BAB II

KAJIAN PUSTAKA

2.1 Teori Dasar

Penelitian ini memiliki teori dasar yaitu mencakup KDD (*Knowledge Discovery In Database*), *Data mining*, *Algoritma K-Means Clustering, Software* pendukung, penelitian terdahulu dan kerangka pemikiran.

2.2 *Knowledge Discovery In Database* (KDD)

Data Mining sering disebut juga sebagai *Knowledge Discovery In Database* (KDD) menemukan pengetahuan baru yang didapatkan dari hasil pengolahan data. Menurut (Handoko, 2016) Konsep *Data Mining* pada saat ini adalah seperti mengumpulkan, menggunakan sebuah data besar untuk menemukan ketergantungan atau pola data yang berhubungan satu sama lain. Menurut (Sari & Harman, 2020) KDD *(knowledge discovery in database)* merupakan sebuah tahapan yang tidak mudah dalam pengidentifikasian pola pada sebuah data dimana pola tersebut juga bersifat baru, juga dapat berguna. *Knowledge Discovery in Database* (KDD) adalah proses penemuan sebuah informasi baru yang berguna dalam sebuah set *database* yang terdiri dari pemahaman di bidang aplikasi, kemudian membuat data target dalam *database*, *cleaning* data dan *preprocessing* data (Fiandra et al., 2017).

Menurut (Sinaga & Handoko, 2021) tahap-tahap dalam *Knowledge Discovery in Database* (KDD) terdiri dari:

1. *Cleaning* Data

   Tahap yang dilakukan adalah membersihan data berupa pengurangan data yang tidak tetap.

2. Data *Integration*

   Tahap ini dilakukan untuk menggabungan data secara keseluruhan dari sumber-sumber data.

3. Data *Selection*

   Pada tahap ini, yang dilakukan dalam penyeleksian data adalah pemilihan data dari *database* yang sesuai dengan tujuan peneliti.

4. Data *Transformation*

   Tahapan ini melakukan perubahan data sesuai dengan metode atau teknik data mining yang diterapkan.

5. *Data Mining*

   Tahap ini, dilakukan implementasi terhadap metode *data mining* yang sesuai agar mendapatkan suatu pola dari data.

6. *Patten Evaluation*

   Tahap ini dilakukan untuk mengidentifikasi data dan pola data.

7. *Knowledge Presentation*

   Tahap ini, *data mining* menghasilkan informasi yang dapat dipresentasikan lalu dapat dijadikan informasi bagi tempat objek penelitian.

2.3 *Data Mining*

Menurut (Santoso, 2017) *Data Mining* merupakan sebuah metode pengolahan data guna menemukan pola baru yang terdapat pada data tersebut. Pemanfaatan *Data Mining* memang berguna sebagai bahan untuk menambahkan informasi dalam berbagai kalangan mulai dari bisnis hingga medis, ini dibuktikan juga setelah mengkaji kembali pengertian *data mining* menurut para ahli. Menurut (Kurnia et al., 2020) *Data Mining* adalah gabungan sejumlah disiplin ilmu komputer.

Selain itu, beberapa bidang dalam kehidupan sehari-hari yang mengaplikasikan *data mining* diantaranya adalah (Kurnia et al., 2020):

1. Marketing dan Bisnis perusahaan memiliki data yang berguna dalam strategi marketing dan bisnisnya. Seperti strategi dalam pemasaran produk agar menghasilkan diagram penjualan yang tinggi, pemilihan vendor yang tepat. Berikut contoh aplikasi *data mining* dalam marketing dan bisnis:
    a) *Market Basket Analysis* atau analisis keranjang belanja dimana konsumen akan ditampilkan pada jenis belanjaan yang biasa dikonsumsinya. MBA juga dikenal dengan *association rule* (aturan asosiasi) yaitu salah satu konsep dalam *data mining* yang berusaha menemukan asosiasi atau keterkaitan data.
    b) *Recommender System*, adalah sistem yang merekomendasikan beberapa variabel dengan tingkatan tertinggi sehingga dapat memilih dengan lebih tepat seperti dalam pemilihan rekomendasi

supplier mana yang menunjukkan performansi baik. Teknik yang digunakan aplikasi ini adalah teknik kalsterisasi ataupun klasifikasi.

