OPTIMASI *HYPERPARAMETER* MENGGUNAKAN *GRIDSEARCHCV* PADA *K-NEAREST NEIGHBOR CLASSIFIER* UNTUK KLASIFIKASI KANKER PAYUDARA

SKRIPSI

Oleh:
MOHAMMAD ALFI MASYKUR NAZEMI
NIM. 19650006

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

OPTIMASI *HYPERPARAMETER* MENGGUNAKAN *GRIDSEARCHCV* PADA *K-NEAREST NEIGHBOR* CLASSIFIER UNTUK KLASIFIKASI KANKER PAYUDARA

SKRIPSI

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

Oleh:
MOHAMMAD ALFI MASYKUR NAZEMI
NIM. 19650006

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

HALAMAN PERSETUJUAN

OPTIMASI *HYPERPARAMETER* MENGGUNAKAN *GRIDSEARCHCV* PADA *K-NEAREST NEIGHBOR CLASSIFIER* UNTUK KLASIFIKASI KANKER PAYUDARA

SKRIPSI

Oleh:
MOHAMMAD ALFI MASYKUR NAZEMI
NIM 19650006

Telah Diperiksa dan Disetujui untuk Diuji:
Tanggal: 25 Oktober 2023

Pembimbing I,	Pembimbing II,
Prof. Dr. Suhartono, M.Kom	Fajar Rohman Hariri, M.Kom
NIP. 19680519 200312 1 001	NIP. 19890515 201801 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

OPTIMASI *HYPERPARAMETER* MENGGUNAKAN *GRIDSEARCHCV* PADA *K-NEAREST NEIGHBOR CLASSIFIER* UNTUK KLASIFIKASI KANKER PAYUDARA

SKRIPSI

Oleh:
MOHAMMAD ALFI MASYKUR NAZEMI
NIM 19650006

Telah Dipertahankan di Depan Dewan Penguji Skripsi
dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
Untuk Memperoleh Gelar Sarjana Komputer (S.Kom)
Tanggal: 27 November 2023

Susunan Dewan Penguji

Ketua Penguji	: Syahiduz Zaman, M.Kom NIP. 19700502 200501 1 005	( )
Anggota Penguji I	: Okta Qomaruddin Aziz, M.Kom NIP. 1911019 201903 1 013	( )
Anggota Penguji II	: Prof. Dr. Suhartono, M.Kom NIP. 19680519 200312 1 001	( )
Anggota Penguji III	: Fajar Rohman Hariri, M.Kom NIP. 19890515 201801 1 001	( )

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

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

PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan di bawah ini:

Nama	: Mohammad Alfi Masykur Nazemi
NIM	: 19650006
Fakultas / Jurusan	: Sains dan Teknologi / Teknik Informatika
Judul Skripsi	: Optimasi *Hyperparameter* menggunakan *GridSearchCV* pada *K-Nearest Neighbor Classfier* untuk Klasifikasi Kanker Payudara.

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, 25 Oktober 2023

Yang membuat pernyataan,

Mohammad Alfi Masykur Nazemi

NIM. 19650006

HALAMAN MOTTO

”Tidak semua kebiasaan itu baik,

akan tetapi segala bentuk kebaikan harus dibiasakan”

HALAMAN PERSEMBAHAN

Alhamdulillahirabbil'alamin, Dengan rasa syukur yang sebesar-besarnya, saya menyampaikan dedikasi saya atas skripsi ini kepada dua insan yang senantiasa memberikan kasih sayang, bantuan, dan motivasi yang tak tergoyahkan sepanjang proses penyelesaian tugas akhir ini, khususnya kepada ibu dan ayah yang saya hormati. Saya mengucapkan terima kasih atas doa, kesabaran, dan kasih sayang yang telah menjadi sumber dukungan yang teguh sepanjang perjalanan akademis saya.

Tak lupa, ucapan terima kasih yang sebesar-besarnya saya sampaikan kepada seluruh pihak yang telah membantu terselesaikannya skripsi ini, baik dosen pembimbing, teman-teman, dan semua orang yang telah memberikan bantuan dan dukungan selama proses berlangsung. Karya ini saya persembahkan sebagai wujud rasa syukur dan penghargaan, dengan harapan dapat bermanfaat dan memberikan sumbangan kecil bagi kemajuan ilmu pengetahuan. Saya mengucapkan terima kasih atas dukungan dan doa yang terus menerus diberikan.

KATA PENGANTAR

Assalamualaikum Warahmatullahi Wabarakatuh.

Tidak ada kata-kata yang pantas penulis ungkapkan selain rasa syukur dan terima kasih kepada Allah *Subhanahu Wa Ta'ala* atas karunia dan rahmat-Nya sehingga dapat menyelesaikan skripsi ini. Shalawat dan salam kami panjatkan kepada Nabi Agung Nabi Muhammad Rasulullah *Shallallahu 'Alaihi Wasallam* yang telah memimpin kita semua. Skripsi ini dimaksudkan untuk memenuhi salah satu persyaratan sarjana komputer (S.KOM) di Universitas Islam Negeri Maulana Malik Ibrahim Malang program studi Teknik Informatika Fakultas Sains dan Teknologi.

Penulis menemui beberapa masalah dan kendala selama melakukan penelitian dan pembuatan skripsi ini. Namun berkat kegigihan dan kesabaran penulis, skripsi ini dapat diselesaikan dengan baik. Hal ini karena dorongan dan arahan yang diberikan oleh berbagai pihak.

Melalui kesempatan ini, penulis ingin menyampaikan rasa terima kasih dan penghargaan yang setinggi-tingginya kepada kedua orang tua yang selalu memberikan doa, kasih sayang, serta dukungan moral dan material. Kasih sayang penulis kepada Ibu dan Ayahnya tidak pernah bisa diungkapkan dengan kata-kata. Beberapa dukungan lainnya juga penulis ucapkan kepada:

1. Prof. Dr. H. M. Zainuddin, M.A., selaku rektor Universitas Islam Negeri Maulana Malik Ibrahim Malang.
2. Prof. Dr. Sri Hariani, M.Si., selaku dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang.

3. Dr. Fachrul Kurniawan, M.MT., selaku ketua program studi Teknik Informatika Universitas Islam Negeri Maulana Malik Ibrahim Malang.
4. Prof. Dr. Suhartono, M.Kom., selaku dosen pembimbing I dan Fajar Rahman Hariri, M.Kom., selaku dosen pembimbing II yang telah membimbing penulis untuk mengembangkan pemikiran dalam penyusunan skripsi ini hingga selesai.
5. Syahiduz Zaman, M.Kom., selaku dosen penguji I dan Okta Qomaruddin Aziz, M.Kom selaku dosen penguji II yang telah menguji, menasehati, serta memberikan saran untuk menjadikan penyusunan skripsi ini lebih baik lagi.
6. Seluruh dosen dan segenap staff program studi Teknik Informatika Universitas Islam Negeri Maulana Malik Ibrahim Malang yang telah memberikan segala ilmu yang peneliti mulai dari nol hingga menjadi tahu sehingga dapat memudahkan peneliti dalam penyelesaian skripsi ini.
7. Pemilik Nomor Induk Mahasiswa 19310005 yang telah memberikan segala dukungannya dalam berbagai bentuk selama penyusunan skripsi ini.
8. Teman-teman Musyrif/ah Pusat Mahad Al-Jamiah pada umumnya dan khususnya Musyrif Mabna Al-Faraby 2022-2023, terima kasih atas bantuan dan motivasi kalian semua.
9. Teman-teman Jurusan Teknik Informatika Angkatan 2019 “ALIEN” yang telah menghabiskan waktu, memotivasi, dan berjuang bersama penulis semasa kuliah.
10. Denis Erlangga, Yusabbih Barqu, Rifqi Mufiddin, Syahrul Mubin, Edy Hyto, Ahmad Ramadhani, Yusral Ruslin, Arib Akram, Gus Addien, serta teman-teman dekat penulis yang senantiasa memberikan dukungannya.

11. Semua pihak yang tidak dapat penulis sebutkan satu persatu yang telah membantu penulis dalam penyusunan skripsi ini.

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

Wassalamualaikum Warahmatullahi Wabarakatuh.

Malang, 25 Oktober 2023

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

ABSTRAK

Nazemi, Mohammad Alfi Masykur. 2023. **Optimasi *Hyperparameter* menggunakan *GridSearchCV* pada *K-Nearest Neighbor Classifier* untuk Klasifikasi Kanker Payudara.** Skripsi. Jurusan Teknik Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (I) Prof. Dr. Suhartono, M.Kom (II) Fajar Rohman Hariri, M.Kom.

Kata kunci: *GridSearchCV, Hyperparameter, K-Nearest Neighbor Classifier, Klasifikasi Kanker Payudara.*

Kanker payudara merupakan salah satu jenis kanker yang menyumbangkan angka kematian terbesar di dunia. Setidaknya mengacu pada data *Global Cancer Observatory* pada tahun 2020 hampir 10 juta kasus kematian yang diakibatkan oleh kanker payudara di dunia. Di Indonesia penyakit ini menjadi penyumbang kematian pertama dengan kasus kematian pada tahun 2020 mencapai lebih dari 22 ribu kasus kematian. Melihat jumlah kematian yang disebabkan oleh kanker payudara, dianggap sebagai ancaman besar bagi dunia medis. Salah satu cara untuk menekan angka kematian tersebut adalah dengan cara deteksi dini sel kanker, karena sel kanker dapat dideteksi lebih awal. Adanya deteksi sejak dini merupakan kunci utama untuk meningkatkan kemungkinan kelangsungan hidup, mengurangi dampak penyakit, dan meningkatkan standar hidup bagi mereka yang menderita kanker payudara. Penelitian ini memanfaatkan algoritma *machine learning* untuk memprediksi sel kanker dengan cepat dan efisien. Tujuan penelitian ini untuk mengetahui performa model *K-Nearest Neighbor* dalam mengklasifikasi penyakit kanker payudara pada *dataset Wisconsin Diagnostic Breast Cancer* (WDBC). Sebelumnya data melalui proses *preprocessing* data yang meliputi seleksi fitur, proses *encoder*, *standrization* data serta *split* data. Pada penelitian ini *split* data terbagi menjadi 3 model, yaitu model A dengan perbandingan 80% data *training*: 20% data *testing*, model B dengan perbandingan 70% data *training*: 30% data *testing*, model C dengan perbandingan 60% data *training*: 40% data *testing*. Selanjutnya, pada masing-masing pembagian model tersebut akan dilakukan perbandingan antara model *K-Nearest Neighbor* tanpa adanya penyetelan *hyperparameter* atau secara default dengan model *K-Nearest Neighbor* dengan penyetelan *hyperparameter* menggunakan *GridSearchCV* untuk mencari nilai parameter yang optimal serta dievaluasi menggunakan *confusion matrix*. Pada penelitian ini didapatkan nilai akurasi terbaik pada model B dengan nilai akurasi tanpa adanya penyetelan *hyperparameter* sebesar 94.7% sedangkan dengan adanya penyetelan *hyperparameter* nilai akurasi sebesar 97.6% dengan menggunakan *Euclidean distance* sebagai metode perhitungan jarak.

ABSTRACT

Nazemi, Mohammad Alfi Masykur. 2023. **Hyperparameter Optimization using GridSearchCV on K-Nearest Neighbor Classifier for Breast Cancer Classification.** Undergraduate Thesis. Department of Informatics Engineering Faculty of Science and Technology Maulana Malik Ibrahim State Islamic University Malang. Supervisor: (I) Prof. Dr. Suhartono, M.Kom (II) Fajar Rohman Hariri, M.Kom.

Breast cancer is one type of cancer that contributes to the largest mortality rate in the world. At least referring to Global Cancer Observatory data in 2020 almost 10 million cases of death caused by breast cancer in the world. In Indonesia, this disease became the first contributor to death with death cases in 2020, reaching more than 22 thousand cases of death. Looking at the number of deaths caused by breast cancer, it is considered a major threat to the medical world. One way to reduce the mortality rate is by early detection of cancer cells because cancer cells can be detected early. Early detection is key to increasing the chances of survival, reducing the impact of disease, and improving the standard of living for those with breast cancer. The research leverages machine learning algorithms to predict cancer cells quickly and efficiently. The purpose of this study was to determine the performance of the K-Nearest Neighbor model in classifying breast cancer in the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. Previously, data went through a data preprocessing process which included feature selection, encoder process, data standrization and data split. In this study, split data is divided into 3 models, namely model A with a ratio of 80% training data: 20% testing data, model B with a ratio of 70% training data: 30% testing data, model C with a ratio of 60% training data: 40% testing data. Furthermore, in each of these model divisions, a comparison will be made between the K-Nearest Neighbor model without hyperparameter adjustment or by default with the K-Nearest Neighbor model with hyperparameter adjustment using GridSearchCV to find optimal parameter values and evaluated using a confusion matrix. In this study, the best accuracy value was obtained in model B with an accuracy value without hyperparameter adjustment of 94.7%, while with hyperparameter adjustment, an accuracy value of 97.6% was obtained using Euclidean distance as a distance calculation method.

Keywords: *Breast Cancer Classification, GridSearchCV, Hyperparameter, K-Nearest Neighbor Classifier.*

الملخص

نزمي، محمد ألفي مشكور. ٢٠٢٣. **تحسين المعلمة الفائقة باستخدام GridSearchCV على مصنف K-Nearest Neighbor لتصنيف سرطان الثدي**. أُطرُوحَة. قسم الهندسة المعلوماتية. كلية العلوم والتكنولوجيا ، جامعة مولانا مالك إبراهيم الإسلامية الحكومية مالانج. مشرفا : (١) الأستاذ الدكتورسوهرطانو، الماجستير(٢) فجر رحمن حربرى، الماجستير.

الكلمات المفتاحية: المعلمة الفائقة, GridSearchCV , K-Nearest Neighbor , تصنيف سرطان الثدي.

كانسر الثدي هو نوع من السرطان الذي يساهم في أكبر معدل وفيات في العالم. على الأقل بالإشارة إلى بيانات المرصد العالمي للسرطان، في عام ٢٠٢٠ كان هناك ما يقرب من ١٠ ملايين حالة وفاة بسبب سرطان الثدي في العالم. وفي إندونيسيا، يعد هذا المرض المساهم الأول في الوفيات، حيث وصلت الوفيات عام ٢٠٢٠ إلى أكثر من ٢٢ ألف حالة وفاة. ونظرًا لعدد الوفيات الناجمة عن سرطان الثدي، فهو يعتبر تهديدًا كبيرًا لعالم الطب. إحدى الطرق لتقليل معدل الوفيات هي الكشف المبكر عن الخلايا السرطانية، لأنه يمكن اكتشاف الخلايا السرطانية في وقت مبكر. فالكشف المبكر هو المفتاح الأساسي لزيادة فرص البقاء على قيد الحياة، وتقليل تأثير المرض، وتحسين المستوى المعيشي لمن يعانين من سرطان الثدي. تستفيد هذه الدراسة من خوارزمية تعلم الآلة لتوقع الخلايا السرطانية بسرعة وكفاءة. هدف هذه الدراسة هو تقييم أداء نموذج الجار الأقرب *K-Nearest Neighbor* في تصنيف مرض سرطان الثدي باستخدام مجموعة البيانات *Wisconsin Diagnostic Breast Cancer* (WDBC). تم معالجة البيانات مسبقاً من خلال عمليات تجهيز البيانات، بما في ذلك اختيار الميزات، وعمليات الترميز، وتوحيد البيانات، وتقسيم البيانات. في هذه الدراسة تم تقسيم البيانات المقسمة إلى ٣ نماذج وهي النموذج (أ) بنسبة ٨٠٪ بيانات تدريب: ٢٠٪ بيانات اختبار، النموذج (ب) بنسبة ٧٠٪ بيانات تدريب: ٣٠٪ بيانات اختبار، النموذج (ج) بنسبة ٦٠٪ من بيانات التدريب: ٤٠٪ من بيانات الاختبار. بعد ذلك، لكل قسم نموذج، سيتم إجراء مقارنة بين نموذج *K-Nearest Neighbor* بدون إعدادات المعلمة الفائقة أو بشكل افتراضي مع نموذج *K-Nearest Neighbor* مع إعدادات المعلمة الفائقة باستخدام *GridSearchCV* للعثور على قيم المعلمة المثالية وتقييمها باستخدام الارتباك مصفوفة. في هذا البحث تم الحصول على أفضل قيمة دقة في النموذج (ب) بقيمة دقة بدون تعديلات المعلمة الفائقة بنسبة ٩٤.٧٪، بينما مع تعديلات المعلمة الفائقة كانت قيمة الدقة ٩٧.٦٪ باستخدام المسافة الإقليدية كطريقة لحساب المسافة.

BAB I

PENDAHULUAN

1.1. Latar Belakang

Kanker payudara adalah kategori penyakit di mana sel-sel jaringan payudara berubah dan membelah tak terkendali yang mengakibatkan munculnya suatu benjolan. Sebagian besar kanker payudara dimulai di lobulus (kelenjar susu) atau di saluran yang menghubungkan lobulus ke puting (Houssein et al., 2021). Kanker payudara merupakan tumor ganas yang berkembang dari sel-sel di payudara. Duktus termasuk anatomi payudara yang pada umumnya kanker berasal darinya dan beberapa diantaranya berasal dari kelenjar payudara atau yang biasa dikenal dengan lobulus.

Kanker payudara dapat menyerang pria dan wanita, namun diketahui bahwa wanita adalah korban utama dari penyakit ini. Kanker payudara adalah jenis kanker yang paling umum di seluruh dunia. Berdasarkan data *Global Cancer Observatory* yang dihasilkan oleh *International Agency for Research on Cancer* pada tahun 2020 setidaknya ada 19.3 juta kasus kanker baru dan hampir 10 juta kematian akibat kanker di dunia. Menurut data kanker payudara menjadi kasus kematian terbesar setelah kanker paru-paru dengan total 2.3 juta kasus (Sung et al., 2021). Dilansir dari Kementerian Kesehatan RI bahwasannya kanker payudara menempati urutan pertama jumlah kanker terbesar di Indonesia serta menjadi penyumbang kematian pertama akibat kanker. Bersumber dari data *Global Cancer Observatory* pada tahun 2020 di Indonesia ada 65.858 kasus baru kanker payudara dari total 396.914 kasus. Sementara, untuk kasus kematian mencapai lebih dari 22 ribu jiwa (*Kanker*

Payudara Paling Banyak Di Indonesia, Kemenkes Targetkan Pemerataan Layanan Kesehatan, 2022).

Ada beberapa faktor penyebab terjadinya kanker payudara. Dalam buku karangan Ashariati (2019) disebutkan bahwa ada 4 faktor penyebab terjadinya kanker payudara, antara lain faktor umur, faktor hormonal, faktor keturunan dan faktor gaya hidup (Ashariati, 2019). Ashariati juga menjelaskan bahwa risiko wanita terkena kanker payudara yang disebabkan oleh keturunan dalam hal ini ibu atau saudarinya menderita kanker payudara maka risiko kejadian meningkat sebesar 3 kali. Dalam penelitian lain yang ditulis oleh Azmi et al (2020) juga menjelaskan bahwasannya risiko terjadinya kanker payudara yang disebabkan oleh riwayat keluarga kanker payudara meningkat sebesar 10 kali daripada wanita yang tidak mempunyai riwayat keluarga kanker payudara (Azmi et al., 2020).

Kanker payudara adalah salah satu penyakit paling serius yang dialami wanita, dan dianggap sebagai salah satu penyakit paling agresif dalam sejarah medis. Terbukti dengan peningkatan angka kematian yang disebabkan oleh kanker payudara. Melihat jumlah kematian yang disebabkan oleh kanker payudara, dianggap sebagai ancaman besar bagi dunia medis. Salah satu cara untuk menekan angka kematian tersebut adalah dengan cara deteksi dini sel kanker, karena sel kanker dapat dideteksi lebih awal, sehingga angka harapan hidup lebih tinggi (Kartini et al., 2019).

