**REPOSITOR**, Vol. 4, No. 3, Agustus 2022, Pp. 385-396
ISSN : 2714-7975
E-ISSN : 2716-1382


# Penerapan Generalized Association Rule Minning Sebagai Analisa Pasien Penderita Kanker Paru Paru Berdasarkan Data Rekam Medis

**Rangga Kurnia Putra Wiratama[*1], Yufis Azhar[2], Agus Eko Minarno[3]**
[1, 2, 3]IUniversitas Muhammadiyah Malang
ranggakpw@gmail.com[*1], yufis@umm.ac.id[2], aguseko@umm.ac.id[3]

***Abstrak***

*Penggunaan pada data rekam medis yang diambil dari database rumah sakit atau badan kesehatan bertujuan untuk mengetahui secara nyata riwayat pasien yang sebelumnya dinyatakan terkena kanker paru-paru. Namun potensi untuk mengetahui suatu pola penyakit secara manual berupa informasi secara langsung sangat memakan waktu sehingga dilakukan proses kegiatan analisa medis secara manual oleh dokter. Salah satunya adalah pola urutan penyakit pasien yang juga merupakan sebagai atribut pasien. Analisa diagnosa penyakit kanekr paru-paru dengan metode Association Rule menjadi teknik dari data mining yang diperlukan dalam menentukan model atribut dari database rekam medis dari atribut yang sebelumnya dimiliki oleh seorang pasien. Berdasarkan data rekam medis yang akan diperiksa terdapat 1000 data dengan 3 kategori yaitu risiko rendah, risiko sedang dan risiko tinggi. batasan dari aturan asosiasi tersebut tak terhingga sehingga nilai support dan nilai confidence akan terdeteksi oleh program secara otomatis dan nilai tersebut akan dijadikan sebagai variabel evaluasi untuk melakukan klasifikasi untuk menentukan kategori yang mendekati.*

## 1. Pendahuluan

Kanker paru-paru adalah penyakit ketika sel-sel ganas terbentuk di paru-paru seseorang. Penyakit ini ditandai dengan pertumbuhan sel yang tidak terkendali di jaringan paru-paru manusia [1]. Kanker paru-paru sudah dialami oleh banyak orang ketika ada kebiasaan dan aktivitas yang dapat menyebabkan masalah paru-paru, termasuk merokok yang merupakan salah satu penyebab kanker paling umum di Indonesia selama ini.[2]. Berdasarkan berbagai kasus yang terjadi bahwa Kanker paru-paru juga dapat terjadi di lingkungan sekitar, misalnya orang yang rutin terpapar bahan kimia di lingkungan kerja atau terpapar asap rokok orang lain. [3]. Penyakit kanker juga merupakan penyakit yang sangat di waspadai oleh semua orang dan bisa menyebabkan kematian pasien di rumah sakit jika tidak di tangani dengan benar.

Dalam dunia medis, dokter memerlukan sebuah analisa dari pasien mulai dari memeriksa diagnosa penyakit kanker pada paru paru hingga status pasien yang merupakan faktor penyebab terjadinya kanker paru paru. Sedangkan untuk melakukan penelitian tersebut memerlukan sebuah teknologi dan sebuah analisa yang cukup besar untuk bertahan lama. Penelitian ini bertujuan untuk membuat sistem aplikasi yang mampu mendeteksi kanker paru secara dini pada pasien dan mengklasifikasikan paru-paru kedalam tipe kanker dengan resiko yang rendah sampai ke resiko tinggi serta menganalisa performansi sistem yang digunakan untuk membantu dokter menangani pasien yang teridap penyakit kanker paru paru sedini mungkin dengan menggunakan Association Rule dengan cara mencari keterkaitan antara satu atribut pasien dalam data rekam medis dengan atribut yang lain yang termasuk dalam kategori resiko rendah sampai ke resiko tinggi.

Untuk mengetahui pola penyakit kanker paru – paru pada pasien yang disebabkan oleh berbagai faktor penyebab dengan menentukan model analisis penyakit dan menganalisis data pasien di rumah sakit menjadi sangat penting. [4]. Seluruh data aktivitas pasien telah di rekap dalam dunia medis dalam bentuk data rekam medis dari pasien yang telah dirawat sebelumnya[5]. Data rekam medis menyimpan semua data yang berisikan catatan dan dokumen mulai dari identitas lengkap dari pasien hingga berupa hasil pemeriksaan, pengobatan, tindakan dan pelayanan lainnya[6]. Maka sesuai dengan hasil data riwayat pasien tersebut akan dijadikan sebuah dataset yang di gunakan sebagai pengujian pada metode penelitian,

Untuk mengidentifikasi sebuah status penyakit kanker pasien dengan kategori ketertaitan tertentu dibutuhkannya pencocokan data secara satu per satu untuk mengetahui berapa persen akurasi support[7]. Dan juga terdapat faktor lain seperti waktu, akurasi identifikasi, dan tenaga

ahli juga menjadi pertimbangan saat mencari ketertaikan penyakit kanker tersebut. Untuk itu dibutuhkan sebuah teknologi yang dapat melakukan pencarian sebuah rule yang mengidentifikasi pola penyakit pada data rekam pasien [8]. Pada permasalahan ini akan dilakukan sebuah penelitian yang menggunakan salah satu bagian dari metode Data Minning yang biasa di gunakan untuk menganalisa suatu pola dalam data adalah metode Association Rule.

Pada peneltian sebelumnya yang melakukan sebuah analisa menggunakan algoritma apriori telah mendapatkan hasil identifikasi pola data dari hasil penelusuran pada data historis [9]. Dan terdapat sebuah Penelitian lainnya berkata bahwa prosedur pemecahan Apriori jua adalah bentuk terapan berdasarkan data mining yg membangun sebuah contoh pengetahuan berupa pola & aturan dengan nilai confidence[10]. Model pengetahuan terlatih sering digunakan untuk memprediksi tren masa depan dalam data. Implementasi logika aturan yang terkait dengan algoritma Apriori dilakukan dalam dua langkah, yaitu dengan menemukan semua himpunan elemen yang besar dan membentuk aturan yang nilainya berdasarkan nilai confidence dan nilai support.

