LAPORAN PENELITIAN KELOMPOK

Analisis Persentase Penduduk yang Mempunyai Keluhan Kesehatan Menurut Provinsi, Tipe Daerah, dan Jenis Kelamin dengan Menggunakan Algoritma Clustering K-Means (2020-2022)

Disusun untuk memenuhi tugas mata kuliah “Big Data”

Dosen Pengampu :

Dr.Ir.Ananto Tri Sasongko, M.Sc.,

Disusun Oleh :

Abdul Aziz Firdaus (312110262)

Muhammad Akbar (312110483)

Muhammad Arya Dipanegara Gunawan (312110396)

Ridho Pikriyansyah (312110169)

FAKULTAS TEKNIK

PROGRAM STUDI TEKNIK INFORMATIKA

UNIVERSITAS PELITA BANGSA

BEKASI

2023

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB I

PENDAHULUAN

1.1 Latar Belakang

Keluhan kesehatan merupakan gangguan terhadap kondisi fisik dan jiwa, dapat disebabkan oleh berbagai faktor seperti kecelakaan atau hal lain yang menghambat kegiatan sehari-hari. Di antara keluhan kesehatan yang umum dialami oleh penduduk termasuk panas, sakit kepala, batuk, pilek, diare, asma/sesak nafas, dan sakit gigi. Meskipun seseorang dengan penyakit kronis tidak mengalami gejala selama survei satu bulan terakhir, tetap dianggap memiliki keluhan kesehatan.

Upaya untuk menanggulangi masalah keluhan kesehatan mencakup berbagai strategi seperti peningkatan gizi, penambahan fasilitas kesehatan, pelaksanaan imunisasi, pelayanan kesehatan gratis, pengadaan obat generik, peningkatan jumlah tenaga medis, serta penyuluhan mengenai kebersihan dan pola hidup sehat. Kualitas pelayanan kesehatan, termasuk penanganan keluhan masyarakat, dipengaruhi oleh ketersediaan sarana dan prasarana, termasuk alokasi dana.

Undang-Undang Nomor 36 Tahun 2009 tentang Kesehatan memberikan landasan hukum bagi upaya kesehatan yang melibatkan pemerintah dan masyarakat. Upaya kesehatan, seperti pencegahan penyakit, peningkatan kesehatan, pengobatan penyakit, dan pemulihan kesehatan, harus dilakukan secara terpadu, terintegrasi, dan berkesinambungan. Selain itu, setiap individu memiliki hak untuk lingkungan yang sehat, informasi dan edukasi kesehatan yang seimbang, serta akses informasi tentang data kesehatan diri termasuk tindakan dan pengobatan yang diterima dari tenaga kesehatan.

Penelitian ini menggunakan data yang diambil dari dokumen Badan Pusat Statistik Nasional melalui situs resmi mereka. Metode analisis yang diterapkan adalah *clustering*, yang bertujuan memisahkan penduduk berdasarkan keluhan kesehatan menurut provinsi. Hasil dari proses *clustering* dibagi menjadi tiga kelompok, yaitu keluhan kesehatan tinggi, sedang, dan rendah. Informasi ini dapat menjadi masukan penting bagi pemerintah, sehingga provinsi yang masuk dalam

kelompok tinggi mendapatkan perhatian lebih guna meningkatkan kualitas pelayanan kesehatan di wilayah tersebut. Analisis cluster menjadi alat statistik yang efektif untuk memahami perbedaan antar kelompok dalam konteks keluhan kesehatan penduduk.[1]

1.2 Tujuan Penelitian

Dengan merinci latar belakang yang telah dibahas sebelumnya, tujuan penelitian ini dirumuskan sebagai berikut:

1. Menganalisis keluhan kesehatan penduduk.
2. Mengidentifikasi provinsi dengan keluhan kesehatan tinggi.
3. Mengelompokkan penduduk menurut tingkat keluhan kesehatan.
4. Mengidentifikasi keluhan antar jenis kelamin.
5. Mengidentifikasi keluhan antar kota dan desa.
6. Mengidentifikasi keluhan kenaikan dan penurunan dalam waktu 3 tahun.

1.3 Manfaat Penelitian

Berikut adalah manfaat penelitian dari tujuan-tujuan yang disebutkan sebelumnya:

1. Memberikan pemahaman mendalam tentang jenis keluhan kesehatan yang paling umum di masyarakat.
2. Memungkinkan pengalokasian sumber daya kesehatan secara lebih efektif dengan fokus pada provinsi-provinsi yang membutuhkan perhatian lebih besar untuk menangani keluhan kesehatan tinggi.
3. Memberikan gambaran yang lebih rinci tentang tingkat keluhan kesehatan di berbagai kelompok penduduk.
4. Mendukung pengembangan program kesehatan yang lebih terpersonalisasi dengan memperhatikan perbedaan keluhan kesehatan antara laki-laki dan perempuan.
5. Memahami perbedaan keluhan kesehatan antara masyarakat perkotaan dan pedesaan.
6. Memberikan gambaran dinamis tentang perubahan dalam keluhan kesehatan dari waktu ke waktu.

BAB II
TINJAUAN PUSTAKA

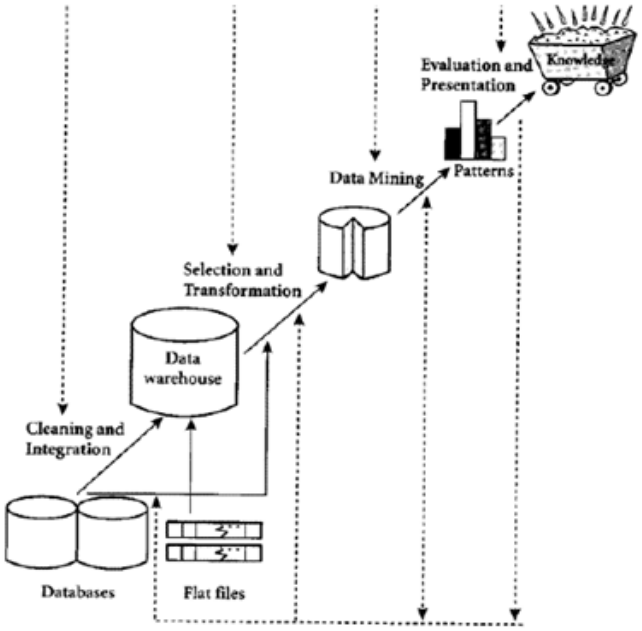
2.1 Data Mining

Menurut Wahyudi, Dkk (2022 : 9) data mining adalah proses untuk mendapatkan informasi yang berguna dari basis data yang besar dan perlu diekstraksi agar menjadi informasi baru dan dapat membantu dalam pengambilan keputusan. Menurut Ma'rifatin (2020 : 286) data mining merupakan metode yang digunakan dalam pengolahan data berskala besar oleh karena itu data mining memiliki peranan yang sangat penting dalam beberapa bidang kehidupan diantaranya yaitu bidang industri, bidang keuangan, cuaca, ilmu dan teknologi. Berdasarkan dari definisi diatas dapat disimpulkan bahwa pengertian data mining adalah gabungan dari beberapa disiplin ilmu yang menyatukan teknik dari pembelajaran mesin, pengenalan pola, statistik, database, dan visualisasi untuk penanganan permasalahan pengambilan informasi daribasis data yang besar.Data mining bukanlah suatu bidang yang sama sekali baru. Salah satu kesulitan untuk mendefinisikan data mining adalah kenyataan bahwa data mining mewarisi banyak aspek dan teknik dari bidang-bidang ilmu yang sudah mapan terlebih dahulu. Berawal dari beberapa disiplin ilmu, data mining bertujuan untuk memperbaiki teknik tradisional sehingga bisa menangani :

