Terbit online pada laman web jurnal: **http://journal.ittelkom-pwt.ac.id/index.php/dinda**

Journal of Dinda

Kelompok Keahlian Rekayasa Data
Institut Teknologi Telkom Purwokerto

Vol. 2 No. 1 (2022) 11 - 20 ISSN Media Elektronik: 2809-8064

Perbandingan Performa Antara Algoritma Naïve Bayes dan K-Nearest Neighbour Pada Klasifikasi Kanker Payudara

Annisa Nugraheni[1], Rima Dias Ramadhani[2], Amalia Beladinna Arifa[3], Agi Prasetiadi[4]
[1, 3, 4]Program Studi S1 Teknik Informatika, Fakultas Informatika, Institut Teknologi Telkom Purwokerto
[2] Program Studi S1 Sains Data, Fakultas Informatika, Institut Teknologi Telkom Purwokerto
17102075@ittelkom-pwt.ac.id, rima@ittelkom-pwt.ac.id, amalia@ittelkom-pwt.ac.id, agi@ittelkom-pwt.ac.id

Abstract

Breast cancer is the second most common cause of death from cancer after lung cancer is in the first place. Breast cancer occurs when cells in breast tissue begin to grow uncontrollably and can disrupt existing healthy tissue. Therefore, there is a need for a classification to distinguish breast cancer patients and healthy people. Based on previous research, the Naïve Bayes and K-Nearest Neighbor algorithms are considered capable of classifying breast cancer. In the research process using the breast cancer dataset from the Breast Cancer Coimbra dataset in 2018 UCI Machine Learning Repository with a total of 116 data, while for the calculation of the feasibility of the method using the Confusion Matrix (Accuracy, Precision, and Recall) and the ROC-AUC curve. The purpose of this study is to compare the performance of the Naïve Bayes and K-Nearest Neighbor algorithms. In testing using the Naïve Bayes algorithm and the K-Nearest Neighbor algorithm, there are several test scenarios, namely, data testing before and after normalization, model testing based on a comparison of training data and testing data, model testing based on K values in K-Nearest Neighbors, and model testing. based on the selection of the strongest attribute with the Pearson correlation test. The results of this study indicate that the Naïve Bayes algorithm has the highest average accuracy of 69.12%, healthy precision 64.90%, pain precision 83%, healthy recall 88%, sick recall 61.11% and AUC 0.82 which is included in the good classification category. Meanwhile, the highest average results of the K-Nearest Neighbor algorithm are 76.83% for accuracy, 76% healthy precision, 80.21% pain precision, 74.18% for healthy recall, 80.81% sick recall and 0.91 AUC which is included in the excellent classification category.

Keywords: *Breast Cancer, Performance Test, Naïve Bayes, K-Nearest Neighbor, and Confusion Matrix.*

Abstrak

Kanker payudara merupakan penyakit ke-2 terbanyak yang menjadi penyebab kematian akibat kanker setelah kanker paru diurutan pertama. Kanker payudara terjadi saat sel-sel pada jaringan payudara mulai memiliki pertumbuhan yang tidak dapat dikendalikan serta dapat mengganggu jaringan sehat yang ada. Oleh sebab itu, perlu adanya suatu klasifikasi untuk membedakan pasien kanker payudara dan orang sehat. Berdasarkan penelitian sebelumnya, algoritma Naïve Bayes dan K-Nearest Neighbour dinilai mampu untuk melakukan klasifikasi kanker payudara. Dalam proses penelitian menggunakan dataset penyakit kanker payudara dari dataset *Breast Cancer Coimbra* tahun 2018 *UCI Machine Learning Repository* dengan total 116 data, sedangkan untuk perhitungan kelayakan metode menggunakan *Confusion Matrix* (Akurasi, Presisi, dan Recall) dan kurva ROC-AUC. Tujuan dari penelitian ini adalah membandingkan performansi algoritma Naïve Bayes dan K-Nearest Neighbour. Pada pengujian menggunakan algoritma Naïve Bayes dan algoritma K-Nearest Neighbour, terdapat beberapa skenario pengujian yaitu, pengujian data sebelum dan sesudah normalisasi, pengujian model berdasarkan perbandingan data *training* dan data *testing*, pengujian model berdasarkan nilai K pada K-Nearest Neighbour, dan pengujian model berdasarkan pemilihan atribut terkuat dengan uji korelasi Pearson. Hasil dari penelitian ini menunjukkan bahwa algoritma Naïve Bayes memiliki rata-rata akurasi tertinggi 69.12%, presisi sehat 64.90%, presisi sakit 83%, recall sehat 88%, recall sakit 61.11% dan AUC 0.82 yang termasuk kategori *good classification*. Sedangkan untuk hasil

Diterima Redaksi : 27-10-2021 | Selesai Revisi : 11-01-2022 | Diterbitkan Online : 23-02-2022

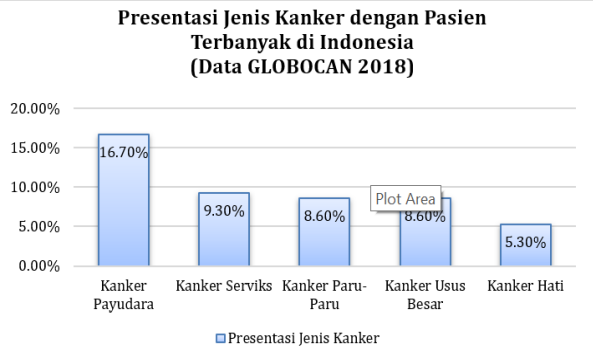
rata-rata tertinggi algoritma K-Nearest Neighbour adalah 76.83% untuk akurasi, presisi sehat 76%, presisi sakit 80.21%, 74.18% untuk recall sehat, recall sakit 80.81% dan AUC 0.91 yang termasuk kategori *excellent classification.*

Kata kunci: Kanker Payudara, Uji Performa, Naïve Bayes, K-Nearest Neighbour, Confusion Matrix.