Seiring dengan perkembangan teknologi yang semakin maju, pemanfaatan teknologi sudah banyak digunakan untuk keperluan medis. Salah satunya adalah penggunaan algoritma *machine learning* yang bisa dimanfaatkan sebagai

memprediksi sel kanker dengan cepat dan efisien. Sebagaimana Al-Qur'an menerangkan bahwa Allah SWT memberikan kemudahan bagi manusia untuk mencapai kebahagiaan di dunia dan di akhirat yang tertuang dalam Al-Qur'an surah Al-A'la ayat 8 yang berbunyi:

وَنُيَسِّرُكَ لِلْيُسْرٰى

"Dan Kami akan memudahkan bagimu ke jalan kemudahan (mencapai kebahagiaan dunia dan akhirat)" (QS. Al-A'la: 8).

Ayat tersebut juga mengacu pada teknologi, yang mana prinsip dasar dan tujuan teknologi adalah untuk memudahkan manusia dalam menjalani kehidupan sehari-hari salah satunya adalah pemanfaatan algoritma *machine learning* dalam memprediksi sel kanker dengan cepat dan efisien. Salah satu metode algoritma *machine learning* adalah metode *K-Nearest Neighbor*. Metode K-NN merupakan salah satu algoritma yang paling popular. Metode K-NN mengklasifikasikan item berdasarkan data pembelajaran yang paling dekat dengan objek (Arifin et al., 2019). Metode K-NN termasuk pendekatan klasifikasi data yang sederhana dan mudah diterapkan, efektif pada dataset yang lebih besar, dan dapat mengklasifikasikan data secara akurat. Penggunaan metode K-NN dalam deteksi medis sangat menarik. Kualitas hasil sangat bergantung pada jarak dan nilai parameter "k" yang mewakili jumlah tetangga terdekat. Metode K-NN memiliki keunggulan menghasilkan data yang kuat atau jelas dan efektif bila digunakan pada data dalam jumlah besar. Dalam penelitian Hashi & Md. Shahid Uz Zaman (2020) menjelaskan bahwasannya penggunaan metode K-NN dengan penyetelan *hyperparameter* memiliki tingkat akurasi yang sangat tinggi sekitar 91.80% dibandingkan dengan metode lainnya.

Penyetelan *hyperparameter* memiliki pengaruh pada hasil akurasi, apabila tanpa penyetelan *hyperparameter* maka akurasi yang didapat hanya 90.16% (Hashi & Md. Shahid Uz Zaman, 2020).

Algoritma *machine learning* diharapkan dapat mempermudah dalam melakukan klasifikasi sel kanker dengan cepat dan efisien, sehingga deteksi kanker sejak dini dapat dilakukan dan dapat menekan angka kematian yang disebabkan kanker payudara. Berdasarkan uraian di atas, maka penelitian ini menggunakan metode *K-Nearest Neighbor* dengan optimasi *hyperparameter* menggunakan *GridSearchCV* dalam mengklasifikasi kanker payudara pada dataset *Wisconsin Diagnostic Breast Cancer*.

1.2. Pernyataan Masalah

Berdasarkan penjelasan latar belakang di atas, rumusan masalah dari penelitian ini adalah bagaimana menerapkan model *K-Nearest Neighbor* dalam mengklasifikasi kanker payudara serta seberapa pengaruh *hyperparameter* menggunakan *GridSearchCV* pada model *K-Nearest Neighbor* dalam mengklasifikasi kanker payudara?

1.3. Tujuan Penelitian

Berdasarkan pernyataan masalah di atas, maka tujuan penelitian ini adalah bertujuan untuk menerapkan model *K-Nearest Neighbor* dalam mengklasifikasi kanker payudara serta mengetahui pengaruh *hyperparameter* menggunakan *GridSearchCV* pada model *K-Nearest Neighbor* dalam mengklasifikasi kanker payudara.

1.4. Batasan Penelitian

Batasan penelitian bertujuan untuk pembahasan dapat di batasi agar tidak keluar dari topik. Adapun batasan penelitian ini adalah:

1. Data yang digunakan merupakan *dataset Wisconsin Diagnostic Breast Cancer* (WDBC) dari *University of Wisconsin Hospital* yang diperoleh dari *UCI Machine Learning Repository*.
2. *Range* nilai parameter (k) yang digunakan dalam pemodelan adalah 1 sampai 10.
3. Perhitungan jarak menggunakan *Euclidean Distance* dan *Manhattan Distance.*
4. Optimasi yang dimaksud sebatas pengaruh *hyperparameter* pada klasifikasi kanker payudara.

1.5. Manfaat Penelitian

Manfaat penelitian yang diharapkan oleh peneliti adalah:

1. Mengetahui seberapa akurat penggunaan model *K-Nearest Neighbor* dalam mengklasifikasi penyakit kanker payudara.
2. Menaruh peran pengetahuan dan wawasan tentang implementasi model *K-Nearest Neighbor* dalam mengklasifikasi awal penyakit kanker payudara.
3. Diharapkan dapat menjadi titik tolak untuk penelitian selanjutnya.

BAB II

STUDI PUSTAKA

2.1. Penelitian Terkait

Dalam penelitian Sharma et al (2018) yang berjudul "*Breast Cancer Detection Using Machine Learning Algorithms*" menyajikan perbandingan antar teknik dan algoritma *machine learning* dalam memprediksi kanker payudara. Penelitian ini membandingkan algoritma K-NN, Random Forest dan Naïve Bayes. Dataset yang dipakai pada penelitian ini adalah *The Wisconsin Diagnosis Breast Cancer data set*. Hasil dari penelitian ini membuktikan bahwasannya algoritma K-NN dan Random Forest dapat menangani masalah klasifikasi dan regresi sedangkan Naïve Bayes hanya bisa menangani klasifikasi saja. Dari perbandingan pada panelitian ini menghasilkan nilai akurasi pada alogritma K-NN 95.90%, algoritma Random Forest 94.74% dan algoritma Naïve Bayes 94.47%. Dapat diketahui bahwasannya algoritma K-NN paling efektif dalam mendeteksi kanker payudara karena memiliki akurasi yang paling baik dibandingkan dengan algoritma lainnya (Sharma et al., 2018).

Penelitian yang berjudul "*Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer*" yang ditulis oleh Henderi et al (2021) mengkaji tentang penggunaan algoritma *K-nearest neighbor* dengan menggunakan normalisasi *Min-Max normalization* dan *Z-Score Normalization* dalam klasifikasi pada dataset *Wisconsin Breast Cancer Diagnostic.* Peneliti membangun model dengan 469 *training data* dan 100 *testing data*. Hasil pengujian *confusion matrix*

menunjukkan bahwasannya nilai akurasi tertinggi didapatkan dengan menggunakan metode normalisasi *Min-Max normalization* dengan nilai k = 5 dan k = 21 dengan tingkat akurasi tertinggi 98%. Sedangkan metode normalisasi *Z-Score normalization* mendapatkan nilai akurasi tertinggi 97% dengan nilai k = 5 dan k = 15 (Henderi et al., 2021).

Dalam penelitian yang dilakukan oleh Assegie (2020) berjudul "*an Optimized K-Nearest Neighbor based Breast Cancer Detection*" menjelaskan tentang penyetelan *hyper-parameter* pada algoritma K-Nearest Neighbors. Peneliti menggunakan data *Breast Cancer Wisconsin Diagnostic Dataset* dengan pengujian model menggunakan 70% *training data* dan 30% *testing data*. Penyetelan *hyper-parameter* ini dengan mencari nilai k terbaik dengan menggunakan *grid search*. Dari hasil miss-classification menunjukkan bahwa nilai k dengan nilai error terendah 0.065 adalah k = 9. Hasil dari penelitian ini menunjukkan bahwa performa algoritma K-Nearest Neighbors *default* adalah 90.10% sedangkan K-Nearest Neighbors dengan penyetelan hyper-parameter menghasilkan performa 94.35% (Assegie, 2020).

Dalam penelitian Angkasa & Junifer Pangaribuan (2022) yang berjudul "Komparasi Tingkat Akurasi Random Forest dan KNN untuk Mendiagnosis Penyakit Kanker Payudara" menganalisis tingkat akurasi performa dari algoritma Random Forest dan KNN untuk mendiagnosis kanker payudara. Dataset yang digunakan pada penelitian ini didapatkan dari *UCI Machine Learning Repository*. Peneliti dalam penelitian ini menggunakan algoritma Random Forest dan KNN serta melakukan analisis dengan ROC *Curve.* Peneliti membangun model dengan

training data 75% dan *testing data* 25%. Dari hasil pengujian ROC *curve* menunjukkan bahwasannya algoritma KNN lebih unggul dibandingkan dengan algoritma Random Forest dengan 0.9959 untuk KNN dan 0.9951 untuk Random Forest (Angkasa & Junifer Pangaribuan, 2022).

Dalam penelitian Jabbar et al (2022) yang berjudul "Komparasi Algoritma Decision Tress, Naïve Bayes, dan K-Nearest Neighbors dalam Klasifikasi Kanker Payudara" menganalisis beberapa metode *machine learning* yang digunakan untuk mengklasifikasi kanker payudara. Dataset yang digunakan pada penelitian ini didapatkan dari *University of California, Irvine Machine Learning Repository*. Penelitian ini menggunakan 3 algoritma yakni Decision Tress, Naïve Bayes, dan K-Nearest Neighbors dengan menggunakan 2 metode *cross-validation* diantaranya *Hold-Out* dan *K*-Fold. Hasil dari pengujian menunjukkan bahwasannya penggunaan metode K-Nearest Neighbors memiliki hasil performa akurasi yang sangat baik dibanding dengan algoritma Decision Tress dan Naïve Bayes (Jabbar et al., 2022).

Tabel 2. 1 Penelitian terkait

No	Objek Penelitian	Sitasi	Metode Penelitian	Hasil Penelitian
1.	*Wisconsin Diagnosis Breast Cancer dataset*	(Sharma et al., 2018)	Metode *K-Nearest Neighbour, Random Forest* dan *Naïve Bayes*	Hasil dari penelitian ini membuktikan bahwasannya algoritma K-NN dan Random Forest dapat menangani masalah klasifikasi dan regresi sedangkan Naïve Bayes hanya bisa menangani klasifikasi saja. Dan algoritma K-NN paling efektif dalam mendeteksi kanker payudara karena memiliki akurasi yang paling baik dibandingkan dengan algoritma lainnya dengan hasil nilai akurasi sebesar 95.90%.

No	Objek Penelitian	Sitasi	Metode Penelitian	Hasil Penelitian
2.	*Wisconsin Breast Cancer Diagnostic*	(Henderi et al., 2021)	K-Nearest Neighbors	Hasil pengujian *confusion matrix* menunjukkan bahwasannya nilai akurasi tertinggi didapatkan dengan menggunakan metode normalisasi *Min-Max normalization* dengan nilai k = 5 dan k = 21 dengan tingkat akurasi tertinggi 98%. Sedangkan metode normalisasi *Z-Score normalization* mendapatkan nilai akurasi tertinggi 97% dengan nilai k = 5 dan k = 15.
3.	*Breast Cancer Wisconsin Diagnostic Dataset*	(Assegie, 2020)	K-Nearest Neighbors	Hasil penelitian ini menunjukkan penggunaan pendekatan *grid search* dalam mencari nilai k yang terbaik pada algoritma K-Nearest Neighbors. Penelitian ini menggunakan 70% data latih dan 30% data testing untuk pengujian model. Dari hasil miss-classification menunjukkan bahwa nilai k dengan nilai error terendah 0.065 adalah k = 9. Hasil dari penelitian ini menunjukkan bahwa performa algoritma K-Nearest Neighbors *default* adalah 90.10% sedangkan K-Nearest Neighbors dengan penyetelan hyper-parameter menghasilkan performa 94.35%.
4.	*Breast Cancer Wisconsin (Diagnostic) Data Set*	(Angkasa & Junifer Pangaribuan, 2022)	Random Forest dan KNN	Hasil penelitian ini menjelaskan bahwasannya algoritma KNN lebih unggul dibandingkan dengan algoritma Random Forest. Dengan hasil ROC *curve* sebesar 0.9959 untuk KNN dan 0.9951 untuk Random Forest.
5.	*Breast Cancer Wisconsin (Diagnostic)*	(Jabbar et al., 2022)	Decision Tress, Naïve Bayes, dan K-Nearest Neighbors	Hasil penelitian ini menunjukkan bahwa metode K-Nearest Neighbors memiliki hasil performa akurasi yang sangat baik dibanding dengan algoritma yang lain. Pada metode *Hold-Out* K-Nearest Neighbors 98%, Decision Tress 94% dan Naïve Bayes 95%. Pada metode *K-Fold* K-Nearest Neighbors 96%, Decision Tress 93% dan Naïve Bayes 95%.

2.2. Kanker Payudara

Kanker payudara merupakan keganasan yang dimulai pada sel payudara dan menyebar ke jaringan payudara (Parannuan, 2016). Kanker dapat mulai tumbuh di kelenjar susu payudara, saluran susu, jaringan lemak, dan jaringan ikat. Kanker payudara termasuk salah satu kanker paling sering dan penyebab utama kematian akibat kanker pada wanita. Akan tetapi tidak menutup kemungkinan terjadi pada pria juga.

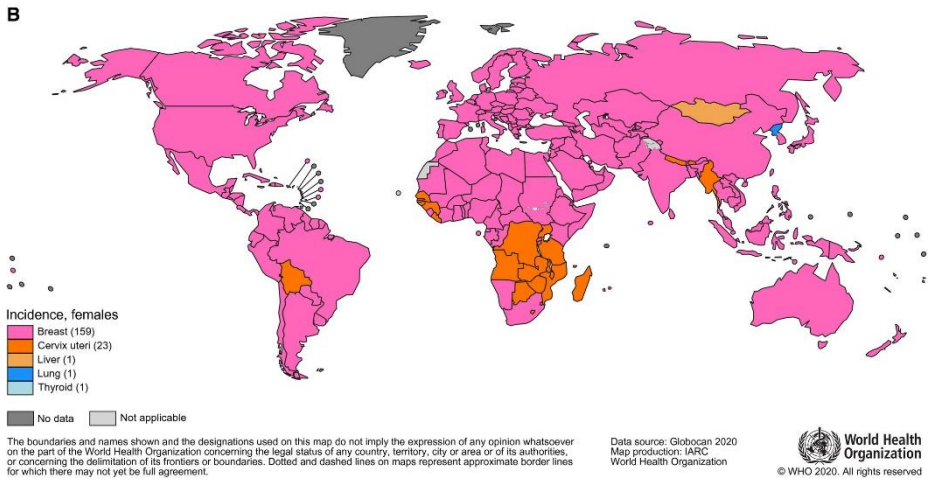
Cancer Site	NO. OF NEW CASES	NO. OF NEW DEATHS
Female Breast	2,261,419	684,996
Lung	2,206,771	1,796,144
Prostate	1,414,259	375,304
Nonmelanoma of skin	1,198,073	63,731
Colon	1,148,515	576,858
Stomach	1,089,103	768,793
Liver	905,677	830,180
Rectum	732,210	339,022
Cervix uteri	604,127	341,831
Esophagus	604,127	544,076
Thyroid	586,202	43,646
Bladder	573,278	212,536
Non-Hodgkin lymphoma	544,352	259,793
Pancreas	495,773	466,003
Leukimia	474,519	311,594
Kidney	431,288	179,368
Corpus uteri	417,367	97,370
Lip, oral cavity	377,713	177,757
Melanoma of skin	324,635	57,043
Ovary	313,959	207,252
Brain, nervous system	308,102	251,329
Larynx	184,615	99,840
Multiple myeloma	176,404	117,077
Nasopharynx	133,354	80,008
Gallbladder	115,949	84,695
Oropharynx	98,412	48,143
Hypopharynx	84,254	38,599
Hodgkin lymphoma	83,087	23,376
Testis	74,458	9334
Salivary glands	53,583	22,778
Anus	50,865	19,293
Vulva	45,240	17,427
Penis	36,068	13,211
Kaposi sarcoma	34,270	15,086
Mesothelioma	30,870	26,278
Vagina	17,908	7995
All Sites	**19,292,789**	**9,958,133**

Gambar 2. 1 Kasus Baru dan Kematian untuk 36 Kanker dan Semua Kanker Gabungan pada tahun 2020

Berdasarkan data *Global Cancer Observatory* yang dihasilkan oleh *International Agency for Research on Cancer* jumlah kasus kanker baru yang didiagnosis pada tahun 2020 adalah 19,3 juta, dan hampir 10,0 juta meninggal

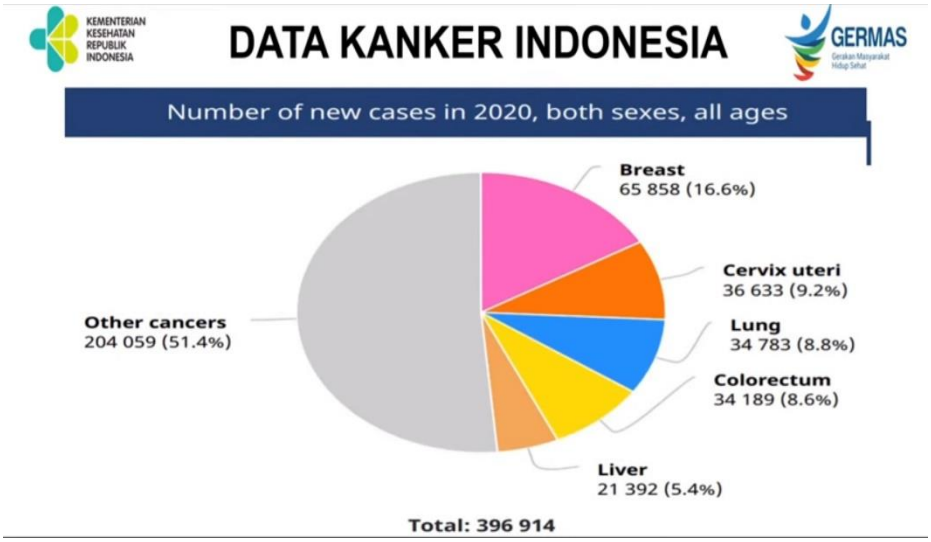
akibat kanker (Gambar 2.1) (Deo et al., 2022). Berdasarkan data tersebut kanker payudara wanita menjadi kasus kanker yang paling umum di dunia dengan presentasi kasus 11.7% disusul dengan kanker paru-paru dengan presentasi kasus 11.4%, kolokletar 10.0%, prostat 7.3%, dan kanker lambung 5.6%.

Pada tahun 2020, kanker payudara menjadi penyebab utama kelima kematian akibat kanker di seluruh dunia, dengan 685.000 kematian (Sung et al., 2021). Di antara wanita, kanker payudara menyumbang 1 dari 4 kasus kanker dan 1 dari 6 kematian akibat kanker, menduduki peringkat pertama untuk kejadian di sebagian besar negara (159 dari 185 negara) (Gambar 2.2).



Gambar 2. 2 Persebaran kanker di dunia

Dilansir dari Kementerian Kesehatan RI bahwasannya kanker payudara menempati urutan pertama jumlah kanker terbesar di Indonesia serta menjadi penyumbang kematian pertama akibat kanker. Bersumber dari data *Global Cancer Observatory* pada tahun 2020 di Indonesia ada 65.858 kasus baru kanker payudara dari total 396.914 kasus (Gambar 2.3). Sementara, untuk kasus kematian mencapai lebih dari 22 ribu jiwa.