Pendekatan sistematis yang digunakan dalam penelitian untuk menentukan model adalah data mining dengan metode association rule. [11]. Data mining adalah proses pengolahan informasi dari database yang besar, termasuk ekstraksi, input, kelengkapan, dan penyajian informasi sehingga dapat digunakan untuk membuat keputusan bisnis yang penting[12]. Metodologi ini akan mengambil semua pola yang dapat diamati dalam database dan mengidentifikasi pola data berdasarkan atribut yang ditentukan.

Penelitian sebelumnya yang dilakukan oleh Juliet Rani Rajan, dkk [13] yang berjudul Multi-Class Neural Networks to Predict Lung Cancer. Pada penelitian ini mengusulkan model Multi Class Neural Network dengan mengikuti Supervised Learning untuk menemukan beberapa struktur yang mendasari data yang memerlukan data pelatihan dan data pengujian. Dengan melakukan klasifikasi berdasarkan pengetahuan yang dipelajari dari data pelatihan selama proses Learning pada mesin. Seperti yang dibahas penggunaan data yang tercantum model ini membantu praktisi medis dalam proses pra diagnosis untuk deteksi dini kanker sehingga meningkatkan tingkat kelangsungan hidup pasien selama 5 tahun. Tetapi dalam model tersebut tidak menggambarkan prediksi akurasi sesuai dengan kelas kategori low – high dengan jelas, sehingga menggunkan prediksi seluruh dataset untuk mengetahui prediksi pasien.

Penelitian lainnya yang dilakukan oleh Radhika P R, dkk [14] yang berjudul A Comparative Study of Lung Cancer Detection using Machine Learning Algorithms. Penelitian ini menggunakan prediksi dan klasifikasi data medis menggunakan metode Regresi Logistik, SVM, Decision Tree dan Naïve Bayes dengan analisis perbandingan tingkat akurasi dari setiap pengklasifikasi yang disajikan. Dalam grafik kinerja, hasil yang dihasilkan berbeda untuk setiap pengklasifikasi pada kumpulan data kanker paru paru. Hasil terbaik diberikan oleh algoritma SVM yang menggunakan dimensi tinggi untuk mengklasifikasikan observasi sehingga performanya paling baik. Kelemahan dalam algoritma ini adalah kurangnya pemrosesan dalam klasifikasi dalam kategori low – high. Peneltian ini masih melanjutakan pengembangan dengan menambahkan pra-pemrosesan ekstra serta tingkat akurasi yang dapat ditingkatkan dengan menggunakan metode yang lainnya.

Berdasarkan permasalahan dari penelitian sebelumnya dalam studi penelitian ini akan di usulkan sebuah metode Assosiation Rule untuk mencari sebuah rule dari data rekam medis pasien yang memiliki riwayat klinis kanker paru – paru berdasarkan dengan pola medis yang mendeteksi keterkaitan penyakit yang diderita pasien tersebut termasuk dalam kategori rendah, sedang, atau tinggi. Perbedaan penelitian ini dengan penelitian sebelumnya adalah dengan mengusulkan perbandingan metode dari Association Rule dengan Algoritma Apriori dan Generalized Sequential Pattern yang di harapkan dapat menemukan rule dengan akurasi yang tinggi serta dapat mendapatkan nilai confidence dan nilai support yang tinggi dari setiap data yang telah di dapatkan.

## 2. Metode Penelitian

Metode analisis ini sering diketahui sebagai metode dengan sistem Analisis keranjang pasar di mana analisis ini biasa digunakan untuk menganalisis isi keranjang belanja konsumen di supermarket [15].

### 2.1 Dataset

Bentuk dari Dataset yang digunakan untuk penelitian ini berupa file excel yang berjumlah 1000 data. Sumber dataset tersebut berasal dari website https://data.world/cancerdatahp/lung-

cancer-data. Data tersbut di klasifikasikan berdasarkan data rekam medis dari perawatan pasien berdasarkan 3 kategori yaitu low, medium, dan high. Pada penelitian ini beberapa atribut lain yang tidak digunakan dalam pengujian akan dihilangkan, seperti Jenis Kelamin dan Usia pada data pasien. Dataset dibagi berdasarkan dari 3 kategori data yaitu 303 total data low, 332 total data medium, dan 365 total data high. Menurut rekam data tersebut sebanyak 21 kolom status pasien yang menjadi penyebab pasien tersebut terkena kanker yang terdiri dari Air Pollution, Alcohol use, Dust Allergy, OccuPational Hazards,'Genetic Risk, chronic Lung Disease, Balanced Diet, Obesity, Smoking, Passive Smoker, Chest Pain, Coughing of Blood, Fatigue, Weight Loss,'Shortness of Breath, Wheezing, Swallowing Difficulty, Clubbing of Finger Nails, Frequent Cold, Dry Cough, Snoring. Seperti Pada Tabel 1 dari 21 kolom tersebut setiap pasien di rekam berdasarkan level tingkat dari status tersebut dari tingkat yang terendah yaitu level 1 sampai dengan level 8.

