BAB III
METODE PENELITIAN

2.1 Object Penelitian

Objek utama dalam penelitian ini adalah data kuantitatif berstruktur waktu yang memuat rekam jejak kondisi kesehatan balita serta determinan sosial-ekonomi di Provinsi Jawa Barat. Data tersebut berupa himpunan informasi historis yang digunakan sebagai dasar fundamental untuk melatih algoritma komputasi dalam memprediksi jumlah kasus stunting. Stunting sendiri merupakan variabel fenomena utama yang diteliti karena signifikansi dampaknya terhadap kualitas sumber daya manusia di masa depan.

Penelitian ini menggunakan beberapa set dataset yang saling terintegrasi, yang berkaitan langsung dengan insiden stunting beserta faktor-faktor demografis, infrastruktur, dan ekonomi di Jawa Barat. Data tersebut dikumpulkan dan disusun berdasarkan dua dimensi utama, yaitu dimensi waktu yang mencakup periode tahunan, dan dimensi spasial (*cross-sectional*) yang mencakup pembagian wilayah administratif pada tingkat 27 kabupaten dan kota. Secara spesifik, karakteristik data yang menjadi objek penelitian ini mencakup dua kelompok variabel.

Data tersebut dikumpulkan berdasarkan dimensi waktu (tahun) dan dimensi lokasi (*cross-sectional*) pada tingkat kabupaten kota. Data yang digunakan mencakup:

1. Variabel Dependen (Target Prediksi)

   Objek target dalam penelitian ini adalah dataset jumlah balita stunting per kabupaten dan kota. Karena tujuan utama dari tugas prediksi dalam penelitian ini adalah pemodelan regresi, maka variabel dependen ini diperlakukan sebagai nilai numerik kontinyu. Model *machine learning* akan mempelajari pola dari variabel independen untuk menghasilkan proyeksi atau tebakan angka aktual (jumlah jiwa) kasus stunting di masa depan.

2. Variabel Independen (Prediktor)

   Objek prediktor mencakup kumpulan data yang bertindak sebagai faktor-faktor determinan yang memengaruhi naik turunnya kasus stunting. Indikator gizi dan kesehatan dalam penelitian ini mencakup data jumlah balita gizi kurang serta cakupan bayi yang mendapatkan ASI eksklusif.

Selain itu, aspek infrastruktur dan lingkungan ditinjau melalui variabel ketersediaan jumlah puskesmas, persentase akses sanitasi yang layak, serta akses terhadap air minum yang layak bagi penduduk di setiap wilayah. Dalam dimensi ekonomi, variabel yang digunakan adalah data garis kemiskinan yang diukur dalam satuan Rupiah per kapita per bulan untuk masing-masing kabupaten dan kota. Terakhir, variabel tahun serta kode kabupaten dan kota dikategorikan sebagai indikator temporal dan spasial yang tidak digunakan sebagai nilai komputasi langsung, melainkan berfungsi sebagai *index keys* untuk melakukan *feature engineering*, khususnya dalam pembentukan *lag features* yang bersumber dari data riwayat tahun sebelumnya

Selain objek data fisik di atas, penelitian ini juga memiliki objek penelitian sekunder, yaitu pengembangan arsitektur model komputasi *machine learning*. Fokus analisis berada pada perbandingan kinerja regresi antara dua algoritma *ensemble learning*, yaitu *Random Forest* dan *Extreme Gradient Boosting* (XGBoost), serta analisis efektivitas optimasi *hyperparameter* dengan metode *Randomized Search*. Kinerja model dinilai berdasarkan kemampuannya memprediksi nilai numerik dengan tingkat kesalahan (*error*) yang seminimal mungkin.

2.2 Prosedur Penelitian

Sub-bab ini menguraikan tahapan, rencana, dan metodologi sistematis yang diimplementasikan dalam penelitian. Prosedur ini dirancang sedemikian rupa agar proses eksperimen dapat direplikasi secara akurat oleh peneliti lain, mencakup seluruh siklus hidup pengembangan model mulai dari akuisisi data hingga penyajian hasil akhir berupa aplikasi (*deployment*). Prosedur penelitian terdiri dari lima tahap utama yang komprehensif.

2.2.1 Rencana Kebutuhan Data

Bagian ini merinci spesifikasi teknis dan infrastruktur teknologi yang digunakan untuk mengolah data dan membangun model prediksi.

