OPTIMASI ALGORITMA C4.5 MENGGUNAKAN SELEKSI FITUR *PARTICLE SWARM OPTIMIZATION* (PSO) DAN TEKNIK *BAGGING* PADA DIAGNOSIS PENYAKIT KANKER PAYUDARA

Skripsi

disusun sebagai salah satu syarat

untuk memperoleh gelar Sarjana Komputer

Program Studi Teknik Informatika

oleh

Raka Hendra Saputra
4611415027

JURUSAN ILMU KOMPUTER

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS NEGERI SEMARANG

2020

PERNYATAAN

Saya menyatakan dengan sebenar-benarnya bahwa skripsi saya yang berjudul "Optimasi Algoritma C4.5 Menggunakan Seleksi Fitur *Particle Swarm Optimization* (PSO) dan Teknik *Bagging* pada Diagnosis Penyakit Kanker Payudara" disusun atas dasar penelitian saya dengan arahan dosen pembimbing. Sumber informasi atau kutipan yang berasal dari karya yang diterbitkan telah disebutkan dalam teks dan dicantumkan dalam daftar pustaka di bagian akhir skripsi ini. Saya menyatakan bahwa skripsi ini bebas plagiat, dan apabila di kemudian hari terbukti terdapat plagiat dalam skripsi ini, maka saya bersedia menerima sanksi sesuai ketentuan peraturan perundang-undangan.

Semarang, 11 Juni 2020

Raka Hendra Saputra

4611415027

PERSETUJUAN PEMBIMBING

Nama : Raka Hendra Saputra

NIM : 4611415027

Program Studi : S-1 Teknik Informatika

Judul Skripsi : Optimasi Algoritma C4.5 Menggunakan Seleksi Fitur *Particle Swarm Optimization* (PSO) dan Teknik *Bagging* pada Diagnosis Penyakit Kanker Payudara

Skripsi ini telah disetujui oleh pembimbing untuk diajukan ke sidang panitia ujian skripsi Program Studi Teknik Informatika FMIPA UNNES.

Semarang, 11 Juni 2020

Pembimbing

Budi Prasetiyo, S.Si., M.Kom.

NIP. 198805012014041001

PENGESAHAN

Skripsi yang berjudul

Optimasi Algoritma C4.5 Menggunakan Seleksi Fitur *Particle Swarm Optimization* (PSO) dan Teknik *Bagging* pada Diagnosis Penyakit Kanker Payudara

Disusun oleh

Raka Hendra Saputra

4611415027

Telah dipertahankan di hadapan sidang panitia ujian skripsi FMIPA UNNES pada tanggal 10 Maret 2020

Panitia:

Ketua

Dr. Sugianto, M.Si.

NIP 196102191993031001

Sekretaris

Dr. Alamsyah, S.Si., M.Kom.

NIP 197405172006041001

Penguji 1

Aji Purwinarko, S.Si., M.Cs.

NIP 198509102015041001

Penguji 2

Dr. Alamsyah, S.Si., M.Kom.

NIP 197405172006041001

Anggota Penguji

Budi Prasetiyo, S.Si., M.Kom.

NIP. 198805012014041001

MOTTO DAN PERSEMBAHAN

MOTTO

- Jangan melihat kesuksesan orang lain sebagai perbandingan tapi lihatlah sebagai acuan.
- Teruslah berbuat kebaikan sampai orang lain lupa akan keburukanmu.

PERSEMBAHAN

Skripsi ini ku persembahkan kepada:

- Kedua Orang Tua saya Bapak Fauzi Johanis dan Ibu Elimarni yang selalu memberikan kasih sayang, doa, dan dukungannnya.
- Kakak saya Andri Nofa dan Febri Fauzi yang telah memberikan dukungan moril dan materiil serta doa yang terus dipanjatkan.
- Teman-teman saya di jurusan Ilmu Komputer, Fakultas MIPA, serta teman-teman di Universitas Negeri Semarang
- Semua pihak yang tidak dapat disebutkan satu persatu yang telah membantu hingga terselesaikannya penulisan skripsi ini.
- Almamaterku, Universitas Negeri Semarang.

PRAKATA

Puji syukur penulis panjatkan kepada Allah *Subhanahu wa ta'ala* atas berkat rahmat dan hidayah-Nya penulis dapat menyelesaikan skripsi yang berjudul "Optimasi Algoritma C4.5 Menggunakan Seleksi Fitur *Particle Swarm Optimization* (PSO) dan Teknik *Bagging* pada Diagnosis Penyakit Kanker Payudara".

Penulis menyadari bahwa penulisan skripsi ini tidak akan selesai tanpa adanya dukungan serta bantuan dari berbagai pihak. Oleh karena itu, penulis ingin menyampaikan ucapan terima kasih kepada:

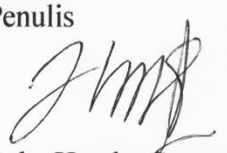
1. Prof. Dr. Fathur Rokhman, M.Hum., Rektor Universitas Negeri Semarang.
2. Dr. Sugianto, M.Si., Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Negeri Semarang.
3. Dr. Alamsyah, S.Si., M.Kom., Ketua Jurusan Ilmu Komputer FMIPA Universitas Negeri Semarang sekaligus dosen penguji yang telah meluangkan waktu, membantu, membimbing, mengarahkan dan memberikan saran sehingga penulis dapat menyelesaikan skripsi ini.
4. Budi Prasetiyo, S.Si., M.Kom., Dosen Pembimbing yang telah meluangkan waktu, membantu, membimbing, mengarahkan dan memberikan saran sehingga penulis dapat menyelesaikan skripsi ini.
5. Aji Purwinarko, S.Si., M.Cs., Dosen Penguji yang telah meluangkan waktu, membantu, membimbing, mengarahkan dan memberikan saran sehingga penulis dapat menyelesaikan skripsi ini.

6. Bapak dan Ibu Dosen Jurusan Ilmu Komputer yang telah memberikan bekal kepada penulis dalam penyusunan skripsi ini.
7. Kedua Orang Tua saya Bapak Fauzi Johanis dan Ibu Elimarni yang telah mencurahkan keringatnya untuk membiayai pendidikan saya, yang selalu memberikan kasih sayang, doa, dan dukungannnya.
8. Kakak saya Andri Nofa dan Febri Fauzi yang telah memberikan dukungan moril dan materiil serta doa yang terus dipanjatkan.
9. Keluarga besar kontrakan *Console House* (Akhsin, Alpin, Broto, Doni, Fachrizal, Farhan, Imron, Iqbal, Jefri, Khakim, Khamim, Salman, Warson) yang telah saling mendukung dalam menyelesaikan skripsi.
10. Teman-teman Boyos (Azmi, Bintang, Cita, Musyafa, Uwis) yang telah menghibur dan menyemangati hingga terselesaikannya penulisan skripsi ini.
11. Teman-teman saya di jurusan Ilmu Komputer, Fakultas MIPA, serta teman-teman di Universitas Negeri Semarang yang telah memberikan semangat dan dukungannya.
12. Semua pihak yang telah membantu terselesaikannya skripsi ini yang tidak dapat penulis sebutkan satu persatu, terimakasih atas bantuannya.

Semoga skripsi ini dapat memberikan manfaat bagi pembaca di masa yang akan datang.

