

Data Mining Prediksi Jumlah Kematian Berjenis Kelamin Perempuan Berdasarkan Usia Dengan Menggunakan Metode Regresi Linear Sederhana

Syahrizal Raksa Negara

Program Studi Teknik Informatika , Universitas Serang Raya

e-mail: syahrizal3452@gmail.com

ABSTRAKSI

Penelitian ini bertujuan untuk menganalisis hubungan antara jumlah kematian berjenis kelamin perempuan berdasarkan usia. Penelitian ini menggunakan data kematian yang dikumpulkan dari instansi terkait dan menerapkan metode regresi linear sederhana untuk memodelkan hubungan antara kedua variabel. Hasil penelitian menunjukkan bahwa terdapat hubungan positif antara usia dan jumlah kematian perempuan. Artinya, seiring dengan bertambahnya usia, risiko kematian bagi perempuan juga meningkat. Model regresi linear sederhana yang dihasilkan dapat digunakan untuk memprediksi jumlah kematian perempuan berdasarkan usia. Penelitian ini memberikan kontribusi dalam memahami pola kematian perempuan berdasarkan usia. Temuan penelitian ini dapat bermanfaat bagi pembuat kebijakan dalam mengembangkan program dan kebijakan kesehatan yang berfokus pada kelompok usia perempuan yang paling berisiko mengalami kematian. Meskipun demikian, penelitian ini hanya menggunakan data dari satu tahun dan satu wilayah saja. Penelitian lanjutan dengan cakupan data yang lebih luas dan menggunakan metode analisis yang lebih kompleks diperlukan untuk mendapatkan pemahaman yang lebih komprehensif mengenai faktor-faktor yang mempengaruhi kematian perempuan berdasarkan usia.

Kata Kunci: data mining, regresi linier, orange, hubungan positif, model prediksi

ABSTRACT

This study aims to analyze the relationship between the number of female deaths and age. It uses mortality data collected from relevant agencies and applies simple linear regression to model the relationship between these two variables. The results show a positive association: as age increases, the risk of death among females also rises. The derived simple linear regression model can be used to predict the number of female deaths based on age. This research contributes to understanding age-related patterns of female mortality and can inform policymakers in designing health programs and policies targeted at the most at-risk female age groups. However, the study is limited by its use of data from a single year and one region; future research with a broader dataset and more sophisticated analytical methods is needed to achieve a more comprehensive understanding of factors influencing female mortality by age.

Keywords: data mining, linear regression, orange, positive association, predictive model

Penulis Korespondensi

Syahrizal Raksa Negara

Tanggal Submit : 26/01/2024

Tanggal Diterima : 11/07/2024

Tanggal Terbit : 30/08/2024

This is an open access article under the [CC-BY-NC-SA](https://creativecommons.org/licenses/by-nc-sa/4.0/) license



Copyright: © 2023 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 (CC BY-NC-SA 4.0) International License (<https://creativecommons.org/licenses/by-nc-sa/4.0/>).

Publisher's Note: JPPM stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

I. PENDAHULUAN

Data mining merupakan serangkaian proses untuk menggali nilai tambah berupa informasi yang selama ini tidak diketahui secara manual dari suatu basis data. Metode ini muncul sejak tahun 1990-an sebagai cara

yang tepat untuk mengekstraksi pola dan informasi guna menemukan hubungan antar data, sehingga objek-objek yang berada dalam satu cluster memiliki kesamaan yang tinggi. Salah satu tujuan utamanya adalah melakukan pengelompokan ke dalam satu atau

lebih cluster sehingga objek-objek serupa terkumpul bersama. Data mining merupakan bagian integral dari proses penemuan pengetahuan dalam basis data (Knowledge Discovery in Databases) [1].

Kesehatan masyarakat menjadi perhatian utama, terutama dalam konteks tingkat kematian perempuan yang merupakan indikator vital bagi suatu populasi. Memahami hubungan antara usia dan jumlah kematian perempuan memungkinkan peneliti mengidentifikasi kelompok usia yang paling rentan. Kesehatan didefinisikan tidak hanya sebagai ketiadaan penyakit atau kelemahan, tetapi juga sebagai kondisi fisik yang baik, mental yang sehat, dan kesejahteraan sosial [2]. Oleh karena itu, analisis pola kematian sangat penting untuk merancang intervensi kesehatan yang tepat sasaran.

