Jurnal Teknologi Dan Sistem Informasi Bisnis
Vol. 5 No. 3 Juli 2023 Hal. 172-182
http://jurnal.unidha.ac.id/index.php/jteksis

E-ISSN : **2655-8238**
P-ISSN : **2964-2132**

Analisis Klaster Data Pasien Diabetes untuk Identifikasi Pola dan Karakteristik Pasien

Ananda Elang Satriatama Setyadji[1], Ari Prasetyo Wibowo[2], I Gusti Ngurah Arnold Matthew D[3], Reyhan Bayu Pratama[4], Tegar Alwinata Masyhuda[5], Yohannes Alexander Agusti Sinaga[6], Endah Purwanti[7*], Indah Werdiningsih[8*]

[1,2,3,4,5,6,7,8]Program Studi Sistem Informasi, Fakultas Sains dan Teknologi, Universitas Airlangga
Email : ananda.elang.satriatama-2020@fst.unair.ac.id[1], ari.prasetyo.wibowo-2020@fst.unair.ac.id[2], i.gusti.ngurah.arnold-2020@fst.unair.ac.id[3], reyhan.bayu.pratama-2020@fst.unair.ac.id[4], tegar.alwinata.masyhuda-2020@fst.unair.ac.id[5], yohannes.alexander.agusti-2020@fst.unair.ac.id[6],
Corresponding Author Email : endahpurwanti@fst.unair.ac.id[7], indah-w-i@fst.unair.ac.id[8]

Submitted: 18-05-2023, Reviewed: 23-05-2023, Accepted 27-06-2023
https://doi.org/10.47233/jteksis.v5i3.828

Abstract
Diabetes is a dangerous health problem both in Indonesia and the world. The aim of the study was to cluster diabetes patient data at the Mojokerto Health Center using the K-Means algorithm to understand the patterns and characteristics of diabetic patients. The steps taken were patient data collection, data preprocessing, data sharing, cluster analysis using K-Means, model performance evaluation, and prediction/classification using the built model. The results of the analysis showed that there were two groups of patients. The first group consisted of 755 female patients aged 20-80 years, while the second group consisted of 404 male patients aged 40-90 years. The most common diagnosis in both groups was Non-insulin-dependent diabetes mellitus, followed by Rheumatoid arthritis in the first group and Respiratory tuberculosis in the second group. In addition, in the two groups, the Modopuro and Kebondalem sub-districts had the highest number of patients.
Keywords: *kmeans, clustering, diabetes, data analysis*

Abstrak
Penyakit diabetes merupakan masalah kesehatan yang berbahaya baik di Indonesia dan dunia. Tujuan penelitian adalah melakukan clustering pada data pasien diabetes di Puskesmas Mojokerto menggunakan metode algoritma K-Means untuk memahami pola dan karakteristik pasien diabetes. Tahapan yang dilakukan adalah pengumpulan data pasien, preprocessing data, pembagian data, analisis klaster menggunakan K-Means, evaluasi performa model, dan prediksi/klasifikasi menggunakan model yang dibangun. Hasil analisis menunjukkan adanya dua kelompok pasien. Kelompok pertama terdiri dari 755 pasien perempuan berusia 20-80 tahun, sedangkan kelompok kedua terdiri dari 404 pasien laki-laki berusia 40-90 tahun. Diagnosa paling umum di kedua kelompok tersebut adalah *Non-insulin-dependent diabetes mellitus*, diikuti oleh *Rheumatoid arthritis* pada kelompok pertama dan *Respiratory tuberculosis* pada kelompok kedua. Selain itu, dalam kedua kelompok tersebut, kelurahan Modopuro dan Kebondalem memiliki jumlah pasien yang paling banyak.
Keywords: kmeans, klastering, diabetes, analisis data

PENDAHULUAN

Diabetes adalah penyakit kronis yang mempengaruhi jutaan orang di seluruh dunia. Penyakit ini merupakan kondisi medis kronis yang ditandai oleh tingginya tingkat glukosa (gula) dalam darah. Keadaan ini terjadi ketika tubuh tidak mampu menghasilkan jumlah insulin yang cukup atau tidak dapat memanfaatkan insulin dengan efisien. Insulin adalah hormon yang diproduksi oleh pankreas dan berperan penting dalam mengatur kadar glukosa dalam darah [1]. Diabetes terbagi menjadi beberapa jenis, termasuk diabetes tipe 1, diabetes tipe 2, diabetes gestasional, dan jenis diabetes lainnya. Diabetes tipe 1 biasanya muncul pada masa anak-anak atau remaja, di mana pankreas tidak dapat menghasilkan insulin sama sekali. Di sisi lain, diabetes tipe 2 adalah jenis yang paling umum terjadi, di mana tubuh tidak dapat menggunakan insulin secara efektif atau tidak memproduksi insulin yang cukup [1]. Salah satu contohnya adalah diabetes *mellitus.* Diabetes *melitus* merupakan salah satu dari banyaknya jenis penyakit diabetes. *diabetes melitus* memerlukan perawatan yang berkelanjutan, terutama dalam mengontrol tingkat glukosa darah guna mencegah atau melambatkan timbulnya masalah komplikasi. Diabetes mellitus menjadi sesuatu yang baru bagi penduduk Indonesia [2].

Menurut Federasi Diabetes Internasional, prevalensi diabetes telah meningkat menjadi 463 juta pada tahun 2019 dan diperkirakan akan mencapai 700 juta pada tahun 2045. Di Indonesia, diabetes adalah salah satu penyakit tidak menular

Jurnal Teknologi Dan Sistem Informasi Bisnis
Vol. 5 No. 3 Juli 2023 Hal. 172-182
http://jurnal.unidha.ac.id/index.php/jteksis

E-ISSN : **2655-8238**
P-ISSN : **2964-2132**

yang paling umum terjadi dengan prevalensi sebesar 6,9% pada tahun 2020, mempengaruhi lebih dari 16 juta orang [3].

Puskesmas adalah organisasi kesehatan fungsional yang berperan sebagai pusat pengembangan kesehatan masyarakat. Selain memberikan pelayanan kesehatan menyeluruh dan terpadu, Puskesmas juga melibatkan partisipasi masyarakat. Puskesmas memiliki satuan penunjang seperti puskesmas pembantu dan puskesmas keliling. Puskesmas pembantu adalah unit pelayanan kesehatan sederhana yang membantu Puskesmas dalam wilayah yang lebih kecil [4].

Puskesmas memiliki tiga fungsi utama. Pertama, sebagai pusat pembangunan kesehatan masyarakat di wilayahnya, Puskesmas bertanggung jawab dalam mengembangkan dan meningkatkan kesehatan masyarakat secara umum. Kedua, Puskesmas juga memiliki peran penting dalam membina partisipasi aktif masyarakat di wilayah kerjanya. Hal ini dilakukan dengan tujuan meningkatkan kesadaran dan kemampuan masyarakat untuk menjalani gaya hidup yang sehat. Ketiga, Puskesmas memberikan pelayanan kesehatan yang menyeluruh dan terpadu kepada masyarakat di wilayahnya. Dalam hal ini, Puskesmas memberikan layanan kesehatan yang meliputi berbagai aspek, seperti pencegahan penyakit, pengobatan, pemantauan kesehatan, serta edukasi kesehatan kepada masyarakat. Dengan fungsi-fungsi tersebut, Puskesmas menjadi lembaga yang berperan penting dalam menjaga dan meningkatkan kesehatan masyarakat di wilayahnya [4].