c) *Churn Prediction* merupakan analisis dari loyal atau tidaknya suatu pelanggan berdasarkan variabel-variabelnya. Sebagai Contoh perusahaan telekomunikasi yang memiliki pelanggan hampir ratusan juta ingin melihat pelanggan apakah tetap loyal atau tidak dengan menggunakan teknik *data mining* sehingga hal tersebut menjadi mudah dan cepat dilakukan. Teknik yang digunakan adalah teknik klasifikasi dan kalsterisasi.

d) *Fraus Decection* digunakan dalam menemukan pelanggan yang mungkin melakukan kecurangan. Sejumlah data yang besar apabila dilakukan secara manual akan membutuhkan biaya dan waktu yang lama sehingga penggunaan teknik *data mining* dapat mempercepat dalam penemuan kecurangan di dalam suatu basis data pelanggan. Sistem ini dibengun menggunakan teknik *anomaly detection*.

2. Sains dan Teknik. Beberapa teknik *data mining* dapat digunakan dalam dunia sains dan teknik untuk menyelesaikan permasalahan yang kompleks, seperti genetika, medis, teknik elektro, dan sebagainya.
3. Seni dan Hiburan. *Data Mining* juga dapat diaplikasikan ke dalam seni dan hiburan, seperti menentukan lagu kesukaan yang sering kali diputar ataupun merekomendasikan jenis lagu ataupun video yang memiliki kemiripan yang sama dengan lagu atau video favorit.

4. Dekripsi

   Deskripsi merupakan penggambaran suatu objek yang bertujuan untuk mengidentifikasi kemudian menganalisis dan mengubah bentuk suatu data yang muncul berulang menjadi bentuk yang bisa dibaca atau dipahami oleh domain aplikasinya.

5. Prediksi

   Prediksi mempunyai kesamaan dengan klasifikasi, namun data yang dikelompokkan sesuai dengan perilaku atau nilai yang diperkirakan untuk waktu yang akan datang. Contoh dari pada prediksi yaitu seperti memprediksi adanya pengurangan jumlah pelanggan dalam waktu dekat atau memprediksi harga saham dalam tiga bulan yang akan datang.

6. Estimasi

   Memiliki pengertian yang hampir serupa dengan prediksi, estimasi memiliki target lebih kearah numerik disbanding kearah kategori. Model dibangun menggunakan record lengkap yang menyediakan nilai dari variabel target dibuat berdasarkan nilai variabel prediksi.

7. Klasifikasi

   Klasifikasi adalah proses dimana ditemukannya sebuah model atau fungsi yang menggambarkan serta membedakan data kepada beberapa kelas. Klasifikasi melibatkan proses pemeriksaan karakteristik dari objek dan memasukkan objek ke dalam salah satu kelas yang sudah didefenisikan sebelumnya.

8. *Clustering*

   Clustering adalah kegiatan mengelompokkan data berdasarkan beberapa kelas dan digabungkan dengan objek yang sama. Tujuannya adalah untuk mengelompokkan objek-objek yang sama atau hampir sama kedalam beberapa kelompok. Semakin besar kemiripan objek dalam suatu cluster dan semakin besar perbedaan tiap cluster maka kualitas analisis cluster semakin baik.

9. Asosiasi

   Asosiasi dalam *data mining* merupakan penemuan atribut yang muncul dalam suatu waktu. Dalam dunia bisnis lebih umum disebut analisis keranjang belanja (*market basket analisys*). Tugas asosiasi adalah untuk menemukan aturan untuk mengukur hubungan antara dua atau lebih atribut yang berhubungan.

2.4 *K-Means Clustering*

Metode *K-Means Clustering* atau dikenal juga *Algoritma K-Means Clustering* adalah metode yang sudah tidak asing lagi dalam pengolahan data serta terkenal praktis. Tujuannya juga untuk pengelompokan data atau objek menjadi beberapa cluster (grup) maka pada setiap cluster akan diisi dengan beberapa data dengan cluster terdekatnya. Menurut (Rochcham, 2020) *Algoritma K-Means* adalah metode *data mining* yang sering digunakan untuk mengidentifikasi dan menganalisis kemiripan dalam pengelompokan data. Menurut (Thabit et al., 2020) *K-Means* yaitu metode algoritma yang menganalisa data dengan menentukan nilai pada data yang

akan dikelompokkan secara acak dan menemukan objek pada satu kelompok yang sama atau memiliki hubungan atau yang tidak berhubungan dengan objek kelompok lainnya.

Langkah-langkah dalam membentuk cluster secara literatif yaitu:

Step 1 : Tentukan dulu jumlah cluster K yang akan dibentuk.

Step 2 : Menentukan titik *clustering* secara acak berdasarkan cluster nya.

