Analisis Perbandingan Klasifikasi Penyakit Jantung Menggunakan Algoritma *Naïve Bayes* dan Algoritma *Logistic Regression*

Annisa Aulia Lestari[1], Lucy Chania Agatha[1], Anita Desiani[1]
[1]Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Sriwijaya,
Indaralaya, Indonesia
Penulis Korespondensi: anita_desiani@unsri.ac.id

***Abstrak*– Penyakit jantung adalah kondisi ketika bagian jantung mengalami kerusakan. Sehingga, diperlukan pendeteksian dini. Salah satunya dengan melakukan klasifikasi data mining menggunakan Algoritma Naïve Bayes dan Logistic Regression. Pada penelitian ini akan membandingkan Algoritma Naïve Bayes dan Logistic Regression melalui metode training percentage split dan k-fold cross validation untuk mendapatkan hasil klasifikasi terbaik dalam mendeteksi penyakit jantung dengan menghitung nilai rata-rata presisi, recall, dan akurasi. Algoritma Naïve Bayes dengan metode training percentage split menghasilkan nilai rata-rata untuk presisi, recall, dan akurasi sebesar 83%, 82.5% dan 81%, sedangkan Algoritma Naïve bayes dengan k-fold cross validation memberikan hasil nilai rata-rata untuk presisi, recall, dan akurasi sebesar 83.5%, 85.5% dan 83%. Algoritma Logistic Regression dengan metode training percentage split menghasilkan nilai rata-rata untuk presisi, recall, dan akurasi sebesar 73.5%, 73.5% dan 73%, sedangkan Algoritma Logistic Regression menggunakan k-fold cross validation menghasilkan nilai rata-rata untuk presisi, recall, dan akurasi sebesar 84%, 83.5% dan 84%. Hal ini menunjukkan bahwa Algortima Naïve Bayes menggunakan percentage split lebih baik dibandingkan Logistic Regression, akan tetapi pada saat menggunakan metode k-fold cross validation Algoritma Logistic Regression mengalami kenaikan signifikan dibandingkan Naïve Bayes. Sehingga untuk melakukan klasifikasi penyakit jantung lebih baik dengan Algoritma Logistic Regression dengan metode k-fold cross validation.**

Kata Kunci — Klasifikasi, Naïve Bayes, Logistic Regression, Penyakit Jantung

***Abstract*– Heart disease is a condition where parts of the heart are damaged. Thus, early detection is needed. One of them is by doing data mining classification using the Naïve Bayes and Logistic Regression algorithms. This research will compare Naïve Bayes and Logistic Regression algorithms through the training percentage split and k-fold cross validation methods to get the best classification results in detecting heart disease by calculating the average value of precision, recall, and accuracy. The Naïve Bayes algorithm with the training percentage split method produces average values for precision, recall, and accuracy of 83%, 82.5% and 81%, while the Naïve Bayes algorithm with k-fold cross validation provides average values for precision, recall, and accuracy of 83.5%, 85.5% and 83%. Logistic Regression algorithm with percentage split training method produces average values for precision, recall, and accuracy of 73.5%, 73.5% and 73%, while Logistic Regression algorithm using k-fold cross validation produces average values for precision, recall, and accuracy of 84%, 83.5% and 84%. This shows that the Naïve Bayes algorithm using percentage split is better than Logistic Regression, but when using the k-fold cross validation method, the Logistic Regression algorithm has a significant increase compared to Naïve Bayes. So that to classify heart disease is better with the Logistic Regression Algorithm with the k-fold cross validation method.**

Keyword — Classification, Naïve Bayes, Logistic Regression, Heart Disease

I. PENDAHULUAN

Penyakit jantung adalah dimana kondisi tubuh mengalami timbunan lemak menyebabkan arteri tersumbat. Kondisi ini menyebabkan berbagai gejala, termasuk nyeri dada dan kejang. Selain itu, kejadian ini disebabkan oleh gangguan aliran darah ke jantung tidak lancar akibat terjadinya sumbatan pada peredaran darah [1]. Faktor penyebab penyakit jantung diantaranya terjadinya penyempitan pembuluh darah di jantung, terjadinya infeksi pada jantung dan kelaianan katup jantung, serta dapat diakibatkan karena penyakit bawaan [2]. Pada tahun 2020, terdapat 11 juta jiwa yang meninggal akibat penyakit jantung dan kerusakan yang terjadi di pembuluh darah [3]. Untuk mengurangi jumlah kasus penyakit jantung, dibutuhkan deteksi dini terhadap kasus penyakit jantung pada pengindap yang rentang risiko penyakit jantung. Sebuah saran yang bisa digunakan dalam deteksi dini yaitu *data mining*.