1. Pendahuluan

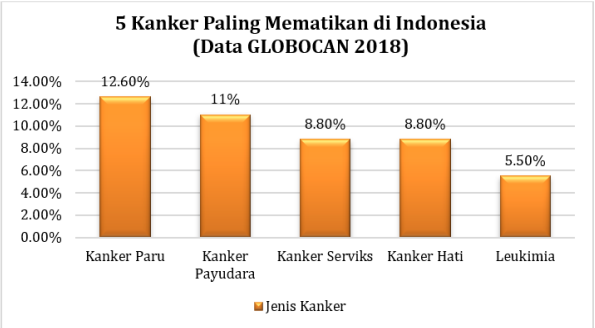
Kanker termasuk jenis penyakit tidak menular yang ditandai dengan adanya suatu sel atau jaringan tidak normal yang bertumbuh terus menerus, bersifat ganas (tidak dapat kendalikan) dan dapat merusak fungsi dari suatu jaringan. Sel kanker dapat terbentuk dari berbagai unsur dalam pembentukan organ yang dapat menjadikan massa tumor akibat sel menggandakan diri, serta dapat menyebar melewati jaringan pembuluh darah maupun pembuluh getah bening [1].

Berdasarkan data *Global Cancer Observatory* (GLOBOCAN) tahun 2018 menjelaskan bahwa ada 18,1 juta kasus kanker baru sedangkan untuk angka kematian berdasarkan penyakit kanker hingga 9,6 juta kasus kematian, serta pertumbuhan kasus kanker mencapai 207.000 kasus per-tahun. Data *Global Cancer Observatory* (GLOBOCAN) tahun 2018 juga menjelaskan peringkat jenis kanker dengan pasien terbanyak, yaitu kanker payudara, kanker serviks, kanker paru-paru, kanker usus besar dan kanker hati yang dapat dilihat pada gambar 1 dibawah ini [2][3]:



Gambar 1. Presentasi Jenis Kanker dengan Pasien Terbanyak [3]

Data *Global Cancer Observatory* (GLOBOCAN) diatas dapat menjelaskan bahwa kanker payudara merupakan jenis kanker yang memiliki pasien terbanyak di Indonesia yaitu sebesar 16,7% atau 58.256 pasien dari 348.809 total kejadian kanker [4]. Kanker payudara dapat dikatakan sebagai tumor ganas yang terdapat pada sel-sel bagian payudara. Kanker payudara terjadi saat sel-sel dalam jaringan payudara mulai memiliki pertumbuhan yang tidak dapat dikendalikan serta dapat mengganggu jaringan sehat yang ada [5][6]. Berdasarkan makalah yang diterbitkan oleh CA berjudul *A Cancer Journal for Clinicians* menyatakan pada tahun 2020 jika kanker payudara pada wanita memiliki 2,3 juta kasus atau 11,7% dari total kasus kanker baru yang didiagnosis, serta memiliki angka resentasi lebih tinggi dibandingkan diagnosis kasus kanker paru yang hanya 11,5%. Hyuna Sung yang merupakan ilmuwan utama serta ahli epidomologi dalam *American Cancer Society* menyatakan bahwa kanker payudara memiliki peringkat tertinggi dalam kasus penyakit kanker dengan 2,3 juta kasus yang meningkat dari pada tahun 2018 yaitu 2.088.849 kasus kanker [7]. Kanker payudara termasuk dalam jenis kanker yang berbahaya, seperti dijelaskan dalam data *Global Cancer Observatory* (GLOBOCAN) tahun 2018 yang menjelaskan 5 kanker dengan jumlah kasus kematian pasien terbanyak di Indonesia, dan kanker payudara menduduki peringkat kedua dengan persentase 11% atau 22.692 kasus kematian dari total 207.210 total data kasus kematian[8]. Gambaran data 5 kanker paling mematikan di Indonesia dapat dilihat pada gambar 2 [8].



Gambar 2. Data Kanker Paling Mematikan di Indonesia

Berdasarkan berbagai survei maupun pendapat ahli yang telah dijelaskan diatas, maka dapat diambil kesimpulan bahwa penyakit kanker payudara termasuk penyakit yang mematikan dan terbanyak terjadi di Indonesia khususnya perempuan. Berdasarkan penelitian sebelumnya, terdapat beberapa metode yang dapat digunakan untuk memprediksi penyakit kanker payudara berdasarkan dataset kanker Coimbra. Penelitian yang dilakukan oleh Miguel Patricio, dkk

tahun 2018 untuk mengembangkan dan mengevaluasi ciri-ciri biologi seseorang mengidap kanker payudara. Variabel yang digunakan pada penelitian ini diantaranya *body mass index* (BMI), *resistin*, *glucose*, dan umur. Penelitian ini menggunakan 3 metode klasifikasi yaitu, *Linear Regression*, *Random Forest* dan *Super Vector Machine* (SVM) serta menggunakan dataset kanker payudara Coimbra sebagai bahan penelitian. Hasil penelitian ini adalah *Linear Regression* memiliki sensitivitas tertinggi dengan nilai 74% hingga 81%, spesifitas tertinggi dengan nilai 81% hingga 89%, *Random Forest* memiliki sensitivitas tertinggi dengan nilai range 85% hingga 90%, spresifitas tertinggi dengan nilai 81% hingga 87%, *Super Vector Machine* (SVM) memiliki nilai sensitifitas tertinggi antara 82% hingga 88%, spesifitas tertinggi memiliki nilai antara 84% hingga 90%, sedangkan nilai AUC ketiga model tersebut antara 0,87 hingga 0,91 [9].

Penelitian kedua yang dilakukan oleh Yixuan Li dan Zixuan Chen tahun 2018 dengan menggunakan dataset kanker payudara Coimbra sebagai data *training* dan dataset kanker payudara wincousin sebagai data *testing* untuk memprediksi kanker payudara termasuk termasuk golongan kanker ganas atau kanker jinak. Dalam penelitian ini menggunakan 5 metode klasifikasi yaitu *Decision Tree*, *Super Vector Machine* (SVM), *Random Forest*, *Linear Regression*, *Neural Network* serta menghasilkan *Random Forest* sebagai metode yang paling sesuai untuk memprediksi kanker payudara dengan hasil akurasi 96,1% dan *F-measure matric* adalah 95,5% sedangkan untuk nilai AUC dataset kanker Coimbra adalah 0,785 dan dataset kanker Wincousin 0,989 [10].