Semarang, 11 Juni 2020

Penulis



Raka Hendra Saputra
4611415027

ABSTRAK

Raka Hendra Saputra. 2020. Optimasi Algoritma C4.5 Menggunakan Seleksi Fitur *Particle Swarm Optimization* (PSO) dan Teknik *Bagging* pada Diagnosis Penyakit Kanker Payudara. Skripsi, Jurusan Ilmu Komputer Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Negeri Semarang. Pembimbing Budi Prasetiyo, S.Si., M.Kom.

Kata kunci: *Data mining*, *Decision tree*, klasifikasi, C4.5, PSO, *Bagging*, Kanker Payudara.

Kanker payudara merupakan penyebab utama kedua kematian akibat kanker pada wanita saat ini dan telah menjadi kanker paling umum di antara wanita di negara maju dan negara berkembang dalam beberapa tahun terakhir. Dengan adanya pendeteksian dini penanganan menjadi lebih cepat sehingga risiko kematian akibat kanker payudara dapat dikurangi. Dalam usaha pendeteksian dini, *data mining* dapat digunakan untuk mendiagnosis kanker payudara. *Data mining* terdiri dari beberapa model penelitian, salah satunya adalah klasifikasi. Metode paling umum digunakan dalam klasifikasi merupakan *decision tree*. C4.5 adalah algoritma dalam *decision tree* yang sering digunakan dalam melakukan proses klasifikasi. Dalam penelitian ini, data yang digunakan adalah *Breast Cancer Wisconsin* (Original) *Data Set* yang diperoleh dari *UCI Machine Learning Repository*. *Breast Cancer Wisconsin* (Original) *Data Set* memiliki 9 fitur dan 1 kelas. *Dataset* ini mengalami ketidakseimbangan kelas dimana *benign* sebanyak 458 (65,5%) data dan *malignant* sebanyak 241 (34,5%) data. Tujuan dari penelitian ini adalah menyeleksi fitur yang akan digunakan dan mengatasi ketidakseimbangan kelas yang terjadi sehingga kinerja algoritma C4.5 menjadi lebih optimal dalam melakukan proses klasifikasi. Metode yang digunakan sebagai seleksi fitur adalah PSO dan *bagging* untuk mengatasi ketidakseimbangan kelas. Klasifikasi diuji menggunakan *confusion matrix* untuk mengetahui akurasi yang dihasilkan. Dari hasil penelitian ini, penerapan PSO sebagai seleksi fitur dan *bagging* untuk mengatasi ketidakseimbangan kelas dengan algoritma C4.5 berhasil meningkatkan akurasi sebesar 5,11%. Dengan akurasi awal 93,43%, setelah penerapan PSO dan *bagging* menjadi 98,54%. Penelitian ini dapat digunakan sebagai acuan peneliti selanjutnya yang berfokus pada perbaikan algoritma sehingga akurasi yang dihasilkan dapat lebih baik. PSO dalam penelitian ini kurang efisien dalam melakukan seleksi fitur sehingga penelitian selanjutnya juga diharapkan dapat mengatasi masalah ini dengan menggunakan metode lain dalam melakukan seleksi fitur.

DAFTAR ISI

Halaman

DAFTAR TABEL

Tabel Halaman

DAFTAR GAMBAR

Gambar Halaman

DAFTAR LAMPIRAN

Lampiran Halaman

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Kanker payudara merupakan penyebab utama kedua kematian akibat kanker pada wanita saat ini dan telah menjadi kanker paling umum di antara wanita di negara maju dan negara berkembang dalam beberapa tahun terakhir (Sumbaly, Vishnusri, & Jeyalatha, 2014: 16). Identifikasi kanker payudara dapat dilakukan secara manual tetapi proses ini sulit dilakukan karena harus mengingat semua informasi yang dibutuhkan untuk setiap keadaan tertentu yang menghasilkan akurasi rendah. Kematian akibat kanker payudara dapat dikurangi jika terdeteksi lebih awal. Ada metode konvensional untuk deteksi kanker payudara tetapi pengklasifikasi *machine learning* perlu dilakukan karena dapat mendapatkan akurasi yang lebih tinggi (Gupta & Kaushik, 2018: 56).

Data mining merupakan teknologi pengenalan pola serta teknik statistik dan matematika untuk menemukan korelasi, pola, dan tren baru yang berarti dengan memilah tumpukan penyimpanan data yang menyimpan data besar (Larose, 2004: 2). Dalam bidang medis, *data mining* dapat digunakan untuk mendiagnosis sebuah penyakit seperti kanker payudara, penyakit jantung, diabetes dan lain sebagainya (Muslim, Rukmana, Sugiharti, Prasetiyo, & Alimah, 2018: 1).

Klasifikasi dalam *data mining* merupakan dua bentuk proses analisis data yang digunakan untuk mengekstraksi model yang menggambarkan kelas data atau

untuk memprediksi tren data di masa depan. Dalam proses klasifikasi terdapat 2 fase; fase pertama adalah *training data* dimana dalam fase ini data dipelajari dan dianalisis dengan algoritma klasifikasi. Model atau pengklasifikasi yang dipelajari disajikan dalam bentuk pola atau aturan klasifikasi; fase kedua adalah penggunaan model untuk klasifikasi, dan *testing data* digunakan untuk memperkirakan akurasi yang dihasilkan berdasarkan aturan klasifikasi (D. Singh, Choudhary, & Samota, 2013: 1).

Masalah yang sering terjadi adalah klasifikasi memiliki sejumlah besar fitur dalam *dataset,* tetapi tidak semuanya akan digunakan. Fitur yang tidak relevan dan berlebihan dapat mengurangi kinerja (Xue, Zhang, & Browne, 2012: 1). Fitur yang tidak diperlukan membuat generalisasi menjadi lebih sulit dan menambah ukuran ruang pencarian yang menjadikan hambatan utama dalam *machine learning* dan *data mining.* Untuk memaksimalkan akurasi pada klasifikasi, dapat menggunakan seleksi fitur untuk melakukan pemilihan fitur yang akan dipakai (Gheyas & Smith, 2010: 5).

Seleksi fitur banyak digunakan untuk mengatasi fitur yang tidak relevan dan berlebihan. Seleksi fitur menyederhanakan sekumpulan data dengan mengurangi dimensi dan mengidentifikasi fitur yang relevan tanpa mengurangi akurasi prediksi (Aghdam & Heidari, 2015: 231). *Particle Swarm Optimization* (PSO) merupakan salah satu optimasi metaheuristik untuk seleksi fitur karena telah terbukti kompetitif dibandingkan dengan algoritma genetika dalam beberapa kasus, terutama dalam bidang optimasi (Muslim *et al*., 2018: 2). Optimasi metaheuristik telah terbukti sebagai metodelogi yang unggul untuk mendaparkan solusi yang baik dalam waktu

yang wajar (Yusta, 2009: 526). Selain terlalu banyaknya fitur yang ada, *dataset* pun sering terjadi ketidakseimbangan data.

Ketidakseimbangan data menjadi salah satu masalah klasik dalam klasifikasi di *machine learning*. Ketidakseimbangan data telah terbukti dapat menurunkan kinerja algoritma *machine learning* (Fanny & Cenggoro, 2018: 60). Ketidakseimbangan dapat diartikan misal salah satu *class* (*majority class*) jauh lebih banyak dari kelas lainnya (*minority class*) (Rout, Mishra, & Mallick, 2018: 431).

