribut.
3. Menyeimbangkan data pada jumlah data pasien kelas positif diabetes dan kelas negatif diabetes sehingga memiliki jumlah yang seimbang untuk diproses.
4. Melakukan *scaling* data atau penskalaan data sehinggga memiliki nilai ukur dengan rentang yang sama.
5. Melakukan pemisahan data latih dan data uji empat kali percobaan dengan rasio perbandingan yang berbeda yaitu 90:10, 80:20, 75:25, dan 70:30.
6. Melakukan klasifikasi menggunakan metode *Naïve Bayes* untuk pemodelan.
7. Mengevaluasi hasil klasifikasi untuk mengetahui hasil akurasi yang didapatkan.
8. Hasil akurasi tertinggi yang diperoleh dari empat pengujian rasio perbandingan terhadap data latih dan data uji digunakan sebagai *output* penelitian.

3.7 Skenario Uji Coba

Skenario uji coba berfungsi sebagai pengujian sistem penelitian yang telah dibuat dengan menghitung nilai akurasi menggunakan bahasa pemrograman Python. Nilai tersebut didapatkan melalui tahapan-tahapan yang sebelumnya dilakukan pada diagram alir penelitian mulai dari akuisisi data, *preprocessing* data, pembagian data, dan pemodelan menggunakan metode *Naïve Bayes*. Pemisahan data latih dan data uji yang digunakan terdapat empat rasio perbandingan yaitu 90:10, 80:20, 75:25, dan 70:30. Pengujian sistem juga menggunakan validasi silang *k-fold cross validation* sebanyak sepuluh kali pengujian atau *fold* = 10. Setelah itu akan di evaluasi terhadap kinerja dari penggunaan metode *Naïve Bayes classifier* pada klasifikasi penyakit diabetes dari data Pima Indians Diabetes Database

menggunakan teknik *confusion matrix* untuk menunjukkan seberapa banyak jumlah prediksi model *Naïve Bayes* yang akurat dan salah pada setiap kelas. *Confusion matrix* menunjukkan nilai dari *true positive* (TP), *false positive* (FP), *true negative* (TN), *false negative* (FN) yang merupakan representasi dari hasil proses klasifikasi. *Confusion matrix* menguji tingkat seberapa baik sistem yang sudah dibuat dalam mengevaluasi kinerja dari model atau sebagai indikator dalam kemampuan dalam estimasi pada kelas dari target. Pengujian akan menghasilkan *output* berupa hasil nilai performa akurasi, presisi, recall, dan f1 score.

BAB IV

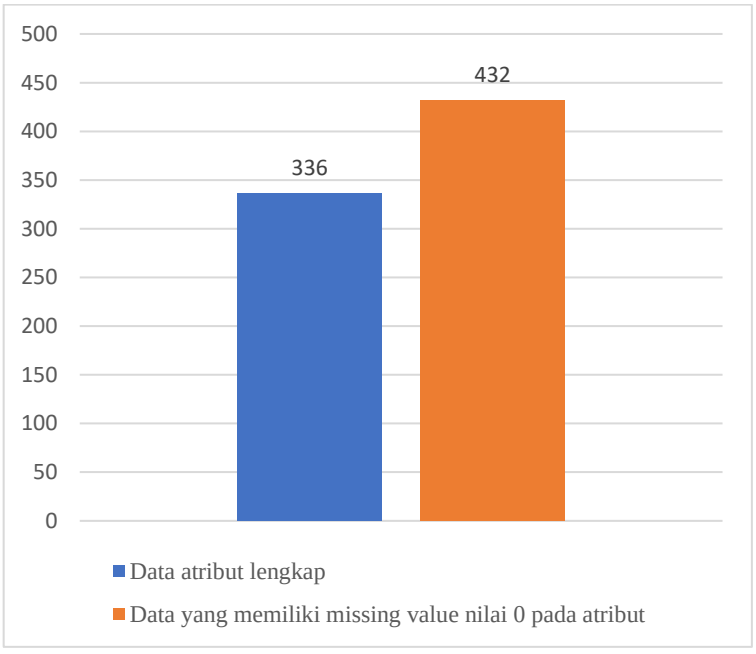
UJI COBA DAN PEMBAHASAN

4.1 Preprocessing

Melalui tahap ini data diolah untuk diproses sebelum dilakukan pemodelan. Persiapan data ini dilakukan untuk memaksimalkan hasil dari implementasi menggunakan metode *Naïve Bayes classifier*.

4.1.1 Missing Value

Permasalahan yang relevan dalam konteks kualitas data adalah keberadaan data yang hilang (*missing value*). Dari keseluruhan 768 data pasien perempuan pada Pima Indians Diabetes Database yang terdiri dari 268 pasien diabetes dan 500 tidak diabetes. Ditemukan 432 data pasien yang memiliki *missing value* pada atribut seperti yang ditampilkan pada gambar 4.1 sebagai berikut.



Gambar 4. 1 Missing Value Atribut

Terdapat jumlah yang tidak sedikit pada kasus *missing value* dalam tiap atribut. Jumlah data yang memiliki *missing value* berkisar hingga ratusan data pasien. Jika terdapat nilai *missing value* dapat dilakukan dengan cara menghilangkan baris yang mengandung *missing value* (Ramadhan, 2021). Sehingga peneliti melakukan eliminasi data.

Dilakukan pemeriksaan pada dataset yang memiliki *missing value*. Dari keseluruhan 768 data pasien perempuan pada Pima Indians Diabetes Database banyak data yang memiliki *missing value* pada masing-masing atribut seperti yang ditampilkan pada tabel 4.1 sebagai berikut.

Tabel 4. 1 Jumlah Missing Value Pada Tiap Atribut

No	Nama Atribut	Jumlah Missing Value
1	Pregnancies	111
2	Glucose	5
3	BloodPressure	35
4	SkinThickness	227
5	Insulin	374
6	BMI	11
7	DiabetesPedigreeFunction	0
8	Age	0

Setelah mengetahui jumlah data yang hilang pada setiap atribut, dilakukan pengecekan pada setiap baris data pasien untuk mengetahui berapa jumlah atribut yang hilang pada setiap data. Ditemukan sebanyak 336 data yang memiliki atribut lengkap atau tidak memiliki *missing value*. Sedangkan jumlah pasien yang memiliki *missing value* sebanyak 432. Terdapat banyak data yang memiliki *missing value* lebih dari satu atribut seperti pada tabel 4.2 sebagai berikut.

Tabel 4. 2 Jumlah Missing Value Pada Tiap Data

No	Keterangan	Missing value atribut	Jumlah data pasien
1	Tidak memiliki missing value	0	336
2	Memiliki missing value di 1 atribut	1	176
3	Memiliki missing value di 2 atribut	2	198
4	Memiliki missing value di 3 atribut	3	42
5	Memiliki missing value di 4 atribut	4	15
6	Memiliki missing value di 5 atribut	5	1

Ditemukan 336 data pasien yang tidak terdapat *missing value* atau memiliki atribut lengkap, 176 data pasien dengan *missing value* di 1 atribut, 198 data pasien dengan *missing value* di 2 atribut, 42 data pasien dengan *missing value* di 3 atribut, 15 data pasien dengan missing value di 4 atribut, 1 data pasien dengan *missing value* di 5 atribut. Data pasien *missing value* dapat memberikan pengaruh pada kualitas data sehingga data pasien yang memiliki jumlah atribut hilang lebih dari satu dilakukan eliminasi. dan tersisa 512 data untuk dianalisis lebih lanjut. Jumlah data pada masing-masing atribut yang masih memiliki *missing value* pada satu atribut masih tetap dipertahankan seperti yang ditampilkan pada tabel 4.3 sebagai berikut.

Tabel 4. 3 Jumlah Missing Value Pada Tiap Atribut Setelah Eliminasi Data

No	Nama Atribut	Jumlah Missing Value
1	Pregnancies	56
2	Glucose	1
3	BloodPressure	0
4	SkinThickness	0
5	Insulin	119
6	BMI	0
7	DiabetesPedigreeFunction	0
8	Age	0

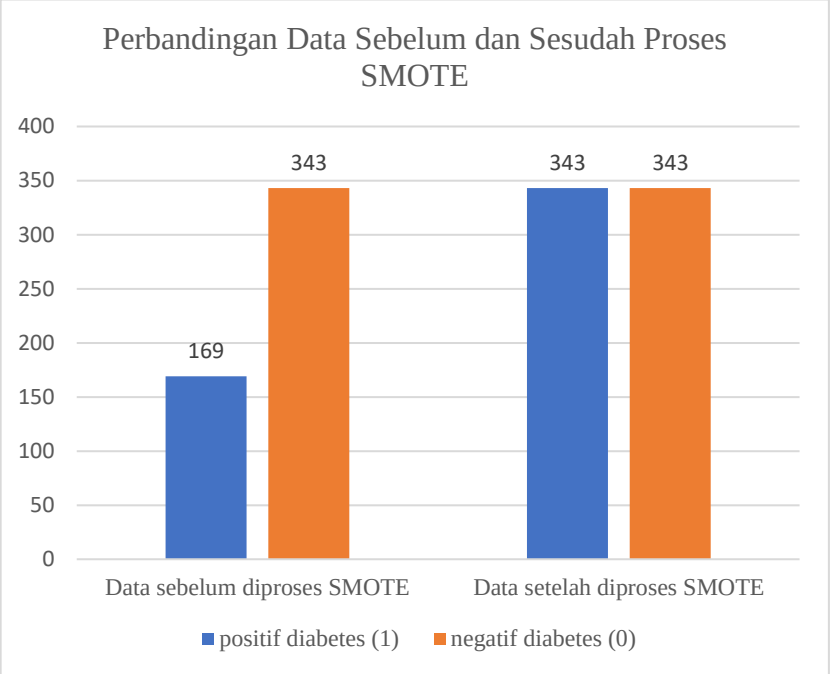
Contoh dataset sebelum dilakukan eliminasi data terdpat pada tabel 3.3. Lalu setelah dilakukan proses eliminasi data ditampilkan pada tabel 4.4 sebagai berikut.

Tabel 4. 4 Contoh Data Setelah Proses Eliminasi

No	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
5	3	78	50	32	88	31	0.248	26	1
…	…	…	…	…	…	…	…	…	…
508	9	170	74	31	0	44	0.403	43	1
509	10	101	76	48	180	32.9	0.171	63	0
510	2	122	70	27	0	36.8	0.34	27	0
511	5	121	72	23	112	26.2	0.245	30	0
512	1	93	70	31	0	30.4	0.315	23	0

4.1.2 Balancing Data

Tahap selanjutnya data akan dilakukan proses penyeimbangan atau oversampling menggunakan metode *synthetic minority oversampling technique* (SMOTE). Hal ini berdasar karena label *outcome* pada dataset memiliki jumlah yang tidak seimbang. Data pasien sebanyak 512 data yang terdiri dari 169 kasus positif diabetes dan 343 kasus negatif diabetes sebelum dilakukan *oversampling* ditampilkan pada gambar 4.2 sebagai berikut.



Gambar 4. 2 Perbandingan Jumlah Data Sebelum dan Sesudah Proses SMOTE

SMOTE digunakan dalam pengelolaan ketidakseimbangan kelas pada dataset mengenai masalah klasifikasi. Model data yang tidak seimbang dapat menyebabkan algoritma *machine learning* akan condong atau mengarah memberikan akurasi lebih besar pada label kelas mayoritas daripada label kelas minoritas sehingga hasilnya kurang maksimal (Gu et al., 2016).

Hasil penyeimbangan dataset menghasilkan kelas minoritas menjadi seimbang sebanyak 343 positif diabetes dan 343 negatif diabetes. Contoh dataset setelah dilakukan proses SMOTE ditampilkan pada tabel 4.5 sebagai berikut.

Tabel 4. 5 Contoh Data Setelah Proses SMOTE

No	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
5	3	78	50	32	88	31	0.248	26	1
…	…	…	…	…	…	…	…	…	…
682	8	176	78	31	0	39.84641	0.379544	42	1
683	3	180	72	24	144	24.24742	0.305431	43	1
684	5	97	73	29	90	39.28636	0.870193	31	1
685	5	132	79	26	0	29.31518	0.449488	56	1
686	6	160	68	28	488	34.50666	0.458061	38	1

Berikut adalah *source code* dari proses *balancing* data yang ditampilkan pada gambar 4.3 sebagai berikut.

```
smote = SMOTE(random_state=42)
X_resampled, y_resampled = smote.fit_resample(X, y)
```

Gambar 4. 3 Source Code Proses SMOTE

4.1.3 Scaling Data

Data selanjutnya akan melalui proses data *scaling* yaitu pengubahan mengubah fitur-fitur atribut dalam suatu dataset sehingga memiliki skala nilai yang teratur. Tujuan dari standarisasi ini adalah untuk membuat semua fitur atribut memiliki skala yang sama sehingga tidak ada fitur atribut yang dominan secara numerik dalam mempengaruhi hasil model pada dataset. Contoh data sebelum diproses dengan *scaling* terdapat pada tabel 4.5. Sedangkan contoh data setelah dilakukan proses *scaling* ditampilkan pada tabel 4.6 sebagai berikut.