Penelitian selanjutnya yang dilakukan oleh Ilham Mubarog, Arief Setyanto, dan Heri Sismoro tahun 2019 untuk mendiagnosa penyakit kanker payudara berdasarkan dataset penyakit kanker payudara Coimbra dengan metode Naïve Bayes. Penelitian ini menghasilkan akurasi terbaik dengan nilai 80%, nilai presisi 83% dan nilai *recall* 83% [11]. Penelitian lainnya yang dilakukan oleh Nur Ghaniaviyanto Ramadhan dan Faisal Dharma Adhinata membahas tentang teknik imbalanced dan klasifikasi pada permasalahan kanker payudara dengan model klasifikasi yang digunakan yaitu naïve bayes dan random forest dengan presisi yang dihasilkan mencapai 99% [12].

Berdasarkan penelitian yang sudah ada, peneliti mengusulkan perbandingan model klasifikasi Naïve Bayes dan K-Nearest Neighbour pada dataset penyakit kanker payudara untuk mengetahui pasien penyakit kanker atau orang sehat. Untuk penelitian klasifikasi dataset penyakit kanker payudara Coimbra sebelumnya belum banyak penelitian yang membandingkan 2 metode tersebut. Pengertian metode klasifikasi yaitu metode yang dapat melakukan proses pengelompokan objek pada suatu kategori yang sebelumnya sudah ditentukan. Tujuan dari penggunaan metode klasifikasi adalah untuk mengetahui kelas objek yang belum memiliki label [13].

Metode *Naïve Bayes* merupakan cara untuk memprediksi hasil data dengan perhitungan probabilitas sederhana yang akan menghitung beberapa probabilitas melalui penjumlahan frekuensi dengan penggabungan nilai dalam dataset. *Naïve bayes* juga merupakan metode klasifikasi yang menilai kategori mempengaruhi segala atribut dan saling bergantung sedangkan metode *K-Nearest Neighbor* merupakan metode klasifikasi *supervised* yaitu memerlukan data latih sehingga dapat melakukan klasifikasi objek, dengan cara kerjanya melakukan evaluasi dengan nilai terdekat dalam data latih untuk menemukan 2 data dengan jarak terdekat [14].

Metode klasifikasi Naïve Bayes memiliki kelebihan dibandingkan metode klasifikasi lain diantaranya dapat digunakan untuk data bersifat kualitatif maupun kuantitatif, dalam jumlah data dan data pelatihan yang dibutuhkan tidak banyak, dalam perhitungan lebih efisien, mudah untuk dipahami dan membuatnya, dalam klasifikasi dokumen dapat diatur sesuai kebutuhan seseorang, dapat melakukan klasifikasi biner maupun lebih dari satu kelas. Namun terdapat juga kekurangan dari metode *naïve bayes* yaitu saat kondisi probabilitas nilai 0 maka prediksi bernilai 0, untuk mendeteksi kata, saat pengambilan keputusan sangat bergantung dengan pengetahuan yang sudah ada sebelumnya, akurasi dapat berkurang bila adanya asumsi variable yang independent [15].

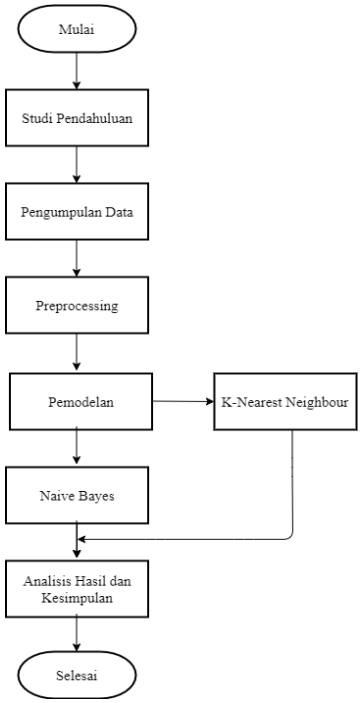
Selain meode Naïve Bayes, metode klasifikasi K-Nearest Neighbour juga memiliki kelebihan dibandingkan metode lain yaitu lebih tangguh terhadap data *training* yang memiliki banyak *noise* serta efektif untuk digunakan pada data *training* yang berukuran besar. Kelemahan dari metode K-Nearest Neighbour adalah perlu penentuan nilai parameter K dan *training* berdasarkan jarak belum jelas mengenai jarak apa yang perlu digunakan [16].

Berdasarkan rumusan masalah di atas, maka pertanyaan yang muncul pada penelitian ini adalah sebagai berikut:

1. Berapa nilai akurasi, presisi, *recall* dan AUC dari algoritma Naïve Bayes dan K-Nearest Neighbour?
2. Algoritma manakah yang lebih baik performanya antara Naïve Bayes atau K-Nearest Neighbour dalam mengklasifikasikan kanker payudara?

2. Metode Penelitian

Dalam penyusunan laporan penelitian ini terdapat beberapa tahapan dalam penelitian dengan diawali melakukan studi pendahuluan, pengumpulan data, preprocessing, pemodelan meggunakan Naïve Bayes dan K-NN, selanjutnya diakhiri dengan analisis hasil dan kesimpulan. Berikut merupakan tahapan yang dilakukan pada proses penelitian yang dapat dilihat pada Gambar 3 dibawah ini:



Gambar 3. Alur Proses Penelitian

2.1 Studi Pendahuluan

Studi pendahuluan merupakan tahapan awal dalam suatu penelitian, yang berfungsi sebagai referensi dasar bagi penulis untuk melakukan penelitian.

2.2 Pengumpulan Data

Pengumpulan data dalam penelitian ini dilakukan dengan mencari data penyakit kanker payudara dari *website UCI Machine Learning Repository* yaitu dataset *Breast Cancer Coimbra* tahun 2018. Dataset tersebut terdiri dari total seratus enam belas data dengan sembian variabel prediktor dan satu variabel kelas seperti Tabel 1.

