BAB I
PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi saat ini sudah sangat pesat termasuk dalam dunia kesehatan. Dalam dunia kesehatan, teknologi memiliki salah satu perang penting dalam menyimpan informasi-informasi mengenai pasien yang berkunjung yang kemudian disimpan dalam bentuk digital.

Data rekam medis merupakan istilah yang merujuk pada informasi-informasi pasien yang berkunjung secara lengkap pada suatu rumah sakit. Data rekam medis sangat penting untuk mengetahui bagaimana riwayat suatu pasien, dari catatan seperti ini, yang memudahkan dokter dalam mengambil suatu keputusan tepat. Pemanfaatan teknologi dalam rekam medis akan memudahkan dokter dan petugas kesehatan dalam mengakses informasi lebih mudah dan mengambil keputusan yang lebih cepat dalam hal seperti keperluan transaksional dan kadang sebagai pembuatan laporan . Sayangnya, penggunaan sistem menejemen informasi pada rumah sakit yang digunakan hanya terbatas pada itu saja, tumpukan data dari tahun ke tahun kemudian tidak dipergunakan untuk apa apa, yang sebenarnya dapat digunakan untuk menggali informasi atau pola pola yang penting dalam kaitannya dengan suatu penyakit (Ali, 2019).

Data mining sendiri bertujuan untuk menemukan suatu informasi yang berguna dari suatu dataset dengan jumlah yang besar (Alsayat & El-Sayed, 2016). Data mining adalah suatu cara untuk melakukan data analisis secara otomatis yang bertujuan untuk mengungkapkan hubungan antar suatu data yang mungkin tidak dapat dikenali (Tomar & Agarwal, 2013). Metode pada data mining dapat dibagi ke dalam dua kategori : *supervised* dan *unsupervised learning*. Tujuan utama pada *supervised learning* biasanya digunakan untuk mengembangkan suatu model yang kemudian akan digunakan untuk melakukan suatu prediksi berdasarkan dari label kelas sedangkan *unsupervised learning* sendiri tidak menggunakan label kelas; tujuannya sendiri untuk mengenali distribusi dari suatu data untuk kemudian mempelajari data tersebut (Ogbuabor & Ugwoke, 2018).

Klustering sendiri merupakan salah satu teknik yang bertujuan untuk menemukan grup grup data pada suatu data yang tidak dilabeli. Digunakan untuk membagi data dalam kelompok-kelompok yang berbeda dengen tujuan data dalam kelompok yang sama memiliki tingkat kesamaan yang tinggi dan perbedaan yang cukup jelas dengan kelompok lainnya (Han et al., 2012). Melalui klusterisasi, suatu data sangat penting untuk dieksplor dan dievaluasi, dengan tujuan menemukan fitur fitur pada data tanpa mengenali data itu sendiri (DeFreitas & Bernard, 2015).

Pada penelitian ini bertujuan untuk mengimplementasikan salah satu metode pada data mining yaitu clustering pada data kesehatan yang didalamnya terdapat data klinis, demografis, dan sosial-ekonomi pada rumah sakit daerah yaitu Rumah Sakit Umum Lakipadada yang terletak di kabupaten Tana Toraja. Jumlah pasien

yang datang sendiri juga tidak sedikit, dikarenakan merupakan satu satunya rumah sakit daerah di kabupaten Tana Toraja. Berdasarkan Badan Pusat Statistik Kabupaten Tana Toraja pada tahun 2023 jumlah penduduk di Tana Toraja sendiri yaitu 291.047 jiwa (Badan Pusat Statistik Kabupaten Tana Toraja, 2023). Bahkan tidak sedikit juga pasien yang datang dari kabupaten dan kota lain untuk melakukan pengobatan pada rumah sakit ini .

Melalui data mining, data dengan sebegitu lengkapnya dapat memberikan wawasan yang lebih luas lagi dan bahkan memungkinkan untuk menemukan pola-pola tersembunyi yang kemudian dapat digunakan untuk memahami suatu penyakit lebih dalam termasuk seperti manajemen penyakit itu sendiri, memahami faktor faktor risiko yang paling berpengaruh, dan mendukung pencegahan yang lebih efektif.

1.2 Rumusan Masalah

1. Bagaimana hubungan kelompok kelompok pasien dengan karakteristik tertentu dan hubungannya terhadap suatu penyakit.
2. Bagaimana hubungan suatu penyakit dan hubungannya dengan berbagai macam faktor.
3. Bagaimana memilih dan menerapkan algoritma klustering yang tepat sesuai dengan struktur dan jumlah dataset untuk digunakan dalam mengelompokkan data kesehatan.
4. Bagaimana menyajikan informasi hasil klusterisasi untuk kemudian memberikan gambaran yang mudah dipahami mengenai pola suatu penyakit dan keterkaitannya dengan berbagai macam variabel.

1.3 Tujuan Penelitian

1. Mengenali kelompok kelompok pasien dengan karakteristik yang serupa dan hubungannya dengan penyakit yang berkaitan.
2. Menganalisis pola penyakit dan menilai keterkaitannya dengan faktor demografis, faktor geografis, dan faktor klinis pasien.

1.4 Manfaaat Penelitian

Penelitian ini bertujuan untuk memberikan wawasan yang lebih luas dan dalam atau bahkan pola pola yang tersembunyi yang mungkin bisa ditemukan melalui analisis data yang begitu kompleks, harapannya hasil penelitian kemudian dapat digunakan untuk meningkatkan kualitas perawatan pasien dan membantu rumah sakit dalam merencanakan pelayanan kesehatan yang lebih baik.

1.5 Batasan Masalah

a. Pengambilan data merupakan data rekam medis dari Rumah Sakit Umum Daerah Lakipadada Makale
b. Data yang diambil terbatas pada 10 penyakit terbanyak untuk pasien rawat inap pada Rumah Sakit Lakipada Tana Toraja pada tahun 2023 dalam rentang waktu 12 bulan.

