Terbit online pada laman web jurnal: **https://ejournal.upm.ac.id/index.php/energy/**

Jurnal Energy

(Jurnal Ilmiah Ilmu-ilmu Teknik)

Vol. 14 No. 1 Edisi Mei 2024 ISSN: 2088-4591 * E-ISSN: 2962-2565

Perbandingan Algoritma K-Nearest Neighbor Dan Logistic Regression Dalam Klasifikasi Penyakit Kanker Serviks

Comparison of K-Nearest Neighbor Algorithm and Logistic Regression in the Classification of Cervical Cancer Disease

Nur Devita Azzahra[1], Ambarwati[2], Anita Desiani[3]*, Sri Indra Maiyanti[4], Indri Ramayanti[5]
[1,2,3,4]Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Sriwijaya, Indralaya
[5]Fakultas Kedokteran, Universitas Muhammadiyah Palembang, Palembang
[1]nurdevitaazzahra@gmail.com, [2]aw9075541@gmail.com, [3]anita_desiani@unsri.ac.id*,
[4]sri_indra_maiyanti@mipa.unsri.ac.id, [5]indri_ramayanti@umpalembang.ac.id

Abstract

Cervical cancer is a serious problem in women's health globally, with high incidence and mortality rates. Early detection is crucial due to its slow progression and often asymptomatic nature in the early stages. One method used to detect cervical cancer is classification through a data mining approach. This study aims to apply the K-Nearest Neighbor (K-NN) and Logistic Regression algorithms in the classification of early detection of cervical cancer, using the percentage spilt and k-fold cross validation testing techniques. The aim is to compare and determine the most appropriate method for cervical cancer classification. Evaluation is done using precision, recall, and accuracy parameters. The results showed that the test results using k-fold cross validation were able to work better than the percentage spilt. Based on algorithm performance, Logistic Regression in its use can work well compared to K-NN, with precision, recall and accuracy values of 83%, 82.5% and 96%, respectively. Thus, it can be concluded that Logistic Regression with k-fold cross validation testing technique is more effective in classifying cervical cancer.

Keywords: *Cervical Cancer, K-Nearest Neighbor, Logistic Regression, Percentage Split, K-Fold Cross Validation*

Abstrak

Kanker serviks merupakan masalah serius dalam kesehatan perempuan secara global, dengan tingkat kejadian dan kematian yang tinggi. Deteksi dini menjadi krusial karena perkembangannya yang lambat dan sering tanpa gejala pada tahap awal. Salah satu metode yang digunakan untuk mendeteksi kanker serviks adalah klasifikasi melalui pendekatan data mining. Penelitian ini bertujuan untuk menerapkan algoritma K-Nearest Neighbor (K-NN) dan Logistic Regression dalam klasifikasi deteksi dini kanker serviks, dengan mengguunakan teknik pengujian percentage spilt dan k-fold cross validation. Tujuannya adalah untuk membandingkan dan menentukan metode yang paling tepat untuk klasifikasi kanker serviks. Evaluasi dilakukan dengan menggunakan parameter presisi, recall, dan akurasi. Hasil penelitian menunjukkan bahwa hasil pengujian dengan menggunakan k-fold cross validation mampu bekerja lebih baik dibandingkan dengan percentage spilt.. Berdasarkan kinerja algoritma, Logistic Regression dalam penggunaannya dapat bekerja dengan baik dibandingkan dengan K-NN, dengan nilai presisi, recall dan akurasi masing-masing sebesar 83%, 82,5% dan 96%. Dengan demikian, dapat disimpulkan bahwa Logistic Regression dengan teknik pengujian k-fold cross validation lebih efektis dalam melakukan klasifikasi penyakit kanker serviks.

Kata kunci: Kanker serviks, *K-Nearest Neighbor, Logistic Regression, Percentage Split, K-Fold Cross Validation*

1. Pendahuluan

Kanker serviks atau kanker leher rahim, adalah jenis kanker yang disebabkan oleh *human papilloma virus* (HPV) dan memiliki tingkat keterikatan yang sangat tinggi, yaitu sebesar 99,7% [1], [2]. Ini adalah jenis kanker wanita paling umum kedua pada wanita berusia 15 tahun hingga 44 tahun dan menjadi penyebab utama keempat kanker wanita didunia dengan sekitar 604.127 kasus kanker serviks baru didiagnosis setiap tahun [3]. Perkembangan kanker serviks cenderung berlangsung secara lambat dan awalnya tidak menimbulkan gejala yang jelas, membuat deteksi pada tahap dini relatif sulit [4]. Akibatnya, insiden kematian akibat kanker serviks meningkat karena kanker sering terdeteksi pada tahap yang lebih parah.