Puskesmas memiliki peran penting dalam memberikan layanan kesehatan kepada masyarakat, termasuk pencegahan, manajemen, dan pengobatan diabetes. Namun, manajemen pasien diabetes di Puskesmas sangatlah sulit karena jumlah pasien diabetes yang semakin meningkat, sumber daya yang terbatas, dan kurangnya sistem perawatan diabetes yang komprehensif.

Penyajian informasi tidak sebanding dengan kebutuhan informasi yang sangat besar. Untuk mengatasi hal tersebut, salah satu cara untuk mengekstrak informasi dari *database* yang besar adalah *data mining* atau penambangan data yang bertujuan untuk mengekstrak informasi abstrak dari *database* yang besar [5]. Klasterisasi adalah teknik penambangan data yang mengelompokkan objek yang serupa ke dalam klaster berdasarkan karakteristik atau atribut. Klasterisasi dapat digunakan di bidang kesehatan untuk mengidentifikasi sub kelompok pasien dengan karakteristik dan faktor risiko yang serupa, yang dapat membantu dalam mengembangkan intervensi yang ditargetkan dan meningkatkan hasil kesehatan. Pendekatan klasterisasi akan diterapkan pada kumpulan data besar pasien diabetes yang telah mengunjungi Puskesmas Mojokerto. Hasil dari pendekatan klasterisasi ini akan memberikan wawasan tentang heterogenitas pasien diabetes dan membantu mengembangkan strategi perawatan dan manajemen diabetes yang lebih efektif di Puskesmas.

Pada penelitian sebelumnya, oleh Parasian D.P Silitonga yang menerapkan metode K-Means Clustering pada data penyakit pasien di Rumah Sakit Haji Adam Malik di Medan menemukan sebuah pola kecenderungan penyakit pada pasien di Rumah Sakit Haji Adam Malik. Kemudian, pada penelitian ini memiliki tujuan utama untuk memberikan panduan dalam mengantisipasi layanan prioritas bagi pasien, terutama bagi pengguna Jaminan Sosial dan Jaminan Kesehatan. Selain itu, penelitian yang sedang dilakukan di Puskesmas Mojokerto memiliki perbedaan dengan penelitian sebelumnya. Penelitian ini fokus pada data penyakit pasien di Puskesmas Mojokerto dan mengadopsi metode kombinasi antara K-Means Clustering dan PCA (Principal Component Analysis). Tujuannya adalah untuk mengidentifikasi pola kecenderungan penyakit pada pasien di Puskesmas Mojokerto dengan menggunakan pendekatan yang lebih kompleks dan integratif [22].

Secara keseluruhan, klasterisasi penyakit diabetes di Puskesmas dapat menjadi cara yang efektif untuk mengidentifikasi sub kelompok pasien diabetes dan mengembangkan intervensi yang ditargetkan untuk meningkatkan hasil perawatan diabetes. Hasil dari pendekatan klasterisasi ini dapat membantu meningkatkan layanan perawatan diabetes di Puskesmas dan mengurangi beban diabetes pada masyarakat.

METODE PENELITIAN

Berikut tahapan penelitian yang dilakukan pada clustering data pasien penyakit diabetes pada Studi Kasus Puskesmas Mojokerto

2.1 Pengumpulan Data

Tahap pengumpulan data merupakan tahap awal dalam penelitian diabetes di Puskesmas Mojokerto. Selama periode 5 bulan., data jumlah pasien diabetes yang terdaftar dan berkunjung ke Puskesmas setiap bulannya dikumpulkan. Dalam pengumpulan data ini, peneliti sangat menjaga kerahasiaan pasien dengan tidak mengambil informasi tentang riwayat medis atau kondisi kesehatan pasien secara detail. Hal ini dilakukan untuk menghormati privasi pasien dan mematuhi etika penelitian yang sesuai dengan standar penelitian ilmiah.

Dalam pengumpulan data ini, peneliti mengambil langkah-langkah yang diperlukan untuk memastikan

Jurnal Teknologi Dan Sistem Informasi Bisnis
Vol. 5 No. 3 Juli 2023 Hal. 172-182
http://jurnal.unidha.ac.id/index.php/jteksis

E-ISSN : **2655-8238**
P-ISSN : **2964-2132**

bahwa identitas pasien tetap terjaga. Data yang dikumpulkan hanya mencakup informasi jumlah pasien diabetes yang terdaftar dan berkunjung ke Puskesmas setiap bulannya, tanpa memuat informasi pribadi yang dapat mengidentifikasi pasien secara individu. Dengan demikian, penelitian ini memastikan bahwa kegiatan pengumpulan data dilakukan dengan penuh kehati-hatian dan memperhatikan aspek privasi pasien serta etika penelitian yang berlaku.

Tabel 1. Deskripsi Dataset

Nama Dataset	Data Pasien Puskesmas Mojokerto
Jumlah Baris	1163 baris
Jumlah Kolom	72 kolom
Nama Kolom	No, Tanggal, Nama Pasien, No. eRM, NIK, No. RM Lama, No. Dokumen RM, Jenis Kelamin, No Telp, Alamat, RT, RW, Pekerjaan, Tanggal Pemeriksaan, Kelurahan, Tempat Lahir, Tgl.Lahir, Umur Tahun, Umur Bulan, Umur Hari, Nama Ayah, Nama Ibu, Jenis Kunjungan, Poli/Ruangan, Asuransi, No. Asuransi, Dokter / Tenaga Medis, Perawat / Bidan / Nutrisionist / Sanitarian, Keluhan Utama, Keluhan Tambahan, Lama Sakit, Merokok, Konsumsi Alkohol, Kurang Sayur/Buah, Terapi Keterangan, RPS, RPD, RPK, Alergi, Kesadaran, Triage, Tinggi, Berat Badan, Lingkar Perut, IMT, Hasil IMT, Sistole, Diastole, Nafas, Detak Nadi, Detak Jantung, Suhu, Aktifitas Fisik, Diagnosa 1, Jenis Kasus 1, Diagnosa 2, Jenis Kasus 2, Diagnosa 3, Jenis Kasus 3, Diagnosa 4, Jenis Kasus 4, Diagnosa 5, Jenis Kasus 5, Tindakan, Resep, Apoteker, Pendaftaran/Rujukan Internal, Lama Antrean, Lama Pemeriksaan, Lama Pelayanan Obat, Petugas Pendaftaran

2.2 Preprocessing

Preprocessing data merupakan langkah penting dalam proses penemuan pengetahuan, karena keputusan-keputusan yang berkualitas harus didasarkan pada data yang berkualitas [11]. Preprocessing data sering kali digunakan untuk mengurangi kesalahan data dan sistematis bias dalam data mentah sebelum analisis apapun terjadi [10].

Preprocessing memiliki beberapa tahap, yaitu seleksi data, cleaning data , dan normalisasi data transformasi data dan training data.