1.6 Data Kesehatan

Data kesehatan merupakan penyimpanan tunggal seluruh data status kesehatan konsumen asuhan kesehatan. Ini mencakup catatan kelahiran, catatan imunisasi, laporan seluruh pemeriksaan fisik, dan catatan seluruh penyakit. Catatan ini harus cukup untu mengidentifikasi pasien, menyokong diagnosa, membenarkan pengobatan, dan mendokumentasikan hasilnya. Peranan data saat ini telah jauh melewati tarif pasien secara individu. Data kesehatan bisa membantu rumah sakit mengambil keputusan untuk memperluas asuhan. Informasi mutu asuhan dapat membantu staf medis memonitor dan mengevaluasi dirinya sendiri (Huffman, E.K, 2018)

Berdasarkan Permenkes No. 24 Tahun 2022 mempertegas definisi yang sama bahwa data kesehatan merupakan dokumen yang berisikan data identitas pasien, pemeriksaan, pengobatan, tindakan, dan pelayanan lain yang telah diberikan kepada pasien.

Lebih lanjut lagi menurut Sakharkar (2009) data kesehatan dapat berfungsi sebagai :

1. Dokumen klinis, berisi riwayat kesehatan, pemeriksaan fisik, catatan perawatan , dsb.
2. Dokumen saintifik, dikarenakan data kesehatan juga dapat digunakan untuk mempelajari kondisi pasien dan mengembangkan pengobatan yang lebih baik, dan juga dapat digunakan untuk penelitian
3. Dokumen administrasi, membantu dalam administasi seperti pelayanan, *budgeting*, dan mengembangkan kualits perawatan, pembuatan statistic kesehatan.

1.7 Data Mining

Menurut Han dan Kamber (2006) data mining dapat diartikan sebagai proses ekstraksi atau “menambang” informasi dari data dengan jumlah yang besar. Definisi lain menurut Aggarwal (2015) data mining merupakan ilmu dalam mengumpulkan, membersihkan, memproses, menganalisa, dan mendapatkan informasi bermafaat dari suatu data.

Hasil dari data mining akan menghasilkan informasi yang valid dengan tingkat keyakinan tertentu. Pola yang juga didapatkan dari data mining dapat memberikan pemahaman yang menghasilkan informasi yang berguna dan bermanfaat (Fayyad et al., 1996).

Secara garis besar, proses data mining biasanya mencakup (Han & Kamber, 2006):

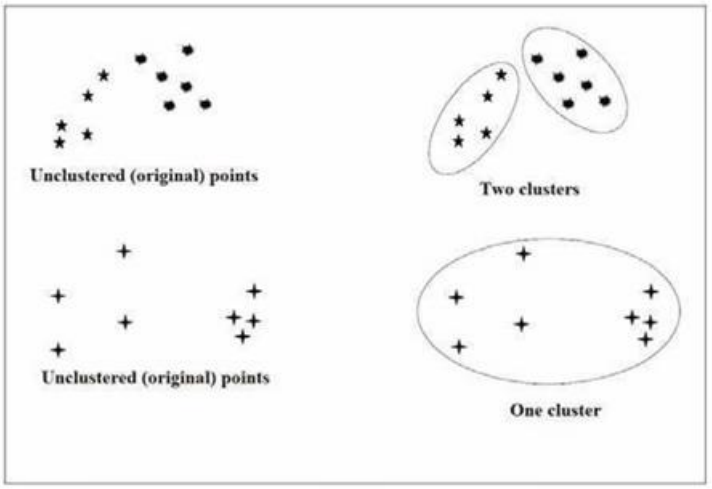
1. *Data cleaning*, untuk membersihkan data yang tidak konsisten dan *noise*
2. *Data integration*, dimana satu atau lebih data dari sumber yang berbeda digabungkan
3. *Data selection*, dimana data-data yang relevan terhadap analisis yang akan dipilih
4. *Data transformation* , dimana data akan ditransformasi atau disesuaikan ke dalam bentuk yang layak atau sesuai untuk dilakukan *data mining*

5. *Data mining*, proses ekstrasi pola pada data menggunakan metode-metode pintar
6. *Pattern evaluation*, untuk mengidentifikasi pola yang menarik mewakili informasi yang betul betul bermanfaat didasarkan pada evaluasi metrik tertentu
7. *Knowledge presentation*, dimana visualisasi dan presentasi informasi dengan teknik yang berbeda beda mengenai hasil ekstrasi informasi yang didapatkan

1.8 Clustering

Menurut Gorunescu (2011) cluster dapat diartikan sebagai mengelompokkan ke dalam grup dari kumpulan data berdasarkan kesamaan dari data-data tersebut, sehingga setiap kelompok memiliki kesamaan pada setiap datanya, namun juga cukup berbeda dengan data pada kelompok lainnya. Pada data yang tidak memiliki label, klastering merupakan model yang cocok untuk menganalisa suata *dataset* yang hasil akhirnya dapat berbentuk rangkuman dari data tersebut atau model generatif (Aggarwal & Reddy, 2014).

Tujuan dari klastering yaitu mengeksplor karakteristik alami dari suatu data yang memiliki sedikit atau tidak ada sama sekali informasi mengenai data tersebut. Selain itu, terkadang proses seperti melabeli suatu data dapat memakan waktu yang cukup banyak, dalam hal ini klastering dapat menjadi pilihan alternatif untuk menghemat waktu (Xu & Wunsch, 2008).



Sumber : (Gorunescu, 2011).

Gambar 1. Contoh pengelompokan klastering

Gambar di atas merupakan salah satu contoh klastering yang paling umum dikenal dengan nama *non-hierarchial clustering*, yaitu proses memisahkan *subset* data dan tidak tumpang tindih, dimana satu data hanya berada pada satu klaster (Gorunescu, 2011).

