KLASIFIKASI PENYAKIT KANKER PAYUDARA BERDASARKAN DATA BREAST CANCER WISCONSIN DENGAN MENGGUNAKAN METODE NAIVE BAYES

SKRIPSI

Oleh:
BRYAN AHSANUL NUR HARITS
NIM. 18650092

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2023

KLASIFIKASI PENYAKIT KANKER PAYUDARA BERDASARKAN DATA BREAST CANCER WISCONSIN DENGAN MENGGUNAKAN METODE NAIVE BAYES

SKRIPSI

Oleh :

BRYAN AHSANUL NUR HARITS
NIM. 18650092

Diajukan kepada:
Universitas Islam Negeri Maulana Malik Ibrahim Malang
Untuk memenuhi Salah Satu Persyaratan dalam
Memperoleh Gelar Sarjana Komputer (S.Kom)

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM MALANG
2023

HALAMAN PERSETUJUAN

KLASIFIKASI PENYAKIT KANKER PAYUDARA BERDASARKAN DATA BREAST CANCER WISCONSIN DENGAN MENGGUNAKAN METODE NAIVE BAYES

SKRIPSI

Oleh :
BRYAN AHSANUL NUR HARITS
NIM. 18650092

Telah Diperiksa dan Disetujui untuk Diuji:
Tanggal: 1 Desember 2023

Pembimbing I,	Pembimbing II,
Prof. Dr. Suhartono, M.Kom	Dr. Totok Chamidy, M.Kom.
NIP. 19680519 200312 1 001	NIP. 19691222 200604 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 KANKER PAYUDARA BERDASARKAN DATA BREAST CANCER WISCONSIN DENGAN MENGGUNAKAN METODE NAIVE BAYES

SKRIPSI

Oleh :
BRYAN AHSANUL NUR HARITS
NIM. 18650092

Telah Dipertahankan di Depan Dewan Penguji Skripsi
dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
Untuk Memperoleh Gelar Sarjana Komputer (S.Kom)
Tanggal: 8 Desember 2023

Susunan Dewan Penguji

Ketua Penguji	: Dr. Irwan Budi Santoso, M.Kom NIP. 19770103 201101 1 004	( )
Anggota Penguji I	: Syahiduz Zaman, M.Kom NIP. 19700502 200501 1 005	( )
Anggota Penguji II	: Prof. Dr. Suhartono, M.Kom NIP. 19680519 200312 1 001	( )
Anggota Penguji III	: Dr. Totok Chamidy, M.Kom NIP. 19691222 200604 1 001	( )

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

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

PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan di bawah ini:

Nama	:	Bryan Ahsanul Nur Harits
NIM	:	18650092
Fakultas / Jurusan	:	Sains dan Teknologi / Teknik Informatika
Judul Skripsi	:	Klasifikasi Penyakit Kanker Payudara Berdasarkan Data Breast Cancer Wisconsin Dengan Menggunakan Metode Naive Bayes

.

Menyatakan dengan sebenarnya bahwa Skripsi yang saya tulis ini benar-benar 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, 14 Desember 2023
Yang membuat pernyataan,

Bryan Ahsanul Nur Harits
NIM.18650092

HALAMAN MOTTO

… SELALU YAKIN …

HALAMAN PERSEMBAHAN

Saya mengucapkan penghargaan dan rasa terima kasih yang tak terhingga kepada:

1. Allah SWT Yang telah memberikan rahmat, petunjuk, serta kekuatan selama perjalanan panjang dalam menyelesaikan skripsi ini.
2. Orangtua tercinta Bapak Hariyadi dan Ibu Sekar Mlati, yang selalu memberikan cinta, dukungan, doa, dan pengorbanan tanpa batas. Kalian adalah sumber inspirasi dan motivasi dalam setiap langkah kami.
3. Bapak Prof. Dr. Suhratono, M.kom selaku dosen pembimbing I serta Bapak Dr. Totok Chamidy, M.Kom selaku dosen pembimbing II, atas ilmu, bimbingan, arahan, dan kesabaran yang diberikan dalam penulisan skripsi ini.
4. Semua pihak yang telah memberikan bantuan, dorongan, serta dukungan dalam berbagai bentuk selama proses penulisan skripsi ini.

Segala jerih payah dan pengorbanan dari setiap individu di atas adalah sebuah anugerah yang tak terhingga bagi kelancaran terselesaikannya skripsi ini. Kami menyadari bahwa tanpa bantuan dan dukungan mereka, skripsi ini tidak akan terwujud.

Semoga hasil dari skripsi ini dapat bermanfaat bagi kita semua dan menjadi langkah awal bagi perubahan yang lebih baik.

KATA PENGANTAR

Puji syukur kehadirat Allah SWT, yang telah melimpahkan rahmat, hidayah, serta karunia-Nya sehingga penulis dapat menyelesaikan penulisan skripsi ini dengan baik. Shalawat serta salam senantiasa tercurah kepada junjungan kita Nabi Muhammad SAW, yang telah menjadi suri tauladan bagi umat manusia.

Penulis mengucapkan terima kasih yang tak terhingga kepada semua pihak yang telah memberikan bantuan, dukungan, serta motivasi dalam penyelesaian skripsi ini. Sehingga penulis dengan segala hormat mengucapkan terimakasih kepada:

1. Prof. Dr. M. Zainuddin, M.A., selaku rektor Universitas Islam Negeri Maulana Malik Ibrahim Malang.
2. Prof. Dr. Hj. Sri Hariani, M.Si., selaku dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang.
3. Dr. Fachrul Kurniawan M.MT., IPM selaku Ketua Prodi Teknik Informatika Universitas Islam Negeri Maulana Malik Ibrahim Malang.
4. Prof. Dr. Suhratono, M.Kom selaku dosen pembimbing I yang telah memberikan arahan, bimbingan, dan masukan berharga sehingga skripsi ini dapat terselesaikan.
5. Dr. Totok Chamidy, M.Kom selaku dosen pembimbing II, yang telah memberikan bimbingan berharga sehingga skripsi ini dapat terselesaikan.

6. Dr. Irwan Budi Santoso, M.Kom selaku penguji I dan Bapak Syahiduz Zaman, M.Kom selaku penguji II yang dengan sabar memberikan arahan dan saran serta meluangkan waktu dalam menyelesaikan skripsi ini.
7. Segenap civitas akademika Jurusan Teknik Informatika, terutama seluruh dosen, terimakasih atas ilmu dan bimbingan yang telah diberikan selama masa perkuliahan ini.
8. Bapak Hariyadi dan Ibu Sekar Mlati selaku orangtua tercinta yang selalu memberikan dukungan, doa dan memberikan nasihat kepada penulis.
9. Kepada diri saya sendiri, terimakasih telah terus berjuang, berusaha, yakin, semangat pantang menyerah dan bertanggung jawab atas apa yang telah dipilih.

Akhir kata, penulis sadar sepenuhnya bahwa skripsi ini masih jauh dari sempurna. Oleh karena itu, diperlukan segala kritik dan saran yang membangun. Penulis berharap bahwa skripsi ini dapat memberikan manfaat, juga dapat menjadi bahan rujukan dan referensi yang berguna bagi pembaca yang ingin melakukan penelitian lebih lanjut.

Malang, 11 Desember 2023

Bryan Ahsanul Nur Harits

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

ABSTRAK

Harits, Bryan Ahsanul Nur, 2023. **Klasifikasi Penyakit Kanker Payudara Berdasarkan Data Breast Cancer Wisconsin Dengan Menggunakan Metode Naïve Bayes. Skripsi.** Program Studi Teknik Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (I) Prof. Dr. Suhartono, M.Kom (II) Dr. Totok Chamidy, M.Kom.

.

Kata kunci: Klasifikasi Kanker Payudara, Naive Bayes, Skenario Pengujian.

Kanker payudara merupakan salah satu masalah kesehatan utama di seluruh dunia dan memiliki dampak yang signifikan terhadap kesejahteraan perempuan. Diagnosa dini dan klasifikasi yang tepat sangat penting dalam manajemen penyakit ini. Dalam konteks ini, penggunaan teknik-teknik analisis data dan pembelajaran mesin telah menjadi sorotan untuk meningkatkan akurasi klasifikasi. Penggunaan metode Naive Bayes yang telah terbukti efektif dalam pengklasifikasian data medis diharapkan dapat memberikan kontribusi penting dalam meningkatkan akurasi klasifikasi kanker payudara.Tujuan dari penelitian ini adalah untuk mengetahui performa metode naïve bayes dalam melakukan klasifikasi kanker payudara berdasarkan data Breast Cancer Wisconsin. Penelitian ini dilakukan dengan menggunakan 4 skenario pengujian perbandingan data training dan data testing, yaitu pengujian A dilakukan dengan rasio perbandingan 90:10, pengujian B dengan rasio perbandingan 80:20, pengujian C dengan rasio perbandingan 75:25, dan pengujian D dengan rasio perbandingan 70:30. Hasilnya pengujian A dengan menggunakan 10- Fold cross validation menghasilkan performa terbaik yaitu dengan nilai akurasi sebesar 98.03%. Berdasarkan scenario perbandingan pengujian A dapat disimpulkan bahwa dapat dikategorikan ke klasifikasi sangat baik.

ABSTRACT

Harits, Bryan Ahsanul Nur, 2023. ***Breast Cancer Disease Classification Based on Breast Cancer Wisconsin Data Using Naïve Bayes Method***. Undergraduate Thesis. Department of Informatics Engineering, Faculty of Science and Technology, Maulana Malik Ibrahim State Islamic University Malang. Advisors: (I) Prof. Dr. Suhartono, M.Kom (II) Dr. Totok Chamidy, M.Kom.

Breast cancer is one of the major health concerns worldwide and significantly impacts women's well-being. Early diagnosis and accurate classification are crucial in managing this disease. In this context, the use of data analysis techniques and machine learning has been highlighted to enhance classification accuracy. The utilization of the Naive Bayes method, which has proven effective in classifying medical data, is expected to contribute significantly to improving breast cancer classification accuracy. The aim of this research is to assess the performance of the Naive Bayes method in classifying breast cancer based on Breast Cancer Wisconsin data. The study conducts four testing scenarios comparing training and testing data: Test A with a 90:10 ratio, Test B with an 80:20 ratio, Test C with a 75:25 ratio, and Test D with a 70:30 ratio. The results show that Test A using 10-Fold crossvalidation achieved the best performance with an accuracy rate of 98.03%. Based on the comparison scenarios, Test A can be categorized as a highly accurate classification.

Keywords : Breast Cancer Classification, Naïve Bayes, Testing Scenarios.

الملخص

هاريتس، برايان أحسان النور، .٢٠٢٣ **تصنيف سرطان الثدي استنادًا إلى بيانات سرطان الثدي في ويسكونسن باستخدام طريقة نايف بايز**. رسالة بكالوريوس. قسم علوم وتكنولوجيا المعلومات، كلية العلوم والتكنولوجيا، جامعة موالنا مالك إبراهيم اإلسالمية الحكومية في ماالنج. المشرفون:)١(البروفيسور الدكتور سهارتونو، ماجستير الحاسوب)٢(.)الدكتور توتوك تشاميدي، ماجستير الحاسوب.

.**كلمات مفتاحية**: تصنيف سرطان الثدي، نايف بايز، سيناريو االختبار

سرطان الثدي هو أحد أهم مشاكل الصحة في جميع أنحاء العالم وله تأثير كبير على رفاهية النساء. التشخيص المبكر والتصنيف الدقيق ضروريان إلدارة هذا المرض. في هذا السياق، أصبح استخدام تقنيات تحليل البيانات وتعلم اآللة محط األلنظار لتحسين دقة التصنيف. من المتوقع أن يُسهم استخدام طريقة النيف بايز المثبتة فعاليتها في تصنيف البيانات الطبية بشكل هام في زيادة دقة تصنيف سرطان الثدي. الهدف من هذه الدراسة هو معرفة أداء طريقة النيف بايز في تصنيف سرطان الثدي استنادًا إلى بيانات سرطان الثدي في ويسكونسن .تمت هذه الدراسة باستخدام 4 سيناريوهات الختبار مقارنة بين بيانات التدريب وبيانات االختبار، حيث تم بنسبة مقارنة C بنسبة مقارنة ،80:20 واالختبار B بنسبة مقارنة ،90:10 واالختبار A إجراء االختبار باستخدام التقييم المتقاطع بـ 10 طيات A بنسبة مقارنة .70:30 أظهرت نتائج اختبار D ، 75:25واالختبار ، يمكن استنتاج أنه يمكن A أداًء ممتازًا بقيمة دقة بنسبة .%98.03 استنادًا إلى سيناريو اختبار االختبار .تصنيفه كتصنيف ممتاز جدًا.

BAB I
PENDAHULUAN

1.1 Latar Belakang

Kanker payudara adalah keganasan sel-sel pada jaringan payudara, bisa berasal dari komponen kelenjarnya (epitel saluran, maupun lobulusnya) seperti jaringan lemak, pembuluh darah, dan persyarafan jaringan payudara (Arifin, 2021). Kanker payudara juga termasuk penyebab nomor dua kematian terbanyak akibat kanker pada wanita setelah kanker serviks, dan cenderung terus meningkat setiap tahunnya (Cahytanti et al., 2020). Sel kanker payudara dapat bersembunyi di dalam tubuh selama beberapa waktu dan menjadi aktif pada tumor ganas. Wanita rentan terhadap kanker, wanita lebih mungkin mengembangkan kanker payudara seiring bertambahnya usia (Dewi, 2020). Beberapa jenis kanker antara lain kanker payudara, kanker serviks, kanker tulang, kanker otak, kanker darah, kanker kelenjar dan berbagai jenis kanker yang terjadi pada berbagai macam jaringan tubuh (Setiawan et al., 2021). Data hasil riset menunjukkan bahwa ada 2,1 juta kasus kanker payudara setiap tahunnya dan pada tahun 2018 diperkirakan sebanyak 627.000 wanita meninggal karena kanker payudara (Nuraini et al., 2021).

Kanker payudara adalah salah satu jenis kanker yang paling umum di kalangan wanita di seluruh dunia. Menurut World Health Organization (WHO), setiap tahunnya terdapat jutaan kasus baru yang terdiagnosis, dengan dampak signifikan terhadap kesehatan global dan kualitas hidup individu. Wanita yang memiliki risiko tinggi terkena kanker payudara adalah wanita yang berusia subur

(Rasjidi, 2010). Wanita yang berusia subur adalah wanita yang berada dalam usia reproduktif antara 15-49 tahun (Kemenkes RI, 2015). Akibat tingginya tingkat insiden kanker payudara salah satu alasanya adalah masih rendahnya pengetahuan dan pemahaman masyarakat akan berbahaya kanker payudara dan kesadaran penting melakukan pemeriksaan dini (Thaha & Widajadnja, 2017).

Penyebab pasti terjadinya kanker payudara belum sepenuhnya dipahami, meskipun beberapa faktor risiko telah diidentifikasi. Faktor-faktor seperti genetika, lingkungan, gaya hidup, serta faktor hormonal, semuanya berperan dalam perkembangan kanker ini. Kanker payudara juga dapat mempengaruhi tidak hanya fisik, tetapi juga aspek psikologis, sosial, dan ekonomi dari individu yang terkena serta keluarga mereka.

Kanker payudara tetap menjadi salah satu masalah kesehatan global yang membutuhkan perhatian serius karena tingkat prevalensinya yang tinggi dan dampaknya terhadap kesejahteraan perempuan di seluruh dunia. *Dataset Breast Cancer Wisconsin (BCW)* telah menjadi salah satu sumber data kritis dalam upaya pemodelan dan klasifikasi kanker payudara. Data ini terdiri dari berbagai fitur, termasuk ukuran inti sel, ketidakteraturan sel, serta parameter-parameter lain yang signifikan dalam diagnosis kanker payudara.

Dalam Al-Quran Surah Yunus ayat 57:

يَٰٓأَيُّهَا ٱلنَّاسُ قَدْ جَآءَتْكُم مَّوْعِظَةٌ رَّبِّكُمْ وَشِفَآءٌ لِّمَا فِى ٱلصُّدُورِ وَهُدًى وَرَحْمَةٌ لِّلْمُؤْمِنِينَ

"*Hai manusia, sesungguhnya telah datang kepadamu pelajaran dari Tuhanmu dan pnyembuh bagi penyakit-penyakit (yang berada) dalam dada dan petunjuk serta rahmat bagi orang-orang yang beriman*" (Qs. Yunus 57).

Menurut Al-Mukhtashar, ayat ini berisi tentang Al-Quran yang merupakan obat penawar untuk penyakit bimbang dan ragu yang bersarang di dalam hati. Salah satu penyebab kanker payudara berbahaya ialah karena penderita yang terkena kanker payudara merasa ragu bisa sembuh.

Metode klasifikasi Naive Bayes, yang didasarkan pada teorema Bayes dengan asumsi independensi antar-fitur, telah terbukti menjadi salah satu pendekatan yang kuat dan efisien dalam klasifikasi kanker payudara berdasarkan data *BCW* (Suprianto, 2020). Keunggulan dari metode ini terletak pada kinerja yang baik dalam menangani dataset berukuran besar dan kemampuannya mengatasi masalah klasifikasi dengan fitur-fitur yang relatif besar (Diana Dumitru, 2009). Metode *Naïve Bayes* digunakan karena medote yang probabilitas dan statistik yaitu dapat memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya (Han *et al*, 2011).

Namun, meskipun Naive Bayes telah menunjukkan hasil yang menjanjikan, pemahaman mendalam tentang kemampuan dan batasannya dalam konteks klasifikasi kanker payudara dengan menggunakan dataset *BCW* masih diperlukan. Penelitian lebih lanjut diperlukan untuk mengevaluasi kinerja metode ini secara lebih mendalam, mengeksplorasi pengoptimalan parameter, dan menganalisis faktor-faktor yang mempengaruhi akurasi klasifikasi. Klasifikasi merupakan cara untuk mengelompokkan data dengan karakteristik yang serupa ke dalam kelas yang sudah ditentukan terlebih dahulu. Label kelas pada klasifikasasi berguna untuk menamai kelompok yang memiliki pola serupa (Muntiari & Hanif, 2022).