Breast Cancer Wisconsin (Original) *Data Set* memiliki 2 buah kelas, yaitu *benign* yang dituliskan 2 pada kelas sebanyak 458 (65,5%) dan *malignant* yang dituliskan 4 pada kelas sebanyak 241 (34,5%). Berdasarkan sampel data yang ada maka dapat dikatakan bahwa *Breast Cancer Wisconsin* (Original) *Data Set* mengalami ketidakseimbangan. Ada tiga pendekatan untuk mengatasi data yang tidak seimbang, yaitu: pendekatan *level data*, *level algoritmik*, dan menggabungkan atau metode *ensemble* (Yap, Rani, Rahman, Fong, Khairudin, & Abdullah, 2014: 14).

Dua metode yang populer digunakan dalam metode *ensemble* adalah *Bagging* dan *Boosting* (Opitz & Maclin, 1999: 169). Teknik *bagging* lebih unggul dibandingkan dengan *boosting* saat mengatasi data yang mengandung *noise* (Khoshgoftaar, Van Hulse, & Napolitano, 2011: 552). Selain itu, teknik *bagging* tidak hanya mudah dikembangkan, tetapi juga kuat ketika berhadapan dengan ketidakseimbangan kelas apabila diimplementasikan dengan benar (Feng, Huang,

& Ren, 2018: 6). Teknik *bagging* dapat diterapkan pada metode berbasis *tree* untuk meningkatkan nilai akurasi yang akan dihasilkan nantinya (Sutton, 2005: 303).

Decision tree banyak digunakan secara luas sebagai alat untuk menghasilkan aturan klasifikasi karena sifatnya sederhana namun sangat kuat (Bramer, 2016: 47). Banyak algoritma *decision tree* yang diusulkan dan digunakan secara luas dalam penelitian, salah satunya adalah C4.5 (Yang & Chen, 2016: 415). Algoritma C4.5 dapat memprediksi dengan hasil terbaik untuk akurasi dan membutuhkan waktu eksekusi paling minimum (Boukenze, Mousannif, & Haqiq, 2016: 8). Algoritma C4.5 mengenali atribut utama dari *set* pelatihan dan menjadikannya simpul akar dari sebuah pohon keputusan. Pada tahap berikutnya membuat simpul daun untuk semua keputusan yang mungkin. Algoritma C4.5 menentukan atribut terbaik dan tepat untuk partisi data menjadi beberapa kelas (Pradeep & Naveen, 2018: 414).

Berdasarkan uraian permasalahan diatas, maka penelitian ini berfokus untuk meningkatkan akurasi algoritma C4.5 menggunakan seleksi fitur *Particle Swarm Optimization* (PSO) dan teknik *bagging* untuk diagnosis kanker payudara dengan judul "Optimasi algoritma C4.5 menggunakan seleksi fitur *Particle Swarm Optimization* (PSO) dan teknik *bagging* pada diagnosis penyakit kanker payudara".

1.2 Rumusan Masalah

Berdasarkan latar belakang di atas, rumusan masalah dalam penelitian ini adalah:

1. Bagaimana PSO menyeleksi fitur penting untuk digunakan dalam proses klasifikasi pada *Breast Cancer Wisconsin* (Original) *Data Set*?
2. Bagaimana penerapan teknik *bagging* dalam melakukan penyelesaian masalah ketidakseimbangan kelas pada *Breast Cancer Wisconsin* (Original) *Data Set?*
3. Bagaimana akurasi yang dihasilkan oleh algoritma C4.5 setelah diterapkan PSO dan teknik *bagging* pada *Breast Cancer Wisconsin* (Original) *Data Set*?

1.3 Batasan Masalah

Pada penelitian ini diperlukan batasan-batasan agar tujuan penelitian dapat tercapai. Adapun batasan masalah yang dibahas pada penelitian ini adalah:

1. Algoritma yang digunakan untuk klasifikasi adalah algoritma C4.5.
2. PSO digunakan untuk mengatasi masalah fitur yang tidak relevan dan berlebihan.
3. Data yang tidak seimbang diatasi menggunakan teknik *bagging*.
4. *Dataset* yang digunakan adalah *Breast Cancer Wisconsin* (Original) *Data Set* yang diperoleh dari *UCI Machine Learning Repository*.
5. Hasil yang diukur dalam penelitian ini merupakan tingkat akurasi dari klasifikasi penyakit kanker payudara.

1.4 Tujuan Penelitian

Tujuan penelitian ini adalah sebagai berikut:

1. Menyeleksi fitur penting dengan PSO pada *Breast Cancer Wisconsin* (Original) *Data Set.*
2. Menerapkan teknik *bagging* dalam mengatasi data tidak seimbang yang terdapat pada *Breast Cancer Wisconsin* (Original) *Data Set.*
3. Mengukur tingkat akurasi yang dihasilkan oleh algoritma C4.5 dengan penerapan PSO dan teknik *bagging* pada *Breast Cancer Wisconsin* (Original) *Data Set.*

1.5 Manfaat Penelitian

Manfaat penelitian ini adalah sebagai berikut:

1. Memperoleh tingkat akurasi yang dihasilkan oleh algoritma C4.5 dengan penerapan PSO dan teknik *bagging* pada *Breast Cancer Wisconsin* (Original) *Data Set.*
2. Menambah wawasan mengenai algoritma C4.5 dengan menerapkan PSO dan teknik *bagging* dalam mendiagnosis penyakit kanker payudara.

1.6 Sistematika Penulisan Skripsi

1.6.1 Bagian Awal Skripsi

Bagian awal skripsi terdiri dari halaman judul, halaman pengesahan, halaman pernyataan, halaman motto dan persembahan, abstrak, kata pengantar, daftar isi, daftar gambar, daftar tabel dan daftar lampiran.

1.6.2 Bagian Isi Skripsi

Bagian isi skripsi terdiri dari lima bab, yaitu sebagai berikut.

1. BAB 1: PENDAHULUAN

   Bab ini berisi tentang latar belakang, rumusan masalah, batasan masalah, tujuan dan manfaat penelitian serta sistematika penulisan skripsi.

2. BAB 2: TINJAUAN PUSTAKA

   Bab ini berisi penjelasan mengenai definisi maupun pemikiran-pemikiran yang dijadikan kerangka teoritis yang menyangkut dan mendasari pemecahan masalah dalam skripsi ini.

3. BAB 3: METODE PENELITIAN

   Bab ini berisi penjelasan mengenai studi pendahuluan, tahap pengumpulan data, dan tahap pengembangan sistem.

4. BAB 4: HASIL DAN PEMBAHASAN

   Bab ini berisi hasil penelitian berserta pembahasannya.

5. BAB 5: PENUTUP

   Bab ini berisi simpulan dari penulisan skripsi dan saran yang diberikan penulis untuk mengembangkan skripsi ini.

1.6.3 Bagian Akhir Skripsi

Bagian akhir skripsi ini berisi daftar pustaka yang merupakan informasi mengenai buku-buku, sumber-sumber dan referensi yang digunakan penulis serta lampiran-lampiran yang mendukung dalam penulisan skripsi ini.

BAB 2

TINJAUAN PUSTAKA

2.1 *Data Mining*

Data mining merupakan sebuah proses berulang dimana kemajuan didefinisikan oleh penemuan, baik secara otomatis maupun manual. *Data mining* adalah proses pencarian informasi baru, berharga, dan penting dalam *volume* data yang besar (Kantardzic, 2011: 2).

Data mining adalah proses menemukan pola dan pengetahuan yang menarik dari sejumlah besar data. Sumber dapat mencakup *databases*, *data warehouses*, *web*, penyimpanan informasi lainnya, atau data yang dialirkan ke sistem secara dinamis (Han, Kamber, & Pei, 2012: 8).

