OPTIMASI AKURASI KLASIFIKASI MENGGUNAKAN K-MEANS DAN ALGORITMA GENETIKA DENGAN MENGINTEGRASIKAN ALGORITMA C4.5 UNTUK DIAGNOSIS KANKER PAYUDARA

Skripsi

disusun sebagai salah satu syarat

untuk memperoleh gelar Sarjana Komputer

Program Studi Teknik Informatika

Oleh

Fachrizal Ahdy Andoyo
4611415040

JURUSAN ILMU KOMPUTER

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS NEGERI SEMARANG

2020

PERNYATAAN

Saya menyatakan bahwa skripsi saya yang berjudul "Optimasi Akurasi Klasifikasi Menggunakan K-Means dan Algoritma Genetika dengan Mengintegrasikan Algoritma C4.5 untuk Diagnosis 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. Dan 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 perundang-undangan.

Semarang, 23 Juni 2020

Fachrizal Ahdy Andoyo
4611415040

PERSETUJUAN PEMBIMBING

Nama	: Fachrizal Ahdy Andoyo
NIM	: 4611415040
Program Studi	: S-1 Teknik Informatika
Judul Skripsi	: Optimasi Akurasi Klasifikasi Menggunakan K-Means dan Algoritma Genetika dengan Mengintegrasikan Algoritma C4.5 untuk Diagnosis Kanker Payudara

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

Semarang, 23 Agustus 2020

Pembimbing

Riza Arifudin, S.Pd., M.Cs.
NIP 198005252005011001

PENGESAHAN

Skripsi yang berjudul

Optimasi Akurasi Klasifikasi Menggunakan K-Means dan Algoritma Genetika dengan Mengintegrasikan Algoritma C4.5 untuk Diagnosis Kanker Payudara

Disusun oleh

Fachrizal Ahdy Andoyo

4611415040

Telah dipertahankan di hadapan sidang panitia ujian skripsi FMIPA UNNES pada tanggal 24 September 2020.

Panitia:

Ketua

Dr. Sugianto, M.Si.
NIP 196102191993031001

Sekretaris

Dr. Alamsyah, S.Si., M.Kom.
NIP 197405172006041001

Ketua Penguji

Endang Sugiharti, S.Si., M.Kom.
NIP 197401071999032001

Anggota Penguji

Dr. Alamsyah, S.Si., M.Kom.
NIP 197405172006041001

Pembimbing

Riza Arifudin, S.Pd., M.Cs.
NIP 198005252005011001

MOTTO DAN PERSEMBAHAN

MOTTO

- Kalem, Tenang, Kuasai. (Fachrizal Ahdy Andoyo).

PERSEMBAHAN

Skripsi ini saya persembahkan kepada:

- Kedua Orang Tua saya Bapak Darsono dan Ibu Siti Khulaelah yang telah mencurahkan keringatnya untuk membiayai pendidikan saya, yang selalu memberikan kasih sayang, doa, dan dukungannnya.
- Kakak-kakak saya, Dedy Ardiansyah, Dony Ferry Anggoro, dan Adik saya, Agil Kusuma Wardhana yang telah memberikan dukungan serta doa yang terus dipanjatkan.
- Teman Hidupku, Trika Dina Mei Lia.
- 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.
- Almamater, 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 Akurasi Klasifikasi Menggunakan K-Means dan Algoritma Genetika dengan Mengintegrasikan Algoritma C4.5 untuk Diagnosis 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:

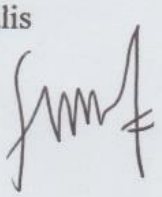
1. Prof. Dr. Fathur Rokhman, M.Hum., Rektor Universitas Negeri Semarang.
2. Dr. Sugianto, M.Si., Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Negeri Semarang.
3. Dr. Alamsyah S.Si., M.Kom., Ketua Jurusan Ilmu Komputer FMIPA Universitas Negeri Semarang yang telah meluangkan waktu, membantu, membimbing, mengarahkan dan memberikan saran sehingga penulis dapat menyelesaikan skripsi ini.
4. Riza Arifudin, S.Pd., M.Cs., Dosen Pembimbing yang telah meluangkan waktu, membantu, membimbing, mengarahkan dan memberikan saran sehingga penulis dapat menyelesaikan skripsi ini.
5. Bapak dan Ibu Dosen Jurusan Ilmu Komputer yang telah memberikan bekal kepada penulis dalam penyusunan skripsi ini.
6. Kedua Orang Tua penulis Bapak Darsono dan Ibu Siti Khulaelah yang telah mencurahkan keringatnya untuk membiayai pendidikan penulis, yang selalu memberikan kasih sayang, doa, dan dukungannnya.

7. Kakak-kakak penulis, Dedy Ardiansyah, Dony Ferry Anggoro dan Adik penulis, Agil Kusuma Wardhana yang telah memberikan dukungan serta doa yang terus dipanjatkan.
8. Teman-teman penulis di jurusan Ilmu Komputer, terutama teman-teman ilkom angkatan 2015, teman-teman kontrakan pondok pink, Alfo, Fajar, Fafa, Dayat, Hamim, Riski, dan Fendi serta teman-teman kontrakan Console House, Imron, Alpin, Akhsin, Raka, Farhan, Iqbal, Arief, Khamim, Khakim, Salman, Rosi, Berly, Alfin, Amanah, Triayana, Uwis dan Cita yang telah memberikan semangat dan dukungannya.
9. 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, 22 Agustus 2020

Penulis



Fachrizal Ahdy Andoyo
4611415040

ABSTRAK

Andoyo, Fachrizal Ahdy. 2020. Optimasi Akurasi Klasifikasi Menggunakan K-Means dan Algoritma Genetika dengan Mengintegrasikan Algoritma C4.5 untuk Diagnosis Kanker Payudara. Skripsi, Jurusan Ilmu Komputer Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Negeri Semarang. Pembimbing Riza Arifudin, S.Pd., M.Cs.