Berdasarkan uraian di atas, penelitian ini akan mengimplementasikan metode *Naïve Bayes* untuk melakukan klasifikasi kanker payudara. Untuk penggunaan data kanker payudara yang akan digunakan diperoleh dari *UCI Machine Learning Repository*. Sehingga, penelitian ini berjudul "**Klasifikasi Penyakit Kanker Payudara Berdasarkan Data Breast Cancer Wisconsin Dengan Menggunakan Metode *Naïve Bayes***".

1.2 Rumusan Masalah

Bagaimana performa metode Naïve Bayes dalam melakukan klasifikasi kanker payudara berdasarkan data dari Breast Cancer Wisconsin Dataset?

1.3 Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk mengetahui performa metode Naïve Bayes dalam melakukan klasifikasi kanker payudara berdasarkan pada data Breast Cancer Wisconsin Dataset yang telah di dapat.

1.4 Batasan Masalah

Agar diperoleh pembahasan yang sesuai dengan rumusan dan tujuan masalah maka diperukan batasan masalah, yaitu data yang akan digunakan dalam penelitian ini adalah data bernama *breast cancer wisconsin* yang dapat diakses secara public melalui *website UCI Machine Learning Repository*.

1.5 Manfaat Penelitian

Dengan melakukan penelitian ini, diharapkan dapat memberikan manfaat untuk kemudian hari. Adapun beberapa memanfaat tersebut ialah sebagai berikut.

1. Diharapkan dapat meningkatkan kesadaran masyarakat tentang kanker payudara, tanda-tanda awal, dan faktor risiko.
2. Memberikan Informasi yang lebih baik tentang kanker payudara untuk membantu keluarga dan teman-teman pasien serta memberikan wawasan kepada pembaca tentang penggunaan metode *Naïve Bayes*.
3. Diharapkan penelitian ini dapat membantu para peneliti yang ingin mengembangkan topik dari penelitian ini.

BAB II
STUDI PUSTAKA

2.1 Metode Naïve Bayes

Pada penelitian yang dilakukan oleh (Safutra, et al., 2016) peneliti mengimplementasikan metode *Naïve Bayes* dalam mendiagnosis penyakit kanker payudara dengan menggunakan *dataset* yang didapat dari UCI *Machine Learning Repository* yaitu *Breast Cancer Wisconsin Diagnostic dataset*. *Dataset* ini berisikan 569 data, 10 atribut dan 2 kelas. Peneliti mengolah data tersebut dengan teknik diskritasi data serta teknik *Maximum Likelihood Estimation* (*MLE*) yang digunakan untuk memperkirakan parameter tiap atribut dalam model statistic. Pada penelitian ini peneliti membagi data yang akan digunakan menjadi dua yaitu data traning dan data uji, hasil yang didapatkan dari penelitian ini menunjukan bahwa metode Naïve Bayes menghasilkan nilai akurasi sebesar 98.6726%.

Pada penelitian (Ramadhan et al., 2021) peneliti mengimplementasikan metode *Naïve Bayes* dalam mengklasifikasi kanker payudara dengan menggunakan *dataset* yang didapat dari UCI *Machine Learning Repository* yaitu *Breast Cancer Wisconsin Diagnostic dataset*. *Dataset* ini berisikan 569 data, 10 atribut dan 2 kelas. Dalam penelitian ini bertujuan untuk menyelesaikan permasalahan pada proses *pre-processing* dengan cara peneliti menggunakan teknik *SMOTE oversampling* untuk masalah data yang tidak seimbang dan *Gini Score* untuk fitur peringkat. Peneliti melakukan pengelompokkan data berdasarkan kelasnya yaitu *Malignant* dan *Benign*, hasil yang didapatkan dari penelitian ini menunjukan bahwa metode *Naïve Bayes* menghasilkan nilai

precision dalam kelas *Malignant* sebesar 88%, dan nilai *recall* sebesar 89%. Sedangkan dalam kelas *Benign* didapatkan nilai *precision* sebesar 93% dan nilai *recall* sebesar 93%.

Pada penelitian yang dilakukan (Diana Dumitru, 2009) peneliti menggunakan metode *Naïve Bayes* untuk melakukan prediksi kanker payudara dengan menggunakan *dataset* yang didapat dari UCI *Machine Learning Repository* yaitu *Breast Cancer Wisconsin Diagnostic dataset*. *Dataset* ini berisikan 569 data, 10 atribut dan 2 kelas. Tujuan dari penelitian ini ialah untuk mencari nilai *sensitivitas* dari model *Naïve Bayes* yang telah dibangun dengan cara membagi data menjadi dua yaitu data traning dan data testing, hasil yang didapatkan dari penelitian ini menunjukan bahwa metode Naïve Bayes menghasilkan nilai akurasi sebesar 74%, dan nilai sensitivitas sebesar 100% yang berarti bahwa model mengenali semua orang yang sakit yang sebenarnya orang yang sakit.

Pada penelitian yang dilakukan oleh (Othman, et al., 2007) peneliti melakukan perbandingan teknik klasifikasi yang berbeda untuk penyakit kanker payudara dengan menggunakan dataset yang didapat dari *UCI Machine Learning Repository* yaitu *Breast Cancer Wisconsin Diagnostic* dataset. Dataset ini berisikan 569 data, 10 atribut dan 2 kelas. Peneliti melakukan pengolahan data dengan menggunakan teknik WEKA *Waikato Environment for Knowledge Analysis*. Pada penelitian ini beberapa teknik atau metode yang dibandingkan oleh peneliti gunakan ialah metode *Naïve Bayes*, *Radial Basis Function, Pruned Tree* dan *Nearest Neighbors algorithm,* penelitian dilakukan dengan cara membagi data

menjadi dua, yaitu 75% dari seluruh data akan digunakan sebagai data pelatihan dan 25% data sisanya akan digunakan sebagai data uji. Hasil yang didapatkan dari penelitian ini menunjukan bahwa metode *Naïve Bayes* menghasilkan nilai akurasi paling tinggi dibandingkan dengan metode *Radial Basis Function*, *Pruned Tree* and *Nearest Neighbors algorithm* yaitu sebesar 89.71%.

Dalam sebuah penelitian yang dilakukan oleh (Kharya et al., 2014) peneliti mengimplementasikan metode *Naïve Bayes* dalam melakukan deteksi probabilistik kanker payudara dengan menggunakan *dataset* yang didapat dari UCI *Machine Learning Repository* yaitu *Breast Cancer Wisconsin Diagnostic dataset*. *Dataset* ini berisikan 569 data, 10 atribut dan 2 kelas. Dalam penelitian ini peneliti mengolah data yang diperoleh dengan sebuah teknik normalisasi data yaitu normalisasi *range of value*. Peneliti melakukan pengelompokkan data berdasarkan kelasnya yaitu 34.5% *Malignant* dan 65.5% *Benign*, hasil yang didapatkan dari penelitian ini menunjukan bahwa metode *Naïve Bayes* menghasilkan nilai akurasi sebesar 93%.

Tabel 2.1 Hubungan Antar Peneliti

No	Paper	Metode	Tahapan Penelitian	Perbandingan
1	Safutra et al., (2016)	*Naïve Bayes*	Data yang digunakan sama, menggunakan data *cleaning*, menggunakan *diskritasi* data, mencari nilai *likelihood* tiap atribut dan hanya mencari nilai akurasi dengan hasil sebesar 98.6726%.	Data yang digunakan sama, melakukan split data, mencari nilai *mean* dan *std deviasi* tiap atribut, mencari 4 nilai evaluasi.

2	Ramadhan et al., (2021)	*Naïve Bayes*	Data yang digunakan sama, menggunakan teknik *SMOTE*, teknik normalisasi *Gini Score*, mengelompokkan data berdasarkan kelasnya, mencari nilai presisi dan recall berdasarkan kelasnya.	Data yang digunakan sama, membuat 4 scenario pengujian.
3	Diana Dumitru (2009)	*Naïve Bayes*	Data yang digunakan sama, menggunakan persamaan *distrubusi Gaussian*, mencari nilai akurasi, *sensitivity* dan *specificity* naïve bayes.	Data yang digunakan sama, menggunakan persamaan *Gausian NB*, mencari 4 nilai evaluasi.
4	Othman et al., (2006)	*Naïve Bayes, Radial Basis Function, Decision tree and pruning, Single Conjunctive Rule Learner dan Nearest Neighbors Algorithm.*	Data yang digunakan sama, membandingkan berbagai metode, menggunakan teknik *WEKA* dan hasil penelitian menunjukkan bahwa *Naïve Bayes* menghasilkan nilai akurasi paling besar yaitu 89.71%.	Data yang digunakan sama, hanya menggunakan metode *Naïve Bayes*, melakukan 4 skenario pengujian.
5	Kharya et al., (2014)	*Naïve Bayes*	Data yang digunakan sama, mengolah data dengan normalisasi *range of values*. Hasil yang didapatkan dalam penelitian menunjukan *Naïve Bayes* dalam memprediksi kanker payudara dengan nilai akurasi sebesar 93%.	Data yang digunakan sama. Tidak melakukan teknik mengolah data memisah data yang akan digunakan menjadi 4 skenario pengujian,

2.2 Identifikasi Kanker Payudara

Salah satu penyakit kanker yaitu kanker payudara menjadi jenis kanker yang sangat menakutkan bagi perempuan di seluruh dunia, juga di Indonesia. Kanker payudara adalah tumor ganas yang terbentuk dari sel-sel payudara yang tumbuh dan berkembang tanpa terkendali sehingga menyebar di antara jaringan atau organ di dekat payudara atau ke bagian tubuh lainnya (Kemenkes RI, 2016).

Ginsburg *et al.* (2020) dalam penelitiannya yang berjudul *Breast cancer early detection: A phased approach to implementation*. Menganalisa bahwa penundaan pengobatan kanker payudara selama lebih dari 3 bulan telah dikaitkan dengan stadium penyakit yang lebih lanjut saat didiagnosis dan kelangsungan hidup yang lebih buruk. Pada saat yang sama, edukasi bagi penyedia layanan kesehatan primer untuk mengenali tanda dan gejala awal kanker payudara diperlukan agar dapat segera dirujuk melalui sistem layanan kesehatan. Faktor–faktor yang bersifat seperti multifaktoral, termasuk faktor structural, sosiolkultural, personal dan finansial merupakan faktor yang dapat mempengaruhi kesempatan perempuan untuk mencari dan menerima perawatan. Bahkan ketika seorang pasien mencari perawatan segera setelah timbulnya gejala yaitu (*Presentasi awal*), hal ini tidak selalu diterjemahkan ke dalam diagnosis dini bisa dikarekan penyedia layanan tidak memiliki pelatihan atau pengetahuan yang sesuai untuk mengenali kanker payudara stadium awal, tidak tahu kemana atau bagaimana merujuk untuk intervensi diagnostic yang diperlukan dan sistem kesehatan terfragmentasi sedemekian rupa sehingga pasien tidak dapat menjalani seluruh jalur perawatan.

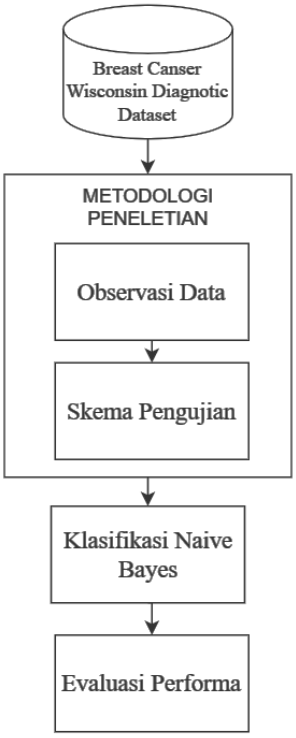
Chazar & Erawan (2020) dalam sebuah penelitiannya yang berjudul *Machine Learning Diagnosis Kanker Payudara Menggunakan Algoritma Support Vector Machine* mengatakan meski tumor tidak selalu kanker, tapi masih banyak orang yang percaya bahwa kanker dan tumor adalah hal yang sama. Untuk menentukan jenis kanker payudara dapat digunakan biopsy dan mamografi untuk pemeriksaan. Apakahn suatu sel kanker payudara bersifat ganas atau jinak dapat diketahui dari hasil pemeriksaan biopsi dengan menggunakan *Fine Needle Aspiration* (*FNA*). Kanker payudara yang berkembang dari tumor dibagi menjadi beberapa fase, mulai dari stadium 0 hingga stadium IV.

Schmid *et al.* (2020) dalam penelitiannya yang berjudul *Pembrolizumab for early triple-negative breast cancer* menguji coba menugaskan (dengan rasio 2:1) pasien dengan kanker payudara triple-negatif stadium II atau stadium III yang sebelumnya tidak diobati untuk menerima terapi neoadjuvan dengan empat siklus pembrolizumab (dengan dosis 200 mg) setiap 3 minggu. Kanker payudara triple-negatif dini yang berisiko tinggi sering dikaitkan dengan kekambuhan dini dan kematian yang tinggi. Kemoterapi adjuvant merupakan pendekatan pengobatan yang lebih disukai, selain berpotensi meningkatkan kemungkinan reseksi tumor dan konservasi payudara, pasien yang memiliki respons lengkap patologis setelah terapi neoadjuvan memiliki kelangsungan hidup bebas kejadian yang lebih lama dan kelangsungan hidup secara keseluruhan.

BAB III
DESAIN DAN IMPLEMENTASI

3.1 Metedologi Penelitian

Untuk menyusun tugas akhir, pengumpulan data dan informasi yang kuat menjadi esensial guna menegaskan validitas dan kelengkapan materi serta pembahasan. Oleh karena itu, bab ini akan menguraikan dengan rinci metode dan langkah-langkah yang akan diterapkan dalam proses klasifikasi penyakit kanker payudara dengan menggunakan metode Naïve Bayes. Dengan demikian, akan dijelaskan dengan terperinci proses yang meliputi pemilihan fitur, pra-pemrosesan data, pembagian dataset, pelatihan model, serta pengujian untuk mengukur performa dan kehandalan model klasifikasi.



Gambar 3.1 Tahapan Penelitian

3.2 Observasi Data

Data yang digunakan dalam penelitian ini adalah data sekunder yang diperoleh melaui data publik yang semua orang dapat mengaksesnya yaitu melaui *UCI Machine Learning Repository*. Lebih tepatnya data public ini yaitu data yang bernama *Breast Cancer Wisconsin* (*Diagnotic*) dari Diagnotic Wisconsin Breast Cancer Database yang diambil dari UCI *machine learning repository* dengan alamat https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic data tersebut didapatkan melalui pencatatan di lapangan atau laporan resmi lainnya. Total *Dataset* ini terdiri dari 569 *instances* dengan 32 atribut, serta karakterisitik *Multivariate*. Dataset tersebut dapat digunakan dengan mengikuti beberapa langkah yaitu dengan melakukan *citation request* dari dataset dan akan merujuk kepada *citation policy* dari *UCI Machine Learning Repository*. Adapun 32 atribut dalam data ini dapat diketahui dalam tabel berikut ini.

Tabel 3.1 Dataset Pima Indians Diabetes

No	Fitur	Tipe
1	*Radius_mean*	Float
2	*texture_mean*	Float
3	*perimeter_mean*	Float
4	*area_mean*	Float
5	*smoothness_mean*	Float
6	*compactness_mean*	Float
7	*concavity_mean*	Float
8	*concave points_mean*	Float
9	*symmetry_mean*	Float
10	*fractial dimension_mean*	Float
11	*radius_se*	Float
12	*texture_se*	Float
13	*perimeter_se*	Float
14	*area_se*	Float
15	*smoothness_se*	Float

16	*compactness_se*	Float
17	*concavity_se*	Float
18	*concave points_se*	Float
19	*symmetry_se*	Float
20	*fractal_dimension_se*	Float
21	*radius_worst*	Float
22	*texture_worst*	Float
23	*perimeter_worst*	Float
24	*area_worst*	Float
25	*smoothness_worst*	Float
26	*compactness_worst*	Float
27	*concavity_worst*	Float
28	*concave points_worst*	Float
29	*symmetry_worst*	Float
30	*fractal_dimension_worst*	Float
31	*Diagnosis*	Binary

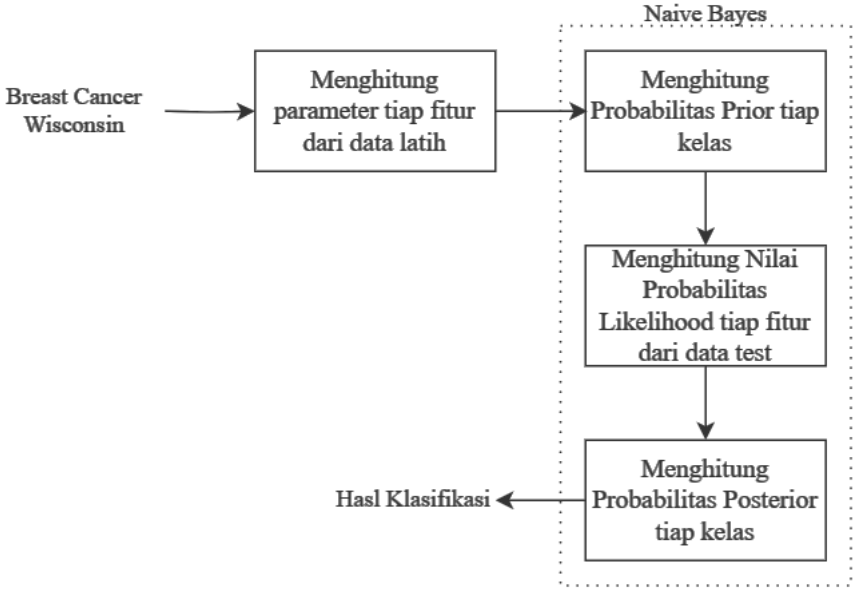
Tahapan berikutnya merupakan tahapan dimana untuk memahami data yang akan digunakan sebelum data diolah pada tahapan proses selanjutnya. Data yang telah didapat dari *UCI Machine Learning Repository* memiliki 32 atribut, jika yang didapat berdasarkan operasi biopsi atau *fine needle aspirate (FNA)* dalam penelitian Zhazar & Erawan (2020) maka akan menghasilkan 10 atribut utama yang dimana dapat diketahui yaitu *radius*, *texture*, *perimeter*, *area*, *smoothness*, *compactness*, *concavity*, *concave points*, *symmetry* dan *fractal dimension*. Masing-masing dari 10 atribut utama tersebut memiliki 3 indikator yaitu *mean*, *standard error/se*, dan *worst*. Terdapat juga 2 atribut tambahan lainnya yaitu *ID number* dan *Diagnosis*. Untuk lebih detailnya dapat dilihat pada tabel 3.2.