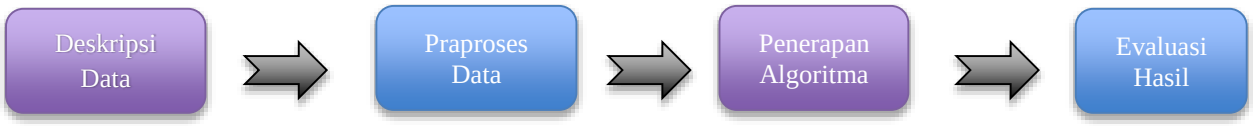
Definisi *Data mining* yakni sebuah proses dimana pengguna mencari pola atau teori terbaru yang berguna serta dapat diterapkan pada basis data yang cukup besar [4]. Salah satu proses *data mining* adalah klasifikasi matematis, klasifikasi sendiri yakni proses evaluasi pada sebuah objek agar bisa diinput pada suatu kategori sesuai dengan jumlah kriteria yang tersedia [5]. Beberapa metode yang digunakan dalam klasifikasi data mining adalah Algoritma *Naïve Bayes* dan *Logistic Regression*. *Naïve Bayes* adalah sebuah algoritma paling sederhana dengan menghitug probabilitas dan menggabungkan kombinasi serta frekuensi nilai sesuai sekumpulan data [6]. Algoritma *Naïve Bayes* memiliki keunggulan yaitu memiliki performa mendeteksi yang tinggi, mudah digunakan, hanya memerlukan satu kali scan data training, mampu menangani data kosong *(missing value)* dan data kontinu [7]. Beberapa penelitian diterapkan dengan algoritma *Naïve Bayes*, seperti yang dilakukan oleh Sri [8] yang menerapkan *Naïve Bayes* untuk pengelompokan status gizi dengan hasil akurasi 93.2%, dan Nungky [9] menerapkan Algoritma *Naïve Bayes* pada pengelompokan diabetes melitus dan memperoleh nilai akurasi 89%. Kekurangan dalam menggunakan metode Algoritma *Naïve Bayes* yaitu probabilitas kurang berjalan secara optimal dan tidak untuk tipe data numerik [10]. Berbeda dengan *Naïve Bayes*. Algoritma *Logistic Regression* dapat memproses probabilitas secara optimal dan dapat menggunakan tipe data numerik.

Logistic Regression adalah analisis *multivariate*, bermanfaat dalam melakuakan prediksi pada variabel terikat sesuai dengan variabel bebasnya [11]. Algoritma *Logistic Regression* ini, digunakan dalam mendeskripsikan data dan menjelaskan hubungan antara satu variabel independen nominal, variabel independen dengan tingkat ordinal, interval atau rasio [12]. Parameter yang diprediksi oleh Algoritma *Logistic Regression* memberikan informasi penting tentang pemahaman hubungan antara satu fitur dengan fitur lainnya [13]. Beberapa penelitian yang menggunakan algoritma *Logistic Regression* seperti penelitian yang dilakukan Jefri [14] yang menggunakan Algoritma *Logistic Regression* sebagai klasifikasi penyakit mata dan perolehan nilai 78.57%, dan Jennie [15] menerapkan Algoritma *Logistic Regression* pada klasifikasi penyakit lambung dan memperoleh nilai akurasi 82 %. Walaupun Algoritma *Logistic Regression* cocok digunakan untuk *dataset* yang besar dan memproses data dalam volume besar dengan kecepatan tinggi. Algoritma ini rentan pada *underfitting* dalam *dataset* yang memiliki kelas tidak ada keseimbanga untuk akurasinya yang kecil.

Berdasarkan kelebihan dan kekurangan algoritma *Naïve Bayes* dan *Logistic Regression* yakni membandingkan hasil dari kedua algoritma tersebut untuk mengidentifikasi metode klasifikasi terbaik yang dapat digunakan dalam klasifikasi penyakit jantung. Dalam pengujian data dilakukan dengan dua metode yaitu *percentage split* dan data traning 80% sedangkan testing 20% dan *k-fold cross validation* melalui k yaitu 10. Hasil kinerja kedua metode akan ditunjukkan melalui nilai akurasi, presisi dan recall. Kedua hasil kinerja tersebut akan dibandingkan untuk mengtahui metode terbaik dalam pendeteksian dini penyakit jantung dengan metode Algoritma *Naïve Bayes* dan *Logistic Regression.*

II. METODOLOGI PENELITIAN

Studi tentang pengelompokan penyakit jantung dilakukan melalui metode Algoritma *Naïve Bayes* dan *Logistic Regression,* terdapat beberapa metode berupa deskripsi data, proses data yaitu meliputi pembagian datanya menggunakan *k-fold cross validation* dan *percentage split*. Dilanjutkan dengan menerapkan metode Algoritma *Naïve Bayes* dan Algoritma *Logistic Regression* diakhiri evaluasi hasil. Alur metode ditunjukkan dalam Gambar 1.



Gambar 1. Alur Penelitian

A. Deskripsi Data

Peneliti menerapkan data terkai dataset penyakit jantung dan didapat dari sumber *website* kaggle (https://www.kaggle.com/datasets/cherngs/heart-disease-cleveland-uci) dengan format csv. Dimana data tersebut terdiri dari 14 atribut, 14 diantaranya sebagai atribut prediksi yaitu sex, age, cp, chol, trestbps, fbs, thalach, restecg, exang, slope, oldpeak, ca, condition dan thal. Data yang tersedia 297 data. Atribut dari data diterapkan ada dalam Tabel I.