Kata kunci: *Data Mining,* K-Means, Algoritma Genetika, *Decision Tree,* Algoritma C4.5, *Wisconsin Diagnotic Breast Cancer.*

Perkembangan era teknologi jaman sekarang mengakibatkan data-data berkembang dengan cepat. *Data mining* adalah teknik yang digunakan untuk mengubah data menjadi informasi yang valid dan berguna. *Data mining* telah banyak digunakan dalam melakukan fungsi prediksi, contohnya dalam bidang ilmu kesehatan dan medis, data mining dapat memprediksi diagnosis penyakit berdasarkan data medis. Teknik *data mining* yang biasa digunakan mendiagnosis penyakit adalah teknik klasifikasi, dimana teknik ini dapat memprediksi sebuah keputusan dan menghasilkan sebuah akurasi yang tinggi. Pohon keputusan atau *decision tree* merupakan metode klasifikasi yang kuat dan terkenal dalam hal prediksi. Salah satu algoritma pohon keputusan adalah algoritma C4.5. Dalam penelitian ini, algoritma C4.5 dapat malakukan diagnosis terhadap kanker payudara karena memiliki struktur yang sederhana dan menghasilkan akurasi yang tinggi. *Wisconsin Diagnotic Breast Cancer* (WDBC) merupakan *dataset* publik yang diambil dari UCI *Machine Learning Repository*, dimana *dataset* ini memiliki 32 atribut dengan 569 sampel. *Dataset* ini memiliki tipe data kontinu dan berdimensi tinggi. Data yang kontinu serta memiliki dimensi tinggi membuat C4.5 membutuhkan waktu komputasi yang lama dan ruang penyimpanan yang besar sehingga mempengaruhi performa akurasi klasifikasi. Untuk menangani masalah tersebut, penerapan kombinasi K-Means dan Algoritma genetika dapat mempercepat kinerja klasifikasi dengan pembentukan *cluster-cluster* dan pemilihan fitur-fitur terbaik berdasarkan nilai *fitness* terbaik. Tujuan dari penelitian ini adalah meningkatkan akurasi algoritma C4.5 dengan kombinasi K-Means dan Algoritma Genetika sebagai fitur seleksi untuk mendiagnosis Kanker Payudara. Hasil penelitian ini merupakan perbandingan akurasi C4.5 sebelum dan sesudah diterapkan kombinasi K-Means dan Algoritma Genetika dalam mendiagnosis Kanker payudara. Akurasi C4.5 adalah 91,228%. Sedangkan, akurasi C4.5 setelah dioptimasi menggunakan K-Means dan Algoritma Genetika sebagai fitur seleksi adalah 94,824%. Dengan demikian, penerapan K-Means dan Algoritma Genetika pada algoritma C4.5 terbukti mampu meningkatkan hasil akurasi dalam mendiagnosis (WDBC) sebesar 3,596%.

DAFTAR ISI

Halaman

DAFTAR TABEL

Tabel Halaman

DAFTAR GAMBAR

Gambar Halaman

DAFTAR LAMPIRAN

Lampiran Halaman

BAB 1

PENDAHULUAN

1.1. Latar Belakang

Perkembangan teknologi informasi dan komunikasi pada era kini mengalami kemajuan yang sangat pesat (Sunge, 2018:27). Teknologi semakin mempermudah dan mempercepat manusia dalam menyelesaikan pekerjaannya. Kebutuhan informasi yang penting dan akurat kini menjadi kebutuhan sehari-hari, namun keakuratan informasi yang diterima belum sepenuhnya akurat, karena informasi yang akurat harus menggunakan data yang valid dan melalui kalkulasi yang tepat. Data yang diolah menjadi sub bidang ilmu pengetahuan kini sering disebut dengan istilah konsep *data mining* (Han *et al*., 2012: 8). Penerapan *data mining* dapat ditemukan dalam berbagai bidang, seperti perbankan, pemasaran, asuransi, perencanaan kota, transportasi, bioinformatika dan tak terkecuali bidang ilmu kedokteran dan medis (Sreedhar *et al*., 2017:1).

Di bidang ilmu kedokteran dan medis, pengolahan *data mining* dapat diterapkan untuk diagnosis penyakit, misalnya kanker payudara, diabetes dan penyakit jantung (Muslim *et al*., 2018: 1). Kanker payudara merupakan penyakit yang cepat berkembang di seluruh dunia. Penyakit ini mayoritas menyerang wanita dan bisa berujung pada kematian. Tidak dilakukannya diagnosis awal dan kurangnya penanganan menjadi penyebab kanker payudara semakin bertumbuh dan bisa berujung kematian. Diagnosis awal sangat diperlukan untuk landasan

penanganan terhadap penyakit kanker payudara (Zamani *et al*., 2012: 222). Dalam pengolahannya, *data mining* juga dibantu dengan Algoritma-algoritma tertentu untuk mempermudah proses diagnosis penyakit (Putranto, 2015: 1007). Algoritma atau teknik yang digunakan dalam pengolahan *data mining* adalah klasifikasi (Listiana & Muslim, 2017: 875).

Teknik klasifikasi yang banyak digunakan adalah *decision tree* yang merupakan algoritma klasifikasi yang memiliki struktur yang sederhana dan mudah ditafsirkan (Karegowda *et al*., 2012 : 46). Salah satunya yaitu algoritma C4.5 yang dapat memprediksi dengan hasil terbaik dalam hal akurasi dan waktu eksekusi minimum (Muslim *et al*., 2017: 1). Namun, akurasi prediksi *decision tree* juga dapat dipengaruhi oleh data kontinu (Rajeshinigo *et al*., 2017: 2755). Untuk mengatasinya diperlukan teknik *clustering* dalam mengatasi data yang kontinu (Effendy *et al*., 2017: 62)

Clustering merupakan proses pengelompokan data ke dalam kelas atau kelompok sehingga objek dalam suatu kelompok memiliki kesamaan yang tinggi dibandingkan yang lain, tetapi sangat berbeda dengan objek dalam kelompok lain (Karegowda *et al*., 2012: 45). Salah satu algoritma pada metode *clustering* yaitu Algoritma K-Means. K-Means digunakan untuk mengekstrak fitur data untuk menghindari pelatihan berulang pada *subset* yang berbeda, setelah ekstraksi pada fitur data lalu dilakukan teknik pengolahan data menggunakan metode klasifikasi (Zheng *et al*., 2013: 2). Pada klasifikasi yang melibatkan data berdimensi tinggi yang disebabkan oleh masalah *curse of dimensionality* yang berarti bahwa data dimensi tinggi berpengaruh terhadap komputasi waktu dan ruang dari tahap

pemrosesan data membuat akurasi klasifikasi akan sangat berpengaruh. Untuk menangani data dengan dimensi tinggi menggunakan metode pengurangan dimensi data yang biasa disebut fitur seleksi (Talita, 2016: 47).