Data mining terdiri dari empat kelas tugas, yaitu: (1) Klasifikasi, mengatur data kedalam kelompok yang telah ditentukan; (2) *Clustering*, seperti klasifikasi tetapi grup tidak ditentukan sebelumnya, sehingga algoritma akan mencoba mengelompokkan item yang serupa menjadi bersama; (3) Regresi, berusaha menemukan fungsi yang memodelkan data dengan kesalahan paling sedikit; dan (4) Asosiasi, yakni mencari hubungan antar variabel (Li, 2010: 2).

2.1.1 Tahapan *Data Mining*

Menurut Xu *et al.*, (2014: 1150), *data mining* sering disamaartikan dengan *Knowledge Discovery in Databases* (KDD) yang berfokus pada proses

penambangan data. Untuk mendapatkan pengetahuan yang berguna dari data, langkah-langkah berikut dilakukan secara berulang:

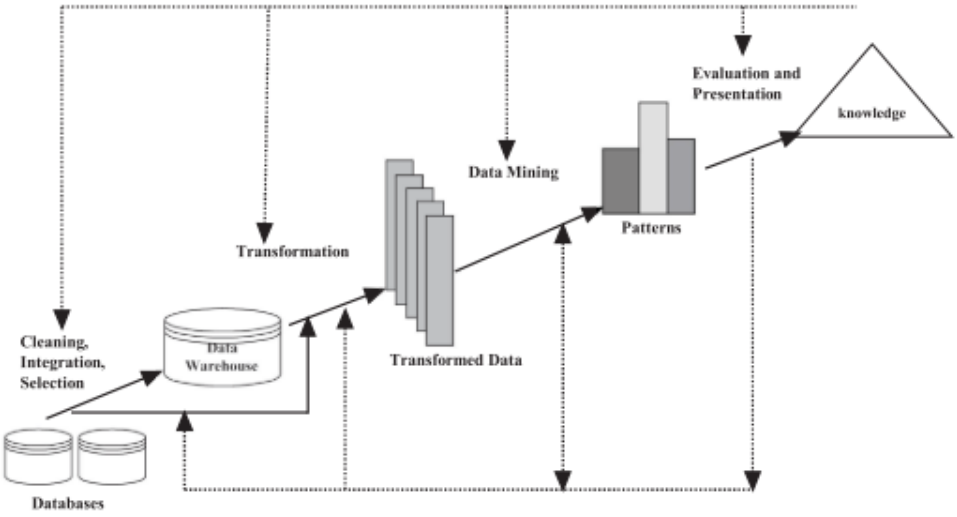
Langkah 1: *Data Prepocessing.* Operasi dasar meliputi pemilihan data (untuk mengambil data yang relevan dengan tugas KDD dari *database*), *data cleaning* (untuk menghilangkan *noise* dan data yang tidak konsisten, untuk menangani bidang data yang hilang, dan lain sebagainya), dan *data integration* (untuk menggabungkan data dari berbagai sumber).

Langkah 2: *Data Transformation.* Tujuannya adalah untuk mengubah data menjadi bentuk yang sesuai untuk penambangan, yaitu, untuk menemukan fitur yang berguna untuk sample data. Seleksi fitur dan transformasi fitur adalah operasi dasar.

Langkah 3: *Data Mining.* Merupakan proses penting yang digunakan untuk mengekstraksi pola (misalnya asosisasi, klasifikasi, pengelompokan, dan lain sebagainya.).

Langkah 4: *Pattern Evaluation and Presentation*. Operasi dasar termasuk mengidentifikasi pola yang benar-benar menarik yang mewakili pengetahuan, dan menyajikan pengetahuan yang ditambang dengan cara yang mudah dipaham.

Untuk lebih jelas gambaran mengenai proses dari *data mining* dapat dilihat pada Gambar 2.1.



Gambar 2.1 Tahapan *Data Mining* (Xu, Jiang, Wang, Yuan, & Ren, 2014: 1150)

2.2 Klasifikasi

Klasifikasi digunakan untuk mengklasifikasikan setiap *item* dalam satu *set* data ke dalam satu *set* kelas atau grup yang telah ditentukan. Tugas klasifikasi adalah menganalisis data dengan membangun model untuk memprediksi label kategori (atribut label kelas). Klasifikasi adalah fungsi *data mining* yang menetapkan *item* dalam *dataset* untuk menentukan kategori atau kelas. Tujuan klasifikasi adalah memprediksi secara akurat kelas target untuk setiap kelas dalam data (Kesavaraj & Sukumaran, 2013: 1).

2.3 Algoritma C4.5

Algoritma C4.5 merupakan algoritma yang banyak digunakan dalam klasifikasi untuk mengambil sebuah keputusan karena dapat menghasilkan pohon keputusan yang mudah untuk diartikan dan dimengerti, memiliki tingkat akurasi

yang dapat diterima, dan efisien untuk mengatasi atribut diskrit dan numerik (Hidayatulloh, Amegia & Susilawati, 2017: 37).

Secara umum, tahapan algoritma C4.5 dalam membuat pohon keputusan adalah sebagai berikut (Muzakir & Wulandari, 2016: 21).

1. Pilih atribut akar.
2. Buat cabang untuk tiap nilai.
3. Bagi kasus dalam cabang.
4. Ulangi proses untuk tiap proses cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Pemilihan atribut akar ditentukan berdasarkan nilai *gain* tertinggi dari seluruh atribut yang tersedia. Untuk menghitung *gain* dapat menggunakan Persamaan 1.

$$Gain\ (A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i) \quad \text{.....(1)}$$

Keterangan:

S = Himpunan Kasus

A = Atribut

n = Jumlah partisi atribut A

$|S_i|$ = Jumlah kasus pada partisi ke-i

$|S|$ = Jumlah kasus dalam S

Adapun untuk menghitung nilai *entropy* dapat dilihat pada Persamaan 2.

$$Entropy(S) = -\sum_{i=1}^{n} p_i * \log_2 p_i \quad \text{.....(2)}$$

Keterangan:

S = Himpunan Kasus

n = Jumlah partisi S

p_i = Proporsi dari Si terhadap S

Sementara itu nilai information *gain* dapat dihitung menggunakan Persamaan 3 (Novianti, Rismawan & Bahri, 2016).

$$Info\ Gain\ (S, A) = Entropy - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i) \quad \text{.....(3)}$$

Keterangan:

S = himpunan kasus

A = atribut

n = jumlah partisi atribut A

$|S_i|$ = jumlah kasus pada partisi ke-i

$|S|$ = jumlah kasus dalam S

Selanjutnya nilai *Split Info* dapat dihitung dengan Persamaan 4.

$$Split\ Info\ (S, A) = -\sum_{i=1}^{n} \frac{S_i}{S}\ log_2 \frac{S_i}{S} \quad \text{.....(4)}$$

Keterangan:

S = himpunan kasus

A = atribut

S_i = jumlah sampel untuk atribut i

2.4 *Particle Swarm Optimization* (PSO)

PSO pertama kali dikenalkan oleh Kennedy dan Eberhart (1995) yang merupakan teknik pencarian berbasis populasi dan dimotivasi oleh perilaku sosial organisme seperti sekelompok burung atau ikan. Fenomena yang mendasari PSO adalah bahwa pengetahuan dioptimalkan oleh interaksi sosial dimana pemikirannya tidak hanya bersifat individu tetapi juga sosial. Partikel pada PSO menyerupai kromosom pada GA. Namun, PSO biasanya lebih mudah diimplementasikan dari pada GA karena tidak ada *crossover* atau *operator* mutasi (Unler & Murat, 2010: 531).