Tabel 3.2 Atribut Utama

No	Atribut	Keterangan
1	*ID number*	Nomor identitas setiap pasien.
2	*Diagnosis*	Identifikasi dalam dataset ini terdapat dua kelas yaitu "B" untuk kanker payudara *benign* (tidak ganas) dan "M" untuk kanker payudara *malignan* (ganas).
3	*Radius*	Rata-rata jarak dari titik tumor pada kanker payudara ke tepi kanker tersebut.
4	*Texture*	Karakteristik dan pola struktur jaringan payudara yang kemungkinan adanya tumor atau kanker.
5	*Perimeter*	Garis atau tepi dari tumor kanker payudara yang menggambarkan batas yang membatasi area kanker di dalam jaringan payudara untuk mengenali sifat pertumbuhan tumor ke dalam jaringan sekitarnya.
6	*Area*	Wilayah atau bagian tertentu dari payudara yang terdampak oleh pertumbuhan tumor atau sel kanker yang tidak normal atau sel kanker tumbuh dan berkembang biak dengan tidak terkendali menyebabkan benjolan yang abnormal.
7	*Smoothness*	Batas massa yang terdeteksi, jika terlihat halus dan teratur berarti tumor bersifat jinak atau *benign* (tidak ganas).
8	*Compactness*	Karakteristik seberapa padat bentuk massa tumor, jika ukuran yang besar atau bentuk lebih padat yang berarti kanker payudara bersifat *malignan* (ganas).
9	*Concavity*	Bentuk dari tepi tumor atau jaringan pada tumor, apabila tepi tumor berbentuk cekung dalam maka menandakan kanker payudara *malignan* (ganas).
10	*Concave Points*	Bentuk titik-titik tertentu pada tepi tumor, tumor yang ganas memiliki titik tepi yang tidak teratur atau bergerigi.
11	*Symmetry*	Kesamaan atau keseimbangan antara dua sisi payudara baik bentuk, ukuran, serta posisi.
12	*Fractial Dimension*	Karakteristik bentuk dari tumor.
13	*Mean*	Nilai rata-rata dari tumor atau massa kanker payudara seperti ukuran, kepadatan, atau parameter lain.
14	*Standard Error/SE*	Niali Ketidakpastian dalam estimasi suatu parameter rata-rata yang dihitung dari sampel data yang digunakan.
15	*Worst*	Perkiraan suatu rata-rata yang mengukur nilai maksimal.

3.3 Naïve Bayes

Penjelasan tentang dasar teori Naive Bayes meliputi konsep probabilitas serta asumsi independensi antar-fitur yang menjadi dasar dari metode ini. Selain itu, tahapan-tahapan penerapan Naive Bayes dalam proses klasifikasi akan diuraikan, termasuk bagaimana model probabilistik ini memproses data untuk memprediksi kelas dari sampel kanker payudara. Langkah-langkah pemilihan parameter yang tepat dan evaluasi kinerja model juga akan dibahas secara rinci dalam sub bab ini. Dalam penelitian (Diana Dumitru, 2009) peneliti menjelaskan bahwa meskipun asumsi indepedensi atribut ini tidak mencerminkan kenyataan di banyak bidang, metode klasifikasi ini terbukti efektif. Tahapan-tahapan penerapan metode ini dalam penelitian ini dijelaskan melalui alur klasifikasi Naïve Bayes yang direpresentasikan dalam bentuk diagram blok pada gambar 3.2



Gambar 3.2 Blok Diagram Naive Bayes

Klasifikasi Naïve Bayes didasarkan pada asumsi bahwa nilai atribut pada suatu kelas tidak bergantung pada kelas lainnya. Metode ini melibatkan penentuan probabilitas prior untuk setiap nilai fitur, probabilitas likelihood untuk setiap kelas, dan perhitungan probabilitas posterior. Hasil dari nilai posterior tertinggi menjadi hasil dari klasifikasi Naïve Bayes.

Pada penelitian kali ini, Gaussian Naïve Bayes dipilih sebagai alternatif yang lebih tepat daripada Multinomial Naïve Bayes. Metode Gaussian Naïve Bayes diadopsi karena mengasumsikan bahwa data dalam setiap kelas mengikuti distribusi Gaussian, yang lebih sesuai untuk data kontinu (N. Rezaeian & G. Novikova, 2020). Sebaliknya, Multinomial Naïve Bayes lebih cocok untuk data diskrit, seperti dalam analisis teks. Berikut merupakan contoh dari penerapan Gaussian Naïve Bayes secara sederhana.

Tabel 3.3 Contoh Dataset

Diagnosis	***Radius_mean***	***Texture_mean***
1	17.99	10.38
1	20.57	17.77
1	19.69	17.89
1	11.42	19.98
0	13.54	14.36
0	13.08	15.71
0	13.03	12.44
0	13.49	18.42
0	15.78	17.89

Berdasarkan pada Tabel 3.3, saya akan mengilustrasikan contoh sederhana yang terbatas pada fitur-fitur tertentu untuk menunjukkan penerapan algoritma Gaussian Naïve Bayes dalam proses klasifikasi. Tabel 3.3 menampilkan penggunaan hanya dua fitur, yaitu radius_mean dan texture_mean, dalam

mengklasifikasi kanker payudara, dengan kelas target diagnosis yang diwakili oleh 0 untuk kanker jinak dan 1 untuk kanker ganas. Berikut merupakan Langkah-langkah penerapan Gaussian Naïve Bayes berdasarkan data pada Tabel 3.3.

Langkah pertama adalah melakukan pemisahan data atau split data. Dalam penelitian ini, penulis membuat contoh pemisahan data dengan proporsi 90 data latih dan 10 data uji. Namun, dari data yang disiapkan, hasilnya menghasilkan 8 data latih dan 1 data uji. Berikut adalah hasil dari proses pemisahan data.

Tabel 3.4 Data Training

Diagnosis	***Radius_mean***	***Texture_mean***
1	17.99	10.38
1	20.57	17.77
1	19.69	17.89
1	11.42	19.98
0	13.54	14.36
0	13.08	15.71
0	13.03	12.44
0	13.49	18.42

Tabel 3.5 Data Test

Diagnosis	***Radius_mean***	***Texture_mean***
0	15.78	17.89

Setelah memisahkan data pada tabel 3.3 menjadi kelas diagnosis 0 untuk kanker jinak dan 1 untuk kanker ganas, langkah berikutnya adalah menemukan subset masing-masing dari kedua diagnosis tersebut. Nilai parameter dari setiap kelas pada tabel 3.4 dapat diidentifikasi melalui persamaan yang selanjutnya dijelaskan.

$$\mu = \frac{jumlah\ semua\ nilai\ n}{n} \tag{3.1}$$

$$\sigma = \frac{\sqrt{\sum_{i=1}^{n}(\mu)^2}}{n} \tag{3.2}$$

Dengan menggunakan persamaan 3.1 dan 3.2 pada tabel 3.4 maka akan diketahui nilai parameter dari kedua fitur tiap kelas. Berikut merupakan hasil persamaan.

Tabel 3.6 Hasil Parameter Kelas 1

Jumlah sampel	4
Rata-rata radius_mean	17.41
Standard deviasi radius_mean	3.632
Rata-rata texture_mean	16.50
Standard deviasi texture_mean	3.632

Tabel 3.7 Hasil Parameter Kelas 0

Jumlah sampel	4
Rata-rata radius_mean	13.28
Standard deviasi radius_mean	0.231
Rata-rata texture_mean	15.23
Standard deviasi texture_mean	2.175

Hasil perhitungan pada tabel 3.6 dan 3.7 merupakan bagian penting dalam proses pra-pemrosesan data dan persiapan untuk analisis statistik atau pembuatan model klasifikasi naïve bayes. Beberapa fungsi perhitungan tersebut dapat diketahui yaitu sebagai berikut.

1. Deskripsi Data

   Memberikan pemahaman tentang distribusi dan statistik deskriptif dari fitur yang diamati dan menunjukkan bagaimana nilai-nilai tersebar di sekitar rata-rata serta seberapa bervariasinya nilai-nilai tersebut.

2. Pemilihan Fitur

   Bisa membantu dalam pemilihan fitur-fitur yang relevan untuk pemodelan klasifikasi naïve bayes. Fitur-fitur dengan variasi yang rendah atau memiliki kontribusi kecil terhadap perbedaan kelas target bisa dihilangkan.

3. Pemrosesan Lanjutan

   Memungkinkan langkah-langkah lanjutan dalam pra-pemrosesan, seperti normalisasi atau penskalaan fitur-fitur, terutama saat menggunakan algoritma yang sensitif terhadap skala.

4. Modeling Statistik

   Berguna dalam membangun model statistik yang mengasumsikan distribusi tertentu untuk fitur-fiturnya. Misalnya, jika distribusi Gaussian digunakan, nilai rata-rata dan standard deviasi digunakan dalam mengasumsikan distribusi tersebut.

5. Inisialisasi Parameter

   Dalam beberapa algoritma machine learning atau teknik optimasi, nilai-nilai ini bisa digunakan sebagai inisialisasi parameter atau untuk mengatur nilai awal.

6. Pemecahan Masalah

   Mengidentifikasi potensi masalah dalam data seperti adanya outliers atau data yang tidak biasa yang mungkin mempengaruhi analisis atau pemodelan klasifikasi naïve bayes.

Jadi, perhitungan nilai-nilai ini merupakan langkah penting dalam pemahaman dan persiapan data sebelum melakukan analisis atau pemodelan lebih lanjut. Dengan pemahaman yang lebih baik tentang distribusi fitur-fitur yang diamati, kita dapat membangun model yang lebih baik dan membuat asumsi yang lebih tepat tentang data yang sedang kita kerjakan.

Langkah selanjutnya adalah menghitung probabilitas prior menggunakan persamaan 3.3 sebagai berikut.

$$P(Y) = \frac{Ny}{N} \tag{3.3}$$

Dimana:

P(Y) : Probabilitas prior

Ny : Jumlah data pada kelas Y

N : Jumlah keseluruhan data

Berdasarkan pada Dataset tabel 3.4 yang digunakan, maka probabilitas prior tiap kelas digambarkan pada tabel 3.8 berikut.

Tabel 3.8 Probabilitas Prior

P(Y)	Nilai
M (1)	4/8
B (0)	4/8

Dengan berdasarkan pada tabel 3.8 maka akan diketaui nilai probabilitas prior dari masing-kelas dengan menggunakan persamaan 3.3 sebagai berikut.

$$P(Y = 0) = \frac{4}{8} = 0.5$$

$$P(Y = 1) = \frac{4}{8} = 0.5$$

Probabilitas prior memiliki peran utama sebagai fondasi awal yang menggambarkan seberapa mungkin suatu kelas (dalam konteks klasifikasi) muncul sebelum data diamati. Selain menjadi pendukung dalam model Naïve Bayes, probabilitas prior juga berfungsi sebagai pengatur bobot informasi, pengendali estimasi, dan peningkatan model. Dengan demikian, probabilitas prior menjadi landasan penting dalam menghitung probabilitas likelihood yang nantinya digunakan bersama-sama dalam Teorema Bayes. Gabungan dari keduanya menghasilkan probabilitas posterior yang menandai prediksi akhir atau estimasi kelas target setelah mempertimbangkan data.

Dengan mengetahui nilai probabilitas prior dari masing-masing kelas langkah selanjutnya adalah menghitung probabilitas likelihood dengan menggunakan persamaan 3.4 sebagai berikut.

$$P(Xi = x|Y = y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad 3.4$$

Dimana:

Xi : fitur ke-i

Y : variable target atau kelas

y : nilai dari variable target atau kelas

μ : mean berdasarkan fitur yang akan dihitung

σ : standard deviasi berdasarkan fitur yang akan dihitung

x : nilai yang diamati dari fitur ke-i

Penerapan persamaan 3.3 berdasarkan data yang terdokumentasi dalam tabel 3.6 dan tabel 3.7 dapat dijelaskan sebagaimana berikut, dengan ketentuan berdasarkan fitur radius_mean = 15.78 dan texture_mean = 17.89 sebagaimana tertera dalam tabel 3.5. Ketika menggunakan dataset yang sama pada tabel 3.5, misalnya kita ingin mengidentifikasi atau mengklasifikasi apakah seorang pasien mengalami kanker payudara ganas atau kanker payudara jinak.

$$P(X_1 = 15.78|Y = 0) = \frac{1}{2\pi \cdot 0.231} e^{-\frac{(15.78-13.28)^2}{2 \cdot 0.231^2}} = 0.198$$

$$P(X_2 = 17.89|Y = 0) = \frac{1}{2\pi \cdot 2.175} e^{-\frac{(17.89-15.23)^2}{2 \cdot 2.175^2}} = 0.086$$

$$P(X_1 = 15.78|Y = 1) = \frac{1}{2\pi \cdot 3.632} e^{-\frac{(15.78-17.41)^2}{2 \cdot 3.632^2}} = 0.0995$$

$$P(X_2 = 17.89|Y = 1) = \frac{1}{2\pi \cdot 3.632} e^{-\frac{(17.89-16.50)^2}{2 \cdot 3.632^2}} = 0.102$$

Dengan diketahui nilai probabilitas prior dari masing-masing kelas dan nilai probablitas likelilihood dari tiap fitur masing-masing kelas maka bisa diketahui persamaan perhitungan posterior. Berikut merupakan persamaan perhitungan probabilitas posterior berdasarkan dari nilai yang diketahui.

$$P\ (Y|X) = P(X|Y) \times P(Y) \qquad 3.5$$

Maka:

$$P\big(Y = 0\big|X = (15.78, 17.89)\big) = 0.198 \times 0.086 \times 0.5 = 0.008514$$

$$P\big(Y = 1\big|X = (15.78, 17.89)\big) = 0.0995 \times 0.102 \times 0.5 = 0.0050745$$

Dari nilai-nilai ini, probabilitas posterior untuk kanker payudara jinak (Diagnosis 0) lebih tinggi ($0.008514 > 0.0050745$) dibandingkan dengan kanker payudara ganas (Diagnosis 1).

Dengan demikian, berdasarkan nilai probabilitas posterior yang dihitung, pasien lebih cenderung terklasifikasi sebagai kanker payudara jinak (Diagnosis 0) daripada kanker payudara ganas (Diagnosis 1).

3.4 Implementasi *Gaussian Naïve Bayes*

Tahap implementasi Gaussian Naïve Bayes ini melibatkan proses pembuatan perangkat lunak yang disesuaikan dengan desain dan rekayasa sistem yang telah dirancang. Hasil dari implementasi *Gaussian Naïve Bayes* ini digunakan untuk melakukan klasifikasi pada kasus penyakit kanker payudara. Pada tahap ini, proses dimulai dengan input data ke dalam sistem yang telah dikembangkan menggunakan bahasa pemrograman Python. Data yang diperoleh dari *UCI Machine Learning Repository* digunakan pada setiap pengujian, di mana skenario pengujian dari proses pelatihan *Gaussian Naïve Bayes* dibentuk dengan menggunakan fungsi dari *train_test_split* dengan ketentuan yang telah ditetapkan sebagai berikut.

1. Digunakan dalam machine learning untuk membagi dataset menjadi dua subset yang lebih kecil: satu untuk pelatihan (training) dan satu lagi untuk pengujian (testing) model. Fungsi ini memungkinkan untuk membagi data dengan proporsi tertentu sesuai kebutuhan. Penggunaan fungsi *train_test_split* ditunjukkan pada tabel berikut.

   Tabel 3.9 Penggunaan *train_test_split*

```
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=test_ratio, train_size=train_ratio, random_state=42)
```

a) Argumen utama :

X: Variabel yang berisi fitur-fitur atau atribut-atribut dari dataset.

y: Variabel yang berisi label atau target yang ingin diprediksi.

B) ***test_size*** adalah parameter yang menentukan ukuran subset yang akan digunakan untuk pengujian. Ini merupakan pecahan dari keseluruhan dataset dan untuk menentukan proporsi data yang akan dijadikan data pengujian.

C) ***train_size*** adalah parameter yang menentukan ukuran subset yang akan digunakan untuk pelatihan. Ini merupakan pecahan dari keseluruhan dataset dan untuk menentukan proporsi data yang akan dijadikan data pelatihan.

D) ***test_size=test_ratio, train_size=train_ratio*** ukuran dari data test dan data training yang akan digunakan dalam skenario pengujian berdasarkan ***ratios.*** Hal tersebut ditunjukkan pada tabel berikut.

Tabel 3.10 Implementasi ke *ratios*

```
for train_ratio, test_ratio in ratios
```

E) Setiap iterasi dari ***for*** akan mengambil satu set rasio dari ***ratios.*** Hal tersebut yang membuat pengujian berjalan.

Tabel 3.11 *ratios*

```
ratios = [(0.9, 0.1), (0.8, 0.2), (0.75, 0.25), (0.7, 0.3)]
```

F) ***random_state*** digunakan untuk mengatur keacakkan (randomness) saat proses pemisahan data dilakukan.