TABEL I. ATRIBUT DATASET

Atribut	Tipe Data	Range
Age/umur	Numerik	Usia 40-80 Tahun
Sex/Jenis Kelamin	Nominal	1 = male; 0 = female
Cp/Jenis Nyeri dada	Nominal	Value 0: angina tipikal Value 1: angina atipikal Value 2: nyeri non-angina Value 3: tanpa gejala
Trestbps/Tekanan Darah Rendah	Numerik	90/60mmHg – 120/180mmHg
Chol/Kolestrol	Numerik	200mg/dL
Fbs/Gula Darah Normal	Numerik	(gulah darah > 120 mg/dl) (1 = true; 0 = false)
Restecg/Elektrokardiografi	Nominal	Value 0: normal Value 1: memiliki kelainan gelombang ST-T (inversi gelombang T dan/atau elevasi atau depresi ST > 0,05 mV) Value 2: menunjukkan hipertrofi ventrikel kiri yang mungkin atau pasti berdasarkan kriteria Estes
Thalach/Detak Jantung Maksimum Tercapai	Numerik	95-170 kali/menit
Exang/ Kestabilan Induksi Angina	Numerik	angina yang diinduksi (1 = yes; 0 = no)
Oldpeak / Depresi ST diberikan induksi dari olahraga relatif pada istirahat	Nominal	0-6.2 ST
Slope/ Kemiringan segmen	Nominal	Value 0: miring Value 1: datar Value 2: menurun
Ca/ Nomor Pembuluh darah Utama	Numerik	0-3 diwarnai dengan flourosopi
Thal	Nominal	0 = normal; 1 = cacat tetap; 2 = cacat yang dapat dibalik dan label
Condition	Nominal	0 = tidak ada penyakit, 1 = penyakit

B. Praproses data

Praproses data diterapkan dalam memperoleh hasil akurat, mengurangi waktu perhitungannya sebagai *large scale problem,* serta membentuk nilai kecil dan tidak melakuukan perbuahan pada informasi [16] Dalam praproses data dari 14 atribut di atas tidak ada atribut yang tidak diperlukan atau di hapus karena dari 14 atribut diperlukan dalam melakukan pengkalsifikasian penyakit jantung. Pada atribut dengan range perlu dilakukan metode normalisasi. Metode normalisasi dalam menyelesaikan masalah nilai yang terlalu jauh yakni dengan *Min-Max Normalization*, dimana persamaannya yaitu diterapkan pada persamaan (1).

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

Dimana pada persamaan (1) $normalize\ (x)$ merupakan data setelah normalisasi, x merupakan data sebelum normalisasi, $minValue$ merupakan nilai minimal pada atribut sebelum normalisasi, $maxValue$ adalah nilai maksimal pada atribut sebelum normalisasi, $minRange$ merupakan nilai minimal range 0 dan $maxRange$ merupakan nilai maksimal range 1.

Pada tahap selanjutnya, penerapan metode Algoritma *Naïve Bayes* dan Algoritma *Logistic Regression* dengan membagi data menjadi dua dengan menggunakan *percentage split* dan *k-fold cross validation* dengan parameter ukuran data training 80% serta data testing 20% dalam Algoritma *Naïve Bayes* dan *Logistic Regression*. lalu dilakukan dengan *k-fold cross validation* yang mana menggambil nilai K 10 ke pada kedua algoritma.

C. Penerapan Algoritma

Dalam penerapan algoritma tersebut menerapkan beberapa metode yakni Algoritma *Naïve Bayes* dan *Logistic Regression*, penelitian ini membandingkan dua model tersebut.

1) Naïve Bayes

Algoritma *Naïve Bayes* adalah algoritma metode prediksi. *Naïve Bayes* yakni model probabilistik serta metode statistik yang diusulkan para ilmuan Inggris Thomas Bayes memberikan prediksinya pada kemungkinan di waktu mendatang sesuai pengalaman masa lalu, dinamakan dengan teori Bayes [17]. Dalam teorema Algoritma Naïve Bayes menetukan peluang bersyarat. [18], metodenya menerapkan *Naïve Bayes Classifier* sebagai perhitungan bobot kesempatan masing-masing atribut [19]. Upaya menerapkan klasifikasi dengan algoritma *Naïve Bayes* yakni:

1. Hitung banyaknya kasus dan banyaknya kejadian.

2. Hitung peluang semua variabel berdasarkan setiap kondisi yang ada kasus dengan persamaan (2).

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)} \quad (2)$$

Yang mana $P(A|B)$ merupakan probabilitas bersyarat A yang diberikan oleh B. Pada $P(B|A)$ dimana probabilitas bersyarat B yang diberikan oleh A. Untuk $P(A)$ adalah probabilitas kejadian A, sedangkan $P(B)$ merupakan probabilitas kejadian B [20].

3. Menghitung semua peluang dari masing-masing kelompok pada setiap atribut.

4. Menguji hasil yang diperoleh dari *Naïve Bayes* akan diambil data dan selanjutnya data akan diprediksi hasil peluang untuk setiap kejadian dengan menggunakan persamaan (3).

$$P(C|x_1 \dots x_n) = \frac{P\ (C)\ P\ (x_1 \dots x_n \mid C\ )}{P\ (x_1 \dots x_n\ )} \quad (3)$$

Pada persamaan (3) variabel C merepresentasikan kelas, sementara variabel $x_1 \dots x_n$ merepresentasikan karakteristik petunjuk yang dibutuhkan untuk melakukan klasifikasi atau kriteria. Maka rumus tersebut menjelaskan bahwa peluang masuknya sampel karakteristik tertentu dalam kelas C (*Posterior*) adalah peluang munculnya kelas C (sebelum masuknya sampel tersebut, seringkali disebut *prior*), dikali dengan peluang kemunculan karakteristik-karakteristik sampel pada kelas C (disebut juga *likelihood*), dibagi dengan peluang kemunculan karakteristik-karakteristik sampel secara global (disebut juga *evidence*).