1. **Seleksi data** Seleksi data adalah proses memilih dan mempertahankan subset dari data yang relevan untuk analisis atau model yang akan dibangun. Tujuan dari seleksi data adalah untuk mengurangi kompleksitas data dan meningkatkan kualitas analisis atau model yang dihasilkan. dilakukan dengan memilih feature atau variabel yang relevan dan sesuai dengan tujuan analisis data. Hal ini penting dilakukan untuk menghindari terjadinya overfitting pada model yang dibangun dan memastikan keakuratan hasil analisis.
2. **Cleaning data** (pembersihan data) adalah proses mengidentifikasi, mengatasi, dan memperbaiki suatu masalah data atau ketidakakuratan dalam dataset. Tujuan utama dari cleaning data adalah untuk memastikan kebersihan, kekonsistenan, dan kevalidan data yang akan digunakan dalam analisis atau pemodelan. dilakukan untuk membersihkan data dari noise, outlier, dan missing value. Noise dan outlier dapat menyebabkan analisis yang tidak akurat, sedangkan missing value dapat mengganggu proses analisis data. Beberapa metode yang sering digunakan untuk membersihkan data adalah penghapusan data, imputasi data, dan interpolasi data.
3. **Normalisasi data** adalah proses mengubah data menjadi bentuk yang terstandarisasi atau ternormalisasi. Tujuan dari normalisasi data adalah untuk menghilangkan perbedaan skala atau ukuran yang dapat mempengaruhi analisis atau pemodelan data. Normalisasi merupakan teknik pengukuran ulang atau pemetaan dalam tahap pra-pemrosesan . Teknik ini berguna untuk tujuan prediksi atau peramalan . Terdapat banyak cara untuk melakukan prediksi atau peramalan yang dapat berbeda satu sama lain. Oleh karena itu, teknik normalisasi diperlukan untuk membuat variasi prediksi dan peramalan lebih dekat satu sama lain. Terdapat beberapa teknik normalisasi yang umum digunakan, yaitu Min-Max.
4. **Transformasi data** dilakukan untuk mengubah data yang akan dianalisis sehingga dapat menghasilkan nilai yang lebih bermakna. Beberapa metode transformasi data yang sering digunakan adalah normalisasi, standarisasi, dan encoding. Normalisasi dilakukan untuk mengubah data pada rentang nilai tertentu,

Jurnal Teknologi Dan Sistem Informasi Bisnis
Vol. 5 No. 3 Juli 2023 Hal. 172-182
http://jurnal.unidha.ac.id/index.php/jteksis

E-ISSN : **2655-8238**
P-ISSN : **2964-2132**

sedangkan standarisasi dilakukan untuk mengubah data sehingga memiliki nilai mean = 0 dan standar deviasi = 1. Encoding dilakukan untuk mengubah data kategori menjadi nilai numerik.

5. **Training Data** pada tahap preprocessing digunakan untuk mempersiapkan data sebelum dilakukan training model. Proses preprocessing meliputi tahapan cleaning, transforming, dan feature selection. Setelah preprocessing, data dibagi menjadi training set dan validation set untuk menghindari overfitting dan memastikan akurasi model.

Tahapan preprocessing data memainkan peran yang krusial dalam penelitian, karena tahap ini akan menghasilkan dataset yang siap digunakan untuk analisis data selanjutnya. Dalam tahap ini, data mentah yang dikumpulkan dari pengumpulan data pasien diabetes di Puskesmas Mojokerto akan diperlakukan melalui serangkaian langkah pemrosesan. Langkah-langkah tersebut meliputi pembersihan data, seperti mengatasi missing values dan outliers, normalisasi data untuk menghilangkan perbedaan skala, serta seleksi fitur untuk memilih fitur yang relevan dan signifikan. Dengan melakukan tahap preprocessing data ini, dataset akan menjadi lebih terstruktur, bersih, dan siap untuk digunakan dalam proses analisis data selanjutnya.

Dataset yang telah diproses melalui tahapan preprocessing data akan menjadi landasan yang kuat dalam membangun model dan melakukan analisis data. Dalam penelitian diabetes di Puskesmas Mojokerto, dataset yang sudah diproses akan digunakan untuk mengaplikasikan algoritma machine learning dan metode analisis statistik yang relevan. Dengan melakukan preprocessing data secara teliti, peneliti dapat memastikan bahwa dataset yang digunakan memiliki kualitas yang baik, mengurangi risiko bias atau kesalahan dalam hasil analisis, serta meningkatkan keakuratan dan keandalan temuan penelitian. Oleh karena itu, tahapan preprocessing data tidak boleh diabaikan dan perlu dilakukan dengan cermat dan teliti dalam rangka mencapai hasil analisis data yang akurat dan reliable.

2.3 Menentukan Nilai K

Dalam pengajuan untuk melakukan penelitian, Metode Elbow digunakan. Metode Elbow adalah metode yang digunakan untuk menghasilkan informasi dalam menentukan jumlah cluster terbaik dengan cara melihat persentase hasil perbandingan antara jumlah cluster yang akan membentuk siku pada suatu titik [8].

Tujuan dari metode elbow adalah untuk memilih nilai k yang kecil dan masih memiliki nilai withinss yang rendah [9].

Pada metode Elbow, dilakukan memvariasikan jumlah klaster dari 1 hingga k (sejumlah titik data yang ada). Kemudian, menghitung withinss untuk setiap nilai k tersebut. Dalam grafik elbow, sumbu x menunjukkan nilai k, sedangkan sumbu y menunjukkan nilai withinss.

Proses selanjutnya adalah memeriksa pola grafik dan mencari titik di mana penurunan withinss mulai menurun secara signifikan, atau secara visual terlihat seperti "siku" pada siku tangan. Titik ini menandakan jumlah klaster yang optimal, di mana penambahan klaster setelahnya memberikan manfaat yang relatif lebih kecil dalam mengurangi withinss.

Pemilihan jumlah klaster yang optimal bergantung pada persepsi peneliti dan konteks masalah yang diteliti. Metode Elbow memberikan panduan visual yang dapat membantu dalam pengambilan keputusan, tetapi tetap memerlukan penilaian subjektif dari peneliti untuk menentukan titik siku yang paling sesuai sebagai jumlah klaster yang optimal.

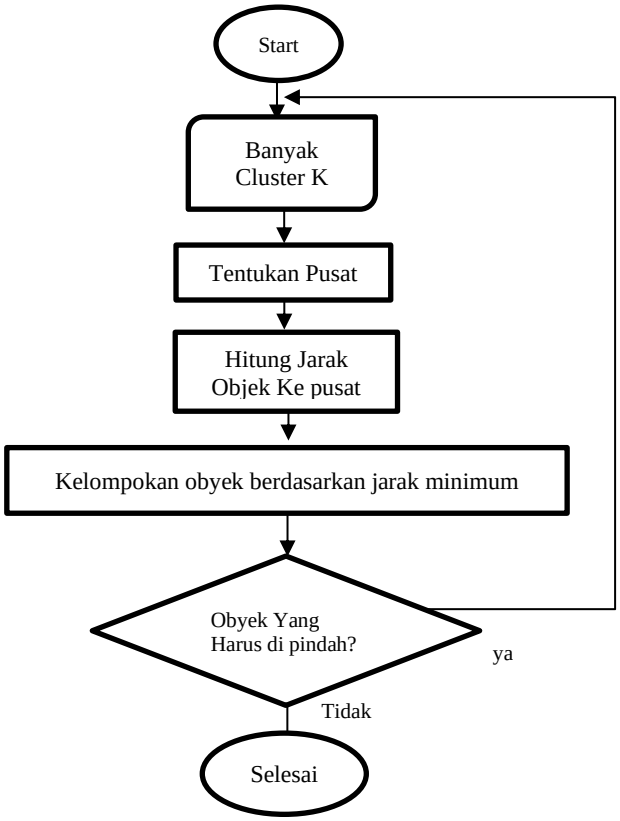
2.4 K-Means Clustering

K-Means Clustering adalah metode pengelompokan data non-hirarki yang membagi data ke dalam beberapa kelompok (cluster) sehingga data dengan karakteristik yang serupa dikelompokkan bersama dalam satu cluster, sementara data dengan karakteristik yang berbeda dikelompokkan ke dalam cluster yang berbeda [12].

K-Means Clustering adalah salah satu metode pengelompokan atau *clustering* yang paling umum digunakan pada implementasi data mining dalam melakukan pengklasteran atau pengelompokkan data menjadi beberapa kelompok [6].

