ISSN: 2614-1205

Optimasi Algoritma *Naïve Bayes* dengan Menerapkan *Discretization* dan *Particle Swarm Optimization* untuk Diagnosis *Breast Cancer Disease*

Fiscall Saktiyana Aditama[1], Endang Sugiharti[2]

[1,2]Jurusan Ilmu Komputer, Fakultas Matematika dan IPA,
Universitas Negeri Semarang, Indonesia
Email: [1]fiscalladitama@students.unnes.ac.id, [2]endangsugiharti@mail.unnes.ac.id

Abstrak

Data mining adalah proses mempekerjakan satu atau lebih *machine learning* untuk menganalisis dan mengekstraksi pengetahuan secara otomatis dengan pendekatan matematis. *Naïve Bayes* merupakan Algoritma yang sering digunakan dalam *data mining*. Data *Breast Cancer Coimbra Disease* yang digunakan diambil dari *UCI Machine Learning Repository*. *Discretization* digunakan untuk melakukan *preprocessing* data pada penelitian. Untuk meningkatkan akurasi, Algoritma *Naïve Bayes* dapat digabungkan dengan *feature selection Particle Swarm Optimization.* Penelitian bertujuan untuk mengetahui bagaimana penerapan Algoritma *Naïve Bayes* digabungkan dengan *Discretizationdan* dan *Particle Swarm Optimization* pada diagnosis *Breast Cancer Coimbra* Disease dan mengetahui peningkatan akurasinya. Berdasarkan hasil sepuluh *k-fold cross validation* hasil akurasi yang didapat diperoleh akurasi tertinggi *fold* ke 2 sebesar 91.67%, akurasi terendah pada *fold* ke 3 sebesar 41.67%, dan rata-rata akurasi sebesar 66.59%. Penerapan *Discretization* dan *Particle Swarm Optimization* mampu meningkatkan akurasi dari diagnosis *Breast Cancer Coimbra Disease* pada akurasi terbaik sebesar 8.34%, akurasi terendah sebesar 23.49%, dan rata-rata akurasi sebesar 11.82%.
Kata Kunci: Algoritma *Naïve Bayes, Discretization, Particle Swarm Optimization, Breast Cancer Coimbra Disease*

Abstract

Data mining is the process of employing one or more machine learning to analyze and extract knowledge automatically with a mathematical approach. Naïve Bayes is an algorithm that is often used in data mining. Data on Breast Cancer Coimbra Disease used were taken from the UCI Machine Learning Repository. Discretization is used to preprocess data in research. To improve accuracy, the Naïve Bayes Algorithm can be combined with the Particle Swarm Optimization feature selection. This study aims to determine how the application of the Naïve Bayes Algorithm is combined with Discretization and Particle Swarm Optimization in the diagnosis of Breast Cancer Coimbra Disease and to find out the increase in its accuracy. Based on the results of ten k-fold cross validation, the accuracy results obtained were the highest accuracy for the second fold of 91.67%, the lowest accuracy on the 3rd fold was 41.67%, and the average accuracy was 66.59%. The application of Discretization and Particle Swarm Optimization was able to increase the accuracy of the diagnosis of Breast Cancer Coimbra Disease with the best accuracy of 8.34%, the lowest accuracy of 23.49%, and the average accuracy of 11.82%.
***Keywords**: Naïve Bayes Algorithm, Discretization, Particle Swarm Optimization, Breast Cancer Coimbra Disease*

1. PENDAHULUAN

Perkembangan teknologi yang semakin cepat di era globalisasi pada dunia kesehatan, tidak menutup kemungkinan bahwa pengambilan suatu keputusan merupakan sesuatu yang sangat vital dalam menentukan keputusan yang harus diambil dalam menentukan dignosis suatu penyakit. Perkembangan teknologi informasi pada masa kini tumbuh dengan sangat pesat, hal tersebut menyebabkan terkumpulnya data dalam jumlah besar. Tersedianya dalam jumlah besar data tersebut menjadi informasi yang berguna jika dapat diolah dengan baik dan benar. Pengolahan dari kumpulan banyak data ini disebut dengan metode *data mining.*

Data mining adalah proses yang mempekerjakan satu atau lebih teknik pembelajaran komputer (*machine learning*) untuk menganalisis dan mengekstraksi pengetahuan (*knowledge*) secara otomatis[1]. Salah satu metode dalam *data mining* adalah klasifikasi[2]. Klasifikasi adalah proses untuk menemukan model yang menggambarkan dan membedakan kelas data[3]. Proses klasifikasi dengan berbagai metode dapat dimanfaatkan untuk mendiagnosis suatu penyakit, salah satunya penyakit kanker payudara. Dalam bidang kedokteran, data mining dapat digunakan untuk mendiagnosis beberapa penyakit seperti kanker payudara, penyakit jantung, diabetes, dll [20].