Gambar 2. 3 Data kanker di Indonesia

2.3. Dataset

Bahan utama dari penelitian ini adalah data. Dataset merupakan objek dalam memori yang mewakili data dan hubungannya (Suhartini et al., 2020). Dengan kata lain dataset merupakan kumpulan data sistematis yang umumnya disajikan dalam bentuk tabel, memiliki baris, dan kolom. Setiap baris dan kolom biasanya mewakili variabel yang berbeda. Contohnya, suatu kolom mewakili jumlah skor mata kuliah, sedangkan barisnya mewakili nilai huruf dari mata kuliah. Fungsi dataset adalah untuk memperhatikan keterkaitan antar variabel. Apalagi jika jumlah data dan variabel yang dinilai beragam.

Dataset yang dipakai pada penelitian ini adalah dataset *Wisconsin Diagnostic Breast Cancer* (WDBC) dari *University of Wisconsin Hospital* yang diperoleh dari repositori *UCI Machine Learning Repository*. Dataset memiliki 569 pola (357 untuk jinak, 212 untuk ganas) dengan tiga kelas (nomor ID, jinak, ganas) dan 32 kolom untuk fitur.

2.4. *K-Nearest Neighbor*

Algoritma *K-Nearest Neighbor* (K-NN) adalah metode untuk mengidentifikasi item berdasarkan data pembelajaran yang paling dekat dengan objek (Kafil, 2019). Algoritma K-NN merupakan salah satu algoritma dalam kasus klasifikasi yang paling dasar. Algoritma K-NN digambarkan dengan namanya, dimana K menandakan sejumlah nilai *Nearest Neighbours* yang digunakan untuk menentukan kesamaan suatu titik baru dengan tetangganya. Algoritma K-NN bersifat *instance based* dan termasuk jenis algoritma *supervised learning*. Algoritma K-NN biasanya digunakan untuk permasalahan kasus regresi dan klasifikasi.

Langkah-langkah dalam menerapkan algoritma K-NN menurut (Daqiqil, 2021) dalam bukunya yang berjudul "MACHINE LEARNING Teori, Studi Kasus dan Implementasi Menggunakan Python" adalah sebagai berikut:

1. Menentukan nilai K (jumlah tetangga yang dekat).
2. Menghitung jarak antara titik baru dengan semua data training.
3. Memilih K titik terdekat dengan titik yang baru.
4. Jumlah titik terdekat dihitung berdasarkan kategori/kelas dalam hal klasifikasi. Titik baru akan ditambahkan ke kategori dengan jumlah titik terdekat terbanyak. Dalam hal regresi, titik baru akan menjadi rata-rata dari K tetangga terdekat.

Perhitungan jarak sangat penting untuk keberhasilan algoritma ini. Jarak menunjukkan tingkat kesamaan antara data uji dan data latih. Semakin besar jaraknya, semakin besar perbedaannya, dan sebaliknya. *Euclidean Distance* dan

Manhattan Distance merupakan metode yang sering dipakai dalam perhitungan jarak. *Euclidean Dictance* merupakan salah satu metode perhitungan jarak yang digunakan untuk menghitung jarak antara dua titik dalam ruang *Euclidean* (mencakup dua dimensi, tiga dimensi, atau bahkan lebih) (Nishom, 2019). Berikut adalah rumus dari *Euclidean Distance*:

$$d(x,y) = |x - y| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad \textbf{(2.1)}$$

Keterangan:

d = jarak antara *x* dan *y*

x = data pusat klaster

y = data pada atribut

i = setiap data

n = jumlah data

x_i = data pada pusat klaster ke *i*

y_i = data pada setiap data ke *i*

Manhattan Distance merupakan satu metode perhitungan jarak yang digunakan untuk menghitung jarak dengan menerapkan konsep selisih mutlak (Baharuddin et al., 2019). Berikut adalah rumus dari *Manhattan Distance*:

$$d(x,y) = \sum_{i=1}^{n}|x_i - y_i| \quad \textbf{(2.2)}$$

Keterangan:

d = jarak antara *x* dan *y*

x = data pusat klaster

y = data pada atribut

Metode *Euclidean Distance* ini memiliki mempunyai hasil yang lebih maksimal dibandingkan dengan metode yang lain (Pribadi Wahyu et al., 2022). Selain itu metode perhitungan jarak ini lebih umum digunakan dalam metode K-NN sehingga lebih banyak referensi dari *Euclidean Distance* untuk menghitung jarak.

Algoritma K-NN memiliki kelebihan dan kekurangan dalam penerapannya. Kelebihan dari algoritma ini adalah tahan terhadap data latih yang *noisy* dan efektif saat data latih berukuran besar (Qiudandra et al., 2022). Sedangkan kekurangannya adalah perlu menentukan *hyperparameter* K (jumlah tetangga terdekat) karena *hyperparameter* ini bisa mempengaruhi kinerja dari algoritma K-NN. Biaya komputasinya cukup besar juga menjadi kekurangan dari algoritma K-NN karena jarak dari setiap sampel uji ke sampel pelatihan lengkap harus dihitung.

2.5. *GridSearchCV*

GridSearchCV adalah proses pemilihan optimal yang menemukan *hyperparameter* terbaik dalam *hyperparameter* yang ditentukan untuk mengoptimalkan performa model (Ramdhani, 2023). *GridSearchCV* bekerja dengan menggabungkan *hyperparameter* dan menghitung rata-rata dari *cross validation* dari setiap kombinasi, yang kemudian diterapkan ke model (Ailiyya, 2020).

2.6. *Confesion Matrix*

Confesion Matrix merupakan salah satu jenis teknik untuk menilai kinerja model yang dikembangkan selama proses *training* atau yang sering disebut dengan evaluasi model. *Confesion Matrix* adalah sebuah teknik untuk meninjau hasil implementasi algoritma klasifikasi menggunakan tabel (Grandis et al., 2021). Tabel 2.2 merupakan evaluasi dari *Confesion Matrix.*

Tabel 2. 2 Confesion Matrix

Kelas	Prediksi	
	Positif	Negatif
Positif	TP	FN
Negatif	FP	TN

Dimana TP (*True Positif*) merupakan ukuran berapa banyak titik data positif yang terklasifikasi secara akurat oleh sistem. Sedangkan FP (*False Positif*) merupakan ukuran berapa banyak titik data positif yang terklasifikasi namun tidak akurat dengan kata lain salah oleh sistem. FN (*False Negative*) adalah ukuran banyaknya titik data negatif yang terklasifikasi namun tidak akurat dengan kata lain salah oleh sistem. Sedangkan TN (*True Negative*) adalah ukuran banyaknya titik data negatif yang terklasifikasi secara akurat oleh sistem. *Accuracy, precision, recall* dan *f-measure* akan ditentukan selanjutnya dengan menggunakan angka-angka dalam tabel. Berikut persamaan dari *Accuracy, precision, recall* dan *f-measure* (Baharuddin et al., 2019).

$$Accuracy = \frac{(\mathrm{TP} + \mathrm{TN})}{(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})} \quad \textbf{(2.3)}$$

$$Precision = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FP})} \quad \textbf{(2.4)}$$

$$Recall = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})} \quad \textbf{(2.5)}$$

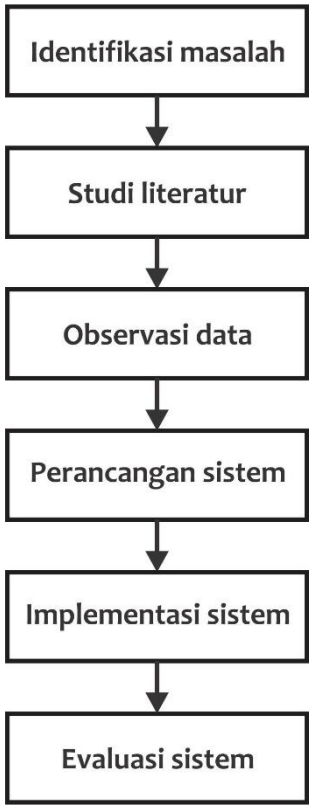
$$F - Measure = \frac{2 \text{ x } \mathrm{P} \text{ x } \mathrm{R}}{\mathrm{P} + \mathrm{R}} \quad \textbf{(2.6)}$$

BAB III

METODOLOGI PENELITIAN

3.1. Alur Penelitian

Alur penelitian berisi tahapan-tahapan yang akan dilakukan oleh peneliti dalam melakukan penelitian ini. Fungsi dari adanya alur penelitian ini adalah agar peneliti dapat terstruktur dengan baik dalam melakukan penelitian ini, dimulai dari identifikasi masalah yang ada, melakukan studi literatur, observasi data, merancang sistem, lalu dapat mengimplementasikan model *K-Nearest Neighbor* dalam mengklasifikasi kanker payudara setelah itu model di evaluasi. Berikut diagram alur penelitian yang dilakukan oleh peneliti.



Gambar 3. 1 Diagram alur penelitian

3.2. Tahap Awal

3.2.1. Identifikasi Masalah

Pada tahap ini, peneliti mengidentifikasi masalah yang akan diteliti dan dibahas dalam penelitian ini. Data dan informasi dianalisis dan dievaluasi agar peneliti dapat mengidentifikasi permasalahan yang berkaitan dengan kanker payudara. Identifikasi masalah dapat membantu membuat perencanaan dan penelitian menjadi lebih efektif, serta membantu menentukan arah dan tujuan penelitian yang akan dilakukan.

3.2.2. Studi Literatur

Pada tahap ini, peneliti melakukan kajian terhadap berbagai literatur yang ada, baik melalui buku, jurnal penelitian yang sudah ada, maupun website yang mendukung penelitian ini. Topik-topik yang dikaji seperti metode *K-Nearest Neighbor*, Kanker Payudara, *Machine Learning*, *Confusion Matrix* serta dataset. Studi literatur dapat membantu peneliti dalam pemahaman yang lebih lanjut terkait penelitian yang akan dilakukan.

3.3. Observasi Data

3.3.1. Data dan Sumber Data

Data yang digunakan merupakan dataset *Wisconsin Diagnostic Breast Cancer* (WDBC) dari *University of Wisconsin Hospital* yang diperoleh dari repositori *UCI Machine Learning Repository*. Dataset berisi 569 data dengan 32 atribut, terdiri dari kolom id dan diagnosis serta terdapat 10 atribut yang bernilai riil yang dihitung dari setiap inti sel yaitu *radius, texture, perimeter, area, smoothness,*

compactness, concavity, concave points, symmetry dan *fractal dimension*. Dataset juga memiliki 3 indikator yaitu *mean, standard error/se* dan *worst* seperti yang disajikan pada tabel 3.1.

Tabel 3. 1 Atribut dataset

No.	Atribut	Keterangan
1.	*Radius*	Rata-rata jarak dari pusat ke titik-titik pada *perimeter*
2.	*Texture*	Standar deviasi nilai dari *gray-scale*
3.	*Perimeter*	Merupakan ukuran dari panjang total yang mengelilingi tumor yang menunjukkan batas antara tumor dengan jaringan disekitarnya
4.	*Area*	Ukuran dari suatu tumor yang menunjukkan luas dari tumor tersebut
5.	*Smoothness*	*Smoothness* atau kehalusan tumor dihitung dengan membandingkan luasnya dengan rata-rata panjang garis yang mengelilinginya
6.	*Compactness*	Ukuran dari konsentrasi massa yang dimiliki oleh tumor (perimeter^2 / area - 1.0)
7.	*Concavity*	Merupakan keparahan bagian cekung dari kontur tumor
8.	*Concave points*	Merupakan jumlah bagian cekung dari kontur tumor
9.	*Symmetry*	Simetri dari tumor
10.	*Fractal dimension*	Dimensi fraktar dari tumor
11.	*Mean*	Ukuran yang megukur nilai rata-rata dari suatu fitur
12.	*Standard error/se*	Ukuran yang megukur tingkat ketidakpastian nilai rata-rata dari suatu fitur
13.	*worst*	Ukuran yang megukur nilai maksimal dari suatu fitur
14.	*Id*	Nomor unik yang dimiliki tiap pasien
15.	*Diagnosis*	Kolom yang terdiri dari salah satu 2 kelas, “B” atau “M”. “B” atau *Benign* artinya tumor tidak ganas. “M” atau *Malignant* artinya tumor itu ganas

3.3.2. Contoh Dataset Kanker Payudara

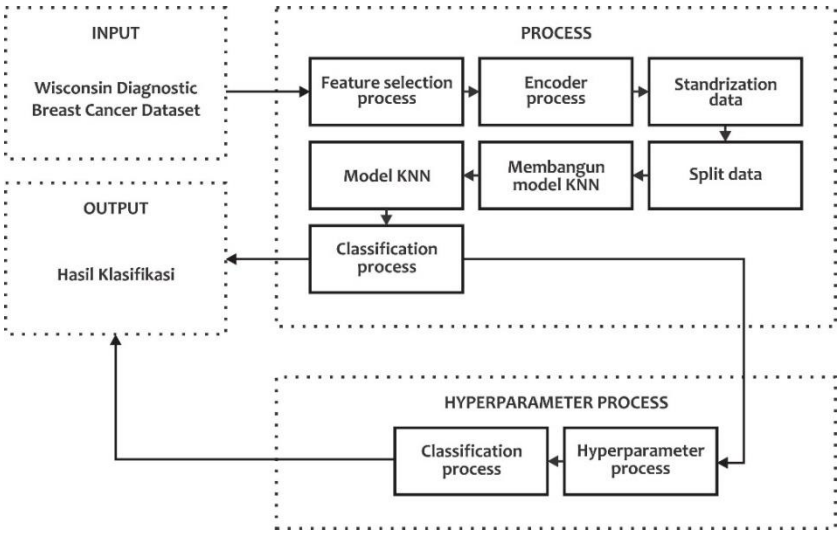
Tabel 3. 2 Contoh dataset kanker payudara

diagnosis	area_mean	radius_se	compactness_se	texture_worst	radius_worst
M	1001	1.095	0.04904	25.38	17.33
B	520	0.1852	0.01898	14.5	20.49

diagnosis	area_mean	radius_se	compactness_se	texture_worst	radius_worst
B	273.9	0.2773	0.01432	10.23	15.66
M	704.4	0.4388	0.05328	18.07	19.08
M	1404	0.6917	0.01259	29.17	35.59
M	1076	0.6289	0.03994	22.82	21.32
B	201.9	0.1563	0.01646	8.964	21.96
M	534.6	0.2871	0.02336	15.67	27.95
B	449.3	0.2636	0.01427	13.76	20.7
B	561	0.2338	0.01382	15.15	31.82
M	588.9	0.6191	0.03348	16.39	34.01
M	1024	0.6362	0.05296	19.76	24.7
M	1148	0.4357	0.03698	23.36	32.06

3.4. Perancangan Sistem

Perancangan sistem diperlukan dalam suatu penelitian agar tahapan penelitian dapat dilakukan secara terstruktur. Perancangan sistem dalam penelitian ini dapat dilihat pada gambar 3.2.



Gambar 3. 2 Desain sistem

Proses yang pertama kali dilakukan adalah input data *Wisconsin Diagnostic Breast Cancer*. Setelah data kanker diinputkan maka dilakukan *preprocessing* data

yang terdiri dari seleksi fitur, proses *encoder*, *standrization* atau normalisasi data, serta membagi data menjadi data *training* dan data *testing*. Setelah data siap, maka lanjut ke proses klasifikasi menggunakan model K-Nearest Neighbor. Dari proses tersebut akan diperoleh hasil klasifikasi yang nantinya akan dievaluasi menggunakan *confusion matrix* untuk mengetahui nilai akurasi dari model klasifikasi yang digunakan. Data yang siap tadi juga melalui proses klasifikasi namun sebelumnya terjadi proses penyetelan *hyperparameter* menggunakan *GridSearchCV* untuk mencari parameter yang optimal. Dari proses tersebut akan diperoleh hasil klasifikasi dengan menggunakan parameter yang optimal dan nantinya akan dievaluasi menggunakan *confusion matrix* untuk mengetahui nilai akurasi dari model klasifikasi yang digunakan.

3.5. *Preprocessing*

Preprocessing merupakan tahap awal dari proses pengolahan data sebelum proses utama dilakukan. Data disiapkan untuk dianalisa pada tahapan ini, tujuannya adalah untuk memaksimalkan hasil. Tahapan ini sangat penting untuk diperhatikan, karena teknik apa saja yang akan dilakukan pada tahapan ini memberikan pengaruh kepada bagaimana cara model melakukan pembelajaran terhadap data yang diberikan. Penelitian ini menggunakan 4 tahapan yakni seleksi fitur, proses *encoder*, *standrization* atau normalisasi data, serta membagi data menjadi data *training* dan data *testing*.

3.5.1. Proses seleksi fitur

Seleksi fitur merupakan proses penting dalam klasifikasi. Tujuan dari proses ini adalah memilih fitur-fitur penting dari sekumpulan fitur yang ada pada dataset. Pada dataset *Wisconsin Diagnostic Breast Cancer* terdapat beberapa fitur yang tidak digunakan dalam proses klasifikasi. Seperti fitur "id" dan "Unnamed: 32". Kedua fitur tersebut tidak penting dalam proses klasifikasi, sehingga dua fitur tersebut tidak dipilih atau digunakan dalam proses klasifikasi kanker payudara.

3.5.2. Proses *encoder*

Pada proses ini merupakan proses merubah fitur data yang bernilai kategori menjadi fitur data yang bernilai numerik. fitur "diagnosis" sendiri memiliki 2 level, yaitu "M" dan "B" yang mana 2 level tersebut dirubah menjadi "1" dan "0".

3.5.3. *Split data*

Pada tahapan ini data dibagi menjadi 2 bagian, yakni data latih (*training data*) dan data uji (*testing data*). Hal ini penting untuk dilakukan karena model yang akan dibangun harus diuji dengan data yang berbeda. Peneliti pada tahap ini menggunakan *library* yang disediakan Python untuk memudahkan dalam *split* data yakni *Scikit-Learn*. Sebelum itu, peneliti mengimport *library* tersebut serta menggunakan fungsi *train_test_split()*. Pada tahap ini peneliti membagi data dengan perbandingan 80%:20%, 70%:30% dan 60%:40%.