Algoritma fitur seleksi merupaan bagian dari proses *preprocesing* dalam pengolahan data. Fitur seleksi juga berguna untuk mempermudah olah data yang berdimensi tinggi (Wahyuni, 2016: 284). Umumnya, fitur seleksi dikategorikan menjadi 3 kategori, yaitu *wrapped based, embedded method,* dan *filter based method.* Metode *wrapped* memerlukan waktu komputasi dan memori ruang yang besar serta algoritma tambahan untuk menghasilkan *subset* terbaik (Talita, 2016: 48). Metode *filter* membutuhkan waktu komputasi yang cepat untuk memilih fitur berasarkan karakteristik data, sehingga belum tentu menemukan fitur terbaik. Metode *embedded* merupakan gabungan dari *wrapped* dan *filter* untuk menemukan kombinasi *subset* fitur yang terbaik (Boomert, 2020: 2).

Salah satu algoritma dalam *wrapped* yang dapat mengoptimasi akurasi adalah Algoritma Genetika (Zamani, 2012: 224). Algoritma genetika adalah algoritma untuk mengatasi solusi terhadap permasalahan yang berdasarkan prinsip seleksi alam dalam ilmu genetika (Arifudin, 2012: 4). Saat pemilihan fitur, Algoritma Genetika digunakan sebagai algoritma pemilihan acak yang mampu menjelajahi ruang pencarian yang besar (Kumar *et al*., 2014: 273). Tujuan Algoritma Genetika yaitu untuk memilih nilai optimal untuk bobot dengan cara mempertahankan populasi yang memiliki nilai *fitness* baik agar menghasilkan keturunan dan membentuk populasi yang baru (Alalayah *et al*., 2018: 43).

Penelitian ini berfokus dalam mendiagnosis penyakit kanker payudara. *Dataset* yang digunakan dalam penelitian ini adalah *Wisconsin Diagnotic Breast Cancer* (WDBC) yang diambil dari UCI *Machine Learning Repository*. *Dataset* WDBC terdiri dari 569 data sampel, 32 atribut dengan kelas *malignant* (ganas) dan *benign* (jinak). *Dataset* WDBC tidak memiliki *missing value* sehingga data dapat diolah langsung menggunakan metode-metode *data mining.*

Berdasarkan uraian permasalahan di atas, maka penelitian ini berfokus untuk meningkatkan akurasi Algoritma C4.5 menggunakan K-Means yang dikombinasikan dengan seleksi fitur Algoritma Genetika untuk diagnosis Kanker Payudara dengan judul **"Optimasi Akurasi Klasifikasi Menggunakan K-Means dan Algoritma Genetika dengan Mengintegrasikan Algoritma C4.5 untuk Diagnosis Penyakit Kanker Payudara".**

1.2. Rumusan Masalah

Berdasarkan latar belakang diatas, rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana penerapan kombinasi K-Means dan Algoritma Genetika sebagai fitur seleksi pada Algoritma C4.5 dalam mendiagnosis kanker payudara?
2. Bagaimana hasil akurasi dari klasifikasi dari Algoritma C4.5 dengan menerapkan kombinasi K-Means dan Algoritma Genetika sebagai fitur seleksi dalam mendiagnosis kanker payudara?

1.3. Batasan Masalah

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

1. Algoritma klasifikasi yang digunakan dalam penelitian ini adalah Algoritma C4.5.
2. Fitur seleksi yang digunakan adalah Algoritma Genetika.
3. Data yang digunakan dalam penelitian ini adalah *Wisconsin Diagnotic Breast Cancer Dataset* (WDBC) yang diambil dari UCI *Machine Learning Repository*.

1.4. Tujuan Penelitian

Tujuan penelitian ini adalah sebagai berikut:

1. Mengetahui penerapan K-Means dan Algoritma Genetika sebagai fitur seleksi pada Algoritma C4.5 dalam mendiagnosis kanker payudara.
2. Untuk meningkatkan akurasi dari Algoritma C4.5 dengan menerapkan K-Means dan Algoritma Genetika sebagai fitur seleksi dalam mendiagnosis kanker payudara.

1.5. Manfaat Penelitian

Manfaat penelitian ini adalah sebagai berikut:

1. Mengetahui perbandingan akurasi dari Algoritma C4.5 sebelum dan sesudah diterapkan K-Means dan Algoritma Genetika sebagai fitur seleksi dalam mendiagnosis kanker payudara.
2. Mengetahui hasil peningkatan akurasi dari Algoritma C4.5 sebelum dan sesudah diterapkan K-Means dan Algoritma Genetika sebagai fitur seleksi.

1.6. Sistematika Penulisan

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 landasan teori maupun pemikiran-pemikiran yang dijadikan kerangka teoritis yang mendasari pemecahan penelitian.

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 penelitian dan saran yang diberikan penulis untuk mengembangkan penelitian agar lebih baik.

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. Tinjauan Pustaka

Penelitian ini dikembangkan berdasarkan berdasarkan beberapa referensi yang mempunyai keterkaitan metode dan objek penelitian. Penggunaan referensi ini ditujukan untuk memberikan batasan-batasan terhadap metode dan sistem yang nantinya akan dikembangkan lebih lanjut. Berikut adalah hasil dari penelitian sebelumnya.