5. Hasil prediksi yang diperoleh sangat baik ketika data uji dapat ditebak dengan benar.

2) Algoritma Logistic Regression

Logistic Regression adalah algoritma analisis prediksi. Penerapannya efisien karena variabelnya tidak bebas dari dataset adalah biner. *Logistic Regression* digunakan dalam mendeskripsikan serta menganalisis data untuk menetukan keterkaitan satu variabel biner tidak bebas serta sutu atau lebih variabel bebas [21]. Adapun langkah dalam pengerjaan Algoritma *Logistic Regression* sebagai berikut :

1. Menetukan nilai prediksi menggunakan persamaan (4).

$$\boldsymbol{b} = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} \quad (4)$$

Dimana pada persamaan (4) $\sum y$ merupakan jumlah seluruh nilai variabel terikat (y), untuk $\sum x^2$ adalah jumlah seluruh nilai variabel bebas (x) yang dikuadratkan, $\sum x$ jumlah seluruh nilai variabel bebas (x), $\sum xy$ jumlah seluruh nilai variabel bebas yang dikalikan dengan variabel terikat (y), sedangkan untuk n adalah banyak data.

2. Menentukan nilai α pada persamaan (5).

$$\boldsymbol{\alpha} = \frac{\sum y - \boldsymbol{b} \sum x}{n} \quad (5)$$

Pada persamaan (5) $\sum y$ adalah jumlah seluruh nilai variabel terikat (y), $\boldsymbol{b}$ koefisien regresi, pada $\sum x$ merupakan jumlah seluruh nilai variabel bebas (x), sedangkan pada n adalah banyak data.

3. Menghitung nilai Y prediksi menggunakan persamaan (6).

$$Y = a + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_n x_{nj} \quad (6)$$

Dimana a merupakan *intercept*, $\beta_1, \ldots, \beta_n$ yakni atribut bebas penurunan, n yakni banyaknya atribut bebas, dan j yakni banyaknya *record* pada *dataset*.

4. Mengeksponensialkan nilai y prediksi yang didapat dari persamaan.
5. Menghitung nilai peluang eksponensial yang diperoleh dari nilai Y menggunakan persamaan (7).

$$P(y) = \frac{\exp(Y)}{1+\exp(Y)} \quad (7)$$

Dimana Y merupakan nilai prediksi dan exp meurapakan nilai eksponen.

6. Menghitung maksimum log likelihood dengan persamaan (8).

$$l(\beta) = \sum[y_i ln\hat{y}_i + (1 - y_i)\ln(1 - \hat{y}_i) \quad (8)$$

Dimana y_i merupakan nilai *independent variabel* berdasarkan kelas yang sudah ditentukan, $\hat{y}_i$ variabel terikat yang diprediksikan.

D. Evaluasi Hasil

Confusion Matrix merupakan matriks yang memiliki fungsi untuk menampilkan penaksiran kinerja dari algoritma, serta digunakan untuk menghitung kinerja perfomansi dari suatu model algoritma dalam suatu prediksi aktual dengan berbentuk *False Positif* (FP), *True Positif* (TP), *False Negative* (FN), dan *True Negative* (TN) dari informasi. Penjelasan lengkap dari *confusion matrix* [22]. Adapun bentuk Counfusion matrix untuk klasifikasi dua kelas terlihat dalam Tabel II.

TABEL II. CONFUSION MATRIX

Kelas		**Nilai Akurasi**	
		Positif	**Negatif**
Nilai Prediksi	**Positif**	True Positif (TP)	False Negatif (FN)
	Negatif	False Positif (FP)	True Positif (TP)

Penjelasan : *True Positive* (TP) yakni banyaknya data positif yang pengkasifikasiannya juga positif. False Negative (FN) adalah banyaknya data negatif pada pengklasifikasiannya positif. *False Positive* (FP) yakni banyaknya data positif yang pengklasifikasiannya negatif. *True Negative* (TN) yakni banyaknya data negatif yang pengklasifikaisannya juga negatif [23]. Dari *Confusion Matrix* bisa dihitung presisi, akurasi, dan recall, sebagai pengukuran hasil pengelompokan data yang sesuai pada semua populasi dapat ditentukan menggunakan akurasi [24]. Berikut rumus untuk menentukan akurasi ditunjukkan pada persamaan (9).

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

Presisi digunakan untuk pengukuran rasio berupa seberapa tepatnya hasil pada output dari sistem. *Confusion Matrix* mengidentifikasi banyaknya jumlah dari hasil yang sesuai serta juga sedikitnya hasil yang tidak benar [25]. Rumus dalam menentukan presisi ditunjukkan pada persamaan (10).

$$Presisi = \frac{TP}{TP+FP} \quad (10)$$

Recall digunakan untuk pengukuran rasio berupa hasil yang sesuai dan diserahkan sistem untuk perbandingan pada semua hasil. *Confusion Matrix* mengidentifikasi yakni banyanya hasil yang benar serta sedikitnya jumlah hasil yang sesuai dan diberikan diagnosis yang salah, maka dilakukannya perhitungan recall [26]. Berikut rumus untuk menentukan recall ditunjukkan pada persamaan (11).

$$Recall = \frac{TP}{TP+FN} \quad (11)$$

III. HASIL DAN PEMBAHASAN

A. Hasil Algoritma Logistic Regression

Penerapan Algoritma *Logistic Regression* pada penyakit jantung. *Confusion matriks* dari *Algoritma Logistic Regression* pada dataset penyakit jantung dengan penerapan *percentage split* dan *k-fold cross validation* dapat diperhatikan pada Tabel III.