1. Jumlah data yang sangat besar.
2. Dimensi data yang tinggi.
3. Data yang heterogen dan berbeda sifat.[2]

2.1.1 Tahap - Tahap Data Mining

Istilah *data mining* dan *knowledge discovery in database* (KDD) seringkali digunakan secara bergantian untuk menjelaskan proses penggalian informasi tersembunyi dalam suatu basis data yang besar. Walaupun sebenarnya *data mining* sendii adalah bagian dari tahapan proses dalam KDD. Proses KDD secara garis besar dapat dilihat pada Gambar 1 berikut : [3]



Gambar 1. Tahap - Tahap Data Mining

a. *Data cleaning*, untuk membersihkan data dari noise data dan data yang tidak konsisten. Proses *cleaning* mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan mempebaiki kesalahan pada data
b. *Data integration*, mengkombinasikan atau mengintegrasikan beberapa sumber data.
c. *Data selection*, mengambil data-data yang relevan dari database untuk dianalisis.
d. *Data transformation*, mentransformasikan data *summary* ataupun operasi agregasi.
e. *Data mining*, merupakan proses yang esensial dimana metode digunakan untuk mengekstrak pola data yang tersembunyi dengan menggunakan Teknik atau metode tetentu. Teknik, metode, atau algoritma dalam *data mining* sangat bevariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.
f. *Interpretation / Evaluasi*, pola informasi yang dihasilkan dari proses *data mining* perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini mencakup pemeriksaan

apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya.[3]

2.1.2 Teknik Data Mining

Terdapat beberapa teknik data mining yang sering disebut-sebut dalam literatur. Namun ada 3 teknik data mining yaitu:

1. Association

   Rule MiningAssociation Rule mining adalah teknik mining untuk menemukan asosiatif antara kombinasi atribut. Contoh dari aturan asosiatif dari analisa pembelian disuatu pasar swalayan dapat mengatur penempatan barangnya atau merancang strategi pemasaran dengan memakai kupon diskon untuk kombinasi barang tertentu.
2. Clustering

   Berbeda dengan association rule mining dan klasifikasi dimana kelas data telah ditentukan sebelumnya, clustering dapat dipakai untuk memberikan label pada kelas data yang belum diketahui. Karena itu clustering sering digolongkan sebagai metode unsupervised learning. Prinsip clustring adalah memaksimalkan kesamaan antar cluster. Clustering dapat dilakukan pada data yang memiliki beberapa atribut yang dipetakan sebagai ruang multidimensi.
3. Klasifikasi

   alam klasifikasi, terdapat target variabel kategori. Sebagai contoh, penggolongan pendapatan dapat dipisahkan dalam tiga kategori, yaitu pendapatan tinggi, pendapatan sedang, pendapatan rendah.[2]

2.2 Algoritma Klasifikasi K-Means

K-Means merupakan algoritma clustering yang berulang-ulang. Algoritma K-Means dimulai dengan pemilihan secara acak *K*, *K* disini merupakan banyaknya Cluster yang ingin dibentuk. Kemudian tetapkan nilai - nilai *K* secara random, untuk sementara nilai tersebut menjadi pusat dari cluster atau biasa disebut dengan

centroid, meanatau“means”. Hitung jarak setiap data yang ada terhadap masing - masing centroid menggunakan rumus Euclidian hingga ditemukan jarak yang paling dekat dari setiap data dengan centroid. Klasifikasikan setiap data berdasarkan kedekatannya dengan centroid. Lakukan langkah tersebut hingga nilai centroid tidak berubah (stabil).[4]

2.3 Clustering

Menurut Widodo (2013:9) Clustering atau klasifikasi adalah metode yang digunakan untuk membagi rangkaian data menjadi beberapa group berdasarkan kesamaan-kesamaan yang telah ditentukan sebelumnya. Cluster adalah sekelompok atau sekumpulan objek - objek data yang similar satu sama lain dalam cluster yang sama dan dissimilar terhadap objek - objek yang berbeda cluster. Objek akan dikelompokkan kedalam satu atau lebih cluster sehingga objek - objek yang berada dalam satu cluster akan mempunyai kesamaan yang tinggi antara satu dengan yang lainnya.

Objek - objek dikelompokkan berdasarkan prinsip memaksimalkan kesamaan objek pada cluster yang sama dan memaksimalkan ketidaksamaan pada cluster yang berbeda. Kesamaan objek biasanya diperoleh dari nilai - nilai atribut yang menjelaskan objek data, sehingga objek - objek data biasanya dipresentasikan sebagai sebuah titik dalam ruang multidemensi. Dengan menggunakan clustering ini, kita dapat mengkalsifikasikan daerah yang padat, menemukan pola - pola distribusi secara keseluruhan, dan menemukan keterkaitan yang menarik antara atribut data. Dalam data mining, usaha difokuskan pada metode - metode penemuan untuk cluster pada basis data berukuran besar secara efektif dan efisien.

Beberapa kebutuhan clustering dalam data mining meliputi skalabilitas, kemampuan untuk menangani tipe atribut yang berbeda mampu menangani dimensionalitas yang tinggi, menangani data yang mempunyai noise, dan dapat diterjemakan dengan mudah. Adapun tujuan dari data clustering ini adalah untuk meminimalisasikan objektif function yang di-set dalam proses clustering, yang pada umumnya berusaha meminimalisasikan variasi dalam suatu cluster. Dan meminimalisasikan variasi antar cluster. Secara garis besar, terdapat beberapa

metode klasifikasi data.Pemilihan metode clustering tergantung pada tipe data dan tujuan clustering itu sendiri.[4]

2.4 Metode K-Means

K-Means merupakan salah satu metode data clustering non hierarki yang berusaha mempartisi data yang ada ke dalam bentuk satu atau lebih cluster atau kelompok sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam satu cluster yang sama dan data yang mempunyai karakteristik yang berbeda dikelompokkan kedalam kelompok yang lainnya. K-Means adalah metode clustering berbasis jarak yang membagi data ke dalam sejumlah cluster dan algoritma ini hanya bekerja pada atribut numeric.

Algoritma K-Means termasuk partitioning clustering yang memisahkan data ke daerah bagian yang terpisah. Algoritma K-Means sangat terkenal karena kemudahan dan kemampuannya untuk meng-cluster data yang besar dan data outlier dengan sangat cepat. Dalam algoritma K-Means, setiap data harus termasuk ke cluster tertentu dan bisa dimungkinkan bagi setiap data yang termasuk cluster tertentu pada suatu tahapan proses, pada tahapan berikutnya berpindah ke cluster lainnya.

Pada dasarnya penggnaan algoritma dalam melakukan proses clustering tergantung dari data yang ada dan konklusi yang ingin dicapai. Untuk itu digunakan algoritma K-Means yang didalamnya membuat aturan sebagai berikut

1) Jumlah Cluster perlu diinputkan.
2) Hanya memiliki atribut bertipe numeric.

