Indonesian Journal of Health Science
Volume 5 No.2, 2025

PERBANDINGAN METODE K-NEAREST NEIGHBORS (K-NN) DAN REGRESI LOGISTIK BINER DALAM MEMPREDIKSI KANKER

Christina Amanda Surbakti [1)], Albert Samuel Sinaga [2)], Agnes Monica Simorangkir [3)]*, Auta Shinta Sarah [4)], Clara Jocelyn Harefa [5)], Syairal Fahmy Dalimunthe [6)]

[1)] christinaamanda86@gmail.com, *Universitas Negeri Medan*
[2)] albertsamuels02@gmail.com, *Universitas Negeri Medan*
[3)] agnesmosimorangkir@gmail.com, *Universitas Negeri Medan*
[4)] shintaauta@gmail.com, *Universitas Negeri Medan*
[5)] clarajocelynharefa@gmail.com, *Universitas Negeri Medan*
[6)] fahmy@unimed.ac.id, *Universitas Negeri Medan*
*penulis korespondensi

Abstarck
Background: *Cancer is one of the diseases with a high mortality rate, so an accurate classification method is needed to support the diagnosis process. This study compares the performance of the K-Nearest Neighbors (KNN) method and Binary Logistic Regression in classifying cancer as malignant or benign.* ***Methods:*** *This study used a secondary dataset from Kaggle consisting of 569 cancer patient data with 11 independent variables covering tumor characteristics. The model was developed using data normalization, training and testing data division, and the K-Fold Cross Validation technique to optimize the K parameter in KNN. Model evaluation was carried out based on accuracy, precision, recall, and the McNemar and ANOVA tests to test the significance of differences in model performance.* ***Results:*** *The KNN model with K=13 showed an accuracy of 95.58%, a precision of 95.83%, and a recall of 97.18%, while Binary Logistic Regression had an accuracy of 94.69%, a precision of 92.86%, and a recall of 92.86%. The McNemar test results showed that there was no significant difference between the two models (p-value = 1), while the ANOVA results showed that all independent variables contributed to the model.* ***Conclusion:*** *Both methods performed well in cancer classification, but KNN with K=13 had a slight advantage in accuracy and recall compared to Binary Logistic Regression. The implementation of this model can support decision support systems in cancer diagnosis to improve the accuracy of classification results.*
Keywords: *Binary Logistic Regression, Cancer, K-Nearest Neighbors, Prediction*

Abstrak
Latar Belakang: Kanker merupakan salah satu penyakit yang memiliki tingkat kematian tinggi, sehingga dibutuhkan metode klasifikasi yang akurat untuk mendukung proses diagnosis. Penelitian ini membandingkan performa metode K-Nearest Neighbors (KNN) dan Regresi Logistik Biner dalam mengklasifikasikan kanker sebagai ganas atau jinak. **Metode:** Penelitian ini menggunakan dataset sekunder dari Kaggle yang terdiri dari 569 data pasien kanker dengan 11 variabel independen yang mencakup karakteristik tumor. Model dikembangkan dengan menggunakan normalisasi data, pembagian data training dan testing, serta teknik K-Fold Cross Validation untuk optimasi parameter K dalam KNN. Evaluasi model dilakukan berdasarkan akurasi, presisi, recall, serta uji McNemar dan ANOVA untuk menguji signifikansi perbedaan performa model. **Hasil:** Model KNN dengan K=13 menunjukkan akurasi 95,58%, presisi 95,83%, dan recall 97,18%, sementara Regresi Logistik Biner memiliki akurasi 94,69%, presisi 92,86%, dan recall 92,86%. Hasil uji McNemar menunjukkan bahwa tidak terdapat perbedaan signifikan antara kedua model (p-value = 1), sedangkan hasil ANOVA menunjukkan bahwa semua variabel independen berkontribusi terhadap model. **Kesimpulan:** Kedua metode menunjukkan performa yang baik dalam klasifikasi kanker, tetapi KNN dengan K=13 memiliki sedikit keunggulan dalam akurasi dan recall dibandingkan Regresi Logistik Biner. Implementasi model ini dapat mendukung sistem pendukung keputusan dalam diagnosis kanker untuk meningkatkan ketepatan hasil klasifikasi.
Kata Kunci: Kanker, K-Nearest Neighbors, Prediksi, Regresi Logistik Biner

PENDAHULUAN

Kanker adalah kondisi dimana sel di bagian tubuh tertentu mengalami pertumbuhan yang tidak wajar (normal), sel membelah terus-menerus dan di luar kendali. Kanker merupakan penyakit yang tidak menular, namun penderitanya meningkat setiap tahun dan sangat berbahaya sehingga menyebabkan tingginya angka kematian (Alfiani, Widayanti, & Putri, 2024). Tumor jinak biasanya tidak berkembang menjadi kanker. Namun, ada juga tumor yang bersifat ganas

yang bisa berkembang menjadi kanker. Tidak semua tumor bisa menyebabkan kanker. Berdasarkan sel-sel yang terkandung di dalamnya, tumor dikelompokan menjadi tiga jenis, yaitu tumor jinak, tumor pra-kanker, dan tumor ganas (Akri, 2024). Klasifikasi data kanker dapat membantu dalam meramalkan hasil penyakit atau menemukan genetika tumor pada pasien kanker. Dalam ilmu kedokteran, salah satu masalah yang paling menantang adalah menentukan penyakit pasien berdasarkan beberapa tes. Akibatnya, diagnostik medis semakin mengandalkan sistem pengklasifikasian (Solikin, 2021). Untuk mengatasi tantangan dalam diagnosis kanker, berbagai metode klasifikasi data telah dikembangkan dalam bidang kecerdasan buatan dan statistika. Dua pendekatan yang umum digunakan dalam klasifikasi kanker adalah *K-Nearest Neighbors* (KNN) dan Regresi Logistik Biner.