Tabel 4. 6 Contoh Data Setelah Proses Scaling

No	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
1	0.647172	0.685586	0.015788	0.499526	-0.95699	0.008564	0.299327	1.644739	1
2	-0.86908	-1.25125	-0.48796	-0.10431	-0.95699	-1.05315	-0.53326	-0.17154	0
3	-0.86908	-1.12827	-0.48796	-0.70814	-0.25187	-0.82564	-1.08832	-1.12747	0
4	-1.17234	0.347409	-2.67088	0.499526	0.30322	1.449461	5.309953	0.019648	1
5	-0.26258	-1.46645	-1.8313	0.19761	-0.29688	-0.38579	-0.84398	-0.64951	1
…	…	…	…	…	…	…	…	…	…
682	1.253674	1.546399	0.519538	0.096971	-0.95699	0.955978	-0.44715	0.879991	1
683	-0.26258	1.669373	0.015788	-0.6075	0.123191	-1.40997	-0.67073	0.975584	1
684	0.34392	-0.88233	0.099746	-0.10431	-0.28188	0.871033	1.032952	-0.17154	1
685	0.34392	0.193692	0.603496	-0.40622	-0.95699	-0.64133	-0.23616	2.218301	1
686	0.647172	1.054506	-0.32005	-0.20494	2.703615	0.146079	-0.2103	0.497616	1