Hermawanti (2012: 57-64) melakukan sebuah penelitian dengan judul "Penerapan Algoritma Klasifikasi C4.5 untuk Diagnosis Penyakit Kanker Payudara". Tujuan penelitian ini yaitu mengetahui tingkat akurasi yang dihasilkan dengan menggunakan algoritma C4.5 dalam mendiagnosis kanker payudara. Untuk membuat kategori klasifikasi, peneliti menggunakan kurva ROC dan melakukan pengujian menggunakan *confusion matrix*. Hasil pengujian didapatkan yaitu nilai AUC sebesar 0,941 dengan kategori klasifikasi sangat baik dan akurasi sebesar 94,56%.

Zamani *et al*., (2012: 1-6) dengan penelitian berjudul "Implementasi Algoritma Genetika pada Struktur *Backpropagation Neural Network* untuk Klasifikasi Kanker Payudara" menggunakan Algoritma Genetika untuk mengoptimasi *Neural Network* sebagai klasifikasinya. Parameter pertama yang dioptimasi yaitu jumlah unit *hidden layer* dan parameter keduanya yaitu *learning rate*. Hasil uji yang pertama menghasilkan 98,6% untuk kriteria generasi

maksimum. Dari pengujian 10 *fold* validasi ditemukan rata-rata akurasi sebesar 97%.

Elouedi *et al*., (2014: 266-231) melakukan penelitian dengan usulan teknik *clustering* dalam jurnalnya *International Conference of Soft Computing and Pattern Recognition, 6th International Conference of IEEE*, yang berjudul "*A Hybrid Approach Based on Decision Trees and Clustering for Breast Cancer Classification*". Penelitian ini juga menggunakan klasifikasi C4.5 untuk menilai jumlah *cluster* yang telah dibuat. Akurasi yang diperoleh sebelum dilakukan *cluster* sebesar 91,56% dan setelah dilakukan *cluster* sebesar 95,14%.

Rajeshinigo *et al*., (2017: 2755-2758), dalam penelitiannya yang berjudul "*Accuracy Improvement of C4.5 using K-Means Clustering*" pada jurnal *International Journal of Science and Research (IJSR)*. Penelitian ini mengusulkan algoritma K-Means untuk mengatasi data kontinu sehingga dapat meningkatkan akurasi algoritma C4.5. Hasil akurasi yang diperoleh pada algoritma C4.5 sebesar 73% dan setelah dioptimasi dengan K-Means sebesar 92%.

Muslim *et al*., (2018: 1-7) melakukan penelitian dengan judul "*Optimization of C4.5 algorithm-based Particle Swarm Optimization for breast cancer diagnosis*" dalam jurnal *International Conference on Mathematics, Science and Education (ICMSE)*. Penelitian ini menggunakan Metode Algoritma C4.5 dan *Particle Swarm Optimization* (PSO) yang bertujuan untuk meningkatkan akurasi dari diagnosis kanker payudara dan hasil akurasi yang diperoleh dengan proses klasifikasi Algoritma C4.5 yaitu 95,61%, sementara hasil akurasi C4.5 yang sudah dimodifikasi menggunakan *Particle Swarm Optimization* yaitu 96,49%.

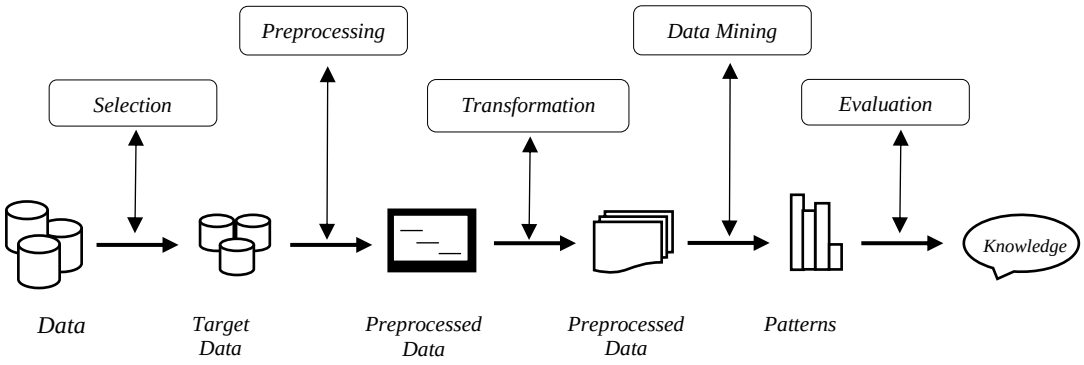
2.2. Landasan Teori

2.2.1 Data Mining

Suatu data dapat diolah menjadi sebuah informasi yang kemudian informasi-informasi tersebut menjadi sebuah ilmu pengetahuan atau disiplin ilmu. (Han *et al*., 2012:6) pengolahan sebuah data menggunakan metode-metode untuk menemukan pola atau bentuk sebuah data menjadi sebuah ilmu pengetahuan disebut dengan istilah *data mining*. *Mining* berarti tambang yang dikembangkan menjadi sebuah konsep dalam melihat pengetahuan yang tersimpan dalam *database* (Sunge, 2018: 27). Penelitian dengan *data mining* memungkinkan seseorang dapat mengambil sebuah keputusan dalam waktu yang akan datang (Putranto, 2015: 1007).

Data mining dapat diartikan masing-masing data sebagai informasi yang belum diolah, dan *mining* dilambangkan sebagai penambangan informasi (Han *et al*.,2012: 6). Banyak orang yang memberikan istilah lain untuk *data mining*, seperti penambangan pengetahuan dari data, analisis data, ekstraksi pengetahuan. Istilah yang paling popular yaitu *Knowledge Discovery from Data* (KDD) atau penemuan pengetahuan dari data. Menurut Han *et al*., (2012:7) langkah-langkah KDD dimulai dari *data cleaning, data integration, data selection, data transformasion, data mining, evaluation, representation knowledge*.

Proses *data mining* Han *et al.* (2012: 7) secara detail dapat ditunjukkan pada Gambar 2.1.



Gambar 2.1. Proses *Data Mining*

Menurut Han *et al.*, (2012: 17) Tahap *data mining* ada 7, yaitu:

1. Pembersihan data *(Data cleaning)*

   Merupakan proses awal sebelum mengolah data. Proses *cleaning* dapat dilakukan dengan cara membuang duplikasi data, memeriksa data yang tidak konsisten, dan memperbaiki kesalahan data seperti *missing value* atau nilai yang hilang.