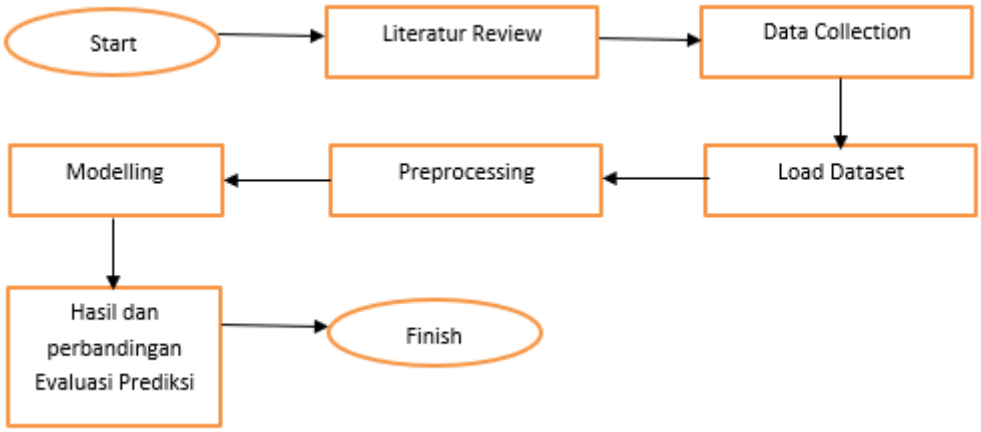
K-Nearest Neighbor (KNN) adalah metode klasifikasi terhadap objek baru berdasarkan data training yang memiliki jarak tetangga terdekat (*nearest neighbor*) dengan objek baru tersebut (Fasnuari, Yuana, & Chulkamdi, 2022). Regresi logistik yaitu bentuk regresi yang digunakan untuk memodelkan hubungan antara variabel dependen dan variabel independen, ketika variabel adalah sebuah data dengan ukuran biner/dichotomous misalnya ya atau tidak, sukses atau gagal, puas atau tidak puas, bagus atau rusak, mati atau hidup (Putri, Titaley, & Salaki, 2022).

Beberapa penelitian sebelumnya telah membahas perbandingan algoritma *K-Nearest Neighbor* (K-NN) dan Regresi Logistik dalam klasifikasi deteksi dini kanker serviks oleh Nur Devita Azzahra Dengan menggunakan teknik pengujian *percentage split* dan *k-fold cross validation*, penelitian ini menemukan bahwa Regresi Logistik dengan teknik *k-fold cross validation* lebih efektif dalam mengklasifikasikan kanker serviks, mencapai nilai akurasi sebesar 96 (Azzahra, Ambarwati, Desiani, Maiyanti, & Ramayani, 2024). Berikutnya penelitian yang menerapkan metode KNN pada dataset pasien kanker payudara oleh Dewi Cahyani dengan nilai k antara 3 hingga 5, serta menggunakan *cross-validation* dengan k-fold=5. Hasilnya menunjukkan bahwa metode KNN mencapai akurasi tertinggi sebesar 93% pada subset data tertentu (Cahyanti, Rahmayani, & Husniar, 2020). Kemudian penelitian yang membandingkan kinerja algoritma KNN dan *Decision Tree* dalam mendeteksi dini kanker payudara oleh Fahrurrozi (Fahrurrozi & Wasilah., 2023). Hasilnya menunjukkan bahwa algoritma Decision Tree memiliki akurasi sebesar 96,49%, sedangkan KNN memiliki akurasi 94,73%. Berikutnya penelitian menggunakan dataset METABRIC yang terdiri dari 1.904 catatan pasien untuk memprediksi kelangsungan hidup 5 tahun pasien kanker payudara menggunakan pendekatan machine learning oleh Khaola Chtouki Studi ini membandingkan tujuh model klasifikasi, termasuk *Logistic Regression* dan *K-Nearest Neighbor* (KNN), untuk mengevaluasi kinerja mereka dalam memprediksi tingkat kelangsungan hidup pasien (Chtouki, Rhanoui, Mikram, & Yousfi, 2023). Dan penelitian yang terakhir ini mengevaluasi kinerja algoritma *Logistic Regression, Naïve Bayes,* dan *Random Forest* dalam memprediksi kanker payudara menggunakan dataset Coimbra oleh Cecep Wahyu Cahyana.Hasilnya menunjukkan bahwa *Logistic Regression* mencapai akurasi sebesar 80%, sedangkan *Random Forest* mencapai akurasi 85% (Cahyana & Nurlayli, 2023). Sedangkan penelitian kali ini akan membandingkan performa KNN dan Regresi Logistik Biner dalam mengklasifikasikan suatu kanker tumor berdasarkan tipe kanker yaitu ganas atau jinak, sehingga dapat membantu dalam proses diagnosis yang lebih akurat dan efisien.

METODE

Jenis Penelitian yang dilakukan adalah penelitian kuantitatif. Penelitian kuantitatif adalah penelitian yang menghasilkan temuan-temuan baru yang dapat dicapai dengan menggunakan prosedur-prosedur statistik (Ali, Hariyati, Pratiwi, & Afifah, 2022). Dalam melakukan penelitian tentu saja tidak terlepas dari tahapan tahapan yang dilakukan untuk

memperoleh hasil dari penelitian itu sendiri. Begitu juga dengan penelitian kali ini, terdapat beberapa tahapan yang dilakukan yang dapat dilihat pada flowchart berikut ini:



Gambar 1. Flowchart Alur Penelitian

Literatur Review

Langkah selanjutnya adalah melakukan literatur *review* dengan tujuan untuk menemukan landasan teori yang digunakan dan mencari literatur-literatur ilmiah yang relavan untuk mendukung penelitian yang dilakukan.

Data Collection dan Load Dataset

Dataset yang digunakan adalah dataset yang berisi karakteristik pasien yang didiagnosis kanker. Data ini merupakan data sekunder yang diperoleh dari kaggle (https://www.kaggle.com/datasets/erdemtaha/cancer-data). Data sekunder adalah data yang diperoleh lewat pihak lain, tidak langsung diperoleh oleh peneliti dari subjek penelitiannya (Pratama, Naila, & Faradita, 2024).