Berikut merupakan *source code* dari proses *scaling* data yang ditampilkan pada gambar 4.4 sebagai berikut.

```
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

Gambar 4. 4 Source Code Proses Scaling

4.2 Split Data

Prosedur yang melibatkan inisialisasi test_size untuk masing-masing dari empat skenario pengujian dan membagi data menjadi set pelatihan dan pengujian. Data uji dibagi menggunakan ukuran nilai berikut: 0,1, 0,2, 0,25, dan 0,3. Untuk menginisialisasi random_state, yang menerima angka acak, peneliti juga menggunakan nilai input 42, seperti yang ditunjukkan pada Gambar 4.5.

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.1, random_state=42)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=42)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

Gambar 4. 5 Source Code Proses Split Data

Rasio yang digunakan untuk split data yaitu perbandingan 90:10 total data latih 617 data dan data uji 69 data. Perbandingan 80:20 total data latih 548 data dan 138 data uji. Perbandingan 72:25 total dari data latih 514 dan 172 data uji, serta perbandingan 70:30 total data latih 480 data dan 206 data uji yang ditunjukkan tabel 4.7 sebagai berikut.

Tabel 4. 7 Rasio Perbandingan Split Data

No	Rasio perbandingan	Jumlah Data Latih	Jumlah Data Uji
1	90:10	617	69
2	80:20	548	138
3	75:25	514	172
4	70:30	480	206

4.3 Hasil Uji Coba

Bagian hasil uji coba membahas dan menganalisis hasil pengujian dari sistem setelah melewati proses *preprocessing*, dengan fokus pada evaluasi metode *Naive Bayes* untuk klasifikasi penyakit diabetes dengan skenario uji coba yang sudah ditentukan. Dataset awal yang tidak seimbang antara pasien positif dan negatif telah diseimbangkan menjadi 686 entri. Selanjutnya, dataset dibagi menjadi empat skenario rasio perbandingan yang berbeda untuk menilai akurasi prediksi terbaik dari *output* yang diprediksi. Hasil uji coba menunjukkan akurasi yang dihitung berdasarkan nilai y_pred dan y_test dari implementasi metode algoritma *Naive Bayes*. Evaluasi kinerja sistem dilakukan menggunakan *confusion matrix* untuk menganalisis klasifikasi antara data *actual* dan *predicted* yang sudah dilakukan.

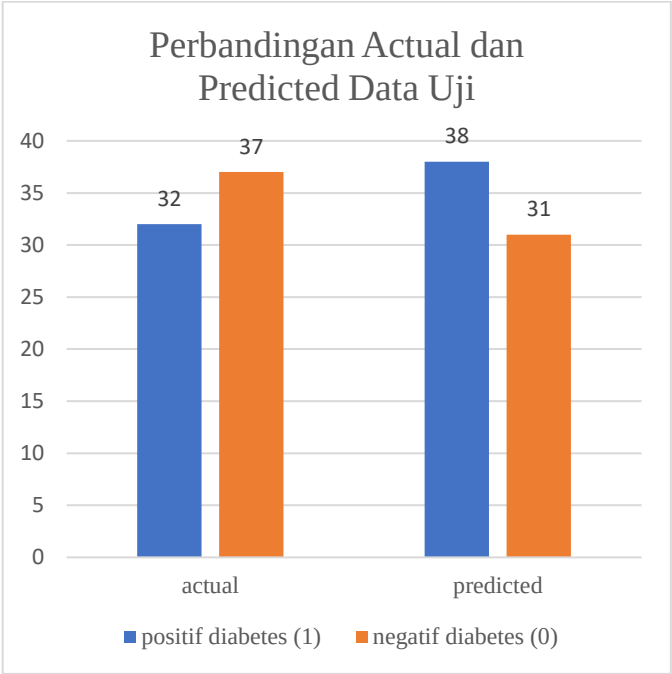
4.3.1 Pengujian Rasio Perbandingan 90 : 10

Data *training* sebanyak 617 data dan data uji sebanyak 69 data. Nilai *input* test_size 0,1, dan nilai *input* random_state 42 digunakan sedangkan pengujian menggunakan rasio perbandingan 90:10. Hasil pengujian yang menampilkan data aktual dan data antisipasi disajikan pada tabel 4.8. telah dilakukan pengujian.

Tabel 4. 8 Hasil Pengujian Prediksi Pengujian 90 : 10

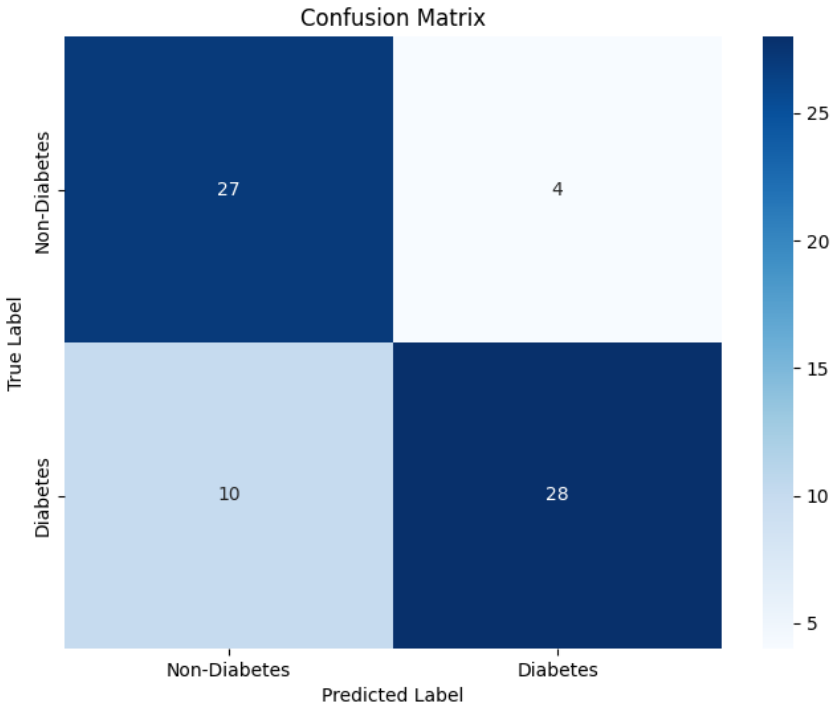
No	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	y_pred	Outcome
1	-0.26258	-1.12827	0.183705	-1.41261	-0.31938	-0.47679	0.070063	0.497616	0	0
2	1.556925	0.562612	1.359121	0.298248	0.265714	-0.42346	0.621928	1.835927	1	1
3	-0.26258	-0.7901	-0.32005	-0.70814	-0.34939	-0.29478	1.270683	-0.45832	0	0
4	-1.17234	1.669373	1.527038	-0.40622	-0.28188	0.448417	-0.64488	0.210835	1	1
5	-0.26258	0.993019	-0.15213	-0.00367	1.503418	0.296743	-0.55438	0.210835	1	1
…	…	…	…	…	…	…	…	…	…	…
65	0.34392	0.747072	1.44308	1.304635	-0.95699	2.466911	-0.55814	0.019648	1	1
66	2.76993	1.300453	0.015788	1.103358	-0.06434	0.65836	0.76376	0.879991	1	1
67	-0.26258	-0.82084	0.687455	-1.9158	-0.47691	-2.16037	-0.73538	-0.26713	0	0
68	-0.86908	-0.57489	-0.32005	-1.11069	-0.95699	-1.06832	-1.09436	-0.84069	0	0
69	-0.56583	-1.37422	-0.99171	-0.80878	-0.95699	-0.88631	-0.71728	-0.7451	0	0

Dari pengujian perbandingan 90 : 10 dengan nilai *actual* 1 (positif diabetes) sebanyak 32 data dan nilai *actual* 0 (negatif diabetes) sebanyak 37 data, menghasilkan nilai dari prediksi 1 (positif diabetes) sebanyak 38 data dan nilai *actual* 0 (negatif diabetes) sebanyak 31 data seperti yang ditampilkan gambar 4.6 dibawah ini.



Gambar 4. 6 Perbandingan Data Actual dan Predicted dari Pengujian 90 : 10

Gambar 4.6 menunjukkan data dari hasil pengujian dengan rasio perbandingan 90:10, di mana terdapat prediksi 27 data sebagai diabetes dan hasilnya benar-benar terdeteksi sebagai diabetes (TP), 4 data diprediksi tidak diabetes tetapi sebenarnya diabetes (FN), 10 data diprediksi diabetes tetapi sebenarnya tidak (FP), dan 28 data diprediksi tidak diabetes dan benar-benar tidak terdeteksi diabetes (TN). Informasi ini direpresentasikan dalam bentuk tabel *confusion matrix* seperti yang ditunjukkan pada gambar 4.7.



Gambar 4. 7 Confussion Matrix Pengujian 90 : 10

Berdasarkan gambar 4.7 *confusion matrix* pengujian rasio perbandingan 90 : 10 menghasilkan nilai akurasi, presisi, recall dan f1 score yang ditunjukkan pada tabel 4.9 sebagai berikut.

Tabel 4. 9 Hasil Confusion Matrix Pengujian 90:10

No	Confusion Matrix	Nilai dalam persen
1	Akurasi	80%
2	Presisi pasien positif diabetes	88%
3	Presisi pasien negatif diabetes	73%
4	Recall pasien positif diabetes	74%
5	Recall pasien negatif diabetes	87%
6	F1 score pasien positif diabetes	80%
7	F1 score pasien negatif diabetes	79%

Pengujian rasio 90 : 10 menggunakan validasi silang dengan nilai *k*=10 untuk evaluasi algoritma *Naïve Bayes*. Data dibagi sesuai dengan pembagian proporsi yang digunakan yaitu 90 : 10. Sehingga diperoleh hasil kinerja akurasi dalam presentase untuk dataset diabetes yang disajikan pada tabel 4.10 sebagai berikut.

Tabel 4. 10 10-Fold Cross Validation Pengujian 90 : 10

Parameter K	Pengujian 90 : 10
K = 1	63.77%
K = 2	72.46%
K = 3	68.12%
K = 4	75.36%
K = 5	79.71%
K = 6	66.67%
K = 7	84.06%
K = 8	75.36%
K = 9	68.12%
K = 10	73.91%

Pengujian akurasi data pasien diabetes pada pengujian dengan rasio perbandingan 90 : 10 mendapatkan akurasi nilai tertinggi pada *k*=7 dengan nilai 84.06% yang ditunjukkan pada label kuning tabel 4.10 diatas.

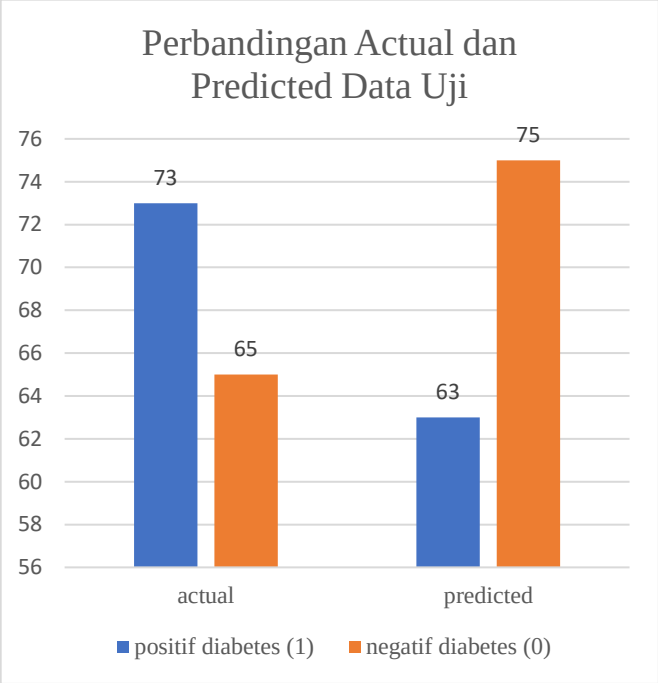
4.3.2 Pengujian Rasio Perbandingan 80 : 20

Pengujian menggunakan rasio perbandingan 80:20 dilakukan dengan 548 data untuk pelatihan dan 138 data untuk pengujian, dengan test_size 0.2 serta random_state 42. Pengujian ini mengasilkan deteksi, yang menampilkan data aktual dan hasil prediksi, dapat dilihat pada tabel 4.11.

Tabel 4. 11 Hasil Pengujian Prediksi Pengujian 80 : 20

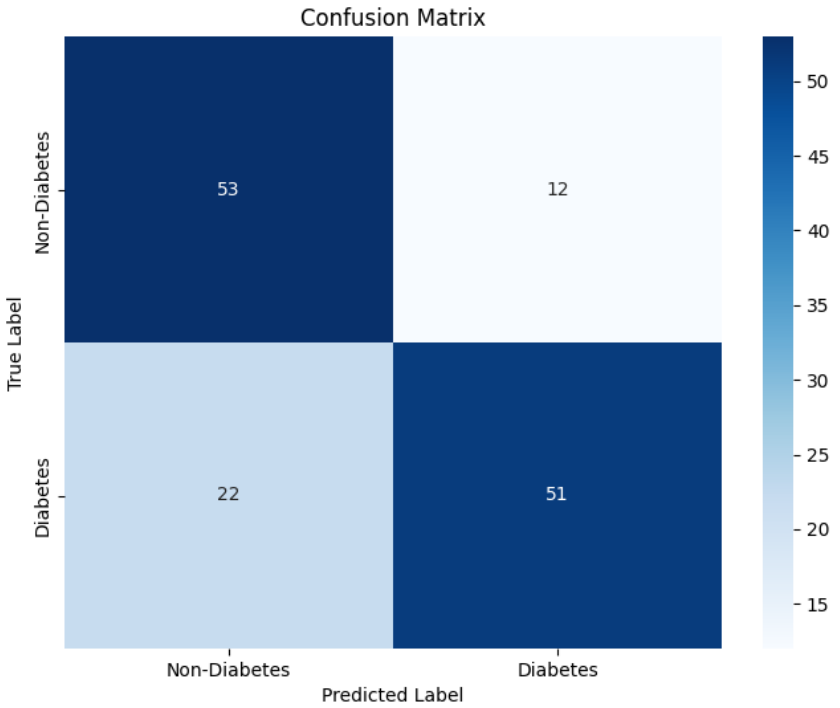
No	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	y_pred	Outcome
1	-0.26258	-1.12827	0.183705	-1.41261	-0.31938	-0.47679	0.070063	0.497616	0	0
2	1.556925	0.562612	1.359121	0.298248	0.265714	-0.42346	0.621928	1.835927	1	1
3	-0.26258	-0.7901	-0.32005	-0.70814	-0.34939	-0.29478	1.270683	-0.45832	0	0
4	-1.17234	1.669373	1.527038	-0.40622	-0.28188	0.448417	-0.64488	0.210835	1	1
5	-0.26258	0.993019	-0.15213	-0.00367	1.503418	0.296743	-0.55438	0.210835	1	1
…	…	…	…	…	…	…	…	…	…	…
134	-0.26258	-1.28199	0.015788	0.19761	-0.95699	0.554588	-0.78666	-0.45832	0	0
135	-0.56583	-1.15902	-1.15963	-0.40622	-0.83697	-0.78014	0.718639	-1.03188	0	0
136	0.040669	-0.45192	0.015788	1.70719	0.595769	0.539421	2.601018	2.218301	1	1
137	-0.56583	0.624099	0.351621	0.499526	0.498253	0.706261	-0.59963	-0.36273	1	0
138	-0.86908	1.023762	-1.32755	0.298248	3.176193	0.944704	-0.46429	-0.7451	1	1

Dari pengujian perbandingan 80 : 20 dengan nilai *actual* 1 (positif diabetes) sebanyak 73 data dan nilai *actual* 0 (negatif diabetes) sebanyak 65 data, menghasilkan nilai dari prediksi 1 (positif diabetes) sebanyak 63 data dan nilai *actual* 0 (negatif diabetes) sebanyak 75 data seperti yang ditampilkan gambar 4.8 dibawah ini.



Gambar 4. 8 Perbandingan Data Actual dan Predicted dari Pengujian 80 : 20

Gambar 4.8 menampilkan data hasil pengujian dengan rasio perbandingan 80:20. Terdapat prediksi 53 data sebagai diabetes yang benar-benar terdeteksi sebagai diabetes (TP), 12 data diprediksi tidak diabetes tetapi sebenarnya diabetes (FN), 22 data diprediksi diabetes tetapi sebenarnya tidak terdeteksi diabetes (FP), dan 51 data diprediksi tidak diabetes dan benar-benar tidak terdeteksi diabetes (TN). Hasil ini disajikan dalam tabel *confusion matrix* seperti yang terlihat pada gambar 4.9.



Gambar 4. 9 Confussion Matrix Pengujian 80 : 20

Berdasarkan gambar 4.9 *confusion matrix* pengujian rasio perbandingan 80 : 20 menghasilkan menghasilkan nilai akurasi, presisi, recall dan f1 score yang ditunjukkan pada tabel 4.12 sebagai berikut.

Tabel 4. 12 Hasil Confusion Matrix Pengujian 80:20

No	Confusion Matrix	Nilai dalam persen
1	Akurasi	75%
2	Presisi pasien positif diabetes	81%
3	Presisi pasien negatif diabetes	71%
4	Recall pasien positif diabetes	70%
5	Recall pasien negatif diabetes	82%
6	F1 score pasien positif diabetes	75%
7	F1 score pasien negatif diabetes	76%

Pengujian rasio 80 : 20 menggunakan validasi silang dengan nilai *k*=10 untuk evaluasi algoritma *Naïve Bayes.* Data dibagi sesuai dengan pembagian proporsi yang digunakan yaitu 80 : 20. Sehingga diperoleh hasil kinerja akurasi dalam presentase untuk dataset diabetes yang disajikan pada tabel 4.13 sebagai berikut.