TABEL III. *CONFUSION MATRIX* ALGORITMA *LOGISTIC REGRESSION*

Percentage Split				***K-Fold Cross Validation***			
Kelas		**Nilai Aktual**		**Kelas**		**Nilai Aktual**	
		Ada Penyakit	**Tidak Ada Penyakit**			**Ada Penyakit**	**Tidak Ada Penyakit**
Nilai Prediksi	**Ada Penyakit**	*23*	*9*	**Nilai prediksi**	**Ada Penyakit**	143	17
	Tidak Ada Penyakit	*7*	*21*		**Tidak Ada Penyakit**	31	106

Dari Tabel 2 terlihat Algoritma *Logistic Regression* dengan *Percentage Split* memprediksi 23 orang terkena penyakit juga sebagai orang terkena penyakit, 7 orang yang terkena penyakit sebagai orang tidak menderita penyakit, 9 orang tidak menderita penyakit dan menjadi orang yang terkena penyakit, 21 orang tidak menderita penyakit dan menjadi orang yang tidak terkena penyakit. Sesuai *K-Fold Cross Validation* memberikan prediksinya 143 orang menderita penyakit menjadi orang yang terkena penyakit, 31 orang terkena penyakit sebagai orang yang tidak terkena penyakit, 17 orang tidak terkena penyakit sebagai orang terkena penyakit dan 106 orang tidak terkena penyakit sebagai orang yang tidak terkena penyakit. Dari Algoritma *Logistic Regression* dengan *Percentage Split* memperoleh akurasi sebesar 84%. untuk Nilai presisi pada kelas terkena penyakit sebesar 70% serta kelas yang tidak terkena penyakit 77%. Nilai recall terkena penyakit 75% serta tidak terkena penyakit 72%. Sedangkan dengan *K-Fold Cross Validation* memperoleh akurasi sebesar 84%. Nilai presisi yang memiliki penyakit 86% serta tak terkena penyakit 82%. Nilai recall terkena penyakit 77% serta tidak terkena penyakit ada 89%.

B. Hasil Algoritma Naïve Bayes

Penggunaan algoritma *Naïve Bayes* pada penyakit jantung. *Confusion matriks* dari algoritma Naïve Bayes dalam dataset penyakit jantung dengan *training persentase split* dan *k-fold validation* sesuai Tabel IV.

TABEL IV. *CONFUSION MATRIX* ALGORITMA *NAÏVE BAYES*

Percentage Split				***K-Fold Cross Validation***			
Kelas		**Nilai Aktual**		**Kelas**		**Nilai Aktual**	
		Ada Penyakit	**Tidak Ada Penyakit**			**Ada Penyakit**	**Tidak Ada Penyakit**
Nilai Prediksi	**Ada Penyakit**	26	2	**Nilai prediksi**	**Ada Penyakit**	140	20
	Tidak Ada Penyakit	9	23		**Tidak Ada Penyakit**	30	107

Dari Tabel IV Algoritma *Naïve Bayes* dengan *Percentage Split* memberikan prediksinya 26 orang terkena penyakit sebagai terkena penyakit, 9 orang terkena penyakit sebagai tidak terkena penyakit, 2 orang tidak terkena penyakit sebagai terkena penyakit, 23 orang tidak terkena penyakit sebagai tidak terkena penyakit. Sedangkan berdasarkan *K-Fold Cross Validation* memprediksi 140 orang terkena penyakit sebagai terkena penyakit, 30 orang terkena penyakit sebagai tidak terkena penyakit, 20 orang tidak terkena penyakit sebagai terkena penyakit dan 107 orang tidak terkena penyakit sebagai tidak terkena penyakit. Pada Algoritma *Naïve Bayes* dengan *Percentage Split* memperoleh akurasi sebesar 81%. Nilai presisi untuk terkena penyakit sebesar 92% dan untuk tidak terkena penyakit sebesar 74%. Nilai recall untuk terkena penyakit sebesar 72% dan untuk tidak terkena penyakit sebesar 93%. Sedangkan dengan *K-Fold Cross Validation* memperoleh akurasi sebesar 83%. Nilai presisi untuk terkena penyakit sebesar 84% dan untuk tidak terkena penyakit sebesar 82%. Nilai recall untuk terkena penyakit sebesar 78% dan untuk tidak terkena penyakit sebesar 88%.

C. Perbandingan Hasil Kedua Metode

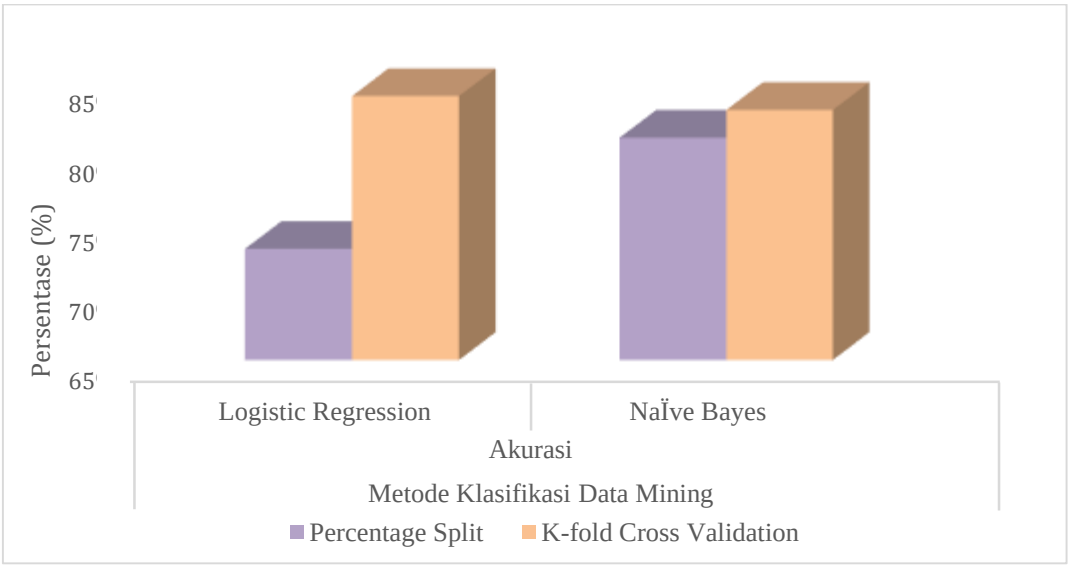
Hasil prediksi dari kedua metode Algoritma *Naïve Bayes* dan *Algoritma Logistic Regression* melalui latih *k-fold cross validation* dan *percantage split* menjelaskan algoritma *Naïve Bayes* dan *Algoritma Logistic*