Algoritma K-Means merupakan metode non-hierarki yang pada awalnya mengambil sebagian banyaknya komponen populasi untuk dijadikan pusat cluster awal. Pada tahap ini pusat cluster dipilih secara acak dari sekumpulan populasi data. Berikutnya K-Means menguji masing - masing komponen di dalam populasi data dan menandai komponen tersebut kesalahsatu pusat cluster yang telah didefinisikan tergantung dari jarak minimum antar komponen dengan tiap - tiap cluster. Posisi pusat cluster akan dihitung kembali sampai semua komponen data digolongkan ke

dalam tiap - tiap pusat cluster dan terakhir akan terbentuk posisi pusat cluster yang baru.

Algoritma K-Means pada dasarnya melakukan dua proses, yakni proses pendeteksian lokasi pusat tiap cluster dan proses pencarian anggota dari tiap - tiap cluster. Cara kerja algoritma K-Means:

1. Tentukan *K* sebagai jumlah cluster yang ingin di bentuk.
2. Bangkitkan *K* centroid (titik pusat cluster) awal secara random.
3. Hitung jarak setiap data ke masing - masing centroid.
4. Setiap data memilih centroid yang terdekat.
5. Tentukan posisi centroid yang baru dengan cara menghitung nilai rata - rata dari data-data yang terletak pada centroid yang sama.
6. Kembali kelangkah-3 jika posisi centrroid baru dengan centroid yang lama tidak sama.[4]

BAB III

METODE PENELITIAN

3.1 Jenis Penelitian

Penelitian ini merupakan jenis penelitian terapan di bidang kesehatan yang bertujuan untuk menganalisis pola keluhan kesehatan penduduk berdasarkan provinsi, tipe daerah, dan jenis kelamin dalam rentang waktu 2020-2022. Dalam konteks ini, penelitian menggunakan metode clustering K-Means sebagai pendekatan analisis data. Fokus utama penelitian ini adalah mengidentifikasi kelompok atau klaster penduduk yang memiliki pola keluhan kesehatan serupa, dengan variabel provinsi, tipe daerah, dan jenis kelamin sebagai faktor utama dalam pembentukan klaster. Data keluhan kesehatan akan dikumpulkan dari sumber-sumber yang dapat dipercaya, kemudian diolah dan disiapkan untuk analisis.

Langkah selanjutnya melibatkan implementasi algoritma clustering K-Means untuk mengelompokkan penduduk berdasarkan pola keluhan kesehatan mereka. Hasil clustering akan dianalisis untuk memahami perbedaan pola keluhan kesehatan antar klaster, dengan mempertimbangkan faktor geografis dan demografis yang menjadi fokus penelitian.

3.2 Pengumpulan Data

Penelitian ini dimulai dengan mengumpulkan data keluhan kesehatan penduduk selama periode tahun 2020-2022. Sumber data mencakup lembaga kesehatan, survei kesehatan, dan statistik resmi. Variabel yang diambil meliputi jenis kelamin, provinsi, tipe daerah (kota dan desa), dan detail keluhan kesehatan. Data ini kemudian diproses dan dibersihkan untuk memastikan keakuratan dan konsistensi. Proses ini mencakup langkah-langkah pembersihan data untuk menangani nilai-nilai yang hilang atau tidak valid.

3.3 Metode Analisis Data

Setelah mengumpulkan data keluhan kesehatan penduduk dan menerapkan analisis deskriptif untuk mengelompokkan berdasarkan pola keluhan, metode analisis dilakukan dengan menghitung nilai rata-rata keluhan kesehatan untuk setiap tahun (2020-2022) dalam setiap klaster. Analisis melibatkan perbandingan rata-rata (mean) antar provinsi, jenis kelamin, kota, dan desa. Pengujian tren kenaikan/penurunan juga dilakukan untuk mengevaluasi perubahan statistik dalam pola keluhan kesehatan selama periode tersebut.

BAB IV

HASIL DAN ANALISA

4.1 Data Penelitian

Data dalam penelitian ini diambil dari hasil presentase data penduduk keluhan kesehatan menurut provinsi, tipe daerah dan jenis kelamin dari tahun 2020 - 2022 yang terdiri dari 34 provinsi. Data ini diambil dari badan pusat statistik indonesia https://www.bps.go.id/id untuk dapat mengidentifikasi pola - pola penting dan perubahan signifikan dalam kondisi kesehatan masyarakat.

Adapun data penelitian yang didapatkan dari badan pusat statistik indonesia, dinataranya:

Tabel 1. Data Penelitian

P R O V I N S I	2020		
	LAKI-LAKI		
	Perkotaan	Perdesaan	Perkotaan+Perdesaan
Aceh	24.68	24.91	24.84
Sumatera Utara	23.21	26.4	24.64
Sumatera Barat	30.46	27.02	28.65
Riau	25.3	22.42	23.6
Jambi	19.75	20.78	20.44
Sumatera Selatan	29.01	27.28	27.93
Bengkulu	28.76	28.94	28.88
Lampung	28.82	30.07	29.68
Kepulauan Bangka Belitung	31.13	26.27	28.99
Kepulauan Riau	16.37	18.85	16.61
DKI Jakarta	31.85	0	31.85
Jawa Barat	30.04	31.96	30.46
Jawa Tengah	33.94	33.74	33.84
DI Yogyakarta	35.6	36.46	35.82
Jawa Timur	31.14	31.48	31.29
Banten	28.37	34.56	30.08
Bali	23.18	29.07	24.94
Nusa Tenggara Barat	41.21	42.69	41.96
Nusa Tenggara Timur	33.19	32.95	33.01
Kalimantan Barat	20.76	26.23	24.3

Kalimantan Tengah	26.29	23.8	24.82
Kalimantan Selatan	31.9	33.77	32.88
Kalimantan Timur	22.31	19.89	21.52
Kalimantan Utara	26.02	20.64	23.93
Sulawesi Utara	22.44	25.08	23.69
Sulawesi Tengah	25.19	24.82	24.93
Sulawesi Selatan	28.03	26.44	27.14
Sulawesi Tenggara	28.58	27.58	27.98
Gorontalo	30.08	29.52	29.76
Sulawesi Barat	24.82	24.69	24.72
Maluku	15.84	17.47	16.75
Maluku Utara	15.41	14.48	14.75
Papua Barat	17.74	22.23	20.32
Papua	16.79	15.85	16.13
INDONESIA	29.36	29.2	29.29
P R O V I N S I	PEREMPUAN		
	Perkotaan	Perdesaan	Perkotaan+Perdesaan
Aceh	28.95	29.61	29.4
Sumatera Utara	26.54	28.81	27.59
Sumatera Barat	36.32	31.08	33.52
Riau	28.32	24.11	25.78
Jambi	21.88	23.35	22.87
Sumatera Selatan	33.54	29.1	30.75
Bengkulu	32.22	31.36	31.64
Lampung	31.24	33.93	33.1
Kepulauan Bangka Belitung	35.91	31.05	33.69
Kepulauan Riau	19.81	20.41	19.87
DKI Jakarta	35.75	0	35.75
Jawa Barat	33.25	34.94	33.66
Jawa Tengah	37.14	37.64	37.39
DI Yogyakarta	39.54	42.15	40.27
Jawa Timur	34.22	34.34	34.28
Banten	32.86	38.32	34.45
Bali	24.09	30.13	26.03
Nusa Tenggara Barat	45.14	46.65	45.93
Nusa Tenggara Timur	33.62	36.53	35.85
Kalimantan Barat	23.28	29.65	27.43
Kalimantan Tengah	30.63	26.63	28.22
Kalimantan Selatan	36.71	37.14	36.94
Kalimantan Timur	25.31	20.99	23.89
Kalimantan Utara	29.25	27.15	28.41
Sulawesi Utara	25.13	27.61	26.33
Sulawesi Tengah	28.91	26.84	27.45