Step 3 : Menghitung jarak antar data dengan titik clustering menggunakan rumus *Euclidean Distance:*

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Rumus 2 1 *Euclidean Distance*

Step 4 : Setelah data dikelompokkan berdasarkan jarak yang terdekat dengan setiap titik *clustering*, untuk menentukan ataupun menghitung titik clustering yang baru ditemukan yaitu dengan menghitung nilai rata-rata dari titik data yang ada di cluster masing-masing dengan menggunakan rumus:

$$\mu_k = \frac{1}{N_k}\sum_{q=1}^{N_k} x_q$$

Rumus 2 2 Nilai rata-rata

Step 5 : Lakukan proses literasi sampai selesai. Literasi terakhir berakhir apabila nilainya sama dengan iterasi sebelumnya.

K-Means adalah metode yang membagi data menjadi beberapa cluster berbasis jarak dengan atribut numeric. *Algoritma K-Means* terkenal karena kemudahannya dalam mengolah data besar dan cepat. *Algoritma K-Means clustering* adalah

algoritma yang membutuhkan parameter input sebanyak k dan membagi sekumpulan n objek kedalam k cluster membentuk grup dengan tingkat kemiripan yang tinggi sedangkan pada anggota cluster lain memiliki tingkat kemiripan yang rendah.

2.5 *Software* Pendukung

Berikut beberapa *software* pendukung yang dapat digunakan dalam pengolahan *data mining* dengan menggunakan metode *K-Means Clustering*.

2.5.1 *RapidMiner*

Rapidminer merupakan sebuah perangkat lunak yang difungsikan dalam membantu analisis *data mining, text mining* serta analisis prediksi. Sebagai *software* yang bersifat *Open source* atau terbuka *RapidMiner* ini telah menempati peringkat pertama sebagai *Software data mining* pada pemilihan oleh KDnuggets, sebuah portal *data mining* pada tahun 2010-2011. Dikenal sebagai alat bantu dalam membuat keputusan yang baik, *RapidMiner* menggunakan teknik *deskriptif* dan prediksi terhadap wawasan penggunanya.

Gambar 2. 1 Aplikasi *RapidMiner*

Sumber: *RapidMiner* Studio (2019)

2.5.2 *Tanagra*

Tanagra adalah salah satu *software* dalam *data mining* yang memiliki *User interface* sederhana sehingga dikenal mudah dalam pengoperasiannya. Aplikasi Tanagra juga memberikan akses terhadap algoritma dalam teknik *data mining,* penggunaan aplikasi Tanagra juga di set mengidentifikasi data dengan *database* dalam bentuk *txt* pada notepad dan *xls* pada Microsoft excel. Akan tetapi perangkat lunak ini tidak memasukkan set sumber data yang luas, akses langsung ke data warehouses, dan *database*, data *cleansing* dan *interactive utilization* seperti yang ada pada *software* komersil saat ini.

Gambar 2. 2 Aplikasi *Tanagra*

Sumber: *Tanagra Data Mining*

2.6 Penelitian Terdahulu

Ada beberapa penelitian terdahulu yang dijadikan sebagai acuan dalam penelitian ini, berikut adalah beberapa penelitian tersebut:

1. Menurut (Swastati, 2017) Penelitian yang berjudul **"Pengenalan Penyakit Pada Manusia Berbasis Android Menggunakan Metode *Sequential Search*"** membahas tentang pengenalan penyakit pada manusia yang sering diderita yaitu berupa penyakit tidak menular, penyakit menular, dan kronis dengan menggunakan metode *sequential search* dimana penggunaan metode ini adalah unruk mencari sebuah data dari kumpulan data dari awal sampai akhir fungsinya adalah agar dapat membantu masyarakat dalam mencari data. Kemudian aplikasi yang digunakan adalah android 4.4 (KitKat), android (5.0) Lollipop, android (5.1) Lollipop dengan hasil cukup baik artinya masih perlu penambahan dan hanya dapat dijalankan pada android.
2. Menurut Penelitian (MURTI, 2017) yang berjudul **"Penerapan Metode *K-Means Clustering* untuk mengelompokkan potensi produksi buah-buahan di Provinsi Daerah Istimewa Yogyakarta"** membahas tentang pengelompokan hasil produksi buah-buahan dalam bebrapa daerah dengan menggunakan metode *K-Means* seperti berdasarkan luas panen, produksi dan tahun panen berdasarkan data di beberapa daerah tujuannya adalah untuk memudahkan pengelompokan suatu daerah dengan hasil produksi buah yang paling banyak, sedang dan rendah. Hasilnya adalah akan ditemukan pengelompokan daerah dengan potensial produksi buah yang paling tinggi.