Jurnal Teknologi Dan Sistem Informasi Bisnis
Vol. 5 No. 3 Juli 2023 Hal. 172-182
http://jurnal.unidha.ac.id/index.php/jteksis

E-ISSN : **2655-8238**
P-ISSN : **2964-2132**

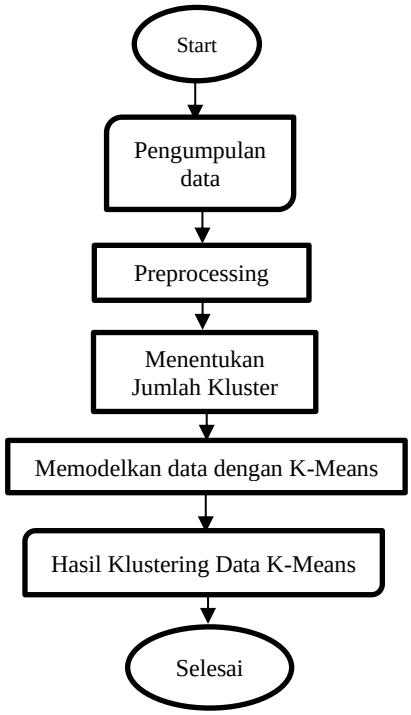


Gambar 1 Alur K-Means Clustering

2.5 Alur Penelitian

Tahap Alur Penelitian merupakan tahapan penting dalam sebuah penelitian, di mana desain yang telah dibuat akan direalisasikan dalam bentuk pemrograman menggunakan bahasa Python. Dalam pengembangan program, dua library populer yang sering digunakan adalah Scikit-learn dan Pandas.

Dalam mengimplementasikan alur program yang dijelaskan dalam Gambar 2 menggunakan Python, peneliti akan menggunakan Scikit-learn dan Pandas untuk membaca dataset, melakukan *preprocessing data* seperti *cleaning*, normalisasi, dan seleksi fitur, membagi data menjadi data latih dan data uji, membangun model menggunakan algoritma machine learning yang sesuai dengan tujuan penelitian, melatih model dengan data latih, mengevaluasi performa model dengan metrik yang relevan seperti akurasi, presisi, dan recall, serta melakukan prediksi atau klasifikasi menggunakan model yang telah dibangun. Gambar 2 merupakan alur jalannya program yang akan diimplementasikan menggunakan bahasa pemrograman *Python*.



Gambar 2. Gambar Alur Penelitian

HASIL DAN PEMBAHASAN

Hasil dari penelitian ini berupa pengolahan dataset tersebut berdasarkan langkah-langkah yang sudah tertera di tahapan metodologi penelitian dengan menggunakan metode K-means Clustering. Alat yang digunakan untuk membantu pengolahan dataset ini adalah menggunakan bahasa pemrograman Python, Jupyter Notebook, dan beberapa library terkait, seperti Numpy, Pandas, Seaborn, Matplotlib, Missingno.

Untuk memastikan kualitas data yang baik, tahap awal adalah melakukan data *preprocessing*. Dataset tersebut tidak mengandung baris yang mengandung duplikat. Selanjutnya, dari 72 kolom, sejumlah 35 kolom memiliki *missing value* dan sejumlah 37 kolom tidak memiliki *missing value.*

Tahap selanjutnya adalah melakukan seleksi data untuk melakukan pemilihan fitur yang berguna untuk pengolahan data selanjutnya dan menghapus fitur yang tidak berguna. Kriteria fitur yang dihapus adalah fitur yang memiliki *missing value* yang banyak, fitur yang memiliki nilai yang sama untuk setiap barisnya, dan fitur yang berupa identitas atau yang tidak ada hubungannya dalam membuat model *machine learning*, seperti NIK, Tanggal, Nama Pasien, dan sebagainya. Hasil akhir dari seleksi data adalah 1163 baris dan 11 kolom.

Jurnal Teknologi Dan Sistem Informasi Bisnis
Vol. 5 No. 3 Juli 2023 Hal. 172-182
http://jurnal.unidha.ac.id/index.php/jteksis

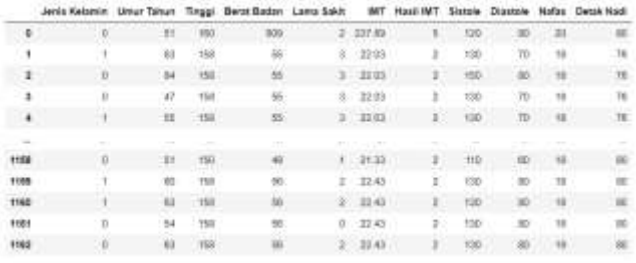
E-ISSN : **2655-8238**
P-ISSN : **2964-2132**

Gambar 3. Hasil Seleksi Data

Dataset juga harus dipastikan memiliki format yang benar untuk setiap kolom dan tidak ada *noise* yang mengganggu kualitas data. Gambar 3 ditunjukkan bahwa data belum memiliki format yang benar dan terdapat *noise* pada setiap kolom, seperti kolom Umur Tahun memiliki satuan 'Tahun'. Lalu, kolom yang memiliki sifat data kategorikal harus diubah menjadi data numerik. Kolom Jenis Kelamin termasuk ke dalam data nominal yang berarti tidak memiliki tingkatan untuk setiap nilai. Kolom Hasil IMT termasuk ke dalam data ordinal yang merupakan setiap nilai harus diperhatikan bobot atau tingkatannya. Tabel 2 menunjukkan hasil dari transformasi data untuk mengubah data kategorikal menjadi data numerik.

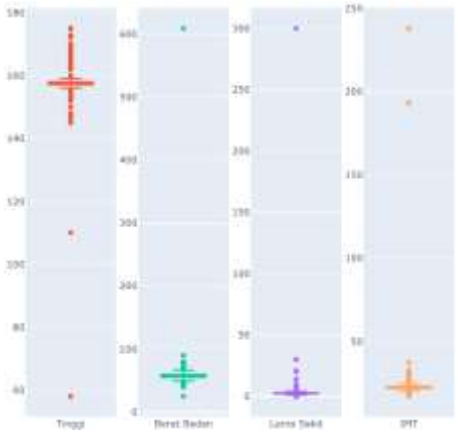
Tabel 2. Transformasi Data

Jenis Kelamin	P = 0 L = 1
Hasil IMT	KURANG = 1 IDEAL = 2 LEBIH = 3 GEMUK = 4 SANGAT GEMUK = 5



Gambar 4. Hasil Cleaning Data

Tahap selanjutnya adalah mengecek adanya data yang mengandung *outlier* dan normalisasi data. *Outlier adalah* kasus atau data dengan karakteristik unik yang sangat berbeda dengan observasi lain dan muncul sebagai nilai ekstrim untuk satu variabel atau kombinasi variabel yang disebabkan oleh kesalahan dalam memasukkan data [7]. Oleh karena itu, tahap selanjutnya adalah melakukan deteksi *outlier* dengan menggunakan metode IQR (*Inter Quartile Range*). Gambar 5 menunjukkan beberapa kolom yang memiliki *outlier* yang jauh. Oleh karena itu, data yang terdeteksi sebagai *outlier* dilakukan penghapusan dari dataset. Hasil dari penghapusan data *outlier* adalah 1159 baris dan 11 kolom. Selanjutnya, normalisasi dilakukan untuk membuat nilai setiap kolom mempunyai skala yang sama. Normalisasi data menggunakan metode MinMax. Hasil dari skala untuk semua data adalah antara 0 dan 1, seperti gambar 6.



