

Implementation of the K-Nearest Neighbor Algorithm for Birth Rate Prediction

Akhyar Alhafiz^{1*}, Rakhmat Kurniawan R.²
Sistem Informasi, Universitas Islam Negeri Sumatera Utara Medan
akhayaralhafiz879@gmail.com¹, rahmat.kr@uinsu.ac.id²

Article Info

Article history:

Received 2025-06-19

Revised 2025-07-08

Accepted 2025-07-19

Keyword:

Birth Rate,
Data Mining,
Family Planning,
K-Nearest Neighbor,
Population Prediction.

ABSTRACT

This study aims to predict the monthly birth rate using the K-Nearest Neighbor (KNN) regression algorithm. The dataset consists of historical data from 2010 to 2020, covering six districts and including variables such as total population, number of couples of reproductive age, family planning participation rate, and monthly birth rate as the prediction target. Data preprocessing involved handling missing values and applying Min-Max normalization. To maintain the time-series nature of the data, a chronological split was used, with 576 records from 2010 to 2018 for training and 216 records from 2019 to 2020 for testing. The model was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2). The best performance was achieved at $K = 7$, with $MAE = 19.94$, $RMSE = 30.91$, and $R^2 = 0.34$. Additionally, the KNN model was compared with Linear Regression and Decision Tree, where KNN outperformed both alternatives. The final model was implemented in a web-based application to facilitate demographic data management and automatic birth rate prediction per district. This system is expected to support policy planning in the fields of population control and public health.



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

I. PENDAHULUAN

Tingkat kelahiran bayi di Indonesia menunjukkan kecenderungan peningkatan setiap bulannya. Kondisi ini turut mendorong pertumbuhan jumlah penduduk dari waktu ke waktu. Salah satu penyebab dari fenomena ini adalah kemajuan teknologi yang memungkinkan proses kelahiran berlangsung lebih cepat dan efisien. Apabila tren ini terus berlanjut, maka akan berpotensi menimbulkan permasalahan kepadatan penduduk, tidak hanya di Indonesia tetapi juga di tingkat global [1].

Tingginya angka kelahiran memerlukan perhatian khusus dalam pengendaliannya. Berdasarkan data terakhir, Indonesia berada di peringkat keempat negara dengan jumlah populasi terbesar di dunia, yaitu sebanyak 272,20 juta jiwa, berada setelah India, China, dan Amerika Serikat [2]. Selain pengaruh teknologi, rendahnya partisipasi pria dalam program keluarga berencana turut menjadi faktor pendorong tingginya angka kelahiran. Partisipasi pria hanya sekitar 4,4%, jauh tertinggal dibandingkan perempuan yang telah mencapai 61,4% [3]. Ketimpangan ini mencerminkan bahwa

keterlibatan pria dalam program KB belum menjadi kebiasaan umum di Indonesia.

Khususnya di Kota Tanjung Balai, jumlah kelahiran bayi tercatat meningkat setiap tahunnya. Pada tahun 2020, tercatat sebanyak 2.723 bayi lahir, melanjutkan tren penurunan yang sempat terjadi selama tiga tahun sebelumnya. Sebagai perbandingan, pada tahun 2018 terdapat 3.832 kelahiran, kemudian menurun menjadi 2.980 pada tahun 2019. Penurunan tersebut kemungkinan dipengaruhi oleh efektivitas pelaksanaan program KB, meningkatnya kesadaran masyarakat terhadap kesehatan reproduksi, serta adanya perubahan kondisi sosial ekonomi yang memengaruhi keputusan keluarga dalam memiliki anak [4].

Namun demikian, angka kelahiran di Kota Tanjung Balai masih tergolong tinggi dan cenderung sulit dikendalikan. Hal ini menimbulkan kekhawatiran karena pertumbuhan penduduk yang terus meningkat dapat berdampak besar terhadap sektor-sektor publik seperti pendidikan, layanan kesehatan, lapangan kerja, hingga ketersediaan sumber daya lainnya. Rendahnya partisipasi pria dalam program KB

menjadi salah satu faktor yang membuat upaya pengendalian kelahiran belum optimal.

Penelitian ini bertujuan untuk memberikan kontribusi dalam membantu pemerintah menyusun kebijakan dan rencana pembangunan terkait pertumbuhan penduduk dan aspek sosial ekonomi. Proses peramalan (forecasting) menjadi elemen penting dalam perencanaan, karena dapat memberikan gambaran untuk memilih alternatif keputusan yang tepat dalam merespons isu-isu kependudukan dan pelaksanaan program KB. Saat ini, salah satu pendekatan yang berkembang pesat dan memberikan banyak manfaat adalah penemuan informasi dari pola tersembunyi dalam jumlah data yang besar, atau dikenal dengan istilah data mining. Data mining merupakan serangkaian tahapan untuk mengekstraksi nilai tambah berupa pengetahuan yang sebelumnya sulit ditemukan secara manual [5]. Proses ini juga dikenal sebagai Knowledge Discovery in Database (KDD), yaitu serangkaian aktivitas mulai dari pengumpulan data historis hingga identifikasi pola atau hubungan dalam kumpulan data berskala besar [6].

Dalam memodelkan prediksi angka kelahiran, pemilihan metode algoritma yang tepat sangat penting untuk memperoleh hasil yang akurat dan dapat diandalkan. Pada penelitian ini, penulis memilih untuk menggunakan algoritma K-Nearest Neighbor (KNN) regresi dibandingkan metode prediksi deret waktu lainnya seperti ARIMA, LSTM, atau SVR. Pemilihan ini didasarkan pada beberapa pertimbangan. Metode ARIMA lebih cocok untuk data univariat dan mengasumsikan stasionaritas, yang kurang sesuai dengan karakteristik data kelahiran yang bersifat multivariabel dan tidak sepenuhnya stasioner. Di sisi lain, algoritma LSTM memerlukan jumlah data yang jauh lebih besar serta arsitektur model yang kompleks, sedangkan Support Vector Regression (SVR) memiliki sensitivitas tinggi terhadap parameter dan skala data. Sementara itu, KNN memiliki keunggulan karena bersifat non-parametrik, tidak memerlukan asumsi distribusi data, dan mampu menangani hubungan non-linear antar fitur. Karakteristik ini sangat cocok dengan struktur dataset yang digunakan, yaitu data kelahiran berdasarkan jumlah penduduk, pasangan usia subur, dan tingkat partisipasi KB, yang tidak memiliki pola musiman yang kuat, namun menunjukkan variasi lokal yang kompleks.