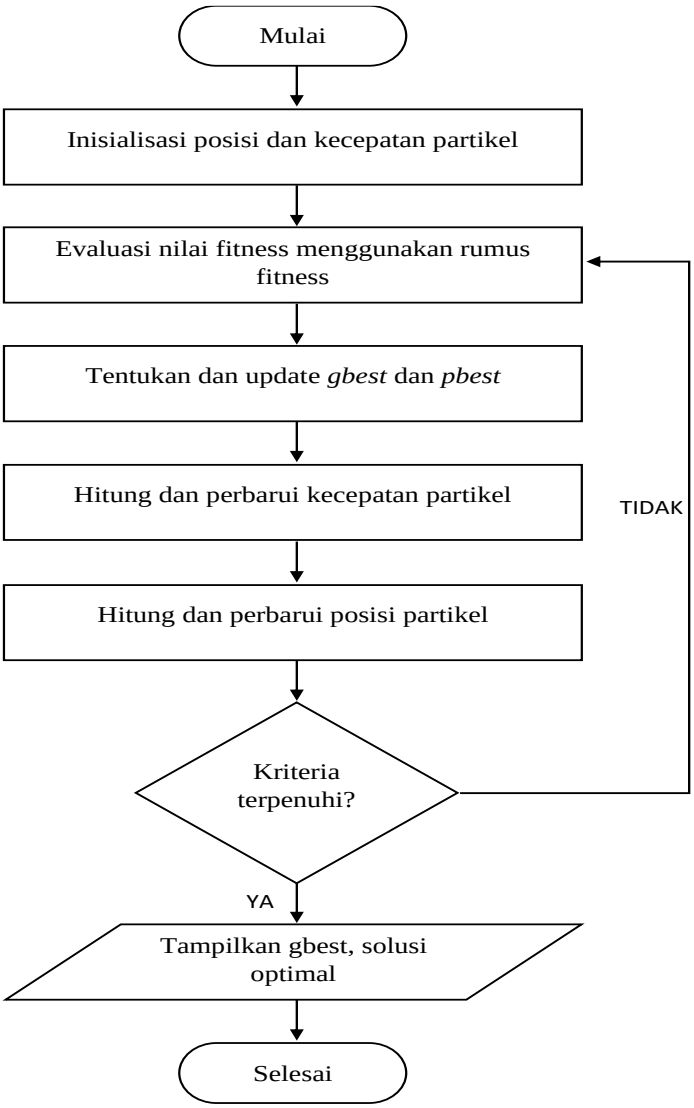
2.4.1 *Proses Particle Swarm Optimization* (PSO)

PSO dimulai dengan inisialisasi acak dari populasi partikel. Partikel bergerak ke ruang pencarian untuk mencari solusi optimal dengan memperbarui posisi masing-masing partikel berdasarkan keahliannya sendiri dan partikel di sekitarnya. Selama bergerak, posisi terkini dari partikel i diwakili oleh vektor $x_i = (x_{i1}, x_{i2}, \dots, x_{iD})$, dimana D adalah dimensi ruang pencarian. Kecepatan partikel i direpresentasikan sebagai $v_i = (v_{i1}, v_{i2}, \dots, v_{iD})$, yang dibatasi oleh kecepatan maksimum yang telah ditentukan sebelumnya, v_{max} dan $v_{id}^t \in [-v_{max}, v_{max}]$. Posisi partikel terbaik sebelumnya dicatat sebagai *pbest* atau *personal best* dan posisi terbaik yang diperoleh dari kelompok disebut sebagai *gbest* atau *global best*. PSO mencari solusi optimal dengan memperbarui posisi dan kecepatan setiap partikel dengan Persamaan 5 dan 6.

$$x_{id}^{t+1} = x_{id}^t + v_{id}^{t+1} \quad \text{.....(5)}$$

$$v_{id}^{t+1} = w \times v_{id}^{t} + c_1 \times r_{1i} \times (p_{id} - x_{id}^{t}) + c_2 \times r_{2i} \times \left(p_{gd} - x_{id}^{t}\right) \text{.....(6)}$$

Dimana *t* menunjukkan iterasi, *w* adalah berat inersia, c_1 dan c_2 adalah konstanta percepatan, r_{1i} dan r_{2i} adalah nilai acak antara [0-1], p_{id} dan p_{gd} mewakili elemen *pbest* dan *gbest*. Langkah PSO dapat dilihat seperti pada Gambar 2.2.



Gambar 2.2 *Flowchart* PSO (Armaghani, Hajihassani, Mohamad, Marto, & Noorani, 2014: 4)

2.5 *Binary Particle Swarm Optimization* (BPSO)

BPSO adalah salah satu metode turunan dari PSO. PSO merupakan algoritma yang banyak digunakan untuk mengatasi masalah optimisasi yang dirancang oleh Kennedy dan Eberhart (1995). Tidak seperti algoritma lainnya (seperti GA, DE, dll), PSO cenderung sederhana dan tidak memiliki operasi *crosssover* dan mutasi. *Basic* PSO diusulkan sebagai teknik optimisasi yang diterapkan pada ruang nyata (Bin, Qinke, Jing & Xiao, 2012: 224). Sementara *Binary Particle Swarm Optimization* (BPSO) yang diusulkan oleh Kennedy dan Eberhart memperluas PSO ke ruang pencarian *binary*. BPSO diterapkan untuk pemilihan fitur, setiap *bit* pada posisi vektor partikel menunjukkan fitur tersebut terpilih atau tidak untuk digunakan. Secara umum, *subset* yang dipilih oleh BPSO akan mempengaruhi akurasi yang dihasilkan dalam hal klasifikasi (Haixiang, Yijing, Yanan, Xiao, & Jinling, 2016: 180).

Menurut Kennedy dan Eberhart juga Mirjalili dan Lewis, PSO dapat dikonversi menjadi versi biner dengan menggunakan fungsi *transfer* yang memetakan nilai kecepatan ke bilangan *real* antara 0 dan 1. Oleh karena itu, *personal best (pbest)* dan *global best* (*gbest*) dibatasi menjadi "0" dan "1" (Mafarja & Sabar, 2018: 2). Dalam teknik BPSO, probabilitas partikel menjadi 0 atau 1 ditentukan oleh nilai kecepatan menggunakan fungsi sigmoid, dapat dilihat pada Persamaan 7.

$$x(t+1) = \begin{cases} 1, & jika\ rand\ < S\big(v(t+1)\big) \\ 0, & jika\ tidak \end{cases} \quad \text{.....(7)}$$

Dimana *rand*() adalah angka acak yang terdistribusi secara seragam antara 0 dan 1. Fungsi S() adalah fungsi sigmoid yang dihitung menggunakan Persamaan 8 (Babaoglu, Findik, & Ülker, 2010: 3179).

$$S\left(v_{ij}(t+1)\right) = \frac{1}{1+e^{-v_{jt}(t+1)}} \quad \text{.....(8)}$$

2.6 Teknik *Bagging*

Teknik *bagging* merupakan salah satu metode *ensemble* yang paling banyak digunakan dalam penelitian. Metode *ensemble* telah berhasil mengatasi ketidakseimbangan data walaupun tidak secara khusus dirancang untuk mengatasi hal tersebut (Laradji, Alshayeb, & Ghouti, 2015: 3). Metode *ensemble* yang paling banyak digunakan adalah *AdaBoost* dan *Bagging* yang dalam beberapa masalah klasifikasi mengarah ke peningkatan yang signifikan (Galar, Fernandez, Barrenechea, Bustince, & Herrera, 2012: 468). *Bagging* adalah metode *ensemble* yang efektif dan sederhana, karena mampu meningkatkan akurasi klasifikasi (Liang & Zhang, 2010: 31).