Tabel 4. 13 10-Fold Cross Validation Pengujian 80 : 20

Parameter K	Pengujian 80 : 20
K = 1	68.12%
K = 2	68.84%
K = 3	72.46%
K = 4	73.91%
K = 5	78.26%
K = 6	71.01%
K = 7	79.71%
K = 8	77.54%
K = 9	73.91%
K = 10	74.64%

Pengujian akurasi data pasien diabetes pada pengujian dengan rasio perbandingan 80 : 20 mendapatkan akurasi nilai tertinggi pada *k*=7 dengan nilai 79.71% yang ditunjukkan pada label kuning tabel 4.13 diatas.

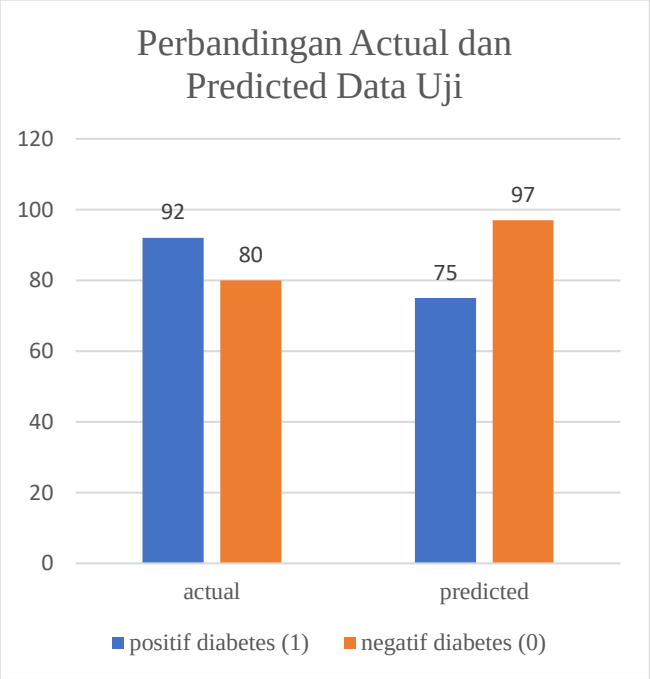
4.3.3 Pengujian Rasio Perbandingan 75 : 25

Pada pengujian dengan rasio perbandingan 75:25, digunakan 514 data untuk pelatihan dan 172 data untuk pengujian, dengan test_size 0.25 serta random_state 42. Pengujian ini menghasilkan deteksi, yang memperlihatkan data aktual dan hasil prediksi, ditampilkan dalam tabel 4.14.

Tabel 4. 14 Hasil Pengujian Prediksi Pengujian 75 : 25

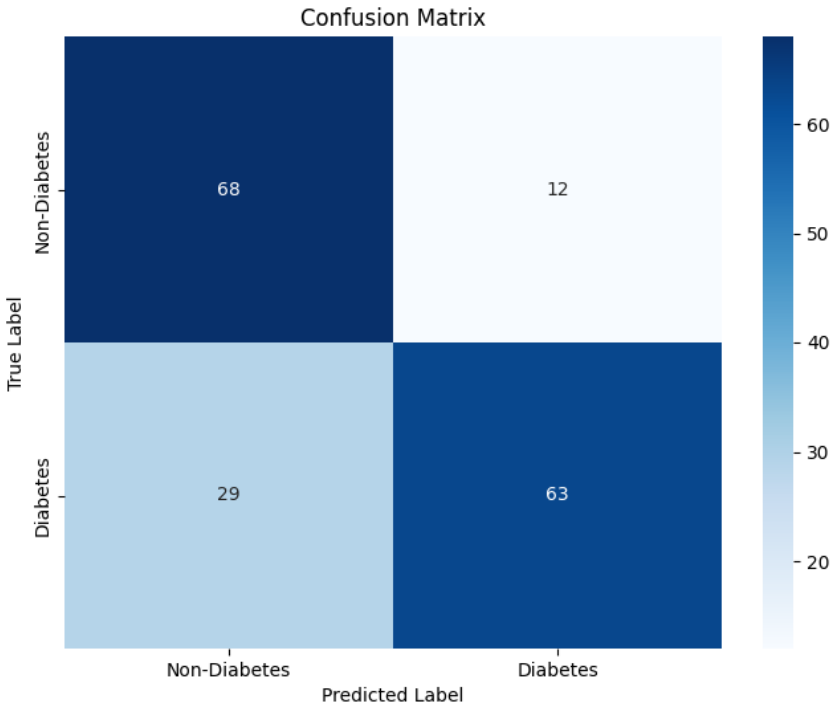
No	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	y_pred	Outcome
1	-0.26258	-1.12827	0.183705	-1.41261	-0.31938	-0.47679	0.070063	0.497616	0	0
2	1.556925	0.562612	1.359121	0.298248	0.265714	-0.42346	0.621928	1.835927	1	1
3	-0.26258	-0.7901	-0.32005	-0.70814	-0.34939	-0.29478	1.270683	-0.45832	0	0
4	-1.17234	1.669373	1.527038	-0.40622	-0.28188	0.448417	-0.64488	0.210835	1	1
5	-0.26258	0.993019	-0.15213	-0.00367	1.503418	0.296743	-0.55438	0.210835	1	1
…	…	…	…	…	…	…	…	…	…	…
168	-0.86908	1.085249	-0.99171	0.096971	3.03367	0.950838	-0.88124	-0.7451	1	1
169	-0.26258	0.070719	0.015788	-0.50686	0.468248	-0.17344	0.06403	-0.55391	0	1
170	-0.56583	-0.42117	0.267663	0.19761	-0.95699	0.327078	-1.14564	-1.12747	0	0
171	1.556925	-0.48266	0.771413	-0.50686	-0.95699	-0.70299	1.994587	1.549146	1	1
172	-0.86908	-1.0053	-1.32755	-1.9158	-0.95699	-1.67501	-0.33417	-1.03188	0	0

Dari pengujian perbandingan 75 : 25 dengan nilai *actual* 1 (positif diabetes) sebanyak 92 data dan nilai *actual* 0 (negatif diabetes) sebanyak 80 data, menghasilkan nilai dari prediksi 1 (positif diabetes) sebanyak dari 75 data dan nilai dari *actual* 0 (negatif diabetes) sebanyak dari 97 data seperti yang ditampilkan gambar 4.10 dibawah ini.



Gambar 4. 10 Perbandingan Data Actual dan Predicted dari Pengujian 75 : 25

Gambar 4.10 menunjukkan data dari pengujian dengan rasio 75:25, di mana 68 data diprediksi sebagai diabetes dan benar-benar terdeteksi (TP), 12 data diprediksi tidak diabetes namun sebenarnya diabetes (FN), 29 data diprediksi sebagai diabetes tetapi sebenarnya tidak (FP), dan 63 data diprediksi tidak diabetes dan benar-benar tidak terdeteksi (TN). Hasil ini ditampilkan dalam tabel *confusion matrix* seperti terlihat pada gambar 4.11.



Gambar 4. 11 Confussion Matrix Pengujian 75 : 25

Berdasarkan tabel 4.11 *confusion matrix* pengujian rasio perbandingan 75 : 25 menghasilkan nilai akurasi, presisi, recall dan f1 score yang ditunjukkan pada tabel 4.15 sebagai berikut.

Tabel 4. 15 Hasil Confusion Matrix Pengujian 75:25

No	Confusion Matrix	Nilai dalam persen
1	Akurasi	76%
2	Presisi pasien positif diabetes	84%
3	Presisi pasien negatif diabetes	70%
4	Recall pasien positif diabetes	68%
5	Recall pasien negatif diabetes	85%
6	F1 score pasien positif diabetes	75%
7	F1 score pasien negatif diabetes	77%

Pengujian rasio 75 : 25 menggunakan validasi silang dengan nilai *k*=10 untuk evaluasi algoritma *Naïve Bayes.* Data dibagi sesuai dengan pembagian proporsi yang digunakan yaitu 75 : 25. Sehingga diperoleh hasil kinerja akurasi dalam presentase untuk dataset diabetes yang disajikan pada tabel 4.16 sebagai berikut.

Tabel 4. 16 10-Fold Cross Validation Pengujian 75 : 25

Parameter K	Pengujian 75 : 25
K = 1	68.60%
K = 2	70.35%
K = 3	73.26%
K = 4	73.84%
K = 5	74.42%
K = 6	71.51%
K = 7	76.74%
K = 8	76.74%
K = 9	72.67%
K = 10	76.16%

Pengujian akurasi data pasien diabetes pada pengujian dengan rasio perbandingan 75 : 25 mendapatkan akurasi nilai tertinggi pada *k*=7 dan *k*=8 dengan nilai 76.74% yang ditunjukkan pada label kuning tabel 4.16 diatas.

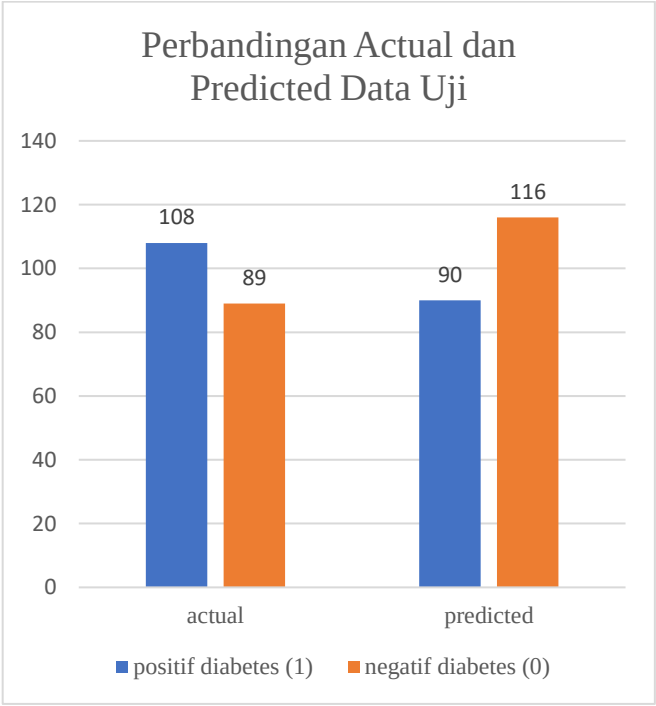
4.3.4 Pengujian Rasio Perbandingan 70 : 30

Pengujian menggunakan rasio 70:30 dilakukan dengan 480 data *training* dan 206 data *testing*, digunakan input test_size 0.3 serta random_state 42. Pengujian ini menghasilkan deteksi, yang memperlihatkan data aktual dan prediksi, dapat dilihat pada tabel 4.17.

Tabel 4. 17 Hasil Pengujian Prediksi Pengujian 70 : 30

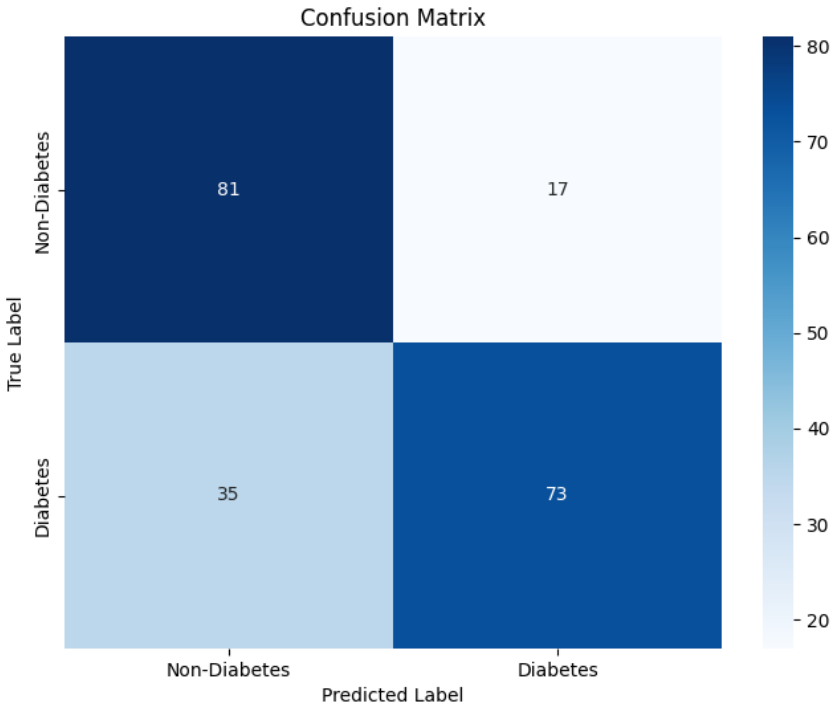
No	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	y_pred	Outcome
1	-0.26258	-1.12827	0.183705	-1.41261	-0.31938	-0.47679	0.070063	0.497616	0	0
2	1.556925	0.562612	1.359121	0.298248	0.265714	-0.42346	0.621928	1.835927	1	1
3	-0.26258	-0.7901	-0.32005	-0.70814	-0.34939	-0.29478	1.270683	-0.45832	0	0
4	-1.17234	1.669373	1.527038	-0.40622	-0.28188	0.448417	-0.64488	0.210835	1	1
5	-0.26258	0.993019	-0.15213	-0.00367	1.503418	0.296743	-0.55438	0.210835	1	1
…	…	…	…	…	…	…	…	…	…	…
202	-0.56583	-0.42117	-0.32005	-0.80878	-0.25187	0.084401	-0.64186	-0.64951	0	0
203	-0.86908	-0.39043	-0.73984	0.499526	-0.95699	0.010123	-0.04647	-1.03188	0	1
204	0.040669	0.562612	-1.15963	-0.20494	0.093186	-0.6133	-0.72633	0.402023	0	0
205	-0.26258	0.009232	1.359121	1.103358	0.805803	0.873102	0.531608	-0.55391	1	0
206	0.34392	-0.5134	-0.48796	1.002719	-0.02683	-0.01424	0.071533	-0.64951	0	1

Dari pengujian perbandingan 70 : 30 dengan nilai *actual* 1 (positif diabetes) sebanyak 108 data dan nilai *actual* 0 (negatif diabetes) sebanyak 98 data, menghasilkan nilai dari prediksi 1 (positif diabetes) sebanyak dari 90 data dan dari nilai *actual* 0 (negatif diabetes) sebanyak 116 data seperti yang ditampilkan gambar 4.12 dibawah ini.



Gambar 4. 12 Perbandingan Data Actual dan Predicted dari Pengujian 70 : 30

Berdasarkan gambar 4.12, hasil pengujian dengan rasio 70:30 menunjukkan 81 data yang diprediksi diabetes dan benar-benar terdeteksi (TP), 17 data yang diprediksi tidak diabetes tetapi sebenarnya diabetes (FN), 35 data yang diprediksi diabetes namun sebenarnya tidak (FP), dan 73 data yang diprediksi tidak diabetes dan benar-benar tidak terdeteksi (TN). Hasil ini ditampilkan dalam tabel *confusion matrix* seperti terlihat pada gambar 4.13.



Gambar 4. 13 Confussion Matrix Pengujian 70 : 30

Berdasarkan gambar 4.13 *confusion matrix* pengujian rasio perbandingan 70 : 30 menghasilkan nilai akurasi, presisi, recall dan f1 score yang ditunjukkan pada tabel 4.18 sebagai berikut.

Tabel 4. 18 Hasil Confusion Matrix Pengujian 70:30

No	Confusion Matrix	Nilai dalam persen
1	Akurasi	75%
2	Presisi pasien positif diabetes	81%
3	Presisi pasien negatif diabetes	70%
4	Recall pasien positif diabetes	68%
5	Recall pasien negatif diabetes	83%
6	F1 score pasien positif diabetes	74%
7	F1 score pasien negatif diabetes	76%

Pengujian rasio 70 : 30 menggunakan validasi silang dengan nilai *k*=10 untuk evaluasi algoritma *Naïve Bayes*. Data dibagi sesuai dengan pembagian proporsi yang digunakan yaitu 70 : 30. Sehingga diperoleh hasil kinerja akurasi dalam presentase untuk dataset diabetes yang disajikan pada tabel 4.19 sebagai berikut.

Tabel 4. 19 10-Fold Cross Validation Pengujian 70 : 30

Parameter K	Pengujian 70 : 30
K = 1	72.82%
K = 2	70.39%
K = 3	73.30%
K = 4	75.73%
K = 5	73.79%
K = 6	70.87%
K = 7	76.70%
K = 8	79.61%
K = 9	72.82%
K = 10	77.18%

Pengujian akurasi data pasien diabetes pada pengujian dengan rasio perbandingan 70 : 30 mendapatkan akurasi nilai tertinggi pada *k*=8 dengan nilai 79.61% yang ditunjukkan pada label kuning tabel 4.19 diatas.

4.4 Pembahasan

Pada pembahasan ini berisi penyajian hasil dari penelitian yang telah diujikan. Pengujian yang telah dilaksanakan pada program yang telah dibuat untuk klasifikasi penyakit pada pasien diabetes menggunakan *machine learning* dengan algoritma metode *Naïve Bayes classifier*. Data yang digunakan didapatkan melalui

website atau situs publik yaitu Pima Indians Diabetes Database yang berisi pasien perempuan sebanyak 768 data. 268 positif dan 500 negatif.

Data dilakukan penyisihan terhadap data pasien yang memiliki *missing value* lebih dari 1 atribut agar memiliki hasil penelitian yang baik sehingga tidak terdapat *noise*. Dataset memiliki 8 atribut diantaranya seperti *pregnancies* (kehamilan), *glucose* (glukosa), *blood pressure* (tekanan darah), *skin thickness* (ketebalan kulit), *insulin, BMI* (massa ideal tubuh), *diabetes pedigreem function* (keturunan keluarga diabetes), *age* (umur), dan memiliki dua label *outcome*. Label dari kelas Pima Indians Diabetes Database memiliki dua kelas yang pertama nilai 0 sebagai tidak memiliki diabetes dan yang kedua nilai 1 sebagai pasien yang memiliki penyakit diabetes. Terlihat ada ketidakseimbangan atau *imbalance* data jumlah antar label kelas 0 negatif diabetes dan kelas 1 positif diabetes, sebanyak 343 pasien tidak diabetes dan 169 pasien diabetes.