*Tabel 1. Sampel dataset rekam medis*

| Pasien | Atribut 1 | Atribut 2 | Atribut 3 | Class |
|---|---|---|---|---|
| P1 | 2 | 4 | 5 | Low |
| P10 | 3 | 1 | 5 | Medium |
| P100 | 4 | 5 | 6 | High |
| P1000 | 7 | 7 | 7 | High |
| P101 | 6 | 8 | 9 | High |
| P102 | 4 | 5 | 6 | High |
| P103 | 2 | 4 | 5 | Low |
| P104 | 3 | 1 | 4 | Low |

### 2.2 Data Mining

Data mining secara sederhana dapat dipahami sebagai mengekstraksi atau "mengekstraksi" pengetahuan dari sejumlah besar data, meskipun data mining sering disebut sebagai "ekstraksi pengetahuan dari data", sehingga namanya menjadi lebih panjang. Dari nama-nama tersebut, beberapa kata yang dimaksud yaitu "menggali pengetahuan" tidak dapat diungkapkan dengan kata-kata yang lebih pendek ketika penekanannya pada big data. Namun, penambangan adalah istilah eksplisit yang mengacu pada proses yang menggambarkan atau menghasilkan informasi kecil dan berharga dari data dalam jumlah besar[16]. Pada data mining terdapat langkah-langkah yang dilakukan untuk memperoleh pengetahuan yang terdiri dari serangkaian iterasi untuk mendapatkan hasil yang diinginkan. Langkah-langkah dilakukan secara berurutan mulai dari Data Cleaning, Data Intergration, Data Selection, Data Transformation, Data Evaluation, dan Knowledge Reprentation.

#### 2.2.1 Data Cleaning

Pada proses ini terhadap data besar dilakukan sebelum memulai memproses sebuah data yang bertujuan untuk membersihkan suatu noise atau data yang tidak jelas atau konsisten. Prosesnya adalah data yang mungkin tidak diperlukan akan dihapus dan hanya data yang penting saja yang digunakan untuk implementasi lebih lanjut[17]. Dan setelah setelah itu adalah menyimpan di database, di mana data disimpan dengan tujuan untuk menjaga semua dari sumber asli nya dan akan menciptakan sebuah generator yang menghasilkan laporan setelah memproses semua data dengan menentukan apakah pasien terkena kanker atau tidak[18].

#### 2.2.2 Data Intergration

Tujuan utama dari metode Intergration data adalah untuk mengekstrak data knowledge tambahan dari beberapa dataset yang tidak dapat diperoleh dari satu set data saja. Untuk mencapai tujuan ini, Intergration data harus memenuhi banyak tantangan komputasi. Tantangan ini muncul karena berbagai ukuran, format, dan dimensi data yang terintegrasi, serta karena kompleksitas, noisiness, konten informasi, dan kesesesan bersama[19].

#### 2.2.3 Data Selection

Proses ini mengambil data yang relevan untuk dilakukannya sebuah proses analisis yang diambil dari database sehingga data yang secara umum mempunyai kesamaan jenis akan di

jadikan sebagai satu data untuk menghindari terjadinya duplikasi data yang tidak di perlukan lagi dalam proses kedepan.

### 2.2.4 Data Evaluation

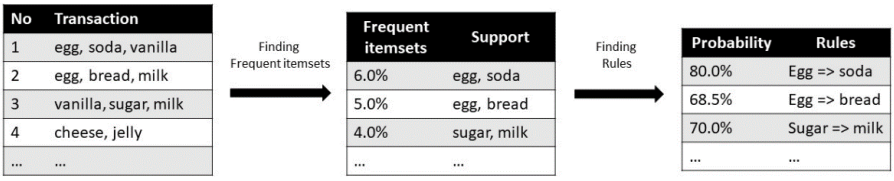
Pada sebuah tahap processing data, data yang di telah di proses akan dilakukan sebuah evaluasi yang di gunakan untuk mengidentifikasi model yang mewakili data pengetahuan dan berguna saat mengidentifikasi model yang sangat menarik yang mewakili data pengetahuan berdasarkan beberapa tindakan menarik dari pola yang sudah di proses sebelumnya pada tahap processing data[17].

### 2.2.5 Knowledge Reprentation

Proses ini dilakukan dimana pada data yang sudah di lakukan processing selanjutnya akan di tampilkan dalam bentuk visual dan gambaran dari suatu data knowledge yang digunakan untuk menampilkan sebuah representasi data sebagai evaluasi hasil akhir dari suatu data yang telah di lakukan proses panjang[20]. Sehingga pada tahap ini seluruh output akan di gambarkan dalam bentuk visual berupa diagram dan chart sebagai tampilan dari hasil akhir data yang di proses.

## 2.3 Aturan Asosiasi

Association rule mining atau analisis asosiasi merupakan suatu teknik data mining dengan melakukan sebuah proses yang menemukan sebuah aturan asosiasi antara suatu kombinasi item dari dalam database yang ada. Penerapan dari algoritma ini adalah sebagai contoh pada analisis pasar pembelian dapat menentukan seberapa besar kemungkinan pelanggan membeli barang pertama sekaligus item kedua sehingga akan membentu sebuah kombinasi yang disebut dengan itemset. Dengan ini pemilik pasar tersebut dapat mengatur segala kombinasi dan penempatan semua barang atau item dengan merancang sebuah teknik pemasaran baru dengan mengetahui stok produk yang harus di lakukan pembelian besar-besaran atau bisa dengan melalui pembuatan kupon diskon untuk kombinasi barang tertentu. Ada dua langkah didalam algoritma ini meliputi proses dari cara menentukan sebuah kombinasi yang cocok dari berbagai itemset seperti yang di jelaskan Pada Gambar 1, yaitu sebagai Langkah pertama adalah melakukan perhitungan untuk mencari himpunan elemen frequent dan kemudian langkah kedua mencari aturan asosiasi himpunan frequent itemsets dari kumpulan tersebut. yang telah di kombinasikan sebelumnya.




*Gambar 1. Dua Langkah Proses Proses Association Rules*

## 2.4 Algoritma Apriori

Algoritma apriori digunakan untuk mencari frequent itemset yang memenuhi minsup kemudian mendapatkan rule minconf dari frequent itemset. Algoritma ini mengontrol perluasan himpunan elemen dari kandidat itemset hasil frequent itemset dengan support-based pruning untuk menghapus kumpulan elemen yang tidak menarik dengan mengatur nilai minsup. Prinsip apriori ini adalah bahwa jika himpunan elemen diklasifikasikan sebagai himpunan elemen beraturan, yang memiliki dukungan lebih dari yang ditentukan sebelumnya, maka semua himpunan bagiannya juga termasuk dalam kelompok himpunan elemen beraturan begitu juga sebaliknya.

