JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 4, Nomor 2, April 2020, Page 437-444
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v4i2.2080

Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung

Dito Putro Utomo*, Mesran

Program Studi Teknik Informatika, STMIK Budi Darma, Medan, Indonseisa
Email: 1,*ditopuro12@gmail.com, mesran.skom.mkom@gmail.com
Email Penulis Korespondensi: ditopuro12@gmail.com

Abstrak−Penyakit jantung merupakan salah satu penyakit dengan tingkat kematian yang tinggi, tercatat 12 juta jiwa kematian setiap tahunnya di seluruh dunia. Hal ini yang menyebabkan perlu diagnosa dini untuk mengetahui penyakit jantung tersebut. Tetapi proses diagnosa merupakan hal yang cukup menantang dikarenakan keterkaitan yang cukup kompleks diantara atribut–atribut penyakit jantung. Sehingga kiranya perlu diketahui atribut utama yang digunakan sebagai proses pengambilan keputusan atau proses klasifikasi pada penyakit jantung. Pada penelitian ini dataset yang digunakan memiliki 57 jenis atribut didalamnya. Sehingga perlu dilakukan reduksi untuk memeprsingkat proses diagnosa, proses reduksi dapat dilaukan dengan menggunakan metode Principal Componen Analysisi (PCA). Metode PCA itu sendiri dapat dikombinasikan dengan teknik kalsifikasi data mining untuk mengukur tingkat akurasi pada dataset. Penelitian ini melakukan perbandingan tingkat akurasi dengan menggunakan algortima C5.0 dan algoritma Naïve Bayes Classifier (NBC), hasil yang didapatkan baik sesudah dan sebeulm reduksi adalah algoritma Naïve Bayes Classifier (NBC) memiliki kinerja yang lebih baik dari pada algoritma C5.0.

Kata Kunci: Penyakit Jantung, Data Mining, Reduksi, C5.0, Naïve Bayes Classifier

Abstract−Heart disease is a disease with a high mortality rate, there are 12 million deaths each year worldwide. This is what causes the need for early diagnosis to find out the heart disease. But the process of diagnosis is quite challenging because of the complex relationship between the attributes of heart disease. So it is important to know the main attributes that are used as a decision making process or the classification process in heart disease. In this study the dataset used has 57 types of attributes in it. So that reduction is needed to shorten the diagnostic process, the reduction process can be carried out using the Principal Component Analysis (PCA) method. The PCA method itself can be combined with data mining calcification techniques to measure the accuracy of the dataset. This study compares the accuracy rate using the C5.0 algorithm and the Naïve Bayes Classifier (NBC) algorithm, the results obtained both after and before the reduction are Naïve Bayes Classifier (NBC) algorithms that have better performance than the C5.0 algorithm.

Keywords: Heart Disease, Data Mining, Reduction, C5.0, Naïve Bayes Classifier

1. PENDAHULUAN

Penyakit Jantung merupakan sebutan umum yang digunakan untuk menggambarkan gangguan terhadap fungsi kerja jantung. Penyakit atau gangguan jantung sendiri memiliki banyak jenis dan macam nama penyakitnya sepertia kardiovaskuler, jantung koroner dan serangan jantung. Penyakit Jantung merupakan salah satu penyakit yang paling sering terjadi kasusnya pada kalangan masyarakat, dimana penyakit jantung ini dapat menimpa dan menyerang siapapun tanpa memandang usia, jenis kelamin dan haya hidup. Menurut WHO (Organisasi Kesehatan Dunia) dan CDC, penyakit jantung adalah penyebab utama kematian di Inggris, Amerika Serikat, Kanada dan Australia. Jumlah orang dewasa yang didiagnosis dengan penyakit jantung terdiri dari 26,6 Juta Jiwa (11,3%) dari populasi orang dewasa [1].

Penyakit jantung salah satu penyakit yang memiliki angka kematian dengan tingkat yang tinggi lebih dari 12 juta jiwa kematian yang terjadi pada seluruh dunia dikaernakan penyakit jantung ini [2]. Dengan demikian diagnosa secara dini sangat penting untuk dilakukan, diagnose pada penyakit jantung merupakan hal yang sangat menantang dikarenakan saling ketergantung yang kompleks dari beberapa faktor atribut. Permasalahan yang sering kali dihadapi adalah kurangnya akurasi pada proses klasifikasi [3].

Atribut merupakan suatu hal yang penting pada keakuratan hasil proses, sehingga perlu diketahui atribut – atribut utama pada penyakit terkhususnya penyakit jantung. Didalam proses pengambilan keputusan pada penyakit sering didapatkan hasil yang berbeda sehingga perlu mengetahui atribut – atribut utama didalam pengambilan keputusan. Pada penelitian ini menggunakan dataset dengan atribut sebanyak 57, dengan jumlah atribut yang terlalu banyak akan memakan waktu didalam proses pengambilan keputusan. Sehingga kiranya atribut yang terlalu banyak perlu dilakukan reduksi.