Penelitian ini menggunakan regresi linear sederhana sebagai metode analisis untuk mendalami faktor-faktor demografis yang memengaruhi pola kematian perempuan. Dengan regresi linear sederhana, hubungan antara variabel usia dan jumlah kematian dapat diukur secara kuantitatif dan diuji signifikansinya. Penelitian, yang merupakan penyelidikan terorganisir dan kritis dalam mencari fakta, menyediakan kerangka kerja yang sistematis untuk studi ini [8]. Melalui pemodelan statistik tersebut, diharapkan terungkap determinan-determinan kunci yang dapat menunjang perumusan kebijakan kesehatan.

Hasil penelitian ini diharapkan dapat membantu perumusan kebijakan kesehatan yang lebih efektif dengan merinci determinan-determinan kematian perempuan. Dengan mengidentifikasi variabel-variabel demografis penting, pembuat kebijakan dapat merancang program yang ditujukan pada kelompok usia berisiko tinggi. Selain memberikan kontribusi pada tinjauan lokal, studi ini juga relevan dalam pengembangan ilmu data mining, khususnya dalam penerapan regresi linear sederhana di bidang kesehatan masyarakat. Dengan demikian, penelitian ini tidak hanya memberikan wawasan kontekstual tetapi juga memperkaya metode analisis data.

Penelitian ini memiliki beberapa tujuan utama. Pertama, menganalisis dan memahami hubungan antara usia dan jumlah kematian perempuan menggunakan regresi linear sederhana. Kedua, membangun model prediktif untuk meramalkan jumlah kematian perempuan berdasarkan variabel usia. Ketiga, memberikan kontribusi pada pemahaman pola kematian perempuan dengan fokus pada faktor umur. Keempat, menyajikan temuan yang dapat digunakan pembuat kebijakan dalam merancang program kesehatan yang lebih terarah untuk kelompok usia berisiko tinggi.

Manfaat penelitian ini mencakup peningkatan akurasi identifikasi kelompok usia berisiko tinggi dan penyusunan intervensi yang lebih tepat sasaran. Model prediktif yang dihasilkan dapat membantu memproyeksikan tren kematian di masa mendatang dan mengalokasikan sumber daya kesehatan secara efektif.

Temuan penelitian juga diharapkan menjadi bahan pertimbangan bagi praktisi kesehatan di tingkat fasilitas layanan hingga dinas kesehatan. Dengan demikian, studi ini mendukung upaya pencegahan kematian perempuan melalui pendekatan berbasis data.

Dengan memanfaatkan teknik data mining dalam kerangka kesehatan masyarakat, penelitian ini menggabungkan kemajuan metodologi dan aplikasi praktis untuk menurunkan angka kematian perempuan. Pendekatan ini menekankan pentingnya kolaborasi lintas sektor—mulai dari fasilitas layanan primer hingga pengambil kebijakan—dalam menerjemahkan hasil analisis menjadi program nyata. Pengintegrasian regresi linear sederhana dalam studi kesehatan masyarakat membuka peluang baru untuk penelitian lanjutan. Akhirnya, diharapkan hasil penelitian dapat mengurangi angka kematian perempuan dan meningkatkan kualitas hidup masyarakat.

II. METODE PENELITIAN

2.1 Proses Penelitian

Penelitian ini dilaksanakan melalui serangkaian tahapan yang dijabarkan pada Gambar 1.



Gambar 1. Proses Penelitian

2.2 Sumber Data

Data kematian perempuan berdasarkan kelompok usia diambil dari portal resmi Pemerintah Provinsi DKI Jakarta, yaitu <https://data.jakarta.go.id/dataset/>. Data mentah tersebut berisi catatan jumlah kematian per kelompok umur, yang selanjutnya diunduh dalam format CSV untuk keperluan analisis lebih lanjut. “Data adalah kumpulan kejadian yang diangkat dari suatu kenyataan berupa angka-angka, huruf-huruf, atau simbol-simbol khusus atau gabungan darinya yang

masih belum bisa bercerita banyak, sehingga perlu diolah lebih lanjut” [6].