Deteksi dini kanker serviks merupakan langkah penting untuk meningkatkan prognosis dan kesempatan penyembuhan. Dalam penelitian ini, deteksi dini dilakukan menggunakan pendekatan data mining. Data mining

DOI: 10.51747/energy.v14i1.1843

adalah proses menemukaan informasi tersembunyi dalam kumpulan data besar dan mengorganisasi informasi tersebut menjadi struktur yang dapat digunakan dengan mudah [5]. Banyak penelitian sebelumnya yang telah menggunakan teknik data mining dalam klasifikasi kanker serviks, termasuk diantaranya penelitian oleh Tembusi et al. [6], yang mengimplementasikan K-NN dengan metode pengujian K-fold cross validtation dan memperoleh tingkat akurasi sebesar 95%.

Beberapa metode data mining yang dapat diaplikasikan dalam klasifikasi kanker serviks termasuk didalamnya algoritma *K-Nearest Neighbor* (K-NN) dan *Logistic Regression*. Algoritma K-NN adalah algoritma generalisasi untuk aturan tetangga terdekat, offset induktifnya adalah label kelas k-sampel dengan label kelas yang akan diuji paling mirip dengan yang terdekat [7]. Kelebihan K-NN meliputi kemampuannya dalam menangani data dengan dimensi yang tinggi dan mampu mengatasi data outlier [7], [8]. Sebelumnya, banyak penelitian yang telah menggunakan K-NN dalam klasifikasi salah satunya penelitian yang dilakukan Kurniawan & Barokah [9], mereka mengaplikasikan metode K-NN untuk menentukan pengajuan kartu kredit dengan menggunakan *data training* dan *data testing* masing-masing berjumlah 250 dan 100 data, dan berhasil mencapai tingkat akurasi sebesar 93%. Namun, K-NN juga memiliki beberapa kelemahan, seperti membutuhkan pertimbangan dalam memilih nilai k (jumlah tetangga terdekat) dan memiliki kompleksitas komputasi yang tinggi [10], [11].

Logistic Regression adalah sebuah metode pembelajaran terbimbing yang digunakan untuk masalah regresi dan klasifikasi. Metode ini menggunakan fungsi logistik atau sigmoid untuk memodelkan hubungan antara atribut independen dengan probabilitas klasifikasi data kategorikal [12]. Salah satu kelebihan utama dari *Logistic Regression* adalah kemudahannya dalam penggunaan, efisiensi waktu, biaya, serta interpretasi yang mudah oleh praktisi dan peneliti [13]. Metode ini juga memiliki keunggulan lainnya, yaitu komputasi yang lebih cepat dan kemampuan dalam mengatasi masalah dengan data berdimensi tinggi [14]. Beberapa penelitian telah mengaplikasikan *Logistic Regression* dalam klasifikasi. Sebagai contoh, Tyasnurita & Pamungkas [15] menggunakan *Logistic Regression* untuk diagnosis diabetik retinopati dengan tiga skenario pemodelan. Hasil pemodelan menunjukkan bahwa skenario pemodelan pertama memiliki performa yang lebih baik dengan rata-rata nilai akurasi pada data pelatihan, data pengujian, dan data validasi masing-masing sebesar 76,10%, 74,28%, dan 80,17%. Meskipun memiliki kelebihan, *Logistic Regression* juga memiliki beberapa kelemahan. Sensitivitas terhadap atribut independen dan dependen, serta adanya data pencilan atau outlier, dapat mempengaruhi hasil model. Selain itu, *Logistic Regression* memiliki batasan dalam memprediksi hasil berkelanjutan, dan untuk hasil yang lebih stabil, diperlukan ukuran sampel yang besar [16].

Pada penelitian ini akan digunakan algoritma *K-Nearest Neighbor* (K-NN) dan *Logistic Regression* dalam klasifikasi deteksi dini kanker serviks. Kedua algoritma ini akan digunakan, karena mudah dan sederhana untuk diimplementasikan [17]. Penggunaan kedua metode ini akan memberikan pemahaman lebih mendalam tentang potensi penggunaan data mining dalam mendukung deteksi dini kanker serviks. Penelitian ini dilakukan dengan teknik pengujian *percentage split* dan *k-fold cross validation*. Dalam mengukur kinerja K-NN dan *Logistic Regression* dilihat berdasarkan nilai akurasi, presisi dan recall sehingga didapatkan algoritma terbaik dalam melakukan klasifikasi kanker serviks.