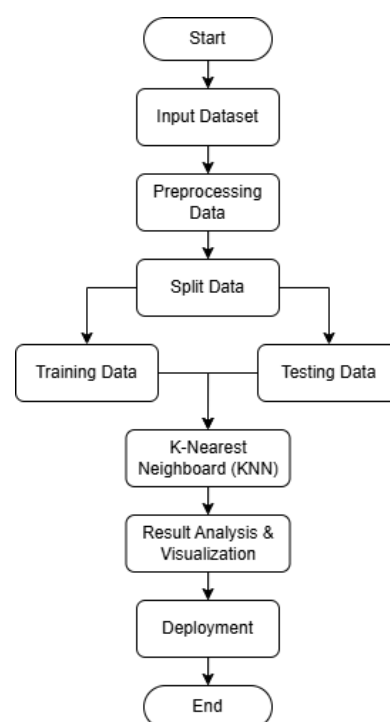
Sejumlah penelitian terdahulu telah memanfaatkan algoritma ini untuk berbagai tujuan. Syafana et al. (2024) menerapkan KNN untuk memprediksi angka kelahiran berdasarkan kelompok umur ibu, namun hasil akurasi hanya mencapai 45%, menunjukkan masih adanya keterbatasan pada data dan fitur yang digunakan [7]. Sutrimo (2022) mengembangkan prediksi jenis persalinan (normal atau caesar) menggunakan KNN berbobot berdasarkan data kesehatan ibu hamil, dan hasilnya menunjukkan tingkat ketepatan yang cukup baik [8]. Qibtiyah et al. (2024) menggunakan algoritma Naïve Bayes untuk memprediksi angka kelahiran di Kabupaten Bojonegoro dengan variabel pasangan usia subur, yang menghasilkan hubungan yang signifikan antara fitur dan target [9]. Desti Puspita Sari et al.

(2023) berhasil menerapkan KNN dalam sistem prediksi kelulusan siswa SMP dengan hasil yang akurat, menunjukkan efektivitas algoritma ini dalam konteks pendidikan [10]. Sedangkan Habibi et al. (2025) menggunakan KNN untuk memprediksi jumlah kasus HIV di Provinsi Jawa Barat berdasarkan data dari Open Data Jabar; meskipun nilai R^2 yang diperoleh hanya 0,2397, penelitian ini mampu mengidentifikasi wilayah prioritas penyebaran [11].

Berdasarkan latar belakang dan studi terdahulu, penelitian ini bertujuan membangun sistem prediksi angka kelahiran di Kota Tanjung Balai dengan menerapkan algoritma K-Nearest Neighbor, serta mengevaluasi hasilnya menggunakan confusion matrix. Sistem ini diharapkan dapat mendukung pengambilan keputusan dalam pengendalian populasi secara lebih tepat dan efisien.

II. METODE

Penelitian ini menggunakan pendekatan kuantitatif dengan menerapkan algoritma K-Nearest Neighbor (KNN) regresi untuk memprediksi angka kelahiran per bulan di Kota Tanjung Balai. KNN bekerja dengan cara mencari sejumlah titik data tetangga terdekat dari data uji, kemudian menghitung rata-rata dari nilai targetnya sebagai prediksi. Evaluasi dilakukan menggunakan metrik Mean Absolute Error (MAE) untuk menilai akurasi model prediktif terhadap data actual [12]. Diagram alur penelitian ditunjukkan pada Gambar 1.



Gambar 1. Tahapan Algoritma

Uraian mendalam terkait masing-masing tahap diatas dijabarkan pada bagian berikut.

1) Input Dataset

Dataset yang digunakan dalam penelitian ini merupakan data historis angka kelahiran di Kota Tanjung Balai dengan cakupan periode Januari 2010 hingga Desember 2020. Data ini bersifat bulanan dan dikumpulkan dari enam kecamatan yang ada, sehingga setiap bulan terdiri dari 6 entri (satu untuk masing-masing kecamatan). Dengan cakupan 11 tahun \times 12 bulan \times 6 kecamatan, total jumlah data yang digunakan adalah 792 baris data. Distribusi angka kelahiran pada dataset bervariasi antara 40 hingga 250 kelahiran per bulan. Rata-rata angka kelahiran per bulan berada pada kisaran 110–130 kelahiran, tergantung kecamatan dan waktu. Informasi ini penting sebagai konteks evaluasi model prediksi, khususnya dalam interpretasi nilai MAE. Setiap baris data memuat nilai dari beberapa variabel, yaitu:

- Jumlah Penduduk: jumlah total penduduk per kecamatan per bulan,
- Pasangan usia subur: jumlah pasangan usia subur yang tercatat
- Tingkat Partisipasi KB: tingkat keikutsertaan dalam program keluarga berencana,
- Angka Kelahiran: jumlah bayi lahir pada bulan tersebut (sebagai target prediksi).

2) Preprocessing Data

Sebelum dilakukan pemodelan, data melalui tahap praproses agar siap digunakan [13]. Tahapan preprocessing meliputi:

- Menghapus data kosong (missing values)
- Normalisasi min-max untuk mengubah nilai-nilai dataset agar berada dalam rentang tertentu

$$x' = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (1)$$

3) Split Data (Pembagian Data)

Pembagian data pada penelitian ini dilakukan secara berdasarkan urutan waktu (chronological split), bukan secara acak. Hal ini sesuai dengan karakteristik data deret waktu (time series), di mana prediksi terhadap masa depan harus didasarkan pada data masa lalu secara berurutan. Dataset terdiri atas data bulanan dari Januari 2010 hingga Desember 2020, yang mencakup total 792 baris data (132 bulan \times 6 kecamatan). Pembagian data dilakukan dengan menggunakan periode:

- Data Latih (Training Set): Januari 2010 – Desember 2018
- Data Uji (Testing Set): Januari 2019 – Desember 2020

Dengan metode ini, diperoleh:

- 576 baris data untuk pelatihan model (9 tahun \times 12 bulan \times 6 kecamatan), dan

- 216 baris data untuk pengujian model (2 tahun \times 12 bulan \times 6 kecamatan)

Pendekatan ini dipilih untuk mensimulasikan skenario nyata prediksi, di mana sistem mempelajari pola dari data historis untuk memprediksi nilai pada periode waktu berikutnya secara berurutan. Pembagian berbasis waktu ini juga penting agar tidak terjadi data leakage dari masa depan ke masa lalu dalam proses pelatihan model.