Sulawesi Selatan	31.78	29.71	30.6
Sulawesi Tenggara	30.77	30.34	30.51
Gorontalo	32.8	35.96	34.63
Sulawesi Barat	29.7	27.32	27.87
Maluku	20.29	20.49	20.41
Maluku Utara	18.33	16.82	17.25
Papua Barat	21.28	23.65	22.65
Papua	17.51	16.02	16.43
INDONESIA	32.79	32.46	32.65

P R O V I N S I	2021		
	LAKI-LAKI		
	Perkotaan	Perdesaan	Perkotaan+Perdesaan
Aceh	24.29	21.81	22.65
Sumatera Utara	18.08	20.15	18.99
Sumatera Barat	25.04	20.39	22.7
Riau	31.57	15.8	22.19
Jambi	12.44	17.35	15.73
Sumatera Selatan	33.59	22.68	26.78
Bengkulu	36.05	20.77	25.86
Lampung	36.48	23.11	27.38
Kepulauan Bangka Belitung	29.46	20.66	25.69
Kepulauan Riau	13.95	11.86	13.8
DKI Jakarta	25.35	0	25.35
Jawa Barat	30.47	24.07	29.13
Jawa Tengah	32.06	25.16	28.79
DI Yogyakarta	31.64	23.16	29.51
Jawa Timur	30.91	22.89	27.35
Banten	30.37	20.64	27.66
Bali	24.18	20.84	23.23
Nusa Tenggara Barat	38.06	41.7	39.87
Nusa Tenggara Timur	25.3	29.13	28.16
Kalimantan Barat	18.31	20.74	19.86
Kalimantan Tengah	26.61	19.19	22.29
Kalimantan Selatan	31.21	29.5	30.34
Kalimantan Timur	24.04	13.73	20.82
Kalimantan Utara	28.71	18.21	24.76
Sulawesi Utara	25.45	16.26	21.24
Sulawesi Tengah	32.06	21.41	24.73
Sulawesi Selatan	34.82	19.42	26.44
Sulawesi Tenggara	31.27	20.61	24.51
Gorontalo	33.67	25.23	28.93
Sulawesi Barat	31.49	22.54	24.38
Maluku	15.08	14.47	14.74

Maluku Utara	12.87	15.35	14.63
Papua Barat	25.98	15.59	20.08
Papua	12.34	12.33	12.33
INDONESIA	28.91	22.43	26.15
P R O V I N S I	PEREMPUAN		
	Perkotaan	Perdesaan	Perkotaan+Perdesaan
Aceh	29.78	26.62	27.66
Sumatera Utara	20.87	23.18	21.92
Sumatera Barat	28.92	23.85	26.31
Riau	31.38	18.19	23.4
Jambi	14.46	19.66	17.96
Sumatera Selatan	36.53	24.61	29.08
Bengkulu	35.87	23.31	27.44
Lampung	37.37	25.87	29.54
Kepulauan Bangka Belitung	29.69	22.85	26.65
Kepulauan Riau	15.66	15.63	15.66
DKI Jakarta	26.61	0	26.61
Jawa Barat	31.32	27.16	30.37
Jawa Tengah	33.56	27.96	30.8
DI Yogyakarta	32.26	27.08	30.88
Jawa Timur	33.15	25.73	29.72
Banten	31.27	24.08	29.19
Bali	25.4	20.88	24.02
Nusa Tenggara Barat	42	46.64	44.36
Nusa Tenggara Timur	29.19	33.01	32.09
Kalimantan Barat	22.13	23.22	22.83
Kalimantan Tengah	27.32	19.8	22.88
Kalimantan Selatan	35.12	32.45	33.73
Kalimantan Timur	24.96	15.07	21.85
Kalimantan Utara	29.33	19.84	25.65
Sulawesi Utara	26.97	18.31	22.97
Sulawesi Tengah	33.11	23.88	26.73
Sulawesi Selatan	36.85	23.76	29.56
Sulawesi Tenggara	32.97	23.91	27.09
Gorontalo	38.39	29.15	33.19
Sulawesi Barat	34.21	25.63	27.36
Maluku	19.25	17.28	18.14
Maluku Utara	15.15	17.35	16.73
Papua Barat	27	17.34	21.46
Papua	14.2	12.8	13.2
INDONESIA	30.58	25.43	28.32

P R O V I N S I	2022		
	LAKI-LAKI		
	Perkotaan	Perdesaan	Perkotaan+Perdesaan
Aceh	23.9	31.03	28.6
Sumatera Utara	26.24	28.96	27.42
Sumatera Barat	15.2	36.02	25.92
Riau	22	28.68	26.05
Jambi	15.66	29.25	24.81
Sumatera Selatan	30	31.1	30.7
Bengkulu	35.25	21.77	26.15
Lampung	22.29	34.66	30.65
Kepulauan Bangka Belitung	32.95	33.52	33.2
Kepulauan Riau	16.39	24.78	17.44
DKI Jakarta	14.66	0	14.66
Jawa Barat	27.28	36.84	29.32
Jawa Tengah	30.67	34.08	32.31
DI Yogyakarta	31.17	38.34	32.98
Jawa Timur	27.31	33.87	30.23
Banten	18.66	30.85	21.73
Bali	17.15	24.18	19.14
Nusa Tenggara Barat	38.47	45.55	41.98
Nusa Tenggara Timur	35.96	25.63	28.35
Kalimantan Barat	15.52	28.99	24.15
Kalimantan Tengah	20.6	24.68	23
Kalimantan Selatan	23.48	33.82	28.76
Kalimantan Timur	22.08	30.19	24.67
Kalimantan Utara	33.37	36.85	34.66
Sulawesi Utara	19.18	26.77	22.73
Sulawesi Tengah	23.02	21.31	21.84
Sulawesi Selatan	25.29	36.49	31.38
Sulawesi Tenggara	25.6	31.28	29.23
Gorontalo	34.36	33.44	33.84
Sulawesi Barat	31.72	27.45	28.29
Maluku	12.62	14.29	13.56
Maluku Utara	11.28	16.45	14.98
Papua Barat	16.8	20	18.67
Papua	13.27	9.45	10.52
INDONESIA	25.15	31.67	27.92
P R O V I N S I	PEREMPUAN		
	Perkotaan	Perdesaan	Perkotaan+Perdesaan
Aceh	30.42	36.83	34.66

