SISTEM PAKAR KLASIFIKASI KANKER PAYUDARA MENGGUNAKAN ALGORITMA *K-NEAREST NEIGHBOR*

Randy Ginanjar[1], Erfian Junianto[2]

[1]Universitas Adhirajasa Reswara Sanjaya
Jl. Sekolah International No.1-2 Antapani, Bandung, 022-7100124
e-mail: erfian.ejn@ars.ac.id

[2]Universitas Adhirajasa Reswara Sanjaya
Jl. Sekolah International No.1-2 Antapani, Bandung, 022-7100124
e-mail: randyginanjarr@gmail.com

Abstract

Sistem pakar merupakan sistem yang berusaha mengadopsi kepakaran manusia sehingga komputer bisa melakukan hal-hal yang dapat dikerjakan oleh seorang pakar untuk memecahkan permasalahan yang bersifat spesifik. Pakar dalam hal ini adalah seorang yang ahli dibidangnya. Sistem pakar dapat digunakan untuk semua bidang ilmu termasuk dunia medis/kedokteran. Dataset yang digunakan dalam penelitian ini adalah *breast cancer coimbra* yang merupakan dataset sekunder yang didapat dari UCI *Machine Learning Repository*. Dataset ini terdiri dari 116 sampel data dengan 10 atribut yaitu *age, Body Mass Index* (BMI)*, glucose, insulin, Homeostatic model assessment* (HOMA), *leptin, adiponectin, resistin* dan *monocyte chemoattractant protein-1* (MCP-1) Dimana keseluruhan atribut tersebut bernilai numerik. Ada banyak teknik untuk meningkatkan keakuratan keputusan yang dapat digunakan salah satunya dengan menggunakan algoritma sehingga dapat meningkatkan keakuratan keputusan yang diambil. Salah satu algoritma yang paling baik dalam menangani dataset dengan nilai numerik adalah *k-nearest neighbor* dan *neural network*. Namun algoritma *k-nearest neighbor* merupakan algoritma yang paling mudah dipahami dan diimplementasikan serta paling sederhana diantara algoritma lain. Sementara *neural network* merupakan algoritma yang kompleks dan sulit dipahami, dalam beberapa kasus *k-nearest neighbor* dianggap sebanding dengan algoritma yang lebih kompleks seperti *neural network* dan *support vector machine*. Hasil dari penelitian ini adalah suatu rekayasa inferensi kepakaran dengan tujuan untuk memperoleh keputusan klinis penyakit kanker payudara berdasarkan tingkat stadiums ebagai upaya meningkatkan pelayanan praktek kedokteran pada pasien untuk penanganan medis sedini mungkin.

Keywords*:* sistem pakar, Coimbra, algoritma, klasifikasi

1. Pendahuluan

Isu kesehatan masih menjadi masalah utama yang dialami oleh masyarakat Indonesia. Menurut Direktorat Jenderal Pencegahan dan Pengendalian Penyakit Kementerian Kesehatan Anung Sugihantono, Angka kejadian penyakit kanker di Indonesia menempati urutan ke-8 di Asia Tenggara dengan perbandingan 136,2 kasus setiap 100.000 penduduk (Kementerian Kesehatan, 2019).

Angka kejadian tertinggi di Indonesia untuk laki-laki adalah kanker paru-paru sebanyak 19,4 kasus setiap 100.000 penduduk. Sementara kasus kanker tertinggi di Indonesia untuk perempuan adalah kanker payudara sebanyak 42,1 kasus setiap 100.000 pendudukan. (Kementerian Kesehatan, 2019)

Data Global Cancer Observatory (Globocan) menyebutkan, pada tahun 2018 menempatkan kanker payudara sebagai kasus kanker paling banyak dengan angka 58.256 kasus (16,7%) dari total 348.809 kasus kanker (Global Cancer Observatory, 2018). Mengingat jumlah kasus kanker payudara yang banyak muncul di Indonesia, perlu sebuah upaya untuk mencegah atau mengatasi kanker payudara.

Screening kanker payudara merupakan strategi paling penting untuk deteksi awal dan untuk memastikan kemungkinan yang lebih

besar untuk mendapat hasil yang baik dalam pengobatan (Crisostomo, Matafome, & Santos-Silva, 2016). Jika kebanyakan upaya pencegahan dilakukan melalui diagnosa luar, maka pada *screening* ini dilakukan analisa terhadap sampel darah, beberapa elemen yang dianalisa pada setiap sampel darah adalah glukosa, insulin, usia dan indeks masa tubuh (BMI) (Patricio, et al., 2018).