3. Menurut Penelitian (Bastian et al., n.d.) pada tahun 2018 Penelitian yang berjudul **"Penerapan Algoritma *K-Means clustering* Analisis Pada Penyakit Menular Studi kasus Kabupaten Majalengka)"** membahas tentang penerapan Algoritma *K-Means Clustering* dalam pengelompokan data penyakit menular pada manusia berdasarkan data yang diperoleh dari puskesmas di Kabupaten Majalengka yang terdapat ada 32 kantor Puskesmas dengan mengangkat 6 jenis data penyakit menular yang dikumpulkan dari sejumlah Puskemas di Kabupaten Majalengka. Hasil dari penelitian tersebut akan diketahui Puskesmas yang mendominasi dengan tingkat tertinggi penderita penyakit menular serta jenis penyakitnya sehingga tiap-tiap puskesmas dari kabupaten Majalengka dapat mengendalikan persediaan obat serta penanganan yang lebih intensif sesuai dengan hasil data yang diperoleh.
4. Menurut Penelitian (Handoko & Lesmana, 2018) yang berjudul **"*Data Mining* pada jumlah penumpang menggunakan metode *Clustering*"** membahas tentang pengelompokan jumlah penumpang di bandar udara hang Nadim dengan menggunakan metode *clustering* dengan beberapa variabel yaitu variabel pertama penumpang yang datang, kedua penumpang yang berangkat dan ketiga penumpang yang transit dengan data banyak, sedang dan sedikit yaitu di tahun 2015 hingga tahun 2017 juga dilakukan pengujian dengan menggunakan Aplikasi *RapidMiner*. Hasil yang diperolah adalah pengetahuan tentang jadwal padat pelanggan bandar udara hang nadim perbulannya sehingga dapat membantu dalam mengantisipasi petugas bandar udara hang

Nadim terhadap penumpang yang berangkat, penumpang yang datang serta penumpang yang transit di bulan-bulan tertentu.

5. Menurut Penelitian (Dhuhita, 2015)yang berjudul **"*Clustering* Menggunakan Metode *K-Means* untuk menentukan Status Gizi Balita"** membahas tentang pengelompokan antara tinggi badan balita(TB) dan berat badan balita (BB), dikelompokkan menjadi status gizi balita kedalam 5 cluster status gizi yaitu gizi buruk, gizi kurang, gizi baik, gizi lebih dan obesitas menggunakan tabel *Growth Chart* dengan menggunakan data balita dengan jumlah 50 balita di Desa Karang Songo dengan usia < 3 tahun. Kemudian dilakukan perhitungan cluster dengan menggunakan SPSS, Analisa hasil data output lalu melakukan pengujian dengan membandingkan hasil pengelompokan *algoritma K-Means* dan tabel *Growth Chart*. Hasilnya adalah didapatkan 17 data yang memiliki kelompok yang sama sehingga disimpulkan *algoritma K-Means* memiliki nilai akurasi 34% benar dan dapat berubah sesuai dengan data yang ditambahkan.
6. Menurut Penelitian (Prediction & Syndromes, 2019) yang berjudul ***"Disease Prediction through syndromes using K-Means algorithm"***. *This study describe a research work aiming to find out how much efficient K-Means can be build an expert system to detect human disease by evaluating symptoms data to improve the quality of health evaluation. This research collected 61 symptoms and there certain 20 diseases, that is Anemia, Angina, Asthma, Bacillary Dysentery, Bronchiolitis, Chickenpox, Dengue Fever, Diabetes Mellitus, Diarrhea, Jaundice, Leukemia, Malaria, Myocardial Infarction (MI), Peptic Ulcer, Pneumonia, Rheumatic Fever, Scurvy, Stroke, Tuberculosis,*