3.5 Skenario Uji COba

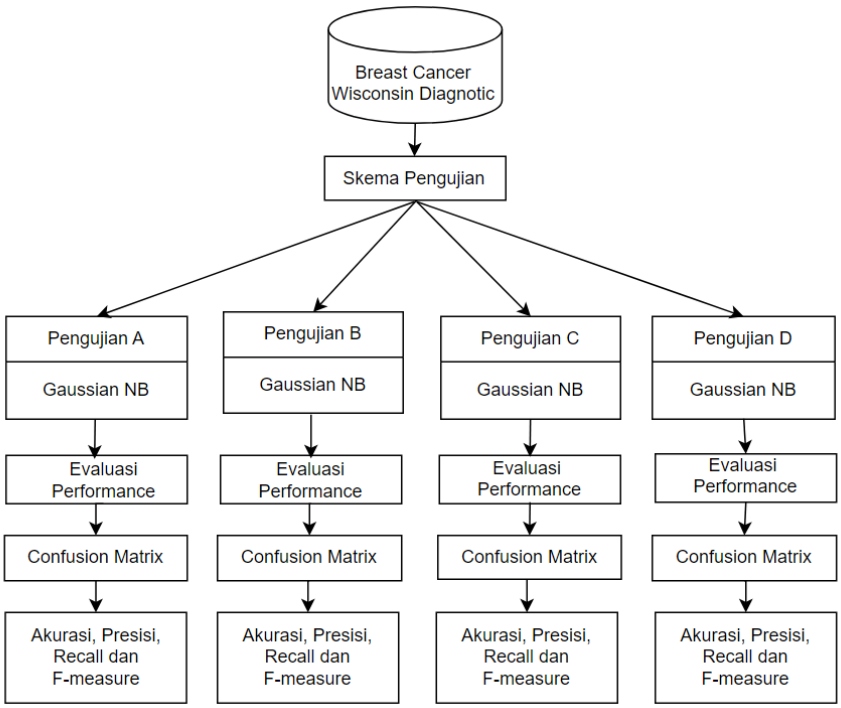
Data yang diperoleh dari *UCI Machine Learning Repository* akan di pisahkan menjadi beberapa skenario pengujian hal ini diperlukan karena pada tahap skenario pengujian, perbaikan struktur data input dilakukan dengan tujuan untuk memaksimalkan hasil. Proses dalam skema pengujian disesuaikan dengan karakteristik data yang akan diinputkan, mengikuti kebutuhan yang spesifik dari setiap jenis data.

Skenario pengujian adalah rangkaian langkah atau situasi yang dirancang secara terencana untuk menguji fungsionalitas, kinerja, atau karakteristik suatu sistem, aplikasi, atau perangkat lunak. Manfaat dari skema itu sendiri adalah data mudah dimengerti, mengurangi waktu proses datamining, memudahkan proses machine learning dan analisa data (Indrayuni, 2019). Ada berbagai macam tahapan yang dapat dilakukan dalm proses skema pengujian, namun pada penelitian ini peneliti akan melakukan 4 skema pengujian yang dilakukan dengan meggunakan *split* data.

Proses pemisahan atau *splitting* data menjadi data latih (*training* data) dan data uji (*testing* data) sangatlah penting dalam pemodelan data *machine learning*. Langkah ini krusial karena model harus dibangun dan dievaluasi dengan dataset yang berbeda. Pembagian ini diperlukan untuk mencegah *overfitting*, di mana model mungkin memperoleh hasil yang sangat baik saat diuji dengan data latih namun memiliki kinerja yang buruk saat diuji dengan dataset yang tidak terlibat

dalam proses pelatihan. Data latih digunakan untuk membangun model, sedangkan data uji berperan dalam mengevaluasi performa model dengan memastikan kecocokan dan generalisasi yang tepat dari model yang telah dibuat dari split data yang sesuai.

Perbandingan pembagian data latih dan data uji pada sistem ini diimplementasikan melalui empat skenario pengujian, masing-masing dengan rasio perbandingan yang berbeda, yaitu pengujian A dengan rasio 90% data *train* : 10% data *test*, pengujian B dengan perbandingan 80% data *train* : 20% data *test*, pengujian C dengan perbandingan 75% data *train* : 25% data *test*, dan pengujian D dengan perbandingan 70% data *train* : 30% data *test*. Hasil dari pembagian data untuk pengujian A adalah 513 data *train* dan 56 data *test*, pengujian B menghasilkan 456 data *train* dan 113 data *test*, sedangkan pengujian C menghasilkan 427 data *train* dan 142 data *test*. Pengujian D seharusnya menghasilkan 399 data *train* dan 170 data *test*.



Gambar 3.3 Skenario Uji Coba

Dengan melakukan pembagian dataset menjadi beberapa subset yang berbeda, dapat diperoleh nilai prediksi yang paling akurat. Setiap pengujian akan dievaluasi berdasarkan modelnya masing-masing yang akan menghasilkan tabel confusion matrix. Tabel 3.12 selanjutnya akan menampilkan representasi dari confusion matrix tersebut.

Tabel 3.12 Confusion Matrix

Aktual	Prediksi	
	Positif	Negatif
Positif	TP	FP
Negatif	FN	TN

Dimana:

TP (*True Positive*) : prediksi kanker dan aslinya kanker

TN (*True Negative*) : prediksi non-kanker dan sebenarnya bukan kanker

FP (*False Positive*) : prediksi kanker dan aslinya bukan kanker

FN (*False Negative*) : prediksi non-kanker dan aslinya merupakan kanker

Confusion Matrix digunakan untuk membandingkan kelas data sebenarnya dengan hasil prediksi. Hasil dari confusion matrix digunakan untuk menghitung tingkat akurasi, presisi, *recall*, dan *f-measure* sebagai evaluasi performa model klasifikasi.

Akurasi adalah salah satu metrik evaluasi dasar yang digunakan untuk mengukur seberapa baik model klasifikasi dalam membuat prediksi yang benar secara keseluruhan dari keseluruhan kasus yang diamati. Akurasi memberikan persentase prediksi yang benar secara keseluruhan dari seluruh prediksi yang dilakukan oleh model. Presisi adalah metrik evaluasi yang mengukur seberapa tepat model dalam mengidentifikasi kasus positif dari semua kasus yang diprediksi sebagai positif oleh model. Metrik ini berguna ketika fokus utama adalah pada seberapa banyak dari prediksi positif yang benar-benar relevan. Berikut perhitungan performa untuk mendapatkan hasil prediksi. Presisi memberikan informasi tentang seberapa baik model dalam menghindari membuat kesalahan dengan memprediksi kasus negatif sebagai kasus positif. Recall, yang juga dikenal sebagai sensitivitas, adalah metrik evaluasi yang mengukur kemampuan model dalam menemukan semua kasus positif yang sebenarnya dalam dataset. Recall memberikan informasi tentang seberapa baik model dapat mengenali kasus positif. F-measure (F1-score) adalah metrik gabungan yang

menggabungkan presisi dan recall menjadi satu nilai tunggal. Ini berguna saat Anda ingin menemukan keseimbangan antara presisi dan recall dalam model klasifikasi. Berikut perhitungan performa untuk mendapatkan hasil prediksi.

$$Akurasi = \frac{TP+TN}{TP+FP+TN+FN} \quad (3.6)$$

$$Presisi = \frac{TP}{TP+FP} \quad (3.7)$$

$$Recall = \frac{TP}{TP+FN} \quad (3.8)$$

$$F - measure = \frac{2\ x\ Presisi\ x\ Recall}{Presisi+Recall} \quad (3.9)$$

BAB IV
HASIL DAN PEMBAHASAN

4.1 Hasil Uji Coba

Pada sub bab ini, dijelaskan dengan rinci analisis hasil dari pengujian sistem berdasarkan skenario pengujian yang telah disusun, bertujuan untuk menilai performa metode *Naïve Bayes* dalam klasifikasi kanker payudara menggunakan 569 dataset yang telah dipisahkan ke dalam empat skenario pengujian. Masing-masing pengujian perlu di evaluasi, ini mencakup penjabaran nilai akurasi dari setiap pengujian dalam menggunakan metode *Naïve Bayes*. Dengan demikian, evaluasi sistem yang terbangun menjadi krusial untuk memahami secara mendalam bagaimana kinerja metode *Naïve Bayes* dalam konteks klasifikasi kanker payudara.

4.1.1 Pengujian A

Pada tahap ini, sesuai dengan skenario pengujian yang telah disusun penggunaan data yang terdiri dari 569 dataset akan dibagi menjadi subset data latih dan data uji dengan perbandingan rasio 90:10, yang menghasilkan 513 data untuk latihan dan 56 data untuk pengujian. Hasil dari pengujian ini kemudian mengungkapkan hal-hal sebagai berikut.

Tabel 4.1 Hasil Parameter Pengujian A Kelas 0

Atribut/Fitur	Mean Kelas 0	Std Deviasi Kelas 0
Radius_mean	12.173842	1.793027
Texture_mean	17.894921	3.985937
Perimeter_mean	78.248265	11.894030
Area_mean	465.089590	135.785508

Smoothness_mean	0.091880	0.013062
Compactness_mean	0.079642	0.034333
Concavity_mean	0.046449	0.044886
Concave Points_mean	0.025466	0.015716
Symmetry_mean	0.173081	0.024648
Fractial Dimension_mean	0.062783	0.006781

Tabel 4.2 Hasil Parameter Pengujian A Kelas 1

Atribut/Fitur	Mean Kelas 1	Standard Deviasi Kelas 1
Radius_mean	17.420872	3.269915
Texture_mean	21.572359	3.797949
Perimeter_mean	115.079128	22.250138
Area_mean	975.335897	376.680538
Smoothness_mean	0.102842	0.012760
Compactness_mean	0.145239	0.054032
Concavity_mean	0.160762	0.074483
Concave Points_mean	0.087377	0.033938
Symmetry_mean	0.193436	0.027937
Fractial Dimension_mean	0.062737	0.007754

Dalam Tabel 4.1, terdapat hasil parameter yang dihitung dari data pelatihan untuk setiap fitur yang terkait dengan kelas 0 (Kanker Jinak), sedangkan Tabel 4.2 menampilkan hasil parameter yang dihitung dari data pelatihan untuk setiap fitur yang terkait dengan kelas 1 (Kanker Ganas). Kedua tabel tersebut didasarkan pada perbandingan rasio 90:10.

Tabel 4.3 Confusion Matrix Pengujian A

	Data Aktual	
Prediksi	**B**	**M**
B	40	0
M	2	15

Dari Tabel 4.3, terlihat bahwa pada pengujian A dengan menggunakan metode *Gaussian Naïve Bayes*, terdapat prediksi sebanyak 40 data kanker payudara yang diprediksi sebagai jinak dan sesuai dengan hasil aktualnya yang juga terdeteksi sebagai kanker payudara jinak. Model *Gaussian Naïve Bayes* juga memprediksi 15 data kanker payudara sebagai ganas dan hasilnya benar terdeteksi sebagai kanker payudara ganas. Namun, terdapat kesalahan di mana model *Gaussian Naïve Bayes* memprediksi tidak ada data kanker payudara jinak yang ternyata terdeteksi sebagai kanker payudara ganas, serta model tersebut juga memprediksi 2 data kanker payudara ganas yang ternyata adalah kanker payudara jinak.

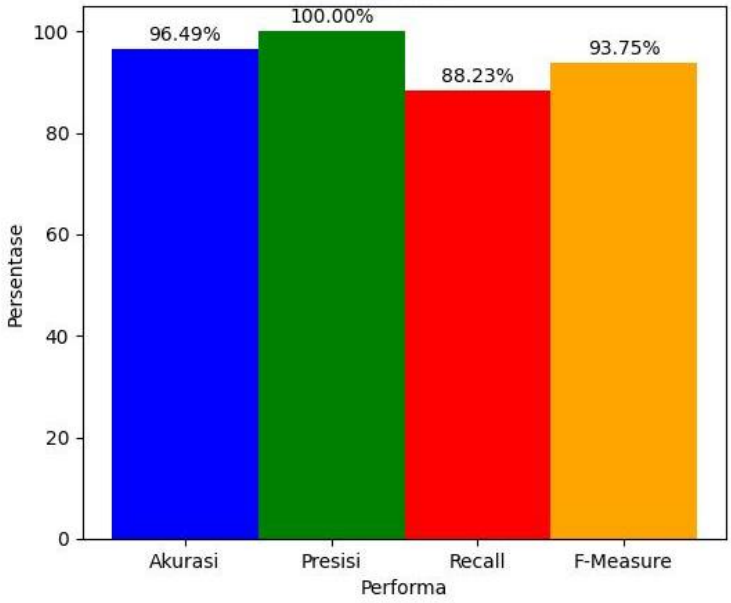
Perhitungan performa dari pengujian A berdasarkan tabel 4.3 dapat diketahui sebagai berikut.

$$Akurasi = \frac{40 + 15}{40 + 2 + 15 + 0} = 96.49\%$$

$$Presisi = \frac{15}{15 + 0} = 100\%$$

$$Recall = \frac{15}{15 + 2} = 88.23\%$$

$$F - measure = \frac{2 \times 100 \times 88.23}{100 + 88.23} = 93.75\%$$



Gambar 4.1 Grafik Performa Pengujian A

4.1.2 Pengujian B

Pada tahap ini, sesuai dengan skenario pengujian yang telah disusun, penggunaan data yang terdiri dari 569 dataset akan dibagi menjadi subset data latih dan data uji dengan perbandingan rasio 80:20, yang menghasilkan 456 data untuk latihan dan 113 data untuk pengujian. Hasil dari pengujian ini kemudian mengungkapkan hal-hal sebagai berikut..

Tabel 4.4 Hasil Parameter Pengujian B Kelas 0

Atribut/Fitur	Mean Kelas 0	Std Deviasi Kelas 0
Radius_mean	12.168056	1.810039
Texture_mean	17.821643	3.895270
Perimeter_mean	78.214091	11.981704
Area_mean	464.910839	137.338990
Smoothness_mean	0.091733	0.013056
Compactness_mean	0.079826	0.035065
Concavity_mean	0.047205	0.046283

Concave Points_mean	0.025540	0.015585
Symmetry_mean	0.173751	0.024742
Fractial Dimension_mean	0.062836	0.006967

Tabel 4.5 Hasil Parameter Pengujian B Kelas 1

Atribut/Fitur	Mean Kelas 1	Standard Deviasi Kelas 1
Radius_mean	17.416923	3.287343
Texture_mean	21.492308	3.862681
Perimeter_mean	115.012959	22.334525
Area_mean	975.013609	379.380612
Smoothness_mean	0.102531	0.012687
Compactness_mean	0.143885	0.052531
Concavity_mean	0.159455	0.073925
Concave Points_mean	0.086764	0.033651
Symmetry_mean	0.193533	0.027493
Fractial Dimension_mean	0.062623	0.007623

Dalam Tabel 4.4, terdapat hasil parameter yang dihitung dari data pelatihan untuk setiap fitur yang terkait dengan kelas 0 (Kanker Jinak), sedangkan Tabel 4.5 menampilkan hasil parameter yang dihitung dari data pelatihan untuk setiap fitur yang terkait dengan kelas 1 (Kanker Ganas). Kedua tabel tersebut didasarkan pada perbandingan rasio 80:20.

Tabel 4.6 Confusion Matrix Pengujian B

	Data Aktual	
Prediksi	**B**	**M**
B	70	1
M	5	38

Dari Tabel 4.6, terlihat bahwa pada pengujian B dengan menggunakan metode *Gaussian Naïve Bayes*, terdapat prediksi sebanyak 70 data kanker payudara yang diprediksi sebagai jinak dan sesuai dengan hasil aktualnya yang

juga terdeteksi sebagai kanker payudara jinak. Model *Gaussian Naïve Bayes* juga memprediksi 38 data kanker payudara sebagai ganas dan hasilnya benar terdeteksi sebagai kanker payudara ganas. Namun, terdapat kesalahan di mana model *Gaussian Naïve Bayes* memprediksi 1 data kanker payudara jinak yang ternyata terdeteksi sebagai kanker payudara ganas, serta model tersebut juga memprediksi 5 data kanker payudara ganas yang ternyata adalah kanker payudara jinak.

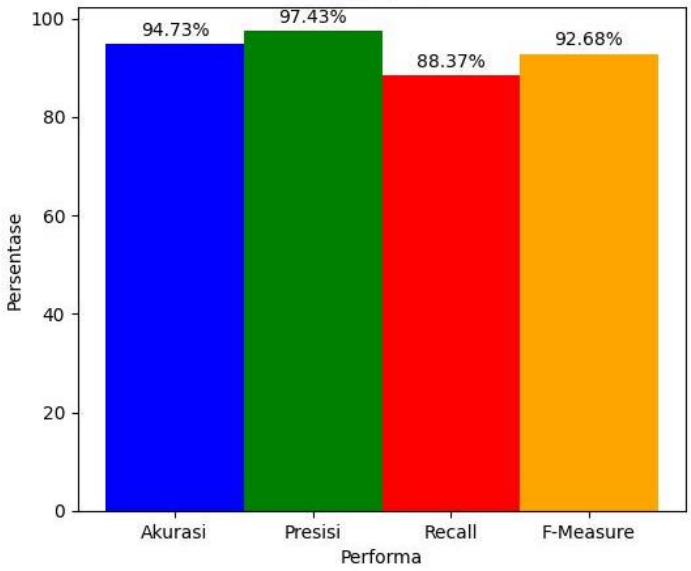
Perhitungan performa dari pengujian B berdasarkan tabel 4.6 dapat diketahui sebagai berikut.

$$Akurasi = \frac{38 + 70}{38 + 70 + 1 + 5} = 94.73\%$$

$$Presisi = \frac{38}{38 + 1} = 97.43\%$$

$$Recall = \frac{38}{38 + 5} = 88.37\%$$

$$F - measure = \frac{2 \times 97 \times 88}{97 + 88} = 92.68\%$$



Gambar 4.2 Grafik Performa Pengujian B

4.1.3 Pengujian C

Pada tahap ini, sesuai dengan skenario pengujian yang telah disusun, penggunaan data yang terdiri dari 569 dataset akan dibagi menjadi subset data latih dan data uji dengan perbandingan rasio 75:25, yang menghasilkan 427 data untuk latihan dan 142 data untuk pengujian. Hasil dari pengujian ini kemudian mengungkapkan hal-hal sebagai berikut.