Untuk memudahkan proses analisa terhadap sampel darah tersebut, perlu dikembangkan sebuah sistem pakar (*expert system*) untuk membantu mengklasifikasi sampel darah tersebut apakah pasien perlu segera mendapat perawatan atau hanya perlu kontrol kesehatan secara rutin saja. Sistem pakar dapat merekam bahkan menerapkan kecerdasan buatan dalam melakukan diagnosis (Rahayu & Sandi, 2018).

Kecerdasan buatan akan menganalisa pola dan pengetahuan yang didapat dari sejumlah sampel data. Dalam hal ini diperlukan sebuah algoritma untuk melakukan komputasi. Beberapa algoritma telah diimplementasikan untuk upaya *screening* kanker payudara ini, diantaranya *neural network* (Nugraha, Shidiq, & Rahayu, 2019), *support vector machine* (Purwaningsih, 2018), *decision tree* (Islam & Poly, 2019). Berdasarkan beberapa algoritma yang telah digunakan sebelumnya, hasil yang didapat belum optimal.

Dataset yang digunakan dalam penelitian ini adalah *breast cancer coimbra* yang merupakan dataset sekunder yang didapat dari UCI *Machine Learning Repository*. Dataset ini terdiri dari 116 sampel data dengan 10 atribut yaitu *age, Body Mass Index* (BMI)*, glucose, insulin, Homeostatic model assessment* (HOMA), *leptin, adiponectin, resistin* dan *monocyte chemoattractant protein-1* (MCP-1) (Patricio, et al., 2018). Dimana keseluruhan atribut tersebut bernilai numerik.

Salah satu algoritma yang paling baik dalam menangani dataset dengan nilai numerik adalah *k-nearest neighbor* dan *neural network* (Larose, 2007). Namun algoritma *k-nearest neighbor* merupakan algoritma yang paling mudah dipahami dan diimplementasikan (Prasetio & Riana, 2015) serta paling sederhana diantara algoritma lain (Gorunescu, 2011). Sementara *neural network* merupakan algoritma yang kompleks dan sulit dipahami, dalam beberapa kasus *k-nearest neighbor* dianggap sebanding dengan algoritma yang lebih kompleks seperti *neural network* dan *support vector machine* (Wu, Kumar, Ross, Ghosh, & Yang, 2008).

2. Metode Penelitian

Desain Penelitian

Penelitian didefinisikan oleh *Higher Education Funding Council for England* (HECFE) sebagai penyelidikan yang dilakukan untuk mendapatkan pengetahuan dan pemahaman (Dawson, 2009), serta merujuk pada aktifitas penyelidikan sistematis atau investigasi di suatu bidang, dengan tujuan menemukan atau merevisi fakta, teori, aplikasi dan sebagainya (Berndtsson, Hansson, Olsson, & Lundell, 2008).

Menurut (Dawson, 2009) ada empat metode penelitian yang paling umum digunakan yaitu: *action research*, *experiment*, *case study*, dan *survey*. Pada penelitian ini menggunakan metode *experiment*, yaitu penelitian yang melibatkan penyelidikan kepada beberapa variable menggunakan tes tertentu yang dikendalikan sendiri oleh peneliti. Pada penelitian ini dilakukan beberapa tahapan penelitian sebagai berikut:

1. Pengumpulan Data (*Data Gathering*)
   Pada tahapan ini dijelaskan tentang bagaimana dan darimana data dalam penelitian ini didapatkan. Pada tahap ini juga ditentukan data yang akan diproses.
2. Pengolahan Data Awal (*Data Pre-processing*)
   Pengolahan data awal meliputi pembersihan data, pentransformasian data ke dalam bentuk yang dibutuhkan serta pengelompokan dan penentuan atribut data.
3. Metode yang Diusulkan (*Proposed Method*)
   Setelah pengolahan data awal, lalu dibuatkan model yang sesuai dengan jenis data. Pembagian data ke dalam data pelatihan (training dataset) dan data pengujian (testing dataset) juga diperlukan untuk pembuatan model.
4. Eksperimen dan Pengujian Model (*Model Test and Experiment*)
   Pada tahapan ini, dilakukan eksperimen dan pengujian model terhadap data yang sebelumnya sudah diolah. Perhitungan dengan masing-masing algoritma akan diulang beberapa kali untuk mendapatkan besaran parameter terbaik.