Sumatera Utara	30.6	34.26	32.17
Sumatera Barat	19.38	41.7	30.69
Riau	24.6	30.92	28.33
Jambi	17.63	32.34	27.4
Sumatera Selatan	34.58	33.54	33.94
Bengkulu	37.54	23.18	28.07
Lampung	27.19	38.55	34.76
Kepulauan Bangka Belitung	37.65	37.74	37.69
Kepulauan Riau	18.06	29.91	19.34
DKI Jakarta	17.55	0	17.55
Jawa Barat	31.19	40.39	33.17
Jawa Tengah	36.5	39.13	37.75
DI Yogyakarta	36.25	45.11	38.59
Jawa Timur	30.92	37.48	33.88
Banten	22.24	35.8	25.58
Bali	18.82	26.06	20.92
Nusa Tenggara Barat	41.71	50.83	46.21
Nusa Tenggara Timur	42.57	28.63	32.25
Kalimantan Barat	17.26	31.85	26.3
Kalimantan Tengah	23.95	27.74	26.1
Kalimantan Selatan	29.64	38.84	34.31
Kalimantan Timur	26.05	35.22	28.78
Kalimantan Utara	35.5	38.48	36.53
Sulawesi Utara	20.87	28.5	24.3
Sulawesi Tengah	23.07	24.09	23.75
Sulawesi Selatan	28.96	41.92	36.01
Sulawesi Tenggara	30.08	35.23	33.32
Gorontalo	39.49	39.07	39.26
Sulawesi Barat	37.35	32.61	33.6
Maluku	17.41	16.55	16.94
Maluku Utara	15.82	19.99	18.74
Papua Barat	21	22.01	21.59
Papua	14.58	9.93	11.21
INDONESIA	29.15	35.79	31.94

4.2 Analisis Data

Setelah data yang diperoleh sebagai hasil amatan yang disajikan dalam format tabel, maka akan dilakukan proses klasifikasi data dengan menggunakan nilai rata - rata (mean). Berikut hasil perhitungannya:

Nilai Mean Laki - Laki

Tahun 2020 = Perkotaan + Perdesaan + (Perkotaan + Perdesaan)

$$\text{Rata - Rata} = \frac{Nilai\ 1 + Nilai\ 2 + Nilai\ 3}{N}$$

$$\text{Rata - Rata} = \frac{29.36 + 29.2 + 29.29}{3}$$

$$\text{Rata - Rata} = \frac{87.85}{3} = 29.28$$

Tahun 2021 = Perkotaan + Perdesaan + (Perkotaan + Perdesaan)

$$\text{Rata - Rata} = \frac{Nilai\ 1 + Nilai\ 2 + Nilai\ 3}{N}$$

$$\text{Rata - Rata} = \frac{28.91 + 22.43 + 26.15}{3}$$

$$\text{Rata - Rata} = \frac{77.49}{3} = 25.83$$

Tahun 2022 = Perkotaan + Perdesaan + (Perkotaan + Perdesaaan)

$$\text{Rata - Rata} = \frac{Nilai\ 1 + Nilai\ 2 + Nilai\ 3}{N}$$

$$\text{Rata - Rata} = \frac{25.15 + 31.67 + 27.92}{3}$$

$$\text{Rata - Rata} = \frac{84.74}{3} = 28.25$$

Nilai Mean Perempuan

Tahun 2020 = Perkotaan + Perdesaan + (Perkotaan + Perdesaan)

$$\text{Rata - Rata} = \frac{Nilai\ 1 + Nilai\ 2 + Nilai\ 3}{N}$$

Rata - Rata = $\frac{32.79 + 32.46 + 32.65}{3}$

Rata - Rata = $\frac{97.9}{3}$ = 32.63

Tahun 2021 = Perkotaan + Perdesaan + (Perkotaan + Perdesaan)

Rata - Rata = $\frac{Nilai\ 1 + Nilai\ 2 + Nilai\ 3}{N}$

Rata - Rata = $\frac{30.58 + 25.43 + 28.32}{3}$

Rata - Rata = $\frac{84.33}{3}$ = 28.11

Tahun 2022 = Perkotaan + Perdesaan + (Perkotaan + Perdesaan)

Rata - Rata = $\frac{Nilai\ 1 + Nilai\ 2 + Nilai\ 3}{N}$

Rata - Rata = $\frac{29.15 + 35.79 + 31.94}{3}$

Rata - Rata = $\frac{96.88}{3}$ = 32.29

Tabel 2. Rata - Rata Antar Jenis Kelamin 2020 - 2022

Tahun/Kategori	2020	2021	2022
Laki - Laki Rata - Rata	29.29	25.83	28.25
Perempuan Rata - Rata	32.63	28.11	32.29
Total Keseluruhan	61.92	53.94	60.54

4.3 Pengujian

4.3.1 Pengujian Ranking Antar Provinsi

```
df = df.replace(0, None)

selected_columns = df.columns[1:]

for col_name in selected_columns:

    df = df.withColumn(col_name, df[col_name].cast("double"))

# Calculate mean values

mean_values = df.agg(*(mean(column).alias(column) for column in
df.columns)).collect()[0].asDict()

# Replace missing values with mean

for column in df.columns:

    df = df.withColumn(column, when(df[column].isNull(),
mean_values[column]).otherwise(df[column]))


df_clustering = df.select(selected_columns + ["provinsi"])

vec_assembler = VectorAssembler(inputCols=selected_columns,
outputCol="features")

final_data = vec_assembler.transform(df_clustering)

final_data.select('provinsi','features').show(5)
```

Tabel 3. Ranking Antar Provinsi Top 5

provinsi	features
Aceh	[24.68,24.91,24.8...]
Sumatera Utara	[23.21,26.4,24.64...]
Sumatera Barat	[30.46,27.02,28.6...]
Riau	[25.3,22.42,23.6,...]
Jambi	[19.75,20.78,20.4...]

```
scaler = StandardScaler(inputCol="features",
                        outputCol="scaledFeatures",
                        withStd=True,
                        withMean=False)

# Compute summary statistics by fitting the StandardScaler
scalerModel = scaler.fit(final_data)

# Normalize each feature to have unit standard deviation.
final_data = scalerModel.transform(final_data)

final_data.select('provinsi', 'features' ,'scaledFeatures').show(30)
```

```
+--------------------+--------------------+--------------------+
|            provinsi|            features|      scaledFeatures|
+--------------------+--------------------+--------------------+
|                Aceh|[24.68,24.91,24.8...|[4.05565383671143...|
|      Sumatera Utara|[23.21,26.4,24.64...|[3.81408936588624...|
|      Sumatera Barat|[30.46,27.02,28.6...|[5.00547876281322...|
|                Riau|[25.3,22.42,23.6,...|[4.15753817134519...|
|               Jambi|[19.75,20.78,20.4...|[3.24550904680109...|
|    Sumatera Selatan|[29.01,27.28,27.9...|[4.76720088342782...|
|            Bengkulu|[28.76,28.94,28.8...|[4.72611849043034...|
|             Lampung|[28.82,30.07,29.6...|[4.73597826474974...|
|Kepulauan Bangka ...|[31.13,26.27,28.9...|[5.11557957604647...|
|      Kepulauan Riau|[16.37,18.85,16.6...|[2.69007509347513...|
|         DKI Jakarta|[31.85,26.6163636...|[5.23389686787922...|
|          Jawa Barat|[30.04,31.96,30.4...|[4.93646034257745...|
|         Jawa Tengah|[33.94,33.74,33.8...|[5.57734567333817...|
|       DI Yogyakarta|[35.6,36.46,35.82...|[5.85013276284145...|
|          Jawa Timur|[31.14,31.48,31.2...|[5.11722287176637...|
|              Banten|[28.37,34.56,30.0...|[4.66202995735427...|
|                Bali|[23.18,29.07,24.9...|[3.80915947872654...|
| Nusa Tenggara Barat|[41.21,42.69,41.9...|[6.77202166170495...|
| Nusa Tenggara Timur|[33.19,32.95,33.0...|[5.45409849434573...|
|    Kalimantan Barat|[20.76,26.23,24.3...|[3.41148191451091...|
|   Kalimantan Tengah|[26.29,23.8,24.82...|[4.32022444761522...|
|  Kalimantan Selatan|[31.9,33.77,32.88...|[5.24211334647872...|
|    Kalimantan Timur|[22.31,19.89,21.5...|[3.66619275109530...|
|    Kalimantan Utara|[26.02,20.64,23.9...|[4.27585546317794...|
|      Sulawesi Utara|[22.44,25.08,23.6...|[3.68755559545399...|
|     Sulawesi Tengah|[25.19,24.82,24.9...|[4.13946191842630...|
|    Sulawesi Selatan|[28.03,26.44,27.1...|[4.60615790287769...|
|   Sulawesi Tenggara|[28.58,27.58,27.9...|[4.69653916747215...|
|           Gorontalo|[30.08,29.52,29.7...|[4.94303352545705...|
|      Sulawesi Barat|[24.82,24.69,24.7...|[4.07865997679002...|
+--------------------+--------------------+--------------------+
only showing top 30 rows
```