Regression terebut bekerja baik dalam melakukan prediksi penyakit jantung pada dataset *kaggle*. Bandingan pada hasil ukur algoritma ini terlihat di Tabel V.

TABEL V. NILAI PRESISI, RECALL DAN AKURASI *Naïve Bayes* DAN *ALGORITMA LOGISTIC REGRESSION*

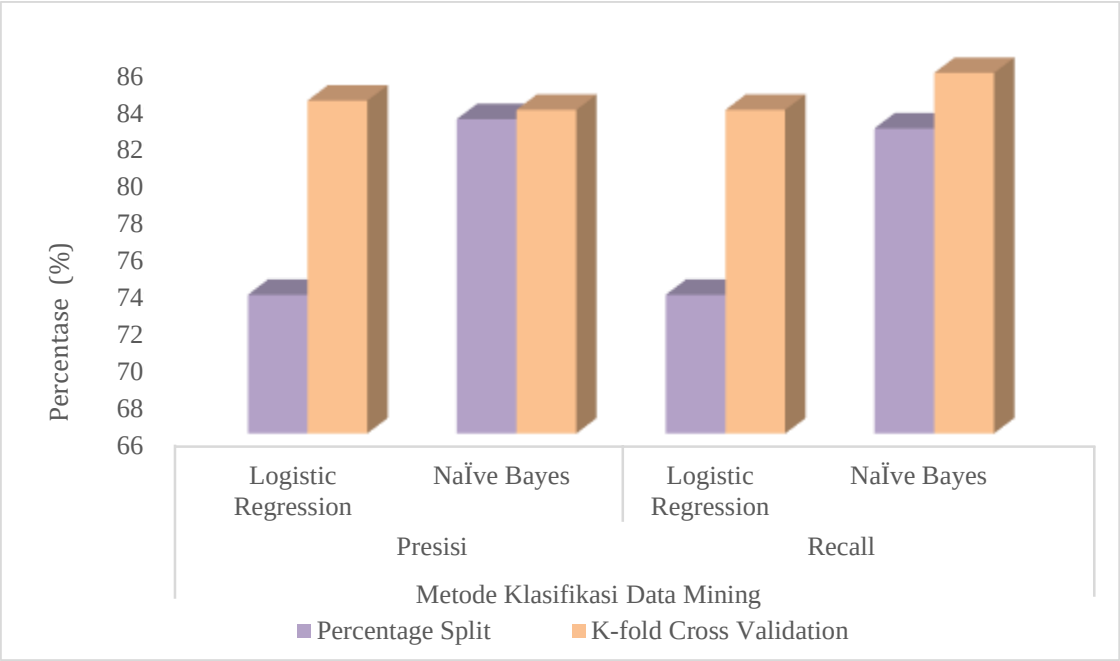
Algoritma	Pemodelan	Label	Presisi	Recall	Akurasi
Naive Bayes	***Percentage split***	**Ada Penyakit**	83%	82,5%	81%
		Tidak Ada Penyakit			
	K-Fold cross Validation	**Ada Penyakit**	83%	85%	83%
		Tidak Ada Penyakit			
Algoritma Logistic Regression	***Percentage split***	**Ada Penyakit**	73,5%	73,5%	73%
		Tidak Ada Penyakit			
	K-Fold cross Validation	**Ada Penyakit**	84%	83,5%	84%
		Tidak Ada Penyakit			

Dari Tabel V dapat dilihat pada teknik pengujian *precentage split* algoritma *Logistic Regression* menghasilkan nilai presisi, recall dan akurasi lebih rendah dibanding algoritma *Naïve Bayes* dengan nilai masing-masing berturut 73,5%, 73,5%, dan 73%. Dan algoritma *Naïve Bayes* hanya memberikan hasil nilai akurasi, presisi, dan recall masing-masing berturut 83%, 82,5% dan 81%. Pada teknik uji *k-fold cross validation* algoritma *Naïve Bayes* memiliki nilai presisi, akurasi, dan recall rendah dibanding algoritma *Logistic Regression* dengan nilai masing-masing berturut 84%, 83,5% dan 84%. Dan algoritma *Naïve Bayes* hanya memiliki akurasi, presisi, dan recall masing-masing berturut 83%, 83% dan 83%. Berdasarkan hasil kedua teknik pengujian metode algoritma *Naïve Bayes* mempunyai akurasi, presisi dan recall yang lebih besar dari *Logistic Regression*. Meskipun pada teknik pengujian percentage split yang paling baik dalam melakukan pendeteksian penyakit jantung yaitu menggunakan Algoritma *Naïve Bayes* dibandingkan *Logistic Regression*. Akan tetapi pada teknik pengujian *K-Fold cross validation* menghasilkan nilai yang lebih besar dari *precentage split* untuk kedua metode klasifikasi, dimana pada nilai akurasi mengalami kenaikan signifikan yang terjadi pada Algoritma *Logistic Regression* walapun selisih nilai yaitu 1% dari metode Algoritma *Naïve Bayes.*