Reduksi merupakan pemilihan fitur atau atribut untuk mengurangi dimensi data dengan menghapus atribut yang tidak relevan pada dataset, dimana atribut yang tidak relevan akan sangat menggangu proses dan tingkat akurasi pada klasifikasi. Hal ini menginsprasi penulis untuk melakukan reduksi pada dataset penyakit jantung yang memiliki jumlah atribut yang banyak. Principal Compinen Analysis (PCA) merupakan salah satu teknik yang dapat digunakan untuk mengurangi dimensi pada dataset tetapi meminimalisir hilangnya informasi yang terdapat ada dataset tersebut. PCA sendiri dapat digabung dengan metode atau teknik – teknik lainnya untuk mengukur tingkat akurasi pada dataset.

Untuk mengukur tingkat akurasi pada dataset dapat dilakukan dengan teknik klasifikasi datamining, klasifikasi merupakan suatu pekerjaan menilai objek data untuk memasukannya kedalam kelas tertentu dari jumlah kelas yang tersedia. Klasifikasi melakukan pembangunan model berdasarkan data latih yang ada, kemudian

Submitted: **12/03/2020**; Accepted: **10/04/2020**; Published: **25/04/2020**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 4, Nomor 2, April 2020, Page 437-444
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v4i2.2080

menggunakan model tersebut untuk mengklasifikasikan pada data yang baru. Klasifikasi dapat didefinisikan sebagai pekerjaan yang melakukan pelatihan/pembelajaran terhadap fungsi target yang memetakan setiap set atribut (fitur) ke satu jumlah label kelas yang tersedia.

Algoritma C5.0 merupakan bagian dari pada teknik klasifikasi datamining, algoritma C5.0 memiliki kemampuan untuk mengklasifikasikan sebagai pohon keputusan atau seperangkat aturan. C5.0 adalah classifier yang mengklasifikasikan data dalam waktu yang lebih singkat dibandingkan dengan classifier lain. Untuk menghasilkan pohon keputusan penggunaan memori minimum dan juga meningkatkan akurasi [4].

Naïve Bayes Classifier merupakan algoritma yang berlandaskan Teorema Bayes menggunakan teknik probabilitas dan statistik untuk meperkirakan ataupun memprediksi peluang yang akan terjadi berdasarkan dengan pengalaman sebelumnya. Penelitian – penelitian lainnya Algoritma Naïve Bayes Classifier memiliki kinerja yang cukup tinggi untuk proses klasifikasi dengan melakukan pengujian sebanyak 13106 terhadap data sampel menghasilkan nilai akurasi terkecil sebesar 78% dan nilai akurasi terbesar 91,60% [5].

Pada penelitian sebelumnya yang dilakukan oleh Siti Masripah dengan melakukan komparasi algoritma klasifikasi data mining antara algortima C5.0 dan algoritma naïve bayes hasil yang didapatkan adalah dari perbandingan kedua algoritma bahwa tingkat akurasi yang lebih baik adalah menganalisa menggunaka algoritma klasifikasi C4.5 yaitu 88.90 % sedangkan untuk tingkat akurasi menggunakan algortima klasifikasi Naïve Bayes yaitu 80.00% [6].

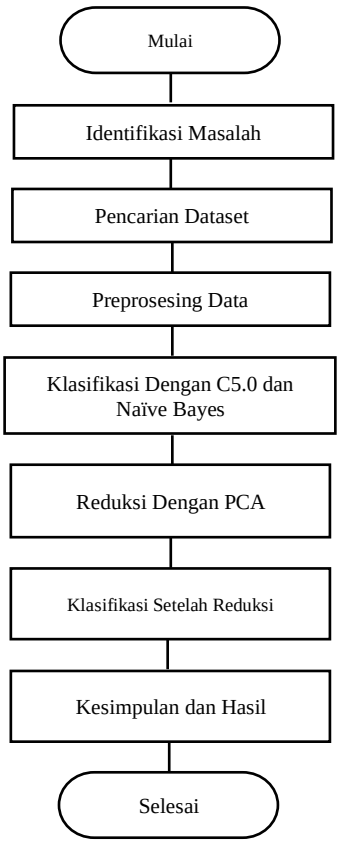
Penelitian lain yang dilakukan oleh Gusni Rahayu dan Mustakim menggunakan metode Principal Component Analysis (PCA) untuk mereduksi dimensi data clustering menghasilkan bahwa Principal Component Analysis (PCA) merupakan pendekatan fitur selection untuk pengurangan dimensi tanpa pengawasan teknik. Penelitian ini menghasilkan sebuah metode yang lebih efektif dengan menggunakan PCA yang dimodifikasi dengan metode lainnya sebagai pengukur tingkat akurasi [7].