1.9 K-Prototype *Clustering*

K-Prototype merupakan algortima dari kombinasi algoritma *k-means* dan *k-modes*. Penggunaan *k-means* digunakan untuk menangani data numerikal dan versi modifikasi *k-modes* digunakan untuk menangani data kategorikal dalam suatu dataset dengan jenis tipe data campuran (Satapathy et al., 2014). Diambil dari Huang (1998) berikut cara kerja dari algoritma berikut :

1. Pilih jumlah kluster yang akan terbentuk
2. Pilih centroid untuk setiap kluster, dimana nilai mean untuk setiap fitur numerikal dan mode(nilai yang paling sering muncul) untuk kategorikal.
3. Ambil setiap data lalu lakukan perhitungan ke setiap kluster berdasarkan selisih dan ketidaksamaan yang dihitung sebagai berikut :

$$d_2(\mathrm{X},\mathrm{Y}) = \sum_{j=1}^{p}\left(x_j - y_j\right)^2 + \gamma \sum_{j=p+1}^{m} \delta\left(x_j, y_j\right) \qquad (1)$$

   a) x_j, y_j , merupakan nilai fitur untuk data poin X dan Y pada atribut j dalam hal ini X merupakan data dan Y merupakan centroid
   b) $\sum_{j=1}^{p}\left(x_j - y_j\right)^2$ merupakan *Euclidean distance* untuk fitur numerikal dari atribut ke-1 hingga ke- p(jumlah fitur numerikal) . Nilai ini dihitung dengan mengkuadratkan selisih antara nilai data dan centroid, lalu menjumlahkan seluruh hasilnya
   c) .$\delta\left(x_j, y_j\right)$, merupakan *similarity difference* untuk fitur kategorikal dimana :

$$\delta\left(x_j, y_j\right) = \begin{cases} 0\ jika\ x_j = y_j \\ 1\ jika\ x_j \neq y_j \end{cases}$$

   Jika nilai antara data kategori dan centroid sama akan bernilai 0 dan jika berbeda akan bernilai 1
   d) $\gamma \sum_{j=p+1}^{m} \delta\left(x_j, y_j\right)$ adalah total ketidaksamaan untuk fitur kategorikal dari atribut ke- p+1 hingga ke-m . Setia ketidaksamaan akan bernilai 1 . Dengan γ memberikan pengaruh nilai terhadap jarak total.
4. Kelompokkan data pada klaster dengan nilai selisih jarak total paling rendah .

Perhitungan total *cost function* dari Huang (1998) pada K-Prototype digunakan perhitungan sebagai berikut :

$$P(\mathrm{X},\mathrm{Y}) = \sum_{l=1}^{k}(P_l^r + P_l^c) \qquad (2)$$

Untuk $\boldsymbol{P_l^r}$ merupakan *cost* untuk numerikal dimana :

$$\sum_{j=1}^{p}\left(x_{i,j} - q_{l,j}\right)^2 \qquad (3)$$

   a. Hitung nilai *Euclidean distance* ke klaster numerikal centroid $q_{l,j}$ untuk setiap fitur j
   b. Jumlahkan untuk setiap perhitungan data i dalam cluster l

$$P_l^r = \sum_{i=1}^{n} w_{i,l} \sum_{j=1}^{p} \left(x_{i,j} - q_{i,j}\right)^2 \quad (4)$$

Dimana $w_{i,l}$= 1 jika data i berada pada klaster l , dan sebaliknya 0 jika tidak

Untuk $\boldsymbol{P_l^c}$ merupakan *cost* untuk kategorikal dimana =

$$\delta(x_{i,j}, q_{l,j}) = \begin{cases} 0\ jika\ x_{i,j} = q_{l,j} \\ 1\ jika\ x_{i,j} \neq q_{l,j} \end{cases}$$

a. Untuk setiap data i pada cluster l untuk setiap fitur kategori dimana jika data $x_{i,j} = q_{l,j}$ maka beri nilai 0 dan jika $x_{i,j} \neq q_{l,j}$ beri nilai 1
b. Jumlahkan untuk setiap data i pada cluster l dan kalikan dengan γ

$$P_l^r = \gamma \sum_{i=1}^{n} w_{i,l} \sum_{j=p+1}^{m} \left(x_{i,j}, q_{i,j}\right) \quad (5)$$

1.10 UMAP: Uniform Manifold Approximation and Projection

UMAP merupakan salah satu teknik dalam mengurangi dimensi namun dengan tujuan tetap mempertahankan struktur global dan lokal dari data saat ditaruh pada dimensi yang lebih rendah(de Zarzà et al., 2023).

Penempatan titik dengan metode ini akan mempertimbangkan semua fitur baik itu dari fitur kategori maupun fitur numerik. Hal ini dicapai dengan cara kerja dari UMAP disendiri. Pada dasarnya UMAP akan membangun suatu *graph network relationship* dengan menghitung tetangga-tetangga dari setiap titik dengan jarak terdekat sesuai dengan jumlah tetangga dengan jarak yang dekat dihitung dari penentuan n_neighbours paramater. Tetangga-tetangga berjumlah *n_neighbours* dengan jarak yang paling dekat kemudian akan diambil dan dibuatkan *fuzzy simplical set*. Kemudian setiap tetangga untuk titik yang dihitung jarak ke tetangganya akan diberi *weight* dengan range 0-1. Dimana bobot semakin mendekati 1 artinya jaraknya semakin jauh dan bobot mendekati 0 artinya jaraknya semakin dekat. Perhitungan di atas akan dihitung dengan rumus sebagai berikut (McInnes et al., 2018):

$$weight\left(x_i, x_{i_j}\right) = exp\left(\frac{-max\left(0, d\left(x_i, x_{i_j}\right) - p_i\right)}{\sigma_i}\right) \quad (6)$$

Dimana :

$d\left(x_i, x_{i_j}\right)$: jarak antara point x_i dan tetangganya x_{i_j}

p_i : jarak ke tetangga yang paling dekat dari x_i

σ_i : faktor skala untuk x_i yang disesuikan dengan kepadatan tetangga di sekitar titik tersebut

Hasil di atas akan dilakukan proses simetrasasi dikarenakan untuk setiap x_i akan memiliki graf lokal tersendiri dengan tetangganya. Graf ini memiliki satu arah dari titik x_i ke tetangga yang menggambarkan hubungan x_i ke tetangganya x_{i_j}. Maka untuk membentuk graf secara global diperlukan proses simetrisasi dengan persamaan dibawah ini :

$$\mu_{ij} = w_{ij} + w_{ji} - w_{ij} \, . \, w_{ji} \qquad (7)$$

Proses simetrisasi ini akan menghasilkan graf yang mencerminkan gabungan dua arah dari titik $x_i \rightarrow x_{i_j}$ dan $x_{i_j} \rightarrow x_i$

Perhitungan di atas kemudian akan digunakan untuk membangun *fuzzy simplical set* yang kemudian nantinya akan menggambarkan bagaimana hubungan antara satu titik dengan titik lainnya sebagai tetangganya pada dimensi yang lebih tinggi. Hasil dari *fuzzy simplicat set* ini kemudian akan digunakan untuk menempatikan data-data pada dimensi yang lebih rendah dengan mempertahankan hubungan pada *fuzzy simplicat set* sebaik mungkin.