Kanker Payudara atau *Carcinoma Mamae* dalam bahasa inggrisnya disebut *Breast Cancer* merupakan kanker pada jaringan payudara[4]. Kanker payudara adalah pertumbuhan sel yang tidak terkendali pada kelenjar penghasil susu (*lobular*), saluran kelenjar dari *lobular* ke puting payudara (*duktus*), dan jaringan penunjang payudara yang mengelilingi *lobular*, *duktus*, pembuluh darah dan pembuluh limfe, tetapi tidak termasuk kulit payudara[5]. Dewasa ini terdapat penelitian yang mengembangkan teknik komputasi cerdas untuk mendiagnosis penyakit kanker payudara, salah satu yang pernah diusulkan adalah dengan menggunakan metode *Naïve Bayes*. Metode *Naïve Bayes*[6] diusulkan untuk mendiagnosis penyakit kanker payudara.

Metode *Naïve Bayes* dipilih karena banyak diimplementasikan dalam berbagai bidang ilmu. Salah satu Algoritma *data mining* yang terbaik dan banyak digunakan untuk klasifikasi *dataset* nominal adalah *Naïve Bayes*[7]. Meski demikian faktor negatif dari *dataset* seperti *noise*, nilai-nilai yang hilang, dan data yang tidak konsisten sangat mempengaruhi keberhasilan metode yang digunakan. Dengan demikian *preprosesing* data menggunakan metode *discretization* digunakan pada *dataset* untuk mendapatkan set data akhir yang dapat dianggap benar dan berguna untuk Algoritma penambangan data lebih lanjut[8]. *Discretization* adalah metode yang bertujuan mengurangi jumlah nilai yang berbeda untuk variabel kontinu yang diberikan dengan membagi rentangnya menjadi seperangkat interval terpisah yang terbatas, dan kemudian menghubungkan interval ini dengan label yang bermakna sehingga dapat mengurangi permintaan memori sistem dan meningkatkan efisiensi Algoritma[9]. Selain menggunakan metode *preprocessing Discretization, feature selection* menggunakan *Particle Swarm Optimization* untuk optimasi atribut *dataset* dapat dikombinasikan dengan *classifier* Algoritma yang lain untuk meningkatkan performa klasifikasi.

Penelitian ini bertujuan untuk mengetahui peningkatan akurasi Algoritma *Naïve Bayes* sebelum dan sesudah penambahan *Discretization* dan *Particle Swarm Optimization* dalam diagnosis *Breast Cancer Coimbra Disease*. Metode *Discretization* dan *Particle Swarm Optimization* diusulkan untuk meningkatkan akurasi Algoritma *Naïve Bayes* dalam diagnosis *Breast Cancer Coimbra Disease* menggunakan *dataset* yang diperoleh dari *Machine Learning Repository of UCI*.

2. METODE PENELITIAN

Penelitian ini menggabungkan metode *Discretization* dan *Particle Swarm Optimization* dengan Algoritma *Naïve Bayes* untuk meningkatan akurasi diagnosis *Breast Cancer Coimbra Disease*. *Dataset Breast Cancer Coimbra Disease* diperoleh dari *machine learning repository of UCI* (https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra). *Dataset* ini berasal dari *Laboratory of Biostatistics and Medical Informatics* and *IBILI - Faculty of Medicine, University of Coimbra, Azinhaga Santa Comba, Celas*, 3000-548 *Coimbra, Portugal*. Data yang dikumpulkan berisikan data 64 wanita sakit dan 52 wanita sehat. Jadi, *dataset* berisi 116 contoh. Data pasien dikumpulkan sebelum operasi dan perawatan [10].

Variabel kategoris menunjukkan nilai 1 dan 2, yang sesuai, masing-masing untuk wanita sehat dan sakit. *Dataset* lengkap, tidak mengandung nilai yang hilang. Deskripsi *dataset* dapat dilihat pada Tabel 1. [10] [11].