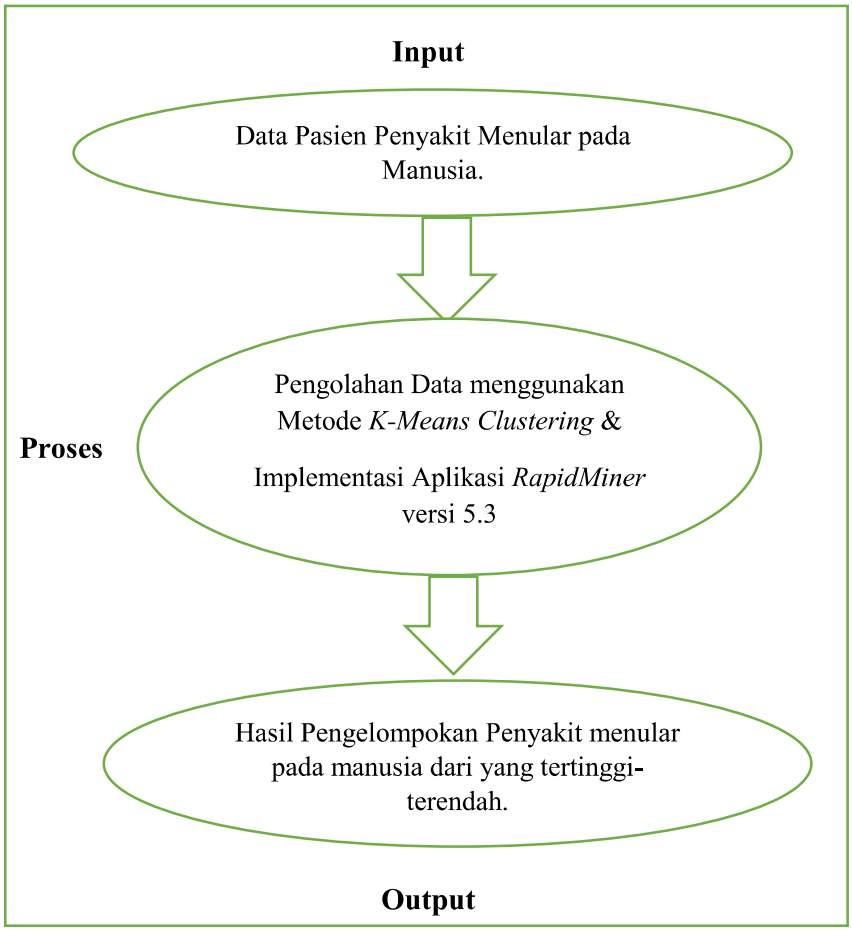
Typhoid Fever. Then applied K-Means to the dataset to predict the outcome of each point based on selected independent data and evaluated the performance of the model 'f1-score' does it is combines precision can give a better intuition about the performance. Conclusion of this research is analize disease in the healthcare domain to discover a new range of patterns and information using K-Means clustering and to make K-Means more effective it can be used in combination with other algorithms to produce accurate, relevant and useful results.

7. Menurut Penelitian (Li, 2019) yang berjudul ***"Study on the Grouping of Patients with Chronic Infectious Diseases Based on Data Mining".*** *This study describes data mining technology research that is used to classify chronic infectious disease patients in order to predict patients according to the level of infectious disease suffered. K-Means clustering algorithm was used to classify chronic infectious disease patients, then C5.0 decision tree algorithm was used to predict the state of chronic infectious disease patients with 170,246 outpatient data, 43,448 data formed after cleaning data. The C5.0 decision tree algorithm was used to predict the treatment situation of patients with chronic infectious diseases, 99.94% accuracy rate verified by the confusion model. The conclusion of this research study is that medical institutions should be better at socializing chronic infectious diseases to patients and their communities, providing solutions to help them improve medication adherence. To accelerate the development of hospital information as well as in handling patients to build a database of chronic infectious diseases to determine the level of ups and*

downs of chronic infectious diseases from time to time, strengthen the blocking of transmission from mother to child, to effectively curb chronic infectious diseases, reduce the burden of disease and death.

2.7 Kerangka Pemikiran

Kerangka Pemikiran merupakan sebuah diagram atau tabel yang menggambarkan secara garis besar berjalan nya suatu penelitian. Adapun kerangka pemikiran dari penelitian ini adalah sebagai berikut:



Gambar 2.3 Kerangka Pemikiran

Sumber: Data Peneliti (2021)

Sebagai input dalam penelitian ini adalah data pasien penyakit menular pada manusia di UPT Puskesmas Sei Langkai yang meliputi 3 kelurahan/desa yaitu Sei Langkai, Tembesi, dan Sei Pelunggut yang akan diproses dengan menggunakan Metode *K-Means Clustering* kemudian di Implementasikan juga menggunakan *Software* Aplikasi *data mining RapidMiner* sebagai pengujian data serta akan menghasilkan pengelompokan Penyakit menular pada manusia dari yang tertinggi sampai terendah di UPT Puskesmas Sei Langkai sebagai Outputnya.