2. Integrasi data (*data integration*)

   Integrasi data merupakan penggabungan data dari beberapa *database* ke dalam *database* yang baru. Integrasi dilakukan pada atribut yang mengdentifikasi entitas yang unik. Dalam penggunaannya, integrasi data harus dilakukan secara cermat karena kesalahan sekecil apapun dapat membuat hasil yang menyimpang.

3. Seleksi data (*data selection*)

   Data dalam *database* tentunya tidak dipakai semuanya. Oleh karena itu, dilakukan proses pengambilan data yang relevan dari *database* untuk dianalisis.

4. Transformasi data (*data transformation*)

   Transformasi data merupakan proses untuk pengubahan atau penggabungan data kedalam format yang sesuai untuk dapat diproses oleh *data mining.*

5. Proses *mining*

   Proses *mining m*erupakan proses utama dalam menerapkan sebuah metode untuk mendapatkan pola dan hasil yang sesuai.

6. Evaluasi pola (*evaluation pattern*)

   Evaluasi digunakan untuk mengidentifikasi pola yang unik sebelum menjadi *knowledge* atau pengetahuan yang ditemukan.

7. Presentasi pengetahuan (*Knowledge presentation*)

   Merupakan tahap terakhir dalam prose *data mining.* proses ini digunakan untuk menyajikan pengetahuan atau *knownledge* yang didapat kepada pengguna dalam bentuk visual.

2.2.2 Klasifikasi

Klasifikasi adalah suatu proses menggambarkan dan membedakan kelas data atau konsep yang digunakan untuk menemukan model (atau fungsi) dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang kelasnya tidak

diketahui. Terdapat beberapa algoritma klasifikasi data salah satunya *decision tree* antara lain Algoritma C4.5 dan ID3. Algoritma C4.5 merupakan pengembangan dari algoritma ID3. Algoritma yang merupakan pengembangan dari ID3 ini dapat mengklasifikasikan data dengan metode pohon keputusan (Han *et al*., 2012:328).

Tahapan dari klasifikasi dalam *data mining* terdiri dari :

1. Pembangunan model, dalam tahapan ini dibuat sebuah model untuk menyelesaikan masalah klasifikasi *class* atau *attribut* dalam data, model ini dibangun berdasarkan *training set* sebuah contoh data dari permasalahan yang dihadapi, *training set* ini sudah mempunyai informasi yang lengkap baik atribut maupun kelasnya.
2. Penerapan model, pada tahapan ini model yang sudah dibangun sebelumnya digunakan untuk menentukan *attribut / class* dari sebuah data baru yang atribut / kelasnya belum diketahui sebelumnya.
3. Evaluasi, pada tahapan ini hasil dari penerapan model pada tahapan sebelumnya dievaluasi menggunakan parameter terukur untuk menentukan apakah model tersebut dapat diterima

2.2.3 Algoritma C4.5

Algoritma C4.5 merupakan metode berbasis pohon keputusan atau *Decision Tree* yang dikembang melalui penyempurnaan algoritma IDE3 pada tahun 1986 oleh Quinlann Ross (Kathija, 2017: 15). Dalam algoritma C4.5 pemilihan atribut dilakukan dengan menggunakan *Gain, Ratio*, dengan mencari nilai *Entropy*. Secara

umum, langkah-langkah algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut :

1. Memilih atribut sebagai akar.
2. Membuat cabang untuk tiap-tiap nilai.
3. Membagi kasus yang ada dalam cabang.
4. Mengulang proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Menurut Sunge (2018: 28) dalam penggunaanya, algoritma C4.5 menggunakan konsep *gain ratio* untuk penyeleksian variabel. *Gain* (S,A) merupakan perolehan informasi dari atribut A relatif terhadap *output* data S. Perolehan informasi didapat dari *output data* atau variabel S yang dikelompokkan berdasarkan atribut A, dinotasikan dengan *gain* (S,A).

2.2.4 K-Means

K-means merupakan salah satu algoritma *clustering* . Tujuan algoritma ini yaitu untuk membagi data menjadi beberapa kelompok. Algoritma ini menerima masukan berupa data tanpa label kelas. Hal ini berbeda dengan *supervised learning* yang menerima masukan berupa vektor (x_1 , y_1) , (x_2 , y_2) , …, (x_i , y_i), dimana x_i merupakan data dari suatu data pelatihan dan y_i merupakan label kelas untuk x_i .

Pada algoritma pembelajaran ini, komputer mengelompokkan sendiri data-data yang menjadi masukannya tanpa mengetahui terlebih dulu target kelasnya. Pembelajaran ini termasuk dalam *unsupervised learning*. Masukan yang diterima

adalah data atau objek dan k buah kelompok *(cluster)* yang diinginkan. Algoritma ini akan mengelompokkan data atau objek ke dalam k buah kelompok tersebut. Pada setiap *cluster* terdapat titik pusat *(centroid)* yang merepresentasikan *cluster* tersebut. Efektivitas ini tergantung pada sifat datanya (Lavanya, 2013: 345).

Menurut Santhanam (2015:78) Algoritma untuk melakukan K-Means *clustering* adalah sebagai berikut:

1. Pilih K buah titik *centroid* secara acak.
2. Kelompokkan data sehingga terbentuk K buah *cluster* dengan titik centroid dari setiap *cluster* merupakan titik *centroid* yang telah dipilih sebelumnya.
3. Perbaharui nilai titik *centroid.*
4. Ulangi langkah 2 dan 3 sampai nilai dari titik *centroid* tidak lagi berubah.

2.2.5 Algoritma Genetika

Algoritma Genetika pertama kali dipublikasikan oleh John Holland pada tahun 1975 di Amerika Serikat. Saat itu, algoritma genetika hanya menitikberatkan pada proses *crossover* sehingga dijuluki algoritma genetika sederhana (Noordiansyah *et al*., 2016: 3758). Algoritma Genetika merupakan algoritma yang berbasis evolusi biologi. Berdasarkan teori evolusi, individu akan bersaing untunk bertahan hidup di alam yang memiliki sumber daya terbatas. Proses adaptasi setiap individu akan mempengaruhi kelangsungan hidup individu tersebut, apakah akan bertahan ataupun akan musnah (Noordiansyah *et al*., 2016: 3757).