Tabel 1. *Dataset Breast Cancer*

Atribut	Deskripsi
Usia	Usia pasien: 24 hingga 89 dalam tahun
BMI	Indeks massa tubuh: 18,37 to 38,58 kg/m
Glukosa	Jumlah gula dalam darah: 60 hingga 201 mg / dL
Insulin	Hormon yang diproduksi oleh pankreas untuk mengurangi kadar glukosa dalam darah: 2,432 hingga 58,46 μU / mL
HOMA	*Homeostatic Model Assessment*, adalah metode yang digunakan untuk mengukur resistensi insulin[12]: 0,467 hingga 25,05
Leptin	Protein yang bertanggung jawab untuk mengendalikan makanan yang dicerna, mengirim informasi ke otak[13]: 4,3 hingga 90,3 ng / mL
Adiponektin	Protein yang bertanggung jawab untuk pengaturan glukosa dalam darah: 1,66 hingga 38,04 ng / mL
Resistin	Protein yang bertanggung jawab untuk memblokir aksi utama leptin 3,21 hingga 82,1 ng / mL
MCP-1	*Monocyte Chemoattractant Protein* 1, merekrut *monocytes* dan sel-sel spesifik ke tempat-tempat peradangan.

2.1 *Discretization*

Discretization atau diskritisasi adalah proses *transfer* fungsi berkelanjutan, model dan persamaan nilai diskrit ekuivalen. Diskritisasi memainkan peran penting pada proses *data preprocessing* dalam *machine learning*[14]. Metode diskritisasi tidak hanya dapat mengurangi permintaan memori sistem dan meningkatkan efisiensi Algoritma *data mining* dan *machine learning*, tapi juga membuat pengetahuan yang diambil dari *dataset Discretized* lebih ringkas, mudah dipahami dan mudah digunakan[15].

Proses dari *discretization* adalah menemukan jumlah interval diskrit, dan kemudian lebar, atau batas untuk interval, memberi rentang nilai atribut kontinu. Pada penelitian ini data dibagi menjadi dua interval. Pada interval pertama data diberi label 0, dan pada interval kedua diberi label 1.

Discretization dipandang sebagai partisi dari atribut bernilai kontinu ke dalam atribut diskrit berurutan dengan sejumlah interval diskrit, yang setara dengan proses pengurangan jumlah keadaan variabel acak diskrit berurutan dengan menggabungkan beberapa keadaannya bersama-sama.

2.2 Algoritma *Naïve Bayes*

Naïve Bayes merupakan salah satu Algoritma yang terdapat pada teknik klasifikasi. *Naïve Bayes* merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukan oleh ilmuwan Inggris *Thomas Bayes*, yaitu memprediksi peluang di masa depan berdasarkan pengalaman dimasa sebelumnya sehingga dikenal sebagai Teorema *Bayes*. Teorema tersebut dikombinasikan dengan *Naïv*e dimana diasumsikan kondisi antar atribut saling bebas. Penggunaan *Naive Bayes* akan lebih baik jika lebih banyak data pelatihan. Diperlukan data latih seakurat mungkin dan hasilnya akan lebih baik [21]. Klasifikasi *Naive Bayes* diasumsikan bahwa ada atau tidak ciri tertentu dari sebuah kelas tidak ada hubungannya dengan ciri dari kelas lainnya [16].

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \quad (1)$$

Keterangan:
X : Data dengan class yang belum diketahui
H : Hipotesis data X merupakan suatu class spesifik
P(H|X) : Probabilitas hipotesis H berdasar kondisi X (posteriori probability)
P(H) : Probabilitas hipotesis H (prior probability)
P(X|H) : Probabilitas X berdasarkan kondisi pada hipotesis H
P(X) : Probabilitas X Karena dalam klasifikasi nilai P(X) selalu konstan untuk semua kelas, maka:

$$P(H|X) = P(X|H).P(H) \quad (2)$$

2.3 *Particle Swarm Optimization*

Particle Swarm Optimization (PSO) merupakan Algoritma pencarian berbasis populasi yang diinisialisasi dengan populasi solusi acak, dan digunakan untuk memecahkan masalah optimasi. PSO diperkenalkan oleh Kennedy dan Eberhart pada

tahun 1995 berdasarkan penelitian terhadap perilaku kawanan burung dan ikan. *Particle Swarm Optimization* (PSO) sering digunakan dalam penelitian, karena PSO memiliki. Algoritma PSO dapat meningkatkan bobot atribut dan meningkatkan akurasi suatu Algoritma dan klasifikasi data yang lebih besar [17]. Oleh karena itu, partikel memiliki kecenderungan untuk terbang menuju daerah pencarian yang lebih baik dan lebih baik selama proses pencarian [18]. PSO sebagai fitur seleksi dapat meningkatkan performa algoritma klasifikasi untuk diagnosis kanker payudara[20]