Berdasarkan penjelasan yang sudah dijabarkan sehingga penelitian ini akan melakukan reduksi pada dataset penyakit jantung dengan mengetahui kombinasi setiap atribuat pada dataset dan kemudian melakukan perbandingan tingkat akurasi pada dataset menggunakan teknik klasifikasi datamining.

2. METODE PENELITIAN

2.1 Metodologi Penelitian

Metodologi penelitian merupakan tahapan-tahapan yang sistematis dilakukan pada penelitian ini sehingga penelitian ini terarah dengan baik. Berikut adalah metodologi penelitian yang dilakukan



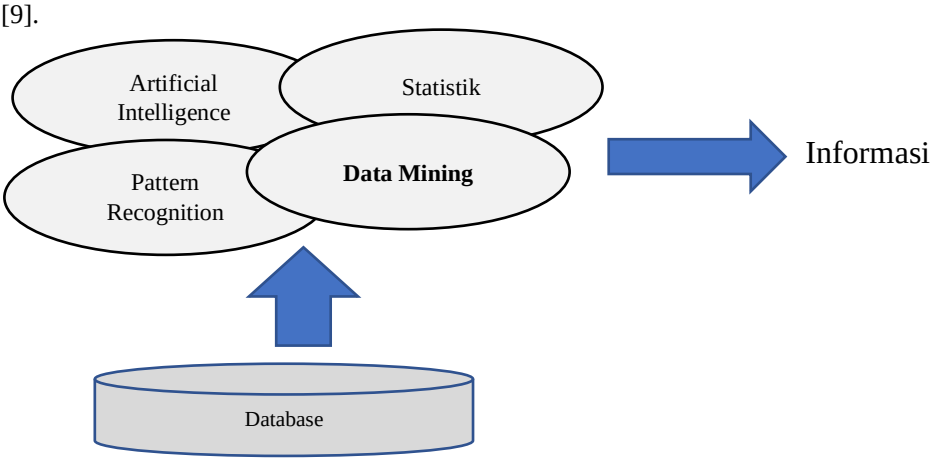
Gambar 1. Metodologi Penelitian

Submitted: **12/03/2020**; Accepted: **10/04/2020**; Published: **25/04/2020**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 4, Nomor 2, April 2020, Page 437-444
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v4i2.2080

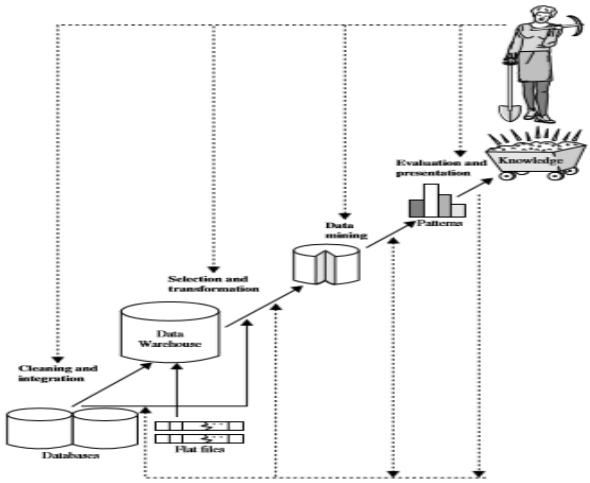
2.2 Data Mining

Data mining merupakan analisis dari peninjauan kumpulan data untuk menemukan hubungan yang tidak diduga dan meringkas data dengan cara berbeda dengan cara yang berbeda dengan sebelumnya, yang dapat dipahami dan bermanfaat bagi pemilik data. Data mining merupakan bidang dari beberapa bidang keilmuan yang menyatukan teknik dari pembelajaran mesin, pengenalan pola, statistic, database, dan visualisasi untuk penanganan permasalahan pengambilan informasi dari database yang besar [8]. Data Mining adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan di dalam database. Data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terikat dari berbagai database besar. Berdasarkan pengertian data mining yang telah dijelaskan di atas, maka data mining merupakan pengetahuan yang tersembunyi di dalam database yang di proses untuk menemukan pola dan teknik statistik matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi pengetahuan dari database tersebut [9].