Algoritma Genetika merupakan suatu metode optimasi untuk mencari solusi yang optimal dari suatu permasalahan. Menurut Arifudin (2012:4) Algoritma Genetika banyak digunakan untuk mencari solusi masalah optimasi penjadwalan dan pencarian. Algoritma genetika termasuk algoritma pengoptimalan pencarian yang menirukan proses evolusi alami (seleksi alam dan genetika alami). Beda halnya Algoritma pecarian lain yang melakukan pencarian dengan lingkup lokal, Algoritma Genetika justru mampu melakukan pencarian dalam lingkup yang besar secara efektif (Karegowda, 2011:16). Algoritma Genetika merupakan metode paling *representative* menyajikan teknik pengoptimalan yang cerdas dalam penelitian Ren (2010:1161).

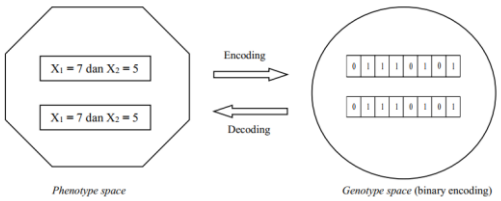
Menurut Arifudin (2012:4) Algoritma Genetika mempunyai komponen dasar sebagai berikut:

1. Representasi genetik dari solusi permasalahan
2. Populasi awal yang terbentuk dari solusi yang ada.
3. Fungsi evaluasi berdasarkan rating fitness.
4. Adanya operator-operator genetik
5. Nilai-nilai parameter Algoritma Genetika.

Algoritma genetika terdiri dari berbagai himpunan, seperti populasi atau solusi yang dihasilkan secara acak, kemudian kromosom atau individu didalam populasi. Secara umum Algoritma Genetika memiliki 3 operator genetik yaitu reproduksi, *crossover* atau kawin silang, dan mutasi (Arifudin, 2012:4).

2.2.5.1 Representasi Kromosom

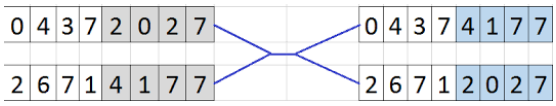
Representasi kromosom merupakan suatu proses pengkodean, pengkodean disini artinya menyelesaikan suatu permasalahan yang ada. Pengkodean kandidat penyelesaian ini disebut dengan kromosom. Kode tersebut meliputi penyandian gen, dimana sat gen mewakili sebuah variabel. Misalkan dalam suatu masalah, seperti proses *binary encoding*, kromosom dari *fenotip space* akan dikodekan menjadi kode biner atau direpresentasikan ke dalam kromosom biner yang ada di *genotip space.* Representasi kromosom digambarkan seperti Gambar 2.2.



Gambar 2.2 Representasi Kromosom

2.2.5.2 Crossover

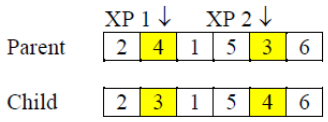
Proses Rekombinasi atau *crossover* digunakan Coley (1999:23) untuk menghasilkan himpunan solusi baru (*offspring*) dari proses rekombinasi dua individu (*parents*). Jumlah kromosom dalam sebuah populasi yang diberi rekombinasi ditentukan oleh paramater yang disebut dengan *crossover rate* (probabilitas persilangan). Terdapat 2 jenis Rekombinasi yaitu Rekombinasi 1 Titik dan Rekomendasi 2 Titik. Hal ini ditunjukkan pada Gambar 2.3.



Gambar 2.3 *Crossover*

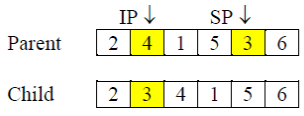
2.2.5.3 Mutasi

Operator mutasi digunakan untuk menghasilkan anak dengan melakukan perubahan pada satu individu. Dua metode mutasi ini dipilih secara acak pada setiap generasi untuk menghasilkan populasi yang lebih beragam. *Reciprocal exchange mutation* bekerja dengan memilih dua posisi (*exchange point* / XP) secara random kemudian menukarkan nilai pada posisi tersebut seperti ditunjukkan pada Gambar 2.4.



Gambar 2.4 *Reciprocal exchange mutation*

Insertion mutation bekerja dengan memilih satu posisi (*selected point* / SP) secara random kemudian mengambil dan menyisipkan nilainya pada posisi lain (*insertion point* / IP) secara random seperti ditunjukkan pada Gambar 2.5.



Gambar 2.5 *Insertion mutation*

2.2.5.4 Seleksi

Menurut Coley (1999: 23) sejauh ini seleksi dilakukan untuk memilih individu dari himpunan populasi dan *offspring* yang dipertahankan hidup pada generasi berikutnya. Semakin besar nilai *fitness* sebuah kromosom maka semakin besar peluangnya untuk terpilih. Hal ini dilakukan agar terbentuk generasi berikutnya yang lebih baik dari generasi sekarang. Terdapat berbagai metode seleksi yang sering digunakan seperti *Roulette wheel, binary tournamen, elitism,* dan *replacement.*

2.2.6 Evaluasi

Penelitian membutuhkan proses memvisualkan kinerja dari algoritma *machine learning*, maka digunakan *confusion matrix* atau matriks kebingungan. *Confusion matrix* menggunakan 4 evaluasi matriks yaitu nilai sensitivitas, spesifitas, prediksi positif, dan prediksi negatif (Santhanam *et al*., 2015: 80). Sensitivitas adalah kemampuan mengidentifikasi secara benar terhadap sebuah kriteria. Beda halnya dengan spesifitas yang merupakan kemampuan dalam mengidentifikasi ketidakbenaran terhadap suatu kriteria. Sedangkan nilai poitif dan negatif diartikan sebagai proporsi nilai positif dan negatif yang diprediksikan (Noordiansyah *et al*., 2016: 3758).