3.5.4. Proses normalisasi

Sebelum data-data tersebut di gunakan, sebuah data sebelumnya harus melalui proses normalisasi. Maksud dari normalisasi pada tahap ini adalah

menyamakan semua variabel yang ada pada dataset dengan rentang nilai yang sama dengan interval 0,1 yang semula semua variabel memiliki nilai interval yang berbeda satu sama lain. Peneliti pada tahap ini menggunakan *library Scikit-Learn*. Sebelum itu, peneliti mengimport *library* tersebut serta menggunakan salah satu fungsi scaler yang tersedia yaitu *MinMaxScaler.* Pada tahapan ini peneliti menggunakan rumus persamaan *Min-Max Scaling* untuk normalisasi data. Berikut rumus persamaan *Min-Max Scaling*.

$$x_{i'} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad \textbf{(3.1)}$$

Keterangan:

$x_{i'}$ = data standarisasi ke-i

x_i = data aktual ke-i

x_{min} = nilai minimal dari data ke-x

x_{max} = nilai maksimal dari data ke-x

Tabel 3. 3 Contoh dataset sebelum normalisasi

radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean
17.99	10.38	122.8	1001	0.1184
20.57	17.77	132.9	1326	0.08474
19.69	21.25	130	1203	0.1096
11.42	20.38	77.58	386.1	0.1425
20.29	14.34	135.1	1297	0.1003

Contoh hasil normalisasi untuk atribut radius_mean:

$$x_{11} = \frac{x_i - x_{min}}{x_{max} - x_{min}} = \frac{17.99 - 6.981}{28.110 - 6.981} = \frac{11.009}{21.129} = 0.521$$

$$x_{12} = \frac{x_i - x_{min}}{x_{max} - x_{min}} = \frac{20.57 - 6.981}{28.110 - 6.981} = \frac{13.589}{21.129} = 0.643$$

⋮

$$x_{15} = \frac{x_i - x_{min}}{x_{max} - x_{min}} = \frac{20.29 - 6.981}{28.110 - 6.981} = \frac{13.309}{21.129} = 0.629$$

Contoh hasil normalisasi untuk atribut texture_mean:

$$x_{21} = \frac{x_i - x_{min}}{x_{max} - x_{min}} = \frac{10.38 - 9.710}{39.280 - 9.710} = \frac{0.67}{29.57} = 0.022$$

$$x_{22} = \frac{x_i - x_{min}}{x_{max} - x_{min}} = \frac{17.77 - 9.710}{39.280 - 9.710} = \frac{8.07}{29.57} = 0.272$$

⋮

$$x_{25} = \frac{x_i - x_{min}}{x_{max} - x_{min}} = \frac{14.34 - 9.710}{39.280 - 9.710} = \frac{4.63}{29.57} = 0.156$$

Contoh hasil normalisasi untuk atribut perimeter_mean:

$$x_{31} = \frac{x_i - x_{min}}{x_{max} - x_{min}} = \frac{122.8 - 43.79}{188.5 - 43.79} = \frac{79.01}{144.71} = 0.545$$

$$x_{32} = \frac{x_i - x_{min}}{x_{max} - x_{min}} = \frac{132.9 - 43.79}{188.5 - 43.79} = \frac{89.11}{144.71} = 0.615$$

⋮

$$x_{35} = \frac{x_i - x_{min}}{x_{max} - x_{min}} = \frac{135.1 - 43.79}{188.5 - 43.79} = \frac{91.31}{144.71} = 0.630$$

Contoh hasil normalisasi untuk atribut area_mean:

$$x_{41} = \frac{x_i - x_{min}}{x_{max} - x_{min}} = \frac{1001 - 143.5}{2501 - 143.5} = \frac{857.5}{2357.5} = 0.363$$

$$x_{42} = \frac{x_i - x_{min}}{x_{max} - x_{min}} = \frac{1326 - 143.5}{2501 - 143.5} = \frac{1182.5}{2357.5} = 0.501$$

⋮

$$x_{45} = \frac{x_i - x_{min}}{x_{max} - x_{min}} = \frac{1297 - 143.5}{2501 - 143.5} = \frac{1153.5}{2357.5} = 0.489$$

Contoh hasil normalisasi untuk atribut smoothness_mean:

$$x_{51} = \frac{x_i - x_{min}}{x_{max} - x_{min}} = \frac{122.8 - 0.052}{0.163 - 0.052} = \frac{0.066}{0.111} = 0.598$$

$$x_{52} = \frac{x_i - x_{min}}{x_{max} - x_{min}} = \frac{132.9 - 0.052}{0.163 - 0.052} = \frac{0.032}{0.111} = 0.294$$

$$\vdots$$

$$x_{55} = \frac{x_i - x_{min}}{x_{max} - x_{min}} = \frac{135.1 - 0.052}{0.163 - 0.052} = \frac{0.048}{0.111} = 0.435$$

Tabel 3. 4 Contoh dataset sesudah normalisasi

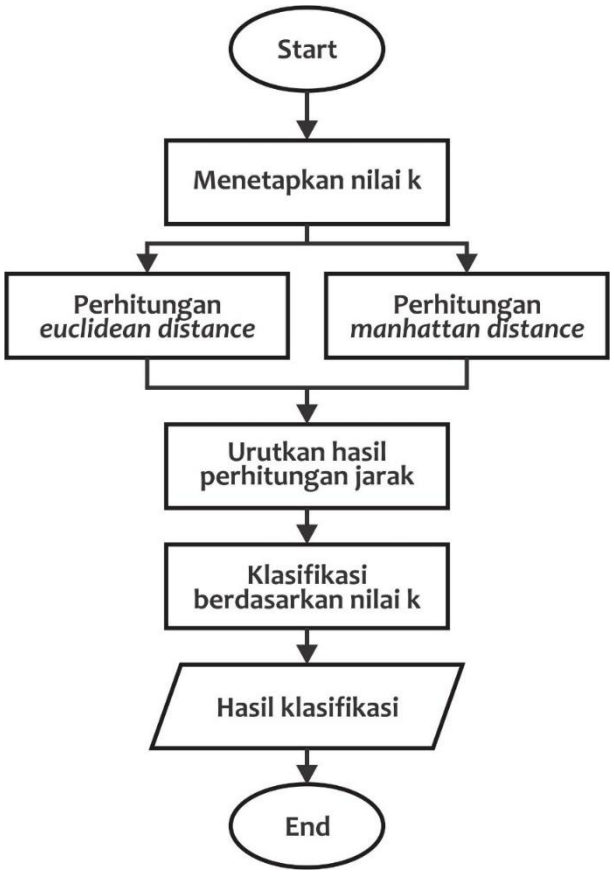
radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean
0.521	0.022	0.545	0.363	0.593
0.643	0.272	0.615	0.501	0.294
0.601	0.390	0.595	0.449	0.514
0.210	0.360	0.233	0.102	0.811
0.629	0.156	0.630	0.489	0.435

3.6. Implementasi Metode *K-Nearest Neighbor*

Adapun langkah-langkah dalam implementasi metode *K-Nearest Neighbor* pada penelitian ini adalah sebagai berikut:

1. Menetapkan nilai k yang terbaik.
2. Setelah menentukan nilai k yang terbaik langkah selanjutnya adalah klasifikasi menggunakan model K-NN. Pada tahapan ini klasifikasi dilakukan dengan menghitung jarak antar data menggunakan metode *Euclidean Distance* dan *Manhattan Distance.*
3. Mengurutkan hasil perhitungan langkah ke-2 secara *ascending*.
4. Klasifikasi *nearest neighbor* berdasarkan nilai k.
5. Hasil klasifikasi menggunakan model *K-Nearest Neighbor.*

Adapun *flowchart* dari implementasi metode *K-Nearest Neighbor* pada penelitian ini adalah sebagaimana gambar 3.3.



Gambar 3. 3 Flowchart K-Nearest Neighbor

3.6.1. Contoh Implementasi *K-Nearest Neighbor*

Adapun contoh dari implementasi model *K-Nearest Neighbor* pada penelitian ini adalah sebagai berikut. Pada contoh ini peneliti akan menggunakan 10 data dan 5 atribut tidak termasuk atribut diagnosis yang digunakan sebagai label.

Tabel 3. 5 Contoh dataset

diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean
M	16.13	20.68	108.1	798.8	0.117
M	19.81	22.15	130	1260	0.09831
B	13.54	14.36	87.46	566.3	0.09779

diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean
B	13.08	15.71	85.63	520	0.1075
B	10.49	19.29	67.41	336.1	0.09989
M	15.34	14.26	102.5	704.4	0.1073
M	21.16	23.04	137.2	1404	0.09428
M	16.65	21.38	110	904.6	0.1121
B	13.03	18.42	82.61	523.8	0.08983
M	14.99	25.2	95.54	698.8	0.09387

Dari data tersebut akan dilakukan *preprocesing* mulai dari *data cleaning, transformasi data, outlier data* dan *split data* menjadi *data training* dan *data testing* setelah itu dilakukan standarisasi data. Pada contoh implementasi *K-Nearest Neighbor*, peneliti membagi data *train* dan *test* sebesar 90:10. Berikut adalah tabel *data training* dan *data testing.*

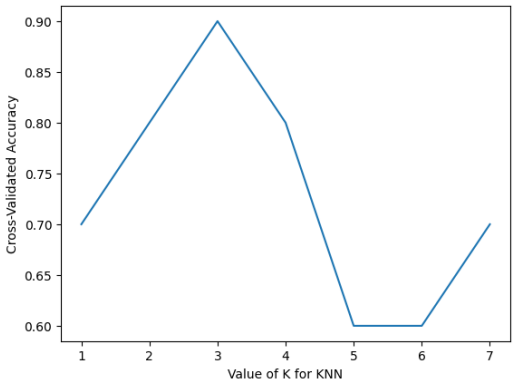
Tabel 3. 6 Data training

radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean
0.873	0.721	0.896	0.865	0.312
0.454	0	0.502	0.344	0.642
0.528	0.586	0.583	0.433	1
0.577	0.650	0.610	0.532	0.81
0.285	0.009	0.287	0.215	0.292
0.421	1	0.403	0.339	0.148
0	0.459	0	0	0.370
0.242	0.132	0.261	0.172	0.650
1	0.802	1	1	0.163

Tabel 3. 7 Data testing

radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean
0.238	0.380	0.217	0.175	0

Setelah data terbagi menjadi *data training* dan *data testing* maka selanjutnya adalah menentukan nilai parameter k yang terbaik, dengan menggunakan metode *GridSearchCV*. Range k yang digunakan peneliti adalah 1-10. Berikut adalah hasil dari *cross validation accuracy* dari pencarian parameter terbaik menggunakan metode *GridSearchCV*.



Gambar 3. 4 Cross validation accuracy

Dapat dilihat pada gambar 3.4. bahwasannya nilai parameter k yang memiliki *accuracy* tertinggi adalah nilai k = 3. Sehingga pada contoh implementasi *K-Nearest Neighbor* peneliti menggunakan nilai parameter k = 3. Setelah nilai k ditentukan, maka langkah selanjutnya adalah melakukan perhitungan jarak. Perhitungan jarak ini menggunakan metode *Euclidean dan Manhattan distance.* Perhitungan jarak metode *Euclidean distance* menggunakan persamaan 2.1.

Perhitungan *Euclidean data training* 1 dengan *data testing*:

$$E = \sqrt{\begin{matrix}(0.873-0.238)^2+(0.721-0.380)^2 \\ +(0.896-0.217)^2+(0.865-0.175)^2+(0.312-0)^2\end{matrix}} = 1.246$$

Perhitungan *Euclidean data training* 2 dengan *data testing*:

$$E = \sqrt{\begin{matrix}(0.454 - 0.238)^2 + (0.721 - 0)^2 \\ + (0.502 - 0.217)^2 + (0.344 - 0.175)^2 + (0.642 - 0)^2\end{matrix}} = 0.845$$

Perhitungan *Euclidean data training* 3 dengan *data testing*:

$$E = \sqrt{\begin{matrix}(0.528 - 0.238)^2 + (0.721 - 0.586)^2 \\ + (0.583 - 0.217)^2 + (0.433 - 0.175)^2 + (1 - 0)^2\end{matrix}} = 1.151$$

⋮

Perhitungan *Euclidean data training* 9 dengan *data testing*:

$$E = \sqrt{\begin{matrix}(1 - 0.238)^2 + (0.802 - 0.380)^2 \\ + (1 - 0.217)^2 + (1 - 0.175)^2 + (0.163 - 0)^2\end{matrix}} = 1.441$$

Perhitungan jarak metode *Euclidean distance* menggunakan persamaan 2.2.

Perhitungan *Manhattan data training* 1 dengan *data testing*:

$$E = (0.873 - 0.238) + (0.721 - 0.380) + (0.896 - 0.217) + (0.865 - 0.175) + (0.312 - 0) = 2.656$$

Perhitungan *Manhattan data training* 2 dengan *data testing*:

$$E = (0.454 - 0.238) + (0 - 0.380) + (0.502 - 0.217) + (0.344 - 0.175) + (0.642 - 0) = 0.933$$

Perhitungan *Manhattan data training* 3 dengan *data testing*:

$$E = (0.528 - 0.238) + (0.586 - 0.380) + (0.583 - 0.217) + (0.433 - 0.175) + (1 - 0) = 2.119$$

⋮

Perhitungan *Manhattan data training* 9 dengan *data testing*:

$$E = (1 - 0.238) + (0.802 - 0.380) + (1 - 0.217) + (1 - 0.175) + (0.163 - 0) = 2.954$$

Tabel 3. 8 Perolehan jarak data uji euclidean distance

Data ke-	*Eucledian Distance*	Rangking	Diagnosis
1	1.246	8	1
2	0.845	5	1
3	1.151	7	1
⋮	⋮	⋮	⋮
9	1.441	9	1

Tabel 3. 9 Perolehan jarak data uji manhattan distance

Data ke-	*Manhattan Distance*	Rangking	Diagnosis
1	2.656	8	1
2	0.933	4	1
3	2.119	6	1
⋮	⋮	⋮	⋮
9	2.954	9	1

Langkah selanjutnya setelah menghitung jarak antar *data training* dengan *data testing* adalah mengurutkan data secara *ascending*. Berikut adalah data dari hasil perhitungan jarak dengan *eucledian dan Manhattan distance* yang diurutkan secara *ascending*.

Tabel 3. 10 Hasil urutan jarak euclidean distance

Data ke-	*Eucledian Distance*	Rangking	Diagnosis
5	0.481	1	0
7	0.527	2	0
8	0.697	3	0
⋮	⋮	⋮	⋮
9	1.441	9	1

Tabel 3. 11 Hasil urutan jarak manhattan distance

Data ke-	*Manhattan Distance*	Rangking	Diagnosis
5	0.078	1	0
7	0.181	2	0
8	0.447	3	0
⋮	⋮	⋮	⋮

Data ke-	*Manhattan Distance*	Rangking	Diagnosis
9	1.441	9	1

Berdasarkan tabel 3.11 dan tabel 3.12 menunjukkan perolehan jarak antar *data training* dengan *data testing* yang diurutkan secara *ascending*. Karena pada contoh implementasi *K-Nearest Neighbor* menggunakan nilai k = 3 maka diambil 3 terkecil. Maka pada tabel 3.11 dan tabel 3.12 label mayoritas yang muncul adalah 0 maka dapat disimpulkan bahwasannya *data testing* termasuk dalam klasifikasi kanker payudara dengan label 0 atau termasuk kanker payudara berjenis jinak.

3.7. Proses *Hyperparameter*

Pada penelitian ini dalam proses *Hyperparameter* untuk menemukan parameter yang optimal menggunakan GridSearchCV. GridSearchCV digunakan untuk mencari *hyperparameter* terbaik untuk membentuk model yang optimal sehingga menghasilkan nilai akurasi yang terbaik dengan *cross validation*. Hasil dari *cross validation* berupa *best_score* yang merupakan skor rata-rata dari akurasi silang tertinggi (Priya C, 2021). Berikut parameter penting peneliti.

Tabel 3. 12 Parameter penting

Parameter	Keterangan
`n_neighbors=list(range(1,11))`	Nilai jarak k yang dicari mulai dari 1-10
`weight_options=["uniform","distance"]`	Berfungsi mengontrol bagaimana tetangga diberi "bobot" saat melakukan prediksi

Sebelumnya peneliti menentukan parameter apa saja yang akan dioptimalkan. Peneliti menggunakan parameter yang sesuai dengan tabel xx.xx. GridSearchCV dirancang untuk menentukan hyperparameter optimal yang akan menghasilkan performa terbaik untuk model K-Nearest Neighbors. Dalam

penelitian ini, nilai *cross validation* ditetapkan sebagai 10 yang menandakan bahwa model dan parameter terkaitnya akan divalidasi sebanyak sepuluh kali. Data akan dibagi sebanyak 10 bagian sama besar dengan (9 bagian untuk training dan 1 bagian untuk testing).

3.8. Skema Evaluasi

Tahapan terakhir dalam penelitian ini adalah melakukan evaluasi terhadap kinerja dari penggunaan *K-Nearest Neighbor* dalam melakukan klasifikasi kanker payudara berdasarkan data *Wisconsin Diagnostic Breast Cancer*. Pada tahapan ini peneliti menggunakan *confusion matrix* untuk menunjukkan berapa banyak prediksi model *K-Nearest Neighbor* yang akurat dan salah pada setiap kelas. Hasil dari evaluasi akan dimunculkan dalam bentuk *confusion matrix* 3 x 3 pada tabel 2.2. *confusion matrix* menunjukkan jumlah nilai True Positive (TP), False Positive, True Negative, False Negative yang merupakan representasi hasil proses klasifikasi (Fauziningrum & Suryaningsih, 2021). Berdasarkan evaluasi *confusion matrix* juga dapat diketahui berbagai parameter pengukuran kinerja model, yaitu *Akurasi, Precision, Recall dan F-Measure*.

BAB IV

UJI COBA DAN PEMBAHASAN

4.1. Skenario Uji Coba

Penelitian ini mengusulkan penerapan 3 model pembagian data *training* dan data *testing*, yaitu model A dengan perbandingan 80% data *training*: 20% data *testing*, model B dengan perbandingan 70% data *training*: 30% data *testing*, model C dengan perbandingan 60% data *training*: 40% data *testing*. Pembagian ini juga berfungsi untuk mengurangi permasalahan *overfitting* (Yuliany et al., 2022).

Tabel 4. 1 Pembagian dataset

Model	Perbandingan	Jumlah data	Data *training*	Data *testing*
A	80%:20%	569	455	114
B	70%:30%		398	171
C	60%:40%		341	228

Skenario uji coba pada penelitian ini terbagi menjadi 3 model dengan menggunakan semua fitur yang ada pada *dataset* dengan jumlah 31 variabel. Selanjutnya, pada masing-masing pembagian model tersebut akan dilakukan perbandingan antara model *K-Nearest Neighbor* tanpa adanya penyetelan *hyperparameter* atau secara *default* dengan model *K-Nearest Neighbor* dengan penyetelan *hyperparameter* menggunakan *GridSearchCV* untuk mencari nilai k atau jarak yang terbaik.

4.2. Hasil Uji Coba

Sub bab ini menjelaskan tentang hasil analisis dari pengujian sistem berdasarkan skenario uji coba pada sub bab 4.1. Pada sub bab ini berguna untuk

mengetahui performa dari metode *K-Nearest Neighbor* dalam mengklasifikasi kanker payudara pada *dataset Wisconsin Diagnostic Breast Cancer* yang telah terbagi menjadi 3 model pengujian. Setiap model pengujian nantinya dapat diketahui nilai akurasi dari setiap model yang ada.

4.2.1. Pengujian Model A

Pada model kedua perbandingan yang digunakan antara data *training* dengan data *testing* sebesar 80:20 atau 455 data *training* dan 114 data *testing*. Pada pengujian model ini terdapat 2 skenario pengujian utama yaitu *default* dan adanya penyetelan *hyperparameter* menggunakan *GridSeacrhCV*.

4.2.1.1. Euclidean *Default*

Pada skenario pengujian pertama yakni secara *default* atau tanpa adanya penyetelan *hyperparameter* dengan menggunakan nilai n_neighbors = 5 serta menggunakan *euclidean distance* sebagai metode perhitungan jarak maka didapatkan hasil sebagai berikut.

Tabel 4. 2 Hasil prediksi euclidean default model A

radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	aktual	prediksi
0.322732	0.205614	0.3223	0.186893	0.378531	M	B
0.271617	0.269192	0.256997	0.151771	0.163049	B	B
0.290075	0.433886	0.298735	0.16369	0.373575	B	M
⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.357281	0.325668	0.348697	0.218961	0.211815	B	B
0.385205	0.235712	0.380001	0.243097	0.260184	M	B
0.430167	0.336152	0.416765	0.285981	0.289127	M	M

Tabel 4. 3 Confusion matrix euclidean default model A

Data aktual	Prediksi	
	B	M
B	71	1
M	4	38

Tabel 4.3. menunjukkan bahwa model *K-Nearest Neighbor* memprediksi terdapat 71 jenis kanker jinak dan hasil sebenarnya terdeteksi kanker jinak, model *K-Nearest Neighbor* memprediksi terdapat 4 jenis kanker jinak akan tetapi hasil sebenarnya terdeteksi kanker ganas, model *K-Nearest Neighbor* memprediksi terdapat 1 jenis kanker ganas akan tetapi hasil sebenarnya terdeteksi kanker jinak, model *K-Nearest Neighbor* memprediksi terdapat 38 jenis kanker ganas dan hasil sebenarnya terdeteksi kanker ganas.