Tabel 1. Atribut Data

No. Atribut	Atribut	Domain Nilai	Keterangan
1.	Age	Angka numerik	Umur pasien (tahun)
2.	BMI	Angka numerik	*Body Mass Index* (berat badan) pasien (kg/m^2)
3.	Glucose	Angka numerik	Kadar gula dalam tubuh pasien (mg/dL)
4.	Insulin	Angka numerik	Kadar insulin (hormon polipeptida) dalam tubuh pasien (µU/mL)
5.	HOMA	Angka numerik	Pengukuran resistensi insulin dan fungsi sel beta dalam tubuh pasien
6.	Leptin	Angka numerik	Kadar leptin (hormon yang dibuat sel lemak) dalam tubuh pasien (ng/mL)
7.	Adiponectin	Angka numerik	Kadar adiponectin (hormon protein dan adipokin) dalam tubuh pasien(µg/mL)
8.	Resistin	Angka numerik	Kadar resistin (protein kaya asam amino) pada tubuh pasien (ng/mL)
9.	MCP.1	Angka numerik	Kadar MCP.1 pada tubuh (pg/dL)
10.	Classification	Angka numerik	Label yang menunjukkan seseorang sehat atau sakit kanker. 1 : sehat 2 : sakit kanker

2.3 Preprocessing

Preprocessing dalam penelitian ini dilakukan untuk menyiapkan dataset yang akan dianalisis lebih lanjut, antara lain:

a. Analisis Distribusi Data

Analisis distribusi data bertujuan untuk mengetahui keseimbangan jumlah kelas sehat dan sakit kanker. Apabila proporsi jumlah kelas sangat berbeda jauh (tidak seimbang) maka perlu dilakukan penyeimbangan data dengan metode *oversampling* atau *undersampling*.

b. Analisis Atribut *Missing Value*

Analisis *missing value* bertujuan untuk mengetahui atribut yang hilang atau kosong. Penanganan *missing value* bisa menggunakan metode *imputation*.

c. Pemilihan Atribut Yang Akan Digunakan

Pada scenario pengujian pertama atribut yang akan diteliti yaitu sebagai parameter menentukan klasfikasi pasien kanker atau sehat adalah usia, resistin, HOMA, insulin, leptin, adiponectin, dan MCP1, sedangkan untuk pengujian skenario kedua memilih atribut yang memiliki korelasi pearson tertinggi untuk dihapus.

d. Normalisasi Data

Normalisasi data bertujuan untuk mengubah data dalam rentang nilai tertentu. Metode yang digunakan adalah *min-max*, seperti persamaan 1.

$$X_{new} = \frac{Xold - Xmin}{Xmax - Xmin} \quad (1)$$

2.4 Membangun Model

Model klasifikasi dalam penelitian ini menggunakan K-Nearest Neighbour (K-NN) dan Naïve Bayes.

a. Naïve Bayes

Dalam perhitungan data dengan *Naïve Bayes* terdapat beberapa proses, diawali dengan memasukkan dataset penyakit kanker payudara, dilanjutkan dengan membagi data *training* dan data *testing*, menghitung nilai *mean* dari data tersebut, dilanjutkan dengan mencari nilai standar deviasi dan mengurutkan nilai terdekat dari data *training* ke data *testing* sehingga diperoleh hasil. Untuk mengetahui cara kerja klasifikasi *Naïve Bayes* menggunakan persamaan 2 [17].

$$P(H|E) = \frac{P(E|H) \text{ X } P(H)}{P(E)} \quad (2)$$

Dimana:

P(E) = probabilitas awal (priori) bukti E tanpa melihat hipotesis yang lain.

P(H) = probabilitas awal (priori) hipotesis H tanpa melihat bukti apapun.

P(H|E) = probabilitas akhir bersyarat *(conditional probability)* yang terjadi pada suatu hipotesis H jika terdapat bukti E.

P(E|H) = probabilitas sebuah bukti E yang terjadi dan dapat mempengaruhi hipotesis H.

H = hipotesis atau peristiwa.

E = evidence atau bukti.

b. K-Nearest Neighbour

Perhitungan data menggunakan metode K-Nearest Neighbour memiliki beberapa proses, dimulai dengan memasukkan dataset penyakit kanker payudara, dilanjutkan dengan *membagi* data *training* dan data *testing*, proses ketiga adalah menentukan nilai K pada, selanjutnya menghitung nilai *ecludian* (jika data numerik), urutkan hasil perhitungan jarak dari yang terkecil ke terbesar, pilih mayoritas jarak terdekat, dan proses terakhir adalah menentukan hasil prediksi. Perhitungan dalam algoritma K-Nearest Neighbour dapat dijelaskan menggunakan persamaan *eucledian distance* (3) [18].

$$d\ (xi\ , yi\ ) = \sqrt{\sum_{i=0}^{n} (x_{i-}\ y_i)^2} \quad (3)$$

Keterangan :

Xi = data *training*

Yi = data *testing*

D (xi, yi) = jarak data x ke y

i = variabel data

n = dimensi data

2.5 Analisis Hasil dan Kesimpulan

Analisis hasil dilakukan dengan cara membandingkan nilai akurasi, presisi, recall dan AUC, dari masing-masing metode. Nilai akurasi, presisi, recall didapat dari matrik konfusi, sedangkan nilai AUC didapatkan dari persamaan AUC.

Tabel 2. Klasifikasi Kategori Kelas AUC

Rentang Kelas	Kategori
0.90 – 1.00	*Excellent classification*
0.80 – 0.90	*Good classification*
0.70 – 0.80	*Fair classification*
0.60 – 0.70	*Poor classification*
0.50 – 0.60	*failure*

Kesimpulan diambil berdasarkan hasil pengujian metode, sehingga kesimpulan ini nantinya yang akan menjawab rumusan masalah penelitian.

3. Hasil dan Pembahasan

3.1 Analisis Distribusi Data

Analisis distribusi data bertujuan untuk mengetahui keseimbangan jumlah kelas sehat dan sakit kanker. Distribusi kelas data *Breast Cancer Coimbra*, dapat dilihat pada tabel 3.

Tabel 3. Distribusi Data *Breast Cancer Coimbra*

No	Kelas	Jumlah

1.	1 (sehat)	52
2.	2 (sakit kanker)	65

3.2 Analisis Atribut Missing Value

Analisis *missing value* bertujuan untuk mengetahui atribut yang hilang atau kosong. Jumlah *missing value* dalam setiap atribut data *Breast Cancer Coimbra* dapat dilihat pada tabel 4.