2. Metodologi

2.1. Pengumpulan Data

Data klasifikasi kanker serviks yang digunakan dalam penelitian ini diperoleh dari situs web Kaggle (https://www.kaggle.com/datasets/loveall/cervical-cancer-risk-classification) dimana pada data klasifikasi kanker serviks ini mengarah ke pemeriksaan biopsi. Jumlah dari keseluruhan data yaitu sebanyak 858 dan 36 atribut. Atribut beserta tipe data pada dataset ini akan ditunjukkan pada tabel 1.

Tabel 1. Atribut pada dataset

Atribut	Range	Tipe Data
Umur	13-84 tahun	Numerik
Jumlah pasangan seksual	1-28 orang	Numerik
Usia awal melakukan hubungan seksual	10-32 tahun	Numerik
Jumlah kehamilan	0-11 kehamilan	Numerik
Merokok	0: Tidak, 1: Ya	Kategori
Merokok (tahun)	0-37 tahun	Numerik

DOI: 10.51747/energy.v14i1.1843

Merokok (bungkus/tahun)	0-37 bungkus/tahun	Numerik
Kontrasepsi hormonal	0: Tidak, 1: Ya	Kategori
Kontrasepsi hormonal (tahun)	0-30 tahun	Numerik
IUD (Intrauterine Device)	0: Tidak, 1: Ya	Kategori
IUD (tahun)	0-19 tahun	Numerik
PMS (Penyakit Menular Seksual)	0: Tidak, 1: Ya	Kategori
PMS (jumlah)	0-4 penyakit	Numerik
PMS:Kondilomatosis	0: Tidak, 1: Ya	Kategori
PMS: Kondilomatosis Serviks	0: Tidak, 1: Ya	Kategori
PMS: Kondilomatosis Vagina	0: Tidak, 1: Ya	Kategori
PMS: Kondilomatosis vulva-perineal	0: Tidak, 1: Ya	Kategori
PMS: sifilis	0: Tidak, 1: Ya	Kategori
PMS: penyakit radang panggul	0: Tidak, 1: Ya	Kategori
PMS: Herpes genital	0: Tidak, 1: Ya	Kategori
PMS: Moluskum kontagiosum	0: Tidak, 1: Ya	Kategori
PMS:AIDS	0: Tidak, 1: Ya	Kategori
PMS:HIV	0: Tidak, 1: Ya	Kategori
PMS:Hepatitis B	0: Tidak, 1: Ya	Kategori
PMS:HPV	0: Tidak, 1: Ya	Kategori
PMS: Jumlah diagnosis	0-3 diagnosis	Numerik
PMS: Waktu sejak diagnosis pertama	1-22 tahun	Numerik
PMS: Waktu sejak diagnosis terakhir	1-22 tahun	Numerik
Dx:Kanker	0: Tidak, 1: Ya	Kategori
Dx:CIN	0: Tidak, 1: Ya	Kategori
Dx:HPV	0: Tidak, 1: Ya	Kategori
Dx (Hasil diagnosis akhir)	0: Tidak, 1: Ya	Kategori
Hinselmann	0: Tidak, 1: Ya	Kategori
Schiller	0: Tidak, 1: Ya	Kategori
Citology	0: Tidak, 1: Ya	Kategori
Biopsy	0: Tidak, 1: Ya	Kategori

2.2. Persiapan Data

1) *Missing Data* (Data Hilang)

Dari tabel 1, akan ada 2 atribut yang ditiadakan yaitu atribut PMS: Waktu sejak diagnosis pertama dan PMS: Waktu sejak diagnosis terakhir. Sehingga total atribut yang digunakan yaitu 34 atribut. Kemudian, akan dilakukan penghapusan baris pada data yang bernilai kosong. Setelah menghapus baris yang bernilai kosong, tersisa 668 data yang valid dan siap untuk digunakan.

2) Normalisasi Data

Pada dataset klasifikasi kanker serviks terdapat atribut yang rangenya cukup jauh yaitu umur. Pada atribut dengan range yang jauh perlu dilakukan metode normalisasi. Metode yang digunakan untuk perrmasalahan nilai yang terlampau jauh, yaitu *Min-Max Normalization*. Rumus umum *Min-Max Normalization* ditunjukkan oleh persamaan (1)

$$Normalized(x) = \frac{minRange\ x + (x\ minValue)(maxRange - minRange)}{maxValue - minValue} \quad (1)$$

3) Pengujian Data

DOI: 10.51747/energy.v14i1.1843

Dataset akan dimodelkan dengan dua teknik pengujian, yaitu *Percentage Split* dan *K-fold cross validation*. Pada *Percentage Split*, data dibagi menjadi 80% sebagai data latih dan 20% sebagai data uji. Sedangkan pada *K-fold cross validation*, data dibagi ke dalam 10 bagian (k=10) dan secara bergiliran masing-masing bagian akan menjadi data latih maupun data uji.