Tabel 4.7 Hasil Parameter Pengujian C Kelas 0

Atribut/Fitur	Mean Kelas 0	Standard Deviasi Kelas 0
Radius_mean	12.187929	1.800149
Texture_mean	17.843918	3.941545
Perimeter_mean	78.334291	11.909995
Area_mean	466.252612	136.649498
Smoothness_mean	0.091588	0.013041
Compactness_mean	0.079486	0.035213
Concavity_mean	0.046227	0.044189
Concave Points_mean	0.025253	0.015400
Symmetry_mean	0.172574	0.023839
Fractial Dimension_mean	0.062761	0.007022

Tabel 4.8 Hasil Parameter Pengujian C Kelas 1

Atribut/Fitur	Mean Kelas 1	Standard Deviasi Kelas 1
Radius_mean	17.404367	3.329199
Texture_mean	21.497089	3.685571
Perimeter_mean	114.869810	22.642985
Area_mean	974.936709	385.156188
Smoothness_mean	0.102288	0.012909
Compactness_mean	0.141554	0.051217
Concavity_mean	0.157428	0.074693
Concave Points_mean	0.085893	0.033897

Symmetry_mean	0.192663	0.026925
Fractial Dimension_mean	0.062345	0.007578

Dalam Tabel 4.7, terdapat hasil parameter yang dihitung dari data pelatihan untuk setiap fitur yang terkait dengan kelas 0 (Kanker Jinak), sedangkan Tabel 4.8 menampilkan hasil parameter yang dihitung dari data pelatihan untuk setiap fitur yang terkait dengan kelas 1 (Kanker Ganas). Kedua tabel tersebut didasarkan pada perbandingan rasio 75:25.

Tabel 4.9 Confusion Matrix Pengujian C

	Data Aktual	
Prediksi	**B**	**M**
B	85	4
M	5	49

Dari Tabel 4.9, terlihat bahwa pada pengujian C dengan menggunakan metode *Gaussian Naïve Bayes*, terdapat prediksi sebanyak 85 data kanker payudara yang diprediksi sebagai jinak dan sesuai dengan hasil aktualnya yang juga terdeteksi sebagai kanker payudara jinak. Model *Gaussian Naïve Bayes* juga memprediksi 49 data kanker payudara sebagai ganas dan hasilnya benar terdeteksi sebagai kanker payudara ganas. Namun, terdapat kesalahan di mana model *Gaussian Naïve Bayes* memprediksi 4 data kanker payudara jinak yang ternyata terdeteksi sebagai kanker payudara ganas, serta model tersebut juga memprediksi 5 data kanker payudara ganas yang ternyata adalah kanker payudara jinak.

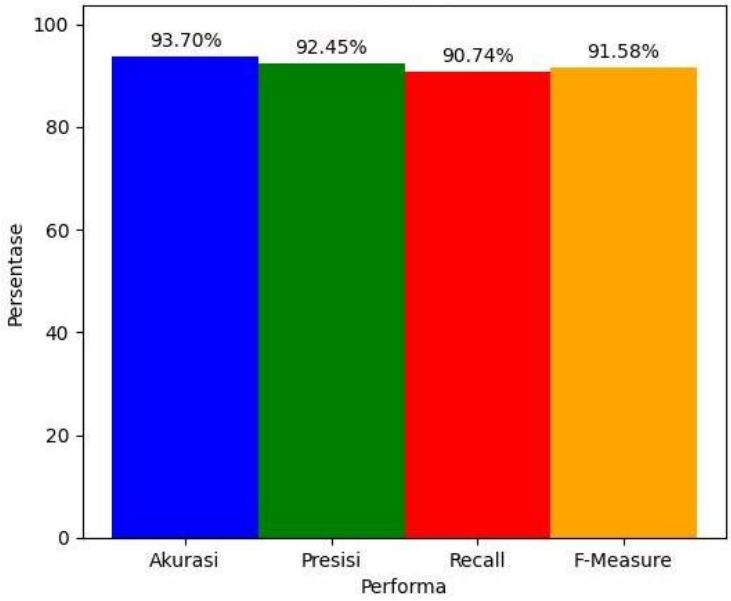
Perhitungan performa dari pengujian C berdasarkan tabel 4.9 dapat diketahui sebagai berikut.

$$Akurasi = \frac{49 + 85}{49 + 4 + 85 + 5} = 93.70\%$$

$$Presisi = \frac{49}{49 + 4} = 92.45\%$$

$$Recall = \frac{49}{49 + 5} = 90.74\%$$

$$F - measure = \frac{2 \times 92 \times 90}{92 + 90} = 91.58\%$$



Gambar 4.3 Grafik Pengujian C

4.1.4 Pengujian D

Pada tahap ini, sesuai dengan skenario pengujian yang telah disusun, penggunaan data yang terdiri dari 569 dataset akan dibagi menjadi subset data latih dan data uji dengan perbandingan rasio 70:30, yang menghasilkan 399 data untuk latihan dan 170 data untuk pengujian. Hasil dari pengujian ini kemudian mengungkapkan hal-hal sebagai berikut.

Tabel 4.10 Hasil Parameter Pengujian D Kelas 0

Atribut/Fitur	Mean Kelas 0	Standard Deviasi Kelas 0
Radius_mean	12.228590	1.775947
Texture_mean	17.836948	3.959325
Perimeter_mean	78.619880	11.740349
Area_mean	468.995582	135.771438
Smoothness_mean	0.092163	0.012925
Compactness_mean	0.080261	0.035408
Concavity_mean	0.046494	0.044404
Concave Points_mean	0.025704	0.015453
Symmetry_mean	0.173019	0.024196
Fractial Dimension_mean	0.062844	0.007067

Tabel 4.11 Hasil Parameter Pengujian D Kelas 1

Atribut/Fitur	Mean Kelas 1	Standard Deviasi Kelas 1
Radius_mean	17.430604	3.347324
Texture_mean	21.368792	3.675687
Perimeter_mean	115.044765	22.714357
Area_mean	978.583221	388.594533
Smoothness_mean	0.102310	0.012789
Compactness_mean	0.141810	0.050471
Concavity_mean	0.157552	0.071965
Concave Points_mean	0.086434	0.033040
Symmetry_mean	0.192489	0.026777
Fractial Dimension_mean	0.062364	0.007556

Dalam Tabel 4.10, terdapat hasil parameter yang dihitung dari data pelatihan untuk setiap fitur yang terkait dengan kelas 0 (Kanker Jinak), sedangkan Tabel 4.11 menampilkan hasil parameter yang dihitung dari data pelatihan untuk

setiap fitur yang terkait dengan kelas 1 (Kanker Ganas). Kedua tabel tersebut didasarkan pada perbandingan rasio 70:30.

Tabel 4.12 Confusion Matrix Pengujian D

	Data Aktual	
Prediksi	**B**	**M**
B	**104**	**4**
M	**8**	**55**

Dari Tabel 4.12, terlihat bahwa pada pengujian D dengan menggunakan metode *Gaussian Naïve Bayes*, terdapat prediksi sebanyak 104 data kanker payudara yang diprediksi sebagai jinak dan sesuai dengan hasil aktualnya yang juga terdeteksi sebagai kanker payudara jinak. Model *Gaussian Naïve Bayes* juga memprediksi 55 data kanker payudara sebagai ganas dan hasilnya benar terdeteksi sebagai kanker payudara ganas. Namun, terdapat kesalahan di mana model *Gaussian Naïve Bayes* memprediksi 4 data kanker payudara jinak yang ternyata terdeteksi sebagai kanker payudara ganas, serta model tersebut juga memprediksi 8 data kanker payudara ganas yang ternyata adalah kanker payudara jinak.

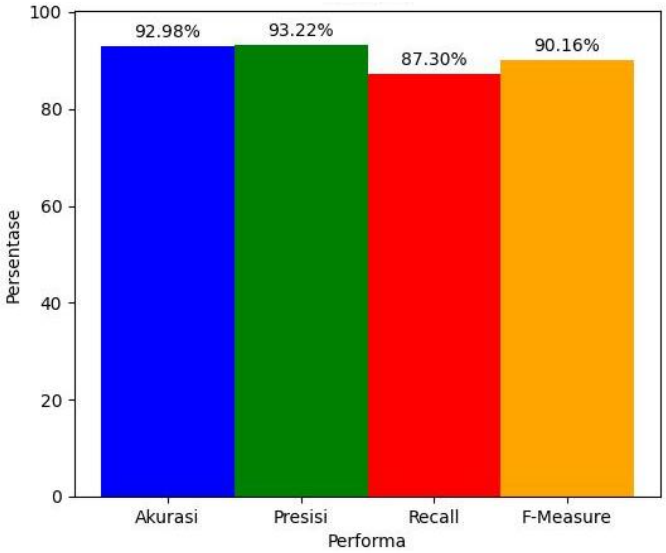
Perhitungan performa dari pengujian D berdasarkan tabel 4.12 dapat diketahui sebagai berikut.

$$Akurasi = \frac{55 + 104}{55 + 4 + 104 + 8} = 92.98\%$$

$$Presisi = \frac{55}{55 + 4} = 93.22\%$$

$$Recall = \frac{55}{55 + 8} = 87.30\%$$

$$F - measure = \frac{2 \times 93 \times 87}{93 + 87} = 90.16\%$$



Gambar 4.4 Grafik Performa Pengujian D

4.1.5 Pengujian 10-Fold Cross Validation

Data sampel dibagi menjadi K subdivisi yang berbeda, di mana setiap subdivisi berfungsi sebagai subset pelatihan atau subset pengujian. Dalam setiap iterasi K-fold cross-validation, subdivisi bergantian digunakan sebagai subset pengujian dan pelatihan. Misalnya, pada iterasi pertama, satu subdivisi digunakan sebagai subset pengujian dan subdivisi lainnya digunakan sebagai subset pelatihan. Proses ini diulang hingga semua subdivisi berfungsi sebagai subset pengujian sekali, dengan demikian, data dipartisi dengan cara ini untuk melakukan validasi silang K-Fold. Secara khusus, proses pelatihan model diulang sebanyak "k" kali, dan kinerja model dihitung sebagai rata-rata dari keseluruhan proses pelatihan (Oktafiani et al., 2023).

4.1.5.1 Pengujian A Dengan 10-Fold

Didapatkan nilai akurasi pengujian A dengan menggunakan K-Fold cross-validation, dengan ketentuan K=10, yang terperinci tercantum pada tabel berikut.

Tabel 4.13 Hasil 10-Fold Pengujian A

Iterasi	Nilai Akurasi (%)
K-1	94.23
K-2	86.53
K-3	92.15
K-4	94.11
K-5	90.19
K-6	90.19
K-7	76.47
K-8	98.03
K-9	90.19
K-10	94.11

Berdasarkan pada tabel 4.13, telah tercatat bahwa nilai akurasi terbaik untuk pengujian A menggunakan metode 10-Fold adalah 98.03%. Analisis mendalam terhadap hasil ini menunjukkan dari nilai akurasi yang tercatat pada setiap iterasi, terdapat variasi yang signifikan dalam performa klasifikasi. Sebagian besar iterasi menunjukkan hasil yang stabil, dengan nilai akurasi berkisar antara 86.53% hingga 94.23%. Namun, terdapat perbedaan drastis pada iterasi ke-7, di mana nilai akurasi menurun secara tajam menjadi 76.47%. Iterasi ke-8 menonjol sebagai titik puncak, mencapai nilai akurasi tertinggi sebesar 98.03%. Analisis atas variasi ini dapat mencakup penyebab dari fluktuasi yang signifikan pada iterasi tertentu, seperti distribusi data yang tidak seimbang atau potensi adanya outlier yang mempengaruhi kinerja model. Kesimpulan dari analisis ini akan membantu dalam memahami faktor-faktor yang memengaruhi

hasil klasifikasi serta mengidentifikasi strategi untuk meningkatkan konsistensi dan performa keseluruhan model.

4.1.5.2 Pengujian B Dengan 10-Fold

Didapatkan nilai akurasi pengujian B dengan menggunakan K-Fold *cross-validation*, dengan ketentuan K=10, yang terperinci tercantum pada tabel berikut.

Tabel 4.14 Hasil 10-Fold Pengujian B

Literasi	Nilai Akurasi (%)
K-1	95.65
K-2	86.95
K-3	93.47
K-4	91.30
K-5	82.60
K-6	86.66
K-7	86.66
K-8	93.33
K-9	91.11
K-10	93.33

Berdasarkan pada tabel 4.14, telah diperoleh nilai akurasi tertinggi pada pengujian B menggunakan metode 10-Fold, yakni sebesar 95.65%. Analisis atas hasil ini mengindikasikan bahwa model yang dikembangkan mampu melakukan klasifikasi dengan tingkat akurasi yang sangat baik, menunjukkan potensi untuk penggunaan lebih lanjut dalam mendukung deteksi dini dan penanganan kanker payudara secara efektif.

4.1.5.3 Pengujian C Dengan 10-Fold

Didapatkan nilai akurasi pengujian C dengan menggunakan K-Fold cross-validation, dengan ketentuan K=10, yang terperinci tercantum pada tabel berikut.

Tabel 4.15 Hasil 10-Fold Pengujian C

Literasi	Nilai Akurasi (%)
K-1	90.69
K-2	95.34
K-3	81.39
K-4	83.72
K-5	96.57
K-6	97.67
K-7	88.09
K-8	90.47
K-9	88.09
K-10	90.47

Berdasarkan tabel 4.15, nilai akurasi terbaik yang diperoleh pada pengujian C dengan menggunakan metode 10-Fold adalah sebesar 97.67%. Analisis dari hasil eksperimen menunjukkan bahwa nilai akurasi untuk pengujian C pada liputan data k-6 mencapai nilai tertinggi, menunjukkan performa yang sangat baik dalam klasifikasi. Sementara itu, variasi nilai akurasi dari setiap liputan data memberikan gambaran tentang fluktuasi performa model pada subset data yang berbeda. Dengan nilai terendah sebesar 81.39% pada k-3, dan nilai tertinggi pada k-6, hal ini menunjukkan variasi yang signifikan dalam kemampuan model ketika diuji pada liputan data yang berbeda.

4.1.5.4 Pengujian D Dengan 10-Fold

Didapatkan nilai akurasi pengujian D dengan menggunakan K-Fold cross-validation, dengan ketentuan K=10, yang terperinci tercantum pada tabel berikut.

Tabel 4.16 Hasil 10-Fold Pengujian D

Literasi	Nilai Akurasi (%)
K-1	87.5
K-2	90
K-3	90
K-4	97.5
K-5	90
K-6	85
K-7	97.5
K-8	92.5
K-9	82.05
K-10	94.87

Berdasarkan tabel 4.16, diperoleh nilai akurasi pada pengujian D menggunakan metode 10-Fold sebesar 97.5%. Analisis terhadap data menunjukkan variasi yang signifikan dalam nilai akurasi antara pengujian yang berbeda (seperti K-1, K-2, K-3, dst.). Terdapat kecenderungan terjadinya fluktuasi yang cukup besar dari 82.05% hingga 97.5%, menunjukkan variasi yang signifikan dalam performa model pada setiap iterasi pengujian. Nilai tertinggi akurasi yang mencapai 97.5% pada pengujian K-4 dan K-7 menandakan konsistensi performa yang tinggi dalam pengenalan dan klasifikasi data. Namun, penting untuk memperhatikan penurunan signifikan pada pengujian K-9 yang

mencapai 82.05%, yang dapat menandakan kemungkinan adanya data outlier atau perlu dilakukannya penyesuaian model pada kondisi tersebut.

4.1.6 Menghitung Bobot Nilai Evaluasi

Analytical Hierarchy Process (AHP) atau Proses Hirarki Analitik (PHA) adalah metode pengukuran yang digunakan untuk menemukan skala rasio terbaik dari perbandingan berpasangan yang diskrit maupun kontinu. AHP sangat cocok dan fleksibel digunakan untuk menentukan keputusan yang efisien dan efektif berdasarkan segala aspek yang dimilikinya. AHP dikembangkan untuk menyusun suatu permasalahan ke dalam suatu hirarki yang selanjutnya dilakukan pembobotan (menentukan prioritas) untuk memilih keputusan terbaik (supriadi, 2018). Dalam penelitian ini metode AHP digunakan untuk menentukan bobot dari nilai akurasi, presisi, recall dan f-mesure. Hasil dari pembototan AHP dapat dilihat pada tabel berikut.

Tabel 4.17 Bobot Prioritas

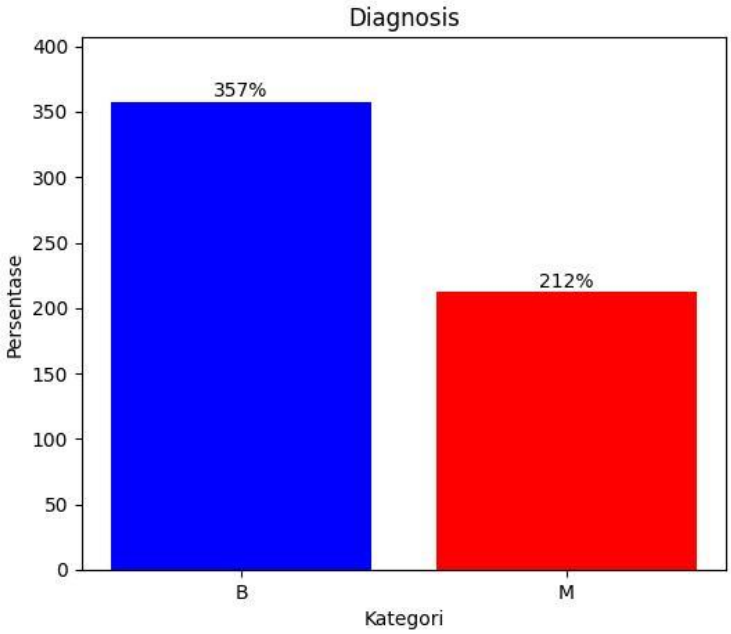
Kriteria	Bobot
Akurasi	0.593
Presisi	0.225
Recall	0.112
F-measure	0.068
N	1

4.2 Pembahasan

Pembahasan pada sub bab ini didasarkan pada hasil uji coba dari sistem yang telah dikembangkan. Data yang diperoleh dari *UCI* terkait *Breast Cancer Wisconsin Diagnostic* menjadi fondasi utama dalam penelitian ini. Dataset ini

terdiri dari 569 sampel dengan 32 atribut yang mencakup berbagai informasi penting terkait karakteristik sel-sel kanker payudara. Atribut-atribut ini termasuk parameter-parameter seperti *radius*, *texture*, *perimeter*, *area*, *smoothness*, *compactness*, *concavity*, *concave points*, *symmetry*, dan *fractal dimension*. Masing-masing atribut utama ini memiliki tiga nilai indikator, yakni *mean*, *standard error* (SE), dan nilai terburuk (*worst*), yang memberikan gambaran yang komprehensif tentang karakteristik sel kanker payudara.