2.3 Data Mining

Data mining merupakan proses menemukan informasi tersembunyi dari sekumpulan data yang tersimpan dalam database atau datasheet [7]. Pada penelitian ini, proses data mining mencakup tahap eksplorasi data, pemilihan variabel, dan penerapan algoritma clustering serta analisis statistik deskriptif. Berbagai teknik statistik, matematika, dan pembelajaran mesin digunakan untuk mengidentifikasi pola serta mengubah data mentah menjadi informasi yang bermakna bagi pengambilan keputusan di bidang kesehatan masyarakat [3].

2.4 Regresi Linier Sederhana

Regresi linier sederhana adalah teknik analisis statistik yang digunakan untuk memodelkan dan memprediksi hubungan antara satu variabel bebas (X) dan satu variabel terikat (Y) [4]. Model ini digambarkan oleh persamaan:

$$Y = b_0 + b_1X$$

Keterangan:

b_0 : konstanta (intersep)

b_1 : koefisien regresi (slope)

Y : variabel terikat (jumlah kematian perempuan)

X : variabel bebas (usia)

Estimasi parameter b_0 dan b_1 dilakukan dengan metode kuadrat terkecil (least squares), di mana:

$$b_1 = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}, \quad b_0 = \bar{y} - b_1 \bar{x}$$

III. HASIL DAN PEMBAHASAN

Analisis data mencakup kegiatan membedakan, mengurai, dan memilah elemen-elemen informasi untuk dikelompokkan ke dalam kategori tertentu demi tujuan tertentu, kemudian mengkaji relasi antar kelompok tersebut dan menafsirkan maknanya [5]. Dalam penelitian ini, analisis dimulai dengan penetapan variabel—usia (X) sebagai variabel bebas dan jumlah kematian perempuan (Y) sebagai variabel terikat—kemudian data dihitung secara manual untuk memperoleh parameter regresi sebelum diverifikasi menggunakan aplikasi Orange.

Dari sampel 100 data, dibuat tabel tambahan berisi kolom X^2 , Y^2 , dan XY , kemudian dihitung $\sum X$, $\sum Y$, $\sum X^2$, $\sum Y^2$, dan $\sum XY$. Berdasarkan metode kuadrat terkecil, koefisien regresi b_1 dan konstanta b_0 dihitung dengan rumus:

$$b_1 = \frac{n \sum XY - (\sum X)(\sum Y)}{n \sum X^2 - (\sum X)^2}, \quad b_0 = \bar{Y} - b_1 \bar{X}.$$

$$a = \frac{(21329)(255022500) - (5050)(1077145)}{(100)(25502250) - (5050)^2}$$

$$a = \frac{543942822500 - 5439463322500}{2550225000 - 25502500} = 2$$

$$b = \frac{(100)(1077145) - (5050)(21329)}{(100)(25502250) - (5050)^2}$$

$$b = \frac{107714500 - 107711450}{1287863625000 - 25502500} = 1,19$$

Substitusi nilai menghasilkan $b_0=2$ dan $b_1=1,19$. Sehingga persamaan regresi linier sederhana menjadi: $Y^{\wedge} = 2 + 1,19X$. Sebagai contoh prediksi, untuk $X=0,5$ (usia setengah tahun), diperoleh $Y^{\wedge} = 2 + 1,19 \times 0,5 = 2,595$.

3.1 Validasi dengan Orange

Hasil perhitungan manual selanjutnya dipindahkan ke format CSV dan diimpor ke Orange. Pada platform ini, modul “Linear Regression” digunakan untuk membangun model yang sama, kemudian dievaluasi. Tabel data muncul pada jendela “Data Table” (Gambar 2), di mana variabel X dan Y telah terdeteksi sebagai predictor dan response. Orange secara otomatis menghitung ulang koefisien b_0 dan b_1 , serta menampilkan nilai R^2 dan statistik uji signifikansi model.