Gambar 2. Akar Ilmu Data Mining

Di dalam *data mining* terdapat istilah lain yang mempunyai makna serupa dengan *data mining* yakni *Knowledge Discovery in Database* (KDD). *Data mining* serta KDD memiliki tujuan yang sama yakni menggunakan data yang tersedia pada basis data kemudian mengolah data untuk mendapatkan sebuah informasi baru yang bermanfaat. Selain itu, banyak istilah lain yang memiliki makna serupa dengan data mining misalnya, *Knowledge Discovery in Database*, ekstraksi pengetahuan, analisis pola / data, arkeologi data, dan pengerukan data. Banyak orang memperlakukan data mining sebagai sinonim untuk istilah lain yang populer digunakan, penemuan pengetahuan dari data, atau KDD, sementara yang lain melihat penambangan data hanya sebagai langkah penting dalam proses penemuan pengetahuan [8].



Gambar 3. Tahapan Proses *Knowledge Discovery in Database* (KDD)

2.3 Metode Principal Component Analysis (PCA)

Principal Component Analysis (PCA) melakukan pemetaan/transformas set data dari dimensi lama ke dimensi baru dengan memanfaatkan teknik dalam aljabar linear, tanpa memerlukan parameter tertentu dalam memberikan keluaran hasil pemetaanya [9].

Submitted: **12/03/2020**; Accepted: **10/04/2020**; Published: **25/04/2020**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 4, Nomor 2, April 2020, Page 437-444
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v4i2.2080

PCA memerlukan masukan data yang mempunyai sifat zero-mean pada setiap fiturnya. Sifat zero-mean pada setiap fitru data bisa didapatkan dengan mengurangkan semua nilai dengan rata-ratanya. Set data X dangan dimensi MxN, dimana M adalah jumlah data dan N adalah jumlah fitur.

$$X = \begin{vmatrix} X_{11} & X_{12} & X_{1j} & X_{1N} \\ X_{21} & \dots & \dots & X_{2N} \\ X_{31} & \dots & \dots & X_{3N} \\ X_{i1} & \dots & \dots & X_{iN} \\ X_{M1} & X_{M2} & X_{Mj} & X_{MN} \end{vmatrix}$$

Untuk fitur ke-j, semua nilai pada kolom tersebut dikurangi dengan rata-ratanya, diformulasikan dengan:

$$X_{ij} = X_{ij} - \bar{X}_j \quad (1)$$

i = 1,2 …, M, dan j adalah kolom ke-j

Selanjutnya dilakukan perhitungan matriks kovarian dari matriks X, yaitu C_X. Formula yang digunakan adalah dot-product pada setiap fitur.

$$C_X = \frac{1}{M} X^T.X \quad (2)$$

N adalah jumlah fitur, sedangkan X^T adalah matriks transpos dari X.

$$C_X = \frac{1}{M} \text{ x } \begin{bmatrix} X_{11} & X_{21} & X_{i1} & X_{M1} \\ X_{12} & \dots & \dots & X_{M2} \\ X_{1i} & \dots & \dots & X_{Mi} \\ X_{1N} & X_{2N} & X_{iN} & X_{NM} \end{bmatrix} \text{ x } \begin{bmatrix} X_{11} & X_{12} & X_{1j} & X_{1N} \\ X_{21} & \dots & \dots & X_{2N} \\ X_{i1} & \dots & \dots & X_{iN} \\ X_{M1} & X_{M2} & X_{Mj} & X_{MN} \end{bmatrix}$$

$$= \frac{1}{M} \begin{bmatrix} X_{11} & X_{12} & X_{1j} & X_{1N} \\ X_{21} & \dots & \dots & X_{2N} \\ X_{i1} & \dots & \dots & X_{iN} \\ X_{N1} & X_{N2} & X_{Nj} & X_{NN} \end{bmatrix}$$

Pada matriks C_X, elemen ke-ij adalah inner-product antara baris matriks X_T dengan kolom matriks X. Sifat-sifat yang dimiliki oleh matriks C_X

Cara yang umum digunakan untuk mendapatkan C_Y adalah dengan eigenvalue dan eigenvector. Eigenvalue dan eigenvector dari matriks X berturut-turut adalah nilai skala ƛ dan vektor u yang memenuhi persamaan berikut

$$Xu = ƛu \quad (3)$$

Dengan mencari matriks ortonormal P dimana Y = PX dan $C_Y = \frac{1}{M} YY^T$ adalah matriks diagonal, dan kolom dari P adalah komponen utama (principal components) dari X, persamaan C_Y bisa dijabarkan sebagai berikut:

$$\begin{aligned} C_Y &= \frac{1}{M} YY^T \\ &= \frac{1}{M} (PX)(PX)^T \\ &= \frac{1}{M} PXX^TP^T \\ &= P(\frac{1}{M} XX^T)P^T \end{aligned}$$