1. Data Kuantitatif

   Data yang digunakan adalah data berstruktur waktu yang mencakup aspek temporal (tahun) dan lokasi dari 27 Kabupaten atau Kota di Jawa Barat.

Sumber data diperoleh secara legal dan tervalidasi melalui situs web pemerintah provinsi, yaitu Open Data Jabar.

2. Perangkat Keras (*Hardware*)

   Penelitian dilakukan menggunakan satu unit laptop dengan spesifikasi prosesor *multi-core*, media penyimpanan (*Storage*), jaringan internet yang stabil, dan memori (RAM) minimal 8GB.

3. Perangkat Lunak (*Software*) dan *Environment*

   a. Sistem Operasi Windows 11

      Berfungsi sebagai perangkat lunak sistem utama (*core system software*) yang bertugas mengelola seluruh sumber daya perangkat keras (*hardware resources*) laptop, seperti alokasi memori (RAM) dan pemrosesan CPU.

   b. Aplikasi Browser

      Digunakan sebagai portal perantara untuk mengakses infrastruktur komputasi *cloud* dan portal data pemerintah.

   c. *Google Colaboratory*

      Digunakan sebagai *environment* pengembangan berbasis *cloud*. Alasan penggunaan Google Colab adalah kemampuannya menyediakan akses RAM virtual yang stabil secara gratis, serta telah terintegrasi dengan berbagai pustaka Python, sehingga proses pelatihan model *ensemble* yang kompleks tidak membebani memori dan prosesor fisik perangkat.

   d. Bahasa Pemrograman

      Python versi 3.10, dipilih karena memiliki dukungan pustaka (*library*) sains data dan pemodelan regresi yang sangat matang dan komprehensif.

   e. Arsitektur dan *Framework Deployment*

      Visual Studio Code digunakan sebagai *code editor* untuk merancang dan mengintegrasikan arsitektur sistem prediktif yang terdiri dari tiga komponen utama :

      a) Frontend (Antarmuka Web)

         Dibangun menggunakan *framework* Next.js untuk merancang antarmuka visual pengguna (*user interface*) dan visualisasi grafik interaktif.

b) Backend (API Prediksi)

Dibangun menggunakan *framework* Flask (Python) yang bertugas memuat berkas model *machine learning* (.pkl) dan menyediakan layanan REST API untuk memproses data input dari pengguna secara *server-side*.

4. Pustaka (*Library*) Utama Python yang Digunakan

Agar pemrosesan data, pemodelan, hingga visualisasi berjalan optimal, penelitian ini mengimpor beberapa modul spesifik, yaitu :

a. Pandas (pd) & Numpy (np)

Pustaka fundamental yang digunakan untuk memuat data *CSV*, manipulasi struktur *dataframe*, pembersihan data (*data cleaning*), penggabungan tabel (*merging*), serta komputasi matriks numerik.

b. Matplotlib (plt, ticker) & Seaborn (sns)

Digunakan untuk visualisasi *Exploratory Data Analysis* (EDA). Pustaka Seaborn digunakan untuk menghasilkan grafik *Heatmap Correlation*. Sedangkan matplotlib.pyplot digunakan untuk membuat grafik regresi komparatif, dibantu oleh modul matplotlib.ticker untuk memformat skala angka pada sumbu grafik (*axis formatting*) agar visualisasi lebih representatif.

c. SciPy (make_interp_spline)

Pustaka komputasi ilmiah yang digunakan secara spesifik untuk melakukan proses interpolasi *cubic spline*. Modul ini berfungsi untuk menghaluskan kurva garis pada grafik regresi perbandingan aktual dan prediksi, sehingga pergerakan fluktuasi terlihat lebih natural.

d. Scikit-Learn (sklearn)