Data selanjutnya melalui proses *preprocessing* dengan dilakukan proses penyeimbangan data atau balancing data agar data yang akan diujikan memiliki jumlah yang seimbang antara *value* 0 dengan 1. Lalu dilakukan *scalling* data pada dataset agar memiliki cakupan nilai dengan skala yang seimbang antara atribut satu dengan yang lain. Lalu. *Preprocecing* bertujuan agar data dapat memperoleh hasil yang baik.

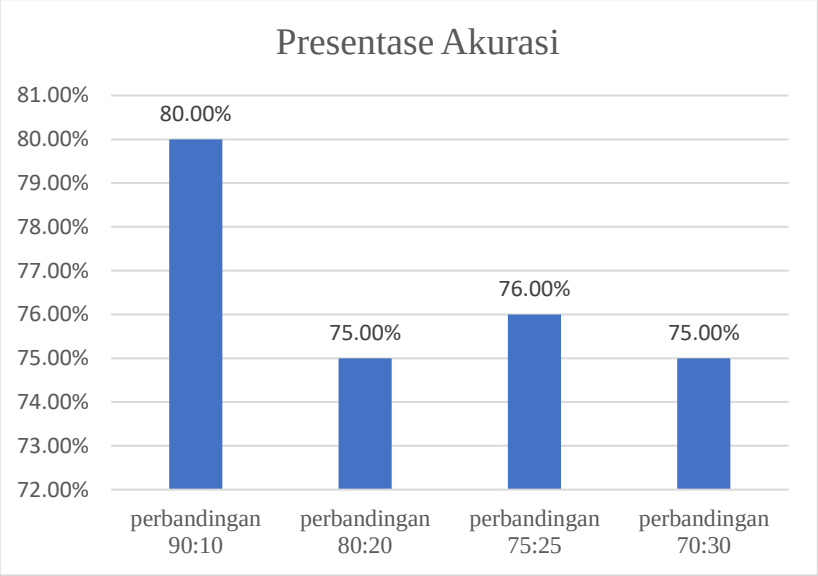
Selanjutnya data dipartisi menjadi empat perbandingan perbandingan pengujian, yaitu perbandingan perbandingan data latih 90%: data uji 10%, 80% data latih dan data uji 20%, 75% data latih dan data uji 25%, serta 70% data latih dan data uji 30% adalah proporsi rasio pembanding. Menggunakan *Naïve Bayes* untuk

pemodelan adalah langkah selanjutnya. Untuk mendapatkan hasil yang tepat, pengukuran dilakukan dengan menggunakan matriks konfusi. Tabel 4.20 menunjukkan variasi nilai akurasi prediksi yang dihasilkan untuk setiap model.

Tabel 4. 20 Perbandingan Hasil Akurasi Tiap Pengujian

Rasio Perbandingan	Jumlah Data = 512				Akurasi (%)
	Data Latih		Data Uji		
	Jumlah	Presentase	Jumlah	Presentase	
90 : 10	617	90%	69	10%	80%
80 : 20	548	80%	138	20%	75%
75 : 25	514	75%	172	25%	76%
70 : 30	480	70%	206	30%	75%

Nilai dari akurasi terbaik yang menampilkan representasi dari nilai hasil 80% diperoleh dari pengujian dengan menggunakan perbandingan 90% data latih (617 data) dan 10% data uji (69 data), seperti pada gambar pada tabel 4.16. Keempat nilai akurasi tersebut ditunjukkan sebagai berikut dalam diagram batang pada gambar 4.14.



Gambar 4. 14 Perbandingan Nilai Akurasi Tiap Pengujian

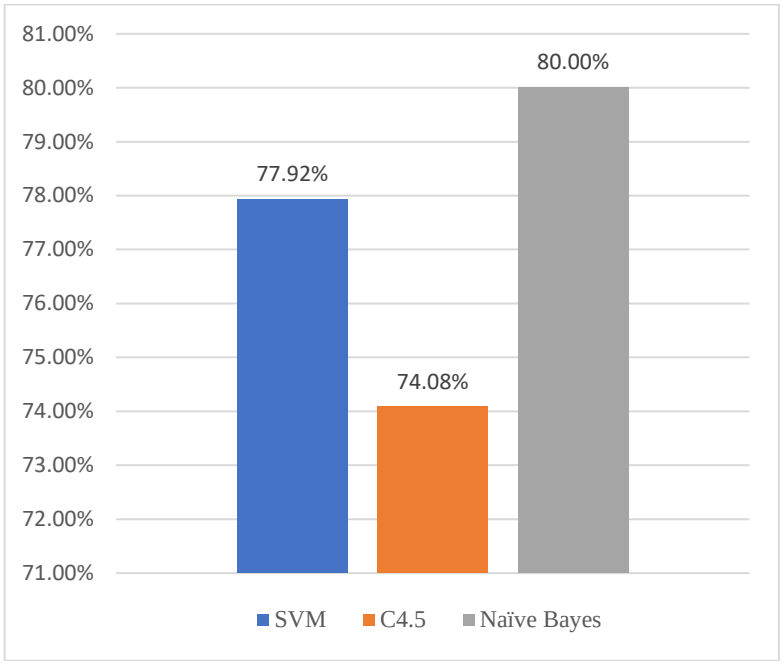
Menurut Gorunescu (2011), rentang nilai 90%-100% tergolong dalam klasifikasi sangat baik, 80%-90% dikategorikan baik, 70%-80% dianggap cukup, 60%-70% tergolong kurang baik, dan 50%-60% masuk dalam kategori gagal. Klasifikasi ini digunakan untuk menilai akurasi. Dalam pengujian, rasio 90:10 dengan akurasi 80% diklasifikasikan sebagai baik, rasio 80:20 dengan akurasi 75% juga cukup, rasio 75:25 dengan akurasi 76% masuk dalam kategori cukup, dan rasio 70:30 dengan akurasi 75% juga diklasifikasikan sebagai cukup.

Dari penelitian lain yang telah dilakukan oleh Agatsa et al (2020), dengan pengujian dari dataset sama Pima Indians Diabetes Database. Dalam penelitiannya menggunakan tahap *preprocessing* data dengan *min-max normalization*. Pembagian *split* data yang digunakan hanya 80% data latih dan 20 % data uji dengan memakai algoritma klasifikasi yaitu teknik metode *support vector machine*. Selain itu menggunakan validasi silang untuk menguji model dengan nilai *k*=5. Dari metode SVM menghasilkan nilai ketepatan prediksi akurasi sebesar 77,92%. Sedangkan pada penelitian yang dilakukan oleh penulis yang sama-sama menggunakan data Pima Indians Diabetes Database. Lalu, pada tahap *preprocessing* data menggunakan eliminasi data untuk mengatasi *missing value*, SMOTE untuk mengatasi ketidakseimbangan data dan *scaling* data agar rentang nilai memiliki ukuran yang tidak berbeda jauh. Data dibagi menjadi empat rasio perbandingan seperti 90:10, 80:20, 75:25, dan 70:30. Untuk pengujian model menggunakan validasi silang *k-fold cross validation* dengan iterasi diujikan sepuluh kali. Metode yang digunakan untuk pemodelan yaitu *Naïve Bayes classifier*. Dari hasil penelitian menghasilkan nilai akurasi 75% saat pembagian data yang sama

yaitu 80:20. Sedangkan hasil akurasi tertinggi sebesar 80% dengan rasio perbandingan 90:10.

Selain itu penelitian lain yang telah dilakukan oleh Robbani et al (2020), dengan pengujian dari data yang sama yaitu Pima Indians Diabetes Database. Tidak terdapat informasi mengenai *preprocessing* data maupun pembagian data yang dilakukan, sehingga data langsung dilakukan pemodelan dan pengujian. Penelitian tersebut menggunakan bagian dari algoritma klasifikasi dengan metode Algoritma C4.5. Dari metode C4.5 menghasilkan nilai ketepatan prediksi akurasi sebesar 74.08%. Sedangkan pada penelitian yang dilakukan oleh penulis sama-sama menggunakan data Pima Indians Diabetes Database. Lalu, pada tahap *preprocessing* data menggunakan eliminasi data untuk mengatasi *missing value*, SMOTE untuk mengatasi ketidakseimbangan data dan *scaling* data agar rentang nilai memiliki ukuran yang tidak berbeda jauh. Data dibagi menjadi empat rasio perbandingan seperti 90:10, 80:20, 75:25, dan 70:30. Untuk pengujian model menggunakan validasi silang *k-fold cross validation* dengan iterasi diujikan sepuluh kali. Metode yang digunakan untuk pemodelan yaitu *Naïve Bayes classifier*. Dari hasil penelitian menghasilkan nilai tertinggi sebesar 80% dengan rasio perbandingan 90:10.

Sehingga telah diketahui bahwa metode atau teknik algoritma *Naïve Bayes* yang digunakan memiliki nilai dari akurasi yang tertinggi dibandingkan metode yang lain terhadap Pima Indians Diabetes Database yang dapat ditampilkan dalam grafik pada gambar 4.15 sebagai berikut.



Gambar 4. 15 Perbedaan Hasil Prediksi Akurasi Antar Metode

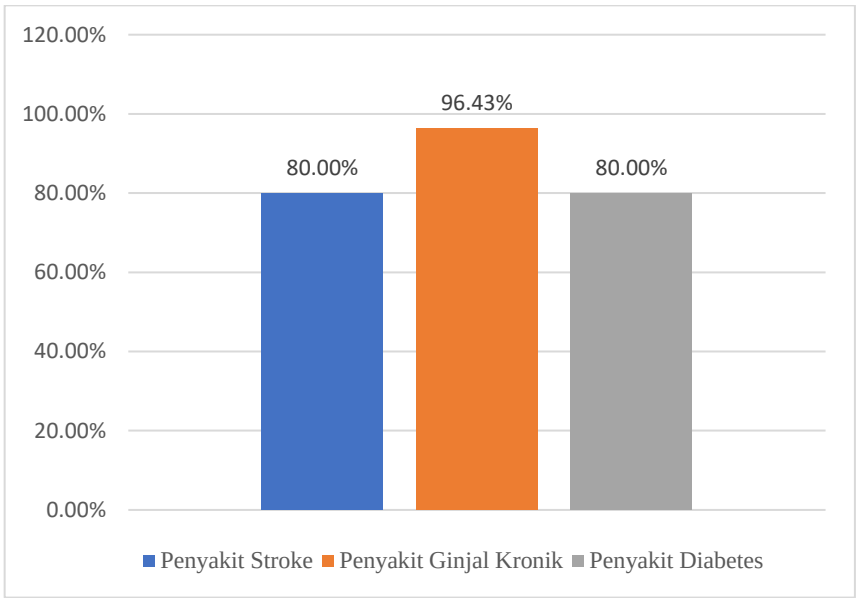
Dari penelitian lain yang telah dilakukan oleh Paramitha et al (2023), dengan pengujian dari dataset Pasien perempuan dan laki-laki Brain Stroke Prediction Dataset dengan menggunakan metode yang sama yaitu *Naïve Bayes*. Data pada tahap *preprocessing* dilakukan perlakuan pengecekan terhadap data yang hilang atau *missing value*, data duplikat, transformasi bentuk data dari data numerik menjadi kategori. Pembagian data yang yang digunakan dengan beberapa rasio perbandingan diantaranya 90:10, 80:20, 70:30, dan 60:40. Penelitian tersebut menghasilkan nilai prediksi akurasi tertinggi sebesar 80% pada proporsi 80:20 dan sebesar 75% pada proporsi data yang sama yaitu 90:10. Sedangkan pada penelitian yang dilakukan oleh penulis sama-sama menggunakan metode *Naïve Bayes classifier*. Namun penggunaan data berbeda yaitu data Pima Indians Diabetes Database. Lalu, pada tahap *preprocessing* data menggunakan eliminasi data untuk mengatasi *missing value*, SMOTE untuk mengatasi ketidakseimbangan data dan

scaling data agar rentang nilai memiliki ukuran yang tidak berbeda jauh. Data dibagi menjadi empat rasio perbandingan seperti 90:10, 80:20, 75:25, dan 70:30. Untuk pengujian model menggunakan validasi silang *k-fold cross validation* dengan iterasi diujikan sepuluh kali. Metode yang digunakan untuk pemodelan yaitu *Naïve Bayes classifier*. Dari hasil penelitian menghasilkan nilai tertinggi sebesar 80% dengan rasio perbandingan 90:10.

Selain itu penelitian lain yang telah dilakukan oleh A'yuniyah, (2022) dengan pengujian dari dataset Penyakit ginjal kronik (PGK) atau Chronic Kidney Disease Dataset dengan menggunakan metode yang sama yaitu *Naïve Bayes classifier*. Data diolah melalui tahap *preprocessing* dengan data cleaning untuk menghilangkan noise seperti *missing value, invalid*, atau salah ketik yang juga memberikan pengaruh pengurangan jumlah data dari 400 data menjadi 280 data. Setelah dilakukan *cleaning* data diproses transformasi data mengubah data menjadi bentuk lebih sederhana. Penelitian tersebut menghasilkan nilai prediksi akurasi sebesar 96.43% dengan proporsi data 70:30. Sedangkan pada penelitian yang dilakukan oleh penulis sama-sama menggunakan metode *Naïve Bayes classifier*. Namun penggunaan data berbeda yaitu data Pima Indians Diabetes Database. Lalu, pada tahap preprocessing data menggunakan eliminasi data untuk mengatasi *missing value*, SMOTE untuk mengatasi ketidakseimbangan data dan *scaling* data agar rentang nilai memiliki ukuran yang tidak berbeda jauh. Data dibagi menjadi empat rasio perbandingan seperti 90:10, 80:20, 75:25, dan 70:30. Untuk pengujian model menggunakan validasi silang *k-fold cross validation* dengan iterasi diujikan sepuluh kali. Metode yang digunakan untuk pemodelan yaitu *Naïve Bayes classifier*. Pada

propordi yang sama 70:30 menghasilkan 76% nilai akurasi dan hasil penelitian menghasilkan nilai tertinggi sebesar 80% dengan rasio perbandingan 90:10.

Sehingga telah diketahui bahwa metode *Naïve Bayes* memiliki nilai akurasi yang lebih tinggi terhadap Penyakit ginjal kronik (PGK) atau Chronic Kidney Disease Dataset sebesar 96.43% daripada dataset penyakit lain yang dapat ditampilkan dalam grafik pada gambar 4.16. Terdapat hal-hal yang bisa memberikan pengaruh dari perolehan hasil akurasi yaitu dari tahap proses *preprocessing* yang belum cukup optimal sehingga dan kualitas dari data itu sendiri. Dikarenakan *Naïve Bayes* mampu melakukan klasifikasi dengan kategori sangat baik dengan rentan presentase 90-100% terhadap penyakit ginjal kronik seperti yang ditampilkan gambar 4.16.



Gambar 4. 16 Perbandingan Akurasi Dari 3 Dataset Dengan Naïve Bayes

Selain itu penelitian ini juga telah melakukan beberapa uji coba lain. Percobaan pertama untuk menguji performa akurasi menggunakan Algoritma *Naïve Bayes* menggunakan dataset Pima Indians Diabetes Database dengan jumlah pasien