Algoritma ini bekerja dengan cara menghasilkan kandidat baru dari k-itemset dari frequent itemset yang dihasilkan dari langkah sebelumnya. Kemudian menghitung nilai support k-itemset dari aturan yang telah di buat sebelumnya. Sebuah Itemset yang mempunyai minimum nilai support di bawah dari minsup akan segera di hilangkan. Algortima ini akan berhenti ketika tidak terdeteksi lagi sebuah frequent itemset baru yang akan dibentuk. Kedua, dari hasil frequent itemset tersebut, selanjutnya adalah menghitung sebuah nilai minconf yang mengikuti rumus sesuai dengan syarat yang ada. Nilai Support tidak perlu dihitung kembali, ketika metode

generate frequent itemset didapatkan dengan melihat minsup-nya. Ketika rule yang dihasilkan dapat memenuhi syarat batasan yang tinggi, maka rule tersebut tergolong strong rules atau aturan baru yang memiliki frequent yang sangat kuat.

Di langkah awal dari algoritma apriori menentukan nilai – nilai yang diperlukan yaitu dengan menganalisis sebuah pola frekuensi yang tinggi y dengan cara menentukan sebuah kombinasi item yang telah memenuhi syarat minimum dari nilai *support* sebuah item dalam basis data yang diperoleh dengan Persamaan berikut.

$$\text{Support (A)} = \frac{\sum \text{Jumlah transaksi mengandung A}}{\sum \text{Total Transaksi}} \text{x } 100\%$$

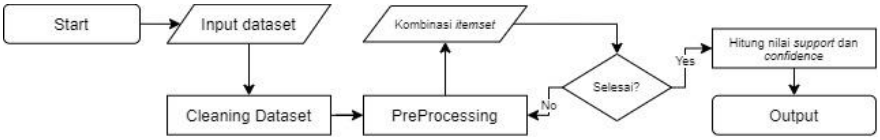
dan nilai *support* pada 2 item akan di peroleh dengan Persamaan berikut.

$$\text{Support (A} \cap \text{B)} = \frac{\sum \text{Jumlah transaksi mengandung A dan B}}{\sum \text{Total Transaks}} \text{x } 100\%$$

Langkah selanjutnya setelah mendapatkan model frekuensi tinggi adalah menentukan aturan asosiasi dengan memenuhi persyaratan kepercayaan minimum dengan menghitung kepercayaan aturan asosiasi A ke B dengan menghitung *confidence* aturan assosiatif A ke B Nilai *confidence* dari aturan A ke B diperoleh dari rumus berikut:

$$\text{Confidence} = \text{P(B\A)} = \frac{\sum \text{Jumlah transaksi mengandung A dan B}}{\sum \text{Jumlah transaksi mengandung A}} \text{x } 100\%$$

Dari item yang dipilih, sistem akan secara otomatis menentukan kombinasi kedua item dan kemudian menghitung setiap nilai *support* dari nomor yang muncul di setiap transaksi, kemudian sistem akan terus menghapus nilai *support* dari dua kombinasi item tersebut yang di bawah minimum support yang dinamakan sebagai pembentukan kombinasi dua item. Proses ini akan terus berlanjut hingga mencapai kombinasi maksimal dari system transaksi item yang bisa didapatkan.




*Gambar 2. Flowchart algoritma apriori*

## 3. Hasil dan Pembahasan

### 3.1 Transform Data

Pada tahap ini seluruh data dalam database rekam medis akan di proses dan di transform menggunakan fungsi One Hot Encoding dan menjadikan sebagai atribut baru sehingga terbentuk dengan menambahkan code angka di setiap atribut yang sebelumnya digunakan untuk menkategorikan sebagai level atribut. seluruh atribut pertama akan dilakukan splitting sehingga membentuk atribut sesuai dengan banyaknya level atribut yang tersedia dalam dataset.

Kemudian setelah di proses dengan menggunakan fungsi tersebut, maka akan di temukan hasil yang berbeda dari dataset asal nya. Perbedaan dari dataset sebelumnya adalah setiap atribut yang memiliki kode level yang berbeda akan dijadikan sebagai bilangan boolean dengan masing-masing kode level yang sudah di tentukan dalam dataset asal. Hasil output dari proses data transform tersebut seperti Pada Tabel 2 yang dimana setiap kode level mencocokan dengan atribut yang sudah di sediakan.

*Tabel 2. Hasil Data Transform*

| Air Pollution_1 | Air Pollution_2 | Air Pollution_3 | Air Pollution_4 | Air Pollution_5 | ... |
|---|---|---|---|---|---|
| 0 | 1 | 0 | 0 | 0 | ... |
| 0 | 1 | 0 | 0 | 0 | ... |
| 0 | 0 | 1 | 0 | 0 | ... |

| | | | | | |
|---|---|---|---|---|---|
| 0 | 0 | 1 | 0 | 0 | ... |
| 0 | 1 | 0 | 0 | 0 | ... |
| ... | ... | ... | ... | ... | ... |

### 3.2 Splitting Data

Untuk melakukan pengujian dalam penggunaan klasifikasi dalam sistem asosiasi diperlukannya sebuah data test yang bertujuan sebagai uji coba dan juga sebagai evaluasi dari keakuratan proses aturan asosiasi yang telah di proses. Sehingga di lakukan splitting data 80% sebagai data train dan 20% sebagai data test. Proses splitting data dilakukan di setiap kategori yaitu dengan membagi 20% dari masing-masing total di setiap kategori seperti Pada Tabel 3.