Bagging adalah metode yang menggabungkan *bootstrapping* dan *aggregating*. Sampel *bootstrap* diperoleh dengan pergantian jumlah elemen atau *resampling* jumlah elemen yang sama dengan dataset asli (Alfaro, Gáamez, & García, 2013: 4).

2.6.1 Proses Teknik *Bagging*

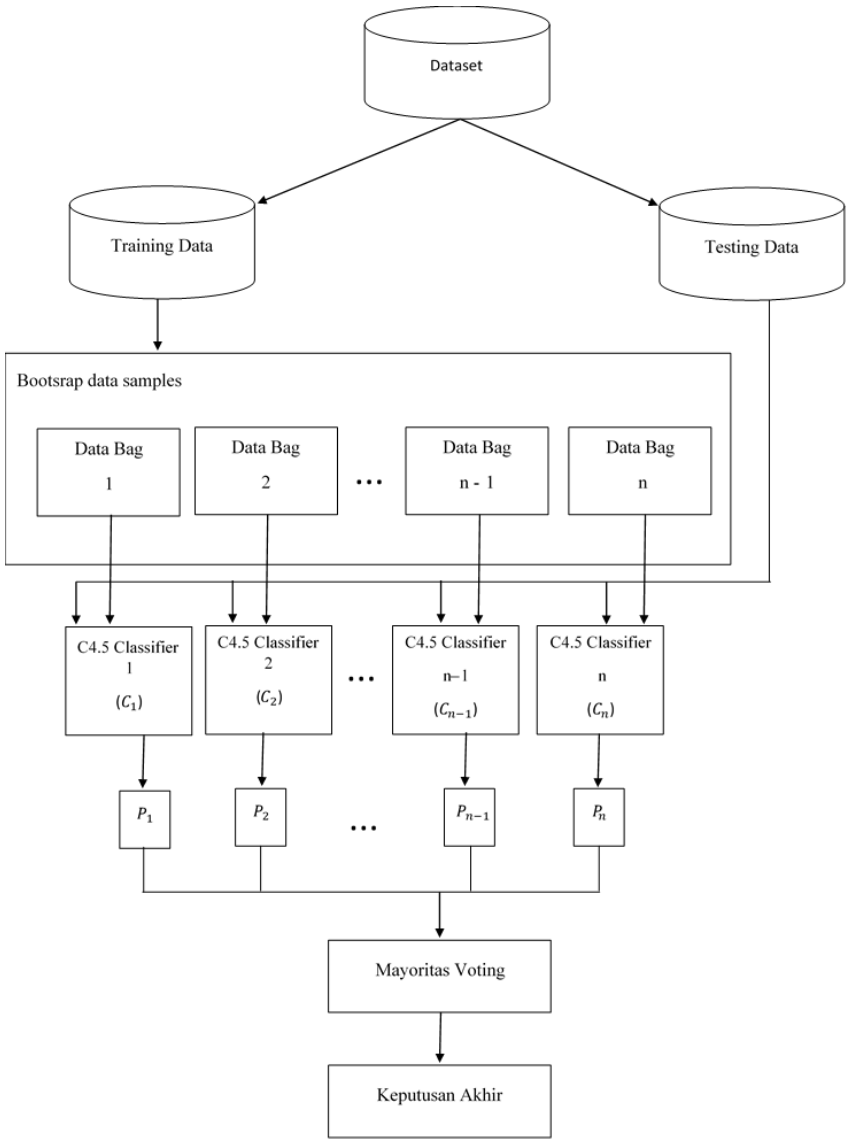
Bagging merupakan metode gabungan *bootstrap* dan *aggregating* seperti pada Gambar 2.3. Jika estimasi *bootstrap* dari parameter distribusi data lebih akurat

dan kuat daripada yang tradisional, maka metode yang sama dapat digunakan untuk mencapai sebuah klasifikasi dengan hasil yang lebih baik setelah menggabungkan keduanya.

1. *Repeat for* b = 1,2,...,B
   a. Ambil replikasi bootstrap T_b dari data training T_n
   b. Buatlah sebuah klasifikasi tunggal $C_b(x_i) = \{1,2,\dots,k\}$dalam T_b
2. Gabungkan klasifikasi *basic* $C_b(x_i)$, $b = 1,2,\dots,B$ berdasarkan *vote* terbanyak (kelas yang paling sering diprediksi) dengan aturan keputusan akhir $C_f(x_i) = argmax_{j \in Y} \sum_{b=1}^{B} I\,(C_b(x_i) = j)$

Gambar 2.3 Algoritma Teknik *Bagging*

Dalam *data training* (T_n), diperoleh sampel *bootsrap* B (T_b) dimana b = 1,2,...,B. Sampel *bootstrap* diperoleh berdasarkan pengambilan dengan pergantian jumlah elemen sama dengan *dataset* aslinya (dalam kasus ini adalah n). Dalam beberapa sampel *bootstrap*, adanya *noisy* akan dihilangkan atau setidaknya dikurangi. Sehingga klasifikasi yang dihasilkan akan lebih baik dibanding menggunakan data aslinya (Alfaro, Gáamez, & García, 2019: 35). Langkah proses *bagging* secara umum dapat dilihat pada Gambar 2.4.



Gambar 2.4 *Flowchart Bagging* (N. Singh & Singh, 2019: 2266)

2.7 *Breast Cancer Wisconsin (Original) Data Set*

Breast Cancer Wisconsin (Original) Data Set merupakan data yang didapat dari *UCI Machine Learning Repository* yang memiliki 699 kasus, 2 kelas (*malignant* dan *benign*), dan 9 atribut bernilai *integer*. Jumlah kelas *Benign*

sebanyak 458 (65.5%) dan *Malignant* sebanyak 241 (34.5%). Rincian atribut yang ada pada *dataset* ini dapat dilihat pada Tabel 2.1.

Tabel 2.1 *Breast Cancer Wisconsin (Original) Data Set* (Sumbaly *et al*., 2014: 16)

No	Atribut	*Domain*
1	*Sample code number*	*id number*
2	*Clump Thickness*	1-10
3	*Uniformity of Cell Size*	1-10
4	*Uniformity of Cell Shape*	1-10
5	*Marginal Adhesion*	1-10
6	*Single Epithelial Cell Size*	1-10
7	*Bare Nuclei*	1-10
8	*Bland Chromatin*	1-10
9	*Normal Nucleoli*	1-10
10	*Mitoses*	1-10
11	*Class*	2 = *benign*, 4 = *malignant*

2.8 *Confusion Matrix*

Confusion Matrix (Kohavi & Provost, 1998) berisi informasi tentang klasifikasi aktual dan prediksi yang dilakukan oleh sistem klasifikasi. Tabel 2.2 *confusion matrix* digunakan untuk dua kelas klasifikasi. Akurasi klasifikasi, sensitivitas, spesifisitas, nilai prediktif positif, dan nilai prediktif negatif dapat didefinisikan dengan menggunakan elemen-elemen dari *confusion matrix* (Akay, 2009: 3243).