Tabel 4. *Missing Value*

No	Atribut	Jumlah Missing Value
1.	Age	0
2.	BMI	0
3.	Glucose	0
4.	Insulin	0
5.	HOMA	0
6.	Leptin	0
7.	Adiponectin	0
8.	Resistin	0
9.	MCP.1	0
10.	Classification	0

3.3 Pemilihan Atribut Yang Akan Digunakan

Pada skenario pengujian pertama menggunakan semua atribut untuk diteliti, sedangkan untuk pengujian skenario kedua memilih atribut yang memiliki nilai korelasi pearson tertinggi untuk dihapus. Hasil pengukuran keterkaitan atribut dengan korelasi pearson menggunakan python dapat dilihat pada gambar 3.

	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1	Classification
Age	1.000000	0.008661	0.230265	0.032259	0.126975	0.102741	-0.219951	0.002390	0.013501	-0.043607
BMI	0.008661	1.000000	0.138727	0.145240	0.114314	0.569426	-0.302843	0.195283	0.223668	-0.132522
Glucose	0.230265	0.138727	1.000000	0.504518	0.696203	0.304775	-0.122004	0.291266	0.265219	0.384453
Insulin	0.032259	0.145240	0.504518	1.000000	0.932141	0.301406	-0.030926	0.146473	0.174334	0.276734
HOMA	0.126975	0.114314	0.696203	0.932141	1.000000	0.327048	-0.056072	0.230779	0.259382	0.284124
Leptin	0.102741	0.569426	0.304775	0.301406	0.327048	1.000000	-0.095491	0.255971	0.013886	-0.001084
Adiponectin	-0.219951	-0.302843	-0.122004	-0.030926	-0.056072	-0.095491	1.000000	-0.252704	-0.200565	-0.019737
Resistin	0.002390	0.195283	0.291266	0.146473	0.230779	0.255971	-0.252704	1.000000	0.366726	0.227204
MCP.1	0.013501	0.223668	0.265219	0.174334	0.259382	0.013886	-0.200565	0.366726	1.000000	0.091486
Classification	-0.043607	-0.132522	0.384453	0.276734	0.284124	-0.001084	-0.019737	0.227204	0.091486	1.000000

Gambar 4. Hasil Korelasi Pearson

Gambar 4 merupakan hasil pengukuran korelasi antara atribut menggunakan korelasi pearson. Pada gambar tersebut diketahui bahwa atribut Insulin dan HOMA memiliki korelasi tertinggi dengan nilai 0.932141 dimana nilai tersebut mendekati 1, sehingga atribut Insulin dan HOMA yang akan di hapus.

d. Normalisasi Data

Normalisasi data bertujuan untuk mengubah data dalam rentang nilai tertentu. Dalam penelitian ini menggunakan metode *min-max.*

3.4 Pengujian Model

Pengujian model yang dilakukan dalam penelitian ini menggunakan beberapa skenario, mulai dari pengujian model sebelum dan seudah normalisasi data, pengujian dengan perbandingan data *testing* dan data *training*, pengujian dengan memilih nilai k pada K-NN, serta pengujian dengan seluruh atribut dan atribut pilihan berdasarkan perhitungan korelasi pearson.

3.4.1 Pengujian Model Naïve Bayes

Pengujian model Naïve Bayes dilakukan menggunakan perbandingan data *training* dan testing dengan rasio 85:25, dan menggunakan perulangan uji 33x, 50x, 70x dan 100x. Dari pengujian tersebut dihasilkan rata-rata terbaik yang diterangkan pada tabel 5.

Tabel 5. Hasil Sebelum dan Sesudah Normalisasi Naïve Bayes

Pengukuran	**Sebelum Normalisasi**	**Sesudah Normalisasi**
Akurasi	64.77%	67.48%
Presisi Sehat	55.96%	55.18%
Presisi Sakit	83%	79.39%
Recall Sehat	88%	85.6%
Recall Sakit	47%	44.48%
ROC-AUC	0.81	0.82

3.4.2 Pengujian Model K-Nearest Neighbour

Pengujian model K-NN dilakukan menggunakan perbandingan data *training* dan testing dengan rasio 85:25, menggunakan k=3, serta menggunakan perulangan pengujian 33x, 50x, 70x dan 100x. Dari pengujian tersebut dihasilkan rata-rata terbaik yang diterangkan pada tabel 6.

Tabel 6. Hasil Sebelum dan Sesudah Normalisasi K-NN

Pengukuran	Sebelum Normalisasi	Sesudah Normalisasi
Akurasi	53.4%	67.68%
Presisi Sehat	51.12%	63.05%
Presisi Sakit	58.51%	71.6%

Recall Sehat	48.78%	67.42%
Recall Sakit	60.61%	68.06%
ROC-AUC	0.89	0.91

3.4.3 Pengujian Model Berdasarkan Perbandingan Data Training dan Testing

a. Pengujian Model Naïve Bayes

Pengujian model Naïve Bayes dilakukan menggunakan data yang sudah di normalisasi dan perbandingan data *training* dan testing dengan rasio 90:10, 85:15, 80:20, 75:25, serta menggunakan perulangan pengujian 33x, 50x, 70x dan 100x. Dari pengujian tersebut dihasilkan rata-rata terbaik yang dijelaskan pada tabel 7.

Tabel 7. Hasil Uji Perbandingan Data Naïve Bayes

Pengukuran	90:10	85:15	80:20	75:25
Akurasi	67.02%	67.48%	64.98%	63.63%
Presisi Sehat	58%	55.18%	59.66%	58.17%
Presisi Sakit	82.49%	79.39%	79.86%	79.91%
Recall Sehat	86.6%	85.6%	87.38%	84.42%
Recall Sakit	46.7%	44.48%	50.2%	47.8%
ROC-AUC	0.81	0.82	0.79	0.8

b. Pengujian Model K-Nearest Neighbour

Pengujian model K-NN dilakukan menggunakan data yang sudah di normalisasi, nilai k=5, dan perbandingan data *training* dan testing dengan rasio 90:10, 85:25, 80:20, 75:25, serta menggunakan perulangan pengujian 33x, 50x, 70x dan 100x. Dari pengujian tersebut dihasilkan rata-rata terbaik yang diterangkan pada tabel 8.