*Tabel 3. Nilai support per item*

| Kategori | Data Train | Data Test |
|---|---|---|
| Low Risk | 273 | 30 |
| Medium Risk | 299 | 33 |
| High Risk | 329 | 36 |

### 3.3 Nilai Support

Dalam sebuah proses pengujian menggunakan dataset asal yaitu data rekam medis yang sudah di lakukan splitting berdasarkan 3 kategori yaitu low, medium, dan high dan yang sudah dilakukan transformasi data sebelumnya. Maka selanjutnya adalah menghitung nilai support setiap kategori dari masing-masing atribut yang tersedia. Dari berbagai atribut akan menghasilkan sebuah nilai support yang berbeda beda.

Seluruh hasil pada setiap katgori akan dilakukan pencarian berdasarkan minimum support yang ditentukan secara manual maka akan didapatkan sebuah hasil nilai support dari setiap atribut yang telah di tentukan dari setiap kategori. Berdasarkan Pada Tabel 4, 5, dan 6 menunjukan sampel dari hasil support pada setiap kategori dengan minimum support yang di tentukan dan juga menunjukan hasil dengan nilai support yang maksimal.

*Tabel 4. Sampel Hasil Nilai Support Kategori Low Risk*

| No | Support | Itemsets |
|---|---|---|
| 0 | 0.366337 | (Air Pollution_2) |
| 1 | 0.369637 | (Air Pollution_3) |
| 2 | 0.303630 | (Alcohol use_1) |
| 3 | 0.363036 | (Alcohol use_2) |
| 4 | 0.231023 | (Alcohol use_3) |
| ... | ... | ... |
| 86 | 0.201320 | (Wheezing_2, Coughing of Blood_4) |
| 87 | 0.201320 | (Shortness of Breath_3, Fatigue_2) |
| 88 | 0.204620 | (Weight Loss_2, Wheezing_4) |
| ... | ... | ... |

*Tabel 5. Sampel Hasil Nilai Support Kategori Medium Risk*

| No | Support | Itemsets |
|---|---|---|
| 0 | 0.391566 | (Dust Allergy_7) |
| 1 | 0.391566 | (Balanced Diet_2) |
| 2 | 0.361446 | (Obesity_4) |
| 3 | 0.361446 | (Smoking_1) |
| 4 | 0.487952 | (Passive Smoker_2) |
| ... | ... | ... |
| 86 | 0.361446 | (Obesity_4, Smoking_1, Wheezing_5, Frequent Co... |
| 87 | 0.361446 | (Weight Loss_7, Obesity_4, Wheezing_5, Frequen... |
| 88 | 0.361446 | (Weight Loss_7, Smoking_1, Wheezing_5, Frequen... |
| ... | ... | ... |

*Tabel 6. Sampel Hasil Nilai Support Kategori High Risk*

| No | Support | Itemsets |
|---|---|---|
| 0 | 0.673973 | (Air Pollution_6) |
| 1 | 0.753425 | (Dust Allergy_7) |
| 2 | 0.753425 | (OccuPational Hazards_7) |
| 3 | 0.594521 | (Genetic Risk_7) |
| 4 | 0.542466 | (chronic Lung Disease_6) |
| ... | ... | ... |
| 86 | 0.515068 | (Obesity_7, Genetic Risk_7, Dust Allergy_7, Oc... |
| 87 | 0.515068 | (chronic Lung Disease_6, Obesity_7, Dust Aller... |
| 88 | 0.515068 | (chronic Lung Disease_6, Obesity_7, Genetic Ri... |
| ... | ... | ... |

Sehingga dihasilkan sebuah maksimum total dan maksimum size yang didapat setelah dilakukan proses dari code yang telah disusun dan Pada Tabel 7 menjelaskan hasil evaluasi dari masing-masing kategori.

*Tabel 7. Evaluasi Nilai Support Setiap Kategori*

| Kategori | Min Support | Maks Combination | Maks Itemset Size |
|---|---|---|---|
| Low | 0,3 | 420 | 2 |
| Medium | 0,35 | 129 | 3 |
| High | 0,5 | 220 | 4 |

### 3.3 Nilai Confidence

Dalam pencarian nilai confidence yaitu berdasarkan dari item A yang mengandung item B dan dibagi dengan total keseluruhan data dan di kali dengan 100%. Maka dari itu akan dilakukan sebuah evaluasi dengan memasangkan kedua jenis item yang berbeda dari setiap itemset yang tersedia dan dilakukan proses perhitungan secara otomatis oleh program. Sehingga di dapat sebuah nilai confidence yang berbeda beda.

Untuk mengetahui besar nilai confidence yang akan menjadi output maka digunakan itemset A sebagai nilai atendance dan item B sebagai nilai consequents. Dan ketika di cari berdasarkan nilai minimum dari nilai confidence yang di inginkan, akan tampil secara keseluruhan kombinasi dari kedua itemset yang berbeda beda dengan nilai support masing-masing.

Hasil evaluasi dari nilai confidence akan terlihat Pada Tabel 8, 9 dan 10 menjelaskan sebuah evaluasi dari masing-masing kategori dengan kombinasi pertemuan antara itemset atecedence dan consequents dengan nilai support masing-masing yang dimana tujuan evaluasi ini adalah mencari nilai support tertinggi yang dihasilkan dari nilai confidence yang mendekati angka 1 atau sama dengan 1.