Berdasarkan analisis yang dilakukan terhadap data kanker pada website kaggle, terdapat 32 variabel yang terdapat pada kaggle, tetapi data yang digunakan pada penelitian ini hanya menggunakan 11 variabel saja. Dikarenakan pada bagian deskripsi website kaggle pada data kanker dijelaskan terdapat 8 variabel yang termasuk dalam kategori mean features yang digunakan dalam analisis kanker yaitu *radius mean, texture mean, perimeter mean, area mean, smoothness mean, compactness mean, concavity mean,* dan *concave points mean* serta tambahan 2 variabel lagi yang termasuk dalam variabel rata -rata yaitu *symmetry mean* dan *fractal dimension mean*. Berdasarkan informasi dari kaggle variabel-variabel ini sudah mewakili karakteristik visual kanker. Adapun 11 variabel yang digunakan yaitu 1 variabel dependen dan 10 variabel independen dengan jumlah data sebanyak 569 seperti yang ditunjukkan pada Tabel 1.

Tabel 1
Deskripsi Dataset Kaggle Jenis Kanker

Nama Variabel	Jenis Variabel	Deskripsi
Diagnosis	Dependen (Y)	Tergolong Kanker Tumor Ganas=M Tergolong Kanker Tumor Jinak=B
Radius Mean	Independen (x_1)	Rata-rata jari-jari tumor
Texture mean	Independen (x_2)	Rata-Rata Tekstur Permukaan Tumor
Perimeter Mean	Independen (x_3)	Rata-Rata Keliling Tumor
Area Mean	Independen (x_4)	Rata-Rata Luas Tumor
Smoothness Mean	Independen (x_5)	Rata-Rata Kehalusan Permukaan Tumor
Compactness Mean	Independen (x_6)	Rata-Rata Kepadatan Bentuk Tumor

Concarity Mean	Independen (x_7)	Rata-Rata Cekungan Pada Tepi Tumor
Concave Points Mean	Independen (x_8)	Rata-Rata Jumlah Titik Cekungan
Symmetry Mean	Independen (x_9)	Rata-Rata Simetri Bentuk Tumor
Fractal Dimension Mean	Independen (x_{10})	Rata-Rata Kompleksitas Bentuk Tumor

Sumber: data diolah (2025)

Preprocessing

Data *preprocessing* adalah proses pengolahan data awal yang digunakan untuk mengubah data mentah yang diperoleh dari berbagai sumber menjadi informasi yang lebih bersih dan dapat digunakan untuk analisis lebih lanjut (Salam, et al., 2023).

Langkah yang diambil dalam penelitian ini meliputi:

a. Pembersihan data, yang bertujuan untuk menyiapkan data agar siap digunakan dalam model. Tahap ini melibatkan pemeriksaan berbagai informasi penting dalam dataset, termasuk identifikasi nilai yang hilang (missing value).
b. Normalisasi data, yaitu proses yang bertujuan untuk mengubah nilai-nilai dalam dataset sehingga berada dalam rentang yang lebih konsisten atau standar.
c. Split Data, yaitu proses pembagian dataset menjadi beberapa bagian yaitu data training untuk keperluan pelatihan dan juga data testing untuk pengujian model.

Modelling

Tahap *Modeling* dalam metodologi merupakan tahap di mana berbagai teknik pemodelan diterapkan pada data yang telah dipersiapkan untuk memenuhi tujuan yang telah ditetapkan (Sugiyono & Hartinah, 2024). Teknik *modelling* data yang digunakan dalam penelitian ini adalah Regresi Logistik Biner dan KNN (K-Nearst Neighbor).

a. Prediksi

Prediksi merupakan bidang pengetahuan yang diterapkan secara sistematis untuk memperoleh informasi berdasarkan data yang ada. Hal ini tentunya melibatkan proses estimasi terhadap peristiwa dimasa depan dengan menggunakan berbagai informasi atau data yang signifikam dari periode sebelumnya.

b. Regresi Logistik Biner

Model regresi logistik biner termasuk dalam kategori sebaran keluarga Eksponensial, yang dalam hal ini merujuk pada sebaran Bernoulli. Sebaran Bernoulli adalah sebaran untuk peubah acak yang memiliki dua kategori, yaitu 0 dan 1. Dalam konteks regresi logistik biner, variabel respon terdiri dari dua kategori, di mana Y = 1 menunjukkan hasil 'sukses' dan Y = 0 menunjukkan hasil 'gagal'. Karena variabel Y hanya memiliki dua kategori, maka variabel tersebut mengikuti distribusi Bernoulli dengan fungsi probabilitas tertentu.

Rumus umum untuk Regresi Logistik adalah:

$$Y = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

Dimana:

π_i = peluang kejadian ke-i

y_i = peubah acak ke-I yang terdiri dari 0 dan 1

Untuk mempermudah memprediksi parameter regresi, maka $\pi(x)$ pada persamaan diatas ditransformasikan sehingga menghasilkan bentuk logit regresi logistik:

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

Untuk mengevaluasi kesesuaian model dan menguji signifikansi variabel independen secara keseluruhan, dapat dilakukan Analisis Ragam (ANOVA). Teknik analisis varians satu jalur (one-way ANOVA) sering digunakan dalam penelitian untuk menguji apakah terdapat perbedaan signifikan antara kelompok dalam suatu dataset. Dalam konteks regresi logistik biner, uji ANOVA dapat digunakan untuk membandingkan model penuh dengan model yang dikurangi guna menentukan apakah variabel independen tertentu memberikan kontribusi

signifikan terhadap model. Penerapan uji ini bertujuan memastikan bahwa model yang dibangun memiliki performa optimal serta tidak mengandung variabel yang tidak memberikan pengaruh signifikan (Fajrin, Pathurahman, & Pratama, 2016)

c. KNN (*K-Nearest Neighbor*)