Di antara atribut-atribut tersebut, terdapat dua atribut tambahan, yaitu ID number dan Diagnosis. Atribut Diagnosis menjadi krusial karena mendefinisikan kelas dari setiap sampel data, dengan nilai M (*Malignant*) menandakan kanker ganas dan B (*Benign*) menunjukkan kanker jinak. Analisis terhadap distribusi kelas menunjukkan adanya ketidakseimbangan jumlah sampel antara kelas M dan B, dengan 212 sampel yang terdiagnosis kanker ganas dan 357 sampel yang terdiagnosis kanker jinak. Ketidakseimbangan ini menjadi faktor penting dalam proses pengembangan model klasifikasi untuk memastikan keakuratan dan reliabilitasnya dalam mengidentifikasi jenis kanker payudara. sebagaimana tergambar pada gambar berikut yang mengilustrasikan ketidakseimbangan ini.



Gambar 4.5 Perbandingan Diagnosis

Penelitian yang dilakukan oleh (Diana Dumitru, 2009) dalam memprediksi kanker payudara dengan menggunakan dataset yang sama yaitu *Breast Cancer Wisconsin* dengan menggunakan metode *Naïve Bayes*. Pada jurnal peneliti hanya menggunakan data sebesar 198 dari total 569 data, serta menambahkan 2 attribute tambahan yaitu *Tumor_size* dan *Lymph_ns* dari total 10 atribute utama. Dengan membagi menjadi data training sebesar 132 : data testing sebesar 66 di penelitian yang telah dilakukan, peneliti menghasilkan nilai akurasi sebesar 74%. Peneliti juga mengatakan teknik *Naive Bayes* khususnya cocok ketika dimensi ruang fitur (input) tinggi. Meskipun sederhana, pengklasifikasi Naive Bayes seringkali dapat mengungguli metode klasifikasi yang lebih canggih di beberapa domain tertentu (Diana Dumitru, 2009). Untuk detailnya dapat dilihat pada tabel berikut.

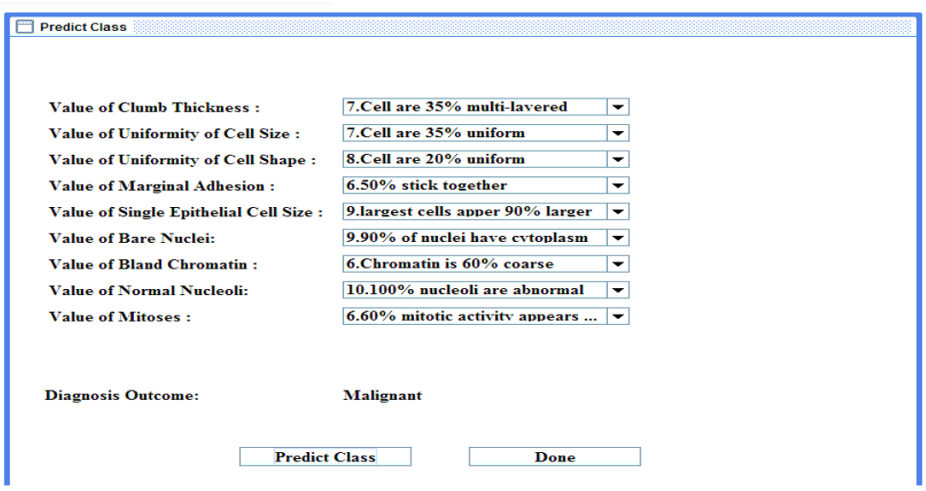
Tabel 4.18 Hasil Penelitian Diana Dumitru

Testing Accuracy (%)	**Testting Specificity (%)**
74.24	91.67

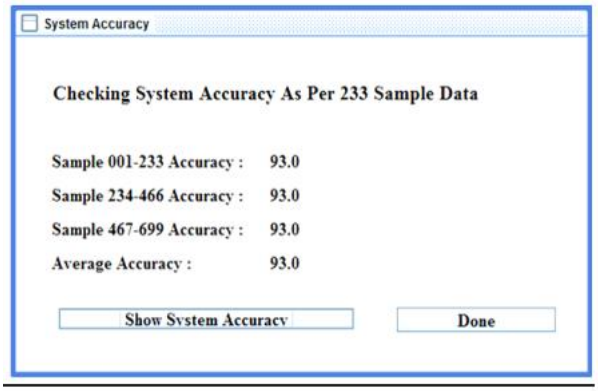
Tabel 4.19 Hasil Confusion Matrix Diana Dumitru

Classification (testing)		**Estimated (Bayesian)**	
Effective classes		Class = R	Class = N
	Class = R	5	13
	Class = N	4	44

Penelitian yang dilakukan Kharya et al., dengan menggunakan dataset yang sama yaitu *Breast Cancer Wisconsin* dengan menggunakan metode Naïve Bayes. Pada jurnal peneliti melakukan normalisasi pada data yang digunakan yaitu normalisasi *range of value* yang berfungsi untuk mengubah rentang nilai dari suatu dataset ke rentang yang diinginkan, juga melakukan *transformation* data di convert ke dalam tabel MySql. Model Klasifikasi dibangun menggunakan Count dan Tabel Probabilitas, untuk pelatihan, seluruh catatan telah digunakan dan pengujian juga dilakukan pada beberapa set data hampir seperempatnya. Dengan melakukan hal tersebut peneliti mendapatkan nilai akurasi sebesar 93% (Kharya et al., 2014). Untuk detailnya dapat dilihat pada gambar berikut.

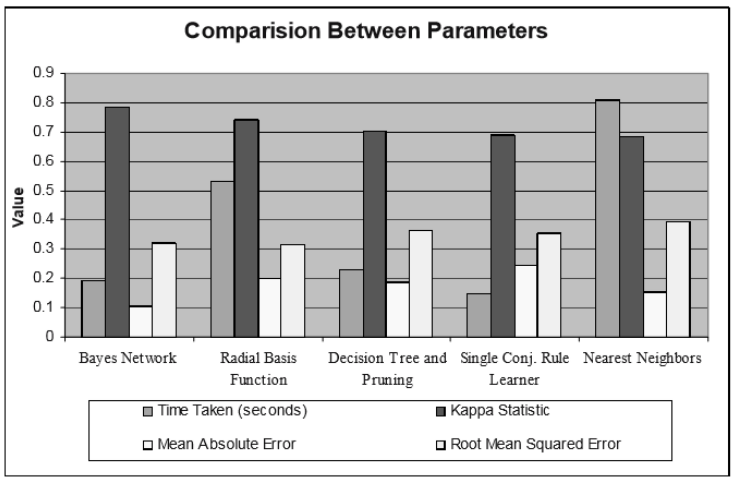


Gambar 4.6 Hasil Penelitian Kharya dkk



Gambar 4.7 Hasil Penelitian Kharya dkk

Penelitian yang dilakukan Othman et al., dengan menggunakan dataset yang sama yaitu *Breast Cancer Wisconsin* dengan menggunakan metode Naïve Bayes, Radial Basis Function, Decision Tree, Single Conjuctive Rule Learner dan Nearest Neighbors Algorithm. Pada jurnal peneliti menggunakan teknik *WEKA* yang merupakan semua data dianggap sebagai instansi, sedangkan fitur-fitur dalam data dikenal sebagai atribut. Dalam penelitian (Othman et al., 2007) ditemukan nilai dari masing-masing metode sebesar 89.71% untuk Naïve Bayes, 87.42% untuk metode Radial Basis Function, 85.71% untuk metode Decision Tree, 85.14% untuk metode SCRL dan 84.57% untuk metode Nearest Neighbors. Untuk detailnya dapat dilihat pada tabel dan gambar berikut.



Gambar 4.8 Hasil Penelitian Diagram Othman dkk

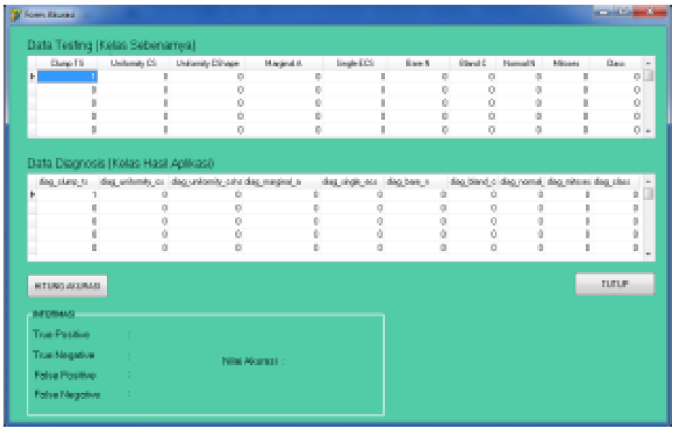
Tabel 4.20 Hasil Penelitian Nilai Akurasi Othman dkk

Algorithm	Correctly Classified Instances % (value)	Incorrectly Classified Instances % (Value)	Time Taken (seconds)	Kappa Statictic
Naïve Bayes	89.7143 (157)	10.2857 (18)	0.19	0.7858
RBF	87.4286 (153)	12.5710 (22)	0.53	0.7404
Decision Tree	85.7143 (150)	14.2857 (25)	0.23	0.7019
SCRL	85.1429 (149)	14.8571 (26)	0.15	0.6893
Nearest Neighbors	84.5714 (148)	15.4286 (27)	0.81	0.9860

Penelitian yang dilakukan oleh Safutra et al., dengan menggunakan dataset yang sama yaitu *Breasr Cancer Wisconsin* serta menggunakan metode Naïve Bayes. Dalam jurnal peneliti melakukan proses diskritisasi data yang dilakukan dengan cara mengkonversi data setiap atribut pada dataset untuk memberikan hasil yang objektif (Safutra et al., 2016). Hasil yang didapatkan pada penelitian ini dengan menggunakan metode Naïve Bayes yaitu nilai akurasi sebesar 98.67%. Untuk detailnya dapat dilihat pada gambar berikut.

Tabel 4.21 Hasil Confusion Matrix Safutra dkk

Kelas Sebenarnya	**Kelas Hasil Prediksi**	
	Kanker Jinak	Kanker Ganas
Kanker Jinak	51	2
Kanker Ganas	1	172

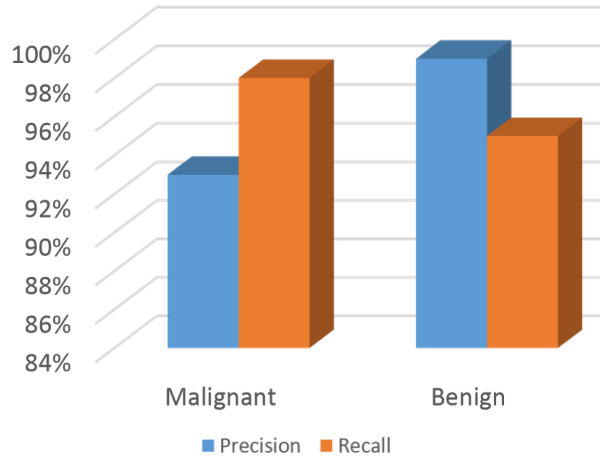


Gambar 4.9 Hasil Penelitian Safutra dkk

Penelitian yang dilakukan oleh Ramadhan et al., dengan menggunakan dataset yang sama yaitu *Breast Cancer Wisconsin* serta menggunakan metode Naïve Bayes. Peneliti menggunakan teknik *SMOTE oversampling* untuk masalah data yang tidak seimbang dan skor *Gini* untuk fitur peringkat. Penelitian ini bertujuan untuk klasifikasi dataset kanker payudara dengan menyelesaikan permasalahan pada proses pre-processing (data tidak seimbang dan pemilihan fitur). Sehingga, dalam penelitian ini menggunakan teknik *SMOTE* dan Gini untuk meningkatkan ketelitian hasil pada kelas jinak (benign) dan juga dapat meningkatkan hasil recall pada kelas ganas (malignan) (Ramadhan et al., 2021). Dan hasil yang diperoleh dalam penelitian ini yaitu nilai akurasi sebesar 96.49% serta nilai presisi sebesar 93% dan nilai *recall* sebesar 95%. Untuk detailnya dapat dilihat pada tabel dan gambar berikut.

Tabel 4.22 Hasil Confusion Matrix Peneliti Ramadhan dkk

	Actual	
Predict	104	4
	2	61



Gambar 4.10 Hasil Penelitian Ramadhan dkk

Penelitian ini telah dilakukan 4 pengujian, diantaranya pengujian akurasi, presisi, *recall* dan *f-measure* dengan menggunakan confusion matrix. Dalam pengujian menggunakan confusion matrix, dilakukan untuk mengevaluasi hasil dari pembagian data atau split data menjadi 4 pengujian, yaitu pengujia A dengan perbandingan 90% data train : 10% data test, pengujian B dengan perbadingan 80% data train : 20% data test, pengujian C dengan perbadingan 75% data train : 25% data test, dan pengujian D dengan perbandingan 70% data train : 30% data test. Dengan pembagian beberapa dataset yang berbeda akan didapatkan nilai ketepatan prediksi terbaik. (Almais et al., 2022). Langkah berikutnya membandingkan metode Naïve Bayes pada data latih dan data uji antara pengujian A dengan pengujian B dan pengujian C serta dengan pengujian D. Untuk

membandingkan antar berbagai model dapat dilakukan berdasarkan dari nilai akurasi, presisi, *recall* dan *f-measure* dari hasil nilai masing-masing pengujian tersebut. Kemudian dibandingkan untuk mencari nilai dari berbagai pengujian mana yang terbaik, setelah model *Naïve Bayes* dilatih, confusion matrix digunakan untuk menunjukkan jumlah prediksi yang benar dan salah dari model Naïve Bayes pada setiap kelas serta untuk mencari nilai akurasi dari sistem, nilai presisi, nilai recall dan nilai f-measure, dibandingkan dengan label kelas aktual dari data uji. Berikut tabel hasil akurasi dari masing-masing pengujian.

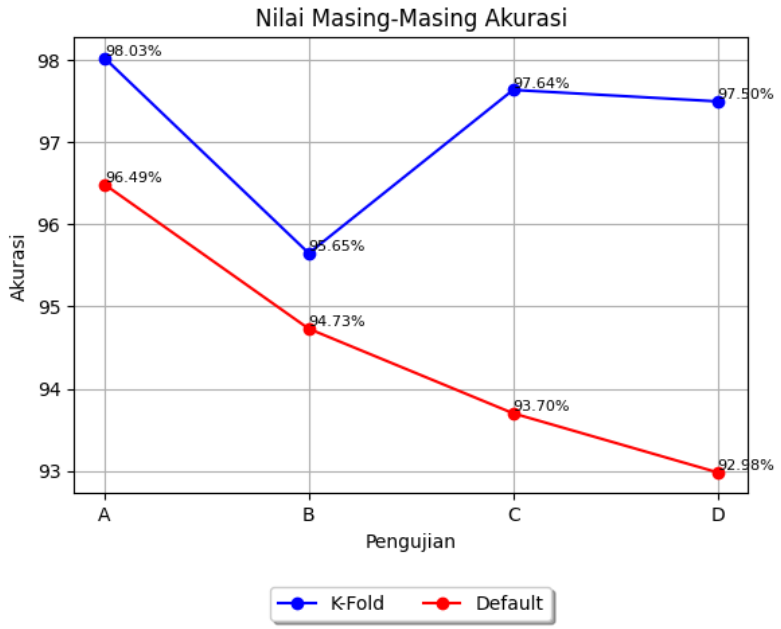
Tabel 4.23 Nilai Akurasi Setiap Pengujian

Pengujian	Jumlah Data = 569				*seed*	Akurasi	
	Train		*Test*				
	Jumlah	precentase	Jumlah	Precentase		Matrix	10-Fold
A	513	90%	56	10%	1234	96.49%	98.03%
B	456	80%	113	20%	1234	94.73%	95.65%
C	427	75%	142	25%	1234	93.70%	97.64%
D	399	70%	170	30%	1234	92.98%	97.5%

Pada Tabel 4.23, model *Naïve Bayes* menunjukkan kinerja yang signifikan dalam dataset yang digunakan, menandakan kemampuan model dalam mengklasifikasikan data secara efektif. Hasil akurasi terbaik, mencapai 98.03%, terlihat pada pengujian A dengan teknik *K-Fold*. Dengan rasio 90% data pelatihan (513 data pelatihan) dan 10% data pengujian (56 data pengujian), model berhasil memberikan hasil yang sangat memuaskan. Penggunaan teknik *K-Fold Cross-Validation* di Tabel 4.23 menunjukkan peningkatan signifikan dalam nilai akurasi. Pendekatan ini menghasilkan estimasi performa model yang lebih stabil dan

meningkatkan generalisasi model karena diuji pada berbagai subset data yang berbeda. Proses ulangan sebanyak K kali dengan subset pengujian yang berbeda dalam setiap iterasi memberikan insight yang lebih kuat terkait kemampuan model. Namun demikian, penting untuk dicatat bahwa *Naïve Bayes* mengasumsikan independensi antar atribut atau ketidaksaling ketergantungan pada nilai atribut kelas lainnya. Visualisasi diagram garis dari masing-masing model memperlihatkan tren yang menarik terkait performa dan peningkatan yang signifikan pada pengujian menggunakan teknik K-Fold.

Analisis ini menyoroti peningkatan performa yang signifikan ketika menggunakan teknik *K-Fold Cross-Validation* serta memberikan pemahaman tambahan tentang asumsi Naïve Bayes terkait independensi antar atribut.



Gambar 4.11 Diagram nilai Akurasi Tiap Pengujian

Berikut merupakan hasil perbandingan nilai akurasi yang didapatkan oleh peneliti dengan peneliti lainnya.