Dengan mendistribusikan persamaan 2, didapatkan matriks C_Y berdimensi NxN:

$$C_Y = PC_XP^T \quad (4)$$

2.4 Teknik Klasifikasi

Klasifikasi merupakan suatu pekerjaan menilai objek data untuk memasukannya kedalam kelas tertentu dari jumlah kelas yang tersedia. Klasifikasi melakukan pembangunan model berdasarkan data latih yang ada, kemudian menggunakan model tersebut untuk mengklasifikasikan pada data yang baru. Klasifikasi dapat didefinisikan sebagai pekerjaan yang melakukan pelatihan/pembelajaran terhadap fungsi target yang memetakan setiap set atribut (fitur) ke satu jumlah label kelas yang tersedia. Sebuah sistem yang melakukan klasifikasi diharapkan dapat melakukan klasifikasi semua data set dengan benar, tetapi tidak dapat dipungkiri bahwa kinerja sistem tidak bisa 100% benar sehinga sebuah sistem klasifikasi juga harus diukur kinerjanya. Umumnya, pengukuran kinerja klasifikasi dilakukan dengan matriks konfusi.

Tabel 1. Matriks Konfusi Untuk Dua Kelas

f_{ij}		Kelas Hasil Prediksi (*j*)	
		Kelas = 1	**Kelas = 0**
Kelas Asli (*i*)	Kelas = 1	F_{11}	F_{10}
	Kelas = 0	F_{01}	F_{00}

Berdasarkan isi matriks konfusi, dapat diketahui jumlah data dari masing-masing kelas yang diprediksi secara benar dan data yang diklasifikasikan secara salah. Kuantitasi matriks konfusi dapat diringkas menjadi dua nilai, yaitu akurasi dan laju eror. Dengan mengetahui jumlah data yang diklasifikasi secara benar, dapat

Submitted: **12/03/2020**; Accepted: **10/04/2020**; Published: **25/04/2020**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 4, Nomor 2, April 2020, Page 437-444
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v4i2.2080

mengetahui akurasi hasil klasifikasi, dan dengan mengetahui jumlah data yang diklasifikasikasn secara salah, dapat mengetahui laju eror yang dilakukan

Untuk mengetahui akurasi digunakan formula :

$$\text{Akurasi} = \frac{Jumlah\ Data\ Yang\ Diprediksi\ Benar}{Jumlah\ Prediksi\ Yang\ Dilakukan} = \frac{f_{11}+f_{00}}{f_{11}+f_{10}+f_{01}+f_{00}} \quad (5)$$

Untuk mengetahui laju eror digunakan formula :

$$\text{Akurasi} = \frac{Jumlah\ Data\ Yang\ Diprediksi\ Salah}{Jumlah\ Prediksi\ Yang\ Dilakukan} = \frac{f_{10}+f_{01}}{f_{11}+f_{10}+f_{01}+f_{00}} \quad (6)$$

2.5 Algoritma C5.0

Algoritma C5.0 adalah perpanjangan dari algoritma C4.5 yang juga merupakan perpanjangan dari ID3. Ini adalah algoritma klasifikasi yang berlaku dalam kumpulan data besar. Lebih baik daripada C4.5 pada kecepatan, memori dan efisiensi. C5.0 bekerja dengan memisahkan sampel berdasarkan pada atribut yang menyediakan perolehan informasi maksimum. C5.0 dapat membagi atribut berdasarkan yang terbesar dari informasi gain. Proses akan berlanjut hingga bagian sampel tidak dapat dibagi. Algoritma C5.0 dapat menangani atribut kontinyu dan diskrit. Pemilihan atribut dalam algoritma ini akan diproses menggunakan information gain. Atribut dengan nilai Gain tertinggi akan dipilih sebagai akar bagi node selanjutnya. Berikut persamaan untuk menghitung entropy atribut [10].:

$$I(S_1,S_2,\ldots,\ldots,S_m) = -\sum_{i=1}^{m} p_i * log_2(p_i) \quad (7)$$

Dengan :
S : Himpunan kasus
m : Jumlah sampel
p_i : Proporsi kelas

Sementara untuk mendapatkan informasi nilai subset tersebut dapat dilihan pada pesamaan sebagai berikut:

$$E(A) = \sum_{j=i}^{y} \frac{S_{ij}+\cdots+S_{mj}}{S} I(S_{1j},\ldots,S_{mj}) \quad (8)$$

Dengan :

$\frac{S_{ij}+\cdots+S_{mj}}{S}$ = Jumlah subset J yang dibagi dengan jumlah sampel S