K-Nearest Neighbor adalah salah satu metode klasifikasi yang mengategorikan data baru berdasarkan kedekatan lokasi (jarak) paling dekat suatu data baru dengan data lain atau kebeberapa data/tetangga (neighbor) terdekat (Martha, Andani, & Rizki, 2022). Menurut (Bakriansyah, Subair, & Setiawan, 2025) menentukan nilai K (jumlah tetangga paling dekat) di dalam metode K-NN dianjurkan ganjil seperti 1,3,5,…. Dalam algoritma *K-Nearest Neighbors* (KNN), *K-Fold Cross Validation* juga digunakan untuk menentukan nilai K yang optimal. Dengan mencoba berbagai nilai K dan mengevaluasi performanya di setiap *fold*, kita dapat memilih K terbaik yang memberikan akurasi tertinggi dan menghindari *overfitting* atau *underfitting*. Cross-validation adalah metode untuk mengevaluasi keefektifan model dengan membagi data menjadi data *training* (untuk melatih model) dan data *testing* (untuk menguji model). Salah satu teknik yang sering digunakan adalah *K-Fold Cross Validation*, di mana data dibagi menjadi K bagian yang bergantian menjadi *training* dan *testing*. Metode ini mengurangi bias karena setiap data memiliki kesempatan menjadi *training* maupun *testing* (Widyaningsih, Arum, & Prawira, 2021).

Melakukan perhitungan nilai jarak (*education distance*) terhadap masing-masing objek data yang diberikan. Rumus untuk menghitung euclidean distance dapat dilihat pada persamaan berikut:

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

Dimana:

$d(x,y) =$ jarak Euclidian

$x_i =$ data $training$

$y_i =$ data $testing$ (Alkhussayid & Ferdiansyah, 2022)

Hasil dan Perbandingan Evaluasi Prediksi

Pada tahap ini akan ditentukan apakah hasil dari tahap sebelumnya mampu memenuhi tujuan yang telah ditetapkan.

a. *False Positive* dan *False Negative*

Menurut (Alkhussayid & Ferdiansyah, 2022), ada dua hal yang harus diperhatikan dalam menentukan akurasi dari suatu model klasifikasi, yaitu *false negative* (F.N) dan *false positive* (F.P). *False negative* dapat terjadi jika sistem mengidentifikasikan suatu kelas yang negatif tetapi pada sistem data tersebut terdeteksi ke dalam kelas tidak negatif. Sebaliknya false positif (F.P) adalah sistem yang mengidentifikasikan suatu kelas positif tetapi sistem mendeteksi sebagai data kelas negatif. Berdasarkan nilai F.P dan F.N inilah dapat dihitung nilai *recall*, presisi dan akurasi sistem dalam model klasifikasi yang dibangun. Adapun pembagian F.N dan F.P dapat dilihat pada matrik dibawah ini:

Tabel 2
Pembagian F.N dan F.P

		Kondisi	
		Layak	Tidak
Pengujian	Layak	Benar (True Positive (T.P))	Salah (false Negatife (F.N))
	Tidak	Salah (False Positive (F.P))	Benar (True Negative (T.N))

Sumber: data diolah (2025)

b. Presisi, *Recall* dan Akurasi

Suatu performa model dapat dilihat melalui tingkat akurasi, Presisi dan *Recall*-nya. Nilai *precision* adalah nilai ketepatan dan juga nilai sensifitas sistem antara informasi yang

diberikan oleh sistem untuk menunjukkan secara benar data kelas negatif atau kelas positif. Sedangkan nilai *recall* merupakan nilai yang memperlihatkan tingkat keberhasilam atau spesifitas untuk mengetahui kembali sebuah informasi secara benar tentang data yang kelas negatif ataupum yang positif. Akurasi sendiri merupakan nilai rasip data yang benar terdeteksi di dalam data pengujian, yang menunjukkan tingkat kedekatan antara nilai prediksi sistem dengan nilai prediksi manusia (Azhari,situmorang, & Rosnelly, 2021) .

Nilai presisi, *recall* dan akurasi dapat dicari dengan persamaan berikut ini

$$\text{Presisi} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\%$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\%$$

$$\text{Akurasi} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \times 100\%$$

c. Perbandingan Akurasi Model

Berdasarkan hasil presisi, recall, dan akurasi dari model regresi logistik biner dan KNN, akan dilakukan perbandingan terhadap ketiga nilai tersebut untuk melihat performa klasifikasi dari kedua model.

- Uji McNemar

Menurut (Kainama, Palit, & Hutabarat, 2022) Tabel Bantu Segiempat Uji McNemar

	Metode 2	
Metode 1	+	-
+	A	B
-	C	D

Sumber: data diolah (2025)

Hipotesis yang dibangun:

H_0 : tidak ada perbedaan signifikan antara metode 1 dengan metode 2

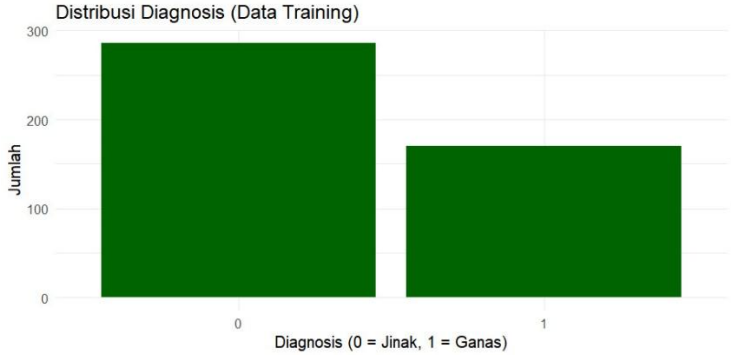
H_1: ada perbedaan signifikan antara metode 1 dengan metode 2

Jika p-value > 0,05 maka gagal tolak H_0, artinya tidak ada perbedaan signifikan antara metode 1 dengan metode 2