Tabel 8. Hasil Uji Perbandingan Data K-NN

Pengukuran	**90:10**	**85:15**	**80:20**	**75:25**
Akurasi	74.65%	75.75 %	72.2%	73%
Presisi Sehat	74%	63.94%	71.29%	70%
Presisi Sakit	73.69%	79.35%	73.79%	77.14%
Recall Sehat	65.62%	86%	68.41%	70.35%
Recall Sakit	82.18%	50.55%	75.44%	77.02%
ROC-AUC	0.9	0.91	0.85	0.88

3.4.4 Pengujian Model Berdasarkan Nilai K Algoritma K-NN

Pengujian model K-NN dilakukan menggunakan data yang sudah di normalisasi, nilai k ganjil 1, 3, 5, 7, 9, dan perbandingan data *training* dan testing dengan rasio 85:25, serta menggunakan perulangan pengujian 33x, 50x, 70x dan 100x. Dari pengujian tersebut dihasilkan rata-rata terbaik yang diterangkan pada tabel 9.

Tabel 9. Hasil Uji Nilai K Pada K-NN

Pengukuran	**K=1**	**K=3**	**K=5**	**K=7**	**K=9**
Akurasi	69.25%	67.68%	75.75%	74.62%	69.63%
Presisi Sehat	64.1%	63.05%	63.94%	67%	69.56%
Presisi Sakit	75.42%	71.6%	79.35%	75%	78.54%
Recall Sehat	71.85%	67.42%	86%	70.46%	72.82%
Recall Sakit	68.25%	68.6%	50.55%	71.35%	76.28%
ROC-AUC	1	0.91	0.91	0.88	0.86

3.4.5 Pengujian Model Berdasarkan Pemilihan Atribut Terkuat Dengan Uji Korelasi Pearson

a. Pengujian Model Naïve Bayes

Pengujian model Naïve Bayes dilakukan menggunakan data yang sudah di normalisasi, melakukan drop 2 atribut dengan korelasi pearson tertinggi yaitu Insulin dan HOMA, menggunakam perbandingan data *training* dan testing dengan rasio 85:25 serta menggunakan perulangan pengujian 33x, 50x, 70x dan 100x. Dari pengujian tersebut dihasilkan rata-rata terbaik yang dijelaskan pada tabel 10.

Tabel 10. Hasil Uji Drop Atribut Terkuat Pada Naïve Bayes

Pengukuran	**Drop Insulin**	**Drop HOMA**	**Drop Insulin&HOMA**
Akurasi	66.31%	67.80%	69.12%
Presisi Sehat	59.68%	60.72%	64.90%
Presisi Sakit	83%	81.97%	80%
Recall Sehat	87.45%	84.66%	82.39%
Recall Sakit	49.59%	53.93%	61.11%
ROC-AUC	0.81	0.81	0.82

b. Pengujian Model K-Nearest Neighbour

Pengujian model K-NN dilakukan menggunakan data yang sudah di normalisasi, nilai k=5, melakukan drop 2 atribut dengan korelasi pearson tertinggi yaitu Insulin dan HOMA, menggunakam perbandingan data *training* dan testing dengan rasio 85:25 serta menggunakan perulangan pengujian 33x, 50x, 70x dan 100x. Dari pengujian tersebut dihasilkan rata-rata terbaik yang diterangkan pada tabel 11.

Tabel 11. Hasil Uji Drop Atribut Terkuat Pada K-NN

Pengukuran	**Drop Insulin**	**Drop HOMA**	**Drop Insulin&HOMA**
Akurasi	76.83%	73.29.80%	76.10%
Presisi Sehat	76%	70%	74.69%
Presisi Sakit	78.84%	77%	80.21%
Recall Sehat	74.18%	71.87%	70.57%
Recall Sakit	79%	75.87%	80.81%
ROC-AUC	0.9	0.89	0.91

3.4.6 Analisis Hasil Pengujian Model

Pada penelitian ini terdapat beberapa skenario pengujian, yaitu pengujian berdasarkan data sebelum dan sesudah normalisasi, pengujian berdasarkan perbandingan data *training* dan data *testing*, pengujian berdasarkan nilai K algoritma K-NN serta pengujian berdasarkan pemilihan atribut terkuat dengan uji korelasi pearson. Berdasarkan skenario pengujian pertama dengan membandingkan data yang belum di normalisasi dan data yang sudang di normalisasi dapat diketahui bahwa setelah data di normalisasi menghasilkan kenaikan akurasi dan AUC pada model Naïve Bayes dan K-NN, dengan nilai akurasi Naïve Bayes sebelum normalisasi adalah 64.77% menjadi 67.48% setelah normalisasi, sedangkan pada K-NN yang sebelum normalisasi akurasi model adalah 53.4% menjadi 67.68% setelah normalisasi. Nilai Presisi dan Recall sebelum normalisasi pada Naïve Bayes lebih tinggi dibandingkan setelah normalisasi, sedangkan untuk K-NN nilai presisi dan recall setelah normalisasi lebih tinggi dibandingkan sebelum normalisasi. Nilai AUC model Naïve Bayes sebelum normalisasi adalah 0.81 sedangkan sesudah normalisasi adalah 0.82 yang termasuk kategori *good classification* dan untuk nilai AUC model K-NN sebelum normalisasi adalah 0.89 termasuk kategori *good classification* sedangkan setelah normalisasi adalah 0.91 termasuk *excellent classification*.