4) Algoritma K-Nearest Neighbor (KNN)

a. Perhitungan Jarak (Euclidean Distance)

Untuk setiap data uji, hitung jarak ke semua data latih menggunakan:

$$d(x, x_1) = \sqrt{(x_1 - x_{i1})^2 + (x_2 - x_{i2})^2 + \dots + (x_n - x_{in})^2} \quad (2)$$

Dimana:

x data uji

x_n data latih ke- i

n jumlah fitur

b. Ambil K Tetangga Terdekat

Pilih K data latih dengan jarak terpendek

c. Prediksi Angka Kelahiran (Regresi)

Nilai prediksi diambil dari rata-rata angka kelahiran K tetangga tersebut:

$$\hat{y} = \frac{1}{K} \sum_{i=1}^K y_i \quad (3)$$

Dimana:

\hat{y} = Hasil prediksi angka kelahiran untuk data uji tertentu
 K = Jumlah tetangga terdekat (neighbors) yang dipilih dalam algoritma KNN

y_i = Nilai angka kelahiran aktual dari tetangga ke- i , yaitu dari data pelatihan (training data) yang memiliki jarak terdekat ke data uji.

5) Result Analysis & Visualization (Hasil Analisis dan Visualisasi)

Tahapan ini melibatkan:

- Menampilkan hasil prediksi dalam bentuk tabel dan grafik
- Melakukan evaluasi akurasi model dengan menggunakan:

Mean Absolute Error (MAE)

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (4)$$

Dimana:

y_i nilai aktual

\hat{y}_i nilai prediksi

6) Deployment

Model dan sistem prediksi diintegrasikan ke dalam aplikasi berbasis web menggunakan Laravel.

III. HASIL DAN PEMBAHASAN

A. Input Dataset

Dataset yang digunakan dalam penelitian ini merupakan data historis angka kelahiran yang diperoleh dari Badan Pusat Statistik di Kota Tanjung Balai. Data ini mencakup periode waktu bulanan dari tahun 2010-2020. Informasi yang dihimpun mencakup beberapa variabel penting yang diyakini memiliki pengaruh terhadap angka kelahiran, yaitu jumlah penduduk, jumlah pasangan usia subur, dan tingkat partisipasi dalam program Keluarga Berencana (KB). Variabel angka kelahiran sendiri dijadikan sebagai target prediksi dalam penelitian ini. Berikut dapat dilihat pada Tabel 1.

TABEL I
DATASET TAHUN 2010-2020

No	Bulan	Kecamatan	Jumlah Penduduk	Pasangan Usia Subur	Tingkat Partisipasi KB	Angka Kelahiran
1	2010-01	Datuk Bandar	2800	478	121	210
2	2010-01	Datuk Bandar Timur	2200	46	787	195
3	2010-01	Tanjung Balai Selatan	36200	260	87	155
4	2010-01	Tanjung Balai Utara	1400	51	61	94
5	2010-01	Sei Tualang Raso	1950	330	135	148
6	2010-01	Teluk Nibung	2300	78	166	145
...
787	2020-12	Datuk Bandar	27946	105	103	95
788	2020-12	Datuk Bandar Timur	30257	149	139	78
789	2020-12	Tanjung Balai Selatan	17689	57	89	88
790	2020-12	Tanjung Balai Utara	29000	89	136	85
791	2020-12	Sei Tualang Raso	16518	69	99	90
792	2020-12	Teluk Nibung	7879	78	73	40

B. Preprocessing Data

Tahap preprocessing dilakukan untuk menyiapkan data agar sesuai dengan kebutuhan model KNN [14]. Fitur numerik seperti jumlah_penduduk, pasangan_usia_subur, dan tingkat_partisipasi_kb dinormalisasi menggunakan metode Min-Max Scaling agar berada pada skala yang seragam, sehingga perhitungan jarak tidak bias. Target yang akan diprediksi adalah angka_kelahiran, sedangkan fitur input meliputi faktor-faktor yang diasumsikan memiliki pengaruh terhadap angka kelahiran.

TABEL II
HASIL NORMALISASI DATA

No	Bulan	Kecamatan	Jumlah Penduduk	Pasangan Usia Subur	Tingkat Partisipasi KB
1	2010-01	Datuk Bandar	0,036216447	0,116369846	0,079397673
2	2010-01	Datuk Bandar Timur	0,023434171	0,010978287	0,535249829
3	2010-01	Tanjung Balai Selatan	0,747763102	0,063186143	0,056125941

4	2010-01	Tanjung Balai Utara	0,006391138	0,012198097	0,038329911
5	2010-01	Sei Tualang Raso	0,018108223	0,080263479	0,088980151
6	2010-01	Teluk Nibung	0,02556455	0,01878507	0,110198494
...
787	2020-12	Datuk Bandar	0,571921602	0,025372042	0,067077344
788	2020-12	Datuk Bandar Timur	0,621154666	0,036106367	0,091718001
789	2020-12	Tanjung Balai Selatan	0,353408607	0,013661869	0,057494867
790	2020-12	Tanjung Balai Utara	0,594375799	0,021468651	0,089664613
791	2020-12	Sei Tualang Raso	0,328461866	0,016589412	0,064339493
792	2020-12	Teluk Nibung	0,144418406	0,01878507	0,046543463

C. Split Data

Setelah proses normalisasi dilakukan, data dibagi menjadi dua bagian, yaitu data latih (training) dan data uji (testing). Pembagian ini bertujuan untuk mengukur kemampuan model dalam melakukan prediksi terhadap data yang belum pernah dilihat sebelumnya [15]. Dalam penelitian ini, pembagian data dilakukan secara berdasarkan urutan waktu (chronological split) untuk mencerminkan karakteristik deret waktu yang digunakan. Data angka kelahiran bulanan dari Januari 2010 hingga Desember 2020 menghasilkan total 792 baris data (132 bulan \times 6 kecamatan). Dataset ini dibagi menjadi data latih (Januari 2010–Desember 2018) sebanyak 576 baris, dan data uji (Januari 2019–Desember 2020) sebanyak 216 baris. Pembagian berbasis waktu ini dipilih agar model hanya belajar dari data masa lalu dan menghindari data leakage, sehingga prediksi terhadap masa depan menjadi lebih realistis. Hasil pembagian ini selanjutnya digunakan untuk proses pelatihan, pengujian, dan evaluasi model prediktif pada tahap berikutnya.