5. Evaluasi dan Validasi Hasil (*Result Evaluation and Validation*)
   Tahap evaluasi merupakan tahap yang terakhir dari kegiatan penelitian, dimana dalam tahap ini hasil dari tahapan eksperimen dan pengujian model sebagai evaluasi.

Pengumpulan Data

Data yang digunakan pada penelitian ini merupakan data sekunder. Data sekunder adalah data yang tidak diperoleh langsung dari obyek penelitian, melainkan telah dikumpulkan oleh pihak lain. Data sekunder yang digunakan pada penelitian ini merupakan kumpulan data dari *Faculty of Medicine of the University of Coimbra* dan juga *University Hospital Center of Coimbra*. Data tersebut merupakan data kuantitatif yang mengindikasikan adanya kanker payudara. Seluruh atribut yang terdapat pada data ini dihasilkan dari analisis darah. Dataset ini diambil dari UCI Machine Learning Repository yang diunduh melalui https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra.

Dataset ini berisi informasi hasil analisa terhadap sampel darah pasien yang diwakili dalam data yang ditetapkan oleh 10 atribut. 10 atribut tersebut berupa numerik. Dataset ini memiliki dua kelas yaitu, perlu adanya kontrol kesehatan rutin (*healthy control*) dan positif menjadi pasien kanker payudara (*patients*). Jumlah sampel pada dataset ini sebanyak 116 sampel data dengan distribusi untuk *healthy control* sebanyak 52 orang dan *patients* sebanyak 64 orang. Penjelasan atribut yang ada pada dataset dapat diihat pada tabel 2.1. Data statistik pada dataset tersebut dapat dilihat pada tabel 2.1.

Tabel 2.1. Atribut Dataset

Atribut	Deskripsi	Tipe Data
AGE	Usia pasien terindikasi kanker payudara.	Numerik
BMI	*Body Mass Index*. Index masa tubuh pasien.	Numerik
GLUCOSE	Kadar glukosa pada darah pasien.	Numerik
INSULIN	Kadar insulum pada darah pasien.	Numerik
HOMA	Hasil pengujian *homeostasis* pada darah pasien (*Homeostasis Model Assesment).*	Numerik
LEPTIN	Kadar hormon Leptin (hormon yang berperan untuk pengaturan masa tubuh, metabolisme dan reproduksi) pada darah.	Numerik
ADIPONECTIN	Kadar hormon Adiponectin (hormon yang mengatur kadar glukosa pada darah) pada darah.	Numerik
RESISTIN	Kadar Resistin (protein yang kaya akan asam amino yang dapat menyebabkan obesitas jika terlalu banyak) pada darah.	Numerik
MCP.1	Kadar *Monocyte Chemoattractant Protein-1* (jenis protein) pada darah.	Numerik
CLASSIFICATION	Klasifikasi apakah pasien tergolong yang perlu kontrol kesehatan atau yang terindikasi pasien positif kanker payudara.	Biner

Pengolahan data Awal

Pengolahan data awal merupakan tindak lanjut dari pengumpulan data. Pada penelitian ini, pengolahan data awal dengan membagi dataset menjadi dua yaitu, *data training* dan *data testing.* Untuk menguji model yang dikembangkan, data akan dibagi menjadi dua bagian, yaitu *data training* dan *data testing*. *Data training* digunakan untuk pengembangan model, sedangkan *data testing* digunakan untuk pengujian model. Pembagian dataset ini menggunakan *split validation* dengan rasio bervariasi mulai 90:10, 80:20, 70:30, 60:40 dan 50:50. Ini dilakukan dengan tujuan agar dapat mengetahui rata-rata performa yang dihasilkan dari algoritma *k-nearest neighbor*.

Evaluasi dan Validasi Hasil

Pada tahap ini akan dilakukan proses pengujian model yang dihasilkan oleh tool Rapidminer dengan mengevaluasi perbandingan hasil akurasi seluruh eksperimen yaitu eksperimen *k-nearest neighbor* pada dataset dengan lima rasio

yang berbeda menggunakan *split validation*. Sementara itu, evaluasi yang digunakan adalah *confussion matrix*.