Dalam penelitian ini, *confusion matrix* atau matriks kebingungan bekerja sebagai alat ukur performa algoritma klasifikasi dengan membandingkan *dataset* dengan hasil klasifikasi yang sesuai data sebenarnya dengan jumlah keseluruhan

data. Hasil akhir dari tahap ini adalah sebuah akurasi dengan persentase (%) (Indrayanti *et al*., 2017: 4).

Evaluasi dari proses klasifikasi yang dilakukan *confusion matix* akan direpresentasikan menjadi tabel yang berisi label klasifikasi dengan label sebenarnya. Label akan dituliskan menjadi kelas ya dan tidak (Indrayanti *et al*., (2017: 4). Tabel *confusion matrix* atau matriks kebingungan secara detail dapat ditunjukan pada pada Tabel 2.1.

Tabel 2.1. *Confusion Matrix*

Klasifikasi		**Kelas Hasil Prediksi**		
		Ya	Tidak	Jumlah
Kelas aktual	Ya	TP	FN	P
	Tidak	FP	TN	N
	Jumlah	*P'*	*N'*	P+N

Perhitungan hasil akurasi menggunakan *confusion matrix* dengan hasil akhir berupa persentase. Perhitungan hasil akurasi dapat dituliskan dengan Persamaan 1.

$$Accuracy = \frac{TP+TN}{P+N} \text{ x } 100\% \quad (1)$$

2.2.7 Breast Cancer

Organisasi Kesehatan Dunia (WHO) menyatakan bahwa kanker payudara masuk dalam lima besar dunia kanker yang paling sering terjadi dan umumnya dialami oleh wanita. WHO mengestimasi bahwa 84 juta orang meninggal akibat kanker dalam rentang waktu 10 tahun. Survei yang dilakukan WHO menyatakan bahwa 8-9 persen wanita mengalami kanker payudara. Pada tahun 2015 sebesar

570.000 penderita kanker payudara meninggal dan seluruhnya adalah wanita (Calle, 2005:3).

Tingkat kelangsungan hidup kanker payudara sangat bervariasi di seluruh dunia, mulai dari 80% atau lebih terjadi di Amerika Utara, Swedia dan Jepang hingga 60% terjadi juga di negara-negara berpenghasilan menengah dan dibawah 40% terjadi di negara-negara berpenghasilan rendah. Kurangnya program deteksi dini mengakibatkan tingginya proporsi perempuan yang mengalami penyakit kanker payudara stadium lanjut, penyebab lainnya yaitu kurangnya diagnosis dan fasilitas pengobatan yang memadai, terutama di negara berkembang. Kanker payudara merupakan penyebab umum kematian akibat kanker di dunia (Hermawanti, 2012:57).

Penelitian ini menggunakan *dataset Wisconsin Diagnotic Breast Cancer* yang diambil dari UCI *machine learning* repository. *Dataset* ini terdiri dari 569 *instances* dan 32 atribut. Deskrpsi atribut dan *dataset* WDBC dapat ditunjukkan pada Tabel 2.2

Tabel 2.2 Deskripsi *dataset* WDBC

Attribute	***Descriptive Error***
Radius	*Mean of distances center to points the perimeter*
Texture	*Standard deviation*
Perimeter	*Perimeter of the cell*
Area	*Area of the cell*
Smoothness	*Local variation in radius*
Compactness	*Perimeter^2 / area*
Concavity	*Severity of concave portions of the contour*
Concave Points	*Number of concave portions of the contour*

Symmetry	*Symmetry of the cell nucleus*
Fractal Dimension	*Coastline approximation*

BAB 5

PENUTUP

5.1. Kesimpulan

Dari hasil penelitian dan pembahasan tentang optimasi akurasi klasifikasi Algoritma C4.5 dengan penerapan kombinasi algoritma K-Means dan Algoritma Genetika sebagai seleksi fitur menggunakan *dataset* WDBC yang diperoleh dari UCI *Machine Learning Repository* dapat ditarik kesimpulan sebagai berikut.

1. Penerapan kombinasi K-Means dan Algoritma Genetika dalam meningkatkan akurasi Algoritma C4.5 untuk mendiagnosis kanker payudara memiliki cara kerja masing-masing. K-Means akan membuat kelompok atau *cluster* pada atribut yang memiliki data kontinu. Dalam penelitian ini, jumlah k-*cluster* = 2 memberikan hasil akurasi yang lebih baik. Setelah proses pembuatan *cluster* oleh K-Means, dilakukan proses pemilihan fitur-fitur terbaik menggunakan Algoritma Genetika. Algoritma Genetika akan memilih fitur terbaik berdasarkan nilai *fitness* terbaik pada generasi maksimum. Hasil dari proses seleksi fitur akan digunakan untuk proses klasifikasi menggunakan Algoritma C4.5 dan menghasilkan sebuah model. Model tersebut akan dievaluasi untuk mengetahui akurasi dari metode yang digunakan.
2. Hasil akurasi yang didapatkan dari proses klasifikasi Algoritma C4.5 sebesar 92,228%. Lalu diterapkan K-Means dengan jumlah k=2 dan

Algoritma Genetika sebagai seleksi fitur didapatkan hasil akurasi sebesar 94,824%. Dengan demikian, dapat disimpulkan bahwa penerapan kombinasi K-Means dan Algoritma Genetika sebagai seleksi fitur pada Algoritma C4.5 dapat meningkatkan hasil akurasi dalam mendiagnosis kanker payudara sebesar 3,596%.

5.2. Saran

Untuk pengembangan lebih lanjut maka penulis memberikan saran sebagai berikut.

1. Perlu adanya uji coba lebih lanjut menggunakan *dataset* yang memiliki bentuk data berbeda.
2. Perlu penerapan algoritma klasifikasi lainnya menggunakan kombinasi K-Means dan Algoritma Genetika sebagai fitur seleksi agar dapat menghasilkan akurasi yang lebih baik.

DAFTAR PUSTAKA

Alalayah, K. M. A., Almasani, S. A. M., & Qaid, W. A. A. (2018). Breast Cancer Diagnosis based on Genetich Algorithm and Neural Network. *International Journal of Computer Applications.* 180(26). 42-44.