TABEL III
PEMISAHAN DATA

Dataset	Persentase	Jumlah
Training	80%	576
Testing	20%	216
Total		792

D. Pemilihan Nilai K Terbaik

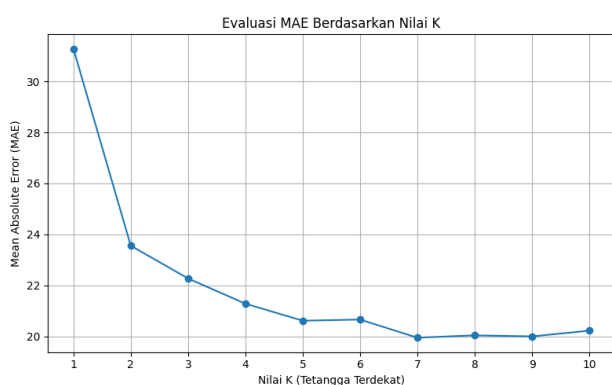
Nilai K optimal dalam penelitian ini ditentukan melalui pengujian terhadap berbagai nilai K dari 1 hingga 10, dengan hasil terbaik diperoleh pada $K = 7$ berdasarkan nilai Mean Absolute Error (MAE) terendah. Hal ini memperkuat alasan pemilihan KNN sebagai algoritma utama dalam membangun model prediksi angka kelahiran. Pemilihan nilai K merupakan tahap krusial dalam implementasi algoritma K-Nearest Neighbor (KNN) karena nilai K menentukan jumlah tetangga terdekat yang akan digunakan dalam proses prediksi. Nilai K yang terlalu kecil dapat menyebabkan model overfitting, sedangkan nilai K yang terlalu besar berpotensi menyebabkan underfitting. Untuk mendapatkan nilai K yang optimal, dilakukan pengujian terhadap berbagai nilai K dari 1 hingga

10 dengan menggunakan metrik evaluasi Mean Absolute Error (MAE). Hasil pengujian disajikan pada Tabel berikut.

TABEL IV
HASIL NORMALISASI DATA

Nilai K	MAE
1	31.28
2	23.55
3	22.27
4	21.28
5	20.61
6	20.66
7	19.94
8	20.04
9	19.99
10	20.22

Berdasarkan hasil pengujian pada Tabel 4, diketahui bahwa nilai MAE terendah diperoleh pada saat $K = 7$, yaitu sebesar 19.94. Artinya, pada saat menggunakan tujuh tetangga terdekat, model mampu memberikan prediksi dengan kesalahan rata-rata terkecil dibandingkan nilai K lainnya. Oleh karena itu, nilai $K = 7$ dipilih sebagai parameter terbaik dan digunakan dalam implementasi akhir model prediksi angka kelahiran.



Gambar 2. Grafik Nilai K Terbaik

Gambar 2 berikut menggambarkan tren nilai MAE terhadap perubahan nilai K . Terlihat bahwa seiring bertambahnya nilai K dari 1 hingga 7, nilai MAE cenderung menurun dan mulai naik kembali setelah $K > 7$, yang mengindikasikan bahwa nilai $K = 7$ merupakan titik optimal untuk dataset ini.

E. Perbandingan Model

Untuk mengukur keefektifan algoritma K-Nearest Neighbor (KNN) dalam memprediksi angka kelahiran, dilakukan eksperimen perbandingan dengan dua algoritma regresi lainnya, yaitu Linear Regression dan Decision Tree Regressor. Tujuan dari eksperimen ini adalah untuk mengevaluasi performa relatif KNN terhadap pendekatan regresi linier klasik dan metode berbasis pohon keputusan, serta untuk melihat sejauh mana KNN dapat memberikan

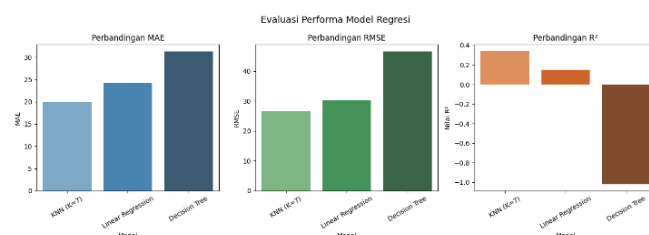
prediksi yang lebih akurat. Evaluasi dilakukan menggunakan metrik yang sama, yaitu Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), dan R-squared (R^2). Seluruh model dilatih menggunakan data yang telah dinormalisasi, dengan rasio pemisahan data training dan testing sebesar 80:20. Hasil evaluasi ditampilkan pada Tabel 5.

TABEL V
PERBANDINGAN MODEL

No	Model	MAE	RMSE	R^2
1	KNN ($K=7$)	19.94	26.54	0.34
2	Linear Regression	24.22	30.12	0.15
3	Decision Tree	31.26	46.51	-1.02

Dari hasil evaluasi pada Tabel 5, terlihat bahwa model KNN memberikan performa terbaik dengan nilai MAE dan RMSE yang paling rendah serta nilai R^2 tertinggi. Hal ini menunjukkan bahwa KNN lebih mampu menangkap pola hubungan antara fitur input dan angka kelahiran dibandingkan Linear Regression dan Decision Tree. Secara khusus, Linear Regression memiliki akurasi terendah dengan MAE tertinggi dan R^2 terendah, yang mengindikasikan bahwa hubungan antara variabel input dan target tidak sepenuhnya linear.

Visualisasi hasil evaluasi disajikan pada Gambar 3, yang menunjukkan performa masing-masing model berdasarkan ketiga metrik evaluasi. Dari grafik tersebut, keunggulan KNN dibandingkan dua model lainnya terlihat secara visual melalui tinggi batang yang lebih rendah pada MAE dan RMSE, serta batang lebih tinggi pada R^2 . Hasil eksperimen ini memperkuat keputusan untuk menggunakan KNN sebagai model utama dalam penelitian, karena mampu memberikan hasil prediksi yang lebih akurat pada dataset angka kelahiran bulanan di Kota Tanjung Balai.

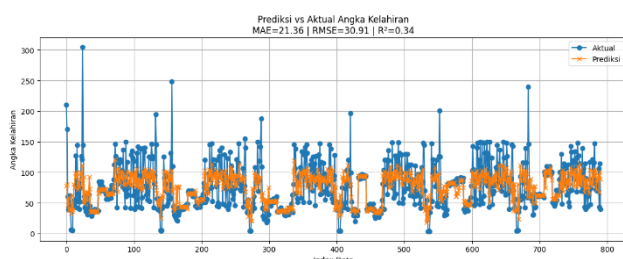


Gambar 3. Grafik Perbandingan Model

F. Evaluasi Model

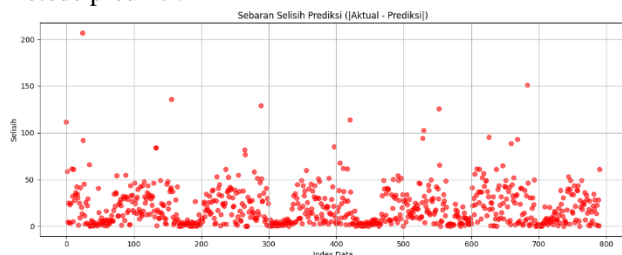
Evaluasi model dilakukan untuk mengetahui sejauh mana kemampuan algoritma K-Nearest Neighbor (KNN) dalam memprediksi angka kelahiran berdasarkan data yang tersedia [16]. Metode evaluasi yang digunakan dalam penelitian ini meliputi Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), dan koefisien determinasi (R-squared atau R^2). MAE mengukur rata-rata kesalahan absolut antara nilai aktual dan hasil prediksi, RMSE memberikan penalti lebih besar terhadap kesalahan yang ekstrem, sedangkan R^2