3. Hasil dan Pembahasan

3.1. Analisis Kebutuhan

Perangkat lunak yang akan dibangun memerlukan berbagai macam kebutuhan yang akan menunjang pembuatan dan pengembangan perangkat lunak. Oleh karena itu diperlukan tahap analisa kebutuhan perangkat lunak yang merupakan proses menganalisis dan mengumpulkan kebutuhan-kebutuhan sistem yang sesuai dengan informasi, sistem kerja, dan tampilan antar muka yang diinginkan, guna menentukan solusi perangkat lunak yang akan dibangun. Berikut ini merupakan tahapan analisa kebutuhan perangkat lunak pada sistem usulan ini :

Kebutuhan Pengguna

Berikut ini merupakan kebutuhan perangkat lunak bagi pengguna yang merupakan *end-user* dari sistem pakar ini:

1. Pengguna dapat masuk ke dalam sistem pakar
2. Pengguna dapat melihat data training
3. Pengguna dapat melihat informasi data traning
4. Pengguna dapat melakukan klasifikasi penyakit
5. Pengguna dapat melihat riwayat klasifikasi penyakit
6. Pengguna dapat registrasi akun

Kebutuhan Sistem

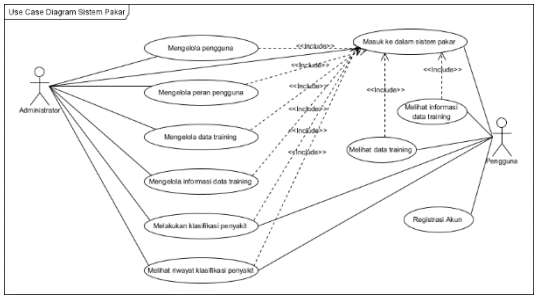
Berikut ini merupakan kebutuhan perangkat lunak bagi pengelola yang merupakan administrator dari sistem pakar ini:

1. Administrator dapat masuk ke dalam sistem pakar
2. Administrator dapat mengelola pengguna
3. Administrator dapat mengelola peran (*role*) pengguna
4. Administrator dapat mengelola data training
5. Administrator dapat mengelola informasi data training
6. Administrator dapat melakukan klasifikasi penyakit
7. Administrator dapat melihat riwayat klasifikasi penyakit

3.2. Perancangan

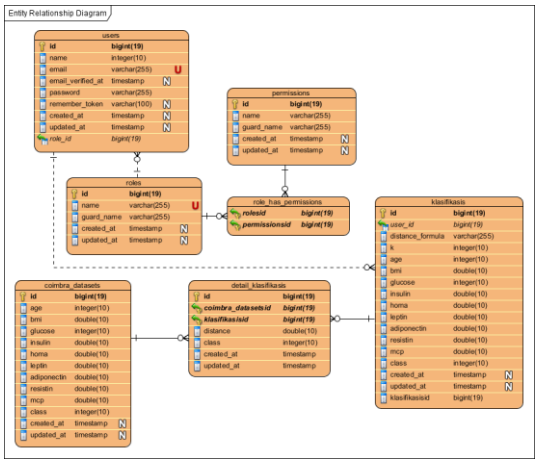
Usecase Diagram

Use Case Diagram menggambarkan fungsionalitas yang diharapkan dari sistem. Berikut ini merupakan *Use Case Diagram* dan deskripsi *Use Case* yang = menggambarkan sistem pakar ini:



Rancangan Database

Desain *database* menggambarkan tabel-tabel serta relasi antar tabel. Penulis menggambarkan tabel beserta relasinya menggunakan *Entity Relationship Diagram* atau biasa disebut dengan ERD.



3.3. Desain Interface

Desain *User Interface* memperlihatkan bagaimanakah bentuk dari perangkat lunak yang akan dibangun nantinya berdasarkan struktur aplikasi yang sudah dibuat. Perancangan antar muka ini meliputi perancangan tampilan input dan perancangan tampilan output.

Halaman awal merupakan halaman yang pertama kali muncul pada saat sistem pakar diakses. Halaman ini hanya menampilkan *cover* website saja dengan navigasi hanya pada login akses ke sistem pakar, daftar akun dan informasi terkait dataset yang digunakan pada skripsi ini. Tampilan antar muka halaman awal dapat dilihat pada gambar 3.1.

Gambar 3.1. Halaman awal

Halaman hasil klasifikasi penyakit muncul ketika sistem selesai memproses klasifikasi yang sebelumnya sudah dimasukan ke dalam sistem pakar. Pada laman ini pengguna hanya dapat melihat hasil klasifikasi dan secara otomatis tersimpan ke dalam sistem untuk dapat dilihat di riwayat klasifikasi. Antar muka hasil klasifikasi penyakit dapat dilihat pada Gambar 3.2.