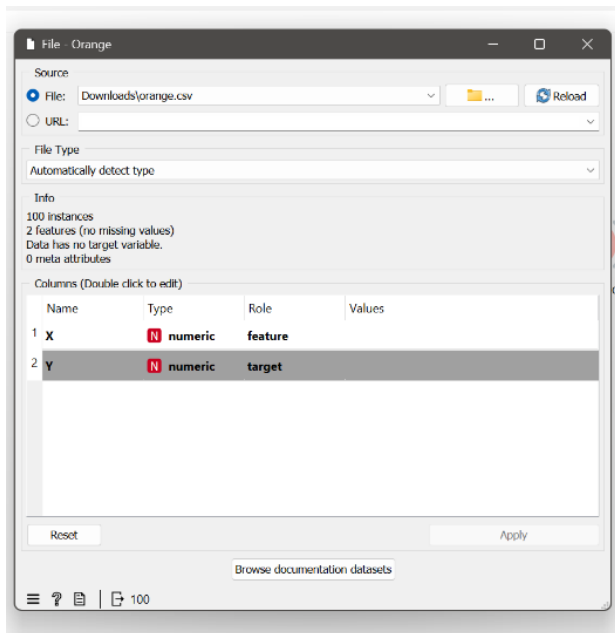
3.2 Visualisasi dengan Scatter Plot

Scatter plot digunakan untuk memvisualisasikan hubungan antara variabel usia (X) dan jumlah kematian (Y) pada kontinum dua dimensi. Setiap titik mewakili satu observasi, sehingga pola linear atau deviasi dari linearitas dapat dilihat dengan jelas [9]. Gambar 3 menampilkan scatter plot hasil data, dilengkapi garis regresi yang menunjukkan arah dan kekuatan hubungan positif antara usia dan risiko kematian perempuan.

Setelah melakukan perhitungan manual dan memperoleh persamaan regresi linier, data selanjutnya diekspor ke format CSV dan diimpor ke dalam perangkat lunak Orange. Proses ini melibatkan tiga tahapan utama, yakni pengelolaan berkas (file), tinjauan tabel data, dan visualisasi melalui scatter plot.

1. Import File

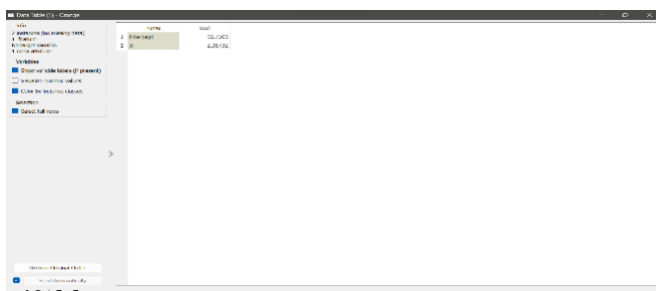
Data awal dari Tabel 2 terlebih dahulu disiapkan di Microsoft Excel, kemudian disimpan sebagai file CSV. File CSV tersebut diunggah ke Orange menggunakan widget **File**, sehingga seluruh variabel—usia (X) dan jumlah kematian (Y)—tersedia di lingkungan Orange. Tampilan antarmuka setelah impor berkas dapat dilihat pada Gambar 2.



Gambar 2. Widget File

2. Data Table & Model Regression

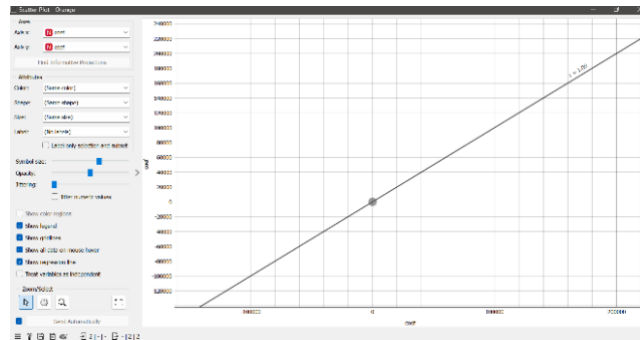
Setelah berkas terunggah, widget **Data Table** menampilkan data mentah dengan kolom-kolom X dan Y beserta sampel sebanyak 100 observasi (Gambar 3). Widget **Linear Regression** kemudian dihubungkan, di mana variabel usia dipilih sebagai *predictor* dan jumlah kematian sebagai *response*. Orange secara otomatis menghitung kembali koefisien intersep (b_0) dan slope (b_1), serta menyajikan nilai koefisien determinasi (R^2) yang mengindikasikan seberapa baik model menerangkan variabilitas data.