*Tabel 8. Sampel Hasil Evaluasi Nilai Confidence Kategori Low Risk*

| No | Antecedents | Consequents | Confidence |
|---|---|---|---|
| 0 | (Clubbing of Finger Nails_1) | (Air Pollution_2) | 0.859155 |
| 1 | (Obesity_3) | (Alcohol use_1) | 0.861111 |
| 2 | (OccuPational Hazards_3) | (Genetic Risk_2) | 0.876543 |
| 3 | (Dry Cough_4) | (Genetic Risk_2) | 1.000000 |
| 4 | (chronic Lung Disease_2) | (Genetic Risk_3) | 0.878049 |
| ... | ... | ... | ... |

*Tabel 9. Sampel Hasil Evaluasi Nilai Confidence Kategori Medium Risk*

| No | Antecedents | Consequents | Confidence |
|---|---|---|---|
| 0 | (Obesity_4) | (Balanced Diet_2) | 1.000000 |
| 1 | (Balanced Diet_2) | (Obesity_4) | 0.923077 |
| 2 | (Balanced Diet_2) | (Smoking_1) | 0.923077 |
| 3 | (Smoking_1) | (Balanced Diet_2) | 1.000000 |
| 4 | (Weight Loss_7) | (Balanced Diet_2) | 1.000000 |
| ... | ... | ... | ... |

T*abel 10. Sampel Hasil Evaluasi Nilai Confidence Kategori High Risk*

| No | Antecedents | Consequents | Confidence |
|---|---|---|---|
| 0 | (Dust Allergy_7) | (OccuPational Hazards_7) | 1.000000 |
| 1 | (OccuPational Hazards_7) | (Dust Allergy_7) | 1.000000 |
| 2 | (Dust Allergy_7) | (Obesity_7) | 0.930909 |
| 3 | (Chest Pain_7) | (Dust Allergy_7) | 0.962406 |
| 4 | (Dust Allergy_7) | (Chest Pain_7) | 0.930909 |
| ... | ... | ... | ... |

### 3.4 Evaluasi Nilai & Confidence

Setelah mengetahui seluruh kombinasi itemset pada sub bab sebelumnya mengenai hasil dari perhitungan untuk mendapatkan nilai support dan confidence yang dimana merupakan nilai dari banyak nya itemset yang muncul serta atribut yang didapatkan dari masing-masing kombinasi, maka Pada Tabel 11 didapatkan sebuah rules yang memiliki nilai Support dan nilai Confidence tertinggi dari ketiga kategori yang sudah di tentukan.

| Kategori | Rules / Itemsets | Evaluasi | |
|---|---|---|---|
| Low Risk | (Frequent Cold_2) | Support<br>Confidence | 0.435644<br>1.000000 |
| Medium Risk | (Wheezing_5) | Support<br>Confidence | 0.515060<br>1.000000 |
| High Risk | (Dust Allergy_7), (OccuPational Hazards_7), | Support<br>Confidence | 0.756849<br>1.000000 |

### 3.5 Pengujian Data Rekam Medis

Pengujian ini dilakukan dengan menghitung data test dengan menggunakan nilai support dari masing-masing itemset yang telah ditentukan. Rumus yang digunakan ialah menggunakan pointwise mutual information yang dimana akan melibatkan hasil dari proses pencarian nilai support dari data train yang sudah di lakukan untuk melakukan evaluasi dan menentukan sebuah klasifikasi dari data test yang sudah di sediakan. Hal pertama yang harus dilakukan adalah mencocokan dari nilai support itemset dari data test dan data train.

Sebagai contoh dalam sebuah sampel data test yang di ambil secara acak dari hasil splitting data memiliki atribut yaitu ap6, au8, da7, oh7, gr7, cld6, bd2, o4, s1, ps2, cp4, cob3, f2, wl7, sob6, w5, sd1, cofn9, fc3, dc4, sn2. Maka beberapa atribut akan dicari berdasarkan nilai confidence yang di uji berdasarkan dari nilai data train kemudian beberapa dari atribut di temukan

sejumlah kombinasi itemset setelah di proses menggunakan asosiasi dengan algoritma apriori. Maka hasil evaluasi dari sampel diatas seperti Pada Tabel 11 dan Tabel 12.

*Tabel 11. Evaluasi Sample Dataset*

| { ap6, au8, da7, oh7, gr7, cld6, bd2, o4, s1, ps2, cp4, cob3, f2, wl7, sob6, w5, sd1, cofn9, fc3, dc4, sn2 } | | | |
|---|---|---|---|
| Antecedents | Consequents | Antecedent Support | Consequent Support |
| (ap6) | (au8) | 0.361204 | 0.391304 |
| (ap6) | (da7) | 0.391304 | 0.361204 |
| (ap6) | (oh7) | 0.391304 | 0.361204 |
| (ap6) | (gr7) | 0.361204 | 0.391304 |
| ... | ... | ... | ... |
| (cob3, fc3, dc4, ...) | (s1, f2, sob6, ...) | 0.511706 | 0.361204 |

*Tabel 12. Evaluasi Nilai Support dan Confidence*

| { ap6, au8, da7, oh7, gr7, cld6, bd2, o4, s1, ps2, cp4, cob3, f2, wl7, sob6, w5, sd1, cofn9, fc3, dc4, sn2 } | | | |
|---|---|---|---|
| Antecedents | Consequents | Support | Confidence |
| (ap6) | (au8) | 0.361204 | 1.000000 |
| (ap6) | (da7) | 0.361204 | 0.923077 |
| (ap6) | (oh7) | 0.361204 | 0.923077 |
| (ap6) | (gr7) | 0.361204 | 1.000000 |
| ... | ... | ... | ... |
| (cob3, fc3, dc4, ...) | (s1, f2, sob6, ...) | 0.923077 | 1.803922 |

Setelah didapat hasil dari nilai support maka dilakukan perhitngan dari nilai support data train dengan data test menggunakan pointwise mutual information serta melakukan evaluasi dalam mencari rata-rata dari semua nilai lift beserta nilai interest dari semua itemset sebagai pembuktian akurasi dalam melakukan klasifikasi dari masing-masing kategori.