HASIL DAN PEMBAHASAN

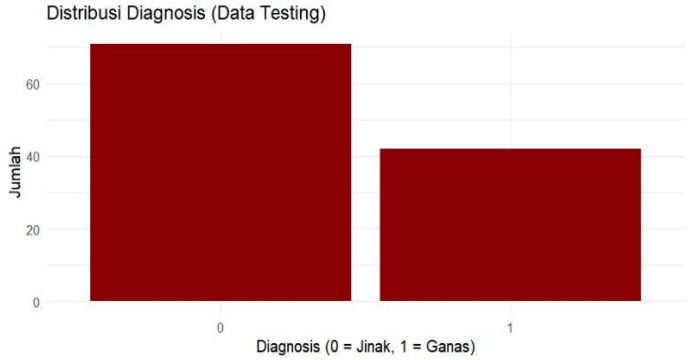
Analisis data

Proses normalisasi bertujuan untuk mengubah data mentah sehingga nilai-nilai dalam dataset berada dalam rentang tertentu yang telah ditentukan (biasanya antara 0 hingga 1, atau -1 hingga 1). Ini penting terutama untuk model-model yang mengandalkan jarak antar data, karena variabel dengan rentang nilai yang lebih besar akan memiliki pengaruh yang lebih besar pada model dibandingkan variabel dengan rentang yang lebih kecil.



Gambar 2. Distribusi Diagnosis Kanker (Data Training)

Grafik batang di atas menunjukkan distribusi diagnosis pada data pelatihan (training data) untuk dua kategori, yaitu jinak (0) dan ganas (1). Dari grafik tersebut, terlihat bahwa jumlah sampel dengan diagnosis jinak jauh lebih banyak dibandingkan dengan diagnosis ganas. Data menunjukkan bahwa jumlah kasus jinak sebesar 67% mendekati 300 kasus, sedangkan jumlah kasus ganas sebesar 33% sekitar 150 kasus.



Gambar 3. Distribusi Diagnosis Kanker (Data Testing)

Diagram batang di atas menunjukkan distribusi diagnosis pada data uji (*testing* data) untuk dua kategori, yaitu jinak (0) dan ganas (1). Dari grafik tersebut, terlihat bahwa jumlah kasus dengan diagnosis jinak lebih banyak dibandingkan dengan diagnosis ganas. Jumlah kasus jinak sebesar 64% berada di sekitar 70 kasus, sedangkan jumlah kasus ganas sebesar 36% sekitar 40 kasus.

Berdasarkan dua diagram batang yang menunjukkan distribusi diagnosis pada data pelatihan (*training* data) dan data uji (*testing* data), dapat disimpulkan bahwa dalam kedua dataset terdapat proporsi yang lebih besar untuk kasus jinak dibandingkan dengan kasus ganas.

Metode Regresi Logistik Biner

Dengan data *training* yang sudah dibuat, nantinya data tersebut akan digunakan untuk membut model regresi logistik biner yang dapat mengidentifikasi tipe kanker tersebut. Model akan dibuat dengan bantuan software RStudio. Hasil yang didapat adalah sebagai berikut.

Tabel 3
Pemodelan regresi logistik biner

	Estimated	Std. error	z value	p-value
Intercept	-10.962	3.772	-2.945	0.00323
Radius_mean	-112.833	92.895	-1.215	0.22450
Texture_mean	12.421	2.291	5.421	5.93 . 10^-8
Perimeter_mean	34.404	88.227	0.390	0.69657
Area_mean	124.357	44.122	2.818	0.00483
Smoothness_mean	10.742	4.475	2.401	0.01637
Compactness_mean	-3.417	7.388	-0.462	0.64374
Concavity_mean	3.536	3.856	0.917	0.35910
Concave point_mean	15.289	7.077	2.160	0.03075
Symmetry_mean	3.528	2.403	1.491	0.13601
Fractal_dimension_mean	-6.168	4.732	-1.303	0.19242

Sumber: data diolah (2025)

Pada tabel hasil pemodelan regresi logistik biner pada tabel 2 terlihat bahwa variabel yang memiliki nilai p-value kurang dari 0.05 merupakan variabel yang signifikan dalam memprediksi tipe kanker. Variabel yang termasuk signifikan ialah *texture mean, area mean, smoothness mean,* dan *concave mean point*. Sementara variabel yang memiliki nilai p-value lebih dari 0.05 adalah variabel yang kurang signifikan dalam memprediksi tipe kanker. Untuk

melihat apakah keenam variabel yang kurang signifikan itu tetap penting atau tidak dalam memprediksi tipe kanker, kita dapat membandingkan model yang dibentuk dengan semua variabel independen dan model yang dibentuk dengan hanya keempat variabel independen yang signifikan dengan uji ANOVA. Hasil uji ANOVA dari kedua metode dapat kita lihat pada tabel berikut.

Tabel 4
Hasil Uji Anova

Model	Residual DF	Residual Deviance	DF	Deviance	p-value
Model full variable	451	124.85	-	-	-
Model reduced variable	445	110.76	6	14.085	0.0287

Sumber: data diolah (2025)

Pada tabel ANOVA terlihat bahwa nilai p-value kurang dari 0.05 yang berarti model dengan semua variabel independen lebih baik daripada model dengan keempat variabel signifikan saja. Hal ini berarti keenam variabel yang kurang signifikan tersebut tetap penting untuk memprediksi tipe kanker. Sehingga model yang akan digunakan dalam pengujian adalah model yang dibuat dengan semua variabel independen

Dari hasil yang sudah diperoleh, selanjutnya akan dibentuk model regresi logistik biner yang nantinya model tersebut akan digunakan untuk memprediksi data test. Model yang terbentuk adalah sebagai berikut.

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right]$$
$$= -10.962 - 112.833x_1 + 12.421x_2 + 34.404x_3 + 124.357x_4 + 10.742x_5 - 3.741x_6 + 3.536x_7 + 15.289x_8 + 3.528x_9 - 6.168x_{10}$$

Setelah mendapatkan model regresi logistik binernya, kita akan menghitung keakuratan model tersebut. Keakuratan model tersebut akan dihitung menggunakan confusion matrix. Dengan bantuan RStudio, confusion matrix yang diperoleh dengan data test adalah sebagai berikut.