Tabel 4.24 Hasil Perbandingan Akurasi

Referensi	Metode	Perbedaan Pengolahan	Hasil Akurasi
Diana Dumitru	*Naïve Bayes*	Menggunakan 132 data sebagai data *training* dan 66 data sebagai data *test.* Total penggunaan data 198 dari 569 data.	74.24%
Kharya et al.	*Naïve Bayes*	Menggunakan teknik Normalisasi pada data yang digunakan yaitu *Range of Values*	93%
Othman et al.	*Naïve Bayes, RBF, Decision Tree, SCRL* dan *Nearest Neighbors*	Menggunakan teknik *WEKA* pada data yang dugunakan, menggunakan rasio perbandingan 75% data *training* : 25% data *test*	89.71% *Naïve Bayes*
Safutra et al.	*Naïve Bayes*	Melakukan teknik *Diskritasi Data*	98.67%
Ramadhan et al.	*Naïve Bayes*	Menggunakan teknik *SMOTE* dan *Gini Score*	96.49%
Peneliti	*Naïve Bayes*	Memisah data menjadi 4 skenario pengujian dengan perbandingan yang berbeda-beda.	98.03%

Pada tabel 4.24, terlihat beragam teknik yang digunakan oleh peneliti lain. Mulai dari teknik Range of Values yang mengubah rentang nilai dataset menjadi rentang yang diinginkan, biasanya 0-1, hingga penggunaan WEKA, Disktritasi Data untuk mengkonversi data setiap atribut dalam dataset, teknik SMOTE untuk menyeimbangkan data, dan G-Score yang digunakan untuk menormalisasi skala dataset dengan rata-rata 0 dan standar deviasi 1. Analisis rinci terkait pengolahan data berikut mencakup perbandingan metode, referensi, perbedaan dalam pengolahan, dan hasil akurasi masing-masing metode dalam tabel.

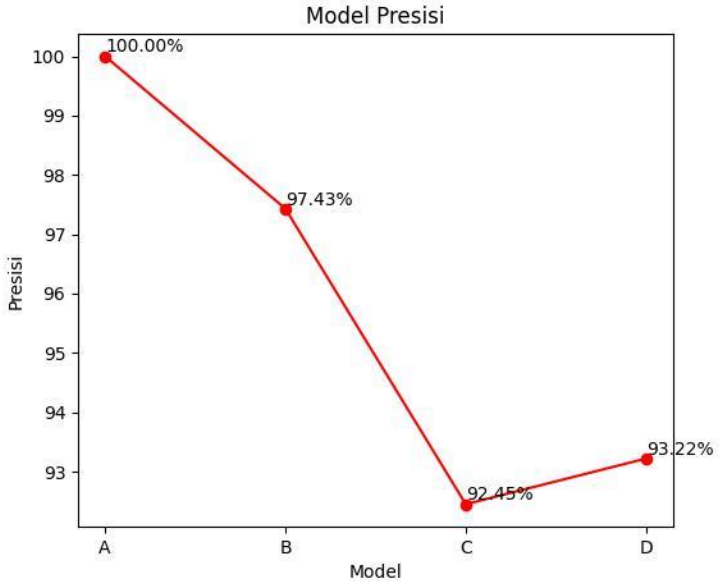
Setelah melatih model Naïve Bayes, confusion matrix digunakan untuk memvisualisasikan jumlah prediksi yang benar dan yang salah dari model tersebut pada setiap kelas, serta untuk menghasilkan nilai evaluasi model yang telah dibuat, dengan membandingkan label kelas aktual pada data uji. Tabel hasil berikut memperlihatkan nilai presisi dari setiap pengujian untuk memberikan gambaran yang lebih jelas.

Tabel 4.25 Nilai Presisi Setiap Pengujian

Pengujian	Banyaknya Data = 569				Seed	Presisi
	Train		Test			
	Jumlah	Presentase	Jumlah	Presentase		
A	513	90%	56	10%	1234	100%
B	456	80%	113	20%	1234	97.43%
C	427	75%	142	25%	1234	92.45%
D	399	70%	170	30%	1234	93.22%

Pada tabel 4.25 model Naïve Bayes menunjukkan performa yang sangat baik pada dataset yang digunakan, menandakan kemampuan model untuk mengklasifikasikan data dengan presisi tinggi. Dalam analisis pengujian yang

dilakukan, terdapat beberapa perbedaan pada proporsi data training dan testing yang digunakan dalam masing-masing pengujian A, B, C, dan D. Dalam pengujian A, proporsi 90% data training dan 10% data testing, menghasilkan nilai presisi yang optimal sebesar 100%. Namun, pada pengujian B hingga D, dengan peningkatan proporsi data testing, terjadi penurunan nilai presisi meskipun tetap dalam kisaran yang tinggi, menunjukkan sensitivitas model terhadap perubahan proporsi data. Pengujian D menunjukkan peningkatan kembali dalam presisi meskipun menggunakan proporsi yang lebih rendah dari data training, mungkin disebabkan oleh distribusi yang lebih seimbang atau keberuntungan tertentu pada seed yang digunakan dalam pengujian tersebut. Perlu diperhatikan bahwa fluktuasi nilai presisi ini dapat diakibatkan oleh skala nilai yang bervariasi antar atribut, sesuai dengan asumsi Naïve Bayes tentang independensi antar atribut. Ini menyoroti pentingnya pemilihan proporsi data training dan testing yang tepat untuk mempertahankan presisi model yang dihasilkan.



Gambar 4.12 Diagram nilai Presisi Tiap Pengujian

Setelah melatih model Naïve Bayes, confusion matrix digunakan untuk memvisualisasikan jumlah prediksi yang benar dan yang salah dari model tersebut pada setiap kelas, serta untuk menghasilkan nilai evaluasi model yang telah dibuat, dengan membandingkan label kelas aktual pada data uji. Tabel hasil berikut memperlihatkan nilai *recall* dari setiap pengujian untuk memberikan gambaran yang lebih jelas.

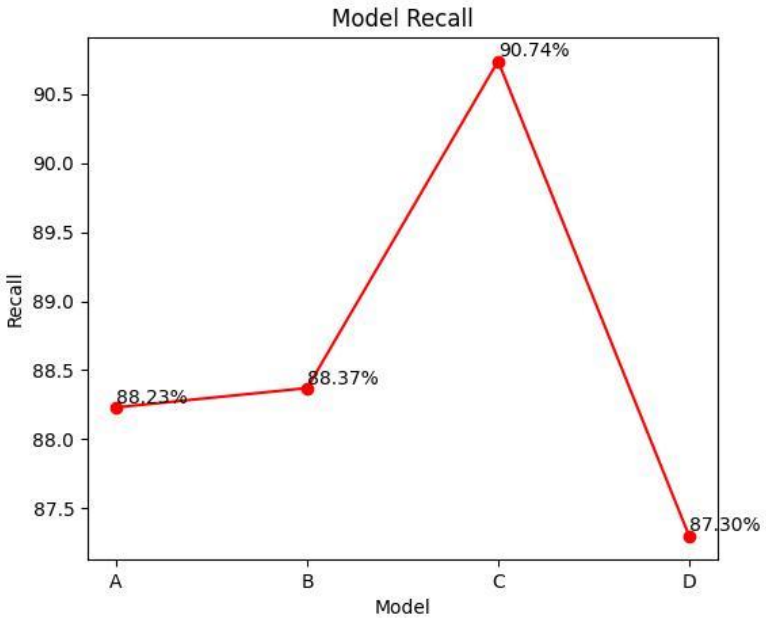
Tabel 4.26 Nilai Recall Setiap Pengujian

Pengujian	Banyaknya Data = 569				*Seed*	*Recall*
	Train		*Test*			
	Jumlah	Presentase	Jumlah	Presentase		
A	513	90%	56	10%	1234	88.23%
B	456	80%	113	20%	1234	88.37%
C	427	75%	142	25%	1234	90.74%
D	399	70%	170	30%	1234	87.30%

Pada Tabel 4.26 model *Naïve Bayes* memberikan performa yang baik pada dataset yang digunakan, hal ini menunjukkan kemampuan model untuk mengklasifikasikan data dengan baik. Hasil nilai Recall tertinggi diperoleh dari pengujian C yang telah mendapatkan nilai berdasarkan confusion matrix, dengan menggunakan rasio perbandingan 75% data train (427 data train) dan data test 25% (142 data test) diperoleh hasil nilai recall sebesar 90.74%.

Dapat dilihat pada Tabel 4.26 hasil dari nilai recal mengalami kenaikan dan pada pengujian D mengalami penurunan nilai. Filosofi di balik *recall* adalah memberikan gambaran tentang seberapa baik model dapat mengidentifikasi semua instance yang benar positif dari suatu kelas tertentu (misalnya, dalam mengidentifikasi semua pasien yang benar-benar menderita penyakit kanker

payudara) relatif terhadap keseluruhan instance yang sebenarnya positif. Ini sangat penting dalam kasus-kasus di mana kesalahan prediksi positif palsu (False Negative) bisa menjadi kritis, seperti dalam kasus diagnosa medis di mana tidak mengidentifikasi penyakit kanker payudara ganas dapat memiliki konsekuensi yang serius. Jadi, *recall* membantu kita memahami seberapa baik model dapat menangkap instance yang benar positif dari kelas yang kita perhatikan. hal ini dikarenakan data memiliki skala nilai yang beragam. Dapat diketahui bahwa dalam Naïve Bayes mengasumsikan semua atribut independet atau tidak saling ketergantungan pada nilai atribut kelas lainnya. Berikut visualisasi diagram garis dari masing-masing pengujian berdasarkan nilai recall.



Gambar 4.13 Diagram nilai Recall Tiap Pengujian

Setelah melatih model Naïve Bayes, confusion matrix digunakan untuk memvisualisasikan jumlah prediksi yang benar dan yang salah dari model tersebut pada setiap kelas, serta untuk menghasilkan nilai evaluasi model yang telah

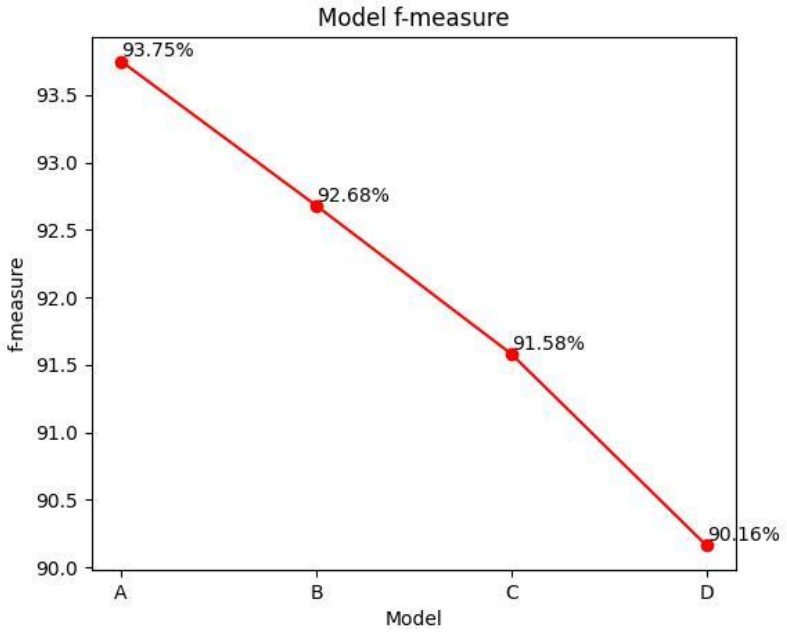
dibuat, dengan membandingkan label kelas aktual pada data uji. Tabel hasil berikut memperlihatkan nilai *f-measure* dari setiap pengujian untuk memberikan gambaran yang lebih jelas.

Tabel 4.27 Nilai F-measure Setiap Pengujian

Pengujian	Banyaknya Data = 569				*Seed*	F-measure
	Train		*Test*			
	Jumlah	Presentase	Jumlah	Presentase		
A	513	90%	56	10%	1234	93.75
B	456	80%	113	20%	1234	92.68
C	427	75%	142	25%	1234	91.58
D	399	70%	170	30%	1234	90.16

Pada Tabel 4.27 model *Naïve Bayes* memberikan performa yang stabil pada setiap skenario pengujian yang dilakukan, hal ini menunjukkan kemampuan model untuk mengklasifikasikan data dengan baik. Dalam analisis pengujian yang dilakukan dengan mempertimbangkan nilai F-measure pada masing-masing pengujian A, B, C, dan D, terlihat pola yang menunjukkan konsistensi performa model. Meskipun proporsi data training dan testing bervariasi antar pengujian, terdapat penurunan nilai F-measure secara bertahap seiring peningkatan proporsi data testing. Pengujian A, dengan proporsi 90% data training dan 10% data testing, menghasilkan nilai tertinggi pada F-measure sebesar 93.75, namun terjadi penurunan seiring dengan peningkatan proporsi data testing pada pengujian B hingga D. Hal ini menandakan bahwa model cenderung lebih baik dalam mengklasifikasikan data pada proporsi yang lebih tinggi dari data training. Penurunan performa ini mungkin disebabkan oleh kurangnya informasi yang relevan yang diperoleh oleh model dari data testing yang lebih sedikit. Meskipun

demikian, nilai-nilai F-measure yang tetap relatif tinggi pada semua pengujian menunjukkan konsistensi model dalam melakukan klasifikasi, meskipun perlu perhatian lebih lanjut terkait peningkatan proporsi data training untuk meningkatkan performa model. Berikut visualisasi diagram garis dari masing-masing model berdasarkan hasil nilai *f-measure*.



Gambar 4.14 Diagram nilai F-measure Tiap Pengujian

Setelah 4 skenario pengujian dilatih serta menemukan nilai dari masing-masing skenario pengujian, selanjutkan akan dihitung bobot nilai dari masing-masing pengujian dengan berdasarkan pada tabel 4.17 untuk menentukan pengujian mana yang terbaik. Berikut tabel perhitungan bobot dari masing-masing pengujian.

Tabel 4.28 Hasil Pembobotan Pengujian A

Pengujian A	Nilai Performa	Nilai bobot	Hasil Pembobotan
Akurasi	98.03%	0.593	58.13
Presisi	100%	0.225	22.5

Recall	88.23%	0.112	9.881
F-measure	93.75%	0.068	6.375
N		1	96.88

Pada tabel 4.28 merupakan hasil akhir dari proses AHP di mana bobot-bobot yang telah ditentukan diterapkan pada metrik-metrik evaluasi untuk memberikan nilai tertimbang atau terbobot. Dengan menggunakan AHP, Anda dapat memberikan penilaian relatif yang lebih berbobot pada setiap metrik berdasarkan prioritas atau pentingnya dari perspektif yang ditentukan sebelumnya.

Tabel 4.29 Hasil Pembobotan Pengujian B

Pengujian B	Nilai Performa	Nilai bobot	Hasil Pembobotan
Akurasi	95.65%	0.593	56.72
Presisi	97.43%	0.225	21.92
Recall	88.37%	0.112	9.897
F-measure	90.16%	0.068	6.130
N		1	94.66

Pada tabel 4.29 matrix-matrix evaluasi performa (akurasi, presisi, recall, dan F-measure) telah diberi bobot berdasarkan tabel 4.17 dari suatu proses penentuan prioritas atau pentingnya setiap matrix tersebut. Hasil pembobotan digunakan untuk memberikan penilaian terhadap performa Pengujian B dengan mempertimbangkan pentingnya setiap metrik sesuai dengan bobot yang telah ditentukan sebelumnya. Dengan demikian, nilai-nilai pembobotan ini memberikan gambaran tentang kontribusi relatif dari setiap metrik dalam evaluasi performa keseluruhan Pengujian B.

Tabel 4.30 Hasil Pembobotan Pengujian C

Pengujian C	Nilai Performa	Nilai bobot	Hasil Pembobotan
Akurasi	97.64%	0.593	57.90
Presisi	92.45%	0.225	20.80
Recall	90.74%	0.112	10.16
F-measure	91.58%	0.068	6.227
N		1	95.08

Pada tabel 4.30 dari hasil tersebut, dapat diamati bahwa nilai performa pada Pengujian C cenderung lebih tinggi secara keseluruhan dibandingkan Pengujian B pada setiap metrik evaluasi. Akurasi pada Pengujian C menunjukkan peningkatan yang signifikan dibandingkan dengan Pengujian B, sementara metrik lainnya juga menunjukkan peningkatan yang konsisten atau sedikit lebih rendah dalam kasus tertentu.

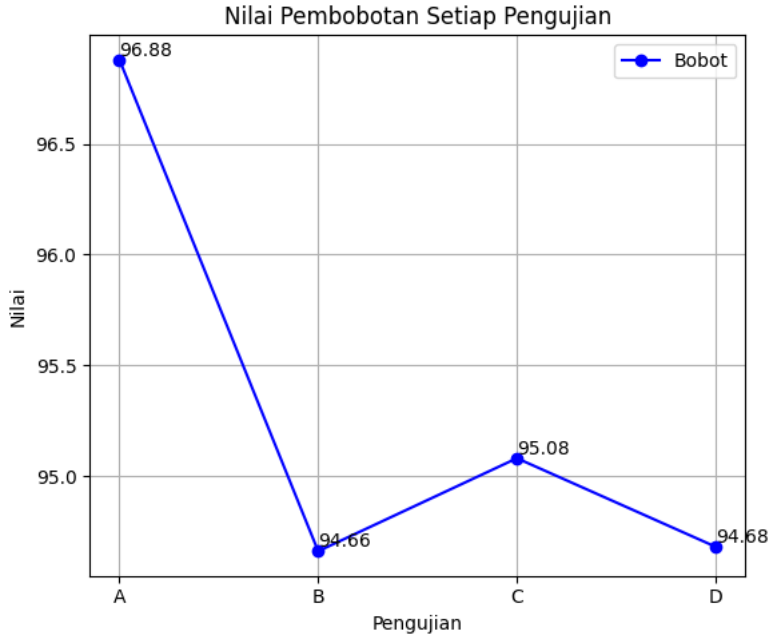
Ini menunjukkan bahwa Pengujian C memiliki performa yang lebih baik secara keseluruhan dibandingkan Pengujian B, meskipun perbedaannya mungkin tidak signifikan dalam beberapa metrik. Dengan demikian, Pengujian C mungkin memiliki model atau sistem yang lebih efektif atau akurat dalam memprediksi data daripada Pengujian B.