Tabel 2.2 Representasi *Confusion Matrix*

Actual	***Predicted***	
	Positive	***Negative***
Positive	*True Positive (TP)*	*False Negative (FN)*
Negative	*False Positive (FP)*	*True Negative (TN)*

$$Accuracy = \frac{TP+TN}{P+N} \text{ x } 100\% \quad \text{.....(9)}$$

2.9 Penelitian Terkait

Penelitian ini dilakukan berdasarkan beberapa referensi yang memiliki keterkaitan dalam hal metode maupun objek yang digunakan. Referensi digunakan untuk memberi batasan mengenai metode dan sistem yang yang akan dikembangkan lebih lanjut ke depannya. Berikut ini merupakan beberapa penelitian yang pernah dilakukan sebelumnya seperti ditunjukkan pada Tabel 2.3.

Tabel 2.3 Penelitian Terkait dan *State of The Art*

Method	***Dataset***	***Feature Selector***	***Meta-learning***	***Classifier***	***Validation Methods***	***Evaluation Methods***
Akay	*Wisconsin Breast Cancer (Original)*	*F-Score*	-	SVM	*10-Fold X Validation*	*ROC & Confusion Matrix*
Lavanya & Rani	*Wisconsin Breast Cancer (Original)*	*Principal Components Attribute Eval*	*Bagging*	CART	*10-Fold X Validation*	-
Muslim MA *et al.*	*Wisconsin Breast Cancer (Original)*	PSO	-	C4.5	*10-Fold X Validation*	*Confusion Matrix*
Shrivas & Singh	*Wisconsin Breast Cancer (Original)*	*Info Gain*	-	6 *Classifier* (C4.5,CART,SVM,Bayes Net, MLP, RBF)	*10-Fold X Validation*	*Confusion Matrix*
Purposed method	*Wisconsin Breast Cancer (Original)*	PSO	*Bagging*	C4.5	-	*Confusion Matrix*

Akay (2009) dalam penelitiannya menunjukkan bahwa pembagian *data training* dan *testing* masing-masing 80% dan 20% merupakan yang paling optimal ketika digunakan untuk klasifikasi penyakit kanker payudara. Lavanya & Rani (2012) dalam penelitiannya menunjukkan penerapan bagging terhadap *decision tree* yang dalam hal ini adalah CART menghasilkan akurasi 97,85%. Muslim *et al.*, (2018) dalam penelitiannya berhasil meningkatkan akurasi sebesar 0,88% dengan memanfaatkan PSO sebagai seleksi fitur pada algoritma C4.5. Shrivas & Singh

(2016) dalam penelitiannya menunjukkan C4.5 menggunakan pembagian *data training* dan *testing* masing-masing 80% dan 20% untuk klasifikasi penyakit kanker payudara menghasilkan akurasi sebesar 92,857%.

Penelitian ini menggunakan objek yang sama seperti penelitian yang sudah pernah dilakukan sebelumnya, yakni *Wisconsin Breast Cancer* (Original) *Dataset*. Yang membedakan adalah penelitian ini menggunakan seleksi fitur *Particle Swarm Optimization* (PSO), *meta learning* menggunakan *bagging*, *classifier* menggunakan C4.5 dan metode dievaluasi menggunakan *confusion matrix*.

BAB 5

PENUTUP

5.1 Kesimpulan

Berdasarkan hasil penelitian dan pembahasan terkait algoritma C4.5 menggunakan seleksi fitur *Particle Swarm Optimization* (PSO) dan teknik *bagging* pada diagnosis penyakit kanker payudara, dapat ditarik kesimpulan sebagai berikut.

1. Penerapan PSO pada *Breast Cancer Wisconsin* (Original) *Data Set* digunakan sebagai seleksi fitur. PSO menyeleksi fitur yang akan digunakan dari sejumlah fitur yang ada pada *dataset*. Dalam hal ini fitur dapat juga disebut sebagai atribut. *Dataset* yang semula memiliki 9 atribut dan 1 kelas menjadi 8 atribut dan 1 kelas setelah diterapkan PSO.
2. Penerapan *bagging* pada *Breast Cancer Wisconsin* (Original) *Data Set* digunakan untuk mengatasi ketidakseimbangan kelas yang terjadi pada *dataset*. *Bagging* menghasilkan *bag* terbaik yang akan digunakan dalam proses klasifikasi algoritma C4.5 sehingga membuat kinerjanya menjadi lebih optimal.
3. Hasil akurasi yang didapatkan ketika diterapkan PSO dan *bagging* pada algoritma C4.5 adalah sebesar 98,54%. Sementara C4.5 tanpa PSO dan *bagging* menghasilkan akurasi sebesar 93,43%. Sehingga dapat diketahui adanya peningkatan sebesar 5,11% berdasarkan perbandingan akurasi yang dihasilkan. Hal ini menunjukkan PSO dan *bagging* berperan penting dalam

mengoptimalkan kinerja algoritma C4.5 sehingga dapat menghasilkan akurasi yang lebih baik.

5.2 Saran

Untuk peneliti yang ingin melakukan pengembangan lebih lanjut, saran yang diberikan adalah sebagai berikut.

1. Menggunakan algoritma lain dengan seleksi fitur PSO dan *bagging* untuk melakukan klasifikasi sehingga akurasi yang dihasilkan dapat lebih baik.
2. Untuk melakukan seleksi fitur dapat menggunakan metode lain yang lebih efisien dibandingkan dengan PSO.
3. Teknik untuk mengatasi ketidakseimbangan kelas dapat menggunakan metode lain yang dapat membuat kinerja algoritma yang digunakan menjadi lebih optimal
4. Melakukan pengembangan metode PSO, *bagging*, atau C4.5 untuk mendapatkan hasil akurasi yang lebih baik.

DAFTAR PUSTAKA

Aghdam, M.H & Heidari, S. 2015. Feature Selection Using Particle Swarm Optimization in Text Categorization. *JAISCR*, 5(4): 231-238.

Akay, M.F. 2009. Support Vector Machines Combined with Feature Selection for Breast Cancer Diagnosis. *Expert Systems with Applications*, 36(2): 3240-3247.

Alfaro, E., Gamez, M, & Garcia, N. 2013. Adabag: An R Package for Classification with Boosting and Bagging. *Journal of Statistical Software*, 54(2): 1-35.

Alfaro, E., Gamez, M, & Garcia, N. 2019. *Ensemble Classification Methods with Applications in R*. New Jersey: John Wiley & Sons Ltd.

Armaghani, D.J., Hajihassani, M., Mohamad, E.T., Marto, A, & Noorani, S.A. 2014. Blasting-Induced Flyrock and Ground Vibration Prediction through Expert Artificial Neural Network Based on Particle Swarm Optimization. *Arabian Journal of Geosciences*, 7(12): 5383-5396.

Babaoglu, I., Findik, O, & Ulker, E. 2010. A Comparison of Feature Selection Models Utilizing Binary Particle Swarm Optimization and Genetic Algorithm in Determining Coronary Artery Disease Using Support Vector Machine. *Expert Systems with Applications*, 37(4): 3177-3183.

Bin, W., Qinke, P., Jing, Z, & Xiao, C. 2012. A Binary Particle Swarm Optimization Algorithm Inspired by Multi-Level Organizational Learning Behavior. *European Journal of Operational Research*, 219(2): 224–233.