2.3. Penerapan *K-Nearest Neighbor* (K-NN)

Tujuan dari metode ini ialah menggunakan data sampel dan atribut untuk mengklasifikasikan objek baru [24]. Kelas hasil klasifikasi ditentukan berdasarkan kelas yang paling banyak muncul pada algoritma ini. Proses perhitungan algoritma K-NN yaitu sebagai berikut:

1) Menentukan jumlah tetangga terdekat (K).
2) Menghitung jarak antara setiap sampel data latih dan data yang diuji berdasarkan persamaan (2) berikut [25]:

$$d(x,y) = \sqrt{\sum_{i=0}^{n}(x_i - y_i)^2} \quad (2)$$

Keterangan :
d = jarak antara ruang atribut
x = data latih
y = data uji
i = indeks atau variabel data
n = jumlah data latih

3) Urutkan jarak dari yang terkecil ke terbesar dan pilih himpunan K yang terkecil dari dataset terkecil.
4) Mengumpulkan label kelas Y (Klasifikasi *K-Nearest Neighbor*).
5) Prediksi nilai *query instance* yang telah didapat sesuai dengan *K-Nearest Neighbor* yang paling banyak.

2.4. Penerapan Algoritma *Logistic Regression*

Algoritma *Logistic Regression* merupakan suatu metode regresi yang mempelajari keterkaitan antara atribut bebas dan atribut terikat yang bersifat biner. Atribut bebas umumnya memiliki dua nilai kemungkinan (0 atau 1) dan sering kali diinterpretasikan sebagai indikator keberhasilan atau kegagalan[26]. Adapun langkah penyelesaian *Logistic Regression* adalah sebagai berikut[27] :

1) Menentukan atribut bebas dan atribut terikat yang terdapat didalam dataset
2) Menentukan nilai acak $\beta_1, \beta_2, \dots, \beta_p$ sebagai asumsi dasar dalam menentukan likelihood
3) Menghitung nilai Y prediksi menggunakan persamaan (3) :
$$Y = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} \quad (3)$$
dimana β_0 adalah *intercept* , $\beta_1, \dots, \beta_p$ adalah parameter regresi logistik, $x_1, \dots, x_p$ adalah nilai peubah bebas, p adalah jumlah atribut bebas dan i adalah jumlah baris dalam dataset.
4) Menghitung nilai $\pi(x)$ dengan menggunakan persamaan (4) :
$$\pi(x) = \frac{\exp(Y)}{1+\exp(Y)} \quad (4)$$
dimana $\pi(x)$ adalah peluang kejadian sukses dengan nilai probabilitas $0 \leq \pi(x) \leq 1$
5) Menghitung nilai maksimum log likelihood dengan persamaan (5) :
$$maks = Y_i \times \ln\big(\pi(x) + (1 - Y_i)\big) \times \ln\big(1 - \pi(x)\big) \quad (5)$$
dimana Y_i adalah nilai atribut bebas berdasarkan kelas yang telah ditentukan.

2.5. Evaluasi

Saat mengevaluasi model klasifikasi, kita akan menghitung jumlah objek pengujian yang diprediksi dengan benar dan tidak benar. Hasil dari perhitungan ini akan disusun didalam tabel *confusion matrix*. *Confusion matrix* merepresentasi visualisasi dari kinerja algoritma klasifikasi yang menggunakan data dalam bentuk matriks untuk membandingkan hasil klasifikasi, termasuk *True Positive* (TP), *False Positive* (FP), *True Negative* (TN), dan *False Negative* (FN). Pada tabel 2 akan diperlihatkan *confusion matrix* untuk dua kelas.

Tabel 2. *Confusion Matrix*

Kelas		***Nilai Aktual***	
		Positif	Negatif
Nilai Prediksi	Positif	TP	FN
	Negatif	FP	TN

Dari hasil *confusion matrix* dapat diukur kinerja metode yang digunakan dengan menggunakan nilai akurasi, presisi, dan *recall*.