Gambar 3 . Data Table

3. Scatter Plot

Untuk memvisualisasikan hubungan linear antara usia dan jumlah kematian, digunakan widget **Scatter Plot**. Setiap titik mewakili satu observasi, sedangkan garis regresi overlay menunjukkan tren positif antara kedua variabel (Gambar 4). Menurut [9], scatter plot semacam ini sangat efektif untuk menguji dan mengilustrasikan sejauh mana variabel prediktor memengaruhi variabel respon. Dari visualisasi tersebut, pola kenaikan jumlah kematian seiring bertambahnya usia perempuan tampak jelas, mengonfirmasi hasil temuan regresi linier sederhana.



Gambar 4. Scatter Plot

V. KESIMPULAN

Berdasarkan scatter plot data kematian perempuan menurut usia, terlihat bahwa tingkat kematian cenderung meningkat seiring bertambahnya usia. Titik-titik data menyebar mengikuti garis tren yang naik, menandakan adanya korelasi positif antara usia dan jumlah kematian. Semakin tinggi usia perempuan, semakin besar pula risikonya untuk meninggal, baik karena penyakit degeneratif maupun gangguan kesehatan lainnya. Temuan ini menegaskan pentingnya pemantauan kesehatan berkelanjutan pada kelompok usia lanjut.

Peningkatan risiko kematian pada usia lanjut dapat dipicu oleh beberapa faktor utama. Pertama, prevalensi penyakit kronis seperti kanker, gangguan jantung, dan stroke umumnya meningkat pada usia yang lebih tua, sehingga memengaruhi angka kematian secara signifikan. Kedua, penurunan fungsi organ tubuh—termasuk fungsi imunitas, paru, dan ginjal—menyebabkan perempuan lanjut usia lebih rentan terhadap infeksi dan komplikasi penyakit. Selain itu, aspek gaya hidup seperti pola makan, aktivitas fisik, dan akses terhadap layanan kesehatan juga berkontribusi pada perbedaan risiko kematian di setiap kelompok usia.

Untuk penelitian selanjutnya, disarankan memperluas cakupan variabel dengan memasukkan faktor-faktor sosio-demografis (misalnya pendidikan, pendapatan, dan status pernikahan) serta variabel klinis seperti indeks massa tubuh dan tekanan darah. Selain itu, penggunaan data longitudinal yang melacak kondisi kesehatan individu dari waktu ke waktu dapat memberikan pemahaman lebih mendalam tentang dinamika kematian berdasarkan usia. Metode analisis multivariat, seperti regresi Cox proportional hazards, juga perlu dipertimbangkan untuk mengevaluasi interaksi antara berbagai determinan kesehatan. Pendekatan ini akan menghasilkan model prediktif yang lebih komprehensif dan akurat untuk mendukung kebijakan kesehatan preventif bagi perempuan di semua rentang usia.

DAFTAR PUSTAKA

- [1] (Alkhairi & Windarto, 2019)
- [2] Riskal Fitri & Syarifuddin Ondeng, 2022;;Lembaga Pembentukan Karakter.
- [3] Arhami & Nasir, 2020; Jollyta et al., 2020.
- [4] Eko putra aperesiasi ekonomi volume 2. Nomor 2 , mei 2014 : 89-94
- [5] Nana Sudjana (2016:27)
- [6] M.Arfa Andika Candra (1) Ika Artahalia Wulandari (2)
- [7] Gede Aditra Pradnyana, S.Kom., M.Kom. Dr. Ketut Agustini,S.Si., M.Si
- [8.] Sandu Siyoto, SKM., 2015)
- [9] Andi Trias Aryanto Tuwanku Aria Auliandri Departemen Manajemen Fakultas Ekonomi dan Bisnis Universitas Airlangg, Jurnal Manajemen Teori dan Terapan Tahun 8. No. 1, April 2015
- [10] INTERNALISASI PENDIDIKAN NILAI DALAM PEMBELAJARAN (UPAYA MENCIPTAKAN SUMBER DAYA MANUSIA YANG BERKARAKTER) Tri Sukitman tri.sukitman@gmail.com Prodi PGSD STKIP PGRI Sumenep