Tabel 4. Ranking Antar Provinsi Top 30

4.3.2 Pengujian Ranking Antar Jenis Kelamin

```
# Create a Window specification for ranking by "avg_pria"
window_spec_pria = Window.orderBy(F.col("avg_pria").desc())
rank_pria = df_result.withColumn("rank", F.rank().over(window_spec_
pria))

# Create a Window specification for ranking by "avg_wanita"
window_spec_wanita = Window.orderBy(F.col("avg_wanita").desc())
rank_wanita = df_result.withColumn("rank", F.rank().over(window_spe
c_wanita))

# Select the required columns for each ranking
rank_pria = rank_pria.select("provinsi", "avg_pria", "rank")
rank_wanita = rank_wanita.select("provinsi", "avg_wanita", "rank")

# Show the results
rank_pria.show()
rank_wanita.show()
```

provinsi	avg_pria	rank
Nusa Tenggara Barat	41.27	1
DI Yogyakarta	32.77	2
Jawa Tengah	31.646666666666665	3
Gorontalo	30.843333333333334	4
Kalimantan Selatan	30.66	5
Nusa Tenggara Timur	29.840000000000003	6
Jawa Barat	29.636666666666667	7
Jawa Timur	29.623333333333335	8
Kepulauan Bangka ...	29.293333333333333	9
Lampung	29.236666666666668	10
Sumatera Selatan	28.47	11
Sulawesi Selatan	28.319999999999997	12
Kalimantan Utara	27.78333333333333	13
Sulawesi Tenggara	27.24	14
Bengkulu	26.963333333333328	15
Banten	26.49	16
Sulawesi Barat	25.796666666666663	17
Sumatera Barat	25.756666666666664	18
Aceh	25.363333333333333	19
DKI Jakarta	23.953333333333333	20

only showing top 20 rows

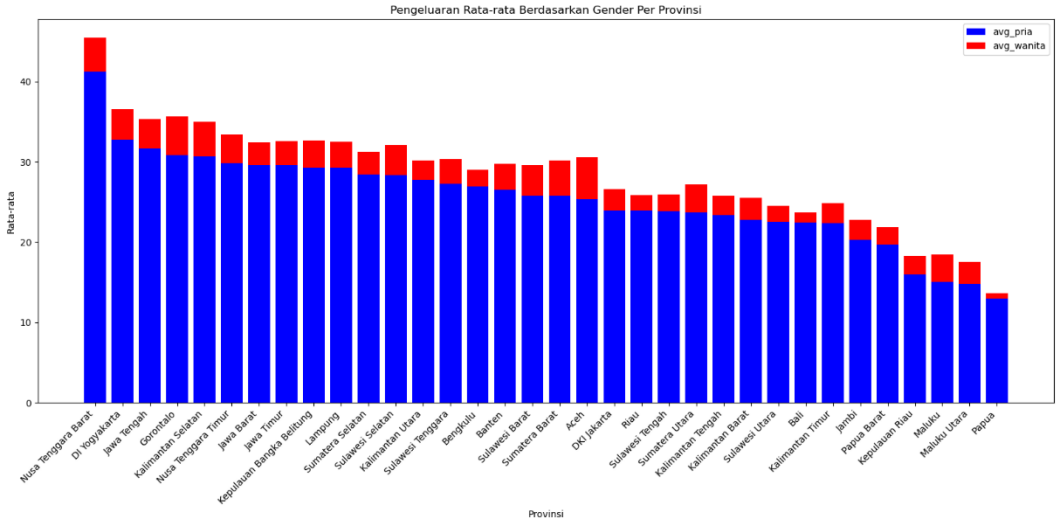
provinsi	avg_wanita	rank
Nusa Tenggara Barat	45.5	1
DI Yogyakarta	36.580000000000005	2
Gorontalo	35.69333333333333	3
Jawa Tengah	35.31333333333333	4
Kalimantan Selatan	34.99333333333333	5
Nusa Tenggara Timur	33.39666666666667	6
Kepulauan Bangka ...	32.67666666666667	7
Jawa Timur	32.626666666666665	8
Lampung	32.46666666666667	9
Jawa Barat	32.4	10
Sulawesi Selatan	32.056666666666665	11
Sumatera Selatan	31.256666666666664	12
Aceh	30.573333333333334	13
Sulawesi Tenggara	30.30666666666667	14
Kalimantan Utara	30.19666666666667	15
Sumatera Barat	30.173333333333332	16
Banten	29.74	17
Sulawesi Barat	29.610000000000003	18
Bengkulu	29.05	19
Sumatera Utara	27.22666666666667	20

only showing top 20 rows

Tabel 5. Ranking Jenis Kelamin Pria

```
# Ambil data yang diperlukan untuk visualisasi
    visualization_data_pria = rank_pria.toPandas()
    visualization_data_wanita = rank_wanita.toPandas()

    # Gabungkan kedua dataframe berdasarkan kolom "provinsi"
    visualization_data = pd.merge(visualization_data_pria,
    visualization_data_wanita, on="provinsi")

    # Set up plot
    plt.figure(figsize=(16, 8))

    colors = {"avg_pria": "blue", "avg_wanita": "red"}

    # Loop untuk setiap kolom yang diinginkan
    for idx, col in enumerate(["avg_pria", "avg_wanita"]):
    plt.bar(visualization_data["provinsi"], visualization_data[col], label=col,
    color=colors[col] ,zorder=2 - idx)  # Menggunakan zorder untuk mengatur
    kedalaman bar

    # Tambahkan label dan judul
    plt.xlabel("Provinsi")
    plt.ylabel("Rata-rata")
    plt.title("Pengeluaran Rata-rata Berdasarkan Gender Per Provinsi")
    plt.xticks(rotation=45, ha="right")
    plt.legend()
    plt.tight_layout()
```

```
# Tampilkan plot

plt.show()
```