menunjukkan seberapa besar variasi target yang dapat dijelaskan oleh model. Pada model KNN dengan parameter $K = 7$, diperoleh hasil evaluasi MAE sebesar 21,36, RMSE sebesar 30,91, dan R^2 sebesar 0,34. Nilai MAE tersebut mengindikasikan bahwa secara rata-rata, prediksi angka kelahiran menyimpang sekitar 21 kelahiran dari nilai aktual setiap bulan. Jika dibandingkan dengan rata-rata angka kelahiran bulanan sebesar ± 180 kelahiran per kecamatan, maka tingkat kesalahan ini berada di kisaran 11%, yang masih dapat diterima untuk aplikasi praktis. Namun demikian, nilai R^2 yang relatif rendah menunjukkan bahwa model belum mampu menangkap keseluruhan variasi data secara optimal, terutama pada kasus lonjakan atau fluktuasi ekstrem [17].



Gambar 4. Perbandingan Prediksi dan Aktual Angka Kelahiran

Gambar 4 memperlihatkan perbandingan antara hasil prediksi dan data aktual untuk seluruh data dalam bentuk grafik garis. Kurva biru menunjukkan angka kelahiran aktual, sedangkan kurva oranye menunjukkan hasil prediksi oleh model KNN. Dari grafik ini, terlihat bahwa secara umum pola prediksi mengikuti tren data aktual, meskipun terdapat deviasi atau perbedaan nilai pada beberapa titik. Fluktuasi data aktual yang cukup tajam di beberapa indeks menyebabkan model tidak sepenuhnya mampu mengikuti perubahan ekstrem, terutama pada nilai-nilai outlier. Namun demikian, pada sebagian besar data, prediksi cenderung mendekati nilai aktual, yang memperkuat validitas penggunaan KNN sebagai metode prediksi.



Gambar 5. Grafik Selisih Prediksi terhadap Nilai Aktual

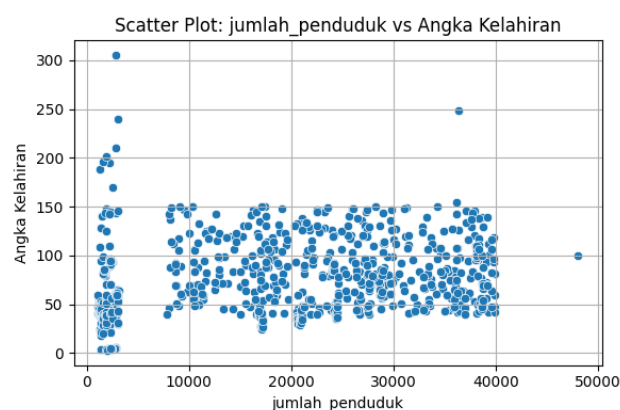
Selain grafik perbandingan langsung, Gambar 5 memperlihatkan distribusi selisih antara nilai aktual dan prediksi ($| \text{aktual} - \text{prediksi} |$) dalam bentuk grafik sebar (scatter plot). Grafik ini memberikan gambaran visual mengenai besar-kecilnya error (kesalahan prediksi) pada setiap data. Sebagian besar titik error berada pada rentang yang relatif rendah (< 50), yang menunjukkan bahwa mayoritas prediksi berada dalam kisaran yang dapat diterima. Titik-titik dengan nilai error tinggi mengindikasikan keberadaan data yang

mungkin mengandung outlier atau variabel yang tidak terobservasi oleh model.

Dengan menggunakan pendekatan ini, evaluasi model menjadi lebih komprehensif karena melibatkan nilai statistik (MAE) dan interpretasi visual dari hasil prediksi. Model KNN dalam penelitian ini dapat dikategorikan cukup baik dalam memodelkan data angka kelahiran, meskipun masih memiliki ruang untuk peningkatan terutama dalam menangani variasi ekstrem.

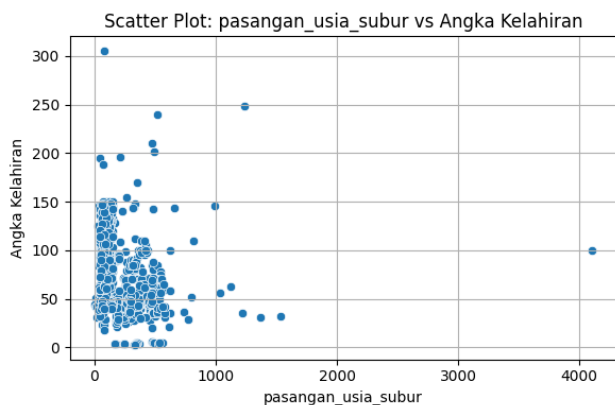
G. Hubungan Fitur Terhadap Target

Dalam tahap eksplorasi data, dilakukan analisis visual terhadap hubungan antara variabel-variabel independen (fitur) dengan variabel target, yaitu angka kelahiran. Tujuannya adalah untuk mengamati pola hubungan yang mungkin memengaruhi hasil prediksi pada model K-Nearest Neighbor (K-NN) [18]. Visualisasi dilakukan menggunakan scatter plot untuk masing-masing fitur terhadap target.



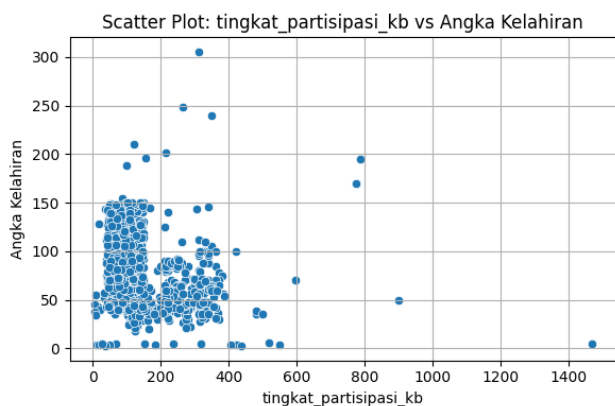
Gambar 6. Jumlah Penduduk vs Angka Kelahiran

Pada Gambar 6 ditampilkan scatter plot antara variabel jumlah_penduduk dan angka kelahiran. Terlihat bahwa meskipun terdapat penyebaran data yang luas, sebagian besar data terkonsentrasi pada kisaran jumlah penduduk antara 5.000 hingga 30.000 jiwa. Namun demikian, tidak ditemukan pola hubungan yang linier maupun non-linier yang signifikan antara jumlah penduduk dengan angka kelahiran. Hal ini mengindikasikan bahwa jumlah penduduk suatu wilayah tidak serta-merta menjadi penentu utama terhadap angka kelahiran, kemungkinan karena faktor lain seperti komposisi usia dan kepesertaan KB turut berperan.