Adapun sebuah rule yang telah di dapat yang di tentukan berdasarkan nilai support tertinggi dan penyesuaian dari nilai confidence di berbagai kategori sesuai dengan contoh di ambil beberapa rule yaitu sebagai berikut:

1. Jika pasien mempunyai jari tabuh (Clubbing of Finger Nails) dengan status level 1 dan pernah terkena polusi udara (Air Polution) di status level 2, kemungkinan pasien memiliki resiko kanker rendah.
2. Jika pasien mempunyai resiko dalam pekerjaan (Occupational Hazard) dengan status level 3 dan mengalami resiko dari genetika (Genetic Risk) di status level 2. kemungkinan pasien memiliki resiko kanker rendah.
3. Jika pasien mengalami batuk kering (Dry Cough) dengan status level 1 dan mengalami resiko dari genetika (Genetic Risk) di status level 2. kemungkinan pasien memiliki resiko kanker rendah.
4. Jika kesehatan diet pasien seimbang (Balanced Diet) dengan status level 2 dan juga memiliki obesitas (Obesity) pada status level 4. kemungkinan pasien memiliki resiko kanker sedang.
5. Jika pasien adalah perokok (Smoking) dengan status level 1 dan mempunyai kesehatan diet yang seimbang (Balanced Diet) pada status level 2 maka kemungkinan untuk terkena resiko kanker sedang
6. Jika pasien kehilangan berat badan (Weight Loss) dengan status level 7 serta pasien tersebut kesehatan dietnya (Balanced Diet) di status level 2 maka kemungkinan untuk terkena resiko kanker sedang.
7. Jika pasien pasien mempunyai resiko dalam bekerja (Occupational Hazards) di status level 7 dan memiliki alergi pada debu (Dust Allergy) pada status level 7, maka pasien tersebut kemungkinan besar memiliki resiko terkena kanker yang cukup tinggi
8. Jika pasien memiliki sejak dulu memiliki obesitas (Obesity) di level 7 dan resiko dalam pekerjaan (Occupational Hazards) pada level 7 serta pernah mengalami sesak nafas (Chest Pain) di status level 7, kemungkinan besar juga pasien tersebut memiliki resiko kanker paru-paru yang cukup tinggi.

Sehingga dapat disimpulkan bahwa dari setiap kategori resiko kanker, beberapa rule telah di dapat dan memenuhi syarat sebagai nilai support dan nilai confidence tertinggi. Jika dituliskan dalam bentuk konsep aturan asosiasi maka setiap atribut dari itemset asal dan itemset yang menjadi acuan adalah sebagai berikut:

1. CFN1 □ AP2 sebuah rule itemset dengan support 20% dan confidence 85%
2. OH3 □ GR2 sebuah rule itemset dengan support 23% dan confidence 86%
3. DC4 □ GR2 sebuah rule itemset dengan support 20% dan confidence 87%
4. BD2 □ O4 sebuah rule itemset dengan support 36% dan confidence 92%
5. S1 □ BD2 sebuah rule itemset dengan support 36% dan confidence 100%
6. WL7 □ BD2 sebuah rule itemset dengan support 36% dan confidence 100%
7. OH7 □ DA7 sebuah rule itemset dengan support 75% dan confidence 100%
8. O7,OH7 □ CP7 sebuah rule itemset dengan support 75% dan confidence 100%

Nilai lift merupakan nilai evaluasi yang menggunakan perhitungan antara dua nilai confidence dari data train dan data test yang juga berfungsi untuk mengetahui kedekatan antara data test yang akan di lakukan klasifikasi dengan data train. Sedangkan pada nilai interest akan melibatkan kombinasi itemset dengan nilai support tertinggi dan juga melibatkan atribut satuan dari data test yang berguna seberapa besar kedekatan antara itemset dari data yang akan di uji ke dalam dataset yang telah diproses dengan aturan asosiasi. Sehingga dua nilai tersebut mempengaruhi hasil dari klasifikasi dengan menentukan data tersebut termasuk dalam salah satu kategori yang telah di tentukan. Sebagai contoh Pada Tabel 13 yaitu perhitungan klasifikasi menggunakan nilai lift dan interest yang melibatkan nilai support dan confidence dari data uji dengan dataset.

*Tabel 13. Evaluasi Nilai Support dan Confidence*

| Kategori | Evaluasi | |
|---|---|---|
| Low Risk | Lift | 0.255556 |
| | Interest | 0.206956 |
| Medium Risk | Lift | 0.923077 |
| | Interest | 0.830434 |
| High Risk | Lift | 0.677812 |
| | Interest | 0.6347826 |

## 4. Kesimpulan

Berdasarkan hasil dari nilai evaluasi sampel dari data test, angka yang paling mendekati dari ketiga kategori menunjukkan kategori medium risk dari rata-rata perhitungan seluruh atribut jika di hitung berdsasarkan nilai support dan nilai confidence. Hasil tersebut menunjukan nilai lift sebesar 0.923077 dan nilai interest sebesar 0.830434. ada kemungkinan juga dalam sampel tersebut memasuki kategori high risk akan tetapi nilai lift hanya sebesar 0.677812 dan nilai interest sebesar 0.6347826 yang lebih kecil dari kategori medium risk.

Dari seluruh nilai evaluasi yang telah ditentukan, maka dapat disimpulkan berdasarkan atribut dari data pasien bahwa pasien tersebut pernah terkena polusi udara (Air Polution) di level 6 dan memiliki obesitas (Obesity) dengan status level 4 dan seterusnya dengan atribut yang menyebabkan resiko pasien terkena kanker paru-paru. Dari data sampel dari data test yang telah di klasifikasi, kombinasi rule terbaik yang di dapat yaitu DC4 □ GR2, BD2 □ O4, AP6 □ CFN2, dan lebih banyak lagi kombinasi lain yang memiliki nilai support 36% dan nilai confidence 100%. Dan seluruh rule tersebut terdapat pada itemset yang dimiliki dari data uji kategori medium risk dengan nilai akurasi interest sebesar 0,8%.

## Referensi

[1] D. Pembimbing and J. Matematika, “Klasifikasi Kanker Paru Dengan Menggunakan Algoritma Classify By Sequence ( Cbs ) Lung Cancer Classification Using Classify By Sequence ( Cbs ) Algorithm,” 2015.