DOI: 10.51747/energy.v14i1.1843

a. Akurasi

Jumlah prediksi benar atau salah yang dibuat oleh model berdasarkan kumpulan data disebut sebagai tingkat akurasi. Untuk menghitung nilai akurasi digunakan persamaan (6) berikut [6]:

$$Akurasi = \frac{TP+TN}{TP+TN+FN+FP} \quad (6)$$

b. Presisi

Presisi merupakan rasio prediksi benar positif terhadap keseluruhan hasil yang diprediksi benar positif. Untuk menghitung nilai presisi digunakan persamaan (7) berikut [6]:

$$Presisi = \frac{TP}{TP+FP} \quad (7)$$

c. *Recall*

Recall merupakan rasio prediksi benar positif yang dibandingkan dengan keseluruhan data yang benar positif. Persamaan (8) digunakan untuk menghitung nilai *recall* [6]:

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

3. Hasil dan Pembahasan

3.1. Algoritma *K-Nearest Neighbor* (K-NN)

Untuk mengimplementasikan algoritma K-NN, akan dilakukan percobaan dengan menggunakan 5 nilai k yang berbeda. Tujuan dari percobaan ini ialah untuk mendapatkan hasil pemodelan yang terbaik. Pada tabel 3 dapat dilihat hasil pemodelan algoritma K-NN dengan menggunakan 5 nilai k, yaitu 1, 2, 3, 4, dan 5.

Tabel 3. Hasil Pemodelan Algoritma K-NN

K	Target Kelas	*Percentage Split*			*K-Fold Cross Validation*		
		Presisi	***Recall***	**Akurasi**	**Presisi**	***Recall***	**Akurasi**
1	NO	92%	96%	89%	94%	96%	90%
	YES	17%	9%		15%	9%	
2	NO	92%	98%	90%	93%	99%	93%
	YES	0%	0%		14%	2%	
3	NO	92%	98%	91%	94%	98%	92%
	YES	33%	9%		21%	7%	
4	NO	92%	99%	91%	93%	100%	93%
	YES	0%	0%		100%	2%	
5	NO	92%	99%	91%	93%	99%	93%
	YES	0%	0%		17%	2%	

Berdasarkan Tabel 3 akurasi keseluruhan sistem, penggunaan K=4 memiliki akurasi keseluruhan yang terbesar dan lebih baik jika dibandingkan dengan K yang lain yaitu sebesar 91% pada *percentage split* dan 93% pada *k-fold cross validation*. Hasil pemodelan menggunakan algoritma K-NN tersebut menunjukkan bahwa algoritma ini cukup baik dalam mengklasifikasikan kanker serviks pada dataset *Kaggle*. Hal ini karena nilai akurasi yang diperoleh pada kedua pemodelan tersebut lebih dari 90%. Hasil klasifikasi algoritma K-NN menggunakan pemodelan *percentage split* dan *k-fold cross validation* akan ditampilkan dalam tabel *confusion matrix*. Pada tabel 4 dapat diperhatikan nilai confusion matrix yang memiliki nilai presisi, recall, dan akurasi terbesar yaitu pada K=4.

Tabel 4. *Confusion Matrix* Algoritma K-NN dengan K=4

Kelas	*Percentage Split*		*K-Fold Cross Validation*	
	YES(Positif Cervical Cancer)	NO(Negatif Cervical Cancer)	YES(Positif Cervical Cancer)	NO(Negatif Cervical Cancer)
YES(Positif Cervical Cancer)	0	11	1	44
NO(Negatif Cervical Cancer)	1	122	0	623

Berdasarkan Tabel 4, terdapat 134 data yang diprediksi menggunakan *percentage split* dan 668 data yang diprediksi menggunakan *k-fold cross validation*. Pada *percentage split*, tidak ada data yang diprediksi positif

DOI: 10.51747/energy.v14i1.1843

secara benar, sedangkan 11 data positif seharusnya diprediksi positif tetapi diprediksi negatif. Sementara itu, 122 data negatif diprediksi secara benar dan 1 data negatif seharusnya diprediksi negatif tetapi diprediksi positif. Pada *k-fold cross validation*, hanya 1 data positif yang diprediksi dengan benar dan 44 data positif seharusnya diprediksi positif tetapi diprediksi negatif. Sedangkan 623 data negatif diprediksi dengan benar dan tidak ada data negatif yang diprediksi sebagai positif. Selanjutnya, nilai presisi, recall dan akurasi *percentage split* dan *k-fold cross validation* pada K-NN dapat dilihat pada Tabel 5.

Tabel 5. Perbandingan *Percentage Split* dan *K-fold Cross Validation* pada K-NN

Teknik Pengujian	**Presisi (%)**	**Recall (%)**	**Akurasi(%)**
Percentage Split	46	49.5	91
K-fold cross validation	96.5	51	93

Berdasarkan Tabel 5, dapat dilihat bahwa pengujian dengan *k-fold cross validation* menghasilkan performa yang lebih baik dibandingkan dengan *percentage split*. Hal ini tercermin dari nilai presisi, *recall*, dan akurasi *k-fold cross validation* yang lebih tinggi daripada *percentage split*. Secara spesifik, nilai presisi *k-fold cross validation* sebesar 50,5% jauh lebih tinggi dibandingkan *percentage split*. Nilai *recall k-fold cross validation* juga 1,5% lebih baik daripada *recall percentage split*. Demikian pula nilai akurasi *k-fold cross validation* yang lebih unggul jika dibandingkan dengan *percentage split*.