Gambar 7. Pasangan Usia Subur vs Angka Kelahiran

Gambar 7 menunjukkan hubungan antara pasangan usia subur (PUS) dengan angka kelahiran. Secara umum terlihat bahwa semakin tinggi jumlah PUS, angka kelahiran cenderung meningkat, terutama pada rentang PUS di bawah 1.000 pasangan. Namun, seiring bertambahnya jumlah pasangan usia subur, penyebaran angka kelahiran justru meluas, menunjukkan adanya variabilitas yang tidak terkontrol oleh hanya jumlah PUS. Meski demikian, terdapat kecenderungan hubungan positif yang lemah antara variabel ini dan target, yang bisa menjadi pertimbangan penting dalam pemodelan.



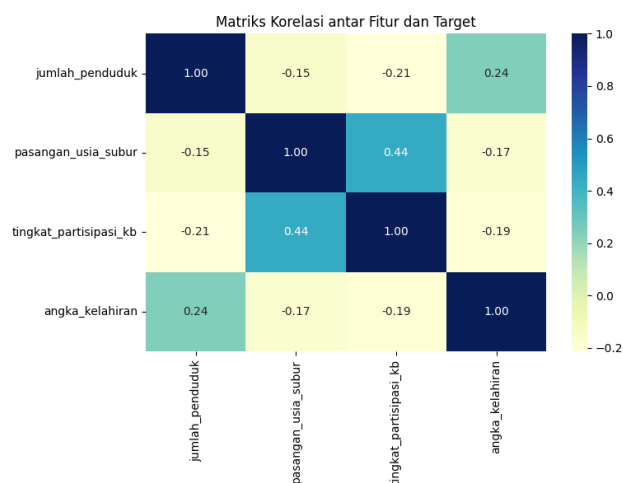
Gambar 8. Tingkat Partisipasi KB vs Angka Kelahiran

Gambar 8 terlihat adanya indikasi hubungan negatif antara tingkat partisipasi KB dengan angka kelahiran. Sebagian besar data terkonsentrasi pada partisipasi KB di bawah 500, dengan penurunan angka kelahiran seiring meningkatnya partisipasi. Korelasi ini mendukung asumsi bahwa program Keluarga Berencana (KB) memiliki peran signifikan dalam menekan angka kelahiran. Hal ini menjadikan tingkat partisipasi KB sebagai salah satu fitur yang potensial untuk meningkatkan performa model prediksi.

Berdasarkan analisis scatter plot, fitur pasangan usia subur dan tingkat partisipasi KB menunjukkan pola yang lebih relevan terhadap angka kelahiran dibandingkan jumlah penduduk.

H. Korelasi antar Fitur dan Target

Analisis korelasi dilakukan untuk mengevaluasi sejauh mana hubungan linier antara masing-masing fitur dengan variabel target, yaitu angka_kelahiran [19]. Pada penelitian ini digunakan koefisien korelasi Pearson, yang menghasilkan nilai antara -1 hingga 1. Nilai mendekati 1 menunjukkan korelasi positif kuat, sedangkan nilai mendekati -1 menunjukkan korelasi negatif kuat. Nilai mendekati 0 menunjukkan tidak adanya korelasi linier yang signifikan.



Gambar 9. Korelasi antar Fitur dan Target

Berdasarkan Gambar 9, diperoleh hasil korelasi sebagai berikut:

- 1) Jumlah Penduduk dan Angka Kelahiran Nilai korelasi sebesar 0.24 menunjukkan adanya hubungan positif yang sangat lemah antara jumlah_penduduk dengan angka_kelahiran. Hal ini mengindikasikan bahwa peningkatan jumlah penduduk memiliki pengaruh kecil terhadap angka kelahiran, dan pengaruhnya tidak cukup kuat untuk menjadi prediktor utama.
- 2) Pasangan Usia Subur dan Angka Kelahiran Koefisien korelasi sebesar -0.17 mengindikasikan hubungan negatif yang lemah. Secara teoritis, semakin banyak pasangan usia subur seharusnya meningkatkan potensi kelahiran. Namun, dalam data ini, korelasi negatif bisa mengindikasikan adanya faktor intervensi lain, seperti program KB atau perilaku reproduksi yang berbeda.
- 3) Tingkat Partisipasi KB dan Angka Kelahiran Terdapat korelasi negatif sebesar -0.19, yang meskipun lemah, konsisten dengan hipotesis bahwa peningkatan partisipasi program KB dapat menurunkan angka kelahiran. Hubungan ini menguatkan relevansi fitur tingkat_partisipasi_kb sebagai salah satu variabel penting dalam pemodelan.

Selain itu, ditemukan korelasi positif sebesar 0.44 antara pasangan_usia_subur dan tingkat_partisipasi_kb, yang

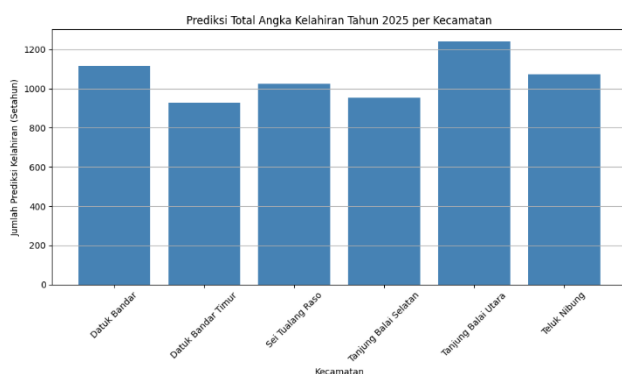
mengindikasikan bahwa wilayah dengan lebih banyak pasangan usia subur juga cenderung memiliki partisipasi KB yang lebih tinggi.