Maka untuk mendapatkan nilai gain, selanjutnya digunakan formula. Maka nilai gain dapat dihitung dengan formula sebagai berikut:

$$Gain(A) = I(S_1,S_2,\ldots,S_m) - E(A) \quad (9)$$

Dengan :
A = Atribut
S = Himpunan Kasus
S_1 = Jumlah Sampel

2.6 Algoritma Naïve Bayes Classifier (NBC)

Merupakan algoritma klasifikasi statistik dimana digunakan untuk melakukan prediksi secara probabilitas (kemungkinan) pada anggota suatu class. Dasar dari NBC merupakan *teorema bayes* dimana mempunyai kemampuan untuk melakukan klasifikasi, NBC hamper sama dengan pohon keputusan serta jaringan syaraf. NBC memiliki tingkatan akurasi yang tinggi jika digunakan pada basis data yang memiliki data yang besar[11-13]. *Teorema Bayes* memiliki persamaan umum :

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (10)$$

X = Data dengan class yang belum diketahui
H = Hipotesis data X merupakan suatu class spesifik
P(H|X) = Probabilitas hipotesis H berdasarkan kondisi x (posteriori prob.)
P(H) = Probabilitas hipotesis H (prior prob.)
P(X|H) = Probabilitas X berdasarkan kondisi tersebut
P(X) = Probabilitas dari X

3. HASIL DAN PEMBAHASAN

Penelitian ini menggunakan dataset yang berasal dari UCI Repository Machine Learning (https://archive.ics.uci.edu/ml/datasets/Z-Alizadeh+Sani#), dimana terdapat 57 atribut dan 2 target kelas pada dataset penyakit jantung tersebut. Penelitian ini melakukan reduksi pada atribut dataset penyakit jantung menggunakan metode Principal Component Analysis (PCA) kemudian membanding hasil akurasi menggunakan teknik klasifikasi datamining (Algoritma C5.0 dan Algoritma Naïve Bayes Classifier (NBC)). Pada dataset terdapat 58 atribut antara lain : *Age, Weight, Length, Sex, BMI, DM, HTN, Current Smoker, Ex-Smoker, FH,*

Submitted: **12/03/2020**; Accepted: **10/04/2020**; Published: **25/04/2020**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 4, Nomor 2, April 2020, Page 437-444
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v4i2.2080

Obesity, CRF, CVA, Airway Disease, Thyroid Disease, CHF, DLP, BP, PR, Edema, Weak Peripheral Pulse, Lung Rales, Systolic Murmur, Diastolic Murmur, Typical Chest Pain, Dyspnea, Function Class, Atypical, Nonanginal, Exertional CP, LowTH Ang, Q Wave, St Elevation, St Depression, Tinversion, LVH, Poor R Progression, BBB, FBS, CR, TG, LDL, HDL, BUN, ESR, HB, K, Na, WBC, Lymph, Neut, PLT, EF-TTE, Region RWMA, VHD, LAD, LCX, RCA. Selain attribute dataset juga memiliki 2 target kelas yaitu: *CAD* dan *Normal*. Dataset yang sudah tersedia kemudian diproses menggunakan metode Principan Component Analysis (PCA) untuk mencari nilai bobot pada setiap atribut, adapaun hasil proses dari metode Principan Component Analysis (PCA) dapat dilihat pada:

Tabel 2. Hasil Proses PCA

No	Eigenvalue	Proporition	Cumulative	Ranking Atribut
1	4,3572	0,07781	0,07781	LAD
2	2,95786	0,05282	0,13063	EF-TTE
3	2,92255	0,05219	0,18282	Age
4	2,31796	0,04139	0,22421	RCA
5	2,18986	0,0391	0,26331	Atipical
6	2,06329	0,03684	0,30016	Region RWMA
7	1,89088	0,03377	0,33392	Typical Chest Pain
8	1,81367	0,03239	0,36631	LCX
9	1,70156	0,03038	0,3967	Lymph
10	1,51847	0,02712	0,42381	Neut
11	1,46035	0,02608	0,44989	ESR
12	1,41198	0,02521	0,4751	Bun
13	1,3925	0,02487	0,49997	VHD
14	1,34324	0,02399	0,52395	FBS
15	1,27849	0,02283	0,54678	DM
16	1,20222	0,02147	0,56825	WBC
17	1,19958	0,021242	0,58967	PR
18	1,16927	0,02088	0,61055	HTN
19	1,14518	0,02045	0,631	Q Wave
20	1,06148	0,01895	0,64996	BP
21	1.04405	0.01864	0.6686	CR
22	1,01796	0,01818	0,68678	K
23	0,97336	0,01738	0,70416	Luog Rales
24	0,94376	0,01685	0,72101	St. Elevation
25	0,89966	0,01607	0,73708	Funciton Class
26	0,86241	0,0154	0,75248	Na
27	0,84482	0,01509	0,76757	HB
28	0,82143	0,01467	0,78223	Poor R Progressio
29	0,79051	0,01412	0,79635	Weight
30	0,78	0,01393	0,81028	BMI
31	0,73581	0,01314	0,82342	Edema
32	0,71835	0,01283	0,83625	Nonanginal
33	0,67783	0,0121	0,84835	Tinversion
34	0,66306	0,01184	0,86019	Sistolic Murmur
35	0,59889	0,01069	0,87089	CRF
36	0,57685	0,0103	0,88119	St. Depression
37	0,54691	0,00977	0,89095	HDL
38	0,524	0,00936	0,90031	CHF
39	0,51443	0,00919	0,9095	Obesitas
40	0,47134	0,00842	0,91791	Thiroid Disease
41	0,45084	0,00805	0,92596	LDL
42	0,44242	0,0079	0,93386	Airwa Disease
43	0,41603	0,00743	0,94129	Weak Peripheral Pulse
44	0,39414	0,00704	0,94833	FH
45	0,37304	0,00666	0,95499	Ex-Smoker
46	0,35826	0,0064	0,96139	Diastolic Murmur
47	0,33825	0,00604	0,96743	Dispoea
48	0,31973	0,00571	0,97314	CVA
49	0,3064	0,00547	0,97861	LVH
50	0,29548	0,00528	0,98389	Current Smoker
51	0,27182	0,00485	0,98874	DLP