3.2. Algoritma *Logistic Regression*

Hasil klasifikasi algoritma *Logistic Regression* pada penyakit kanker serviks juga menggunakan teknik pengujian *percentage split* dan *K-fold cross validation*, dimana tingkat keberhasilan model yang diperoleh akan digunakan sebagai prediksi dengan *confusion matrix* dan akan ditampilkan pada Tabel 6.

Tabel 6. *Confusion Matrix* Algoritma *Logistic Regression*

Kelas	**Percentage Split**		**K-Fold Cross Validation**	
	YES(Positif Cervical Cancer)	NO(Negatif Cervical Cancer)	YES(Positif Cervical Cancer)	NO(Negatif Cervical Cancer)
YES(Positif Cervical Cancer)	6	5	30	15
NO(Negatif Cervical Cancer)	3	120	14	609

Berdasarkan Tabel 6, pada teknik pengujian *percentage split* terdapat 6 data positif yang diprediksi dengan benar dan 5 data positif yang seharusnya diprediksi positif tetapi diprediksi negatif. Pada kondisi negatif, 120 data diprediksi dengan benar dan 3 data negatif yang seharusnya diprediksi negatif tetapi diprediksi positif. Pada *k-fold cross validation*, terdapat 6 data positif yang diprediksi dengan benar dan 15 data positif yang seharusnya diprediksi positif tetapi diprediksi negatif. Pada kondisi negatif, 609 data diprediksi dengan benar dan 14 data negatif yang seharusnya diprediksi negatif tetapi diprediksi positif. Dari *confusion matrix* menggunakan *percentage split* dan *k-fold cross validation* tersebut, dapat diketahui nilai presisi, *recall*, dan akurasi dari masing-masing kondisi yang disajikan pada Tabel 7.

Tabel 7. Perbandingan *Percentage Split* dan *K-fold Cross Validation* pada *Logistic Regression*

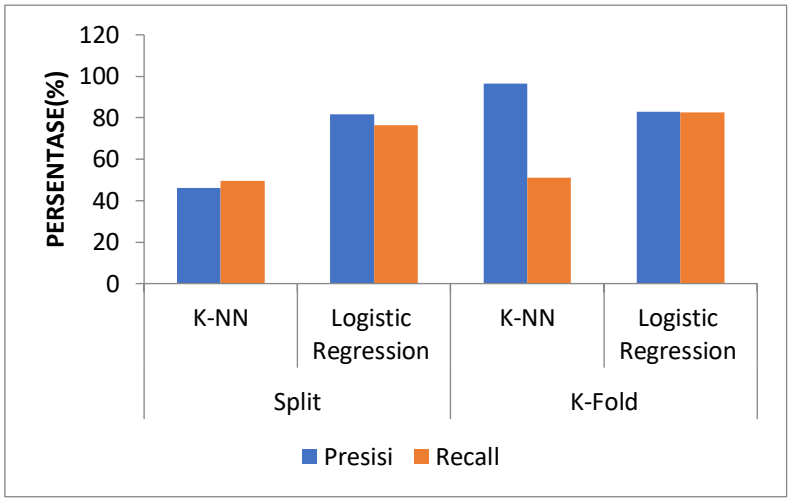
Teknik Pengujian	**Presisi (%)**	**Recall (%)**	**Akurasi(%)**
Percentage Split	81.5	76.5	94
K-fold cross validation	83	82.5	96

Berdasarkan Tabel 7, dapat diketahui bahwa kinerja pengujian *k-fold cross validation* lebih baik daripada *percentage split*. Hal ini terlihat dari nilai presisi, *recall*, dan akurasi hasil *k-fold cross validation* yang lebih tinggi dibandingkan hasil *percentage split*. Secara rinci, presisi *k-fold cross validation* pada *Logistic Regression* 1,5% lebih tinggi dibandingkan presisi *percentage split*. *Recall k-fold cross validation* juga menghasilkan peningkatan sebesar 6% dibandingkan *recall percentage split*. Selanjutnya, akurasi tertinggi 96% dihasilkan oleh *k-fold cross validation*, melampaui akurasi *percentage split*. Dengan demikian, secara keseluruhan kinerja *k-fold cross validation* lebih unggul dibandingkan *percentage split* berdasarkan nilai evaluasi presisi, *recall*, dan akurasi.