Gambar 5. Boxplot Outlier

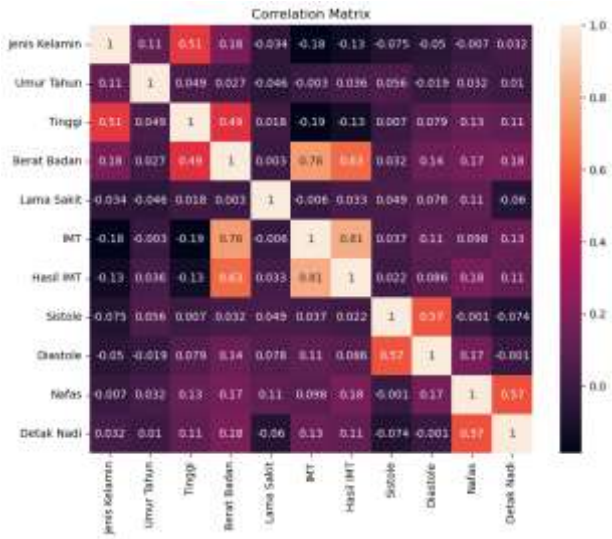
Gambar 6. Hasil Handling Outlier dan Normalisasi Data

Setelah melakukan *preprocessing data*, tahap selanjutnya adalah melihat korelasi atau hubungan antarvariabel. Untuk melihat korelasi antar variabel, gambar 7 merupakan hasil dari perhitungan korelasi antarvariabel menggunakan metode Pearson. Korelasi Pearson menghasilkan koefisien korelasi yang mengukur kekuatan hubungan linier antara dua variabel. Jika hubungan antara kedua variabel tidak linier, koefisien korelasi Pearson tidak mencerminkan kuatnya hubungan antara kedua variabel yang diteliti [20]. Jika koefisien korelasi nilainya mendekati -1 atau 1, hubungan kedua variabel

Jurnal Teknologi Dan Sistem Informasi Bisnis
Vol. 5 No. 3 Juli 2023 Hal. 172-182
http://jurnal.unidha.ac.id/index.php/jteksis

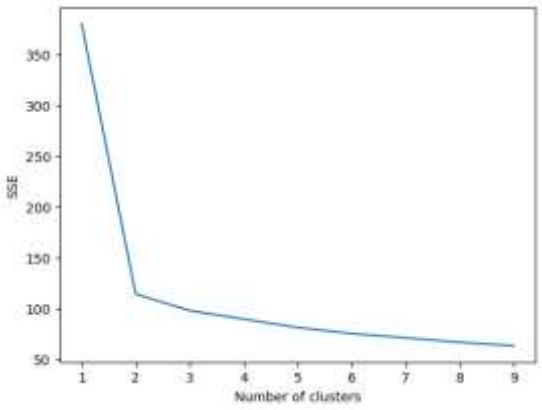
E-ISSN : **2655-8238**
P-ISSN : **2964-2132**

tersebut adalah linier. Namun, jika koefisien korelasi mendekati 0, hubungan kedua variabel tersebut tidak ada atau sangat kecil [21]. Gambar 7 juga menyediakan warna sebagai intensitas hubungan kedua variabel atau koefisien korelasi kedua variabel tersebut. Berdasarkan gambar 7, variabel hasil IMT dengan IMT memiliki koefisien korelasi yang paling tinggi, yaitu 0.81. Hal ini mengartikan bahwa hubungan kedua variabel tersebut linier positif, semakin tinggi nilai IMT semakin tinggi hasil IMT (semakin menuju gemuk). Begitu juga dengan variabel lain, grafik tersebut yang menunjukkan hubungan kedua variabel tersebut kuat adalah variabel jenis kelamin dengan tinggi badan, tinggi badan dengan berat badan, berat badan dengan hasil IMT atau IMT atau tinggi badan, sistole dengan diastole, dan detak nadi dengan nafas.

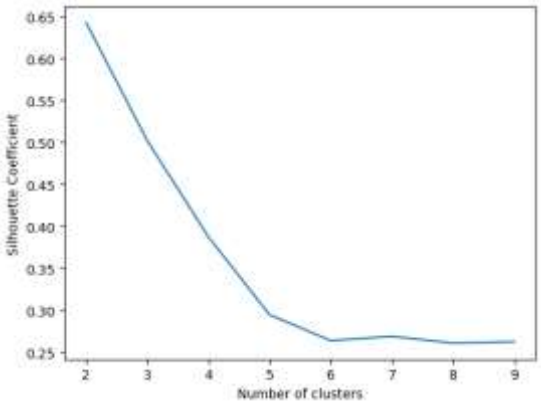


Gambar 7. Correlation Matrix

Tahap selanjutnya adalah melakukan membuat model *machine learning* dengan menggunakan metode K-Means Clustering. Sebelum membuat model, jumlah *cluster* harus ditentukan terlebih dahulu dengan jumlah yang optimal. Untuk menentukan jumlah *cluster* yang optimal, dua metode dapat dipakai, yaitu Elbow Method dan Silhouette Method. Berdasarkan kedua metode tersebut, gambar 8 dan gambar 9 menunjukkan jumlah *cluster* 2 merupakan yang paling optimal. Untuk melihat jumlah *cluster* yang optimal pada metode Elbow, titik yang membentuk "siku" menunjukkan jumlah *cluster* terbaik. Berdasarkan gambar 8, jumlah *cluster* 2 membentuk "siku" maka titik tersebut menunjukkan *cluster* yang paling optimal. Selain itu, untuk melihat jumlah *cluster* yang optimal pada metode Silhouette, titik puncak menunjukkan jumlah *cluster* yang paling optimal. Berdasarkan gambar 9, absis 2 menunjukkan titik puncak pada grafik tersebut. Oleh karena itu, jumlah *cluster* 2 merupakan jumlah yang paling optimal.



Gambar 8. Grafik Elbow Method



Gambar 9. Grafik Silhouette Method

Setelah mendapatkan jumlah *cluster* yang optimal, tahap membuat model K-Means Clustering dilakukan dengan menggunakan dua *cluster*. Jumlah iterasi yang terbentuk sampai pusat *cluster* tidak berganti adalah dua iterasi. Jumlah baris data yang termasuk ke dalam *cluster* 1 adalah 755 baris. Jumlah baris data yang termasuk ke dalam cluster 2 adalah 404 baris. Gambar 10 menunjukkan nilai centroid untuk setiap *cluster* dan kolom. Centroid merupakan titik pusat untuk setiap *cluster* terhadap masing-masing variabel.

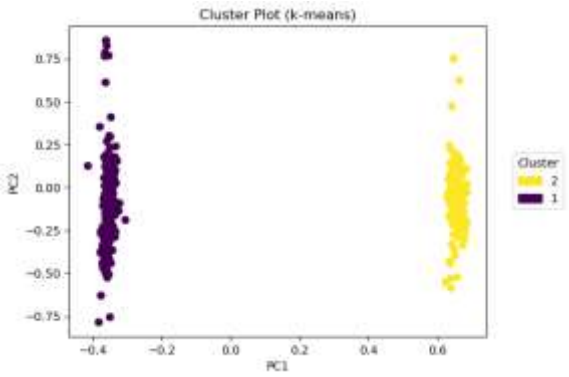
Jurnal Teknologi Dan Sistem Informasi Bisnis
Vol. 5 No. 3 Juli 2023 Hal. 172-182
http://jurnal.unidha.ac.id/index.php/jteksis