Arifudin, R. (2012). Optimasi Penjadwalan Proyek dengan Penyeimbangan Biaya Menggunakan Kombinasi CPM dan Algoritma Genetika. *Jurnal Masyarakat Informatika*. 2(4). 1-14.

Boomert, A., Sun, X., & Bischl, B. (2020). Benchmark for Filter Methods for Feature Selection in High-Dimensional Classification Data. *Computational Satistics and Data Analysis.* 143. 1-19.

Calle, J. (2005). *Breast Cancer Facts and Figures 2005-2006*. Atlanta: American Cancer Society.

Coley, D A. (1999). *An Introduction to Genetic Algorithms for Scientists and Engineers*. Singapore: World Scientific Publishing.

Dubey, A.K., Gupta, U, & Jain, S. (2016). Analysis of K-means Clustering Approach on the Breast Cancer Wisconsin Dataset. *International Journal of Computer Assisted Radiology and Surgery*. 11(11). 2033-2047.

Effendy, D. A., Kursini,. & Sudarmawan. (2017). Algoritma K-Means untuk Diskretisasi Numerik Kontinyu Pada Klasifikasi Intrusion Detetction System Menggunakan Naïve Bayes. *Konferensi Nasional Sistem & Informatika,* Bali: 10 Agustus 2017. 61-66.

Elouedi, H., Meliani, W., Elouedi, Z., & Amor, N. B. (2014). A Hybrid Approach Based on Decision Trees and Clustering for Breast Cancer Classification. *International Conference of Soft Computing and Pattern Recognition (SoCPaR),* 6th International Conference of IEEE. 226-231.

Han, J., & Kamber, M. (2006). *Data Mining Concepts and Techniques (2nd ed).* San Francisco: Morgan Kauffman Publishers.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques(3nd ed).* Waltham: Morgan Kaufmann.

Hermawanti, L. (2012). Penerapan Algoritma Klasifikasi C4.5 untuk Diagnosis Penyakit Kanker Payudara. *Jurnal Teknik Unisfat*. 7(2). 57-64.

Karegowda, A.G., Manjunath, A.S., & Jayaram, M.A. (2011). Application of Genetic Algorithm Optimized Neural Network Connections Weights For Medical Diagnosis of Pima Indians Diabetes. *International Journal on Soft Computing (IJSC).* 2 (2). 15-23.

Kathija, A., Nisha, S. S., & Sathik, M.M. (2017). Classification of Breast Cancer Data Using C4.5 Classifier Algorithm. *International Journal of Recent Engineering Research and Development (IJRERD).* 2(2). 13-19.

Kumar, G. R., Ramachandra, G. A., & Nagamani, K. (2014). An Efficient Feature Selection System to Integrating SVM with Genetic Algorithm for Large Medical Dataset. *International Journal of Advanced Research in Computer Science and Software Engineering.* 4(2). 272-277.

Lavanya, D., &Rani, K. U. (2013). A Hybrid Approach to Improve Classification with Cascading of Data Mining Tasks. *International Journal of Application or Innovation in Engineering & Management.* 2(1). 345-350.

Listiana, E., & Muslim, M. A. (2017). Penerapan Adaboost untuk Klasifikasi Support Vector Machine Guna Meningkatkan Akurasi pada Diagnosa Chronic Kidney Disease. *Prosiding SNATIF ke 4* (pp. 875-881). Kudus: Fakultas Teknik - Universitas Muria Kudus.

Olson, D.L., & Delen, D. (2008). *Advanced Data Mining Techniques*. Verlag: Springer.

Mounika, M., Suganya, S.D., Vijayashanthi, B., & Krishnanand, S. (2015). Predictive Analysis of Diabetic Treatment Using Classification Algorithm. *(IJCSIT) International Journal of Computer Science and Information Technologies.* 6 (3). 2502-2505.

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.* 983. 1-7.

Putranto, A.R., Wuryandari, T, & Sudarno. (2015). Perbandingan Analisis Klasifikasi antara Decision Tree dan Support Machine Vector Multiclass untuk Penentuan Jurusan pada Siswa SMA. *Jurnal Gaussian.* 4(4). 1007-1016.

Rajeshinigo, D., Jebamalar, J.P.A .(2017). Accuracy Improvement of C4.5 using K-means Clustering. *International Journal of Science and Research (IJSR).* 6(6). 2755-2758.

Ren, Y., Bai, G. (2010). Determination of Optimal SVM Parameters by Using Genetic Algorithm/Particle Swarm Optimization. *Journal of Computers*. 5. 1160-1169.

Santhanam, T., Padmavathi, M.S . (2015). Application of K-Means and Genetic Algorithms for Dimension Reduction by Integrating SVM for Diabetes Diagnosis. *Journal Computer Science*. 47. 76 – 83.

Sreedhar, C., Kasiviswanath, N., & Reddy, P. C. (2017). Clustering Large Datasets using K-Means Modifed Inter and Intra Clustering (KM-I2C) in Hadoop. *Journal of Big Data.* 4(27). 1-19.

Sunge, A. S. (2018). Optimasi Algoritma C4.5 dalam Prediksi Web Phising Menggunakan Seleksi Fitur Genetic Algorithm. *Paradigma.* 20(2), 27-32.

Talita, A. S. (2016). Klasifikasi Wisconsin Diagnotic Breast Cancer Data dengan Menggunakan Sequential Feature Selection dan Possibilistic C-Means. *Jurnal Ilmiah Komputasi*. 15(1). 47-52.

Wahyuni, E. S. (2016). Penerapan Metode Seleksi Fitur untuk Meningkatkan Hasil Diagnosis Kanker Payudara. *Jurnal Simetris.* 7(1). 283-294.

Zamani, A. M., Amaliah, B., & Munif, A. (2012). Implementasi Algoritma Genetika pada Struktur Backpropagation Neural Network untuk Klasifikasi Kanker Payudara. *Jurnal Teknik ITS* (1). 222-227..

Zheng, B., Yoon, S.W., & Lam, S.S. (2013). Breast Cancer Diagnosis Based on Feature Extraction Using A Hybrid of K-Means And Support Vec Machine Algorithms. *Expert Systems With Applications.* 41(4). 1476-14