Tabel 5
Confusion matrix model regresi logistik biner

Confusion Matrix		Predict	
		0	1
Actual	0	68	3
	1	3	39

Sumber: data diolah (2025)

Hasil prediksi data test dengan model regresi logistik biner yang telah dibentuk dengan data train dapat dilihat pada *confusion matrix* tersebut. Dari total 113 sampel, model tersebut berhasil memprediksi 68 sampel tipe kanker jinak (kelas 0) dengan benar dan 39 sampel tipe kanker ganas (kelas 1) dengan benar, sementara terjadi 3 kesalahan prediksi tipe kanker jinak sebagai kanker ganas (*false positive*, FP) dan 3 kesalahan prediksi tipe kanker ganas sebagai kanker jinak (*false negative*, FN). Dari *confusion matrix* tersebut dapat diperoleh nilai akurasi, presisi dan *recall* sebagai berikut.

Nilai akurasi $= \frac{68+39}{68+3+3+39} = 0.9469 \times 100\% = 94.69\%$

Nilai presisi $= \frac{39}{39+3} = 0.9286 \times 100\% = 92.86\%$

Nilai *recall* $= \frac{39}{39+3} = 0.9286 \times 100\% = 92.86\%$

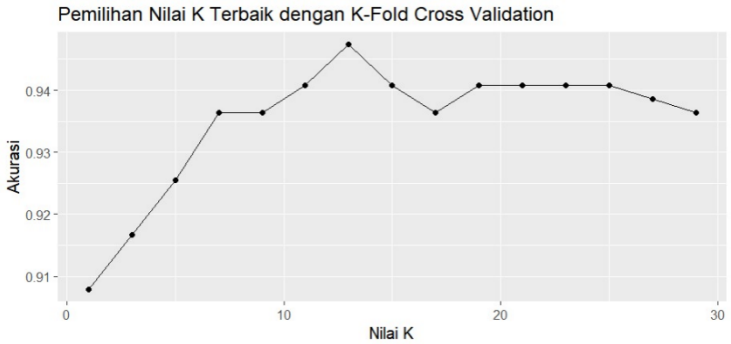
Metode *K-Nearest Neighbors*

Menurut (Prasetyo, 2012) *k-Nearest Neighbor* (k-NN) adalah metode yang melakukan klasifikasi berdasarkan kedekatan lokasi (jarak) suatu data dengan data lain. Nilai k pada K-NN berarti k-data terdekat dari data testing. Dimulai dari melakukan normalisasi data tipe kanker dengan tujuan untuk memperkecil range pada data tersebut, kemudian menghitung jarak data *training* ke data *testing* tipe kanker menggunakan *euclidian distance* dan mengurutkan hasil jarak Euclidian dari yang terkecil ke yang terbesar. Selanjutnya menentukan nilai K dan mencari kelas mayoritas sebanyak nilai K sesudah normalisasi. Penentuan K terbaik akan dilakukan dengan metode *K-Fold Cross Validation*. Dengan bantuan software RStudio untuk mendapatkan K terbaik adalah sebagai berikut:

Tabel 6
Penentuan K dengan metode K-Fold Cross Validation

K	Acuraccy
1	0.9079073
3	0.9166507
5	0.9254419
7	0.9363832
9	0. 9363832
11	0.9407549
13	0.9473483
15	0.9408027
17	0.9364310
19	0.9408027
21	0. 9408027
23	0. 9408027
25	0. 9408027
27	0.9386049
29	0.9364071

Sumber: data diolah (2025)



Gambar 4. Pemilihan Nilai K Terbaik Dengan K-Fold Cross Validation

Hasil menampilkan evaluasi akurasi model *K-Nearest Neighbors* (KNN) untuk berbagai nilai K, yaitu dari 1 hingga 29 dengan kenaikan 2. Dari data yang ditampilkan, akurasi model berkisar antara 90,79% hingga 94,73%, dengan nilai tertinggi dicapai pada K = 13. Maka penerapan model KNN akan dilakukan dengan menggunakan K tetangga sebesar 13.

Hasil prediksi dengan metode KNN dengan K = 13 menunjukkan bahwa model memprediksi kelas 0 sebanyak 71 kali dan kelas 1 sebanyak 42 kali. Berdasarkan *output* diatas didapatkan nilai akurasi sebesar 94,73% yang menandakan bahwa model memiliki performa yang sangat baik dalam memprediksi diagnosa kanker sebesar 94,73%.

Tabel 7
Confusion matrix dengan KNN

Confusion Matrix		Predict	
		0	1
Actual	0	69	2(FP)
	1	3	39(TN)

Sumber: data diolah (2025)

Hasil *Confusion Matrix* dari *model K-Nearest Neighbors* (KNN) menunjukkan bahwa model memiliki performa klasifikasi yang sangat baik. Dari total 113 sampel, model berhasil mengklasifikasikan 69 sampel negatif (kelas 0) dengan benar dan 39 sampel positif (kelas 1) dengan benar, sementara terjadi 2 kesalahan prediksi positif palsu (*false positive*, FP) dan 3 kesalahan negatif palsu (*false negative*, FN). Dari *confusion matrix* tersebut dapat diperoleh nilai akurasi, presisi dan *recall* sebagai berikut.

Nilai akurasi $= \frac{69+39}{69+2+3+39} = 0.9558 \times 100\% = 95{,}58\%$

Nilai presisi $= \frac{39}{40+2} = 0.9523 \times 100\% = 95.23\%$

Nilai *recall* $= \frac{40}{40+2} = 0.9523 \times 100\% = 95.23\%$

Dengan demikian, model memiliki akurasi sebesar 96.46%, yang menunjukkan bahwa sebagian besar prediksi sesuai dengan label sebenarnya. Nilai presisi dan *recall* untuk kelas positif (1) mencapai 95.2%, yang menandakan bahwa model sangat baik dalam mengenali sampel positif dan jarang melakukan kesalahan dalam klasifikasi.