E-ISSN : **2655-8238**
P-ISSN : **2964-2132**

	Cluster 1	Cluster 2
Jenis Kelamin	5.551115e-17	-3.330669e-16
Umur Tahun	5.479452e-01	4.414730e-01
Tinggi	6.341880e-01	7.267934e-01
Berat Badan	4.051282e-01	4.721262e-01
Lama Sakit	8.888889e-02	7.106742e-02
IMT	2.550154e-01	2.587892e-01
Hasil IMT	3.703704e-01	3.333333e-01
Sistole	3.851852e-01	2.970037e-01
Diastole	2.015504e-01	1.731774e-01
Nafas	1.111111e-01	5.721910e-01
Detak Nadi	1.041667e-01	6.686271e-01

Gambar 10. Hasil Centroid Setiap *Cluster*

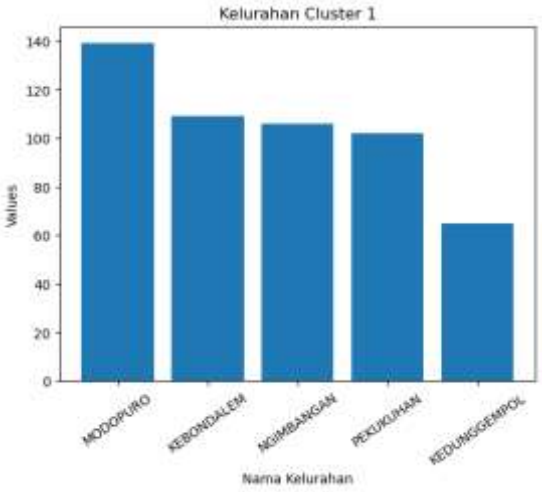
Setelah mendapatkan *cluster* untuk setiap baris, tahap selanjutnya adalah melakukan visualisasi dataset tersebut dengan dipisahkan terhadap *cluster*. Untuk dapat dilakukan visualisasi secara grafis, dibutuhkan dua variabel. Namun, dataset tersebut memiliki 11 kolom yang berarti tidak dapat divisualisasikan secara grafis. Oleh karena itu, metode Principal Component Analysis (PCA) digunakan untuk mengubah dimensi data menjadi dua dimensi. Gambar 11 menunjukkan hasil visualisasi grafis yang merupakan hasil dari reduksi dimensi PCA. Grafik tersebut menunjukkan kedua data dapat dipisahkan secara sempurna antara *cluster* 1 dan 2.



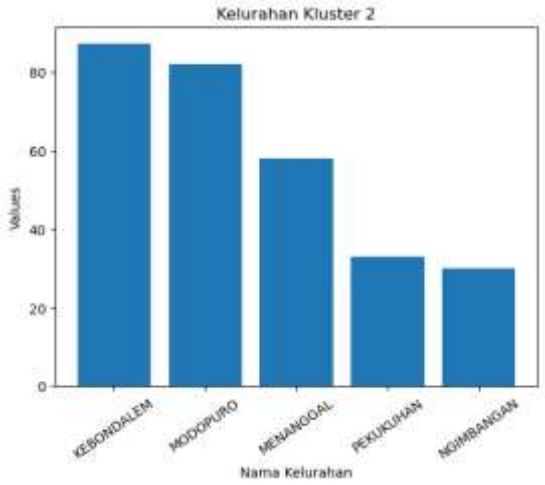
Gambar 11. Visualisasi Plot *Cluster* dengan PCA

Tahap selanjutnya adalah melakukan analisis data terhadap kedua *cluster* yang telah terbentuk. Analisis yang dilakukan adalah melihat 5 kelurahan terbanyak pada setiap *cluster* dan melihat 10 diagnosa terbanyak pada setiap *cluster*.

Kelurahan menunjukkan tempat tinggal pasien tersebut. Gambar 12 menunjukkan 5 kelurahan teratas yang terdapat pada *cluster* 1, yaitu Modopuro, Kebondalem, Ngimbangan, Pekukuhan, dan Kedunggempol. Berbeda dengan *cluster* 2, gambar 13 menunjukkan 5 kelurahan teratas yang terdapat pada *cluster* 2, yaitu Kebondalem, Modopuro, Menanggal, Pekukuhan, dan Ngimbangan. Kelurahan Modopuro adalah kelurahan terbanyak yang terdapat pada *cluster* 1. Namun, Kebondalem adalah kelurahan terbanyak yang terdapat pada *cluster* 2.



Gambar 12. Bar Plot Kelurahan Klaster 1



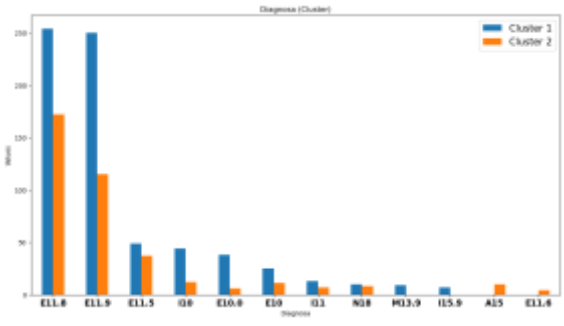
Gambar 13. Bar Plot Kelurahan Klaster 2

Diagnosa menunjukkan hasil diagnosis oleh tenaga medis terhadap pasien tersebut berdasarkan pemeriksaan yang dilakukan. Gambar 14 menunjukkan diagnosa E11.8 atau *Non-insulin-dependent diabetes mellitus with*

Jurnal Teknologi Dan Sistem Informasi Bisnis
Vol. 5 No. 3 Juli 2023 Hal. 172-182
http://jurnal.unidha.ac.id/index.php/jteksis

E-ISSN : **2655-8238**
P-ISSN : **2964-2132**

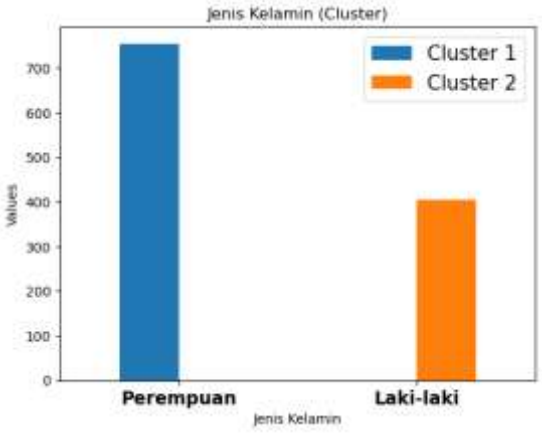
unspecified complications adalah diagnosa terbanyak pada *cluster* 1 dan *cluster 2*. Meskipun diagnosa terbanyak antara kedua *cluster* tersebut sama, grafik tersebut terdapat perbedaan diagnosa antara dua *cluster*. Diagnosa pada *cluster* 1 adalah M13.9 (*Arthritis, unspecified*) dan I15.9 (*Secondary hypertension, unspecified*). Diagnosa pada *cluster* 2 adalah A15 (*Respiratory tuberculosis, bacteriologically and histologically confirmed*) dan E11.6 (*Non-insulin-dependent diabetes mellitus with other specified complications*).