Naïve Bayes* dalam *Particle Swarm Optimization

PSO diterapkan pada pembobotan atribut seperti Algoritma dibawah ini:

a. Identifikasi populasi sampel
b. Hitung $P(C_i)$ pada setiap kelas
c. Inisialisasi posisi setiap partikel atribut ke-j
d. Untuk Setiap Atribut dilakukan
   - Evaluasi nilai fungsi tujuan
   - Cari Pbest dan Gbest
   - *Update* kecepatan dan posisi *particle*
   - Gbest = bobot atribut ke-j
e. hitung $P(X|C_i)$, i=1,2 untuk setiap kelas atau atribut
f. Bandingkan hasil $P(X|C_i)$

Identifikasi populasi sampel dari *dataset Breast Cancer Coimbra Dataset* (BCCD). Hitung $P(C_i)$ untuk setiap kelas, dalam kasus data set pada penelitian ini terdiri dari 2 kelas yaitu pasien sehat dan sakit. Inisialisasi posisi setiap partikel atribut ke-j merupakan awal dari tahap pembobotan atribut dengan PSO. Langkah selanjutnya adalah evaluasi nilai fungsi tujuan dari setiap partikel untuk mendapatkan posisi terbaik (Pbest) dan posisi global terbaik (Gbest), kemudian *update* kecepatan dan posisi partikel.

Ulangi langkah evaluasi nilai fungsi tujuan sampai mencapai konvergen, kemudian Gbest = bobot atribut ke-j. Cek apakah nilai j sudah maksimal, jika belum ulangi langkah-langkah dari inisialisasi posisi setiap partikel atribut ke-j sampai menemukan bobot atribut ke-j. Ulangi langkah tersebut sampai nilai j sudah maksimal atau semua atribut sudah terbobot. Kemudian hitung $P(X|C_i)$, i=1,2 untuk setiap kelas atau atribut. Setelah itu bandingkan, jika $P(X|C_1) > P(X|C_2)$ maka kesimpulannya adalah C_1 atau dalam kasus pada penelitian ini berarti pasien sehat. Jika $P(X|C_1) < P(X|C_2)$ maka kesimpulannya C_2 atau pasien sakit kanker payudara.

3. HASIL DAN PEMBAHASAN

3.1 Hasil

Pada penelitian ini, Sebelum memasuki proses utama *data mining*, data diolah pada proses *preprocessing.* Proses ini sangat penting untuk menyiapkan data yang cocok agar proses *data mining* menghasilkan akurasi yang tinggi. Normalisasi data, *data transformation*, dan *discretization* digunakan untuk melakukan *preprocessing* data pada penelitian. *Dataset Breast Cancer Coimbra* yang digunakan tampak pada Tabel 2.

Tabel 2. *Breast Cancer Coimbra Dataset*

No.	*Age*	*BMI*	*Glu*	*Ins*	*HOMA*	*Lept*	*Adip*	*Res*	*MCP.1*	*Class*
1	48	23,5	70	2,707	0,467409	8,8071	9,7024	7,99585	417,114	1
2	83	20,69049	92	3,115	0,706897	8,8438	5,429285	4,06405	468,786	1
3	82	23,12467	91	4,498	1,009651	17,9393	22,43204	9,27715	554,697	1
4	68	21,36752	77	3,226	0,612725	9,8827	7,16956	12,766	928,22	1
5	86	21,1111	92	3,549	0,805386	6,6994	4,81924	10,57635	773,92	1
6	49	22,85446	92	3,226	0,732087	6,8317	13,67975	10,3176	530,41	1
7	89	22,7	77	4,69	0,890787	6,964	5,589865	12,9361	1256,083	1
	...	...	...	...	...	...	...	...	...	...
110	75	30,48	152	7,01	2,628283	50.53	10,06	11,73	99,45	2
111	54	36,05	119	11,91	3,495982	89,27	8,01	5,06	218,28	2
112	45	26,85	92	3,33	0,755688	54,68	12,1	10,96	268,23	2
113	62	26,84	100	4,53	1,1174	12,45	21,42	7,32	330,16	2
114	65	32,05	97	5,73	1,370998	61,48	22,54	10,33	314,05	2
115	72	25,59	82	2,82	0,570392	24,96	33,75	3,27	392,46	2
116	86	27,18	138	19,91	6,777364	90,28	14,11	4,35	90,09	2