I. Hasil Prediksi

Setelah dilakukan pelatihan dan pengujian model menggunakan algoritma K-Nearest Neighbor (K-NN), dilakukan prediksi terhadap jumlah angka kelahiran pada tahun 2025 di enam kecamatan di Kota Tanjung Balai. Prediksi ini menggunakan nilai-nilai input tahun sebelumnya, seperti jumlah penduduk, pasangan usia subur, dan tingkat partisipasi KB, untuk memperkirakan total angka kelahiran tahunan di masing-masing wilayah. Gambar 8 menunjukkan visualisasi hasil prediksi dalam bentuk diagram batang, sedangkan rincian nilai prediksi disajikan pada Tabel 6 berikut:

TABEL VI
HASIL PREDIKSI TAHUN 2025

No.	Kecamatan	Prediksi (2025)
1	Datuk Bandar	1115
2	Datuk Bandar Timur	929
3	Sei Tualang Raso	1024
4	Tanjung Balai Selatan	953
5	Tanjung Balai Utara	1240
6	Teluk Nibung	1074

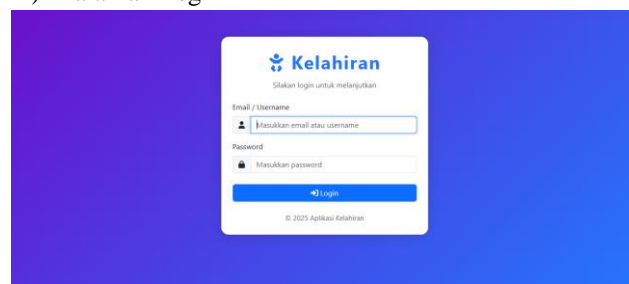


Gambar 10. Hasil Prediksi

Berdasarkan hasil prediksi tersebut, Kecamatan Tanjung Balai Utara diperkirakan memiliki angka kelahiran tertinggi pada tahun 2025, yaitu sebanyak 1.240 kelahiran, disusul oleh Kecamatan Datuk Bandar dengan 1.115 kelahiran, dan Teluk Nibung sebanyak 1.074 kelahiran. Sementara itu, angka kelahiran terendah diperkirakan terjadi di Kecamatan Datuk Bandar Timur, dengan jumlah 929 kelahiran. Perbedaan prediksi ini dapat mencerminkan variasi jumlah pasangan usia subur, kepadatan penduduk, serta efektivitas program KB yang diterapkan di masing-masing kecamatan. Prediksi ini memberikan gambaran yang penting bagi perencanaan sumber daya layanan kesehatan, program keluarga berencana, dan distribusi fasilitas publik di masa mendatang.

J. Deployment

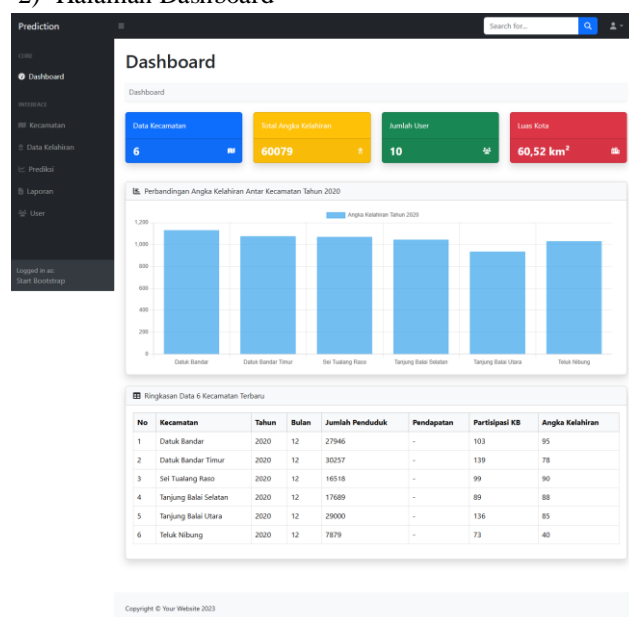
1) Halaman Login



Gambar 9. Halaman Login

Halaman pertama yang ditampilkan pada sistem adalah halaman login, di mana pengguna diminta untuk memasukkan username dan password yang telah terdaftar sebelumnya.

2) Halaman Dashboard



Gambar 10. Halaman Dashboard

Setelah berhasil melakukan proses login, pengguna akan langsung diarahkan ke halaman dashboard. Pada halaman ini ditampilkan grafik perbandingan angka kelahiran berdasarkan masing-masing kecamatan.

3) Halaman Data Kecamatan

No	Nama Kecamatan	Aksi
1	Datuk Bandar	[Edit] [Hapus]
2	Datuk Bandar Timur	[Edit] [Hapus]
3	Sei Tuang Raso	[Edit] [Hapus]
4	Tanjung Balai Selatan	[Edit] [Hapus]
5	Tanjung Balai Utara	[Edit] [Hapus]
6	Teluk Nibung	[Edit] [Hapus]

Gambar 11. Halaman Data Kecamatan

Pada halaman ini, pengguna diberikan akses untuk mengelola data kecamatan.

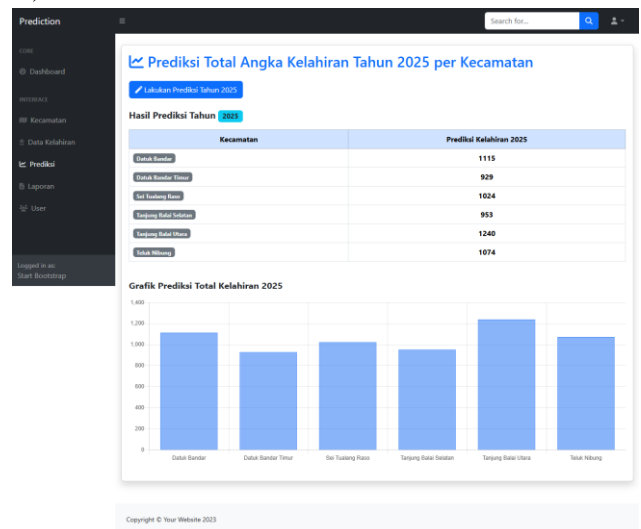
4) Halaman Data Kelahiran

No	Tahun	Bulan	Kecamatan	Jumlah Penduduk	Pasangan Usia Subur	Partisipasi KB	Angka Kelahiran	Aksi
1	2010	01	Datuk Bandar	2800	478	121	210	[Edit] [Hapus]
2	2010	02	Datuk Bandar	2500	350	773	170	[Edit] [Hapus]
3	2010	03	Datuk Bandar	2650	425	338	60	[Edit] [Hapus]
4	2010	04	Datuk Bandar	2700	460	235	39	[Edit] [Hapus]
5	2010	05	Datuk Bandar	2750	480	44	43	[Edit] [Hapus]
6	2010	06	Datuk Bandar	2900	510	44	62	[Edit] [Hapus]
7	2010	07	Datuk Bandar	3100	535	15	43	[Edit] [Hapus]
8	2010	08	Datuk Bandar	2850	470	517	6	[Edit] [Hapus]
9	2010	09	Datuk Bandar	2800	480	236	5	[Edit] [Hapus]
10	2010	10	Datuk Bandar	2750	490	316	5	[Edit] [Hapus]

Gambar 12. Halaman Data Kelahiran

Pada halaman ini, pengguna diberikan akses untuk mengelola data kelahiran, yang mencakup informasi mengenai jumlah penduduk, pasangan usia subur, tingkat partisipasi KB, serta angka kelahiran.