1.11 Asosiasi

Asosiasi atau juga dikenal dengan nama analisis afinitas adalah studi atau analisis terhadap atribut atau karakteristik yang cenderung muncul atau terjadi bersamaan. Tujuan dari analisis ini yaitu untuk menemukan asosiasi atau hubungan antara dua atau lebih atribut atau karakteristik. Aturan asoasi yang terbentuk akan mengambil bentuk “jika antaseden, maka konsekuen” yang kemudian diukur dengan metrik seperti *support* dan *confidence* untuk menentukan kekuatan aturannya (Larose, 2005).

Aturan asosiasi memiliki dua bagian yang dituliskan dalam X → Y yang artinya jika terdapat X maka Y juga cenderung muncul. X dan Y dalam hal ini dapat menjadi item tunggal atau beberapa item yang dikenal dengan nama *itemset*. Disini X dikenal dengan sebutan antaseden dan Y dikenal dengan sebutan konsekuen (Bhatia, 2019).

Itemset merupakan item-item yang terkandung pada dalam satu transaksi, *n-itemset* dapat dikatakan sebagai *itemset* yang mengandung n-*item*. *Frequent itemset* dapat dikatakan sebagai *itemset* yang terjadi setidaknya dalam batas jumlah tertentu (Larose, 2005).

Dua pengukuran *support* dan *confidence* dapat dikatakan sebagai pengukuran mengenai seberapa menarik atau pasti aturan tersebut. Pada dasarnya suatu aturan asosiasi dapat dikatakan menarik dan pasti, jika memenuhi ambang batas *support* yang ditentukan dan ambang batas *confidence* yang ditentukan. Dimana ambang batas ini

ditentukan oleh yang melakukan analisis untuk menemukan pola yang menarik (Han et al., 2011).

Mengacu pada Bhatia (2019) beberapa parameter yang digunakan dalam mengukur seberapa pasti dan menarik suatu aturan diuraikan di bawah berikut :

Support dihitung sebagai berikut :

$$Support(X) = \frac{Jumlah\ transaksi\ yang\ mengandung\ X}{Total\ jumlah\ transaksi} \qquad (8)$$

Secara konvensional aturan pada *supporti* biasanya dituliskan dari antara 0% dan 100%.

Nilai *confidence* sendiri merupakan nilai yang terbentuk dari hasil pembentukan aturan asosiasi yang didapatkan dari *frequent itemset* yang memenuhi *threshold*. Nilai *confidence* pada dasarnya menjelaskan bagaimana kemungkinan kemungkinan bagian konsekuen (B) dari suatu aturan akan muncul ketika bagian antaseden (A) muncul. Setiap *frequent itemset* akan dipisahkan ke dalam dua bagian yaitu **antaseden (A)** dan **konsekuen (B)**, lalu diuji untuk membentuk suatu aturan **A** → **B**.

Lebih lanjut lagi secara matematis nilai *confidence* didefinisikan sebagai berikut :

$$Confidence(A \rightarrow B) = \frac{Support(A \cap B)}{Support(A)} \qquad (9)$$

Lebih lanjut pada rumus di atas maka dapat disimpulkan *confidence* diartikan sebagai saat terdapat item A maka seberapa besar kemungkinannya B juga muncul. Semakin tinggi nilai *confidence* maka dapat diartikan bahwa aturan yang tersebut cukup relevan dan akurat.

Meskipun demikian, perlu diperhatikan bahwa suatu aturan dengan *confidence* yang tinggi belum tentu menjamin aturan tersebut benar-benar akurat, salah satu contoh bisa saja kemunculan B memang sering muncul tanpa tergantung pada A, untuk mengatasi hal ini maka dibutuhkan parameter lainnya yang dikenal dengan nama *lift*.s

Nilai *lift* sederhananya dapat menjadi pendamping untuk nilai *confidence* yang pada dasarnya secara konsep dapat memastikan bagaimana hubungan kemunculan B hanya saat terdapat kemunculan A, dibandingkan dengan kemungkinan muncul B terlepas dari ada tidaknya kemunculan A. Dalam artian lain, *lift* mengukur seberapa besar pengaruh kemunculan A terhadap kemunculan B.

Secara matematis lift dapat dirumuskan sebagai berikut :

$$Lift(A \rightarrow B) = \frac{Support(A \cap B)}{Support(A)\ \times\ Support(B)} \qquad (10)$$

Menurut Larose (2005) secara umum dalam melakukan *mining* aturan asosiasi melalui proses berikut :

1. Temtukan *frequent itemset* , *itemset* yang memenuhi *support*.
2. Dari semua *frequent itemset*, buatkan aturan asosiasi dengan kondisi minimun *support* dan *confidence*.

1.12 Apriori

Algoritma apriori meruapakan salah satu algoritma paling umum dan paling klasik yang digunakan dalam mencari *frequent item set*. Algoritma ini akan menganggap semua dataset dalam bentuk kategorikal. Algoritma ini menggunakan *bread-first search* dimana *k-itemset* kemudian akan digunakan untuk mencari (k+1) *itemset* (Bhargava & Selwal, 2013).

Cara kerja dari algoritma ini pada mencari kombinasi itemset yang memenuhi nilai *support*, itemset set merupakan kombinasi dari item-item antar fitur yang digabungkan lalu dihitung seberapa sering kemunculannya atau proporsi kombinasi item ini dibanding keseluhan data. Untuk cara kerja lebih jelasnya sebagai berikut :

1. Penentuan ambang batas *support* untuk menentukan seberapa sering suatu kombinasi item muncul untuk dianggap relevan.
2. Penentuan itemset tunggal yang memenuhi nilai *support* pada setiap fitur.
3. Proses kemudian akan dilanjutkan untuk kandidat *k* itemset dengan proses sebagai berikut :
    - Ambil dua (k−1) itemset yang memenuhi nilai *support*
    - Gabungkan itemset jika kedua itemset memiliki nilai item yang sama untuk (k−2) item pertama
    - Kombinasikan dengan (k−2) item pertama yang sama dan menggabungkan item yang tidak sama antar itemset
    - Periksa (k−1) subset dari k kandidat itemset yang juga memenuhi *support*. Jika satu saja subset item tidak memenuhi maka buang.
4. Proses akan terus berulang hingga tidak ada kemungkinan kandidat k itemset yang lebih tinggi dapat terbentuk dari kandidat itemset sebelumnya dengan nilai *support* yang ditetapkan
5. Hasil dari setiap kombinasi itemset yang terbentuk dan memenuhi nilai *support* dinamakan *frequent itemset*.