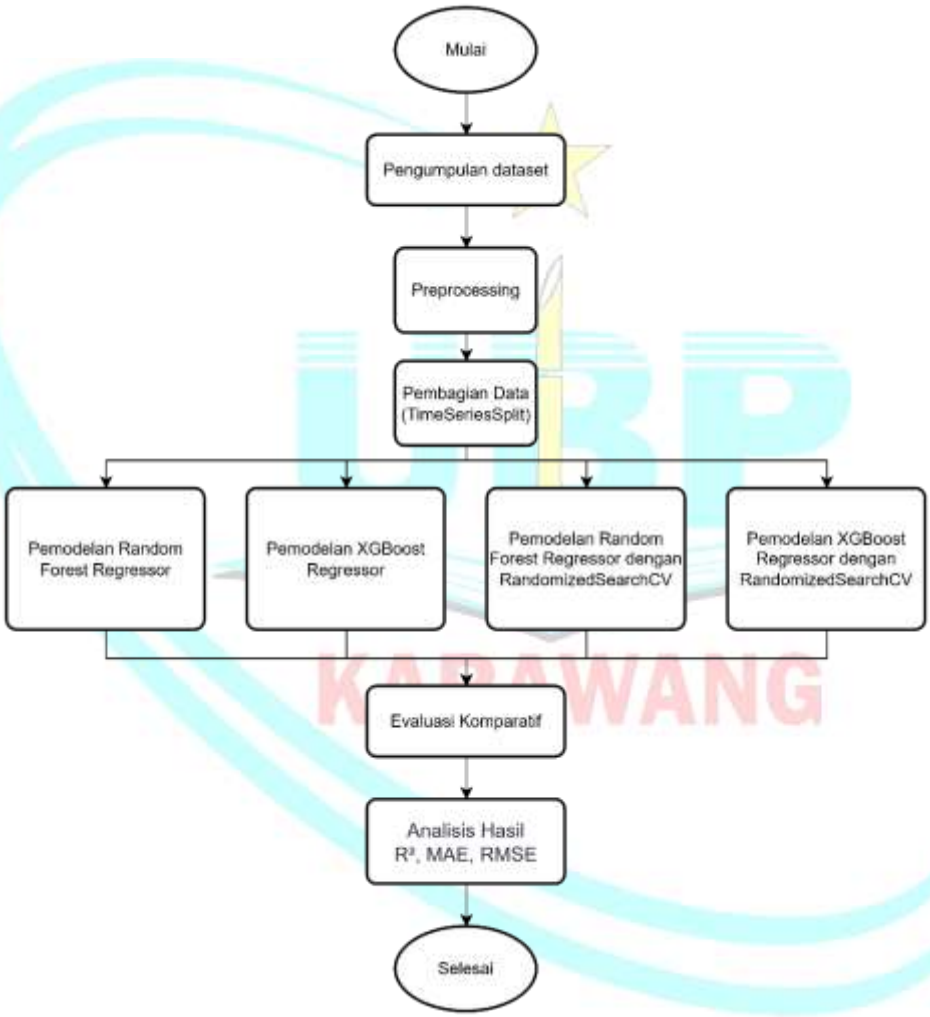
Pustaka *machine learning* komprehensif yang digunakan dalam berbagai tahap. Modul RandomForestRegressor digunakan untuk pemodelan berbasis *bagging*. Modul TimeSeriesSplit dan RandomizedSearchCV digunakan untuk proses validasi silang berstruktur waktu dan optimasi *hyperparameter*. Terakhir, modul *metrics* digunakan untuk memanggil fungsi evaluasi performa regresi,

yaitu *Mean Absolute Error* (MAE), *Mean Squared Error* (MSE atau RMSE), dan *R-squared* (R^2).

e. XGBoost (*XGBRegressor*)

Pustaka independen tingkat lanjut yang diimpor secara khusus untuk membangun algoritma prediksi berbasis *Extreme Gradient Boosting.*

2.2.2 Rancangan Penelitian



Gambar 3. 1 Rancangan Penelitian

Rancangan tahapan penelitian ini dibuat secara sistematis dan bertahap untuk memastikan pencapaian tujuan penelitian, yaitu membangun serta mengevaluasi kinerja model *ensemble learning* (*Random Forest* dan *XGBoost*) dalam memprediksi jumlah kasus stunting di Provinsi Jawa Barat. Secara visual, alur metodologi penelitian ini ditampilkan pada Gambar 3.1 di atas.

3.2.3 Pengumpulan dan Integrasi Data

Tahap ini bertujuan untuk mengumpulkan seluruh data kuantitatif secara komprehensif guna membentuk dataset panel.

1. Sumber dan Deskripsi Dataset

   Pengumpulan data dilakukan dengan memuat keseluruhan dataset berformat .csv ke dalam *environment* komputasi (*Google Collab*). Seluruh data merupakan data sekunder yang bersumber dari Open Data Jabar. Adapun rincian variabel yang digunakan dalam penelitian ini disajikan dalam Tabel 3.1, yang menjadi acuan utama dalam proses pemodelan dan analisis data pada tahapan selanjutnya.