Gambar 2. Pengeluaran Rata - Rata Berdasarkan Gender Provinsi

4.3.3 Pengujian Ranking Antar Kota dan Desa

```
# Menghitung nilai rata-rata di perdesaan

avg_perdesaan = (

    F.col("2020-pria-perdesaan") + F.col("2021-pria-perdesaan") +
F.col("2022-pria-perdesaan") +

    F.col("2020-wanita-perdesaan") + F.col("2021-wanita-perdesaan") +
F.col("2022-wanita-perdesaan")

) / 6


# Menghitung nilai rata-rata di perkotaan

avg_perkotaan = (

    F.col("2020-pria-perkotaan") + F.col("2021-pria-perkotaan") +
F.col("2022-pria-perkotaan") +

    F.col("2020-wanita-perkotaan") + F.col("2021-wanita-perkotaan") +
F.col("2022-wanita-perkotaan")
```

```
) / 6

# Seleksi kolom-kolom yang diperlukan

df_result = predictions.select("provinsi",
avg_perdesaan.alias("nilai_rata_perdesaan"),
avg_perkotaan.alias("nilai_rata_perkotaan"))

df_result.show()
```

```
+--------------------+--------------------+--------------------+
|            provinsi|nilai_rata_perdesaan|nilai_rata_perkotaan|
+--------------------+--------------------+--------------------+
|                Aceh|  28.468333333333334|   27.00333333333334|
|      Sumatera Utara|  26.959999999999997|  24.256666666666664|
|      Sumatera Barat|               30.01|  25.886666666666667|
|                Riau|  23.353333333333335|  27.194999999999997|
|               Jambi|   23.78833333333333|               16.97|
|    Sumatera Selatan|  28.051666666666662|              32.875|
|            Bengkulu|  24.888333333333335|  34.281666666666666|
|             Lampung|  31.031666666666666|  30.564999999999998|
|Kepulauan Bangka ...|   28.68166666666667|   32.79833333333333|
|      Kepulauan Riau|               20.24|  16.706666666666667|
|         DKI Jakarta|                0.0|              25.295|
|          Jawa Barat|               32.56|  30.591666666666665|
|         Jawa Tengah|   32.95166666666667|   33.97833333333333|
|       DI Yogyakarta|   35.38333333333333|  34.410000000000004|
|          Jawa Timur|              30.965|  31.274999999999995|
|              Banten|  30.708333333333332|              27.295|
|                Bali|   25.19333333333333|  22.136666666666667|
| Nusa Tenggara Barat|   45.67666666666667|  41.098333333333336|
| Nusa Tenggara Timur|               30.98|              33.305|
|    Kalimantan Barat|  26.779999999999998|  19.543333333333333|
+--------------------+--------------------+--------------------+
only showing top 20 rows
```

Tabel 6. Pengujian Menghitung Rata - Rata Kota dan Desa

```
# Fungsi untuk menghitung rata-rata di suatu lokasi (perkotaan atau perdesaan) untuk semua tahun

def calculate_avg(location_prefix):

    return sum([F.col(f"{year}-{location_prefix}") for year in range(2020, 2023)]) / 3


# Menghitung rata-rata di perdesaan
```

```
avg_perdesaan = calculate_avg("pria-perdesaan") +
calculate_avg("wanita-perdesaan")

# Menghitung rata-rata di perkotaan

avg_perkotaan = calculate_avg("pria-perkotaan") +
calculate_avg("wanita-perkotaan")

# Menambahkan kolom untuk menentukan apakah di provinsi itu
lebih banyak di perdesaan atau di perkotaan

df_result = predictions.withColumn(

   "dominant_location",

   F.when(avg_perdesaan > avg_perkotaan,
"Perdesaan").otherwise("Perkotaan")

).select("provinsi", "dominant_location")

df_result.show()
```

```
+--------------------+-----------------+
|            provinsi|dominant_location|
+--------------------+-----------------+
|                Aceh|        Perdesaan|
|      Sumatera Utara|        Perdesaan|
|      Sumatera Barat|        Perdesaan|
|                Riau|        Perkotaan|
|               Jambi|        Perdesaan|
|    Sumatera Selatan|        Perkotaan|
|            Bengkulu|        Perkotaan|
|             Lampung|        Perdesaan|
|Kepulauan Bangka ...|        Perkotaan|
|      Kepulauan Riau|        Perdesaan|
|         DKI Jakarta|        Perkotaan|
|          Jawa Barat|        Perdesaan|
|         Jawa Tengah|        Perkotaan|
|       DI Yogyakarta|        Perdesaan|
|          Jawa Timur|        Perkotaan|
|              Banten|        Perdesaan|
|                Bali|        Perdesaan|
| Nusa Tenggara Barat|        Perdesaan|
| Nusa Tenggara Timur|        Perkotaan|
|    Kalimantan Barat|        Perdesaan|
+--------------------+-----------------+
only showing top 20 rows
```

Tabel 7. Menentukan Provinsi Perkotaan atau Pedesaan

4.4.4 Pengujian Kenaikan dan Penurunan Dalam Waktu 3 Tahun

```
# Menghitung perubahan nilai dari tahun ke tahun di perdesaan dan perkotaan

for year in range(2021, 2023):
    for gender in ["pria", "wanita"]:
        for location in ["perdesaan", "perkotaan"]:
            col_name = f"change_{year}_{gender}_{location}"
            predictions = predictions.withColumn(
                col_name,
                F.col(f"{year}-{gender}-{location}") - F.col(f"2020-{gender}-{location}")
            )

# Menentukan kategori kenaikan atau penurunan untuk setiap tahun dan lokasi

for location in ["perdesaan", "perkotaan"]:
    for year in range(2021, 2023):
        trend_col_name = f"trend_{year}_{location}"
        predictions = predictions.withColumn(
            trend_col_name,
            F.when(F.col(f"change_{year}_pria_{location}") + F.col(f"change_{year}_wanita_{location}") > 0, "Kenaikan").otherwise("Penurunan")
        )
```

```
# Menampilkan hasil

result_columns = ["provinsi"] + [f"trend_{year}_{location}" for year
in range(2021, 2023) for location in ["perdesaan", "perkotaan"]]

df_result = predictions.select(result_columns)

df_result.show()
```

provinsi	trend_2021_perdesaan	trend_2021_perkotaan	trend_2022_perdesaan	trend_2022_perkotaan
Aceh	Penurunan	Kenaikan	Kenaikan	Kenaikan
Sumatera Utara	Penurunan	Penurunan	Kenaikan	Kenaikan
Sumatera Barat	Penurunan	Penurunan	Kenaikan	Penurunan
Riau	Penurunan	Kenaikan	Kenaikan	Penurunan
Jambi	Penurunan	Penurunan	Kenaikan	Penurunan
Sumatera Selatan	Penurunan	Kenaikan	Kenaikan	Kenaikan
Bengkulu	Penurunan	Kenaikan	Penurunan	Kenaikan
Lampung	Penurunan	Kenaikan	Kenaikan	Penurunan
Kepulauan Bangka ...	Penurunan	Penurunan	Kenaikan	Kenaikan
Kepulauan Riau	Penurunan	Penurunan	Kenaikan	Penurunan
DKI Jakarta	Penurunan	Penurunan	Penurunan	Penurunan
Jawa Barat	Penurunan	Penurunan	Kenaikan	Penurunan
Jawa Tengah	Penurunan	Penurunan	Kenaikan	Penurunan
DI Yogyakarta	Penurunan	Penurunan	Kenaikan	Penurunan
Jawa Timur	Penurunan	Penurunan	Kenaikan	Penurunan
Banten	Penurunan	Kenaikan	Penurunan	Penurunan
Bali	Penurunan	Kenaikan	Penurunan	Penurunan
Nusa Tenggara Barat	Penurunan	Penurunan	Kenaikan	Penurunan
Nusa Tenggara Timur	Penurunan	Penurunan	Penurunan	Kenaikan
Kalimantan Barat	Penurunan	Penurunan	Kenaikan	Penurunan