Gambar 2. Nilai Akurasi

Pada gambar 2 dapat dilihat nilai akurasi dari metode Algoritma *Naïve Bayes* dan *Algoritma Logistic Regression,* nilai akurasi dengan *k-fold cross validation* lebih besar dari *percentage split.* Pada metode Algoritma *Naïve Bayes* dan *k-fold cross validation* nilai akurasi memiliki nilai selisih 2% lebih tinggi dari *percantage split,* sedangkan pada metode *Algoritma Logistic Regression* dengan *k-fold cross validation* nilai akurasi memiliki nilai dengan selisih yang cukup jauh yaitu 11% lebih tinggi dari *percentage split.*



Gambar 3. Nilai Rata-rata Presisi dan Recall

Pada gambar 3 terlihat bahwa nilai dari rata-rata presisi dengan metode Algoritma *Logistic Regression* dengan *percentage spit* dan *K-fold cross validation* yaitu 78,75%. Hasil dari rata-rata tersebut lebih rendah jika dibandingkan dengan rata rata presisi dari Algoritma *Naïve Bayes* dengan menggunakan *percentage spit* dan *K-fold cross validation* yaitu 83,25%. Sedangkan rata-rata recall dari metode Algoritma *Logistic Regression* dengan *percentage spit* dan *K-fold cross validation* yaitu 78,5%. Hasil rata-rata ini lebih jecil dari presisi dari Algoritma *Naïve Bayes* dan *percentage spit* dan juga *K-fold cross validation* yaitu 84%. Metode Algoritma *Logistic Regression* dengan *K-fold cross validation* lebih baik dibandingkan dengan *percentage spit.* Begitupun dengan Naive Bayes dengan *K-fold cross validation* lebih baik dibandingkan dengan *percentage spit.*

V. KESIMPULAN

Sesuai pada hasil yang diperoleh pada metode Algoritma *Naïve Bayes* dan *Logistic Regression* memiliki kinerja yang baik dalam memprediksi penyakit jantung akan tetapi yang sangat baik dalam memprediksi penyakit jantung yaitu Algoritma *Logistic Regression* dimana pada metode *k-fold cross validation* nilai akurasi Algoritma *Logistic Regression* mengalami kenaikan signifikan dari Algoritma *Naïve Bayes* dengan selisih nilai akurasi sebesar 1%. Algoritma *Logistic Regression* mempunyai hasil baik dalam pendeteksian dini penyakit jantung, meskipun dengan *presisi* dan *recall* pada metode *percentage split* algoritma *Naïve Bayes* memperoleh hasil yang lebih baik. Pada Algoritma *Logistic Regression* dengan metode *percentage split* menghasilkan nilai yang lebih rendah, Namun untuk performa dari presisi, *recall* dan akurasinya masih berada diatas 70%. Maka, kesimpulannya algoritma *Naïve Bayes* dan *Logistic Regression* baik untuk digunakan dalam melakukan pendeteksian dini penyakit jantung.

DAFTAR PUSTAKA

[1] A. Riani, Y. Susianto, and N. Rahman, “Implementasi Data Mining Untuk Memprediksi Penyakit Jantung Mengunakan Metode Naive Bayes,” *J. Innov. Inf. Technol. Appl.*, vol. 1, no. 01, pp. 25–34, 2019.

[2] J. J. Pangaribuan, H. Tanjaya, and Kenichi, “Mendeteksi Penyakit Jantung Menggunakan Mechine Learning Dengan Algoritma Logistic Regression,” *Mach. Learn.*, vol. 45, no. 13, pp. 40–48, 2017.

[3] S. Sahar, “Analisis Perbandingan Metode K-Nearest Neighbor dan Naïve Bayes Clasiffier Pada Dataset Penyakit Jantung,” *Indones. J. Data Sci.*, vol. 1, no. 3, pp. 79–86, 2020.

[4] D. Y. Utami, E. Nurlelah, and F. N. Hasan, “Comparison of Neural Network Algorithms, Naive Bayes and Logistic Regression to predict diabetes,” *J. Informatics Telecommun. Eng.*, vol. 5, no. 1, pp. 53–64, 2021.

[5] Fauziah, M. A. Tiro, and Ruliana, “Comparison of k-Nearest Neighbor (k-NN) and Support Vector Machine (SVM) Methods for Classification of Poverty Data in Papua,” *ARRUS J. Math. Appl. Sci.*, vol. 2, no. 2, pp. 83–91, 2022.

[6] T. Arifin and D. Ariesta, “Prediksi Penyakit Ginjal Kronis Menggunakan Algoritma Naive Bayes Classifier Berbasis Particle

Swarm Optimization," *J. Tekno Insentif*, vol. 13, no. 1, pp. 26–30, 2019.