Submitted: **12/03/2020**; Accepted: **10/04/2020**; Published: **25/04/2020**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 4, Nomor 2, April 2020, Page 437-444
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v4i2.2080

No	Eigenvalue	Proporition	Cumulative	Ranking Atribut
52	0,24763	0,00442	0,99316	TG
53	0,19774	0,00353	0,99669	Length
54	0,12144	0,00217	0,99886	BBB
55	0,0599	0,00107	0,9993	LowTH Aog
56	0,00374	0,00007	1	PLT
57	0,00000	0,0000	0,0000	Exertional CP

Tabel 2. merupakan proses perankingan nilai atribut yang paling berpengaruh, proses perankingan berdasarkan nilai eigenvalue yang paling besar hinggar yang paling kecil. Sehingga dapat diurutkan proses reduksi atribut dimulai dari atribut "Exertional CP" hingga atribut "LAD". Selanjutnya pada penelitian untuk proses klasifikasi menggunakan 10 atribut dengan nilai Eigenvalue terbesar yaitu: *LAD, EF-TTE, Age, RCA, Atipical, Region RWMA, Typical Chest Pain, LCX, Lymph, Neut.*

3.1 Hasil Pengujian

Proses pengujian pada penelitian ini membandingkan tingkat akurasi dengan menggunakan Algoritma C5.0 dan Algoritma Naïve Bayes Classifier, proses pengujian membandingkan kedua algoritma yang digunakan dan membandingkan akurasi yang didapatkan dari sebeleum dilakukan reduksi attribute dan sesudah dilakukan reduksi atribut. Akurasi dataset dapat dilihat dari hasil Recall dan Precision, Recall adalah tingkat keberhasilan dalam menemukan kembali sebuah informasi, Precision merupakan ketepatan tingkat akurasi antara informasi yang tersedia pada dataset dengan jawaban yang diberikan oleh hasil klasifikasi, sedangkan Akurasi adalah tingkat kedekatan antara nilai prediksi dan nilai actual. Hasil proses klasifikasi pada dataset penyakit jantung dapat dilihat dibawah ini:

Tabel 3. Hasil Klasifikasi Algoritma C5.0 Sebelum Reduksi

	True CAD	True Normal	Class Precision
Pred. CAD	203	1	99,51%
Pred. Normal	13	86	86.87%
Class Recall	93,98%	98,85%	**95,38%**

Tabel 4. Hasil Klasifikasi Algoritma Naïve Bayes Classifier Sebelum Reduksi

	True CAD	True Normal	Class Precision
Pred. CAD	215	2	99.08%
Pred. Normal	1	85	98,84%
Class Recall	99,54%	97,70%	**99,01%**

Tabel 5. Hasil Klasifikasi Algoritma C5.0 Sesudah Reduksi

	True CAD	True Normal	Class Precision
Pred. CAD	203	1	99,51%
Pred. Normal	13	86	86,87%
Class Recall	93,98%	98,85%	**95,38%**

Tabel 6. Hasil Klasifikasi Algoritma Naïve Bayes Classifier Sesudah Reduksi

	True CAD	True Normal	Class Precision
Pred. CAD	215	4	98,17%
Pred. Normal	1	83	98,81%
Class Recall	99,54%	95,40%	**98,53%**

Dari hasil proses klasifikasi maka dapat dirumuskan analisis untuk algoritma C5.0 tingkat akurasi pada dataset penyakit jantung yang didapatkan sebelum reduksi atribut dan sesudah reduksi atribut tidak mengalami perubahan baik dari Recall, Precision dan Akurasi. Dimana pada algoritma C5.0 akurasi yang didapatkan sebesar 93,38%. Pada algoritma Naïve Bayes Classifier (NBC) terdapat perbedaan tingkat akurasi yang didapatkan sebelum dilakukan reduksi sebesar 99,01% sedangkan setelah direduksi sebesar 98,53%. Hasil pengujian yang telah dilakukan bahwasannya algoritma Naïve Bayes Classifier (NBC) dapat menangani proses klasifikasi pada dataset penyakit jantung dengan kinerja lebih baik.