Data kemudian dilakukan *preprocessing*, *Discretization* diterapkan untuk mengurangi jumlah nilai yang berbeda pada variabel kontinu yang diberikan dengan membagi rentangnya menjadi seperangkat interval terpisah yang terbatas, dan kemudian menghubungkan interval ini dengan label yang bermakna. Pada penelitian ini, *discretization* membagi nilai dari setiap atribut menjadi dua interval. Tabel 3 menunjukan *dataset Breast Cancer Coimbra* setalah dilakukan *Discretization.*

Tabel 3. *Breast Cancer Coimbra Dataset* Setelah *Discretization*

No.	*Age*	*BMI*	*Glu*	*Ins*	*HOMA*	*Lept*	*Adip*	*Res*	*MCP.1*	*Class*
1	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	1.0	1
2	0.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1
3	0.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	0.0	1
4	0.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	1
5	0.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	0.0	1

No.	*Age*	*BMI*	*Glu*	*Ins*	*HOMA*	*Lept*	*Adip*	*Res*	*MCP.1*	*Class*
6	1.0	1.0	0.0	1.0	1.0	1.0	0.0	1.0	0.0	1
7	0.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	1
	...	...	...	...	...	...	...	...	...	...
110	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	2
111	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	2
112	1.0	1.0	0.0	1.0	1.0	0.0	0.0	1.0	1.0	2
113	0.0	1.0	0.0	1.0	1.0	1.0	0.0	1.0	1.0	2
114	0.0	0.0	0.0	1.0	1.0	0.0	0.0	1.0	1.0	2
115	0.0	1.0	1.0	1.0	1.0	0.0	0.0	1.0	1.0	2
116	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	2

Setelah data dilakukan *preprocessing, feature selection Particle Swarm Optimization* diterapkan untuk mengoptimasikan atribut – atribut pada data *Breast Cancer Disease.* Interval 0 sampai 1 digunakan untuk menentukan atribut yang ada pada *dataset* baik digunakan atau tidak. Atribut menghasilkan nilai 0 menunjukkan bahwa atribut sangat buruk jika akan dilakukan proses *Mining* dengan Algoritma *Naïve Bayes.* Sebaliknya jika atribut menunjukkan nilai 1 maka atribut sangat baik dan layak untuk dilakukan proses *Mining* dengan Algoritma *Naïve Bayes* mengurangi jumlah nilai yang berbeda pada variabel kontinu yang diberikan dengan membagi rentangnya menjadi seperangkat interval terpisah yang terbatas, dan kemudian menghubungkan interval ini dengan label yang bermakna. Pada penelitian ini, *discretization* membagi nilai dari setiap atribut menjadi dua interval. Selanjutnya pada setiap interval diberi label dengan nama 0, dan 1. Tabel 4 menunjukan *dataset Breast Cancer Coimbra Disease* setalah dilakukan *Discretization.*

Tabel 4. *Breast Cancer Coimbra Disease Dataset* Setelah PSO

Age	*BMI*	*Glu*	*Ins*	*HOMA*	*Lept*	*Adip*	*Res*	*MCP.1*
0	0	1	0	1	0	1	0	1

Data yang telah melalui tahap *preprocessing* dan *feature selection* kemudian dilakukan proses *data mining.* Pada tahap ini dilakukan dua kali proses *data mining,* proses pertama yaitu proses klasifikasi Algoritma *Naïve Bayes* dengan menggunakan *dataset* tanpa *Discretization dan Particle Swarm Optimization* dan proses kedua yaitu klasifikasi Algoritma *Naïve Bayes* digabungkan dengan *dataset yang* telah dilakukan proses *Discretization dan Particle Swarm Optimization*.

Pada percobaan pertama, *dataset Breast Cancer Coimbra Disease* tanpa *discretization* dilakukan klasifikasi menggunakan Algoritma *Naïve Bayes* dengan *validation method* yaitu *k-fold cross validation* dengan nilai k=10. *Dataset Breast Cancer Coimbra* dibagi menjadi dua, yaitu data latih dan data uji menggunakan *cross validation*. Data

latih diproses menggunakan Algoritma *Naïve Bayes* untuk menghasilkan pengujian model. Pengujian model diuji menggunakan data uji. *Confusion matrix* digunakan untuk mengukur kinerja Algoritma. Hasil akurasi Algoritma *Naïve Bayes* menggunakan sepuluh *k-fold cross validation* ditunjukan pada Tabel 5.