5) Halaman Prediksi



Gambar 12. Halaman Prediksi

Pada halaman ini, pengguna dapat melakukan prediksi angka kelahiran untuk satu tahun ke depan. Proses prediksi akan dimulai secara otomatis setelah pengguna menekan tombol "Lakukan Prediksi", dan hasilnya akan ditampilkan dalam bentuk grafik.

IV. KESIMPULAN

Berdasarkan hasil penelitian, algoritma regresi K-Nearest Neighbor (KNN) terbukti efektif dalam memprediksi angka kelahiran bulanan berdasarkan data jumlah penduduk, pasangan usia subur, dan tingkat partisipasi KB. Dataset yang digunakan mencakup data bulanan dari enam kecamatan selama periode 2010 hingga 2020, dengan pembagian data dilakukan secara kronologis untuk menjaga integritas deret waktu. Model dievaluasi menggunakan tiga metrik, yaitu Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), dan R-squared (R^2). Nilai K terbaik diperoleh pada $K = 7$, dengan hasil $MAE = 19,94$, $RMSE = 30,91$, dan $R^2 = 0,34$, yang menunjukkan tingkat akurasi prediksi yang cukup baik. Selain itu, performa KNN juga dibandingkan dengan algoritma Linear Regression dan Decision Tree, dan KNN menunjukkan hasil evaluasi yang lebih unggul. Model akhir diimplementasikan ke dalam aplikasi web berbasis Laravel, yang memungkinkan pengguna untuk mengelola data kependudukan serta menghasilkan prediksi angka kelahiran tahunan per kecamatan. Sistem ini diharapkan dapat menjadi alat bantu dalam mendukung kebijakan pemerintah di bidang pengendalian penduduk dan kesehatan masyarakat.

DAFTAR PUSTAKA

- [1] D. S. Seruni, M. T. Furqon, and R. C. Wihandika, "Sistem Prediksi Pertumbuhan Jumlah Penduduk Kota Malang menggunakan Metode K-Nearest Neighbor Regression," 2020. [Online]. Available: <http://j-ptiik.ub.ac.id>
- [2] Sensus Penduduk, "Berita Resmi Statistik Hasil Sensus Penduduk 2020."
- [3] Kementerian Kesehatan, "Rendahnya partisipasi pria dalam program keluarga berencana (KB)."
- [4] BPS, "Kota Tanjung Balai Dalam Angka 2023."
- [5] N. Wulandari, Y. Cahyana, and H. Hikmayanti Handayani, "Sentiment Analysis on the Relocation of the National Capital (IKN) on Social Media X Using Naive Bayes and K-Nearest Neighbor (KNN) Methods," 2025. [Online]. Available: <http://jurnal.polibatam.ac.id/index.php/JAIC>
- [6] S. Triase, "Implementasi Data Mining Dalam Mengklasifikasikan Ukt (Uang Kuliah Tunggal) Pada Uin Sumatera Utara Medan," *Jurnal Teknologi Informasi*, vol. 4, no. 2, 2020.
- [7] V. Syafana, S. S. Hilabi, E. Novalia, and B. Huda, "Prediksi Angka Kelahiran dalam Berbagai Kelompok Umur Ibu Menggunakan Metode K-Nearest Neighbor," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 3, pp. 1096–1103, Jul. 2024, doi: 10.57152/malcom.v4i3.1392.
- [8] Sutrimo and D. Wismarini, "Prediksi Proses Persalinan Menggunakan Algoritma Knn Berobot Pada Monitoring Elektronik Personal Health Record Ibu Hamil," 2022.
- [9] M. Qibtiyah and N. Cahyani, "Prediksi Tingkat Kelahiran Bayi di Kabupaten Bojonegoro dengan Menggunakan Algoritma Naive Bayes," 2024.
- [10] D. Puspita Sari, S. Shofia Hilabi, and Agustia Hananto, "Penerapan Data Mining Metode K-Nearest Neighbor Untuk Memprediksi Kelulusan Siswa Sekolah Menengah Pertama," *SMARTICS Journal*, vol. 9, no. 1, pp. 14–19, Mar. 2023, doi: 10.21067/smartics.v9i1.8088.
- [11] M. A. R. Habibi, S. S. Hilabi, B. Priyatna, and E. Novalia, "Penerapan Metode K-Nearest Neighbor Untuk Prediksi Jumlah Kasus HIV di Provinsi Jawa Barat," *JTIM: Jurnal Teknologi Informasi dan Multimedia*, vol. 7, no. 2, pp. 372–386, May 2025, doi: 10.35746/jtim.v7i2.721.
- [12] S. J. A. B. Bukit and R. K. R., "Prediksi Harga Tandan Buah Segar dengan Algoritma K-Nearest Neighbor," *Jurnal Sistem Komputer dan Informatika (JSON)*, vol. 5, no. 1, p. 92, Sep. 2023, doi: 10.30865/json.v5i1.6818.
- [13] A. Agung, A. Daniswara, I. Kadek, and D. Nuryana, "Data Preprocessing Pola Pada Penilaian Mahasiswa Program Profesi Guru," *Journal of Informatics and Computer Science*, vol. 05, 2023.
- [14] R. Hidayatullah, D. Abdul Fatah, A. Yasid, J. Raya Telang, K. Kamal, and J. Timur, "Penerapan Algoritma K-Nearest Neighbor Pada Minat Beli Mobil Bekas Menggunakan Pendekatan CRISP-DM," 2025.
- [15] M. Diki Hendriyanto and N. Sari, "Penerapan Algoritma K-Nearest Neighbor Penerapan Algoritma K-Nearest Neighbor dalam Klasifikasi Judul Berita Hoax," 2022.
- [16] Z. Hadiansyah, Z. Rozikin, and M. Fatchan, "Implementasi Algoritma K-Nearest Neighbor dalam Klasifikasi Penyakit Kanker Paru Paru," *Journal of Computer System and Informatics (JoSYC)*, vol. 6, no. 1, pp. 96–106, Nov. 2024, doi: 10.47065/josyc.v6i1.6195.
- [17] D. Septiani and A. Susanto, "Implementasi Decision Tree untuk Prediksi Kebutuhan Bahan Kain Pada Usaha Konveksi," 2025.
- [18] M. Radhi, D. Ryan Hamonangan Sitompul, S. Hamonangan Sinurat, and E. Indra, "Analisis Big Data Dengan Metode Exploratory Data Analysis (Eda) Dan Metode Visualisasi Menggunakan Jupyter Notebook," *Jurnal Sistem Informasi dan Ilmu Komputer Prima*, vol. 4, no. 2, 2021.
- [19] M. Husni Mubarak and F. Septian, "Prediksi GDP dengan RF dan XGBoost Berdasarkan Aspek Sosial, Ekonomi, dan Lingkungan," 2025.