[7] I. A. Musdar and P. Informatika, "APLIKASI PREDIKSI KERUSAKAN SMARTPHONE MENGGUNAKAN METODE NAIVE BAYES DAN LAPLACE SMOOTHING," vol. 5, no. 2, pp. 8–16, 2018.

[8] S. Kusumadewi, "Klasifikasi Status Gizi Menggunakan Naive Bayesian Classification," *CommIT (Communication Inf. Technol. J.*, vol. 3, no. 1, p. 6, 2009.

[9] N. Asmiati and Fatmawati, "Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Pengaruh Negatif Game Online Bagi Remaja Milenial," *JTIM J. Teknol. Inf. dan Multimed.*, vol. 2, no. 3, pp. 141–149, 2020.

[10] A. Saputra, R. P. Hasibuan, Renaldi, and Rahmaddeni, "Perbandingan Tingkat Kadar Minuman Beralkohol di Dunia Menggunakan Algoritma Naïve Bayes dan K-Nearest Neighbor," *SENTIMAS Semin. Nas. Penelit. dan Pengabdi. Masy.*, pp. 127–132, 2022.

[11] A. R. C. Adi, "Analisis Kepuasan Pelayanan Rumah Sakir Mitra Husada Pringsewi Menggunakaan Metode Logistic Regression," *J. Ilmu Data*, vol. 2, no. 12, pp. 1–11, 2022.

[12] S. A. Assaidi and F. Amin, "Analisis Sentimen Evaluasi Pembelajaran Tatap Muka 100 Persen pada Pengguna Twitter menggunakan Metode Logistic Regression," *J. Pendidik. Tambusai*, vol. 6, no. 2, pp. 13217–13227, 2022.

[13] F. Reviantika, Y. Azhar, G. I. Marthasari, Wicaksono, and Triyono, "Analisis Klasifikasi SMS Spam Menggunakan Logistic Regression," *J. Sist. Cerdas*, vol. 4, no. 2, pp. 37–43, 2021.

[14] A. Putra Wijaya and H. Santoso, "Komparasi Performansi Algoritma Naive Bayes dan Logistic Regression pada Malware Android," *J. INTEK*, vol. 4, no. 2, pp. 31–40, 2021.

[15] J. Pearce and S. Ferrier, "Evaluating the predictive performance of habitat models developed using logistic regression," *Ecol. Modell.*, vol. 133, no. 3, pp. 225–245, 2000.

[16] A. Desiani, "Perbandingan Implementasi Algoritma Naïve Bayes dan K-Nearest Neighbor Pada Klasifikasi Penyakit Hati," *Simkom*, vol. 7, no. 2, pp. 104–110, 2022.

[17] D. Tien Bui, B. Pradhan, O. Lofman, and I. Revhaug, "Landslide susceptibility assessment in vietnam using support vector machines, decision tree, and nave bayes models," *Math. Probl. Eng.*, vol. 2012, 2012.

[18] R. Ardianto, T. Rivanie, Y. Alkhalifi, F. S. Nugraha, and W. Gata, "Sentiment Analysis on E-Sports for Education Curriculum Using Naive Bayes and Support Vector Machine," *J. Ilmu Komput. dan Inf.*, vol. 13, no. 2, pp. 109–122, 2020.

[19] N. P. Nugraha, R. Azim, S. Z. Daffa, and P. S. Ningayu, "Perbandingan Akurasi Metode Naïve Bayes dan Metode KNN untuk Memprediksi Gagal Ginjal Kronis," *J. Rekayasa Elektro Sriwij.*, vol. 5, no. 1, pp. 1–10, 2023.

[20] D. Marutho, "Perbandingan Metode Naive Bayes , KNN , Decision Tree Pada Laporan Water Level Jakarta," *Manaj. Inform. AMIK JTC Semarang*, vol. 15, no. 2, pp. 90–97, 2019.

[21] M. E. Shipe, S. A. Deppen, F. Farjah, and E. L. Grogan, "Developing prediction models for clinical use using logistic regression: An overview," *J. Thorac. Dis.*, vol. 11, no. Suppl 4, pp. S574–S584, 2019.

[22] A. A. D. Halim and S. Anraeni, "Analisis Klasifikasi Dataset Citra Penyakit Pneumonia menggunakan Metode K-Nearest Neighbor (KNN)," *Indones. J. Data Sci.*, vol. 2, no. 1, pp. 01–12, 2021.

[23] Q. R. Cahyani *et al.*, "Prediksi Risiko Penyakit Diabetes menggunakan Algoritma Regresi Logistik Diabetes Risk Prediction using Logistic Regression Algorithm Article Info ABSTRAK," *JOMLAI J. Mach. Learn. Artif. Intell.*, vol. 1, no. 2, pp. 2828–9099, 2022.

[24] P. G. S. C. Nugraha, G. R. Dantes, and K. Y. E. Aryanto, "Implementasi Metode C4.5 Dan Naive Bayes Berbasis Adaboost Untuk Memprediksi Kelayakan Pemberian Kredit," *Int. J. Nat. Sci. Eng.*, vol. 1, no. 2, p. 65, 2017.

[25] M. D. Purbolaksono, M. Irvan Tantowi, A. Imam Hidayat, and A. Adiwijaya, "Perbandingan Support Vector Machine dan Modified Balanced Random Forest dalam Deteksi Pasien Penyakit Diabetes," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 2, pp. 393–399, 2021.

[26] H. N. Irmanda and Ria Astriratma, "Klasifikasi Jenis Pantun Dengan Metode Support Vector Machines (SVM)," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 5, pp. 915–922, 2020.