BAB II
METODE PENELITIAN

2.1Tahapan Penelitian

Penelitian ini akan mencakup beberapa proses yang digambarkan pada gambar 2 sebagai berikut :

Gambar 2. Tahapan penelitian

a. Studi Literatur

Tahap ini bertujuan untuk mengumpulkan berbagai macam informasi yang terkait dengan penelitian yang akan dilakukan baik itu dari jurnal maupun buku. Hal ini dilakukan untuk mengetahui metode-metode yang sering dilakukan dalam menyelesaikan masalah terkait dengan penelitian serta menemukan kekurangan yang dapat diisi dari penelitian-penelitian sebelumnya.

b. Pengumpulan Data

Tahap ini dilakukan pengumpulan data pada Rumah Sakit Daerah Lakipadada Tana Toraja. Data yang diambil terbatas hanya pada tahun 2023 untuk 10 penyakit terbanyak. Hal ini dilakukan agar data yang diambil juga tidak mengandung noise dan masih layak untuk dilakukan *data mining*. Selain itu data yang diambil juga merupakan data rawat inap dengan pertimbangan proses analisa dapat menyeluruh menggunakan data rawat inap.

c. Pemilihan Metode dan Algoritma Data Mining

Kemudian dilakukan berbagai percobaan menggunakan beberapa algoritma untuk menemukan yang paling cocok dengan dengan struktur dan karakteristik data yang didapatkan. Percobaan kemudian dievaluasi menggunakan beberapa metrik evaluasi dan informasi yang didapatkan dari pecobaan. Hasil yang paling baik kemudian akan ditetapkan sebagai metode algoritma yang akan digunakan nantinya untuk proses data mining. Dalam hal ini penelitian ditetapkan akan menggunakan metode *clustering* dan *asosiasi* dalam proses *data mining*. Proses *clustering* sendiri akan

menggunakan algoritma *K-Prototype* yang kemudian nantninya hasil dari *clustering* akan dianalisa lebih lanjut lagi menggunakan metode asosiasi dalam hal ini menggunakan algoritma *Apriori.*

d. Preprocessing Data

Tahap ini akan melibatkan berbagai proses mulai dari pembersihan data, pemilihan fitur-fitur yang relevan, dan transformasi data. Pada tahap pembersihan data, beberapa hal yang paling umum seperti menghapus data duplikat, menghapus data dengan nilai-nilai ekstrim . Pada pemilihan fitur-fitur yang relevan akan dilakukan pemlihan fitur yang hanya relevan pada penelitian dengan menghapus fitur-fitur yang tidak dibutuhkan. Dan terakhir sebelum melalui proses *data mining* , beberapa fitur juga akan melalui transformasi yang tujuannya agar data dapat diproses melaluia algoritma yang digunakan.

e. Implementasi Metode

Metode final yang sudah dipilih sebelumnya kemudian akan digunakan untuk memproses data yang sudah final. Dalam hal ini akan dilakukan *clustering* menggunakan algoritma *K*-Prototype pada data dan nantinya hasil dari setiap klaster akan dianalisa lebih lanjut menggunakan asosiasi menggunakan algoritma *Apriori*.

f. Evaluasi Hasil

Hasil dari implementasi algoritma kemudian akan dianalisa menggunakan evaluasi metrik yang berbeda-beda tergantung metode yang digunakan untuk menentukan hasil yang paling optimal berdasarkan nilai metrik dan keunikan dari informasi yang didapatkan. Untuk *clustering* sendiri akan digunakan uji coba menggunakan jumlah klaster yang berbeda-beda dengan evaluasi metrik SSE sebagai pemilihan final parameter yang akan digunakan. Lebih lanjut lagi setelah melalui proses *clustering*, akan diaplikasikan algoritma *Apriori* dengan menggunakan evaluasi metrik *support* dan *confidence* untuk memastikan informasi yang didapatkan memiliki kepastian yang cukup kuat.

g. Penyusunan Laporan

Keseluruhan proses akan dibuatkan dalam bentuk laporan yang mencakup teori-teori yang yang digunakan dalam proses implementasi, proses preprocessing data , uji coba dengan berbagai parameter dan pemilihan parameter yang paling optimal hingga hasil final dari proses *data mining* yang berupa temuan-temuan pola yang ditemukan pada setiap *cluster* serta hubungan antar variabel pada setiap *cluster* melalui implementasi asosiasi analisis.

2.2 Waktu dan Lokasi Penelitian

Penelitian dilakukan setelah persetujuan pengajuan judul pada tanggal “sekian” hingga tangga “sekian” pada Rumah Sakit Daerah Lakipadada Tana Toraja. Penelitian berlangsung selama kurang lebih 3 bulan yang mencakup perizinan ke pihak instansi terkait dan pengambilan data yang dilakukan oleh pihak berwenang pada rumah sakit yang dituju. Prosesnya semua dilakukan di daerah Tana Toraja. Selanjutnya hasil dari pengumpulan data kemudian melalui proses penelitian di Laboratorium *Artificial Intelligence (AI)*, Departemen Teknik Informatika, Fakultas Teknik, Universitas Hasanuddin.