Dari tabel 4.3. dapat diketahui nilai dari *accuracy, precision, recall dan f-measure*. Perhitungan dari nilai-nilai tersebut adalah sebagai berikut.

$$Accuracy = \frac{(71 + 38)}{(71 + 38 + 4 + 1)} = 95.6\%$$

$$Precision = \frac{71}{(71 + 4)} = 94.6\%$$

$$Recall = \frac{71}{(71 + 1)} = 98.6\%$$

$$F - Measure = \frac{2 \text{ x } 0.947 \text{ x } 0.986}{0.947 + 0.986} = 96.3\%$$

4.2.1.2. Manhattan *Default*

Pada skenario pengujian kedua yakni secara *default* atau tanpa adanya penyetelan *hyperparameter* dengan menggunakan nilai n_neighbors = 5 serta

menggunakan *manhattan distance* sebagai metode perhitungan jarak maka didapatkan hasil sebagai berikut.

Tabel 4. 4 Hasil prediksi manhattan default model A

radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	aktual	prediksi
0.322732	0.205614	0.3223	0.186893	0.378531	M	B
0.271617	0.269192	0.256997	0.151771	0.163049	B	B
0.290075	0.433886	0.298735	0.16369	0.373575	B	M
⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.357281	0.325668	0.348697	0.218961	0.211815	B	B
0.385205	0.235712	0.380001	0.243097	0.260184	M	B
0.430167	0.336152	0.416765	0.285981	0.289127	M	M

Tabel 4. 5 Confusion matrix manhattan default model A

Data aktual	Prediksi	
	B	M
B	71	1
M	4	38

Tabel 4.5. menunjukkan bahwa model *K-Nearest Neighbor* memprediksi terdapat 71 jenis kanker jinak dan hasil sebenarnya terdeteksi kanker jinak, model *K-Nearest Neighbor* memprediksi terdapat 4 jenis kanker jinak akan tetapi hasil sebenarnya terdeteksi kanker ganas, model *K-Nearest Neighbor* memprediksi terdapat 1 jenis kanker ganas akan tetapi hasil sebenarnya terdeteksi kanker jinak, model *K-Nearest Neighbor* memprediksi terdapat 38 jenis kanker ganas dan hasil sebenarnya terdeteksi kanker ganas.

Dari tabel 4.5. dapat diketahui nilai dari *accuracy, precision, recall dan f-measure*. Perhitungan dari nilai-nilai tersebut adalah sebagai berikut.

$$Akurasi = \frac{(71 + 38)}{(71 + 38 + 4 + 1)} = 95.6\%$$

$$Precision = \frac{71}{(71+4)} = 94.6\%$$

$$Recall = \frac{71}{(71+1)} = 98.6\%$$

$$F - Measure = \frac{2 \text{ x } 0.947 \text{ x } 0.986}{0.947 + 0.986} = 96.3\%$$

4.2.1.3. Euclidean Penyetelan *Hyperparameter*

Pada skenario pengujian ketiga yakni dengan adanya penyetelan *hyperparameter*. Pengujian ini dilakukan dengan menggunakan *GridSearchCV* untuk mencari konfigurasi parameter nilai k atau n_neighbor terbaik.

Tabel 4. 6 Hasil rata-rata skor pengujian model A

mean_test_score	std_test_score	params
0.955942	0.028113	{'n_neighbors': 1, 'weights': 'uniform'}
0.955942	0.028113	{'n_neighbors': 1, 'weights': 'distance'}
0.960435	0.027343	{'n_neighbors': 2, 'weights': 'uniform'}
0.955942	0.028113	{'n_neighbors': 2, 'weights': 'distance'}
0.971304	0.024324	{'n_neighbors': 3, 'weights': 'uniform'}
0.971304	0.024324	{'n_neighbors': 3, 'weights': 'distance'}
0.964783	0.032818	{'n_neighbors': 4, 'weights': 'uniform'}
0.971304	0.024324	{'n_neighbors': 4, 'weights': 'distance'}
0.962464	0.032893	{'n_neighbors': 5, 'weights': 'uniform'}
0.962464	0.032893	{'n_neighbors': 5, 'weights': 'distance'}
0.966908	0.031635	{'n_neighbors': 6, 'weights': 'uniform'}
0.971304	0.024324	{'n_neighbors': 6, 'weights': 'distance'}
0.971304	0.028093	{'n_neighbors': 7, 'weights': 'uniform'}
0.971304	0.028093	{'n_neighbors': 7, 'weights': 'distance'}
0.964734	0.026653	{'n_neighbors': 8, 'weights': 'uniform'}
0.973527	0.023793	{'n_neighbors': 8, 'weights': 'distance'}
0.966908	0.028491	{'n_neighbors': 9, 'weights': 'uniform'}
0.96913	0.026589	{'n_neighbors': 9, 'weights': 'distance'}
0.964734	0.026653	{'n_neighbors': 10, 'weights': 'uniform'}
0.971353	0.022216	{'n_neighbors': 10, 'weights': 'distance'}

Pada tabel 4.6. menunjukkan bahwa nilai k terbaik yang digunakan pada model *K-Nearest Neighbor* adalah k = 8 dengan skor rata-rata data pengujian (validasi silang) untuk setiap kombinasi *hyperparameter* adalah = 0.97. Sehingga pada skenario pengujian ketiga ini menggunakan nilai n_neighbor = 8 serta menggunakan *euclidean distance* maka didapatkan hasil sebagai berikut.

Tabel 4. 7 Hasil prediksi euclidean hyperparameter model A

radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	aktual	Prediksi
0.360594	0.459249	0.346762	0.221082	0.240955	M	M
0.192106	0.240785	0.187478	0.097434	0.447914	B	B
0.267358	0.37369	0.265082	0.142906	0.364952	B	B
⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.357281	0.325668	0.348697	0.218961	0.211815	B	B
0.385205	0.235712	0.380001	0.243097	0.260184	M	B
0.430167	0.336152	0.416765	0.285981	0.289127	M	M

Tabel 4. 8 Confusion matrix euclidean hyperparameter model A

Data aktual	Prediksi	
	B	M
B	71	1
M	3	39

Tabel 4.8. menunjukkan bahwa model *K-Nearest Neighbor* memprediksi terdapat 71 jenis kanker jinak dan hasil sebenarnya terdeteksi kanker jinak, model *K-Nearest Neighbor* memprediksi terdapat 3 jenis kanker jinak akan tetapi hasil sebenarnya terdeteksi kanker ganas, model *K-Nearest Neighbor* memprediksi terdapat 1 jenis kanker ganas akan tetapi hasil sebenarnya terdeteksi kanker jinak, model *K-Nearest Neighbor* memprediksi terdapat 39 jenis kanker ganas dan hasil sebenarnya terdeteksi kanker ganas.

Dari tabel 4.8. dapat diketahui nilai dari *accuracy, precision, recall dan f-measure*. Perhitungan dari nilai-nilai tersebut adalah sebagai berikut.

$$Akurasi = \frac{(71 + 39)}{(71 + 39 + 3 + 1)} = 96.4\%$$

$$Precision = \frac{71}{(71 + 3)} = 95.9\%$$

$$Recall = \frac{71}{(71 + 1)} = 98.6\%$$

$$F - Measure = \frac{2 \text{ x } 0.959 \text{ x } 0.986}{0.959 + \ 0.986} = 97.4\%$$

4.2.1.4. Manhattan Penyetelan *Hyperparameter*

Pada tabel 4.6. menunjukkan bahwa nilai k terbaik yang digunakan pada model *K-Nearest Neighbor* adalah k = 8 dengan skor rata-rata data pengujian (validasi silang) untuk setiap kombinasi *hyperparameter* adalah = 0.97. Sehingga pada skenario pengujian keempat ini menggunakan nilai n_neighbor = 8 serta menggunakan *manhattan distance* maka didapatkan hasil sebagai berikut.

Tabel 4. 9 Hasil prediksi manhattan hyperparameter model A

radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	aktual	Prediksi
0.360594	0.459249	0.346762	0.221082	0.240955	M	B
0.192106	0.240785	0.187478	0.097434	0.447914	B	B
0.267358	0.37369	0.265082	0.142906	0.364952	B	B
⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.357281	0.325668	0.348697	0.218961	0.211815	B	B
0.385205	0.235712	0.380001	0.243097	0.260184	M	B
0.430167	0.336152	0.416765	0.285981	0.289127	M	M

Tabel 4. 10 Hasil prediksi manhattan hyperparameter model A

Data aktual	Prediksi	
	B	M
B	71	1
M	4	38

Tabel 4.10. menunjukkan bahwa model *K-Nearest Neighbor* memprediksi terdapat 71 jenis kanker jinak dan hasil sebenarnya terdeteksi kanker jinak, model *K-Nearest Neighbor* memprediksi terdapat 4 jenis kanker jinak akan tetapi hasil sebenarnya terdeteksi kanker ganas, model *K-Nearest Neighbor* memprediksi terdapat 1 jenis kanker ganas akan tetapi hasil sebenarnya terdeteksi kanker jinak, model *K-Nearest Neighbor* memprediksi terdapat 38 jenis kanker ganas dan hasil sebenarnya terdeteksi kanker ganas.

Dari tabel 4.10. dapat diketahui nilai dari *accuracy, precision, recall dan f-measure*. Perhitungan dari nilai-nilai tersebut adalah sebagai berikut.

$$Akurasi = \frac{(71 + 38)}{(71 + 38 + 4 + 1)} = 95.6\%$$

$$Precision = \frac{71}{(71 + 4)} = 94.6\%$$

$$Recall = \frac{71}{(71 + 1)} = 98.6\%$$

$$F - Measure = \frac{2 \text{ x } 0.946 \text{ x } 0.986}{0.946 + \ 0.986} = 96.3\%$$

4.2.2. Pengujian Model B

Pada model kedua perbandingan yang digunakan antara data *training* dengan data *testing* sebesar 70:30 atau 398 data *training* dan 171 data *testing*. Pada

pengujian model ini terdapat 2 skenario pengujian utama yaitu default dan adanya penyetelan *hyperparameter* menggunakan *GridSeacrhCV*.

4.2.2.1. Euclidean *Default*

Pada skenario pengujian pertama yakni secara default atau tanpa adanya penyetelan *hyperparameter* dengan menggunakan nilai n_neighbors = 2 serta menggunakan *euclidean distance* sebagai metode perhitungan jarak maka didapatkan hasil sebagai berikut.

Tabel 4. 11 Hasil prediksi euclidean default model B

radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	aktual	prediksi
0.439159	0.411566	0.44026	0.290087	0.535137	M	B
0.659709	0.520122	0.685578	0.510932	0.469719	M	M
0.239907	0.439973	0.241587	0.129187	0.067797	B	B
⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.352075	0.34021	0.350287	0.211845	0.347012	M	B
0.234228	0.399729	0.226246	0.125281	0.348597	B	B
0.593923	0.769699	0.581922	0.458289	0.214987	M	M

Tabel 4. 12 Confusion matrix euclidean default model B

Data aktual	Prediksi	
	B	M
B	107	0
M	9	55

Tabel 4.12. menunjukkan bahwa model *K-Nearest Neighbor* memprediksi terdapat 107 jenis kanker jinak dan hasil sebenarnya terdeteksi kanker jinak, model *K-Nearest Neighbor* memprediksi terdapat 9 jenis kanker jinak akan tetapi hasil sebenarnya terdeteksi kanker ganas, model *K-Nearest Neighbor* memprediksi terdapat 0 jenis kanker ganas akan tetapi hasil sebenarnya terdeteksi kanker jinak,

model *K-Nearest Neighbor* memprediksi terdapat 55 jenis kanker ganas dan hasil sebenarnya terdeteksi kanker ganas.

Dari tabel 4.12. dapat diketahui nilai dari *accuracy, precision, recall dan f-measure*. Perhitungan dari nilai-nilai tersebut adalah sebagai berikut.

$$Akurasi = \frac{(107 + 55)}{(107 + 55 + 9 + 0)} = 94.7\%$$

$$Precision = \frac{107}{(107 + 9)} = 92.2\%$$

$$Recall = \frac{107}{(107 + 0)} = 100\%$$

$$F - Measure = \frac{2 \text{ x } 0.922 \text{ x } 1}{0.922 + 1} = 95.9\%$$

4.2.2.2. Manhattan *Default*

Pada skenario pengujian kedua yakni secara *default* atau tanpa adanya penyetelan *hyperparameter* dengan menggunakan nilai n_neighbors = 2 serta menggunakan *manhattan distance* sebagai metode perhitungan jarak maka didapatkan hasil sebagai berikut.

Tabel 4. 13 Hasil prediksi manhattan default model B

radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	aktual	prediksi
0.439159	0.411566	0.44026	0.290087	0.535137	M	M
0.659709	0.520122	0.685578	0.510932	0.469719	M	M
0.239907	0.439973	0.241587	0.129187	0.067797	B	B
⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.352075	0.34021	0.350287	0.211845	0.347012	M	B
0.234228	0.399729	0.226246	0.125281	0.348597	B	B
0.593923	0.769699	0.581922	0.458289	0.214987	M	M

Tabel 4. 14 Confusion matrix manhattan default model B

Data aktual	Prediksi	
	B	M
B	107	0
M	6	58

Tabel 4.14. menunjukkan bahwa model *K-Nearest Neighbor* memprediksi terdapat 107 jenis kanker jinak dan hasil sebenarnya terdeteksi kanker jinak, model *K-Nearest Neighbor* memprediksi terdapat 6 jenis kanker jinak akan tetapi hasil sebenarnya terdeteksi kanker ganas, model *K-Nearest Neighbor* memprediksi terdapat 0 jenis kanker ganas akan tetapi hasil sebenarnya terdeteksi kanker jinak, model *K-Nearest Neighbor* memprediksi terdapat 58 jenis kanker ganas dan hasil sebenarnya terdeteksi kanker ganas.

Dari tabel 4.14. dapat diketahui nilai dari *accuracy, precision, recall dan f-measure*. Perhitungan dari nilai-nilai tersebut adalah sebagai berikut.

$$Akurasi = \frac{(107 + 58)}{(107 + 58 + 6 + 0)} = 96.4\%$$

$$Precision = \frac{107}{(107 + 6)} = 94.7\%$$

$$Recall = \frac{107}{(107 + 0)} = 100\%$$

$$F - Measure = \frac{2 \text{ x } 0.947 \text{ x } 1}{0.947 + 1} = 97.3\%$$

4.2.2.3. Euclidean Penyetelan *Hyperparameter*

Pada skenario pengujian ketiga yakni dengan adanya penyetelan *hyperparameter*. Pengujian ini dilakukan dengan menggunakan *GridSearchCV* untuk mencari konfigurasi parameter nilai k atau n_neighbor terbaik.

Tabel 4. 15 Hasil rata-rata skor pengujian model B

mean_test_score	std_test_score	params
0.955942	0.028113	{'n_neighbors': 1, 'weights': 'uniform'}
0.955942	0.028113	{'n_neighbors': 1, 'weights': 'distance'}
0.960435	0.027343	{'n_neighbors': 2, 'weights': 'uniform'}
0.955942	0.028113	{'n_neighbors': 2, 'weights': 'distance'}
0.971304	0.024324	{'n_neighbors': 3, 'weights': 'uniform'}
0.971304	0.024324	{'n_neighbors': 3, 'weights': 'distance'}
0.964783	0.032818	{'n_neighbors': 4, 'weights': 'uniform'}
0.971304	0.024324	{'n_neighbors': 4, 'weights': 'distance'}
0.962464	0.032893	{'n_neighbors': 5, 'weights': 'uniform'}
0.962464	0.032893	{'n_neighbors': 5, 'weights': 'distance'}
0.966908	0.031635	{'n_neighbors': 6, 'weights': 'uniform'}
0.971304	0.024324	{'n_neighbors': 6, 'weights': 'distance'}
0.971304	0.028093	{'n_neighbors': 7, 'weights': 'uniform'}
0.971304	0.028093	{'n_neighbors': 7, 'weights': 'distance'}
0.964734	0.026653	{'n_neighbors': 8, 'weights': 'uniform'}
0.973527	0.023793	{'n_neighbors': 8, 'weights': 'distance'}
0.966908	0.028491	{'n_neighbors': 9, 'weights': 'uniform'}
0.96913	0.026589	{'n_neighbors': 9, 'weights': 'distance'}
0.964734	0.026653	{'n_neighbors': 10, 'weights': 'uniform'}
0.971353	0.022216	{'n_neighbors': 10, 'weights': 'distance'}

Pada tabel 4.15. menunjukkan bahwa nilai k terbaik yang digunakan pada model K-Nearest Neighbor adalah k = 9 dengan skor rata-rata data pengujian (validasi silang) untuk setiap kombinasi *hyperparameter* adalah = 0.96. Sehingga pada skenario pengujian ketiga ini menggunakan nilai n_neighbor = 9 serta menggunakan *euclidean distance* maka didapatkan hasil sebagai berikut.

Tabel 4. 16 Hasil prediksi euclidean hyperparameter model B

radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	aktual	Prediksi
0.354915	0.397362	0.348697	0.214264	0.316483	M	M
0.240854	0.126141	0.235229	0.128083	0.470711	B	B
0.312793	0.410213	0.299703	0.177245	0.168996	B	B
⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.352075	0.34021	0.350287	0.211845	0.347012	M	B
0.234228	0.399729	0.226246	0.125281	0.348597	B	B
0.593923	0.769699	0.581922	0.458289	0.214987	M	M

Tabel 4. 17 Confusion matrix euclidean hyperparameter model B

Data aktual	Prediksi	
	B	M
B	107	0
M	4	60

Tabel 4.17. menunjukkan bahwa model *K-Nearest Neighbor* memprediksi terdapat 107 jenis kanker jinak dan hasil sebenarnya terdeteksi kanker jinak, model *K-Nearest Neighbor* memprediksi terdapat 4 jenis kanker jinak akan tetapi hasil sebenarnya terdeteksi kanker ganas, model *K-Nearest Neighbor* memprediksi terdapat 0 jenis kanker ganas akan tetapi hasil sebenarnya terdeteksi kanker jinak, model *K-Nearest Neighbor* memprediksi terdapat 60 jenis kanker ganas dan hasil sebenarnya terdeteksi kanker ganas.

Dari tabel 4.17. dapat diketahui nilai dari *accuracy, precision, recall dan f-measure*. Perhitungan dari nilai-nilai tersebut adalah sebagai berikut.

$$Akurasi = \frac{(107 + 60)}{(107 + 60 + 4 + 0)} = 97.6\%$$

$$Precision = \frac{107}{(107 + 4)} = 96.4\%$$

$$Recall = \frac{107}{(107 + 0)} = 100\%$$

$$F - Measure = \frac{2 \text{ x } 0.964 \text{ x } 1}{0.964 + 1} = 98.1\%$$

4.2.2.4. Manhattan Penyetelan *Hyperparameter*

Pada tabel 4.15. menunjukkan bahwa nilai k terbaik yang digunakan pada model K-Nearest Neighbor adalah k = 9 dengan skor rata-rata data pengujian (validasi silang) untuk setiap kombinasi *hyperparameter* adalah = 0.96. Sehingga pada skenario pengujian keempat ini menggunakan nilai n_neighbor = 9 serta menggunakan *manhattan distance* maka didapatkan hasil sebagai berikut.