Tabel 5. Hasil Sepuluh *K-fold Cross Validation* Akurasi Algoritma *Naïve Bayes*

Fold	Akurasi (%)
1	83.33
2	83.33
3	58.33
4	83.33
5	33.33
6	33.33
7	45.45
8	72.73
9	18.18
10	36.36

Hasil dari proses klasifikasi tersebut nantinya akan dibandingkan dengan klasifikasi menggunakan Algoritma *Naïve Bayes* digabungkan *Discretization* dan *Particle Swarm Optimization.* Penerapan Algoritma *Naïve Bayes* pada *dataset Breast Cancer Coimbra Disease* menghasilkan akurasi terbaik pada *fold* ke 2 sebesar 83.33%, akurasi terendah pada *fold* ke 9 sebesar 18.18%, dan rata-rata akurasi sebesar 54.77% yang dihasilkan dari sepuluh *k-fold cross validation*.

Pada percobaan kedua, *dataset Breast Cancer Coimbra Disease* yang telah dilakukan *Discretization* diklasifikasi menggunakan Algoritma *Naïve Bayes* dan *Particle Swarm Optimization* dengan *validation method* yaitu *k-fold cross validation* dengan nilai k=10.

Dataset Breast Cancer Coimbra Disease dibagi menjadi dua, yaitu data latih dan data uji menggunakan *cross validation* secara acak. Kemudian pada data latih dilakukan *feature selection dengan Particle Swarm Optimization*. Data latih tersebut diproses dengan Algoritma *Naïve Bayes* dan didapatkan pengujian model. Pengujian model kemudian diuji dengan menggunakan data uji. *Confusion matrix* digunakan untuk mengukur kinerja Algoritma. Hasil akurasi Algoritma *Naïve Bayes* dengan *Discretization* dan *Particle Swarm Optimization* menggunakan sepuluh *k-fold cross validation* ditunjukan ditunjukan pada tabel 6.

Tabel 6. Hasil Sepuluh *K-fold Cross Validation* Akurasi Algoritma *Naïve Bayes* dengan *Discretization* dan *Particle Swarm Optimization*

Fold	Akurasi (%)
1	58.33
2	91.67
3	41.67
4	50.00
5	75.00
6	58.33
7	81.82
8	63.64
9	63.64
10	81.82

Penerapan klasifikasi menggunakan Algoritma *Naïve Bayes* digabungkan dengan *Discretization* dan *Particle Swarm Optimization* pada *dataset Breast Cancer Coimbra Disease* menghasilkan nilai akurasi terbaik pada *fold* ke 2 sebesar 91.67%, akurasi terendah pada *fold* ke 3 sebesar 41.67%, dan rata-rata akurasi sebesar 66.59% yang dihasilkan dari sepuluh *k-fold cross validation.*

3.2 Pembahasan

Peningkatan akurasi Algoritma *Naïve Bayes* dalam diagnosis *Breast Cancer Coimbra Disease* menggunakan *Discretization* dan *Particle Swarm Optimization* memiliki tiga tahapan. Tahap yang pertama yaitu pengambilan data, tahap kedua yaitu pengolahan data, dan tahap ketiga yaitu proses *data mining*.

Diagnosis *Breast Cancer Coimbra Disease* menggunakan metode klasifikasi Algoritma *Naïve Bayes* memperoleh hasil akurasi terbaik pada *fold* ke 2 sebesar 83.33%, akurasi terendah pada *fold* ke 9 sebesar 18.18%, dan rata-rata akurasi sebesar 54.77% yang didapat dari hasil sepuluh *k-fold cross validation*, sedangkan Diagnosis *Breast Cancer Coimbra Disease* menggunakan metode klasifikasi Algoritma *Naïve Bayes* dengan *Particle Swarm Optimization* dan *dataset* yang telah melalui proses *Discretization* memperoleh nilai akurasi terbaik pada *fold* ke 2 sebesar 91.67%, akurasi terendah pada *fold* ke 3 sebesar 41.67%, dan rata-rata akurasi sebesar 66.59% yang didapat dari hasil sepuluh *k-fold cross validation*. Berdasarkan hasil akurasi dari dua kali proses *data mining* tersebut, penerapan *Discretization* dan *Particle Swarm Optimization* pada Algoritma klasifikasi *Naïve Bayes* dapat meningkatkan akurasi terbaik sebesar 8.34%, akurasi terendah sebesar 23.49%, dan rata-rata akurasi sebesar 11.82% pada Diagnosis *Breast Cancer Coimbra Disease*.