2.3 Instrumen Penelitian

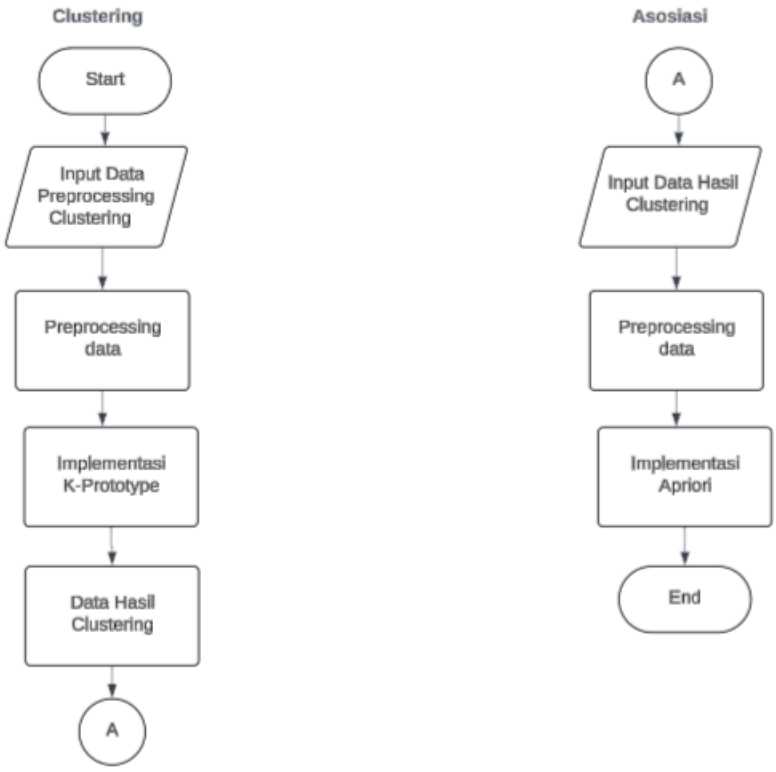
Penelitian dilakukan dengan menggunakan beberapa instrument, untuk menyimpan hasil pengumpulan data dan pemprosesan data disimpan dalam *Microsoft Excel*. Proses implementasi *data* mining menggunakan bahasa *Pyhton* yang pengembangannya melalui *software Jupyter Notebook* untuk kemudahan eksekusi dan presentasi hasil.

2.4Teknik Pengambilan Data

Data yang diambil merupakan data pasien tahun 2023 dan terbatas pada 10 penyakit terbanyak paling atas didasarkan pada data yang diambil merupakan data yang yang masih relevan dengan tren yang terjadi sekarang, data diambil khusus untuk pasien rawat inap dikarenakan proses perawatan dan informasi yang disediakan pada setiap pasien sudah selesai dan lengkap. Hal ini berfungsi untuk memastikan data yang akan digunakan merupakan data yang konsisten dan memungkinkan proses analisa pada berbagai macam variabel . Hasil data yang diambil mencakup 10 penyakit paling umum terjadi di tahun 2023 dengan jumlah data 5239 data mentah yang belum melalui proses *preprocessing*. Adapun fitur-fitur yang terdapat pada data pada setiap pasien yaitu **pekerjaan, status perkawinan, kecamatan, kabupaten, cara bayar, keluhan, diagnosa, status keluar, umur keluar, tanggal keluar, jenis kelamin, tanggal masuk.**

2.5 Perancangan Sistem

Proses pengolahan akan dibagi dalam dua sistem yang masing-masing menunjukkan tahapan dari persiapan data hingga implementasi metode untuk masing masing *clustering* dan asosiasi pada gambar 3.



Gambar 3. Flowchart rancangan sistem

2.5.1 Preprocessing *Clustering*

Tahap ini merupakan tahap awal yang berfungsi untuk mempersiapkan data agar proses pengolahan informasi yang nantinya akan digali dari data dapat tetap relevan pada penelitian dan juga format pada data dapat sesuai dengan kebutuhan algoritma untuk memproses data.

a. *Data Cleaning* dan *Fitur Selection*
   Beberapa proses penting yang dilalui pada proses ini ada sebagai berikut :
   1. Memeriksa nilai null pada setiap baris, untuk baris yang memiliki nilai *null* akan dilakukan penghapusan data.
   2. Menghapus nilai-nilai ekstrim pada data-data numerik dalam hal ini data dengan umur pasien atau lama rawat yang tidak wajar.
   3. Menghapus baris dengan diagnosa yang tidak relevan meskipun masuk dalam dataset kesehatan beberapa diagnosa tersebut diantaranya : "*Fetus and newborn affected by caesarean delivery*", *"Singleton, born in hospital", "Single spontaneous delivery, unspecified"*. Diagnosa diatas merupakan peristiwa kelahiran yang juga tercatat dalam data kesehatan namun tidak dibutuhkan.

4. Melakukan pemilihan fitur-fitur yang hanya benar benar diperlukan, pada tahap ini dihasilkan beberapa fitur yang akan masuk pada pemprosesan *data mining* nantinya yaitu:

- Umur(tahun)
- Lama Rawat
- Jenis Kelamin
- Pekerjaan
- Status Perkawinan
- Kecamatan
- Keluhan
- Diagnosa
- Status Keluar

b. *Data Transformation*

Setelah melakukan data *cleaning*, pemprosesan data akan dilanjutkan dengan melakukan transformasi pada beberapa bagian data yang diperlukan pada proses ini meliputi beberapa bagian seperti proses transformasi teks pada keluhan dengan melakukan standarisasi keluhan dan kemudian dilanjutkan *one-hot encoding* nantinya dan normalisasi pada data numerikal.

1. Standarisasi Keluhan dan *One-Hot Encoding*

Pada tahap ini akan dilakukan proses standarisasi untuk keluhan dikarenakan nilai pada setiap barisnya tidak konsisten, bervariasi, dan mengandung *typo* meskipun mendeskripsikan gejala yang sama.

id	keluhan
81546	MRS DGN KEL NYERI ULUHATI MUAL MUNTASH DEMAM
81569	SESAK
81577	MRS DENGAN KELUHAN BATUK DISERTAI SESAK
81597	MRS DENGAN KELUHAN BAB ENCER, SAKIT PERUT,SAKIT KEPALA
81612	MRS DGN KELUHAN LEMAS,MUNTAH, NYERI ULU HATI, RIW. DM
81624	DEMAM 2 HARI, MENGGIGIL, TEGANG PADA LEHER
81629	MRS DENGAN KEL DEMAM, SAKIT KEPALA, MUAL, SAKIT PERUT
81635	demam,sakit kepala
81653	MRS DENGAN KELUHAN LEMAS, SAKIT KEPALA,KURANG NAFSU MAKAN, NYERI PERUT
81656	MRS DENGAN KELUHAN MUNTAH, SAKIT KEPALA, SUSAH MAKAN

Gambar 4. Contoh dataset sebelum standarisasi keluhan

Proses standarisasi keluhan meliputi proses sebagai berikut :