Tabel 4. 18 Hasil prediksi manhattan hyperparameter model B

radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	aktual	Prediksi
0.354915	0.397362	0.348697	0.214264	0.316483	M	B
0.240854	0.126141	0.235229	0.128083	0.470711	B	B
0.312793	0.410213	0.299703	0.177245	0.168996	B	B
⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.352075	0.34021	0.350287	0.211845	0.347012	M	B
0.234228	0.399729	0.226246	0.125281	0.348597	B	B
0.593923	0.769699	0.581922	0.458289	0.214987	M	M

Tabel 4. 19 Confusion matrix manhattan hyperparameter model B

Data aktual	Prediksi	
	B	M
B	107	0
M	7	57

Tabel 4.19. menunjukkan bahwa model *K-Nearest Neighbor* memprediksi terdapat 107 jenis kanker jinak dan hasil sebenarnya terdeteksi kanker jinak, model *K-Nearest Neighbor* memprediksi terdapat 7 jenis kanker jinak akan tetapi hasil

sebenarnya terdeteksi kanker ganas, model *K-Nearest Neighbor* memprediksi terdapat 0 jenis kanker ganas akan tetapi hasil sebenarnya terdeteksi kanker jinak, model *K-Nearest Neighbor* memprediksi terdapat 57 jenis kanker ganas dan hasil sebenarnya terdeteksi kanker ganas.

Dari tabel 4.19. dapat diketahui nilai dari *accuracy, precision, recall dan f-measure*. Perhitungan dari nilai-nilai tersebut adalah sebagai berikut.

$$Akurasi = \frac{(107 + 57)}{(107 + 57 + 7 + 0)} = 95.9\%$$

$$Precision = \frac{107}{(107 + 7)} = 93.8\%$$

$$Recall = \frac{107}{(107 + 0)} = 100\%$$

$$F - Measure = \frac{2 \text{ x } 0.938 \text{ x } 1}{0.938 + 1} = 96.8\%$$

4.2.3. Pengujian Model C

Pada model ketiga perbandingan yang digunakan antara data *training* dengan data *testing* sebesar 60:40 atau 341 data *training* dan 228 data *testing*. Pada pengujian model ini terdapat 2 skenario pengujian utama yaitu *default* dan adanya penyetelan *hyperparameter* menggunakan *GridSeacrhCV*.

4.2.3.1. Euclidean *Default*

Pada skenario pengujian pertama yakni secara *default* atau tanpa adanya penyetelan *hyperparameter* dengan menggunakan nilai n_neighbors = 5 serta menggunakan *euclidean distance* sebagai metode perhitungan jarak maka didapatkan hasil sebagai berikut.

Tabel 4. 20 Hasil prediksi euclidean default model C

radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	aktual	prediksi
0.387825	0.327021	0.370196	0.252598	0.157498	M	B
0.438758	0.394657	0.441361	0.315297	0.491525	M	M
0.465694	0.500845	0.47125	0.325947	0.441966	M	M
⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.292326	0.318566	0.295189	0.173881	0.278125	B	B
0.348646	0.168752	0.332195	0.219145	0.220141	B	B
0.295754	0.436253	0.324651	0.17517	0.501437	M	M

Tabel 4. 21 Confusion matrix euclidean default model C

Data aktual	Prediksi	
	B	M
B	143	0
M	8	77

Tabel 4.21. menunjukkan bahwa model *K-Nearest Neighbor* memprediksi terdapat 143 jenis kanker jinak dan hasil sebenarnya terdeteksi kanker jinak, model *K-Nearest Neighbor* memprediksi terdapat 8 jenis kanker jinak akan tetapi hasil sebenarnya terdeteksi kanker ganas, model *K-Nearest Neighbor* memprediksi terdapat 0 jenis kanker ganas akan tetapi hasil sebenarnya terdeteksi kanker jinak, model *K-Nearest Neighbor* memprediksi terdapat 77 jenis kanker ganas dan hasil sebenarnya terdeteksi kanker ganas.

Dari tabel 4.21. dapat diketahui nilai dari *accuracy, precision, recall dan f-measure*. Perhitungan dari nilai-nilai tersebut adalah sebagai berikut.

$$Akurasi = \frac{(143 + 77)}{(143 + 77 + 8 + 0)} = 96.4\%$$

$$Precision = \frac{143}{(143 + 8)} = 94.7\%$$

$$Recall = \frac{143}{(143 + 0)} = 100\%$$

$$F - Measure = \frac{2 \text{ x } 0.947 \text{ x } 1}{0.947 + 1} = 97.2\%$$

4.2.3.2. Manhattan *Default*

Pada skenario pengujian kedua yakni secara *default* atau tanpa adanya penyetelan *hyperparameter* dengan menggunakan nilai n_neighbors = 2 serta menggunakan *manhattan distance* sebagai metode perhitungan jarak maka didapatkan hasil sebagai berikut.

Tabel 4. 22 Hasil prediksi manhattan default model C

radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	aktual	prediksi
0.387825	0.327021	0.370196	0.252598	0.157498	M	M
0.438758	0.394657	0.441361	0.315297	0.491525	M	M
0.465694	0.500845	0.47125	0.325947	0.441966	M	M
⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.292326	0.318566	0.295189	0.173881	0.278125	B	B
0.348646	0.168752	0.332195	0.219145	0.220141	B	B
0.295754	0.436253	0.324651	0.17517	0.501437	M	M

Tabel 4. 23 Confusion matrix manhattan default model C

Data aktual	Prediksi	
	B	M
B	143	0
M	7	78

Tabel 4.23. menunjukkan bahwa model *K-Nearest Neighbor* memprediksi terdapat 143 jenis kanker jinak dan hasil sebenarnya terdeteksi kanker jinak, model *K-Nearest Neighbor* memprediksi terdapat 7 jenis kanker jinak akan tetapi hasil sebenarnya terdeteksi kanker ganas, model *K-Nearest Neighbor* memprediksi

terdapat 0 jenis kanker ganas akan tetapi hasil sebenarnya terdeteksi kanker jinak, model *K-Nearest Neighbor* memprediksi terdapat 78 jenis kanker ganas dan hasil sebenarnya terdeteksi kanker ganas.

Dari tabel 4.23. dapat diketahui nilai dari *accuracy, precision, recall dan f-measure*. Perhitungan dari nilai-nilai tersebut adalah sebagai berikut.

$$Akurasi = \frac{(143 + 78)}{(143 + 78 + 7 + 0)} = 96.9\%$$

$$Precision = \frac{143}{(143 + 7)} = 95.3\%$$

$$Recall = \frac{143}{(143 + 0)} = 100\%$$

$$F - Measure = \frac{2 \text{ x } 0.953 \text{ x } 1}{0.953 + 1} = 97.6\%$$

4.2.3.3. Euclidean Penyetelan *Hyperparameter*

Pada skenario pengujian ketiga yakni dengan adanya penyetelan *hyperparameter*. Pengujian ini dilakukan dengan menggunakan *GridSearchCV* untuk mencari konfigurasi nilai k atau n_neighbor terbaik.

Tabel 4. 24 Hasil rata-rata skor pengujian model C

mean_test_score	std_test_score	params
0.967731	0.020602	{'n_neighbors': 1, 'weights': 'uniform'}
0.967731	0.020602	{'n_neighbors': 1, 'weights': 'distance'}
0.967815	0.027591	{'n_neighbors': 2, 'weights': 'uniform'}
0.967731	0.020602	{'n_neighbors': 2, 'weights': 'distance'}
0.970672	0.026308	{'n_neighbors': 3, 'weights': 'uniform'}
0.970672	0.026308	{'n_neighbors': 3, 'weights': 'distance'}
0.967731	0.027757	{'n_neighbors': 4, 'weights': 'uniform'}
0.970672	0.026308	{'n_neighbors': 4, 'weights': 'distance'}
0.967731	0.027757	{'n_neighbors': 5, 'weights': 'uniform'}
0.967731	0.027757	{'n_neighbors': 5, 'weights': 'distance'}

mean_test_score	std_test_score	params
0.967731	0.030716	{'n_neighbors': 6, 'weights': 'uniform'}
0.970672	0.026308	{'n_neighbors': 6, 'weights': 'distance'}
0.970672	0.029413	{'n_neighbors': 7, 'weights': 'uniform'}
0.970672	0.029413	{'n_neighbors': 7, 'weights': 'distance'}
0.964874	0.028684	{'n_neighbors': 8, 'weights': 'uniform'}
0.967731	0.027757	{'n_neighbors': 8, 'weights': 'distance'}
0.967815	0.030566	{'n_neighbors': 9, 'weights': 'uniform'}
0.970672	0.029413	{'n_neighbors': 9, 'weights': 'distance'}
0.961933	0.034829	{'n_neighbors': 10, 'weights': 'uniform'}
0.964874	0.028684	{'n_neighbors': 10, 'weights': 'distance'}

Pada tabel 4.24. menunjukkan bahwa nilai k terbaik yang digunakan pada model K-Nearest Neighbor adalah k = 3 dengan skor rata-rata data pengujian (validasi silang) untuk setiap kombinasi *hyperparameter* adalah = 0.97. Sehingga pada skenario pengujian ketiga ini menggunakan nilai n_neighbor = 3 serta menggunakan *euclidean distance* maka didapatkan hasil sebagai berikut.

Tabel 4. 25 Hasil prediksi euclidean hyperparameter model C

radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	aktual	prediksi
0.22768	0.57998	0.236052	0.131839	0.405293	M	M
0.221313	0.281028	0.215841	0.123679	0.295371	B	B
0.334933	0.325668	0.329277	0.210126	0.211815	B	B
⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.292326	0.318566	0.295189	0.173881	0.278125	B	B
0.348646	0.168752	0.332195	0.219145	0.220141	B	B
0.295754	0.436253	0.324651	0.17517	0.501437	M	M

Tabel 4. 26 Confusion matrix euclidean hyperparameter model C

Data aktual	Prediksi	
	B	M
B	143	0
M	7	78

Tabel 4.26. menunjukkan bahwa model *K-Nearest Neighbor* memprediksi terdapat 143 jenis kanker jinak dan hasil sebenarnya terdeteksi kanker jinak, model *K-Nearest Neighbor* memprediksi terdapat 7 jenis kanker jinak akan tetapi hasil sebenarnya terdeteksi kanker ganas, model *K-Nearest Neighbor* memprediksi terdapat 0 jenis kanker ganas akan tetapi hasil sebenarnya terdeteksi kanker jinak, model *K-Nearest Neighbor* memprediksi terdapat 78 jenis kanker ganas dan hasil sebenarnya terdeteksi kanker ganas.

Dari tabel 4.26. dapat diketahui nilai dari *accuracy, precision, recall dan f-measure*. Perhitungan dari nilai-nilai tersebut adalah sebagai berikut.

$$Akurasi = \frac{(143 + 78)}{(143 + 78 + 7 + 0)} = 96.9\%$$

$$Precision = \frac{143}{(143 + 7)} = 95.3\%$$

$$Recall = \frac{143}{(143 + 0)} = 100\%$$

$$F - Measure = \frac{2 \text{ x } 0.953 \text{ x } 1}{0.953 + 1} = 97.6\%$$

4.2.3.4. Manhattan Penyetelan *Hyperparameter*

Pada tabel 4.24. menunjukkan bahwa nilai k terbaik yang digunakan pada model *K-Nearest Neighbor* adalah k = 3 dengan skor rata-rata data pengujian (validasi silang) untuk setiap kombinasi *hyperparameter* adalah = 0.97. Sehingga pada skenario pengujian keempat ini menggunakan nilai n_neighbor = 3 serta menggunakan *manhattan distance* maka didapatkan hasil sebagai berikut.

Tabel 4. 27 Hasil prediksi manhattan hyperparameter model C

radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	aktual	Prediksi
0.22768	0.57998	0.236052	0.131839	0.405293	M	B
0.221313	0.281028	0.215841	0.123679	0.295371	B	B
0.334933	0.325668	0.329277	0.210126	0.211815	B	B
⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.292326	0.318566	0.295189	0.173881	0.278125	B	B
0.348646	0.168752	0.332195	0.219145	0.220141	B	B
0.295754	0.436253	0.324651	0.17517	0.501437	M	M

Tabel 4. 28 Confusion matrix manhattan hyperparameter model C

Data aktual	Prediksi	
	B	M
B	143	0
M	8	77

Tabel 4.28. menunjukkan bahwa model *K-Nearest Neighbor* memprediksi terdapat 143 jenis kanker jinak dan hasil sebenarnya terdeteksi kanker jinak, model *K-Nearest Neighbor* memprediksi terdapat 8 jenis kanker jinak akan tetapi hasil sebenarnya terdeteksi kanker ganas, model *K-Nearest Neighbor* memprediksi terdapat 0 jenis kanker ganas akan tetapi hasil sebenarnya terdeteksi kanker jinak, model *K-Nearest Neighbor* memprediksi terdapat 77 jenis kanker ganas dan hasil sebenarnya terdeteksi kanker ganas.

Dari tabel 4.28. dapat diketahui nilai dari *accuracy, precision, recall dan f-measure*. Perhitungan dari nilai-nilai tersebut adalah sebagai berikut.

$$Akurasi = \frac{(143 + 77)}{(143 + 77 + 8 + 0)} = 96.4\%$$

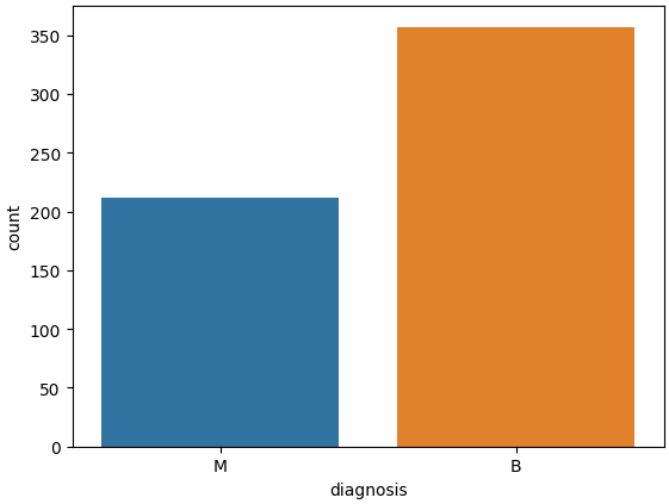
$$Precision = \frac{143}{(143 + 8)} = 94.7\%$$

$$Recall = \frac{143}{(143 + 0)} = 100\%$$

$$F - Measure = \frac{2 \text{ x } 0.947 \text{ x } 1}{0.947 + 1} = 97.2\%$$

4.3. Pembahasan

Pada sub-bab ini merupakan penjelasan dari hasil uji coba dari setiap model yang ada. Penelitian ini menggunakan dataset *Wisconsin Diagnostic Breast Cancer* (WDBC) dari *University of Wisconsin Hospital* yang diperoleh dari repositori *UCI Machine Learning Repository*. Dataset terdiri dari 569 data (357 untuk kanker jinak dan 212 untuk kanker ganas) dengan 10 atribut utama yang setiap atribut memiliki 3 indikator yakni *mean, standard* dan *error/se*.



Gambar 4. 1 Perbandingan diagnosis

Jumlah keseluruhan atribut yang ada pada dataset sebenarnya berjumlah 33 atribut. Namun, dalam proses *preprocessing data* 2 atribut yakni 'Unnamed: 32' dan 'id' tidak dipilih dalam proses seleksi fitur karena 2 atribut tersebut tidak relevan dan tidak dibutuhkan. Sehingga, terdapat 30 atribut yang berfungsi sebagai variabel independent dan 1 atribut yang berfungsi sebagai variabel dependen yakni

diagnosis. Atribut diagnosis memiliki 2 kelas dengan nilai B sebagai kanker jinak dan M sebagai kanker ganas.

Gambar 4.1. menunjukkan perbandingan data dari atribut diagnosis. Sebanyak 212 data yang terdiagnosis kanker ganas dan 357 data terdiagnosis kanker jenis jinak. Gambar di atas juga menunjukkan ketidakseimbangan antar jumlah data dengan perbandingan kanker ganas dan juga kanker jinak sebesar 37% : 63%.

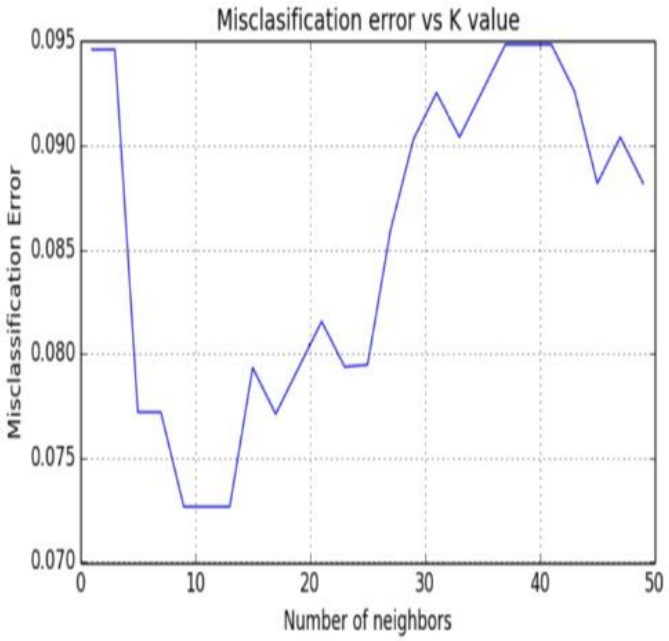
No	Type of Disease	Amount of Test Data	Accuracy		Running time (second)	
			K-NN	Naïve Bayes	K-NN	Naïve Bayes
1	Benign	280	98%	95%	0.0344595909	0.0070559978
2	Malignant	170	92%	91%		
Amount		**310**	**95%**	**93%**		

Gambar 4. 2 Hasil akurasi dari Handayani dan Ikrimach

Penelitian yang di lakukan oleh Handayani dan Ikrimach pada tahun 2020 mempunyai target penelitian yakni menghasilkan algoritma terbaik dalam mengklasifikasikan kanker payudara. Dalam penelitian ini peneliti menggunakan 2 algoritma yakni *K-Nearest Neighbor* dan *Naïve Bayes*. Kedua algoritma tersebut digunakan sebagai klasifikasi kanker payudara pada dataset *Wisconsin Diagnostic Breast Cancer.* Dari hasil penelitian dengan menggunakan perbandingan data *training* dan data *testing* sebesar 80:20 didapatkan hasil akurasi untuk algoritma *K-Nearest Neighbor* sebesar 95.79% dan untuk algoritma *Naïve Bayes* sebesar 93.39% (Handayani & Ikrimach, 2020). Hasil dari detail penelitian dapat dilihat pada gambar 4.2.

Penelitian yang dilakukan oleh Assegie pada tahun 2020 bertujuan untuk mengetahui bahwa *hyperparameter* mempunyai pengaruh yang signifikan dalam

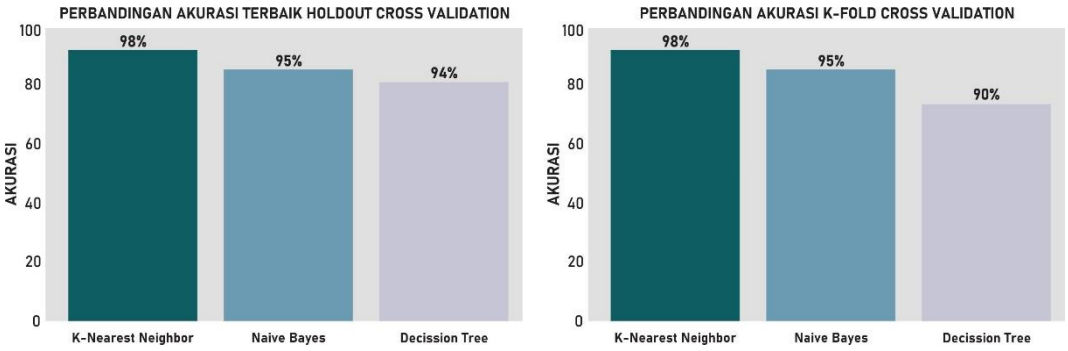
model algoritma *K-Nearest Neighbor* dalam melakukan klasifikasi kanker payudara. Dataset yang digunakan pada penelitian ini menggunakan *Wisconsin Diagnostic Breast Cancer.*



Gambar 4. 3 Hasil *missclasification error* dari Assegie

Peneliti menggunakan *gridsearch* untuk mencari nilai k terbaik dalam melakukan klasifikasi kanker payudara. Peneliti mengungkapkan bahwa *gridsearch* sangat membantu dalam meningkatkan nilai akurasi pada algoritma *K-Nearest Neighbor.* Terbukti dengan adanya penyetelan *hyperparameter* nilai akurasi yang didapat sebesar 94.35% sedangkan nilai akurasi tanpa adanya penyetelan *hyperparameter* hanya sebesar 90.10% (Assegie, 2020). Berikut merupakan hasil penelitian dari *gridsearch* dalam mencari nilai k terbaik.