[2] R. T. Prasetio and S. Susanti, “Prediksi Harapan Hidup Pasien Kanker Paru Pasca Operasi Bedah Toraks Menggunakan Boosted k-Nearest Neighbor,” *J. RESPONSIF Ris. Sains \& …*, vol. 1, no. 1, pp. 64–69, 2019, [Online]. Available: http://ejurnal.ars.ac.id/index.php/jti/article/view/66.

[3] F. T. Waruwu, E. Buulolo, E. Ndruru, K. Kunci, A. Apriori, and R. Penyakit, “KOMIK

(Konferensi Nasional Teknologi Informasi dan Komputer) Implementasi Algoritma Apriori Pada Analisa Pola Data Penyakit Manusia Yang Disebabkan Oleh Rokok," vol. I, pp. 176–182, 2017.

[4] K. Auliasari, Y. Susanti, J. K. Raya Karanglo, and J. Timur, "Analisis Keterkaitan Penyakit Pasien pada Puskesmas Menggunakan Metode Association Rule," *Informatics J.*, vol. 1, no. 2, pp. 59–66, 2016.

[5] X. Zhu, Y. Liu, Q. Li, Y. Zhang, and C. Wen, "Mining effective patterns of Chinese medicinal formulae using top-K weighted association rules for the internet of medical things," *IEEE Access*, vol. 6, pp. 57840–57855, 2018, doi: 10.1109/ACCESS.2018.2873677.

[6] J. K. Abdul Aziz Priatna, Rani Megasari, "Penerapan Association Rules Menggunakan Algoritma Apriori Pada Sistem Rekomendasi Pemilihan Resep Obat Berdasarkan Data Rekam Medis," *… J. Apl. dan …*, vol. 1, no. 2, pp. 55–60, 2018, [Online]. Available: http://jatikom.cs.upi.edu/index.php/jatikom/article/view/18.

[7] J. A. Delgado-Osuna, C. García-Martínez, J. Gómez-Barbadillo, and S. Ventura, "Heuristics for interesting class association rule mining a colorectal cancer database," *Inf. Process. Manag.*, vol. 57, no. 3, p. 102207, 2020, doi: 10.1016/j.ipm.2020.102207.

[8] C. A. Chou, Q. Cao, S. J. Weng, and C. H. Tsai, "Mixed-integer optimization approach to learning association rules for unplanned ICU transfer," *Artif. Intell. Med.*, vol. 103, no. March 2019, p. 101806, 2020, doi: 10.1016/j.artmed.2020.101806.

[9] Z. Abidin, M. Mauladi, and I. Weni, "Penerapan Metode Association Rules Dalam Menentukan Pola Penyakit Dan Usia Pasien Berdasarkan Data Rekam," *JUSS (jurnal sains dan Sist. informasi)*, vol. 1, no. 1, pp. 1–4, 2018.

[10] N. Ramadhani and B. Said, "Analisis Pola Asosiasi Dan Sekuensial Data Rekam Dengan Teknik Data Mining Menggunakan Algoritma," *Semin. Nas. Sist. Inf. Indones.*, no. September, 2014.

[11] A. -, F. Marisa, and D. Purnomo, "Penerapan Algoritma Apriori Terhadap Data Penjualan di Toko Gudang BM," *JOINTECS (Journal Inf. Technol. Comput. Sci.*, vol. 1, no. 1, pp. 1–5, 2016, doi: 10.31328/jointecs.v1i1.408.

[12] D. Listriani, A. H. Setyaningrum, and F. Eka, "Penerapan Metode Asosiasi Menggunakan Algoritma Apriori Pada Aplikasi Analisa Pola Belanja Konsumen (Studi Kasus Toko Buku Gramedia Bintaro)," *J. Tek. Inform.*, vol. 9, no. 2, pp. 120–127, 2018, doi: 10.15408/jti.v9i2.5602.

[13] J. R. Rajan, A. C. Chelvan, and J. S. Duela, "Multi-Class Neural Networks to Predict Lung Cancer," *J. Med. Syst.*, vol. 43, no. 7, 2019, doi: 10.1007/s10916-019-1355-9.

[14] P. R. Radhika, R. A. S. Nair, and G. Veena, "A Comparative Study of Lung Cancer Detection using Machine Learning Algorithms," *Proc. 2019 3rd IEEE Int. Conf. Electr. Comput. Commun. Technol. ICECCT 2019*, pp. 1–4, 2019, doi: 10.1109/ICECCT.2019.8869001.

[15] A. W. Oktavia Gama, I. K. Gede Darma Putra, and I. P. Agung Bayupati, "Implementasi Algoritma Apriori Untuk Menemukan Frequent Itemset Dalam Keranjang Belanja," *Maj. Ilm. Teknol. Elektro*, vol. 15, no. 2, pp. 21–26, 2016, doi: 10.24843/mite.1502.04.

[16] E. Muningsih, "Penentuan rekomendasi produk dengan metode data mining asosiasi generalized sequence pattern (gsp)," pp. 218–224, 2016.

[17] P. R. K. Varma, V. V. Kumari, and S. S. Kumar, *Progress in Computing, Analytics and Networking*, vol. 710, no. Dmd. Springer Singapore, 2018.

[18] S. Beniwal and J. Arora, "Classification and Feature Selection Techniques in Data Mining," *Int. J. Eng. Res. Technol.*, vol. 1, no. 6, pp. 1–6, 2012.

[19] V. Gligorijević and N. Pržulj, "Methods for biological data integration: Perspectives and challenges," *J. R. Soc. Interface*, vol. 12, no. 112, 2015, doi: 10.1098/rsif.2015.0571.

[20] A. Izzah and R. Widyastuti, "Prediksi Kelulusan Mata Kuliah Menggunakan Hybrid Fuzzy Inference System," *Regist. J. Ilm. Teknol. Sist. Inf.*, vol. 2, no. 2, p. 60, 2016, doi: 10.26594/r.v2i2.548.