Tabel 3. 1 Deskripsi Dataset

No	Nama Dataset	Peran	Deskripsi
1	Jumlah Balita Stunting	Target (Y)	Jumlah absolut balita yang mengalami stunting per wilayah.
2	Persentase Balita Stunting	Fitur	Proporsi kasus stunting dibandingkan populasi balita.
3	Status Kelahiran Bayi	Fitur	Jumlah kelahiran hidup/mati berdasarkan jenis kelamin.
4	Prevalensi Stunting	Fitur	Tingkat penyebaran stunting di kabupaten atau kota.
5	Akses Sanitasi Layak	Fitur	Persentase rumah tangga dengan fasilitas sanitasi memadai.
6	Akses Air Minum Layak	Fitur	Persentase rumah tangga dengan akses air bersih.
7	Garis Kemiskinan	Fitur	Standar pengeluaran minimum untuk kebutuhan dasar (Ekonomi).

No	Nama Dataset	Peran	Deskripsi
8	Penerima ASI Eksklusif	Fitur	Jumlah bayi <6 bulan yang mendapat ASI eksklusif.
9	Balita Gizi Kurang	Fitur	Jumlah balita dengan status gizi di bawah standar.
10	Jumlah Puskesmas	Fitur	Ketersediaan fasilitas kesehatan primer di wilayah tersebut.

2. Setelah kesepuluh dataset tersebut dimuat, dilakukan proses penggabungan data menggunakan pustaka Pandas. Penggabungan ini didasarkan pada kunci dimensi waktu (tahun) dan dimensi spasial (kode_kabupaten_kota). Metode outer join digunakan untuk dataset demografi dasar agar seluruh data dimensi waktu terjaga, sementara metode left merge digunakan untuk menyisipkan variabel ekonomi dan kesehatan infrastruktur sehingga menghasilkan satu dataset panel yang terstruktur.

3.2.4 Pra-Pemrosesan Data dan Rekayasa Fitur

Tujuan utama dari fase ini adalah mempersiapkan data mentah menjadi bentuk yang bebas *noise* dan dapat digunakan untuk melatih model *machine learning*.

1. Pemeriksaan Kualitas Data

   Mengidentifikasi dan menangani *missing values* (data kosong). Baris yang tidak memiliki label target (jumlah stunting) dihapus dari *dataframe*. Sedangkan untuk variabel prediktor yang kosong, dilakukan proses imputasi nilai rata-rata (*mean*) yang dihitung secara spesifik berdasarkan karakteristik historis per kabupaten dan kota.

2. Rekayasa Fitur (*Feature Engineering*)

   Tahapan ini sangat esensial untuk data berstruktur waktu guna meningkatkan sinyal prediksi (*predictive signal*). Diterapkan fitur temporal (*lag features*) menggunakan fungsi .shift(1) untuk menarik data

satu tahun sebelumnya (t-1) sebagai variabel *input* tahun berjalan. Selain itu, dibuat fitur turunan matematis berupa Fitur Interaksi (Garis Kemiskinan × Gizi Kurang) untuk menangkap hubungan faktor kesehatan-ekonomi, dan Fitur Rasio (Gizi Kurang dibagi Stunting). Pada fitur rasio, ditambahkan konstanta +1 pada nilai penyebut untuk mencegah kesalahan matematis berupa pembagian dengan angka nol (*division by zero*).

3. Transformasi Data

   Proses standarisasi dilakukan dengan menggunakan MinMaxScaler untuk menyamakan skala seluruh fitur numerik ke dalam rentang [0,1], sehingga tidak ada variabel prediktor yang terlalu mendominasi proses komputasi algoritma.

4. Pembagian Data (*Data Splitting*)

   Dataset dipartisi menjadi dua bagian utama dengan rasio komposisi 80% untuk Data Latih (*Training Set*) dan 20% untuk Data Uji (*Testing Set*). Mengingat karakteristik dataset ini adalah data panel berstruktur waktu, pembagian data mutlak dilakukan secara kronologis, bukan acak. 80% rekaman data historis dari periode tahun awal digunakan untuk melatih model, sementara 20% data dari periode tahun terbaru dikunci (*hold-out*) khusus untuk pengujian. Selama proses pelatihan di dalam *Training Set*, digunakan metode validasi silang berstruktur waktu, yaitu TimeSeriesSplit dengan konfigurasi n_splits=3. Teknik berlapis ini memastikan data pengujian internal selalu berasal dari masa depan relatif terhadap data pelatihan, sehingga kebocoran data (*data leakage*) dapat dihindari sepenuhnya.