768 data. Pendekatan *preprocessing* yang dilakukan yaitu KNN *imputation* untuk menangani *missing value,* SMOTE untuk menangani kelas tidak seimbang dan *scaling* data untuk penskalaan. Menguji rasio perbandingan 90:10, 80:20, 75:25, dan 70:30 antara data pelatihan dan pengujian. Pengujian 90 : 10 menghasilkan nilai akurasi sebesar 74%, pengujian 80 : 20 menghasilkan nilai akurasi sebesar 73%, pengujian 75 : 25 menghasilkan nilai akurasi sebesar 72.80%, dan pengujian 70 : 30 menghasilkan nilai akurasi sebesar 72%. Nilai performa didapatkan tertinggi pada pengujian 90 : 10 sebesar 74%.

Percobaan kedua untuk menguji performa akurasi menggunakan algoritma *Naïve Bayes* menggunakan dataset Pima Indians Diabetes Database dengan jumlah pasien 768 data. Pendekatan *preprocessing* yang dilakukan yaitu *mean imputation* untuk menangani *missing value*, SMOTE untuk menangani kelas tidak seimbang dan *scaling* data untuk penskalaan. Rasio perbandingan data uji dan data latih menggunakan 90:10, 80:20, 75:25, dan 70:30. Pengujian 90 : 10 menghasilkan nilai akurasi sebesar 75%, pengujian 80 : 20 menghasilkan nilai akurasi sebesar 75%, pengujian 75 : 25 menghasilkan nilai akurasi sebesar 72.80%, dan pengujian 70 : 30 menghasilkan nilai akurasi sebesar 73%. Nilai performa didapatkan tertinggi pada pengujian 90 : 10 dan 80 : 20 sebesar 75%.

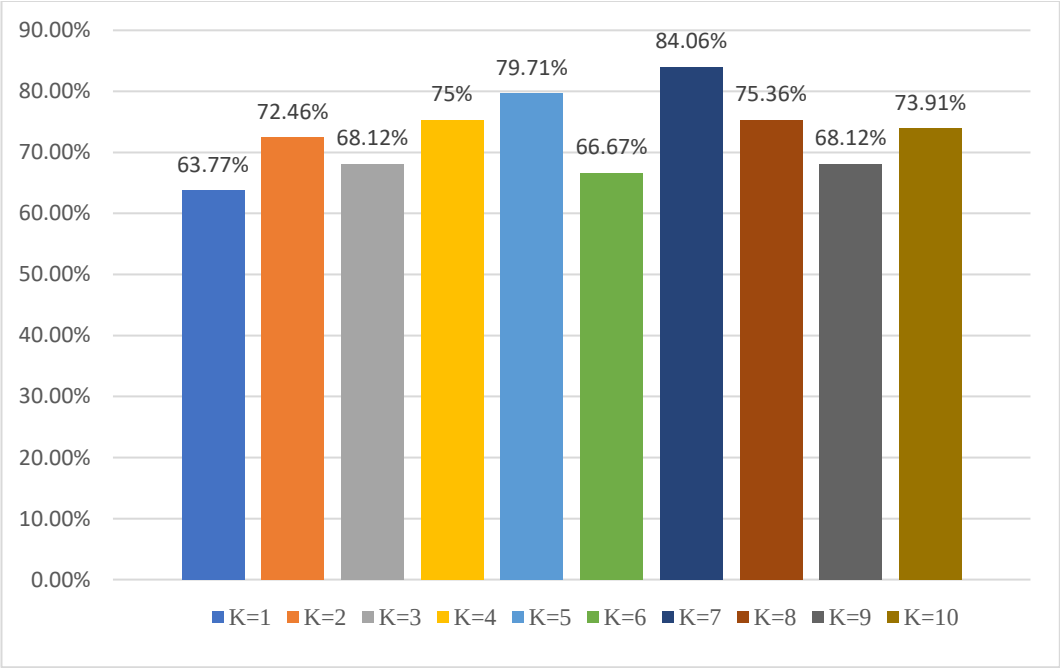
Percobaan ketiga untuk menguji performa akurasi menggunakan algoritma *Naïve Bayes* menggunakan dataset Pima Indians Diabetes Database dengan jumlah pasien 768 data. Pendekatan preprocessing yang dilakukan yaitu *median imputation* untuk menangani *missing value*, SMOTE untuk menangani kelas tidak seimbang dan *scaling* data untuk penskalaan. Rasio perbandingan data uji dan data latih

menggunakan 90:10, 80:20, 75:25, dan 70:30. Pengujian 90 : 10 menghasilkan nilai akurasi sebesar 73%, pengujian 80 : 20 menghasilkan nilai akurasi sebesar 75.50%, pengujian 75 : 25 menghasilkan nilai akurasi sebesar 74.40%, dan pengujian 70 : 30 menghasilkan nilai akurasi sebesar 72.67%. Nilai performa didapatkan tertinggi pada pengujian 80 : 20 sebesar 75.50%.

Percobaan keempat untuk menguji performa akurasi menggunakan algoritma *Naïve Bayes* menggunakan dataset Pima Indians Diabetes Database dengan jumlah pasien 768 data. Pendekatan *preprocessing* yang dilakukan yaitu *mode imputation* untuk menangani *missing value*, SMOTE untuk menangani kelas tidak seimbang dan *scaling* data untuk penskalaan. Rasio perbandingan data uji dan data latih menggunakan 90:10, 80:20, 75:25, dan 70:30. Pengujian 90 : 10 menghasilkan nilai akurasi sebesar 71%, pengujian 80 : 20 menghasilkan nilai akurasi sebesar 74%, pengujian 75 : 25 menghasilkan nilai akurasi sebesar 73.60%, dan pengujian 70 : 30 menghasilkan nilai akurasi sebesar 71.67%. Nilai performa didapatkan tertinggi pada pengujian 80 : 20 sebesar 74%.

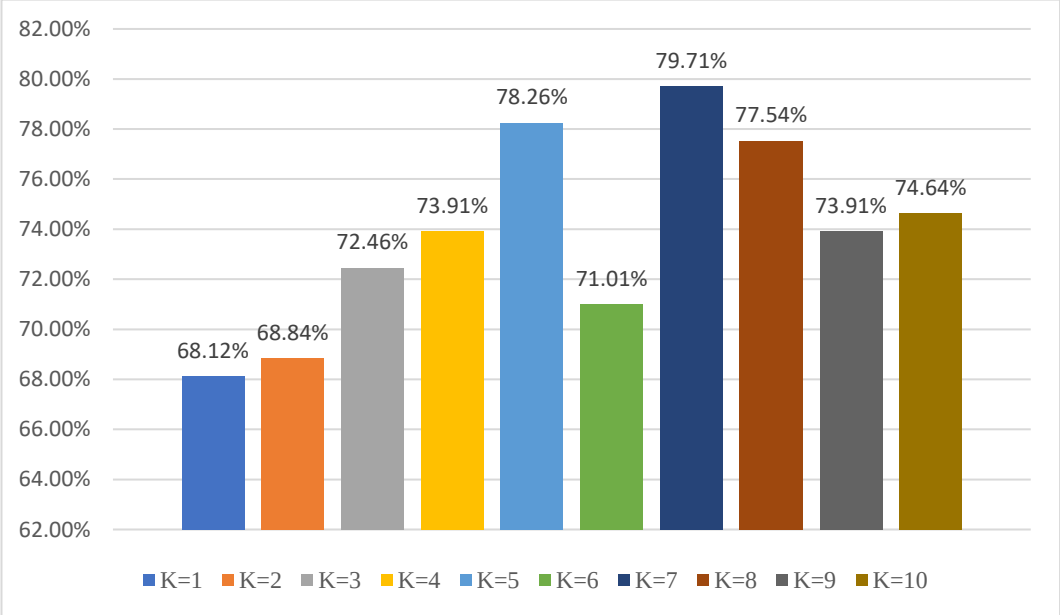
Masing-masing dari percobaan lain yang dilakukan terhadap klasifikasi Pima Indians Diabetes Database menggunakan algoritma *Naïve Bayes* memiliki hasil performa akurasi yang berbeda. Perbedaan terletak pada penanganan *missing value* data. Didapatkan hasil akurasi tertinggi pada penelitian yang melakukan perlakuan eliminasi data pada Pima Indians Diabetes Database sebesar 79.71%.

Pengujian evaluasi algoritma *Naïve Bayes* pada *10-fold cross-validation* dengan rasio perbandingan 90:10 menggunakan parameter *k* dari 1 hingga 10, memberikan berbagai nilai akurasi. Secara rinci, akurasi yang diperoleh adalah: *k*=1 sebesar 63.77%, *k*=2 sebesar 72.46%, *k*=3 sebesar 68.12%, *k*=4 sebesar 75.36%, *k*=5 sebesar 79.71%, *k*=6 sebesar 66.67%, *k*=7 sebesar 84.06%, *k*=8 sebesar 75.36%, *k*=9 sebesar 68.12%, dan *k*=10 sebesar 73.91%. Dari hasil ini, terlihat bahwa akurasi tertinggi dicapai pada *k*=7 dengan nilai 84.06%. Hasil ini ditampilkan secara visual dalam diagram batang pada gambar 4.17, yang dengan jelas memperlihatkan bagaimana performa algoritma *Naïve Bayes* bervariasi dengan perubahan nilai *k*. Penggunaan 10 nilai *k* ini membantu dalam mengidentifikasi parameter optimal untuk mendapatkan hasil akurasi terbaik dalam pengujian model.



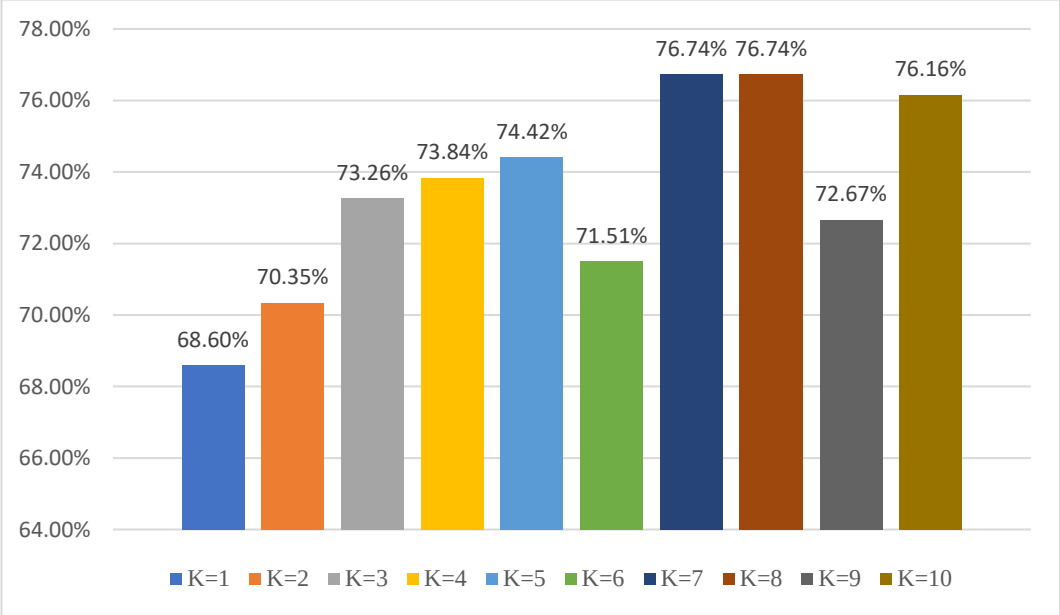
Gambar 4. 17 Perbandingan 10Fold Cross Validation Pengujian 90 : 10

Hasil penggunaan validasi silang 10 kali untuk menilai kinerja algoritma *Naïve Bayes* dalam mengevaluasi perbandingan data 80:20 dengan parameter *k* dari 1 hingga 10 menunjukkan variasi nilai akurasi yang signifikan. Secara terperinci, akurasi yang diperoleh adalah *k*=1 sebesar 68.12%, *k*=2 sebesar 68.84%, *k*=3 sebesar 72.46%, *k*=4 sebesar 73.91%, *k*=5 sebesar 78.26%, *k*=6 sebesar 71.01%, *k*=7 sebesar 79.71%, *k*=8 sebesar 77.54%, *k*=9 sebesar 73.91%, dan *k*=10 sebesar 74.64%. Dari analisis ini, terlihat bahwa akurasi tertinggi dicapai pada *k*=7 dengan nilai 79.71%. Hasil evaluasi ini jika divisualisasikan dalam bentuk diagram batang, seperti yang ditampilkan pada gambar 4.18, menunjukkan secara jelas bagaimana performa algoritma *Naïve Bayes* bervariasi seiring perubahan nilai *k*. Penggunaan 10 nilai *k* ini membantu mengidentifikasi parameter optimal untuk mencapai akurasi terbaik dalam pengujian model, memastikan bahwa model bekerja secara efektif dalam kondisi pengujian yang berbeda.



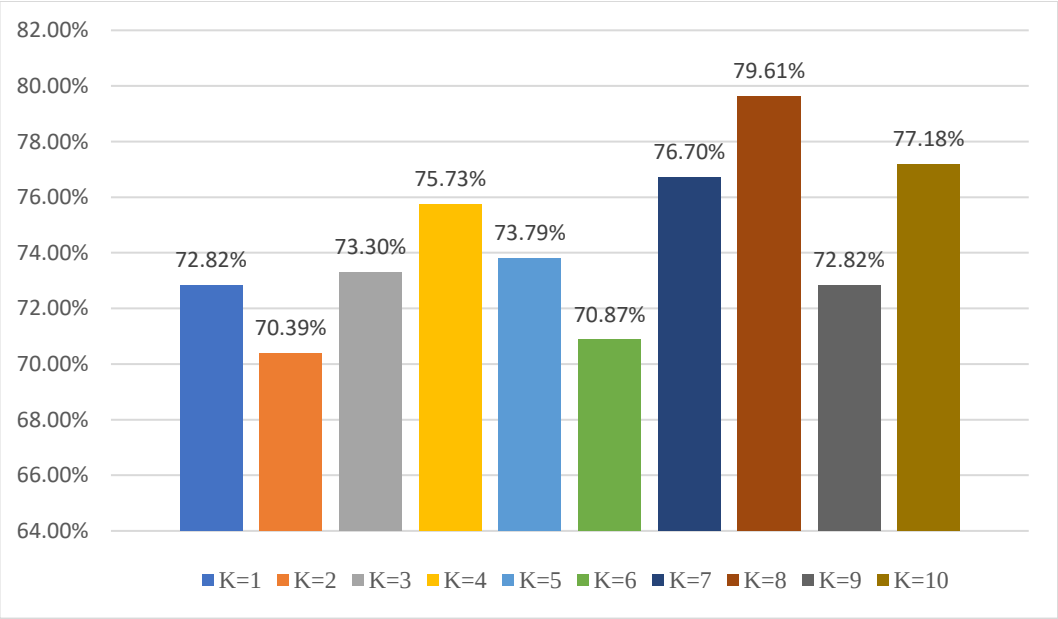
Gambar 4. 18 Perbandingan 10Fold Cross Validation Pengujian 80 : 20

Hasil pengujian evaluasi *10-fold cross validation* menggunakan algoritma *Naïve Bayes* pada pengujian dengan rasio data 75:25 menunjukkan variasi akurasi yang menarik untuk nilai input *k* dari 1 hingga 10. Secara rinci, nilai akurasi yang diperoleh adalah *k*=1 sebesar 68.60%, *k*=2 sebesar 70.35%, *k*=3 sebesar 73.26%, *k*=4 sebesar 73.84%, *k*=5 sebesar 74.42%, *k*=6 sebesar 71.51%, *k*=7 sebesar 76.74%, *k*=8 sebesar 76.74%, *k*=9 sebesar 72.67%, dan *k*=10 sebesar 76.16%. Dari hasil ini, terlihat bahwa akurasi tertinggi dicapai pada parameter *k*=7 dan *k*=8, keduanya dengan nilai 76.74%. Jika 10 nilai k ini divisualisasikan dalam bentuk diagram batang, seperti yang ditampilkan pada gambar 4.19, hasilnya menunjukkan bahwa *k*=7 dan *k*=8 memberikan performa terbaik. Analisis ini membantu mengidentifikasi nilai *k* yang optimal untuk mencapai akurasi maksimal, sehingga dapat memberikan panduan yang lebih baik dalam pemilihan parameter untuk meningkatkan kinerja model pada dataset yang digunakan.



Gambar 4. 19 Perbandingan 10Fold Cross Validation Pengujian 75 : 25

Sedangkan hasil pengujian menggunakan rasio data 70:30 dengan nilai *k* dari 1 hingga 10 melalui evaluasi metode *10-fold cross validation* dari algoritma *Naïve Bayes* menunjukkan variasi nilai akurasi yang signifikan. Secara rinci, akurasi yang diperoleh adalah: *k*=1 sebesar 72.82%, *k*=2 sebesar 70.39%, *k*=3 sebesar 73.30%, *k*=4 sebesar 75.73%, *k*=5 sebesar 73.79%, *k*=6 sebesar 70.87%, *k*=7 sebesar 76.70%, *k*=8 sebesar 79.61%, *k*=9 sebesar 72.82%, dan *k*=10 sebesar 77.18%. Jika hasil akurasi ini divisualisasikan dalam bentuk diagram batang, seperti yang ditampilkan pada gambar 4.20, terlihat bahwa parameter *k*=8 memberikan hasil akurasi terbaik yaitu sebesar 79.61%. Analisis ini menunjukkan bahwa nilai *k*=8 adalah optimal dalam mencapai akurasi tertinggi di antara 10 nilai *k* yang diuji, memberikan panduan yang jelas dalam memilih parameter untuk meningkatkan kinerja model *Naïve Bayes* pada dataset yang digunakan.



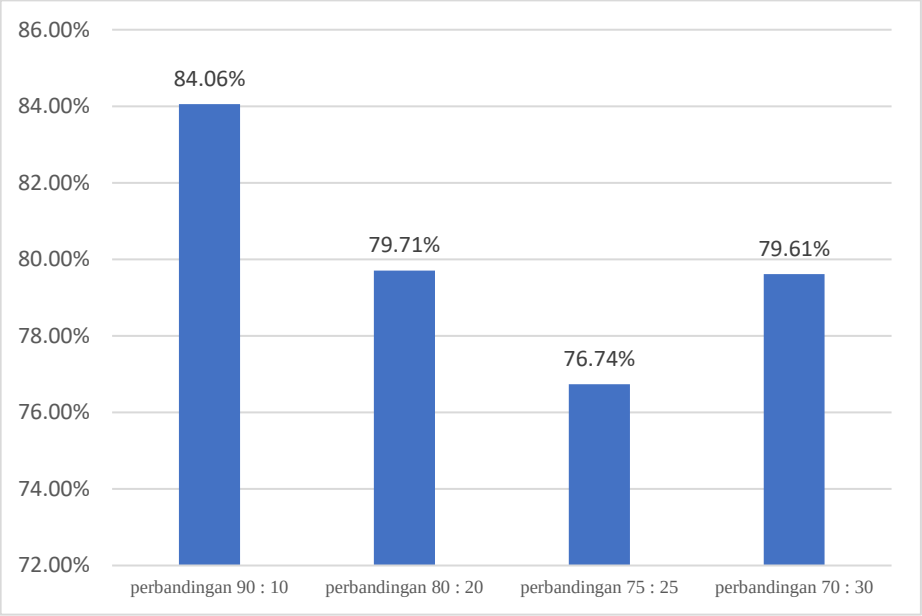
Gambar 4. 20 Perbandingan 10Fold Cross Validation Pengujian 70 : 30

Dari keseluruhan perbandingan pengujian dengan rasio 90:10, 80:20, 75:25, dan 70:30, diketahui bahwa nilai akurasi tertinggi diperoleh dari pengujian dengan rasio 90:10. Dalam pengujian ini, data latih sebanyak 90% atau 617 data digunakan, sementara data uji sebanyak 10% atau 69 data. Pengujian ini menghasilkan nilai akurasi tertinggi dengan parameter *k*=9, mencapai akurasi sebesar 84.06%, yang ditandai dengan label warna kuning. Hasil ini ditampilkan dalam tabel 4.21, menunjukkan bahwa rasio perbandingan 90:10 dengan parameter *k*=9 memberikan performa terbaik dibandingkan dengan rasio perbandingan lainnya. Analisis ini memberikan wawasan penting tentang pengaruh rasio data latih dan uji serta parameter *k* terhadap akurasi model, membantu dalam memilih konfigurasi optimal untuk meningkatkan kinerja algoritma *Naïve Bayes.*

Tabel 4. 21 Perbandingan 10 Fold Cross Validation Tiap Pengujian

Parameter K	Akurasi (%)			
	90 : 10	80 : 20	75 : 25	70 : 30
K = 1	63.77%	68.12%	68.60%	72.82%
K = 2	72.46%	68.84%	70.35%	70.39%
K = 3	68.12%	72.46%	73.26%	73.30%
K = 4	75.36%	73.91%	73.84%	75.73%
K = 5	79.71%	78.26%	74.42%	73.79%
K = 6	66.67%	71.01%	71.51%	70.87%
K = 7	84.06%	79.71%	76.74%	76.70%
K = 8	75.36%	77.54%	76.74%	79.61%
K = 9	68.12%	73.91%	72.67%	72.82%
K = 10	73.91%	74.64%	76.16%	77.18%