Boukenze, B., Mousannif, H, & Haqiq, A. 2016. Performance of Data Mining Techniques to Predict in Healthcare Case Study: Chronic Kidney Failure Disease. *International Journal of Database Management Systems (IJDMS)*, 8(3): 1-9.

Bramer, M. 2016. *Principles of Data Mining*. London: Springer.

Fanny & Cenggoro T.W. 2018. Deep Learning for Imbalance Data Classification Using Class Expert Generative Adversarial Network. *3rd International Conference on Computer Science and Computational Intelligence*. Jakarta: Bina Nusantara University.

Feng, W., Huang, W, & Ren, J. 2018. Class Imbalance Ensemble Learning Based on the Margin Theory. *Applied Sciences*, 8(5): 1-28.

Galar, M., Fernandez, A.,Barrenechea, Bustince,H & Herrera, F. 2012. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-based Approaches. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, 42(4): 463-484.

Gheyas, I.A & Smith, L.S. 2010. Feature Subset Selection in Large Dimensionality Domains. *Pattern Recognition*, 43(1): 5-13.

Gupta, A & Kaushik, B. 2018. Feature Selection from Biological Database for Breast Cancer Prediction and Detection Using Machine Learning Classifier Abhineet. *Journal of Artificial Intelligence*, 11(2): 55-64.

Haixiang, G., Yijing, L., Yanan, L., Xiao, L, & Jinling, L. 2016. BPSO-Adaboost-KNN Ensemble Learning Algorithm for Multiclass Imbalanced Data Classification. *Engineering Applications of Artificial Intelligence*, 49: 176-193.

Han, J., Kamber, M, & Pei, J. 2012. *Data Mining: Concept and Techniques, Third Edition.* Waltham: Morgan Kaufmann Publishers.

Hidayatulloh, T., Amegia, R, & Susilawati, D. 2017. Penerapan Algoritma C4.5 Pada Sistem Pakar Penyakit Aeromonas Hydrophila Ikan Mas Berbasis Mobile. *Jurnal Bianglala Informatika*, 5(1): 37-45.

Indraswari, R & Arifin, A.Z. 2017. RBF Kernel Optimization Method with Particle Swarm Optimization on SVM Using The Analysis of Input Data's Movement. *Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information),* 10(1): 36-42.

Kantardzic, M. 2011. *Data Mining: Concepts, Models, Methods, and Algorithms*. New Jersey: John Wiley & Sons, Inc.

Kesavaraj, G & Sukumaran, S. 2013. A Study On Classification Techniques in Data Mining. *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*. India: Tiruchengode.

Khoshgoftaar, T.M., Hulse, J.V, & Napolitano, A. 2011. Comparing Boosting and Bagging Techniques with Noisy and Imbalanced Data. *IEEE Transactions On Systems, Man, And Cybernetics—Part A: Systems And Humans*, 41(3): 552-568.

Laradji, I.H., Alshayeb, M, & Ghouti, L. 2015. Software Defect Prediction Using Ensemble Learning on Selected Features. *Information and Software Technology*, 58: 388-402.

Larose, D. 2004. *Discovering Knowledge in Data: An Introduction to Data Mining*. New Jersey: John Wiley & Sons, Inc.

Lavanya, D & Rani, K.U. 2012. Ensemble Decision Tree Classifier for Breast Cancer Data. *International Journal of Information Technology Convergence and Services (IJITCS)*. 2(1): 17-24.

Li, Y. 2010. Data Mining: Concepts, Background and Methods of Integrating Uncertainty in Data Mining. *CCSC:SC Student E-Journal*, 3: 2-7.

Liang, G & Zhang, C. 2010. Empirical Study of Bagging Predictors on Medical Data. *Conferences in Research and Practice in Information Technology (CRPIT)*, Australia: New South Wales.

Mafarja, M & Sabar, N.R. 2018. Rank Based Binary Particle Swarm Optimisation for Feature Selection in Classification. *International Conference on Future Networks and Distributed Systems (ICFNDS)*, Jordan: Amman.

Muslim, M.A., Rukmana, S.H., Sugiharti, E., Prasetiyo, B, & Alimah, S. 2018. Optimization of C4.5 Algorithm Based Particle Swarm Optimization for Breast Cancer Diagnosis. *Journal of Physics: Conference Series PAPER*, 983(1): 1-7.

Muzakir, A & Wulandari, R. 2016. Model Data Mining sebagai Prediksi Penyakit Hipertensi Kehamilan dengan Teknik Decision Tree. *Scientific Journal of Informatics*, 3(1): 19-26.

Novianti, B., Rismawan, T, & Bahri, S. 2016. Implementasi Data Mining dengan Algoritma C4.5 untuk Penjurusan Siswa (Studi Kasus: SMA Negeri 1 Pontianak). *Jurnal Coding, Sistem Komputer Untan*, 4(3): 75-84.

Opitz, D & Maclin, R. 1999. Popular Ensemble Methods: An Empirical Study. *Journal of Artificial Intelligence*, 11: 169-198.

Pradeep, K.R & Naveen, N.C. 2018. Lung Cancer Survivability Prediction Based on Performance Using Classification Techniques of Support Vector Machines, C4.5 and Naive Bayes Algorithms for Healthcare Analytics. *Procedia Computer Science*, 132: 412-420.

Rout, N., Mishra, D, & Mallick, M.K. 2018. Handling Imbalanced Data: A Survey. *International Proceedings on Advances in Soft Computing, Intelligent Systems and Applications.* Singapura.

Shrivas, A.K & Singh, A. 2016. Classification of Breast Cancer Diseases using Data Mining Techniques. *International Journal of Engineering Science Invention*, 5(12): 62-65.

Singh, D., Choudhary, N, & Samota, J. 2013. Analysis of Data Mining Classification with Decision Tree Technique. *Global Journal of Computer Science and Technology Software & Data Engineering*, 13(13): 1-6.

Singh, N & Singh, P. 2019. A Novel Bagged Naive Bayes-Decision Tree Approach for Multiclass Classification Problems. *Journal of Intelligent and Fuzzy Systems*, 36(3): 2261-2271.

Sumbaly, R., Vishnusri, N, & Jeyalatha, S. 2014. Diagnosis of Breast Cancer Using Decision Tree Data Mining Technique. *International Journal of Computer Applications*, 98(10): 16-24.

Sutton, C.D. 2005. Classification and Regression Trees, Bagging, and Boosting. *Handbook of Statistics*, 24(4): 303-329.

Unler, A & Murat, A. 2010. A Discrete Particle Swarm Optimization Method for Feature Selection in Binary Classification Problems. *European Journal of Operational Research*, 206(3): 528-539.

Xu, L., Jiang, C., Yuan, J, & Ren, Y. 2014. Information Security in Big Data: Privacy and Data Mining. *IEEE Access*, 2: 1149 - 1176.

Xue, B., Zhang, M, & Browne, W.N. 2012. Particle Swarm Optimization for Feature Selection in Classification: A Multi-Objective Approach. *IEEE Transactions on Cybernetics*,43(6): 1656 - 1671.

Yang, Y & Chen, W. 2016. Taiga: Performance Optimization of the C4.5 Decision Tree Construction Algorithm. *Tsinghua Science and Technology*, 21(4): 415-425.

Yap, B.W., Rani, K.A., Rahman,H.A.A., Fong, S., Khairudin, Z, & Abdullah, N.N. 2014. *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013)*. Singapura: Springer.

Yusta, S.C. 2009. Different Metaheuristic Strategies to Solve the Feature Selection Problem. *Pattern Recognition*, 30(5): 525-534.