3.2.5 Pemodelan dan Optimasi Hyperparameter

Pada tahap ini, dilakukan pembuatan model regresi dan penentuan parameter terbaik untuk mencapai performa komputasi yang optimal.

1. Menginisialisasi model pembelajaran *ensemble* (Random Forest Regression dan XGBoost) yang kemudian dilatih menggunakan pola dari data *Training Set*.

2. Proses optimasi parameter (*Hyperparameter Tuning*) dilakukan menggunakan fungsi RandomizedSearchCV dengan alokasi 30 iterasi eksperimen acak. Fungsi ini diintegrasikan dengan TimeSeriesSplit mempertahankan integritas urutan waktu selama proses pencarian parameter.
3. Pencarian parameter difokuskan pada variabel kunci arsitektur algoritma, seperti n_estimators (jumlah pohon), max_depth (kedalaman maksimal pohon), dan learning_rate (khusus pada XGBoost). Tujuan dari proses optimasi ini adalah menemukan kombinasi parameter yang secara komputasi memberikan tingkat kesalahan absolut paling rendah dibandingkan menggunakan metode pencarian konvensional seperti *Grid Search*.

3.2.6 Evaluasi dan Analisis Komparatif

Tahap pengujian bertujuan untuk mengukur tingkat kemampuan generalisasi model terhadap data yang belum pernah dilihat sebelumnya (*Testing Set*). Performa model *Random Forest* dan *XGBoost* dievaluasi secara komparatif menggunakan tiga metrik regresi statistik :

1. R-squared (R^2)

   Mengukur proporsi variasi pada variabel dependen (jumlah stunting) yang dapat dijelaskan oleh variabel independen. Semakin mendekati angka 1, semakin baik tingkat akurasi kecocokan model.
2. Mean Absolute Error (MAE)

   Menghitung rata-rata kesalahan absolut antara nilai prediksi dari model dengan nilai aktual di lapangan. Semakin kecil nilai MAE, semakin akurat model tersebut dalam memprediksi jumlah jiwa stunting.
3. Root Mean Squared Error (RMSE)

   Mengukur tingkat kesalahan prediksi dengan memberikan bobot penalti yang lebih besar pada kesalahan yang bernilai ekstrem atau melenceng jauh. Semakin kecil nilai RMSE, semakin andal model tersebut.

Model yang memiliki tingkat kesalahan (MAE dan RMSE) terendah serta tingkat akurasi kecocokan data (R^2) tertinggi pada pengujian akhir ditetapkan sebagai model prediksi stunting terbaik untuk penelitian ini.

3.2.6 Implementasi Sistem (*Deployment)*

Tahap akhir yang bersifat aplikatif ini bertujuan membawa model *machine learning* dari lingkungan komputasi eksperimental ke lingkungan produksi berupa sistem pendukung keputusan (*decision support system*) terintegrasi berbasis *web*. Sistem diimplementasikan menggunakan arsitektur *Client-Server* dengan alur kerja sebagai berikut :

1. Serialisasi Model

   Model prediktif regresi yang terbukti paling baik diekspor dan dikonversi menjadi berkas serialisasi (.pkl) menggunakan fungsi joblib agar dapat dieksekusi seketika tanpa perlu melewati proses pelatihan ulang.

2. *Backend* (API Prediksi)

   Dibangun menggunakan *framework* Flask (Python) sebagai lapisan pemroses. Lapisan *server* ini bertugas menerima masukan data interaktif dari pengguna, mengeksekusi otomatisasi rekayasa fitur (*feature engineering*) seperti komputasi rasio dan interaksi, lalu menjalankan berkas .pkl untuk menghasilkan proyeksi kasus stunting.

3. Frontend dan Database

   Antarmuka interaktif (*Client-Side*) dirancang menggunakan *framework* Next.js guna mengakomodasi masukan variabel secara individual per wilayah maupun pengolahan masif secara *batch* berbasis berkas *spreadsheet* (Excel atau CSV). Hasil proyeksi ditampilkan melalui komponen visual tabel dan metrik selisih (*trend badge*). Secara bersamaan, *backend* akan mendokumentasikan setiap hasil prediksi tersebut ke dalam basis data *cloud* Firebase Firestore secara *real-time* untuk menyimpan riwayat analisis secara permanen.