Dengan tingkat akurasi yang diberikan, model ini dapat dibuktikan mampu melakukan Diagnosis *Breast Cancer Coimbra Disease* pada *dataset Breast Cancer Coimbra Disease UCI Machine Learning Repository* dengan baik. Untuk mengetahui bahwa metode ini lebih baik dari metode yang sudah ada, dilakukan perbandingan dengan penelitian sebelumnya yang menggunakan *dataset* dan metode yang sama. Tabel perbandingan akurasi diagnosis *Breast Cancer Coimbra Disease* ditunjukan pada Tabel 7.

Tabel 7. Perbandingan Akurasi *Breast Cancer Coimbra Disease*

Penulis	*Dataset*	Metode	Akurasi (%)
(Wiswandani, 2018)	*UCI*	*Naïve Bayes,* *Decision Tree,* *SVM linear kernel,* *Random Forest,* *SVM kernel rbf*	61,30 46,60 53,40 64,60 69,70
Proposed Method	*UCI*	*Naïve Bayes + Discretization + Particle Swarm Optimization fold 2*	91,67
Proposed Method	*UCI*	*Naïve Bayes + Discretization + Particle Swarm Optimization fold 3*	41,67
Proposed Method	*UCI*	*Naïve Bayes + Discretization + Particle Swarm Optimization*	66,59

Pada penelitian ini, penulis menerapkan *Discretization* dan *Particle Swarm Optimization* pada Algoritma *Naïve Bayes*. Penelitian ini memiliki akurasi yang lebih tinggi dari penelitian sebelumnya yang dilakukan[19] di mana pada penelitian tersebut metode *Naïve Bayes*, *SVM* dengan *rbf kernel* dan metode *kernel linier*, *kNN*, R*andom Forest* dan *Decision Tree* digunakan pada *dataset Breast Cancer Coimbra Disease* menghasilkan akurasi *Naïve Bayes* 61.3%, *SVM kernel rbf* 69.7%, *SVM linear kernel* 53.4%, *Random Forest* 64.6%, *Decision Tree* 46.6%.

Terjadinya peningkatan secara signifikan tersebut karena proses *Discretization* dapat mengurangi permintaan memori sistem dan meningkatkan efisiensi Algoritma *data mining* dan juga membuat pengetahuan yang diambil dari *dataset Discretized* lebih ringkas, mudah dipahami dan digunakan. Hal ini juga terjadi karena penerapan *Particle Swarm Optimization* yang mengoptimasi atribut dari *dataset* untuk mendapatkan model klasifikasi yang lebih kuat dibandingkan dengan model klasifikasi yang didapatkan dari proses Algoritma *Naïve Bayes*. Sehingga metode tersebut dapat meningkatkan akurasi Diagnosis *Breast Cancer Coimbra Disease* dengan menggunakan Algoritma *Naïve Bayes*.

4. SIMPULAN

Berdasarkan hasil penelitian peningkatan akurasi Algoritma *Naïve Bayes* dalam diagnosis *Breast Cancer Coimbra Disease* menggunakan *Discretization* dan *Particle Swarm Optimization* menunjukan bahwa metode *preprocesing Discretization* dan *Particle Swarm Optimization* dapat meningkatkan hasil dari Algoritma *Naïve Bayes* dalam melakukan diagnosis *Breast Cancer Coimbra Disease*. Dari proses klasifikasi *dataset Breast Cancer Coimbra Disease* menggunakan Algoritma *Naïve Bayes* diperoleh hasil akurasi terbaik pada *fold* ke 2 sebesar 83.33%, akurasi terendah pada *fold* ke 9 sebesar 18.18%, dan rata-rata akurasi sebesar 54.77%, sedangkan setelah ditambahkan *Discretization* dan *Particle Swarm Optimization* diperoleh akurasi terbaik pada *fold* ke 2 sebesar 91.67%, akurasi terendah pada *fold* ke 3 sebesar

41.67%, dan rata-rata akurasi sebesar 66.59%. Penggunaan *Discretization* dan *Particle Swarm Optimization* pada algortima *Naïve Bayes* berhasil meningkatkan akurasi terbaik sebesar 8.34%, akurasi terendah sebesar 23.49%, dan rata-rata akurasi sebesar 11.82% dibandingkan dengan hanya menggunakan Algoritma *Naïve Bayes*.