4. KESIMPULAN

Berdasarkan hasil penelitian dan pengujian yang telah dilakukan maka dapat ditarik kesimpulan adalah Kinerja Algoritma Naïve Bayes Classifer (NBC) lebih baik dari algoritma C5.0, hal ini dapat dilihat dari tingkat akurasi yang didapatkan baik sebelum dan sesudah dialukan reduksi. Algoritma Naïve Bayes Classifier (NBC) lebih baik

Submitted: **12/03/2020**; Accepted: **10/04/2020**; Published: **25/04/2020**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 4, Nomor 2, April 2020, Page 437-444
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v4i2.2080

dibandingkan dengan algoritma C5.0 hal ini dikarenakan kemampuan sederhana dalam proses klasifikasi dan membangun model. Untuk algoritma Naïve Bayes Classifier tidak perlu dibentuk sebuah rule, sedangkan pada algoritma C5.0 adanya terbentuk rule dari pohon keputusan yang dihasilkan.

REFERENCES

[1] R. Annisa, "ANALISIS KOMPARASI ALGORITMA KLASIFIKASI DATA MINING UNTUK PREDIKSI PENDERITA PENYAKIT JANTUNG," Jurnal Teknik Informatika Kaputama (JTIK), vol. 3, no. 1, pp. 22 - 28, 2019.
[2] D. B. Umadevi dan M. Snehapriya, "A Survey on Prediction of Heart Disease Using Data Mining Techniques," International Journal of Science and Research (IJSR), vol. 6, no. 4, pp. 2228 - 2232, 2017.
[3] A. Jain, M. Ahirwar dan R. Pandey, "A Review on Intutive Prediction of Heart Disease Using Data Mining Techniques," International Journal of Computer Sciences and Engineering, vol. 7, no. 7, pp. 109 - 113, 2019.
[4] R. Pandya dan J. Pandaya, "C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning," International Journal of Computer Applications, vol. 117, no. 16, pp. 18 - 21, 2015.
[5] S. F. Rodiyansyah dan E. Winarko, "KlasifikasPosting Twitter Kemacetan Lalu Lintas Kota Bandung Menggunakan NaiveBayesian Classification," Indonesian Journal of Computing and Cybernetics System (IJCCS), vol. 6, no. 1, pp. 91 - 100, 2012.
[6] S. Masripah, "Komparasi Algoritma Klasifikasi Data Mining untuk Evaluasi Pemberian Kredit," BINA INSANI ICT JOURNAL, vol. 3, no. 1, pp. 187-193, 2016.
[7] G. Rahayu dan Mustakim, "Principal Component Analysis untuk Dimensi Reduksi Data Clustering Sebagai Pemetaan Persentase Sertifikasi Guru di Indonesia," dalam Seminar Nasional Teknologi Informasi, Komunikasi dan Industri (SNTIKI) 9 , Riau, 2017.
[8] J. Han, M. Kamber dan J. Pei, Data Mining Concepts and Techniques, USA: Morgan Kaufmann, 2012.
[9] E. Prasetyo, Data Mining Mengolah Data Menjadi Informasi Menggunakan Matlab, Yogyakarta: Andi, 2014.
[10] B. R. Patel dan K. K. Rana, "A Survey on Decision Tree Algorithm For Classification," International Journal of Engineering Development and Research, vol. 2, no. 1, pp. 1-5, 2014.
[11] A. Jananto, "Algoritma Naive Bayes untuk Mencari Perkiraan Waktu Studi Mahasiswa," Jurnal Teknologi Informasi DINAMIK, vol. 18, no. 1, pp. 9 - 16, 2013.
[12] E. Buulolo, *Data Mining Untuk Perguruan Tinggi*, 1st ed. Yogyakarta: Deepublish, 2020.
[13] D. Laia, E. Buulolo, and M. J. F. S. Sirait, "Implementasi Data Mining Untuk Memprediksi Pemesanan Driver Go-Jek Online Dengan Menggunakan Metode Naive Bayes (Studi Kasus: Pt. Go-Jek Indonesia)," *KOMIK (Konferensi Nas. Teknol. Inf. dan Komputer)*, vol. 2, no. 1, pp. 434–439, 2018.

Submitted: **12/03/2020**; Accepted: **10/04/2020**; Published: **25/04/2020**