4. Kesimpulan

Setelah melakukan analisis, perancangan, implementasi dan pengujian, maka diperoleh kesimpulan bahwa:

1. Sistem pakar dapat mengklasifikasikan penyakit kanker payudara berdasarkan sampel darah yang diambil dari dataset *coimbra breast cancer*. Sistem pakar dapat mengklasifikasi pasien menjadi dua golongan yakni pasien yang hanya memerlukan kontrol kesehatan saja dan pasien yang positif menderita kanker payudara.
2. Algortima *k-nearest neighbor* dapat diterapkan untuk mengklasifikasi penyakit kanker payudara melalui aplikasi sistem pakar. Prosedur klasifikasi dengan cara menghitung jarak atribut pasien yang dimasukan ke dalam sistem pakar dengan dataset yang sudah ada. Kemudian dicari jarak yang paling dekat sehingga dapat ditemukan klasifikasi yang tepat.

Saran

Berdasarkan kesimpulan di atas, hal yang diharapkan di masa mendatang sebagai bahan penelitian berikutnya adalah:

1. Pada penelitian lanjutan, dapat diimplementasikan formula penghitung jarak lainnya seperti *Manhattan Distance, Minkowsi Distance* atau *Hamming Distance*.
2. Algoritma berbasis peluang seperti *naïve bayes* juga dapat diimplementasikan pada penelitian ini.
3. Hasil dari klasifikasi sistem pakar ini yang dianggap tepat, dapat ditambahkan kedalam basis data dataset agar memperbanyak referensi kasus kanker payudara yang dapat digunakan untuk mengklasifikasi pasien selanjutnya.

Referensi

Crisostomo, J., Matafome, P., & Santos-Silva, D. (2016). Hyperresistinemia and metabolic dysregulation: a risky crosstalk in obese breast cancer. *Endocrine*, 433-442.

Global Cancer Observatory. (2018). *Database Cancer Worldwide*. Retrieved from Global Cancer Observatory: https://gco.iarc.fr/databases.php

Gorunescu, F. (2011). *Data Mining: Concepts, Models and Techniques.* New York: Springer.

Islam, M., & Poly, T. N. (2019). Machine Learning Models of Breast Cancer Risk Prediction. *bioRxiv*.

Kementerian Kesehatan. (2019, Januari 31). *Hari Kanker Sedunia 2019*. (Kementerian Kesehatan Republik Indonesia) Retrieved Maret 28, 2020, from Kementerian Kesehatan Republik Indonesia: https://www.kemkes.go.id/article/view/19020100003/hari-kanker-sedunia-2019.html

Larose, D. T. (2007). *Data Mining Methods and Model.* New Jersey: John Willey & Sons, Inc.

Nugraha, F. S., Shidiq, M. J., & Rahayu, S. (2019). ANALISIS ALGORITMA KLASIFIKASI NEURAL NETWORK UNTUK DIAGNOSIS PENYAKIT KANKER PAYUDARA. *Jurnal Pilar Nusa Mandiri*.

Patricio, M., Pereira, J., Crisostomo, J., Matafome, P., Gomes, M., Seica, R., & Caramelo, F. (2018). Using Resistin, glucose, age and BMI to predict the presence of Breast Cancer. *BMC Cancer*, 18-29.

Prasetio, R. T., & Riana, D. (2015). A Comparison of Classification Methods in Vertebral Column Disorder with the Application of Genetic Algorithm and Bagging. *2015 4th International Conference on Instrumentation, Communications, Information Technology, and Biomedical Engineering (ICICI-BME).* Bandung.

Purwaningsih, E. (2018). Application of the Support Vector Machine and Neural Network Model Based on Particle Swarm Optimization for Breast Cancer Prediction. *SinkrOn*, 1064–1069.

Rahayu, N. S., & Sandi, V. A. (2018). SISTEM PAKAR DIAGNOSA ERROR SISTEM PADA “PT. DANACO GLOBAL SOLUSI-OXY SYSTEM" MENGGUNAKAN METODE CERTAINTY FACTOR BERBASIS WEB. *Sintech Journal*, 61-69.

Wu, X., Kumar, V., Ross, Q. J., Ghosh, J., & Yang, Q. (2008). Top 10 algorithms in data mining. *Knowledge and Information*.