Perbandingan

Setelah memperoleh hasil prediksi dengan model regresi logistik biner dengan model KNN, selanjutnya akan dihitung perbedaan signifikansi akurasi dengan Uji McNemar untuk membandingkan dua model klasifikasi (Logistik vs. KNN) dengan melihat apakah ada perbedaan yang signifikan dalam akurasi mereka pada sampel yang sama.

Tabel 8
Hasil Uji McNemar

	KNN Benar	KNN Salah
Logistik Benar	106	1
Logistik Salah	2	4

Sumber: data diolah (2025)

Hasil uji McNemar menunjukkan bahwa model Regresi Logistik dan KNN memiliki performa yang hampir setara dalam melakukan klasifikasi. Model Regresi Logistik memberikan prediksi yang benar pada 106 kasus, sementara model KNN juga benar pada kasus yang sama. Selain itu, terdapat 1 kasus di mana Regresi Logistik memberikan prediksi yang benar, tetapi KNN salah. Sebaliknya, terdapat 2 kasus di mana Regresi Logistik salah, sedangkan KNN memberikan prediksi yang benar. Kedua model sama-sama mengalami kesalahan pada 4 kasus lainnya. Secara keseluruhan, Regresi Logistik sedikit lebih unggul karena memiliki lebih sedikit kesalahan dibandingkan KNN. Namun, hasil uji McNemar menghasilkan p-value = 1, yang menunjukkan bahwa perbedaan ini tidak signifikan secara statistik, sehingga tidak ada bukti kuat untuk menyimpulkan bahwa salah satu model lebih baik dari yang lain.

Dengan melihat hasil *confusion matrix* pada tabel 2 dan tabel 4, kita dapat membandingkan metode regresi logistik biner dan *k nearest neighbors*. Perbandingan akan dilakukan dengan membandingkan besarnya nilai akurasi, presisi, dan *recall* kedua metode. Perbandingannya dapat dilihat sebagai berikut.

Tabel 9
Perbandingan kedua metode

	Akurasi	Presisi	Recall
Regresi Logistik biner	94.69%	92.86%	92.86%
KNN	95.58%	95.83%	97.18%

Sumber: data diolah (2025)

Akurasi adalah suatu pola tolak ukur untuk memprediksi kelas data dari data yang akan datang. Nilai akurasi menunjukkan seberapa baik model tersebut dalam memprediksi kelas datanya dengan benar. Dari nilai akurasi kedua metode dalam menunjukkan bahwa kedua metode sangat baik dalam memprediksi tipe kanker dengan besar akurasi <90%. Meskipun kedua metode dapat memprediksi dengan sangat baik, diantara kedua metode tersebut, metode KNN lebih baik karena memiliki nilai akurasi yang lebih besar.

Presisi dapat diartikan sebagai rasio antara jumlah prediksi positif yang benar dengan total jumlah prediksi positif. Nilai presisi menunjukkan seberapa baik model memprediksi nilai positif yang benar-benar positif. Dari nilai presisi pada kedua metode menunjukkan bahwa kedua metode sangat baik dalam memprediksi tipe kanker dengan besar presisi <90%. Nilai presisi yang tinggi ini menunjukkan bahwa model jarang membuat kesalahan saat memprediksi nilai positif (*false alarm*). Meskipun kedua metode memiliki nilai presisi yang sangat baik, diantara kedua metode tersebut, metode KNN lebih baik karena memiliki nilai presisi yang lebih baik.

Recall dapat diartikan sebagai proporsi prediksi benar positif dibandingkan dengan keseluruhan data yang sebenarnya positif. Nilai *recall* menunjukkan seberapa baik model memprediksi nilai yang benar-benar positif. Dari nilai *recall* pada kedua metode menunjukkan bahwa kedua metode sangat baik dalam memprediksi tipe kanker dengan besar recall <90%. Nilai *recall* yang tinggi ini menunjukkan bahwa model jarang melewatkan nilai yang benar-benar positif. Meskipun kedua metode memiliki nilai *recall* yang sangat baik, diantara kedua metode tersebut, metode KNN lebih baik karena memiliki nilai *recall* yang lebih besar.

PENUTUP

Simpulan

Berdasarkan hasil penelitian, perbandingan antara metode Regresi Logistik Biner dan *K-Nearest Neighbors* (KNN) dalam mengklasifikasikan kanker menunjukkan bahwa kedua metode memiliki performa yang baik. Namun, metode KNN dengan K=13 menunjukkan akurasi yang lebih tinggi dibandingkan Regresi Logistik Biner, dengan akurasi 95,58%, presisi 95,83%, dan recall 97,18%. Hasil uji McNemar menunjukkan bahwa perbedaan antara kedua model tidak signifikan secara statistik, yang mengindikasikan bahwa kedua metode memiliki performa yang hampir setara dalam klasifikasi kanker.

Meskipun KNN menunjukkan keunggulan dalam hal akurasi, penting untuk mempertimbangkan aspek interpretabilitas dan efisiensi saat menerapkannya di dunia medis. Oleh karena itu, pendekatan hibrida yang menggabungkan regresi logistik sebagai model awal dengan KNN untuk analisis lanjutan dapat menjadi solusi yang lebih seimbang. Dengan pemilihan model yang tepat, teknologi kecerdasan buatan dapat memberikan dukungan kepada dokter dalam membuat diagnosis kanker yang lebih akurat, cepat, dan dapat diandalkan, yang pada gilirannya akan berkontribusi pada peningkatan kualitas layanan kesehatan dan keselamatan pasien.

DAFTAR PUSTAKA

Akri, Y. J. (2024). Potensi Tanaman Obat Indonesia Atasi Tumor Dan Kanker Tinjauan Sistematis. Jurnal Ilmu Kesehatan, 280-287.