Skenario pengujian kedua dengan mengatur perbandingan data *training* dan data *testing* serta menggunakan data yang sudah di normalisasi. Pengaturan perbandingan data *training* dan data *testing* yang digunakan dalam penelitian ini adalah 90:10, 85:15, 80:20 juga 75:35. Dalam penelitian ini diketahui bahwa perbandingan data 85:15 menghasilkan akurasi dan AUC rata-rata tertinggi pada model Naïve Bayes dan K-NN, dengan nilai akurasi Naïve Bayes adalah 67.48%, sedangkan pada K-NN menghasilkan akurasi 75.75%. Pada pengukuran Presisi dan Recall Naïve Bayes yang menghasilkan nilai rata-rata tertinggi adalah perbandingan 80:20 dengan 59.66% presisi sehat, 87.38% recall sehat dan 50.2% recall sakit, sedangkan nilai rata-rata presisi sakit tertinggi adalah 82.49% pada perbandingan data 90:10. Untuk K-NN nilai presisi dan recall tertinggi pada perbandingan data 85:15, dengan nilai presisi sakit 79.35% dan recall sehat 82.18%, sedangkan nilai rata-rata presisi sehat tertinggi adalah 74% dan recall sakit tertinggi adalah 82.18% pada perbandingan data 90:10. Nilai AUC model Naïve Bayes dengan perbandingan data 85:15 adalah 0.82 yang termasuk kategori *good classification* dan untuk nilai AUC model K-NN dengan perbandingan data 85:15 adalah 0.91 termasuk *excellent classification*.

Pengujian menggunakan skenario ketiga, dilakukan dengan memilih nilai K pada K-NN dengan nilai K merupakan bilangan ganjil dari 1 hingga 10 yaitu 1, 3, 5,7, 9 dan perbandingan data 85:25. Pada penelitian ini diketahui bahwa K=5 memiliki akurasi, presisi sakit, dan recall sehat rata-rata tertinggi pada model K-NN, dengan nilai akurasi K-NN adalah 75.75%, presisi sakit 79.35%, dan recall sehat 86%, sedangkan untuk presisi sehat tertinggi dan recall sakit tertinggi pada K=9 dengan 69.56% presisi sehat, 76.28% recall sakit. Nilai AUC tertinggi pada pengujian berdasarkan nilai K adalah 1 termasuk *excellent classification* pada K=1 serta 0.91 pada K=3 dan K=5 yang juga termasuk *excellent classification*.

Pada skenario pengujian keempat dengan mengatur perbandingan data *training* dan data *testing* 85:15 dan pemilihan atribut HOMA dan Leptin sebagai atribut dengan korelasi pearson tertinggi, serta menggunakan K=5 pada KNN. Dalam penelitian ini diketahui pada model Naïve Bayes menghasilkan rata-rata nilai akurasi, presisi sehat, recall sakit dan AUC tertinggi setelah melakukan drop atribut Insulin dan Homa yaitu sebesar 69.12% untuk akurasinya, 64.90% untuk presisi sehat, 61.11% untuk recall sakit dan 0.82 untuk pengukuran AUC yang termasuk *good classification*, sedangkan untuk hasil rata-rata presisi sakit dan recall sehat tertinggi setelah melakukan drop Insulin, dengan 83% presisi sakit dan 87.45% recall sehat. Pengujian dengan K-NN menghasilkan rata rata akurasi, presisi sehat dan recall sehat tertinggi setelah melakukan drop Insulin yaitu 76.83% untuk akurasi, 76% pada presisi sehat dan 74.18% untuk recall sehat, sedangkan untuk hasil rata-rata presisi sakit, recall sakit dan AUC tertinggi

didapatkan setelah melakukan drop Insulin dan Homa, dengan rata-rata presisi sehat sebesar 80.21%, recall skit sebesar 80.81% dan AUC adalaj 0.91 termasuk kategori *excellent classification*.

4. Kesimpulan

Berdasarkan hasil pengujian dan analisis yang telah dilakukan pada penelitian tugas akhir ini, dapat diperoleh beberapa kesimpulan yaitu:

Penelitian ini menggunakan model Naïve Bayes Dan K-NN dengan 4 skenario pengujian untuk menentukan nilai akurasi, presisi, racall dan AUC tertinggi. Adapun hasil dari 4 skenario pengujian adalah sebagai berikut:

a. Skenario pengujian pertama dengan membandingkan data sebelum dan sesudah data dinormalisasi pada pengujian model Naïve Bayes menunjukan nilai akurasi, recall sakit dan AUC lebih tinggi dibandingkan sebelum data di normalisasi, namun untuk nilai presisi sehat, presisi sakit, dan recall sehat lebih tinggi sebelum data di normalisasi. Pada Pengujian K-NN menggunakan K=3 menunjukkan bahwa akurasi, presisi, recall dan AUC lebih tinggi dibandingkan sebelum normalisasi.
b. Skenario pengujian kedua dilakukan dengan membandingkan rasio data *training* dan data *testing* dengan akurasi dan AUC tertinggi pada Naïve Bayes adalah 67.48% dan 0.82 termasuk *good classification* pada perbandingan data 85:15, sedangkan pengukuran presisi sehat,recall sehat dan recall sakit memiliki hasil tertinggi saat perbandingan data 80:20 dan untuk presisi sakit tertinggi pada perbandingan data 90:10. Untuk pengujian menggunakan model K-NN menghasilkan akurasi, presisi sakit, recall sehat dan AUC tertinggi pada perbandingan data 85:15, sedangkan nilai presisi sehat dan recall sakit tertinggi pada perbandingan data 90:10.
c. Skenario pengujian ketiga dilakukan dengan mengubah-ubah nilah K pada Algoritma K-NN yang menghasilkan akurasi, presisi sakit, dan recall sehat rata-rata tertinggi, dengan nilai akurasi K-NN adalah 75.75%, presisi sakit 79.35%, dan recall sehat 86% menggunakan nilai K=5, kemudian untuk presisi sehat tertinggi dan recall sakit tertinggi pada K=9 dengan 69.56% presisi sehat, 76.28%, dan AUC tertinggi adalah 1 pada nilai K=1.
d. Skenario pengujian keempat dilakukan dengan memilih atribut terkuat berdasarkan korelasi pearson untuk di drop, sehingga menghasilkan akurasi, presisi sehat, recall sakit dan AUC tertinggi setelah menghapus atribut Insulin dan Homa yaitu sebesar 69.12% untuk akurasinya, 64.90% untuk presisi sehat, 61.11% untuk recall sakit dan 0.82 untuk pengukuran AUC yang termasuk *good classification*, serta presisi sakit dan recall sehat tertinggi setelah melakukan drop Insulin, dengan 83% dan 87.45% menggunakan model Naïve Bayes. Untuk K=NN akurasi, presisi sehat dan recall sehat tertinggi setelah melakukan drop Insulin yaitu 76.83% untuk akurasi, 76% pada presisi sehat dan 74.18% untuk recall sehat, sedangkan untuk hasil rata-rata presisi sakit, recall sakit dan AUC tertinggi didapatkan setelah melakukan drop Insulin dan Homa, dengan rata-rata presisi sehat sebesar 80.21%, recall skit sebesar 80.81% dan AUC adalaj 0.91 termasuk kategori *excellent classification*.
e. Pengujian model K-NN menghasilkan akurasi, presisi sehat, recall sakit dan AUC lebih tinggi dibandingkan Naïve Bayes. Sedangkan untuk hasil presisi sakit dan recall sehat Naive Bayes. Sehingga dapat ditarik kesimpulan berdasarkan penelitian bahwa K-NN lebih baik dibandingkan Naïve Bayes.