1. Menghapus kata-kata *filler* pada setiap keluhan diantaranya *“mrs”, “dgn”, “dengan”, “keluhan”, “kel”, “dgn keluhan”, “dengan keluhan”, “dgn kel”*

2. Menciptakan kamus standarisasi untuk gejala-gejala umum yang paling sering ditemukan pada dataset. Hasil standarisasi ini menghasilkan sejumlah gejala umum, antara lain: **“bab tidak normal”, “batuk”, “bengkak”, “demam”, “flu”, “kejang”, “lemas”, “menggigil”, “mual”, “muntah”, “nafsu makan menurun”, “nyeri menelan”, “nyeri perut”, “nyeri seluruh badan”, “nyeri ulu hati”, “nyeri/kram anggota gerak”, “perut kembung”, “sakit kepala”, “sesak”, dan “lainnya”.**
3. Kolom gejala pada setiap barisnya kemudian akan disimpan dalam bentuk *list* dimana setiap gejala akan diambil berdasarkan pemisah dengan koma . Selanjutnya, setiap gejala yang mengandung kata yang terdapat dalam kamus akan diubah menjadi bentuk standar sesuai dengan kata yang berkoresponden dalam kamus.
4. Setelah langkah ini, dikarenakan terdapat beberapa gejala yang sebenarnya merupakan variasi dari gejala umum namun tidak terkonversi, dibuatkan pemetaan tambahan berupa list variasi yang mewakili gejala umum pada kamus. Contohnya, berbagai variasi seperti “nyeri seluruh tubuh”, “sakit seluruh badan”, dan “nyeri badan seluruh” akan dipetakan secara manual menjadi “nyeri seluruh badan”. Proses ini dilakukan dengan mencocokkan setiap gejala dalam *list* terhadap daftar variasi, dan menggantinya jika ditemukan kecocokan. Proses ini dilakuan secara iteratif dengan terus melakukan pembaharuan pada daftar variasi hingga proses standarisasi keluhan benar-benar selesai.
5. Untuk gejala yang masih tidak dapat dipetakan ke dalam gejala umum pada kamus atau merupakan gejala-gejala yang ekstrim akan dipetakan pada gejala “lainnnya”.

id	keluhan_transformed
81546	nyeri ulu hati, demam, mual, muntah
81569	sesak
81577	batuk
81597	bab tidak normal, nyeri perut, sakit kepala
81612	lemas, muntah, nyeri ulu hati, lainnya
81624	demam, menggigil, nyeri/kram anggota gerak
81629	demam, sakit kepala, mual, nyeri perut
81635	demam, sakit kepala
81653	lemas, sakit kepala, nafsu makan menurun, nyeri perut
81656	muntah, sakit kepala, nafsu makan menurun

Gambar 5. Contoh dataset setelah standarisasi keluhan

Setelah proses standarisasi keluhan dilakukan, tahap selanjutnya adalah menerapkan teknik one-hot encoding terhadap data gejala. Setiap gejala unik yang terdapat dalam kamus gejala umum akan direpresentasikan sebagai

satu kolom. Kemudian, untuk setiap baris data, nilai 1 akan diberikan pada kolom gejala yang muncul dalam baris tersebut, dan nilai 0 untuk gejala yang tidak muncul.

	keluhan_bab tidak normal	keluhan_batuk	keluhan_bengkak	keluhan_demam	keluhan_flu	keluhan_kejang	keluhan_lainnya	keluhan_lemas	keluhan_menggigil	keluhan_mual	...
0	0	0	0	1	0	0	0	0	0	1	...
1	0	0	0	0	0	0	0	0	0	0	...
2	0	1	0	0	0	0	0	0	0	0	...
3	1	0	0	0	0	0	0	0	0	0	...
4	0	0	0	0	0	0	1	1	0	0	...

Gambar 6. Contoh dataset setelah *one-hot encoding* keluhan

2. Proses *Scaling* untuk Fitur Numerik

Pada tahap ini dilakukan proses scaling yang bertujuan untuk menormalisasi nilai-nilai numerikal dengan tujuan agar bobot pada setiap fitur numerikal tidak berat satu sama lain. Hal ini dilakukan dengan membuat nilai-nilai pada setiap fitur-fitur numerial yang berada pada range yang tidak terlalu besar atau terlalu kecil. Selain itu proses scaling juga bertujuan tidak hanya antar fitur-fitur numerikal namun agar nilainya nantinya tidak terlalu besar yang kemudian menjadi berat sebelah memberi pengaruh yang lebih besar pada hasil kluster dengan nilai-nilai fitur numerikal yang berada pada range yang lebih besar seperti umur. Maka untuk mencapai tujuan ini, metode *scaling* yang paling tepat yaitu menggunakan ***Z-Score Normalization***.

Secara matematis, **Z-Score Normalization** menggunakan persamaan sebagai berikut :

$$z = \frac{x - \mu}{\sigma} \qquad (10)$$

Dimana :

z = merupakan nilai normalisasi

x = merupakan nilai data yang dinormalisasi

μ = merupakan nilai mean dari fitur

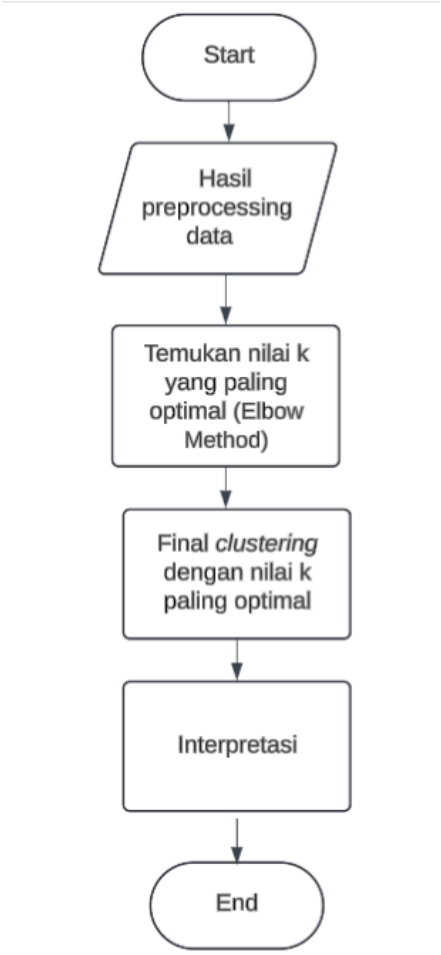
σ = merupakan nilai standar deviasi dari fitur

Proses normalisasi ini pada dasarnya akan membuat setiap fitur numerikal memiliki nilai mean = 0 dan standar deviasi = 1, dengan demikian setiap nilai fitur numerikal akan menjadi nilai yang lebih adil satu sama lain.