Peringkasan hasil *10-fold cross validation* pada perbandingan pengujian dengan rasio 90 : 10, 80 : 20, 75 : 25 dan 70 : 30 ditunjukkan pada gambar 4.21 dibawah ini.



Gambar 4. 21 Perbandingan 10Fold Cross Validation Tiap Pengujian

Dari gambar 4.21 yang menunjukkan nilai-nilai akurasi tertinggi yang didapatkan dari uji coba tiap pengujian, mendapatkan hasil tertinggi akurasi pada pengujian dengan rasio perbandingan 90 : 10 sebesar 84.06%.

4.5 Integrasi Penelitian Dalam Tafsir Al-Qur'an

Setiap hal di dunia ini berkaitan dengan agama yang selalu menjadi poros dalam kehidupan, khususnya agama islam yang berpedoman pada Al-Quran dan juga hadist Rasulullah Shalallaahu Alaihi Wassalaam. Akal sehat menuntun kita dalam menyikapi berbagai hal yang terjadi disekitar kita.

Timbulnya sebuah penyakit atas izin dari Allah, sehingga kesembuhan semua penyakit hanya berhak disembuhkan oleh Allah. Saat seseorang merasakan sakit,

Allah mencabut nikmat diberikan agar dapat menghargai betapa berharganya kesehatan yang ia miliki. Lalu setelah ia sembuh dapat belajar dari kesalahan yang ia perbuat dari melanggar norma-norma kesehatan sesuai anjuran yang berlaku. Ayat ini mengajarkan bahwa manusia harus berusaha menjaga keseimbangan, menerapkan gaya hidup sehat, dan mengambil tindakan pencegahan yang diperlukan. Dengan menjaga kesehatan, manusia dapat mengurangi risiko penyakit dan memperkuat ketahanan tubuh.

Nikmat yang Allah berikan sangat melimpah ada di sekitar kita. Meskipun begitu kita tidak boleh menggunakan nikmat tersebut secara berlebih-lebihan karena untuk mencegah atau menghindari dampak buruk yang akan timbul dari perilaku berlebih-lebihan. Karena kesehatan merupakan hal yang lebih berharga dari apapun tak ternilai harganya. Sehingga sebagai umat islam terdapat larangan untuk berperilaku berlebih-lebihan. Salah satu cara mencegah penyakit diabetes dengan mengonsumsi makan dan minum secukupnya untuk kesehatan dan kesejahteraan tubuh tercantum pada surah Al-A'raf ayat 31.

وَكُلُوا وَاشْرَبُوا وَلا تُسْرِفُوا إِنَّهُ لا يُحِبُّ الْمُسْرِفِينَ

"makan dan minumlah, tetapi jangan berlebihan. Sungguh, Allah tidak menyukai orang yang berlebih-lebihan". (QS. Al-A'raf : 31).

Ayat Al-Qur'an dari surah ini memberikan penjelasan bahwasannya Allah memerintahkan makan dan minum apa saja yang baik yang dihalalkan dan Allah sangat membenci perilaku hambanya yang berlebih-lebihan melampaui batasan yang wajar. Makanan dan minuman yang dikonsumsi lebih baik secukupnya sesuai dengan norma atau anjuran kesehatan yang ada. Karena jika berlebih-lebihan dapat

memberikan pengaruh buruk pada kesehatan sehingga menimbulkan penyakit pada tubuh salah satunya penyakit diabetes.

Terdapat juga sebuah hadits Rasulullah yang memberikan nasihat memanfaatkan lima perkara sebelum lima keadaan mengenai prinsip menjaga kesehatan dan mengambil tindakan pencegahan seperti hadits berikut.

اِغْتَنِمْ خَمْسًا قَبْلَ خَمْسٍ : شَبَابَكَ قَبْلَ هَرَمِكَ وَ صِحَّتَكَ قَبْلَ سَقَمِكَ وَ غِنَاكَ قَبْلَ فَقْرِكَ وَ فَرَاغَكَ قَبْلَ شَغْلِكَ وَ حَيَاتَكَ قَبْلَ مَوْتِكَ

"Manfaatkanlah lima perkara sebelum lima perkara, waktu mudamu sebelum datang waktu tuamu, waktu sehatmu sebelum waktu sakitmu, masa kayamu sebelum datang masa kefakiranmu, masa luangmu sebelum datang masa sibukmu, dan hidupmu sebelum datang matimu." (HR. Al Hakim dalam Al Mustadroknya 4: 341).

Hadist tersebut menekankan pentingnya memanfaatkan nikmat-nikmat yang diberikan oleh Allah dengan sebaik mungkin untuk menginvestasikan diri taat melaksanakan perintah dan menjahui larangan-Nya.

Sebuah penyakit yang dialami oleh manusia, tidak terlepas dari berbagai faktor yang mempengaruhinya. Diantaranya karena pola hidup yang kurang seimbang dan memiliki keterkaitan dengan genetik keturunan dalam keluarga. Dalam menyikapi penyakit yang dialami perlu diyakini bahwa akan memiliki penawar sebagai penyembuh dari sakit untuk kembali sehat. Seperti yang tercantum pada ayat ke-80 dalam Surah Asy-Syuara bahwa Allah adalah Penyembuh sejati yang memiliki kekuasaan untuk menyembuhkan penyakit apa pun.

وَاِذَا مَرِضْتُ فَهُوَ يَشْفِيْنِ

"Dan apabila aku sakit, Dialah yang menyembuhkan aku". (QS. Asy Syuara : 80).

Ayat surah diatas memberikan penjelasan bahwa Allah menyembuhkan seseorang jika mengalami sakit karena Allah berkuasa menyembuhkan semua penyakit. Meski demikian, manusia diminta untuk mencari cara untuk memperoleh kesembuhan. Selain karena genetic keluarga, timbulnya penyakit terkadang disebabkan dari faktor manusia yang kurang menjaga keseimbangan terhadap pola hidup karena melanggar anjuran kesehatan yang berlaku. Hal ini pasti dapat menimbulkan penyakit baik ringan maupun berat mulai dari jangka pendek hingga jangka panjang.

Akal sehat menuntun kita dalam menyikapi berbagai hal yang terjadi disekitar kita. Sehingga kita menemukan solusi untuk memudahkan dalam memyelesaikan urusan di dunia. Allah Subhanahu Wa Ta'ala telah menurunkan Al-Qur'an sebagai wahyu melalui Rasulullah Shalallaahu Alaihi Wassalaam, lalu disebar luaskan kepada umat islam di dunia agar dapat memberikan manfaat satu sama lain. Hal ini menuntun manusia untuk berusaha mencari pertolongan demi kesembuhan penyakit melalui rahmat Allah.

Seseorang yang memiliki penyakit khusus nya diabetes, harus segera berkonsultasi dan berobat kepada instansi kesehatan mulai dari klinik hingga rumah sakit. Hal ini agar mendapat pertolongan untuk proses penyembuhan atau pencegahan suatu penyakit agar tidak muncul dan menyebar sehingga menimbulkan penyakit lain. Oleh karena itu, tenaga medis maupun dokter dapat membantu dalam penanganan hal ini. Rasulullah Shallallahu 'Alaihi Wa Sallam dalam sabdanya menjelaskan bahwasannya semua penyakit pasti diciptakan pula obatnya.

حَدَّثَنَا هَارُونُ بْنُ مَعْرُوفٍ وَأَبُو الطَّاهِرِ وَأَحْمَدُ بْنُ عِيسَى قَالُوا حَدَّثَنَا ابْنُ وَهْبٍ أَخْبَرَنِي عَمْرٌو وَهُوَ ابْنُ الْحَارِثِ عَنْ عَبْدِ رَبِّهِ بْنِ سَعِيدٍ عَنْ أَبِي الزُّبَيْرِ عَنْ جَابِرٍ عَنْ رَسُولِ اللَّهِ صَلَّى اللَّهُ عَلَيْهِ وَسَلَّمَ أَنَّهُ قَالَ لِكُلِّدَاءٍ دَوَاءٌ فَإِذَاأُصِيبَ دَوَاءُ الدَّاءِ بَرَأَ بِإِذْنِ اللَّهِ عَزَّ وَجَلَّ

"Telah menceritakan kepada kami Harun bin Ma'ruf dan Abu Ath Thahir serta Ahmad bin 'Isa mereka berkata; Telah menceritakan kepada kami Ibnu Wahb; Telah mengabarkan kepadaku 'Amru, yaitu Ibnu al-Harits dari 'Abdu Rabbih bin Sa'id dari Abu Az Zubair dari Jabir dari Rasulullah shallallahu 'alaihi wasallam, beliau bersabda: "Setiap penyakit ada obatnya. Apabila ditemukan obat yang tepat untuk suatu penyakit, akan sembuhlah penyakit itu dengan izin Allah 'azza wajalla." (HR Muslim No.4084).

Dalam dunia medis, seseorang bisa dinyatakan diabetes bergantung pada sejumlah faktor yang telah diperhitungkan dan ditetapkan sebelumnya. Dalam konteksnya sebelum adanya sistem yang mempermudah para dokter untuk melakukan klasifikasi penyakit diabetes. Sebelumnya para dokter lebih mengalami kesulitan dalam melakukan klasifikasi diabetes. Sehingga hal tersebut berdampak dalam penanganan penyakit diabetes pada pasien. Dari permasalahan ini banyak penelitian yang sudah dilakukan dalam meningkatkan kualitas dalam dunia kesehatan khususnya dalam proses klasifikasi diabetes sehingga dapat dilakukan deteksi dengan cepat dan penanganan sejak dini. Hal ini menekankan bahwa perubahan positif dimulai dari usaha dan tindakan manusia itu sendiri. Sesuai dengan potongan QS. Ar-Ra'd ayat 11 yang berbunyi:

لَهُ مُعَقِّبَاتٌ مِنْ بَيْنِ يَدَيْهِ وَمِنْ خَلْفِهِ يَحْفَظُونَهُ مِنْ أَمْرِ اللَّهِ ۗ إِنَّ اللَّهَ لَا يُغَيِّرُ مَا بِقَوْمٍ حَتَّىٰ يُغَيِّرُوا مَا بِأَنْفُسِهِمْ ۗ وَإِذَا أَرَادَ اللَّهُ بِقَوْمٍ سُوءًا فَلَا مَرَدَّ لَهُ ۚ وَمَا لَهُمْ مِنْ دُونِهِ مِنْ وَالٍ

"Sesungguhnya Allah tidak mengubah keadaan sesuatu kaum sehingga mereka mengubah keadaan yang ada pada diri mereka sendiri" (QS. Ar-Ra'd: 11)

Ayat di atas menjelaskan bahwa Allah Subhanahu Wa Ta'ala tidak akan mengubah keadaan pada suatu kaum dari satu kondisi ke kondisi yang lain, sebelum mereka dapat mengubah keadaan diri yang menyangkut sikap, mental dan pemikiran mereka sendiri. Harapan besar dengan adanya sistem klasifikasi diabetes yang lebih cepat dan efisien sehingga dapat membantu tenaga kesehatan dalam mendeteksi diabetes, sehingga bisa diberikannya penanganan sejak dini dan dapat menekan angka kematian yang diakibatkan oleh diabetes.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Setelah dilakukan pengujian pada empat proporsi perbandingan dari data latih (*training*) dan data uji (*testing*) dengan rasio 90:10, 80:20 , 75:25, dan 70:30. Perbandingan 90% data latih dan 10% data uji menghasilkan performa model klasifikasi tertinggi dengan akurasi sebesar 80%, presisi pasien positif diabetes sebesar 88%, presisi pasien negatif diabetes sebesar 73%, recall pasien positif diabetes sebesar 74%, recall pasien negatif diabetes sebesar 87%, f1 score pasien positif diabetes sebesar 80%, dan f1 score pasien negatif diabetes sebesar 79%. Tingkat akurasi yang diperoleh menunjukkan kemampuan model dalam memperkirakan kelas target, dan menunjukkan bahwa akurasi antara data target dan prediksi masuk dalam kategori baik. Dalam proses *10-fold cross validation*, akurasi tertinggi didapatkan dengan parameter *k*=7, yaitu sebesar 84.06%.