Tabel 4.31 Hasil Pembobotan Pengujian D

Pengujian D	Nilai Performa	Nilai bobot	Hasil Pembobotan
Akurasi	97.5%	0.593	57.81
Presisi	93.22%	0.225	20.97
Recall	87.30%	0.112	9.777
F-measure	90.16%	0.068	6.130
N		1	94.68

Pada tabel 4.31 dari hasil tersebut, dapat diamati bahwa Pengujian D memiliki tingkat akurasi yang tinggi (97.5%), yang hampir setara dengan Pengujian C. Namun, terdapat perbedaan dalam nilai presisi, recall, dan F-measure. Presisi Pengujian D lebih tinggi daripada Pengujian C, sedangkan recall Pengujian D lebih rendah.

Hal ini menunjukkan bahwa Pengujian D sedikit lebih baik dari pengujian B akan tetapi masih lebih baik pengujian C. Pengujian D memiliki tingkat kesalahan yang lebih rendah dalam prediksi positif (presisi yang lebih tinggi) namun kemampuannya dalam menemukan kembali instance yang relevan (recall) mungkin sedikit lebih rendah dibandingkan Pengujian C.



Gambar 4.15 Diagram Nilai Pembobotan Tiap Pengujian

Pada Gambar 4.15 ini, diketahui perbedaan nilai pembobotan antara setiap pengujian. Hasil pemebobotan untuk setiap pengujian, A, B, C, dan D, menunjukkan tingkat performa relatif dari masing-masing pengujian terhadap kriteria evaluasi yang diberikan. Dalam urutan menurun, nilai pemebobotan untuk

pengujian. Pengujian A mendapat nilai pemebobotan tertinggi, menunjukkan bahwa dalam konteks evaluasi yang diberikan, pengujian A memiliki performa yang paling tinggi dari seluruh pengujian yang dilakukan. Sementara itu, pengujian B menempati peringkat terendah dalam evaluasi berdasarkan bobot yang diberikan.

Terdapat 5 pengelompokan nilai klasifikasi, yaitu dengan rentang nilai antara 90% - 100% dikategorikan klasifikasi sangat baik, rentang 80% - 90% dikategorikan klasifikasi baik, rentang 70% - 80% dikategorikan klasifikasi cukup, rentang 60 % - 70 % dikategorikan klasifikasi kurang baik, dan rentang 50%-60% dikategorikan klasifikasi gagal (Gorunescu, 2011).

4.3 Integrasi Islam

Kanker payudara adalah jenis kanker yang terbentuk dari sel-sel di dalam payudara. Ini bisa terjadi ketika sel-sel payudara mengalami pertumbuhan yang tidak terkendali. Dalam Islam, menjaga kesehatan adalah bagian dari amanah dan tanggung jawab kepada diri sendiri dan masyarakat. Mencegah penyakit dan mengikuti upaya deteksi dini, termasuk kanker payudara, dianjurkan sebagai bagian dari pemeliharaan kesehatan.

Islam menekankan pentingnya mencari pengobatan dalam menghadapi penyakit. Dalam konteks kanker payudara, pengobatan medis dan upaya penyembuhan sangat dianjurkan dalam Islam. Agama Islam menekankan pentingnya dukungan sosial dan emosional kepada individu yang terkena penyakit. Menunjukkan empati, memberikan dukungan moral, serta bantuan

praktis kepada penderita kanker payudara merupakan nilai-nilai yang diakui dalam Islam.

Islam mengajarkan tentang etika dan moralitas dalam menggunakan informasi dan teknologi. Dalam konteks penelitian kanker payudara, penting untuk memastikan bahwa penggunaan data dan informasi dilakukan dengan integritas dan bertanggung jawab., seperti pada surah An-Nahl ayat 69:

ثُمَّ كُلِى مِن كُلِّ ٱلثَّمَرَٰتِ فَٱسْلُكِى سُبُلَ رَبِّكِ ذُلُلًا ۚ يَخْرُجُ مِنۢ بُطُونِهَا شَرَابٌ مُّخْتَلِفٌ أَلْوَٰنُهُۥ فِيهِ شِفَآءٌ لِّلنَّاسِ ۗ إِنَّ فِى ذَٰلِكَ لَءَايَةً لِّقَوْمٍ يَتَفَكَّرُونَ

"Kemudian makanlah dari tiap-tiap (macam) buah-buahan dan tempuhlah jalan Tuhanmu yang telah dimudahkan (bagimu). Dari perut lebah itu ke luar minuman (madu) yang bermacam-macam warnanya, di dalamnya terdapat obat yang menyembuhkan bagi manusia. Sesusungguhnya pada yang demikian itu benar-benar terdapat tanda (kebesaran Tuhan) bagi orang-orang yang memikirkan". (QS. An-Nahl : 69).

Menurut tafsir Al-Muyassar ayat tersebut mengajarkan untuk memilih makanan yang halal dan kemudian makanlah dari setiap buah apa yang kamu sukai lalu tempuhlah jalan-jalan tuhanmu yang telah ditundukan bagimu untuk mendapatkan rizki di gunung-gunung dan sela-sela antara pepohonan, Sesungguhnya Allah telah menjadikannya mudah bagimu kamu tidak akan salah jalan untuk kembali meskipun berjarak jauh. Akan keluar dari perut-peru lebah itu cairan madu dengan berbagi warna yang berbeda-beda seperti putih, kuning, merah, dan warna lainnya. Didalamnya terdapat sumber kesembuhan bagi manusia dari penyakit-penyakit. Sesunggguhnya dalam hal-hal yang dilakukan oleh lebah benar-benar terkandung bukti kuat yang menunjukan kuasa penciptanya bagi orang yang berpikir dan kemudian mengambil pelajaran.

Dalam konteks tafsir di atas, ayat tersebut merupakan bagian dari Al-Qur'an, Surah An-Nahl (16:69) yang menyebutkan tentang keajaiban lebah dan manfaat madu yang dihasilkannya. Meskipun secara spesifik tidak disebutkan "kanker payudara" dalam ayat tersebut, ayat tersebut menunjukkan pada kesembuhan dan manfaat yang terkandung dalam sumber alamiah, seperti madu yang dihasilkan oleh lebah.

4.3.1 ***Muamalah Ma'a Allah***

Penyakit adalah suatu kondisi yang tak terhindarkan dalam kehidupan manusia. Baik itu penyakit yang ringan maupun penyakit yang serius, setiap orang pasti pernah mengalami kondisi kesehatan yang memburuk. Karena dikatakan manusia yang terkena penyakit merupakan ujian dari Tuhan, yang berarti Tuhan sayang dengan orang yang dia uji. Dalam menghadapi berbagai penyakit ini, ayat ke-10 dalam Surah Al-Baqarah menawarkan kandungan berharga bahwa Allah adalah Penyembuh sejati yang memiliki kekuasaan untuk menyembuhkan penyakit apa pun. Surah Al-Baqarah ayat 10:

فِى قُلُوبِهِم مَّرَضٌ فَزَادَهُمُ ٱللَّهُ مَرَضًا ۖ وَلَهُمْ عَذَابٌ أَلِيمٌۢ بِمَا كَانُوا۟ يَكْذِبُونَ

" Dalam hati mereka ada penyakit, lalu ditambah Allah penyakitnya; dan bagi mereka siksa yang pedih, disebabkan mereka berdusta.." (QS. Al-Baqarah : 10).

Al-Muyasassar menafsirkan ayat diatas yaitu Di dalam hati mereka terdapat keraguan dan kerusakan akibatnya mereka diuji Allah dengan berbuat berbagai macam maksiat yang mewajibkan adanya siksaan bagi mereka, sehingga Allah pun menambah keraguan pada hati mereka dan bagi mereka siksaan yang menyedihkan akibat kedustaan dan kemunafikan mereka

Ayat ini menggarisbawahi bahwa keraguan, kerusakan batin, dan penderitaan dapat menjadi konsekuensi bagi orang-orang yang terjerumus dalam perbuatan maksiat dan perilaku yang tidak sesuai dengan nilai-nilai agama. Oleh karena itu, pesan moralnya mungkin menekankan pentingnya menjauhi perbuatan tersebut dan mengikuti ajaran yang baik dalam kehidupan sehari-hari.

4.3.2 *Muamalah Ma'a an-Nas*

Dalam Islam, menekankan pentingnya berinteraksi dan bersikap adil dalam hubungan sosial antara sesama manusia. Hal ini mencakup sikap baik, empati, menghormati hak orang lain, memberikan bantuan kepada yang membutuhkan, dan menjalani hubungan yang berlandaskan kejujuran dan keadilan. Seperti halnya penting bagi individu yang terkena kanker payudara untuk mendapatkan dukungan moral, emosional, dan sosial dari lingkungan sekitarnya. Dalam islam mengajarkan pentingnya memberikan dukungan kepada mereka yang sedang menghadapi tantangan serius seperti kanker payudara, baik melalui kata-kata semangat, bantuan praktis, atau pendampingan secara emosional.

Seperti yang dijelaskan pada surat Al-'Ashr berikut ini tentang saling mendukung dan kepedulian.

إِلَّا ٱلَّذِينَ ءَامَنُواْ وَعَمِلُواْ ٱلصَّٰلِحَٰتِ وَتَوَاصَوْاْ بِٱلْحَقِّ وَتَوَاصَوْاْ بِٱلصَّبْرِ

" Kecuali orang-orang yang beriman dan mengerjakan amal saleh dan nasehat menasehati supaya mentaati kebenaran dan nasehat menasehati supaya menetapi kesabaran." (Qs. Al-'Ashr : 3).

AL-Mukhtashar menafsirkan ayat diatas yaitu kecuali orang yang beriman kepada Allah dan para Rasul-Nya, mengerjakan amal saleh, saling berwasiat di antara mereka dengan kebenaran dan kesabaran dalam menjalani kebenaran. Orang-orang yang mempunyai sifat-sifat ini pasti selamat dalam kehidupan dunia dan akhirat.

Ayat diatas dalam konteks kanker payudara, ayat ini dapat dipahami sebagai dorongan bagi umat beriman untuk saling mendukung, memberikan nasehat yang membangun, dan menolong satu sama lain di dalam menjalani cobaan hidup, termasuk menjaga kesabaran, keyakinan, serta memberikan dukungan kepada mereka yang sedang menghadapi penyakit tersebut.

Dalam Islam, saling menasihati dan mendukung dalam kebenaran serta kesabaran merupakan bagian dari prinsip kebersamaan dan saling membantu antar sesama. Ini dapat diterapkan dalam konteks kanker payudara dengan memberikan dukungan moral, bantuan praktis, serta berbagi informasi yang berguna kepada mereka yang membutuhkan, sejalan dengan nilai-nilai keimanan dan kebaikan yang diajarkan dalam Al-Qur'an.

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil pengujian yang telah dilakukan pada bab sebelumnya, untuk mengetahui performa metode *Naïve Bayes* pada sistem dalam melakukan klasifikasi kanker payudara dapat ditarik kesimpulan pada penelitian ini yatu dengan data sejumlah 569 data dengan 10 atribut utama yaitu *radius*, *texture*, *perimeter*, *area*, *smoothness*, *compactness*, *concavity*, *concave points*, *symmetry*, *fractal dimension*. Serta dengan 2 jumlah kelas, yakni B (*Benign*) kanker payudara jinak dan M (*Malignant*) kanker payudara ganas. Pengujian yang dilakukan pada penelitian ini menggunakan 4 rasio perbandingan dengan masing-masing pengujian dengan perbandingan 90% data train : 10% data test, 80% data train : 20% data test, 75% data train : 25% data test, 70% data train : 30% data test dan dilakukan perbandingan antara model Naïve Bayes. Dari keempat rasio perbandingan tersebut berdasarkan pada tabel 4.17 maka dari nilai akurasi, presisi, *recall* dan *f-measure* terbaik pada rasio perbandingan 90% data train dan 10% data uji yang dimana menghasilkan nilai akurasi tertinggi sebesar 96.49%, presisi sebesar 100%, *recall* sebesar 88.23%, *f-measure* sebesar 93.75%. Berdasarkan dari hasil nilai akurasi, presisi, *recall* dan *f-measure* dari Model A menunjukkan bahwa model dapat mengklasifikasi data target dengan data prediksi dengan kategori sangat baik, kecuali hasil dari nilai *recall* dengan kategori baik.

5.2 Saran

Berdasarkan hasil pengujian yang diperoleh, penulis memahami bahwa penelitian ini jauh dari kata sempurna masih diperlukan beberapa perbaikan dalam penelitian ini. Oleh karena itu, penulis merekomendasikan beberapa saran untuk penelitian selanjutnya:

1. Dapat mencoba menggunakan library *Naïve Bayes* yang lainnya dari bahasa pemrograman python maupun bahasa pemrograman lainnya.
2. Dapat mencoba menggunakan data dengan perbandingan jumlah label yang seimbang.
3. Dapat mencoba menggunakan teknik seleksi fitur atau teknik lainnya.
4. Dapat mencoba metode Machine Learning lain untuk klasifikasi kanker payudara pada dataset *Breast Cancer Wisconsin (Diagnostic*) atau data kanker payudara yang berbeda.

DAFTAR PUSTAKA

Chazar, C., & Widhiaputra, B.E. (2020). Machnie Learning Diagnosis Kanker Payudara Menggunakan Algoritma Support Vector Machine. *Journal Informatika dan Sistem Informasi*,12(1),67-80. https://doi.org/10.37424/informasi.v12i1.48

Dumitru Diana. (2009). Prediction of Recuurrent Events in Breast Cancer Using The Naïve Bayesian Classification. *Annals of University of Craiova, Math. Comp. Sci Ser*, 36(2), 92-96. ISSN: 1223-6934. DOI: 10.1003/2526-3010

Ginsburg, O., Yip, C., Brooks, A., MSEd., Cabanes, A., MPH., Caleffi, M., Yataco, J.A.D., Gyawali, B., McCormack, V., Anderson, M.M.D., Mehrotra, R., Mohar, A., & Murillo, R. (2020). Breast Cancer Early Detection: A Phased Approach to Implementation. *American Cancer Society JOURNALS*,126(10),2379-2393. https://doi.org/10.1002/cncr.32887

Globocan. (2020). Cancer in Indonesia. *Global Cancer Observatory*. https://doi.org/10.1001/jama.247.22.3087

Kharya, S., Agrawal, S., & Soni, S. (2014). Naïve Bayes Classifiers: A Probabilistic Detection Model for Breast Cancer. *International Journal of Computer Application*, 92(10), 26-31. DOI:10.5120/16045-5206

Mohammed, S. A., Darrab, S., Noaman, S. A., & Saake, G. (2020). Analysis of breast cancer detection using different machine learning techniques. *In Communications in Computer and Information Science: Vol. 1234 CCIS*. Springer Singapore. https://doi.org/10.1007/978-981-15-7205-0_10

Moreira, L. B., & Namen, A. A. (2018). A hybrid data mining model for diagnosis of patients with clinical suspicion of dementia. Computer Methods and Programs in Biomedicine, 165, 139–149. https://doi.org/10.1016/j.cmpb.2018.08.016

Othman, F.M., & Yau, T.M.S. (2007). Comparison of Different Classification Tecniques Using WEKA for Breast Cancer. *IFMBE Proceedings*, 520-523. DOI: 10.1029/849-891

Oktafiani, R., Hermawan, A., & Avianto, D. (2023). Pengaruh Komposisi Split Data Terhadapa performa Klasifikasi Penyakit Kanker Payudara Menggunakan Algoritma Machine Learning. *Jurnal Sains dan Informatika*, 9(1), 19-28. DOI: 10.34128/jsi.v9i1.622

Ramadhan, N.G., & Adhinata, F.D. (2021). Teknik SMOTE dan Gini Score Dalam Klasifikasi Kanker Payudara. *Jurnal Peradaban Sains, Rekayasa dan Teknolog,* 9(2), 125-134. https://doi.org/10.37971/radial.v9i2.229

Resmiati, R., & Arifin, T. (2021). Klasifikasi Pasien Kanker Payudara Menggunakan Metode Support Vector Machine dengan Backward Elimination. *Sistemasi,*10(2),381. https://doi.org/10.32520/stmsi.v10i2.1238

Safutra, R.A., & Prabowo, D.W. (2016). Diagnosis Penyakit Kanker Payudara Menggunakan Metode Naïve Bayes Berbasis Dekstop. *Jurnal Penelitian Dosen FIKOM (UNDA)*, 6(1), 1-6.

Saritas, M. M., & Yasar, Ali. (2019). Performance Analysis of ANN and Naïve Bayes Classification Algorithm for Data Classification. *International Journal of Inteligent System and Apllication in Engineering,*7(2),88-91. https://doi.org/10.18201//ijisae.2019252786

Schmid, P., Cortes, J., Pusztai, L., & McArthur, H. (2020). Pembrolizumab for Early Triple-Negative Breast Cancer. *The New England Journal of Medicine*,382,810-821. DOI: 10.1056/NEJMoa1910549

Sugiyarti, E., Jasmi, K.A., Basiron, B., Huda, M., Shankar, K., & Maseleno, A. (2018). Decision Support System of Scholarship Grantee Selection Using Data Mining. *International Journal of Pure and Applied Mathematics,* 119(15),2239-2249. https://acadpubl.eu/hub/2018-119-15/5/842.pdf

Thaha, R., Widajadnja, N., & Hutasoit, G.A. (2017). Hubungan Tingkat Pengetahuan Tentang Kanker Payudara Dengan Perilaku Pemeriksaan Payudara Sendiri (SADARI) Pada Wanita Usia 20-45 Tahun di Desa Sidera Kecamatan Sigi Biromaru. *Healthy Tadulako Journal*,3(2),40-46. https://doi.org/10.22487/htj.v3i2.50

Vangara, R.V.B., Thirupathur, K., & Vangara, S.P. (2020). Opinion Mining Classification using Naïve Bayes Algorithm. *International Journal of Innovative Technology and Exploring Engineering*,9(5),495-498. DOI: 10.35940/ijitee.E2402.039520

Vembandasamy, K., Sasipriya, R., & Deepa, E. (2015). Heart Diseases Detection Using Naïve Bayes Algorithm. *International Journal of Innovative Science, Engineering & Technology*,2(9),441-444.

Wongkar, M., & Angdresey, A. (2019). Sentiment Analysis Using Naïve Bayes Algorithm of The Data Crawler: Twitter. *Departmen of Informatics Engineering*.DOI:10.1109/ICIC47613.2019.8985884