Gambar 14. Bar Plot Diagnosa

Tahap selanjutnya adalah analisis untuk masing-masing gejala unik dan esensial yang telah disebutkan pada masing-masing *cluster*. Pada diabetes tipe 2, hormon insulin tubuh tidak dapat bekerja dengan baik dan dikenal dengan *non-insulin dependent diabetes mellitus* (NIDDM). Hal ini disebabkan oleh berbagai kemungkinan, antara lain kekurangan produksi insulin, resistensi insulin atau kurangnya sensitivitas (daya tanggap) sel dan jaringan tubuh terhadap insulin, yang ditandai dengan peningkatan kadar insulin dalam darah [13]. Diagnosa E11.8 (*Non-insulin-dependent diabetes mellitus with unspecified complications)* adalah diabetes tipe 2 yang tidak memiliki komplikasi yang secara spesifik. Lalu, diabetes tipe 1 (DM), sebelumnya dikenal sebagai E10 atau *insulin-dependent diabetes mellitus* (IDDM), disebabkan oleh kerusakan sel beta pankreas (reaksi autoimun) yang merupakan produksi hormon insulin [14]. *Rheumatoid arthritis* (I15.9) adalah penyakit autoimun peradangan kronis atau reaksi autoimun di mana sistem kekebalan tubuh seseorang dapat menjadi tidak berfungsi dan melemah, menyebabkan kerusakan sendi dan lapisan sinovial, terutama tangan, kaki, dan lutut [18][19]. *Secondary hypertension* (I15.9) atau hipertensi sekunder adalah peningkatan tekanan darah yang diakibatkan oleh penyebab yang mendasari, dapat diidentifikasi, dan seringkali dapat diperbaiki tidak seperti hipertensi primer [15]. Tuberkulosis (TB) merupakan penyakit jenis menular yang disebabkan oleh bakteri TB (*Mycobacterium tuberculosis*) [16]. Dengan demikian, Diagnosa A15 (*Respiratory tuberculosis, bacteriologically and histologically confirmed*) merupakan diagnosis penyakit tuberkulosis yang didasarkan hasil tes bakteriologi dan histologi. Diagnosa E11.6 (*Non-insulin-dependent diabetes mellitus with other specified complications*) termasuk penyakit diabetes tipe II dengan komplikasi lain yang spesifik. Penyakit diabetes melitus tipe II lebih banyak ditemukan dibandingkan dengan diabetes melitus tipe I [18]. Hal ini linier dan serupa dengan hasil penelitian ini. Pada kedua *cluster*, diagnosa *Non-insulin-dependent diabetes mellitus* lebih banyak ditemukan daripada *Insulin-dependent diabetes mellitus*.

Untuk analisis lebih lanjut, setiap *cluster* dipetakan terhadap masing-masing jenis kelamin. Gambar 15 menunjukkan diagram bar untuk pemetaan setiap *cluster* terhadap masing-masing jenis kelamin. Diagram tersebut dapat disimpulkan bahwa seluruh pasien pada *cluster* 1 adalah perempuan dan seluruh pasien pada *cluster* 2 adalah laki-laki. Berdasarkan analisis klaster tersebut, *cluster* 1 tidak ditemukan pasien laki-laki dan tidak ditemukan pasien perempuan pada *cluster* 2.



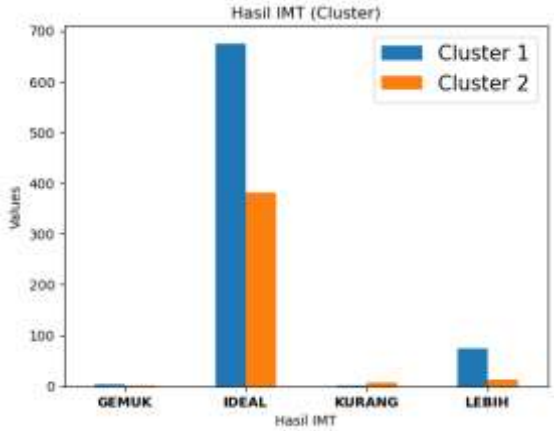
Gambar 15. Bar Plot Jenis Kelamin

Untuk analisis lebih dalam, setiap *cluster* dipetakan terhadap masing-masing hasil IMT. Gambar 16 menunjukkan diagram plot untuk pemetaan setiap *cluster* terhadap masing-masing hasil IMT. Diagram tersebut dapat disimpulkan bahwa seluruh pasien pada *cluster* 1 dan *cluster 2* sama-sama memiliki hasil IMT "ideal" untuk jumlah yang paling tinggi. Hasil IMT yang memiliki kategori "ideal", "gemuk", dan "lebih" adalah frekuensi yang paling tinggi pada *cluster* 1

Jurnal Teknologi Dan Sistem Informasi Bisnis
Vol. 5 No. 3 Juli 2023 Hal. 172-182
http://jurnal.unidha.ac.id/index.php/jteksis

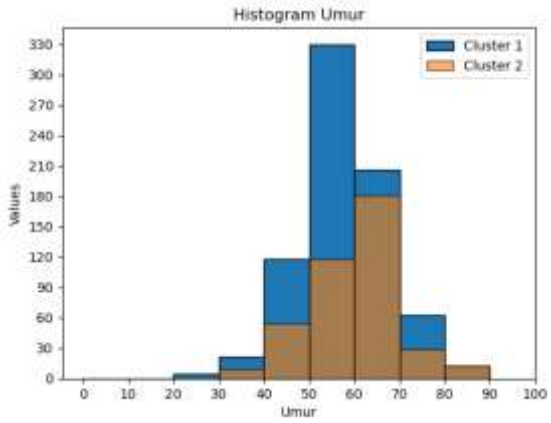
E-ISSN : **2655-8238**
P-ISSN : **2964-2132**

dibandingkan *cluster* 2. Namun, kategori "kurang" memiliki frekuensi yang paling tinggi pada *cluster* 2 dibandingkan *cluster* 1.



Gambar 16. Bar Plot Hasil IMT

Selain itu, untuk melihat distribusi umur pada setiap *cluster*, gambar 17 adalah diagram histogram umur untuk masing-masing *cluster*. Rentang umur 50-60 tahun memiliki frekuensi yang paling tinggi pada *cluster* 1. Namun, rentang umur 60-70 tahun memiliki frekuensi yang paling tinggi pada *cluster* 2. Rentang umur yang didominasi oleh *cluster* 1 adalah rentang umur 30-40 tahun, 40-50 tahun, 50-60 tahun, 60-70 tahun, dan 70-80 tahun. Rentang umur 20-30 tahun hanya terdapat pada *cluster* 1. Namun, rentang umur yang didominasi oleh *cluster* 2 adalah adalah rentang umur 80-90 tahun.



Gambar 17. Histogram Umur

Oleh karena itu, pemerintah setempat dengan puskesmas terkait yang berada di Mojokerto dapat melakukan sosialisasi, pencegahan, dan cara yang lebih spesifik dan khusus terhadap penyakit atau hasil diagnosis yang unik untuk setiap *cluster* untuk menekan jumlah penderita diabetes di wilayah setempat. Secara umum, penyakit diabetes melitus tipe II harus dilakukan pencegahan yang lebih ekstra karena memiliki hasil diagnosis yang terbanyak terhadap kedua *cluster* tersebut. Berdasarkan variabel-variabel pemetaan terkait untuk setiap *cluster*, pemerintah setempat dapat melakukan pertimbangan terhadap hasil pemetaan tersebut. Pemerintah dapat melakukan pertimbangan yang lebih lanjut untuk meningkatkan pelayanan puskesmas berdasarkan lima kelurahan dengan pasien terbanyak di puskesmas Mojokerto, seperti meningkatkan jumlah tenaga medis, fasilitas yang lebih memadai, dan sistem pelayanan yang dilakukan oleh puskesmas tersebut.