Alfiani, D., Widayanti, & Putri, M. (2024). Literature Study: Obesitas Sebagai Faktor Risiko pada Kanker Payudara Triple Negative. Bandung Conference Series: Medical Science, 326-329.

Ali, M. M., Hariyati, Pratiwi, M. Y., & Afifah, S. (2022). Metodologi Penelitian Kuantitatif Dan Penenrapannya Dalam Penelitian. Education Journal, 1-6.

Alkhussayid, M. D., & Ferdiansyah. (2022). Implementasi Algoritma K-Nearest Neighbors Pada Penentuan Jurusan Siswa. Jurnal Sistem Komputer dan Informatika, 25-36.

Azhari, M., Situmorang, Z., & Rosnelly, R. (2021). Perbandingan Akurasi, Recall, dan Presisi Klasifikasi pada Algoritma C4.5, Random Forest, SVM dan Naive Bayes. Jurnal Media Informatika Budidarma, 640-651.

Azzahra, N. D., Ambarwati, A., Desiani, A., Maiyanti, S. I., & Ramayani, I. (2024). Perbandingan Algoritma K-Nearest Nighbor Dan Logistic Regression Dalam Klasifikasi Penyakit Kanker Servik. Jurnal Energy, 1-8.

Bakriansyah, M., Subair, A., & Setiawan, W. (2025). Penerapan Data Mining Untuk Memprediksi Penjualan Produk Terlaris UD Timbul Jaya Menggunakan Metode K-Nearst Neighbor. J-CEKI: Jurnal Cendikia Ilmiah, 327-337.

Cahyana, C. W., & Nurlayli, A. (2023). Analisis Performa Logistic Regression, Naive Bayes, dan Random Forest sebagai Algoritma Pendeteksi Kanker Payudara. Information System and Emerging Technology Journal, 51-64.

Cahyanti, D., Rahmayani, A., & Husniar, S. A. (2020). Analisis performa metode knn pada Datset pasien Pengidap Kanker Payudara. Indonesian Journal of Data and Science, 39-43.

Chtouki, K., Rhanoui, M., Mikram, M., & Yousfi, S. (2023). Supervised Machine Learning For Breast Cancer Risk Factors Analysis And Survival Prediction. A Preprint, 1-10.

Fahrurozi, & Wasilah. (2023). Deteksi Dini Kanker Payudara Menggunakan Algoritma K-Nearest Neighbor (KNN) Dan Decision Tree C-45. Teknika, 427-434.

Fajrin, J., Pathurahman, & Pratama, L. G. (2016). Aplikasi Metode Analysis of Variance (ANOVA) untuk Mengkaji Pengaruh Penambahan Silica Fume Terhadap Sifat Fisik dan Mekanik Mortar. Jurnal Rekayasa Sipil, 11-23.

Fasnuari, H. A., Yuana, H., & Chulkamdi, M. T. (2022). Penerapan Algoritma K-Nearest Neighbor (K-NN) Untuk Klasifikasi Penyakit Diabetes Melitus. ANTIVIRUS: Jurnal Ilmiah Teknik Informatika, 133-142.

Kainama, E. C., Palit, E. I., & Hutabarat, I. M. (2022). Analisis Pengaruh Pandemik COVID-19 Terhadap Laju Pertumbuhan Ekonomi Provinsi Papua Dengan Menggunakan Uji McNemar dan Uji Wilcoxon. Cendera Wasih Jurnal of Statistics and Data Science, 40-48.

Martha, S., Andani, W., & Rizki, S. W. (2022). Perbandingan Metode K-Nearest Neighbor, Regresi Logistik Biner, dan Pohon Klasifikasi pada Analisis Kelayakan Pemberian Kredit. EULER: Jurnal Ilmiah Matematika, Sains dan Teknologi, 262-273.

Prasetyo, E. (2012). K-Support Vector Nearest Neighbor Untuk Klasifikasi Berbasis K-NN. SESINDO - Jurusan Sistem Informasi ITS, 245-250.

Pratama, H. R., Naila, I., & Faradita, M. N. (2024). Analisis Keterampilan Kolaborasi Siswa Sekolah Dasar Menggunakan Media Diorama Pada Pembelajaran Materi Ekosistem. Pendas: Jurnal Ilmiah Pendidikan Dasar, 927-937.

Putri, A. A., Titaley, J., & Salaki, D. T. (2022). Model Regresi Logistik Biner Kecenderungan Gejala Maag Pada Mahasiswa Jurusan Matematika FMIPA UNSRAT. Jurnal Matematika dan Aplikasi, 38-43.

Salam, R. R., Jamil, M. F., Ibrahim, Y., Rahmadden, Soni, & Herianto. (2023). Analisis Sentimen Terhadap Bantuan Langsung Tunai (BLT) Bahan Bakar Minyak (BBM)

Menggunakan Support Vector Machine. MALCOM: Indonesia Journal of Machine Learning and Computer Science, 27-35.

Solikin, I. (2021). Teknik Data Mining untuk Prediksi Kanker Payudara yang Efisien. Fidelity: Jurnal Teknik Elekrto, 63-67.

Sugiyono, & Hartinah, S. S. (2024). Pemodelan Data Mining Transaksi Penjualan Menggunakan Algoritma Apriori (Studi Kasus: Kedai Ngodeng & Smoothies). Jurnal Indonesia: Manajemen Informatika dan Komunikasi, 3080-3098.

Sujana, S., Juwita, A. R., Rahmat, & Faisal, S. (2024). Penerapan Metode Regresi Logistik Untuk Memprediksi Peristiwa Biner Pasien Pasca Operasi Kanker Payudara. Journal of Information System Research, 1095-1101.

Widyaningsih, Y., Arum, G. P., & Prawira, K. (2021). Aplikasi K-Fold Cross Validation Dalam Penentuan Model Regresi Binomial Negatif Terbaik. Barekeng: Jurnal Ilmu Matematika dan Terapan, 315-322.