only showing top 20 rows

Tabel 8. Menghitung Perubahan Nilai Dari Tahun ke Tahun

```
# Menghitung perubahan nilai dari tahun ke tahun di perdesaan dan
perkotaan

for year in range(2021, 2023):
    for gender in ["pria", "wanita"]:
        for location in ["perdesaan", "perkotaan"]:
            col_name = f"change_{year}_{gender}_{location}"
            predictions = predictions.withColumn(
                col_name,
                F.col(f"{year}-{gender}-{location}") - F.col(f"2020-
{gender}-{location}")
```

```
        )

# Menentukan kategori kenaikan atau penurunan untuk setiap tahun dan lokasi
for year in range(2021, 2023):
    col_name = f"total_change_{year}"
    predictions = predictions.withColumn(
        col_name,
        sum([F.col(f"change_{year}_pria_perdesaan") + F.col(f"change_{year}_wanita_perdesaan") + F.col(f"change_{year}_pria_perkotaan") + F.col(f"change_{year}_wanita_perkotaan")])
    )

    trend_col_name = f"trend_{year}"
    predictions = predictions.withColumn(
        trend_col_name,
        F.when(F.col(col_name) > 0, "Kenaikan").otherwise("Penurunan")
    )

# Menghitung total perubahan dari tahun 2020 hingga 2023
col_name_total = "total_change_2020_2023"
predictions = predictions.withColumn(
    col_name_total,
```

```
    sum([F.col(f"total_change_{year}") for year in range(2021, 2023)])
)

# Menentukan kategori kenaikan atau penurunan untuk total tahun 2020-2023
trend_col_name_total = "trend_2020_2023"
predictions = predictions.withColumn(
    trend_col_name_total,
    F.when(F.col(col_name_total) > 0, "Kenaikan").otherwise("Penurunan")
)

# Menampilkan hasil
result_columns = ["provinsi", trend_col_name_total, col_name_total]
df_result = predictions.select(result_columns).distinct()

# Menampilkan hasil
df_result.show()
```

Tabel 9. Menghitung Total Perubahan Nilai 2020-2023

provinsi	trend_2020_2023	total_change_2020_2023
Bali	Penurunan	-35.43000000000001
Sulawesi Selatan	Kenaikan	15.59
Kepulauan Riau	Penurunan	-4.640000000000004
Banten	Penurunan	-54.31000000000001
Jawa Timur	Penurunan	-20.100000000000012
Kepulauan Bangka ...	Penurunan	-4.1999999999999815
Nusa Tenggara Barat	Penurunan	-6.419999999999995
Sulawesi Tengah	Penurunan	-9.570000000000004
Bengkulu	Penurunan	-8.820000000000011
Maluku Utara	Penurunan	-5.82
Sulawesi Utara	Penurunan	-18.209999999999997
Sumatera Selatan	Kenaikan	8.769999999999996
Papua Barat	Penurunan	-4.079999999999995
Kalimantan Timur	Kenaikan	14.340000000000007
Riau	Kenaikan	2.84
Kalimantan Selatan	Penurunan	-24.98
Kalimantan Barat	Penurunan	-21.820000000000007
Kalimantan Utara	Kenaikan	34.17
Jambi	Penurunan	-12.73
Sumatera Barat	Penurunan	-39.25999999999999

only showing top 20 rows

4.4 Kesimpulan Pengujian

Berdasarkan skenario pengujian yang telah dilakukan, maka dapat diperoleh Tabel dan Grafik di bawah ini:

Tabel 10. Kesimpulan Pengujian

Pengujian	Hasil
Ranking Antar Provinsi	Top 5 Provinsi Keluhan Kesehatan: 1. Aceh : 4.05 2. Sumatera Utara : 3.81 3. Sumatera Barat : 5.00 4. Riau : 4.15 5. Jambi : 3.24
Antar Jenis Kelamin	Nilai Rata - Rata Pria : 25.58180555555555 Nilai Rata - Rata Perempuan : 28.661250000000006 Gender Keluhan Rata - Rata Tertinggi : **Perempuan**
Ranking Antar Kota dan Desa	Ranking Top 20 : 8 Perkotaan 12 Perdesaan Toal Paling Banyak : **Perdesaan**
Kenaikan dan Penurunan Dalam Waktu 3 Tahun	Berdasarkan Top 20 : Sulawesi Selatan Penaikan = **15.59** Bali Mencapai Penurunan = **-35.43000000000001**

BAB V

PENUTUP

5.1 Kesimpulan

Berdasarkan hasil penelitian dan pembahasan yang mengacu pada tujuan penelitian ini, maka disimpulkan bahwa :

1. Berhasil menerapkan konsep data mining dengan metode analisis deskriptif (mencari nilai rata-rata) dan menggunakan algoritma clustering k-means.
2. Berdasarkan data penelitian yang diambil dari sumber badan pusat statistik indonesia total keseluruhan rata - rata laki - laki dan perempuan yang tertinggi di tahun 2020.
3. Berdasarkan pengujian keluhan kesehatan indonesia di 34 provinsi Aceh menjadi peringkat pertama dengan keluhan kesehatan terbanyak.
4. Berdasarkan pengujian Perempuan menjadi salah satu peringkat pertama keluhan kesehatan tertinggi di indonesia.
5. Berdasarkan pengujian Perdesaan menjadi salah satu keluhan kesehatan terbanyak di indonesia.
6. Berdasarkan pengujian kenaikan terbanyak terjadi di Sulawesi Selatan dan penurunan terjadi di Bali.

5.2 Saran

Berdasarkan hasil penelitian dan pembahasan yang mengacu pada batasan penelitian ini, maka dapat disarankan bahwa :

1. Bagi peneliti seterusnya diharapkan agar dilanjutkan dengan menggunakan metode lainnya untuk mendukung keakuratan lebih jelas dari pengelompokkan data yang sudah kami lampirkan.
2. Hasil yang didapat dari penelitian ini dapat menjadi masukan kepada pemerintah, provinsi yang menjadi perhatian lebih pada penduduk yang memiliki keluhan kesehatan tertinggi berdasarkan cluster yang telah dilakukan.

DAFTAR PUSTAKA

[1] Nurul Rofiqo, A. P. (2018). PENERAPAN CLUSTERING PADA PENDUDUK YANG MEMPUNYAI KELUHAN KESEHATAN DENGAN DATAMINING K-MEANS. *http://ejurnal.stmik-budidarma.ac.id/index.php/komik*, 216-217.

[2] Mardiansa, H. L. (2023). Penerapan Data Mining Untuk Mengetahui Minat Siswa Pada Pelajaran IPA Mengunakan Metode K-Means Clustering. *Jurnal Multidisiplin Dehasen*, 694 - 695.

[3] Jiawei Han, M. K. (2006). *Data Mining: Concepts and Techniques Third Edition.* San Francisco: Morgan Kaufmann.

[4] Yulia Darmi, A. (2016). PENERAPAN METODE CLUSTERING K-MEANS DALAM PENGELOMPOKAN PENJUALAN PRODUK. *Jurnal Media Infotama*, 149-150.