Tabel 1. Hasil scaling fitur numerikal

Sebelum Normalisasi		Sesudah Normalisasi	
Umur(tahun)	Lama Rawat	Umur(tahun)	Lama Rawat
46	6	0.649474	1.164922
81	5	1.996437	0.680245
81	5	1.996437	0.680245
60	5	1.188259	0.680245
56	4	1.034320	0.195568

2.5.2 Implementasi *Clustering*



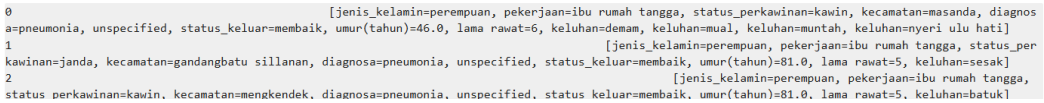
Gambar 7. Flowchart tahapan implementasi *clustering*

Pada tahap ini juga melibatkan beberapa proses penyesuain format dan struktur data sebelum algoritma *clustering* yaitu **K-Prototype** diterapkan. Sebelum menerapkan algoritma ini data yang sudah melalui proses *pre-processing* akan dibagi ke dalam dua kelompok berdasarkan tipe data numerik dan kategorikal , hal ini bertujuan agar algoritma dapat mengenal data numerik yang akan diolah sesuai dengan pengukuran jarak fitur numerik dan data kategori yang akan diolah sesuai dengan pengukuran jarak fitur kategorikal . Kemudian untuk penerapannya akan membutuhkan parameter dan pada implementasinya akan digunakan *elbow method* untuk menentukan jumlah klaster yang akan digunakan dan beberapa parameter lainnya seperti metode inisialisasi centroid pada awal inisialisasi dalam hal ini akan digunakan metode *'Cao'*

2.5.3 Preprocessing Data Asosiasi

Pada tahap ini proses pengolahan data hampir sama dengan pengolahan data pada proses *clustering*. Data yang diolah merupakan data yang telah dilabeli dengan kelompok *cluster* pada setiap data. Proses pengolahan dan persiapan data sebelum implementasi tidak terlalu banyak, mengingat proses pengolahan data sudah dilalui sebelum menerapkan *clustering* pada bagian 2.5.1. Proses persiapan hanya melibatkan persiapan data dengan struktur dan format yang dibutuhkan algoritma *Apriori* untuk memproses data.

Dalam hal ini setiap baris akan dibuatkan kolom baru dengan nama "transaksi" yang mana pada kolom ini mengandung "fitur=nilai" pada setiap fitur untuk baris tersebut termasuk fitur numerik, kategori, dan gejala.

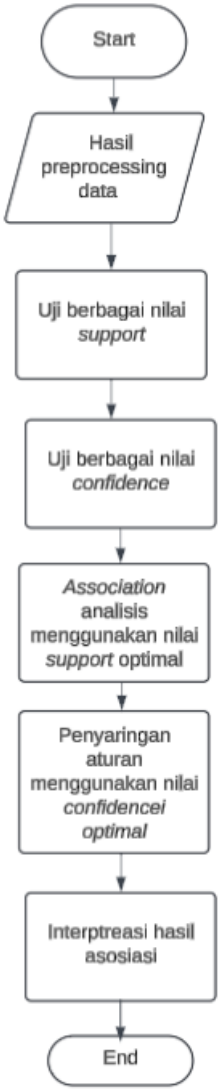


Gambar 8. Contoh *list* pada kolom "transaksi"

Selanjutnya kolom yang berisi *list* transaksi ini akan digunakan untuk membuat *dictionary* yang dikelompokkan berdasarkan label klaster masing-masing baris. Dalam hal ini, transaksi-transaksi tersebut dikelompokkan ke dalam dictionary di mana :

- ***Key*** adalah nilai dari kolom cluster (misalnya: 0, 1, 2, dll)
- ***Value*** adalah kumpulan *list* dari kolom transaksi dari setiap baris dalam klaster tersebut, sehingga membentuk struktur list of list untuk setiap klaster.

2.5.4 Implementasi Asosiasi



Gambar 9. Flowchart tahapan implementasi *association*

Hasil pemprosesan data kemudian akan diterapkan dalam proses asosiasi untuk setiap klaster. Sebelum menerapkan algoritma diperlukan nilai *support* yang optimal dalam penerapan algoritmanya, dan untuk itu dilakukan beberapa percobaan menggunakan nilai *support* yang berbeda beda yang nantinya akan dievaluasi dengana beberapa parameter metrik seperti rata rata jumlah aturan yang tertbentuk, rata rata *confidence*, rata rata *lift*, rata rata panjang aturan. Selain itu proses pembentukan aturan asosiasinya juga perlu disaring nantinya untuk mencari informasi yang relevan namun juga memiliki nilai metrik yang cukup baik untuk memastikan aturan yang dipresentasikan merupakan aturan yang akurat dan pasti, untuk itu akan dilakukan percobaan menggunakan nilai *confidence* yang berbeda-beda yang juga nantinya bersamaan akan

dilihat rata-rata nilai *lift* pada setiap nilai *confidence* yang diuji coba. Nilai *confidence* yang akan dipilih kemudian akan akan dihitung rata-rata nilai *lift*, dan nilai *lift* tersebut akan digunaan untuk menyaring lebih lanjut lagi hasil dari pembentukan aturan asosiasi. Lebih jauh lagi, hasil pembentukan aturan asosiasi yang didapatkan nantinya yang akan dilaporkan hanya akan dipilih beberapa tidak hanya berdasarkan nilai metrik *confidence* dan *lift* yang baik namun juga berdasarkan keunikan dari aturan yang terbentuk dan kompleksitas dari aturan tersebut dalam mengungkapkan hubungan antar variabel pada setiap masing-masing klaster.