Daftar Rujukan

[1] Anonim, "Apa Itu Kanker?," *P2PTM Kemenkes RI*, 2019. http://p2ptm.kemkes.go.id/infographic-p2ptm/penyakit-kanker-dan-kelainan-darah/apa-itu-kanker (accessed Apr. 20, 2021).
[2] I. Namira, "75 Tahun Merdeka, tapi Indonesia Belum Bebas dari 10 Penyakit Ini!," *idntimes.com*, 2020. https://www.idntimes.com/health/medical/izza-namira-1/75-tahun-merdeka-tapi-indonesia-belum-bebas-dari-10-penyakit-ini/10 (accessed Apr. 20, 2020).
[3] Anonim, "Penyakit Kanker di Indonesia Berada Pada Urutan 8 di Asia Tenggara dan Urutan 23 di Asia," *p2p.kemkes.go.id*, 2020. http://p2p.kemkes.go.id/penyakit-kanker-di-indonesia-berada-pada-urutan-8-di-asia-tenggara-dan-urutan-23-di-asia/ (accessed Apr. 20, 2020).
[4] H. Widowati, "Kasus Kanker Payudara Paling Banyak Terjadi di Indonesia," *databoks.katadata.co.id*, 2019. https://databoks.katadata.co.id/datapublish/2019/06/03/kasus-kanker-payudara-paling-banyak-terjadi-di-indonesia (accessed Apr. 20, 2020).
[5] R. Fadli, "Kanker Payudara," *halodoc.com*, 2021. https://www.halodoc.com/kesehatan/kanker-payudara (accessed Apr. 20, 2021).
[6] M. Nareza, "Kanker Payudara," *Alodokter.com*, 2021. https://www.alodokter.com/kanker-payudara (accessed Apr. 20, 2020).
[7] Holy Kartika Nurwigati Sumartiningtyas, "Kanker Payudara Paling Banyak Didiagnosis di Dunia, Studi Jelaskan," *Kompas.com*, 2021.

https://www.kompas.com/sains/read/2021/02/05/192600023/kanker-payudara-paling-banyak-didiagnosis-di-dunia-studi-jelaskan?page=all (accessed Apr. 20, 2021).

[8] D. Andriani, "Ini Jenis Kanker yang Paling Banyak Diderita Masyarakat Indonesia," *Lifestyle.bisnis.com*, 2020. https://lifestyle.bisnis.com/read/20200225/106/1205840/ini-jenis-kanker-yang-paling-banyak-diderita-masyarakat-indonesia (accessed Apr. 20, 2021).

[9] M. Patrício *et al.*, "Using Resistin, glucose, age and BMI to predict the presence of breast cancer," *BMC Cancer*, vol. 18, no. 1, pp. 1–8, 2018, doi: 10.1186/s12885-017-3877-1.

[10] Y. Li, "Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction," *Appl. Comput. Math.*, vol. 7, no. 4, p. 212, 2018, doi: 10.11648/j.acm.20180704.15.

[11] I. Mubarog, A. Setyanto, and H. Sismoro, "Sistem Klasifikasi Pada Penyakit Breast Cancer Dengan Menggunakan Metode Naïve Bayes," *Creat. Inf. Technol. J.*, vol. 6, no. 2, p. 109, 2021, doi: 10.24076/citec.2019v6i2.246.

[12] Ramadhan, Nur Ghaniaviyanto, and Faisal Dharma Adhinata. "Teknik SMOTE dan Gini Score dalam Klasifikasi Kanker Payudara." RADIAL: Jurnal Peradaban Sains, Rekayasa dan Teknologi 9.2 (2021): 125-134. doi: https://doi.org/10.37971/radial.v9i2.229

[13] W. D. Septiani, "Komparasi Metode Klasifikasi Data Mining Algoritma C4.5 Dan Naive Bayes Untuk Prediksi Penyakit Hepatitis," *Jurnal Pilar Nusa Mandiri*, 2017. http://ejournal.nusamandiri.ac.id/index.php/pilar/article/view/149/126 (accessed Apr. 20, 2020).

[14] R. N. Devita, H. W. Herwanto, and A. P. Wibawa, "Perbandingan Kinerja Metode Naive Bayes dan K-Nearest Neighbor untuk Klasifikasi Artikel Berbahasa indonesia," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 4, pp. 427–434, 2018, doi: 10.25126/jtiik.201854773.

[15] H. W. Mochammad, "Algoritma Naive Bayes," *Binus.ac.id*, 2019. https://binus.ac.id/bandung/2019/12/algoritma-naive-bayes/ (accessed Apr. 21, 2020).

[16] Y. G. Prasetyowati, "Klasifikasi menggunakan Metode KNN (K-Nearest Neighbor) dalam Python," *Medium.com*, 2019. https://medium.com/@16611130/klasifikasi-menggunakan-metode-knn-k-nearest-neighbor-dalam-python-a40e79a74101 (accessed Apr. 22, 2020).

[17] A. P. Ayudhitama and U. Pujianto, "Analisa 4 Algoritma Dalam Klasifikasi Penyakit Liver Menggunakan Rapidminer," *J. Inform. Polinema*, vol. 6, pp. 1–9, 2020.

[18] F. Tempola, M. Muhammad, and A. Khairan, "Perbandingan Klasifikasi Antara KNN dan Naive Bayes pada Penentuan Status Gunung Berapi dengan K-Fold Cross Validation," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 5, p. 577, 2018, doi: 10.25126/jtiik.201855983.