5.2 Saran

Dari pengujian penelitian ini yang telah dilakukan, diharapkan pada penelitian selanjutnya dapat meningkatkan hasil klasifikasi dari pengujian model yang lebih optimal akurat. Dibawah ini merupakan saran yang diberikan peneliti dengan harapan dapat menjadi pendukung pada penelitian yang dilakukan yang selanjutnya :

1. Menambahkan jumlah dari data yang akan digunakan, karena jumlah data bisa memberikan pengaruh pada hasil klasifikasi model.

2. Menggunakan teknik atau metode untuk optimasi lainnya agar dapat memberikan hasil klasifikasi model yang maksimal.
3. Memerhatikan label dari kelas yang sama jumlah atau seimbang yang akan digunakan untuk penelitian, karena itu dapat memberikan pengaruh hasil klasifikasi model.
4. Menentukan bilangan *random number generator* (RNG) yang sesuai sebagai inisialisasi random_state karena dapat memengaruhi hasil klasifikasi model.

DAFTAR PUSTAKA

A'yuniyah, Q., Tasia, E., Nazira, N., Pratama, P. F., Anugrah, M. R., Adhiva, J., & Mustakim, M. (2022). Implementasi Algoritma Naïve Bayes Classifier (NBC) untuk Klasifikasi Penyakit Ginjal Kronik. *Jurnal Sistem Komputer Dan Informatika (JSON)*, *4*(1), 72. https://doi.org/10.30865/json.v4i1.4781

Abu Ahmad. (2017). Mengenal Artificial Intelligence, Machine Learning, & Deep Learning. *Jurnal Teknologi Indonesia*, *1*(June), 1–6. https://amt-it.com/mengenal-perbedaan-artificial-inteligence-machine-learning-deep-learning/

Agatsa, D. A., Rismala, R., & Wisesty, U. N. (2020). Klasifikasi Pasien Pengidap Diabetes menggunakan Metode Support Vector Machine. *E-Proceeding of Engineering*, *Vol.7*(No.1), 2517.

Aini, F. N., Wicaksana, A. L., & Pangastuti, H. S. (2020). Tingkat Risiko Kejadian Kardiovaskular pada Penyandang Diabetes Melitus Tipe 2. *Jurnal Persatuan Perawat Nasional Indonesia (JPPNI)*, *4*(3), 182. https://doi.org/10.32419/jppni.v4i3.191

Aryanti, R., Misriati, T., & Hidayat, R. (2023). Klasifikasi Risiko Kesehatan Ibu Hamil Menggunakan Random Oversampling Untuk Mengatasi Ketidakseimbangan Data. *KLIK: Kajian Ilmiah Informatika Dan Komputer*, *3*(5), 409–416. https://djournals.com/klik

Deliana, S. O., Hakim, A. L., Sari, E. O., Apriyanti, H., & Pauziah, S. (2023). Pemanfaatan Media Website "Mantes" untuk Meningkatkan Pengetahuan Masyarakat tentang Cek Kesehatan. *Jurnal Pengabdian Masyarakat Saga Komunitas*, *2*(2), 190–195. https://doi.org/10.53801/jpmsk.v2i2.103

Dinas Kesehatan Riau. (2018). *Profil Kesehatan Provinsi Riau*. https://dinkes.riau.go.id/profil-kesehatan-provinsi-riau

Felice, N., Johan, J., Natthannael, J., Gozal, M. B., Jovannie, C., & Anggreainy, M. susan. (2023). *Brain Stroke Prediction Using Random Forest Method with Tuning Parameter*. https://doi.org/https://doi.org/10.1109/AiDAS60501.2023.10284685

Gu, Q., Wang, X. M., Wu, Z., Ning, B., & Xin, C. S. (2016). An improved SMOTE algorithm based on genetic algorithm for imbalanced data classification. *Journal of Digital Information Management*, *14*(2), 92–103.

Hadna, N. M. S., Santosa, Insap, P., & Winarno, W. W. (2016). Studi Literatur Tentang Perbandingan Metode Untuk Proses Analisis Sentimen Di Twitter.

Seminar Nasional Teknologi Informasi Dan Komunikasi, *7*(2), 57–64.

Heranova, O. (2019). Synthetic Minority Oversampling Technique pada Averaged One Dependence Estimators untuk Klasifikasi Credit Scoring. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *3*(3), 443–450. https://doi.org/10.29207/resti.v3i3.1275

Hermayanti, D., & Nursiloningrum, E. (2017). *Diabetes mellitus (DM) merupakan penyakit disebabkan berkurangnya produksi atau kerja autoimun selektif terhadap sel beta pankreas . Tipe berhubungan dengan banyak gen . Antibodi terhadap gen , disebut diabetes monogenik . Diabetes ataupun resesif , ata*. *Volume 13*(Nomor 1), 25–30.

Indrayanti, Sugianti, D., & Karomi, M. A. Al. (2017). OPTIMASI PARAMETER K PADA ALGORITMA K-NEAREST NEIGHBOUR UNTUK KLASIFIKASI PENYAKIT DIABETES MELLITUS. *Prosiding SNATIF Ke-4 Tahun 2017*, 153–160.

International Diabetes Foundation. (2021). *IDF Diabetes Atlas Tenth Edition 2021*.

Kusuma, P. D. (2020). *Machine Learning Teori, Program, Dan Studi Kasus*. https://books.google.com/books?hl=id&lr=&id=4k3sDwAAQBAJ&oi=fnd&pg=PP1&dq=machine+learning+adalah&ots=E_yn2LRYLL&sig=qr-UdXBTdEazEaksTUgPyUGuqqw

Magnolia, C., Nurhopipah, A., & Kusuma, B. A. (2023). Penanganan Imbalanced Dataset untuk Klasifikasi Komentar Program Kampus Merdeka Pada Aplikasi Twitter. *Edu Komputika Journal*, *9*(2), 105–113. https://doi.org/10.15294/edukomputika.v9i2.61854

Octaviary, S. R. (2022). *DETEKSI AWAL PENYAKIT GAGAL JANTUNG BERDASARKAN FAKTOR RISIKO MENGGUNAKAN METODE NAIVE BAYES*. http://etheses.uin-malang.ac.id/41215/

Paramitha, Y. N., Nuryaman, A., Faisol, A., Setiawan, E., & Nurvazly, D. E. (2023). Klasifikasi Penyakit Stroke Menggunakan Metode Naïve Bayes. *Jurnal Siger Matematika*, *04*(01), 11–16. https://www.kaggle.com/datasets/zzettrkalpakbal/full-filled-

Putra, D., & Wibowo, A. (2020). Prediksi Keputusan Minat Penjurusan Siswa SMA Yadika 5 Menggunakan Algoritma Naïve Bayes. *Prosiding Seminar Nasional Riset Dan Information Science (SENARIS)*, *2*, 84–92.

Rachmawani, N. R., & Oktarlina, R. Z. (2017). Khasiat Pemberian Buncis (Phaseolus vulgaris L.) sebagai Terapi Alternatif Diabetes Melitus Tipe 2. *Jurnal Majority*, *8*(2), 145–150.

Rahmat, B., Agidtama Gafar, A., Fajriani, N., Ramdani, U., Rihin Uyun, F., Purnamasari P., Y., & Ransi, N. (2017). Implemetasi k-means clustering pada rapidminer untuk analisis daerah rawan kecelakaan. *Seminar Nasional Riset Kuantitatif Terapan 2017*, *April*, 58–60. https://ojs.innov-center.org/index.php/snrkt2017/article/download/10/9

Ramadhan, N. G. (2021). Comparative Analysis of ADASYN-SVM and SMOTE-SVM Methods on the Detection of Type 2 Diabetes Mellitus. *Scientific Journal of Informatics*, *8*(2), 276–282. https://doi.org/10.15294/sji.v8i2.32484

Robbani, A. A., Siregar, A. M., & Kusumaningrum, D. S. (2022). Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma C4.5. *Scientific Student Journal for Information, Technology and Science, Vol. III*(No: 1).

Roihan, A., Sunarya, P. A., & Rafika, A. S. (2020). Pemanfaatan Machine Learning dalam Berbagai Bidang: Review paper. *IJCIT (Indonesian Journal on Computer and Information Technology)*, *5*(1), 75–82. https://doi.org/10.31294/ijcit.v5i1.7951

Sholihah, N. N., & Hermawan, A. (2023). Implementation of Random Forest and Smote Methods for Economic Status Classification in Cirebon City. *Jurnal Teknik Informatika (Jutif)*, *4*(6), 1387–1397. https://doi.org/10.52436/1.jutif.2023.4.6.1135

Sihombing, J. (2021). Klasifikasi Data Antroprometri Individu Menggunakan Algoritma Naïve Bayes Classifier. *BIOS : Jurnal Teknologi Informasi Dan Rekayasa Komputer*, *2*(1), 1–10. https://doi.org/10.37148/bios.v2i1.15

Sinaga, S., Sembiring, R. W., & Sumarno, S. (2022). Penerapan Algoritma Naive Bayes untuk Klasifikasi Prediksi Penerimaan Siswa Baru. *Journal of Machine …*, *1*(1), 55–64. https://journal.fkpt.org/index.php/malda/article/view/162%0Ahttps://journal.fkpt.org/index.php/malda/article/download/162/115

Suardana, I. K., Rasdini, I. G. A. A., & Kusmarjathi, N. K. (2015). *HUBUNGAN DUKUNGAN SOSIAL KELUARGA DENGAN KUALITAS HIDUP PASIEN DIABETES MELLITUS TIPE II DI PUSKESMAS IV DENPASAR SELATAN*. *Volume 12*(Nomor 1).

Sudaryanto, A., Setiadi, N. A., & Frankilawati, D. A. (2014). tipe II adalah kombinasi akibat antara jaringan tubuh yang mengalami resistansi terhadap aksi insulin dan ketidakmampuan pankreas untuk menghasilkan cukup insulin ekstra untuk mengatasi kondisi tersebut (Bryer , 2012). Diabetes melitus tipe II merupakan. *Jurnal Kesehatan Vokasional*, 19–24.

Sulistiyowati, N., & Jajuli, M. (2020). Integrasi Naive Bayes Dengan Teknik

Sampling Smote Untuk Menangani Data Tidak Seimbang. *Nuansa Informatika*, *14*(1), 34. https://doi.org/10.25134/nuansa.v14i1.2411

Syuhada, A. S., Simanullang, A. M., Lewa, D. S., & Marthin, S. J. (2021). *Machine Learning*. 1–11.

Wahyuni, N. S. (2022). *Diabetes Pada Anak*. Kementerian Kesehatan Direktorat Jenderal Pelayanan Kesehatan. https://yankes.kemkes.go.id/view_artikel/238/diabetes-pada-anak

World Health Organization. (2023). *Noncommunicable Diseases*. https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases

Yuliska, Y., & Syaliman, K. U. (2020). Literatur Review Terhadap Metode, Aplikasi dan Dataset Peringkasan Dokumen Teks Otomatis untuk Teks Berbahasa Indonesia. *IT Journal Research and Development*, *5*(1), 19–31. https://doi.org/10.25299/itjrd.2020.vol5(1).4688