DOI: 10.51747/energy.v14i1.1843

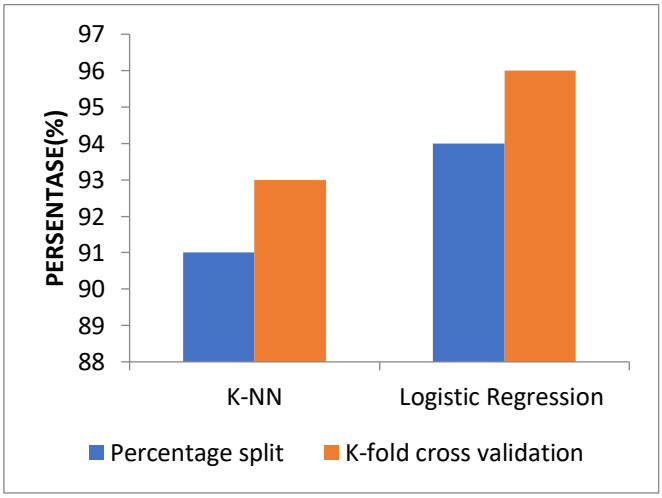
3.3. Perbandingan Hasil Kedua Algoritma

Hasil dari algoritma K-NN dan *Logistic Regression* dengan menggunakan *percentage split* menunjukkan bahwa kedua metode memiliki kinerja yang baik dalam melakukan prediksi penyakit serviks. Nilai rata-rata presisi dan recall seluruh kelas dengan menggunakan *percentage split* dan *K-fold cross validation* ditampilkan pada Gambar 1.



Gambar 1. Rata-Rata Presisi dan Recall Seluruh Kelas pada Kedua Algoritma dengan *Percentage Split* dan *K-Fold Cross Validation*

Pada Gambar 1 dengan menggunakan teknik pengujian *percentage split*, diperoleh nilai rata-rata presisi dan recall untuk K-NN masing-masing sebesar 46% dan 49,5% dan pada *Logistic Regression* memperoleh masing-masing nilai sebesar 81,5% dan 76,5%. Adapun dengan menggunakan teknik pengujian *K-fold cross validation*, diperoleh nilai rata-rata presisi untuk K-NN adalah sebesar 96,5% dan *Logistic Regression* adalah sebesar 83%. Namun, untuk nilai rata-rata recall *Logistic Regression* memperoleh nilai yang lebih besar dibandingkan dengan K-NN dengan masing-masing nilai yang diperoleh sebesar 82,5% dan 51%. Setelah mengetahui nilai rata-rata presisi dan recall dari seluruh kelas, nilai akurasi akan ditampilkan pada Gambar 2.



Gambar 2. Nilai Akurasi pada Kedua Algoritma dengan *Percentage Split* dan *K-Fold Cross Validation*

Pada Gambar 2 baik dengan menggunakan teknik pengujian *percentage split* maupun *K-fold cross validation*, *Logistic Regression* memiliki nilai yang lebih tinggi dibandingkan dengan K-NN. K-NN menghasilkan nilai akurasi sebesar 91% dengan teknik *percentage split* dan 93% dengan *K-fold cross validation*. *Logistic Regression* memperoleh nilai akurasi 94% dengan *percentage split* dan pada *K-fold cross validation* akurasi yang diperoleh sebesar 96%.

4. Kesimpulan

Hasil pengujian menggunakan algoritma K-NN menunjukkan bahwa akurasi terbaik diperoleh dengan menggunakan pemodelan *K-fold cross validation* yaitu diatas 90%. Sementara itu, pada algoritma *Logistic Regression* diperoleh tingkat akurasi yang sama yaitu diatas 90% baik pada pemodelan *percentage split* maupun *K-fold cross validation.* Artinya baik algoritma K-NN maupun *Logistic Regression* dapat bekerja dengan cukup baik dan paling baik menggunakan pemodelan *K- fold cross validation.* Berdasarkan perbandingan hasil akurasi, serta nilai rata-rata presisi dan recall, dapat disimpulkan bahwa algoritma *Logistic Regression* lebih unggul dalam melakukan klasifikasi kanker serviks dibandingkan dengan algoritma K-NN. Saran untuk penelitian selanjutnya adalah untuk mempertimbangkan karakteristik data dan tujuan analisis sebelum memilih metode. Jika data memiliki pola yang jelas dan linearitas yang kuat, *Logistic Regression* bisa menjadi pilihan yang baik. Namun, jika kompleksitas data tinggi dan distribusinya tidak terlalu jelas, K-NN dapat menjadi alternatif yang lebih tepat.