SIMPULAN

Dari penjelasan di atas dapat diambil kesimpulan penelitian ini adalah sebagai berikut:

Terdapat lima kelurahan teratas dan sepuluh diagnosa teratas pada setiap klaster. Kelurahan teratas pada klaster 1 adalah Modopuro, Kebondalem, Ngimbangan, Pekukuhan, dan Kedunggempol, sedangkan kelurahan teratas pada klaster 2 adalah Kebondalem, Modopuro, Menanggal, Pekukuhan, dan Ngimbangan. Hasil menunjukkan bahwa kelurahan Modopuro dan Kebondalem muncul sebagai kelurahan terbanyak pada kedua cluster. Frekuensi hasil Indeks Massa Tubuh (IMT) kategori "ideal" pada kedua cluster adalah yang tertinggi, namun kategori "kurang" lebih banyak ditemukan pada cluster 2. Diagnosa teratas pada klaster 1 dan cluster 2 adalah *Non-insulin-dependent diabetes mellitus with unspecified complications* (E11.8), dengan masing-masing diagnosa yang unik dalam cluster 1 adalah M13.9 dan I15, diagnosa yang unik dalam cluster 2 adalah A15 dan E11.6. Frekuensi hasil Indeks Massa Tubuh (IMT) kategori "ideal" pada kedua cluster adalah yang tertinggi, namun kategori "kurang" lebih banyak ditemukan pada cluster 2. Pemerintah setempat dan puskesmas di Mojokerto dapat melakukan sosialisasi dan pencegahan penyakit diabetes yang lebih spesifik untuk setiap cluster, untuk menekan jumlah penderita diabetes di wilayah tersebut. Pencegahan harus lebih ekstra terhadap diabetes tipe II karena memiliki hasil diagnosis terbanyak.

UCAPAN TERIMAKASIH

Kami ingin mengucapkan terima kasih yang sebesar-besarnya kepada dosen kami, Bu Endah Purwanti dan Bu Indah Werdiningsih sebagai *corresponding author* dalam penelitian ini. Kontribusi dan bantuan mereka dalam

Jurnal Teknologi Dan Sistem Informasi Bisnis
Vol. 5 No. 3 Juli 2023 Hal. 172-182
http://jurnal.unidha.ac.id/index.php/jteksis

E-ISSN : **2655-8238**
P-ISSN : **2964-2132**

pelaksanaan penelitian dan penulisan manuskrip ini sangatlah berarti dan membantu kami dalam menyelesaikan penelitian ini.

DAFTAR PUSTAKA

[1] American Diabetes Association. (2019). Diagnosis and Classification of Diabetes Mellitus. Diabetes Care, 42(Supplement 1), S13–S28. https://doi.org/10.2337/dc19-S002

[2] Catur Mei Astuti & Asih Setiarini. (2013). Faktor Yang Berhubungan Dengan Pengendalian Kadar Glukosa Darah Pasien Diabetes Militus Tipe 2 Rawat Jalan Di Poliklinik Penyakit Dalam RSJ Prof. Dr. Soerojo Magelang.

[3] Kementerian Kesehatan Republik Indonesia. (2020). Penyakit tidak menular di Indonesia 2020. Jakarta: Kementerian Kesehatan Republik Indonesia.

[4] Dinkes Kampung Besar Kota (2023). Tugas dan Fungsi Puskesmas

[5] Mulya, D. P. (2019). Analisa Dan Implementasi Association Rule Dengan Algoritma Fp-Growth Dalam Seleksi Pembelian Tanah Liat (Studi Kasus Di Pt. Anveve Ismi Berjaya). *Jurnal Teknologi Dan Sistem Informasi Bisnis*, *1*(1), 47-57.

[6] Hasanah, N. N., & Purnomo, A. S. (2022). Implementasi Data Mining Untuk Pengelompokan Buku Menggunakan Algoritma K-Means Clustering (Studi Kasus: Perpustakaan Politeknik LPP Yogyakarta). *Jurnal Teknologi Dan Sistem Informasi Bisnis*, *4*(2), 300-311.

[7] Inastri, M. A., & Mimba, N. P. S. H. (2017). Pengaruh Penerapan Good Corporate Governance dan Pengungkapan Corporate Social Responsibility pada Nilai Perusahaan. E-Jurnal Akuntansi Universitas Udayana, 21(2), 1400-1429.

[8] N. P. E. Merliana, Ernawati dan A. J. Santoso, "Analisa Penentuan Jumlah Cluster Terbaik pada Metode KMeans," UNISBANK , 2015.

[9] N. Ahmad, A. Mukharil, T. Informatika, F. Teknik, and U. K. Indonesia, "Jurnal Sistem Informasi (Journal of Information Systems). 2 / 12 (2016), 82-89 DOI : http://dx.doi.org/10.21609/jsi.v12i2.481," J. Sist. Inf. (Journal Inf. Syst., vol. 12, pp. 82–89, 2016.

[10] Tong, D.L., 2011. A Simpler Method Of Preprocessing MALDI-TOF MS Data For Differential Biomarker Analysis : Sistem Cell and Melanoma Cancer Studies. Clinical Proteomics, 8(14), pp.1-18.

[11] Kumar, V., & Chadha, A. (2012). Mining Association Rules in Student's Assessment Data. International Journal of Computer Science Issues , 9, 211-216. M. C., L. C., & D., A. K. (2012). Market Basket Analysis for a Supermarket based on Frequent Itemset Mining . International Journal of Computer Science Issues , 257-264.

[12]Agusta, Yudhi. 2007. 'K-Means penerapan permasalahan dan metode terkait'. Jurnal Sistem dan Informatika, Vol 3.

[13] Manurung, R. (2017). Gambaran Tugas Keluarga di Bidang Kesehatan pada Penanganan Klien Diabetes Melitus Tipe II di Lingkungan I Kelurahan Dolok Tenera Kecamatan Dolok Batu Nanggar Kabupaten Simalungun Tahun 2012. Jurnal Ilmiah Keperawatan Imelda, 3(1), 61-66.

[14]Marzel, R. (2021). Terapi pada DM Tipe 1. Jurnal Penelitian Perawat Profesional, 3(1), 51-62.

[15]Onusko, E. M. (2003). Diagnosing secondary hypertension. American family physician, 67(1), 67-74.

[16]Pratiwi, R. D. (2020). Gambaran Komplikasi Penyakit Tuberkulosis Berdasarkan Kode International Classification of Disease 10. Jurnal Kesehatan Al-Irsyad Vol. 13 No, 2.

[17]Betteng, R. (2014). Analisis faktor resiko penyebab terjadinya Diabetes Melitus tipe 2 pada wanita usia produktif Dipuskesmas Wawonasa. *e-Biomedik*, *2*(2).

[18]Sakti, N. P. R., & Muhlisin, A. (2019). Pengaruh Terapi Komplementer Meditasi terhadap Respon Nyeri pada Penderita Rheumathoid Arthtritis. *The 9th University Research Colloqium (Urecol)*, 9(1).

[19]Masruroh, A. N., & Muhlisin, A. (2020). *Gambaran Sikap dan Upaya Keluarga dalam Merawat Anggota Keluarga yang Menderita Rheumatoid Arthtritis di Desa Mancasan Wilayah Kerja Puskesmas Baki Kabupaten Sukoharjo*. *Universitas Muhammadiyah Surakarta*.

[20]Safitri, W. R. (2016). Analisis Korelasi Pearson Dalam Menentukan Hubungan Antara Kejadian Demam Berdarah Dengue dengan Kepadatan Penduduk di Kota Surabaya Pada Tahun 2012-2014: Pearson Correlation Analysis to Determine The Relationship Between City Population Density with Incident Dengue Fever of Surabaya in The Year 2012-2014. *Jurnal Ilmiah Keperawatan (Scientific Journal of Nursing)*, *2*(2), 21-29.

[21]Roflin, E., & Zulvia, F. E. (2021). *Kupas tuntas analisis korelasi*. Penerbit NEM.

[22] Parasian D.P Silitonga. (2017). Clustering of Patient Disease Data by Using K-Means Clustering. International Journal of Computer Science and Information Security, 15(7). 219-221