Penelitian yang dilakukan oleh Jabbar dkk pada tahun 2022 menjelaskan tentang komparasi algoritma *Decision Tree, Naïve Bayes* dan *K-Nearest Neighbor* dalam melakukan klasifikasi kanker payudara. Dataset yang digunakan pada penelitian ini adalah *Breast Cancer Wisconsin (Diagnostic)* yang diperoleh dari *University of California.* Pada penelitian ini peneliti menggunakan 2 metode *cross-validation* yakni *Hold-Out* dan *K-Fold*. Dari hasil penelitian ini diketahui bahwa algoritma *K-Nearest Neighbor* menghasilkan performa akurasi yang lebih baik dibandingkan algoritma *Decision Tree* dan *Naïve Bayes* dengan 98% pada *Hold-Out* dan 96% pada *K-Fold* (Jabbar et al., 2022). Hasil dari detail penelitian dapat dilihat pada gambar berikut.



Gambar 4. 4 Hasil perbandingan akurasi Jabbar dkk

Pada penelitian ini menggunakan algoritma *K-Nearest Neighbor Classifier* untuk klasifikasi penyakit kanker payudara pada *dataset Wisconsin Diagnostic Breast Cancer* (WDBC) dengan penyetelan *hyperparameter* menggunakan *GridSearchCV* untuk meningkatkan nilai akurasi. Proses pertama kali yang dilakukan yakni *preprocessing data*. *Preprocessing data* merupakan proses mengubah data mentah menjadi data siap pakai untuk proses klasifikasi dengan melalui beberapa tahapan (Kohsasih & Situmorang, 2022). Tahapan pertama yakni

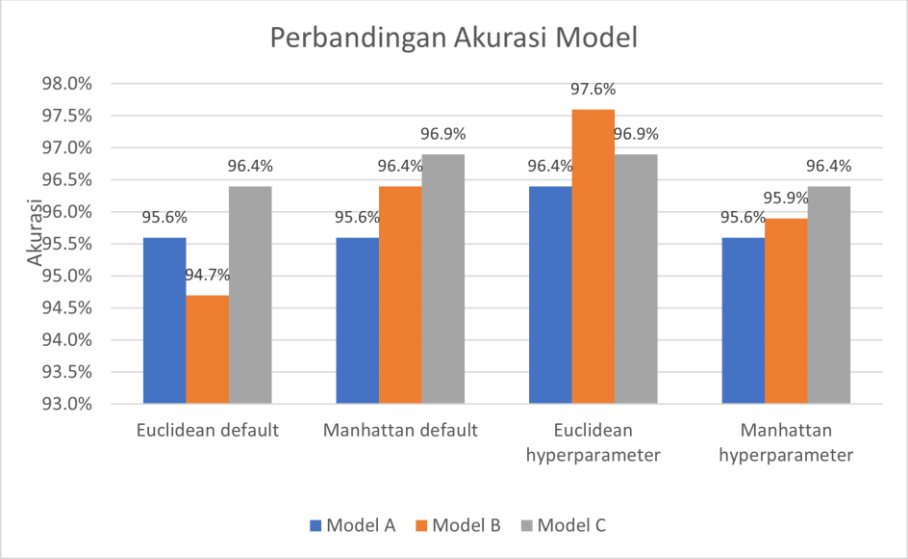
proses seleksi fitur dimana pada proses ini terdapat 2 atribut yang tidak dipilih yakni atribut 'Unnamed: 32' dan 'id'. Kedua, yakni dengan proses *encoder* dengan mengubah tipe data pada atribut diagnosis dari data katerogikal menjadi data numerik.

Setelah proses *encoder,* langkah selanjutnya adalah membagi data *training* dan data *testing* ke dalam 4 model yang telah ditentukan. Model A dengan perbandingan 80%:20%, model B dengan perbandingan 70%:30%, model C dengan perbandingan 60%:40%. Tahapan selanjutnya adalah *standrization* data. Maksud dari standarisasi pada tahap ini adalah menyamakan semua variabel yang ada pada dataset dengan rentang nilai yang sama dengan interval 0,1 yang semula semua variabel memiliki nilai interval yang berbeda satu sama lain. Langkah selanjutnya adalah membandingkan model *K-Nearest Neighbor* tanpa adanya penyetelan *hyperparameter* dengan model *K-Nearest Neighbor* menggunakan penyetelan *hyperparameter* dengan *GridSearchCV*. Perbandingan ini juga membandingkan antara *Euclidean distance* dengan *Manhattan distance* dalam melakukan perhitungan jarak. Berikut tabel hasil akurasi dari tiap model.

Tabel 4. 29 Hasil akurasi tiap model

Model	Banyaknya data = 568				Akurasi			
	Data t*raining*		Data *testing*		*Default*		*Hyperparameter*	
	Presentase	Jumlah	Presentase	Jumlah	Euc.	Manh.	Euc.	Manh.
A	80%	455	20%	114	95.6%	95.6%	96.4%	95.6%
B	70%	398	30%	171	94.7%	96.4%	97.6%	95.9%
C	60%	341	40%	228	96.4%	96.9%	96.9%	96.4%

Pada tabel 4.29. menunjukkan bahwasannya algoritma *K-Nearest Neighbor* memberikan performa yang baik dalam melakukan klasifikasi kanker payudara pada *dataset Wisconsin Diagnostic Breast Cancer*. Hasil akurasi terbaik diperoleh dari model pengujian B dengan penyetelan *hyperparameter* serta menggunakan *Euclidean distance* sebagai mengukur jarak. Model B menggunakan perbandingan data *training* 70% atau 398 data dengan data *testing* 30% atau 171 data dan mendapatkan nilai akurasi terbaik sebesar 97.6%.

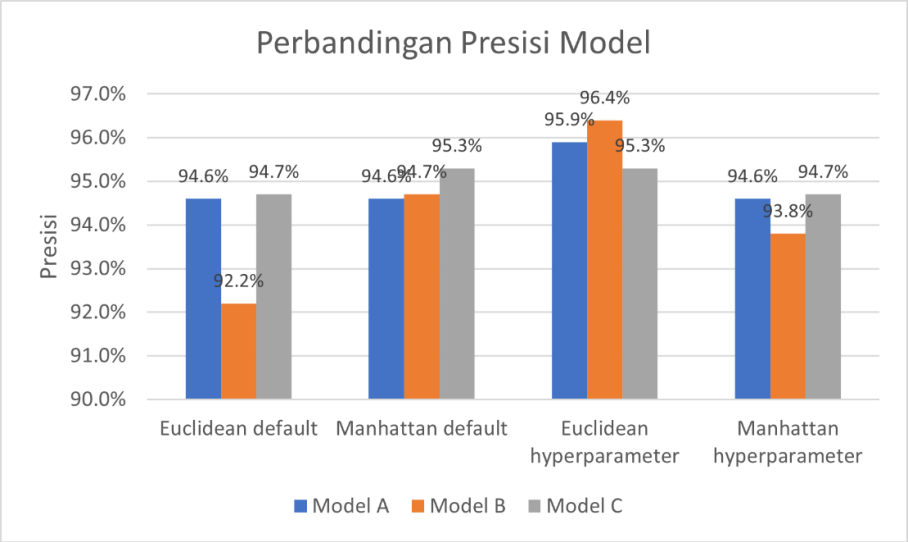


Gambar 4. 5 Perbandingan akurasi model

Tabel 4.29. juga menunjukkan adanya pengaruh dari penyetelan *hyperparameter* menggunakan *gridsearchcv*, ditunjukkan dengan adanya peningkatan nilai akurasi sebesar 3.06% lebih baik jika dibandingkan dengan algoritma *K-Nearset Neighbor* yang diatur secara *default.* Meskipun begitu, terdapat penurunan nilai akurasi yang terjadi pada model C bagian *Euclidean distance*, hal ini bisa saja terjadi karena beberapa akibat. Salah satunya adalah pembagian komposisi antara data *training* dengan data *testing*. Pembagian

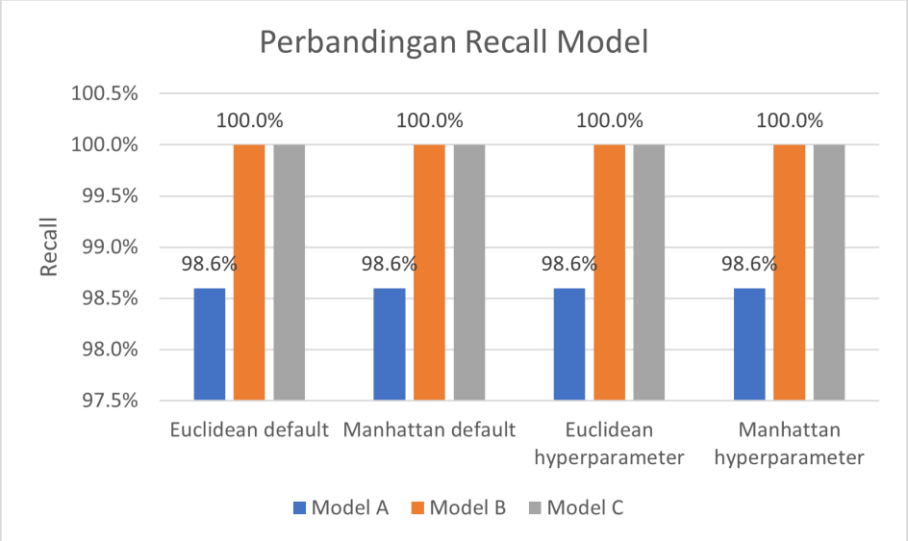
komposisi ini sangat penting karena dapat mempengaruhi nilai akurasi yang diperoleh (Musu et al., 2021). Dengan adanya pembagian komposisi ini kita dapat melihat sejauh mana model klasifikasi dapat memperkirakan hasil berdasarkan data yang sebelumnya tidak terlihat.

Selain nilai akurasi yang didapatkan dari proses evaluasi menggunakan *confusion matrix*, dari *confusion matrix* dapat diketahui juga beberapa parameter pengukuran kinerja model seperti presisi, *recall* dan *F-measure*. Presisi merupakan perbandingan prediksi *True Positives* (TP) dengan total nilai positif. Evaluasi keakuratan model dalam mendeteksi data positif difasilitasi oleh presisi. Dengan meningkatnya nilai presisi, jumlah kesalahan dalam mengidentifikasi data positif semakin berkurang. Dalam konteks *Confusion Matrix*, presisi memberikan pemahaman tentang keakuratan model dalam mengidentifikasi data positif secara akurat. Berikut merupakan perbandingan nilai presisi pada model dalam penelitian ini.



Gambar 4. 6 Perbandingan presisi model

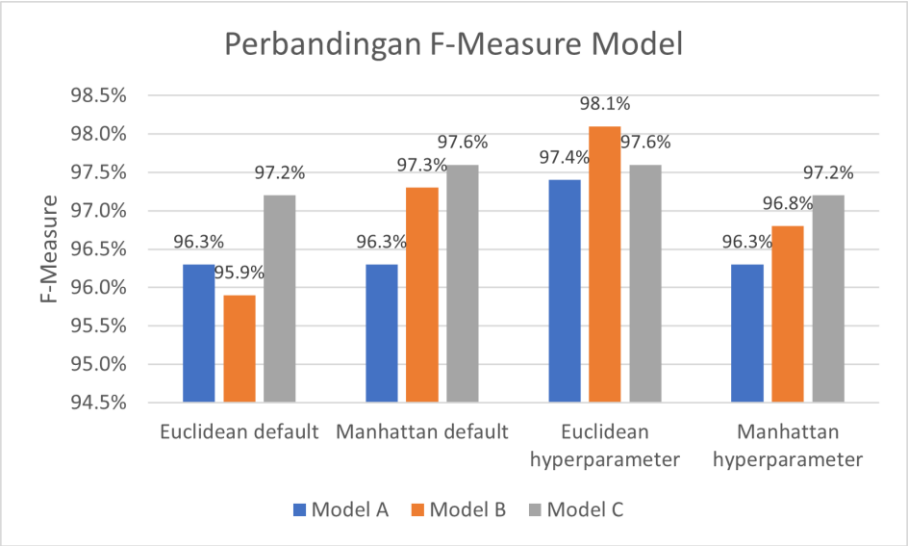
Nilai presisi yang tinggi menunjukkan bahwasannya contoh berlabel positif memang hasil sebenaranya adalah positif (Gelinas et al., 2020). Dari gambar 4.6. dapat diketahui bahwasannya pada semua model yang ada pada penelitian ini menunjukkan nilai presisi yang tinggi. Sehingga semua model bisa mengidentifikasi data positif secara akurat. Model A pada penelitian ini memiliki nilai rata-rata presisi terbaik dari semua model yang ada sebesar 94.9%. Parameter pengukuran kinerja model *confusion matrix* selanjutnya adalah *recall*. *Recall* memberikan pemahaman tentang keakuratan model dalam berupa seberapa sering memprediksi positif ketika kelas aktualnya positif (Saputro & Sari, 2020). Berikut merupakan perbandingan nilai *recall* pada model dalam penelitian ini.



Gambar 4. 7 Perbandingan *recall* model

Apabila presisi fokus pada tingkat kesalahan model yang rendah dengan salah mengklasifikasikan sesuatu sebagai positif padahal seharusnya negatif. Maka, *recall* berfokus pada tingkat kesalahan model yang rendah dengan tidak mengklasifikasikan positif yang seharusnya positif. Dari gambar 4.7. dapat dapat

diketahui bahwasannya pada semua model yang ada pada penelitian ini menunjukkan nilai *recall* yang tinggi. Sehingga semua model menunjukkan semakin banyaknya data positif yang teridentifikasi dengan benar oleh model. Model A dan Model B pada penelitian ini memiliki nilai rata-rata *recall* terbaik dari semua model yang ada sebesar 100%. Parameter pengukuran kinerja model *confusion matrix* selanjutnya adalah *F-Measure*. *F-Measure* merupakan rata-rata harmonik dari tingkat presisi dan perolehan (Baharuddin et al., 2019). *F-Measure* mengintegrasikan nilai presisi dan *recall* menjadi satu nilai, yang mencerminkan kualitas model klasifikasi secara keseluruhan. *F-Measure* Membantu menilai keakuratan model dalam mengidentifikasi data positif dengan benar dan menjaga keseimbangan antara presisi dan perolehan. Pada penelitian ini semua model memberikan nilai *F-Measure* yang baik, sehingga menunjukkan bawasannya model memiliki keseimbangan yang baik untuk menemukan sebagian besar kelas yang tepat (*recall*) dan menghasilkan prediksi yang akurat (presisi). Berikut merupakan perbandingan nilai *F-Measure* pada model dalam penelitian ini.



Gambar 4. 8 Perbandingan *F-Measure* model

Dari penelitian ini didapatkan hasil proses *hyperparameter* yang berbeda pada ketiga model yang telah ditentukan. Model A parameter terpilih adalah `n_neighbors` = 3 dan `weight_options` = "*uniform*". Model B parameter terpilih adalah `n_neighbors` = 6 dan `weight_options` = "*distance*". Model C parameter terpilih adalah `n_neighbors` = 8 dan `weight_options` = "*distance*". Terdapat perbedaan hasil pada tiap model ini dikarenakan perbedaan perbandingan komposisi pada data *training* dengan data *testing*. Hal ini mungkin saja bisa terjadi karena proses *hyperparameter* memerlukan eksperimen dengan berbagai kombinasi nilai parameter untuk menemukan kombinasi yang menawarkan performa tertinggi pada data pengujian. Hal ini juga dapat menghindari pemilihan *hyperparameter* yang berkinerja baik pada data pengujian namun tidak pada data baru dengan memanfaatkan data pengujian terpisah.

Berikut merupakan hasil perbandingan akurasi peneliti dengan peneliti lainnya.

Tabel 4. 30 Hasil perbandingan akurasi

Sitasi	Metode	Akurasi terbaik
(Handayani & Ikrimach, 2020)	*Naïve Bayes* dan *K-Nearest Neighbor* + *K-Fold cross validation*	95.79% *K-Nearest Neighbor*
(Assegie, 2020)	*K-Nearest Neighbor* + *GridSearch*	94.35% *K-Nearest Neighbor*
(Jabbar et al., 2022)	*Decision Tree, Naïve Bayes* dan *K-Nearest Neighbor* + *Hold-Out* dan *K-Fold*	98% *Hold-Out,* 96% *K-Fold* *K-Nearest Neighbor*
Peneliti	***K-Nearest Neighbor* + *GridSearchCV***	**97.6%** ***K-Nearest Neighbor***

Pada tabel 4.30. dapat dilihat bahwa terdapat beberapa teknik yang digunakan oleh peneliti dalam membangun sebuah model klasifikasi pada *dataset*

Wisconsin Diagnostic Breast Cancer. Seperti teknik *cross validation* yaitu *K-Fold* dan *Hold-Out*. *K-Fold* merupakan teknik validasi silang yang membagi data menjadi k (jumlah lipatan) bagian berukuran sama (Azis et al., 2020). Kemudian dilatih dan diuji pada setiap lipatan secara bergantian. Satu lipatan digunakan sebagai data uji pada setiap iterasi, sedangkan lipatan lainnya digunakan sebagai data latih. Metode ini dilakukan sebanyak k kali, dan diperoleh skor kinerja rata-rata dari setiap iterasi. *Hold-Out* merupakan teknik validasi yang paling sederhana dimana data dibagi secara acak menjadi dua bagian yakni data *training* dan data *testing* (Nababan, 2021). Tujuan dari penggunaan *Hold-Out* ini adalah agar hasil evaluasi akhir bisa menjadi lebih baik dengan sesuainya proses pengujian.

Pada tabel 4.29. juga dapat dilihat bahwasannya penggunaan *GridSearchCV* dapat memberikan hasil nilai akurasi yang lebih optimal. Cara kerja *GridSearchCV* yang melalui pendekatan *cross validation* dengan membagi data kedalam beberapa lipatan yang telah di atur kemudian dilatih dan diuji model *K-Nearest Neighbor* dengan kombinasi parameter yang ada menggunakan validasi silang. Setelah proses selesai maka akan menampilkan parameter terbaik untuk membentuk model yang optimal dengan hasil akurasi yang lebih baik, yang mana dalam penelitian ini menghasilkan nilai akurasi sebesar 97.6%. Hasil dari penelitian yang dilakukan oleh Kusuma & Sasongko (2023) juga menunjukkan bahwasannya *GridSearchCV* dapat memberikan parameter yang optimal dengan melalui pendekatan *crossvalidation*. Pada penelitian tersebut juga menunjukkan adanya peningkatan nilai akurasi setelah adanya proses *hyperparameter tuning* menggunakan *GridSearchCV* (Kusuma & Sasongko, 2023). Subarkah dkk. dalam penelitiaannya

menjelaskan bahwa nilai klasifikasi dapat dikelompokkan kedalam beberapa kelompok. Menurut Gorunescu 90%-100% merupakan kategori klasifikasi sangat baik, 80%-90% merupakan kategori klasifikasi baik, 70%-80% merupakan kategori klasifikasi cukup baik, 60%-70% merupakan kategori klasifikasi kurang baik, dan 50%-60% merupakan kategori klasifikasi gagal (Subarkah et al., 2019). Berdasarkan pernyataan tersebut maka keseluruhan model termasuk kategori klasifikasi sangat baik.