5. REFERENSI

[1] Hermawati, F. A. (2013). Data Mining, Yogyakarta: CV. *Andi Offset*.
[2] Muslim, M. A., Nurzahputra, A., & Prasetiyo, B. (2018). Improving Accuracy of C4.5 Algorithm Using Split Feature Reduction Model and Bagging Ensemble for Credit Card Risk Prediction. In *2018 International Conference On* Information *and Communications Technology (ICOIACT)* (Pp. 141-145). IEEE
[3] Han, J., Kamber, M., & Pei, J. (2011). Data Mining Concepts and Techniques Third Edition. *Morgan Kaufmann*.
[4] A. Aswita and D. F. A. Putri, Sistem Pakar Diagnosa Penyakit Kanker Payudara Menggunakan Certainty Factor.
[5] Hutapea, M. (2017). Pengaruh Pelaksanaan Pemeriksaan Payudara Sendiri (Sadari) Terhadap Pengetahuan dan Kemampuan Siswi Dalam Upaya Deteksi Dini Kanker Payudara Sma Swakarya Tahun 2017. *Jurnal Riset Hesti Medan Akper Kesdam I/BB Medan*, *2*(2), 105–116.
[6] D. Dumitru, "Prediction of Recurrent Events in Breast Cancer Using the Naive Bayesian Classification," vol. 36(2), 2009.
[7] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., & Philip, S. Y. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *14*(1), 1–37.
[8] García, S., Gallego, S.R., Luengo, J., Benítez, J. M., & Herrera, F. (2016). Big Data Preprocessing: Methods and Prospects. *Big Data Analytics*, 1(1), 1-25.
[9] Dash, R., Paramguru, R. L., & Dash, R. (2011). Comparative Analysis of Supervised and Unsupervised Discretization Techniques. *International Journal of Advances in Science and Technology*, *2*(3), 29-37.
[10] Patrício, M., Pereira, J., Crisóstomo, J., Matafome, P., Gomes, M., Seiça, R., & Caramelo, F. (2018). Using Resistin, glucose, age and BMI to predict the presence of breast cancer. *BMC Cancer*, *18*(1), 1–8.
[11] Frazão, A. (2018). Exame da Glicose: como é feito e valores de referência. Retrieved from https://www. tuasaude.com/exame-da-glicose/.
[12] Lemos, M. (2018). Para que serve o índice HOMA. Retrieved from https://www.tuasaude.com/para- queserve-o-indice-homa/.
[13] Gunnars, K. (2018). Leptin and Leptin Resistance: Everything You Need To Know. Retrieved from https://www.healthline.com/nutrition/leptin-101/.
[14] Kapoor, P., Arora, D., & Kumar, A. (2017). Implications of discretization towards improving classification accuracy for software defect data. *Journal of Theoretical and Applied Information Technology*, *95*, 6893–6901.
[15] Dash, R., Paramguru, R. L., & Dash, R. (2011). Comparative analysis of supervised and unsupervised discretization techniques. *International Journal of Advances in Science and Technology*, *2*(3), 29–37.
[16] Bustami, Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi.
[17] Cao, J., Cui, H., Shi, H., & Jiao, L. (2016). Big data: A parallel particle swarm

optimization-back-propagation neural network algorithm based on MapReduce. *PloS One*, *11*(6), e0157551.

[18] Grosan, C., Abraham, A., & Chis, M. (2006). Swarm intelligence in data mining. In *Swarm Intelligence in Data Mining* (pp. 1–20). Springer.

[19] Wiswandani, A. (2018). *Analisis Klasifikasi pada Data Breast Cancer Coimbra Menggunakan Metode Machine Learning*.

[20] Muslim, M. A., Rukmana, S. H., Sugiharti, E., Prasetiyo, B., & Alimah, S. (2018, March). Optimization of C4. 5 algorithm-based particle swarm optimization for breast cancer diagnosis. In *Journal of Physics: Conference Series* (Vol. 983, No. 1, p. 012063). IOP Publishing.

[21] Sugiharti, E., Firmansyah, S., & Devi, F. R. (2017). Predictive evaluation of performance of computer science students of unnes using data mining based on naÏve bayes classifier (NBC) algorithm. *Journal of Theoretical and Applied Information Technology*, *95*(4), 902.