Referensi

[1] A. C. D. S. Santos, N. N. T. Silva, C. M. Carneiro, W. Coura-Vital, and A. A. Lima, "Knowledge about cervical cancer and HPV immunization dropout rate among Brazilian adolescent girls and their guardians," *BMC Public Health*, vol. 20, no. 1, pp. 1–11, 2020
[2] M. Atika, Risnawati, H. Apreliasari, and M. C. Hidajat, "Prevalensi Infeksi Human Papilloma Virus(HPV) pada Perempuan Terinveksi Human Immunodeficiency Virus(HIV)," *Jika*, vol. 6, no. 2, pp. 37–42, 2020.
[3] B. L *et al.*, "Human Papillomavirus and Related Diseases in the World- Summary Report," 2023. [Online]. Available: www.hpvcentre.com
[4] B. V. Kumar, B. Nikhila, A. Rajitha, and M. Alekhya, "Cervical Cancer Consignment Adopting Machine Learning Approach," *J. Pharm. Negat. Results*, vol. 13, no. 10, pp. 5949–5955, 2022.
[5] M. T. Sembiring and C. F. Hasibuan, *Data Science Strategi UMKM dalam Pengambilan Keputusan*. 2023.
[6] Z. R. Tembusai, H. Mawengkang, M. Zarlis, A. Info, and A. H. Process, "K-Nearest Neighbor with K-Fold Cross Validation and Analytic Hierarchy Process on Data Classification," vol. 2, no. 1, pp. 1–8, 2021.
[7] W. Xing and Y. Bei, "Medical Health Big Data Classification Based on KNN Classification Algorithm," *IEEE Access*, vol. 8, pp. 28808–28819, 2020.
[8] I. Ratih Anggriyani, D. Kusumawati, and E. I. J. J. Kawulur, "Metode Regresi Logistik Biner dan Metode K-Nearest Neighbor pada Klasifikasi Menopause Dini Wanita Distrik Oransbari Provinsi Papua Barat," in *Seminar Nasional Matematika, Geometri, Statistika, dan Komputasi*, 2022, pp. 228–233.
[9] Y. I. Kurniawan and T. I. Barokah, "Klasifikasi Penentuan Pengajuan Kartu Kredit Menggunakan K-Nearest Neighbor," *J. Ilm. Matrik*, vol. 22, no. 1, pp. 73–82, 2020.
[10] A. Muzakir, A. Desiani, and A. Amran, "Klasifikasi Penyakit Kanker Prostat Menggunakan Algoritma Naïve Bayes Classification of Prostate Cancer Using Naïve Bayes and K-Nearest Neighbor Algorithms," *Komputika*, vol. 12, no. 148, pp. 1–8, 2023.
[11] A. Desiani, "Perbandingan Implementasi Algoritma Naïve Bayes dan K-Nearest Neighbor Pada Klasifikasi Penyakit Hati," *Simkom*, vol. 7, no. 2, pp. 104–110, 2022.
[12] F. Handayani *et al.*, "Komparasi Support Vector Machine, Logistic Regression Dan Artificial Neural Network dalam Prediksi Penyakit Jantung," *J. Edukasi dan Penelit. Inform.*, vol. 7, no. 3, pp. 329–334, 2021.
[13] M. I. Gunawan, D. Sugiarto, and I. Mardianto, "Peningkatan Kinerja Akurasi Prediksi Penyakit Diabetes Mellitus Menggunakan Metode Grid Seacrh pada Algoritma Logistic Regression," *J. Edukasi dan Penelit. Inform.*, vol. 6, no. 3, pp. 280–284, 2020.
[14] M. R. Romadhon and F. Kurniawan, "A Comparison of Naive Bayes Methods, Logistic Regression and KNN for Predicting Healing of Covid-19 Patients in Indonesia," *EIConCIT*, pp. 41–44, 2021.
[15] R. Tyasnurita and A. Y. M. Pamungkas, "Deteksi Diabetik Retinopati menggunakan Regresi Logistik," *J. Ilm. Ilk.*, vol. 12, no. 2, pp. 130–135, 2020.
[16] S. A. T. Al Azhima, D. Darmawan, N. F. A. Hakim, I. Kustiawan, M. Al Qibtiya, and N. S. Syafei, "Hybrid Machine Learning Model untuk Memprediksi Penyakit Jantung dengan Metode Logistic Regression dan Random Forest," *J. Teknol. Terpadu*, vol. 8, no. 1, pp. 40–46, 2022.
[17] A. Indrasetianingsih, F. Fitriani, and P. J. Kusuma, "Klasifikasi Indeks Pembangunan Gender Di Indonesia Tahun 2020 Menggunakan Supervised Machine Learning Algorithms," *Inferensi*, vol. 4, no. 2, p. 129, 2021.