4.4. Integrasi Penelitian dalam Tafsir Al-Qur'an

Kanker payudara merupakan kategori penyakit di mana sel-sel jaringan payudara berubah dan membelah tak terkendali yang mengakibatkan munculnya suatu benjolan. Menurut penelitian yang ditulis oleh Sidrah Nadira dkk. menjelaskan bahwa ada beberapa hal yang mengakibatkan keterlambatan dalam penanganan kanker payudara ini, salah satunya adalah ketidaktahuan dokter atau tenaga medis (*doctor delay*) dan keterlambatan *pre-hospital* (*pre-hospital delay*) (Sidrah Nadira et al., 2023). Kanker payudara merupakan kanker dengan jenis kematian tertinggi. Di Indonesia sendiri pada tahun 2020 setidaknya terdapat 22 ribu jiwa kasus kematian dari total keseluruhan 369.914 kasus. Angka kematian yang tinggi ini dapat ditekan dengan cara mendeteksi kanker payudara sejak dini. Namun, tidak semua dokter ahli bisa dengat cepat membedakan antara kanker jinak maupun ganas dan klasifikasi sel kanker memakan waktu hingga 2 hari.

Kanker jinak dan kanker ganas merupakan suatu kondisi yang berbeda dalam dunia medis. Seseorang bisa saja terkena kanker namun untuk menentukan kondisi tersebut termasuk kanker jinak atau kanker ganas dapat dengan melihat

kembali ciri-ciri yang muncul. Hal ini sesuai dengan firman Allah SWT pada QS. Al-Qamar ayat 49 yang berbunyi:

إِنَّا كُلَّ شَيْءٍ خَلَقْنَٰهُ بِقَدَرٍ

"*Sesungguhnya Kami menciptakan segala sesuatu menurut ukuran.*" (QS. Al-Qamar: 49)

Berdasarkan tafsir yang diterbitkan oleh Lajnah Pentashihan Mushaf Al-Qur'an (2016) QS. Al-Qamar: 49 menjelaskan bahwasannya Allah Subhanahu Wa Ta'ala telah menciptakan segala sesuatu menurut ukuran, yakni suatu sistem dan ketentuan yang telah ditetapkan (Al-Qur'an, 2016). Dalam dunia medis, seseorang bisa dikatakan kanker jinak ataupun ganas bergantung pada sejumlah faktor yang telah diperhitungkan dan ditetapkan sebelumnya. Dalam konteksnya sebelum adanya sistem yang mempermudah para dokter untuk melakukan klasifikasi penyakit kanker payudara, para dokter lebih dulu mengalami kesulitan dalam melakukan klasifikasi kanker payudara. Sehingga hal tersebut berimbas dalam penanganan kanker payudara. Dari permasalahan ini banyak penelitian yang telah dilakukan untuk meningkatkan kualitas dalam dunia kesehatan khususnya dalam proses klasifikasi kanker payudara sehingga dapat dilakukan deteksi dengan cepat dan penanganan sejak dini. Hal ini menekankan bahwa perubahan positif dimulai dari usaha dan tindakan manusia itu sendiri. Sesuai dengan potongan QS. Ar-Ra'd ayat 11 yang berbunyi:

...إِنَّ اللهَ لَا يُغَيِّرُ مَا بِقَوْمٍ حَتَّى يُغَيِّرُوا مَا بِأَنْفُسِهِمْ...

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

Menurut tafsir yang diterbitkan oleh Lajnah Pentashihan Mushaf Al-Qur'an (2016) QS. Ar-Ra'd: 11 menjelaskan bahwa Allah Subhanahu Wa Ta'ala Yang Mahakuasa tidak akan mengubah keadaan suatu kaum dari suatu kondisi ke kondisi yang lain, sebelum mereka mengubah keadaan diri menyangkut sikap mental dan pemikiran mereka sendiri (Al-Qur'an, 2016). Harapan besar dengan adanya sistem klasifikasi kanker payudara yang lebih cepat dan efisien sehingga dapat membantu tenaga kesehatan dalam mendeteksi kanker payudara, sehingga bisa diberikannya penanganan sejak dini dan dapat menekan angka kematian yang diakibatkan oleh kanker payudara. Rasulullah Shallallahu 'Alaihi Wa Sallam dalam sabdanya menjelaskan bahwasannya semua penyakit pasti diciptakan pula obatnya.

Rasulullah Shallallahu 'Alaihi Wa Sallam bersabda:

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

"Telah menceritakan kepada kami Harun bin Ma'ruf dan Abu Ath Thahir serta Ahmad bin 'Isa mereka berkata; Telah menceritakan kepada kami Ibnu Wahb; Telah mengabarkan kepadaku 'Amru, yaitu Ibnu al-Harits dari 'Abdu Rabbih bin Sa'id dari Abu Az Zubair dari Jabir dari Rasulullah shallallahu 'alaihi wasallam, beliau bersabda: "Setiap penyakit ada obatnya. Apabila ditemukan obat yang tepat untuk suatu penyakit, akan sembuhlah penyakit itu dengan izin Allah 'azza wajalla." (HR Muslim).

Menurut Ibnu Qayyim al-Jauziyyah dalam kitabnya yang berjudul Ath-Thibb an-Nabawi hadis di atas mengandung pengabsahan terhadap adanya sebab musahab. Menurutnya setiap penyakit pasti ada lawannya, yaitu pengobatan yang menjadi kebalikan dari penyakitnya, karena segala sesuatu yang diciptakan Allah pasti ada lawannya (Hafil, 2020). Dengan adanya sistem klasifikasi ini merupakan

suatu bentuk pencegahan dini dimana merupakan suatu bentuk langkah pengobatan awal sehingga tidak terjadinya komplikasi penyakit yang lebih parah.

BAB V

KESIMPULAN DAN SARAN

5.5. Kesimpulan

Berdasarkan skenario uji coba yang dilakukan pada penelitian ini dimana percobaan dilakukan menggunakan 3 model yakni model A dengan perbandingan 80% *training* data : 20% *testing* data, Model B dengan perbandingan 70% *training* data : 30% *testing* data, dan Model C dengan perbandingan 60% *training* data : 40% *testing* data. Model B menghasilkan nilai akurasi yang terbaik setelah adanya proses penyetelan *hyperparameter* menggunakan *GridSearchCV* dalam melakukan klasifikasi kanker payudara pada dataset *Wisconsin Diagnostic Breast Cancer* (WDBC) dan menggunakan *Euclidean distance* sebagai metode perhitungan jarak. Nilai akurasi pada model B sebelum adanya penyetelan *hyperparameter* sebesar 94.7%. Namun, ketika model B dioptimalkan dengan adanya penyetelan *hyperparameter* nilai akurasi yang didapatkan sebesar 97.6%. Berdasarkan skenario uji coba yang dilakukan pada penelitian ini juga dapat diketahui bahwasannya metode perhitungan jarak *Euclidean distance* memberikan performa yang lebih baik ketika adanya penyetelan *hyperparameter* dibandingkan dengan metode perhitungan jarak *Manhattan distance*. Dapat disimpulkan bahwa penyetelan *hyperparameter* bisa memberikan parameter yang optimal untuk model klasifikasi *K-Nearest Neighbor* menggunakan *GridSearchCV* sehingga pada penelitian ini dapat meningkatkan nilai akurasi dalam proses klasifikasi kanker payudara pada dataset *Wisconsin Diagnostic Breast Cancer* (WDBC). Tentu perlu adanya penelitian lebih lanjut tentang penggunaan *GridSearchCV* untuk optimasi

hyperparameter untuk model klasifikasi khususnya pada *K-Nearest Neighbor* sehingga dapat meningkatkan nilai akurasi.

5.6. Saran

Peneliti menyadari adanya beberapa kekurangan dalam penelitian ini dan menyadari perlunya kritik dan saran yang membangun untuk meningkatkan upaya penelitian selanjutnya. Perbaikan yang disarankan meliputi:

1. Dapat menjelajahi metode klasifikasi yang berbeda sehingga dapat membandingkan nilai akurasi yang didapatkan agar dapat menemukan metode klasifikasi yang terbaik.
2. Dapat menjelajahi teknik *standrization* yang berbeda dalam proses normalisasi data. Seperti teknik *Z-Score* dan *Decimal scaling*.
3. Dapat menjelajahi metode perhitungan jarak yang berbeda pada *K-Nearest Neighbor*. Seperti metode *Chebyshev distance* dan *Minkowski distance*.
4. Dapat menjelajahi berbagai metode *hyperparameter* yang berbeda untuk *K-Nearest Neighbor*.

DAFTAR PUSTAKA

Ailiyya, S. (2020). Analisis Sentimen Berbasis Aspek pada Ulasan Aplikasi Tokopedia menggunakan Support Vector Machine [UIN Syarif Hidayatullah Jakarta]. In *Institutional Repositroy UIN Syarif Hidayatullah Jakarta*. http://repository.radenintan.ac.id/11375/1

Al-Qur'an, L. P. M. (2016). *Tafsir Ringkas*. Lajnah Pentashihan Mushaf Al-Qur'an.

Angkasa, V., & Junifer Pangaribuan, J. (2022). Komparasi Tingkat Akurasi Random Forest Dan Knn Untuk Mendiagnosis Penyakit Kanker Payudara. *Information System Development*, *7*(1), 49–61.

Arifin, Z., Jafar Shudiq, W., & Maghfiroh, S. (2019). Penerapan Metode Knn (K-Nearest Neighbor) Dalam Sistem Pendukung Keputusan Penerimaan Kip (Kartu Indonesia Pintar) Di Desa Pandean Berbasis Web Dan Mysql. *NJCA (Nusantara Journal of Computers and Its Applications)*, *4*(1), 27–34. https://doi.org/10.36564/njca.v4i1.101

Ashariati, A. (2019). *Manajemen Kanker Payudara Komprehensif*. Airlangga University Press.

Assegie, T. A. (2020). An optimized K-Nearest neighbor based breast cancer detection. *Journal of Robotics and Control (JRC)*, *2*(3), 115–118. https://doi.org/10.18196/jrc.2363

Azis, H., Purnawansyah, Fattah, F., & Putri, I. P. (2020). Performa Klasifikasi K-NN dan Cross Validation Pada Data Pasien Pengidap Penyakit Jantung. *ILKOM Jurnal Ilmiah*, *12*(2), 81–86. https://doi.org/10.33096/ilkom.v12i2.507.81-86

Azmi, A. N., Kurniawan, B., Siswandi, A., & Detty, A. U. (2020). Hubungan Faktor Keturunan Dengan Kanker Payudara DI RSUD Abdoel Moeloek. *Jurnal Ilmiah Kesehatan Sandi Husada*, *9*(2), 702–707. https://doi.org/10.35816/jiskh.v12i2.373

Baharuddin, M. M., Azis, H., & Hasanuddin, T. (2019). Analisis Performa Metode K-Nearest Neighbor Untuk Identifikasi Jenis Kaca. *ILKOM Jurnal Ilmiah*, *11*(3), 269–274. https://doi.org/10.33096/ilkom.v11i3.489.269-274

Daqiqil, I. (2021). *MACHINE LEARNING: Teori, Studi Kasus dan Implementasi Menggunakan Python*. UR PRESS.

Deo, S. V. S., Sharma, J., & Kumar, S. (2022). GLOBOCAN 2020 Report on Global Cancer Burden: Challenges and Opportunities for Surgical Oncologists. *Annals of Surgical Oncology*, *29*(11), 6497–6500.

https://doi.org/10.1245/s10434-022-12151-6

Fauziningrum, E., & Suryaningsih, E. I. (2021). Evaluasi Dan Prediksi Penguasaan Bahasa Inggris Maritim Menggunakan Metode Decision Tree Dan Confusion Matrix (Studi Kasus Di Universitas Maritim Amni). *Angewandte Chemie International Edition, 6(11), 951–952.*, 5–24.

Gelinas, Ulric, Oram, Alan, Wiggins, & William. (2020). Perbandingan Algoritme Klasifikasi Untuk Prediksi Cuaca. *Accounting Information System*, *3*, 17–30.

Grandis, G. F., Arumsari, Y., & Indriati. (2021). Seleksi Fitur Gain Ratio pada Analisis Sentimen Kebijakan Pemerintah Mengenai Pembelajaran Jarak Jauh dengan K-Nearest Neighbor. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *5*(8), 3507–3514.

Hafil, M. (2020). *Ulama Jelaskan Hadist Nabi Soal Setiap Penyakit Ada Obatnya*. Republika. https://islamdigest.republika.co.id/berita/q7ixaj430/ulama-jelaskan-hadist-nabi-soal-setiap-penyakit-ada-obatnya

Handayani, I., & Ikrimach, I. (2020). Accuracy Analysis of K-Nearest Neighbor and Naïve Bayes Algorithm in the Diagnosis of Breast Cancer. *Jurnal Infotel*, *12*(4), 151–159. https://doi.org/10.20895/infotel.v12i4.547

Hashi, E. K., & Md. Shahid Uz Zaman. (2020). Developing a Hyperparameter Tuning Based Machine Learning Approach of Heart Disease Prediction. *Journal of Applied Science & Process Engineering*, *7*(2), 631–647. https://doi.org/10.33736/jaspe.2639.2020

Henderi, Wahyuningsih, T., & Rahwanto, E. (2021). Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer. *IJIIS: International Journal of Informatics and Information Systems*, *4*(1), 13–20. https://doi.org/10.47738/ijiis.v4i1.73

Houssein, E. H., Emam, M. M., Ali, A. A., & Suganthan, P. N. (2021). Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review. *Expert Systems with Applications*, *167*, 114161. https://doi.org/10.1016/j.eswa.2020.114161

Jabbar, M. A., Hasmin, E., Sunardi, Susanto, C., & Musu, W. (2022). Komparasi Algoritma Decision Tree, Naive Bayes, dan K- Nearest Neighbors dalam Klasifikasi Kanker Payudara. *CSRID Journal*, *14*(3), 258–270.

Kafil, M. (2019). Penerapan Metode K-Nearest Neighbors untuk Prediksi Penjualan Berbasis Web pada Boutiq Dealove Bondowoso. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *3*(2), 59–66.

Kanker Payudara Paling Banyak di Indonesia, Kemenkes Targetkan Pemerataan Layanan Kesehatan. (2022). Kementerian Kesehatan Republik Indonesia. https://www.kemkes.go.id/article/view/22020400002/kanker-payudaya-paling-banyak-di-indonesia-kemenkes-targetkan-pemerataan-layanan-kesehatan.html

Kartini, Lamongga Lubis, N., & Moriza, T. (2019). Analisis Faktor Yang Mempengaruhi Keterlambatan Pengobatan Pada Wanita Penderita Kanker Payudara Di Rumah Sakit Umum Daerah Simeulue Tahun 2018. *Jurnal Info Kesehatan*, *17*(1), 16–34.

Kohsasih, K. L., & Situmorang, Z. (2022). Analisis Perbandingan Algoritma C4.5 dan Naïve Bayes Dalam Memprediksi Penyakit Cerebrovascular. *Jurnal Informatika*, *9*(1), 13–17. https://doi.org/10.31294/inf.v9i1.11931

Kusuma, S. T., & Sasongko, T. B. (2023). Optimasi K-Nearest Neighbor dengan Grid Search CV pada Prediksi Kanker Paru-Paru. *Indonesian Journal of Computer Science*, *12*(4), 2162–2171. https://doi.org/10.33022/ijcs.v12i4.3267

Musu, W., Ibrahim, A., & Heriadi. (2021). Pengaruh Komposisi Data Training dan Testing terhadap Akurasi Algoritma C4.5. *Prosiding Seminar Ilmiah Sistem Informasi Dan Teknologi Informasi*, *X*(1), 186–195.

Nababan, J. F. (2021). *Klasifikasi Penderita Stunting Dengan Metode Support Vector Machine (Studi Kasus: Lima Puskesmas Di Kota Bandar Lampung)*.

Nishom, M. (2019). Perbandingan Akurasi Euclidean Distance, Minkowski Distance, dan Manhattan Distance pada Algoritma K-Means Clustering berbasis Chi-Square. *Jurnal Informatika: Jurnal Pengembangan IT (JPIT)*, *4*(1), 20–24. https://doi.org/10.30591/jpit.v4i1.1253

Parannuan, C. (2016). *Pengaruh Obesitas Terhadap Kejadian Kanker Payudara Di Rsup Dr Wahidin Sudiro Husodo Dan Rsud Labuang Baji Makassar*. *4*(1), 1–23.

Pribadi Wahyu, W., Yunus, A., & Sartika Wiguna, A. (2022). Perbandingan Metode K-Means Euclidean Distance Dan Manhattan Distance Pada Penentuan Zonasi Covid-19 Di Kabupaten Malang. *Jurnal Mahasiswa Teknik Informatika (JATI)*, *6*(2), 493–500.

Priya C, B. (2021). *Cross-Validation and Hyperparameter Search in Scikit-Learn - A Complete Guide*. Dev Community. https://dev.to/balapriya/cross-validation-and-hyperparameter-search-in-scikit-learn-a-complete-guide-5ed8

Qiudandra, E., Akram, R., & Novianda. (2022). Sistem Pakar Diagnosa Penyakit Osteoarthritis Dengan Menggunakan Metode K-Nearest Neighbor. *Jurnal Ilmiah Teknik Informatika*, *2*(2), 37–48.

Ramdhani, M. R. (2023). *Analisis Sentimen PPDB DKI Jakarta menggunakan Support Vector Machine*. UIN Syarif Hidayatullah Jakarta.

Saputro, I. W., & Sari, B. W. (2020). Uji Performa Algoritma Naïve Bayes untuk Prediksi Masa Studi Mahasiswa. *Creative Information Technology Journal*, *6*(1), 1. https://doi.org/10.24076/citec.2019v6i1.178

Sharma, S., Aggarwal, A., & Choudhury, T. (2018). Breast Cancer Detection Using Machine Learning Algorithms. *Proceedings of the International Conference on Computational Techniques, Electronics and Mechanical Systems, CTEMS 2018*, *Ml*, 114–118. https://doi.org/10.1109/CTEMS.2018.8769187

Sidrah Nadira, C., Rizka, A., & Humaira, Z. (2023). Faktor Keterlambatan Pada Pasien Kanker Payudara Yang Berobat Di Rsucm Aceh Utara Tahun 2020 - 2021. *Jurnal Ilmiah Manusia Dan Kesehatan*, *6*(1), 88–99. https://doi.org/10.31850/makes.v6i1.1942

Subarkah, P., Marcos, H., Arsi, P., Prediksi, K., & Nasabah, A. (2019). *Perbandingan Algoritme CART dan Naive Bayes Untuk Prediksi Kelayakan Nasabah BMT Khonsa Cilacap*. *November 2019*, 111–116.

Suhartini, Kerta Wijaya, L., & Arini Pratiwi, N. (2020). Penerapan Algoritma K-Means Untuk Pendataan Obat Berdasarkan Laporan Bulanan Pada Dinas Kesehatan Kabupaten Lombok Timur. *Infotek : Jurnal Informatika Dan Teknologi*, *3*(2), 147–156. https://doi.org/10.29408/jit.v3i2.2315

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, *71*(3), 209–249. https://doi.org/10.3322/caac.21660

Yuliany, S., Aradea, & Andi Nur Rachman. (2022). Implementasi Deep Learning pada Sistem Klasifikasi Hama Tanaman Padi Menggunakan Metode Convolutional Neural Network (CNN). *Jurnal Buana Informatika*, *13*(1), 54–65. https://doi.org/10.24002/jbi.v13i1.5022